รายงานวิจัยฉบับสมบูรณ์

โครงการ การศึกษาความสัมพันธ์ระหว่างโครงสร้างและหน้าที่ของฮีโมโกลบินเพื่อเป็นแนวทางในการประยุกต์ใช้ในการรักษาโรค

(Exploring the Structure-Function Relationship of Hemoglobins for Therapeutic Applications)

โดย

รศ.ดร.ชนินทร์ นันทเสนามาตร์

มิถุนายน 2560

สัญญาเลขที่ RSA5780031

รายงานวิจัยฉบับสมบูรณ์

โครงการ การศึกษาความสัมพันธ์ระหว่างโครงสร้างและหน้าที่ของฮีโมโกลบิน เพื่อเป็นแนวทางในการประยุกต์ใช้ในการรักษาโรค
(Exploring the Structure-Function Relationship of Hemoglobins for Therapeutic Applications)

โดย
รศ.ดร.ชนินทร์ นันทเสนามาตร์
ศูนย์เหมืองข้อมูลและชีวการแพทย์สารสนเทศ
คณะเทคนิคการแพทย์ มหาวิทยาลัยมหิดล

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย
และ คณะเทคนิคการแพทย์ มหาวิทยาลัยมหิดล

(ความเห็นในรายงานนี้เป็นของผู้วิจัย สกว. ไม่จำเป็นต้องเห็นด้วยเสมอไป)

Acknowledgements

The accomplishments attained during the 3 year duration of this research project is made possible as a result of the concerted teamwork and contributions of colleagues and collaborators from the Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University; Pure and Applied Biochemistry, Lund University, Sweden; Pharmaceutical Biosciences, Uppsala University, Sweden; and Istituto di Ricerche Farmacologiche Mario Negri, Italy. Special thanks goes to Professor Dr. Leif Bülow from Pure and Applied Biochemistry, Lund University, Sweden for fruitful discussions on the fascinating world of hemoglobins and Professor Dr. Virapong Prachayasittikul for insightful comments and discussions. Acknowledgement also goes to the diligence and persistence of undergraduate and graduate students, research assistants and postdoctoral fellows. We are also grateful for academic licenses of softwares from OpenEyes and ChemAxon as well as several other open source softwares.

บทคัดย่อ

รหัสโครงการ : RSA5780031

ชื่อโครงการ : การศึกษาความสัมพันธ์ระหว่างโครงสร้างและหน้าที่ของฮีโมโกลบิน เพื่อเป็นแนวทางในการประยุกต์ใช้ในการรักษาโรค

ชื่อนักวิจัย : รศ.ดร.ชนินทร์ นันทเสนามาตร์

E-mail Address : chanin.nan@mahidol.edu

ระยะเวลาโครงการ : 16 มิถุนายน 2557 – 16 มิถุนายน 2560

การถ่ายเลือด (Blood transfusion) เป็นกระบวนการรับผลิตภัณฑ์ของเลือดเข้าสู่ระบบไหลเวียน เพื่อทดแทนส่วนประกอบของเลือดที่เสียไป แม้ว่าการถ่ายเลือดจะมีประโยชน์อย่างมากทางการแพทย์ แต่มีข้อจำกัด ที่เป็นสิ่งกีดขวางในการรักษา เช่น เลือดของผู้บริจาคไม่เพียงพอต่อความต้องการ มีอายุการเก็บรักษาที่สั้น ความยุ่งยากและความซับซ้อนในการเลือกเลือด หรือ การเกิดการกระตุ้นของระบบภูมิคุ้มกันของผู้ป่วยหลังจากการได้รับเลือด นอกจากนี้ยังมีข้อจำกัดในการขนส่ง และสถานที่เก็บเลือดไม่เพียงพอ จากปัญหาเหล่านี้นำไปสู่การพัฒนาและการสังเคราะห์เลือดเทียม โดยเฉพาะฮีโมโกลบิน (hemoglobin) ซึ่งเป็นโปรตีนส่วนประกอบสำคัญในเม็ดเลือดแดง ทำหน้าที่ในการขนส่งออกซิเจนไปสู่ส่วนต่างๆ ของร่างกาย แต่อย่างไรก็ตาม พบว่าฮีโมโกลบินเมื่ออยู่นอกเม็ดเลือดแดงจะก่อให้เกิดความเป็นพิษต่อร่างกาย ดังนั้น การศึกษาโครงสร้างและเข้าใจถึงกลไกการทำงานของฮีโมโกลบิน จึงมีความน่าสนใจ เพื่อนำมาประยุกต์ใช้การรักษาทางการแพทย์ในปัจจุบัน

โครงการนี้เป็นการศึกษาเกี่ยวกับความสัมพันธ์ของปัจจัยต่างๆ ที่ส่งผลต่อกลไกการทำงานและคุณสมบัติของฮีโมโกลบินที่เปลี่ยนแปลงไป ประกอบด้วย การศึกษาความสัมพันธ์ของฮีโมโกลบินและการออกฤทธิ์ทางชีวภาพ เพื่อเพิ่มประสิทธิภาพของการจับ-ปล่อยออกซิเจน **(ผลงานตีพิมพ์ที่ 1)** การศึกษาเกี่ยวกับสารที่สามารถกระตุ้นการเกิดการเปลี่ยนแปลงในโปรตีนสายเบต้าของฮีโมโกลบิน **(ผลงานตีพิมพ์ที่ 2)** การศึกษาฐานข้อมูลเกี่ยวกับชนิดของฮีโมโกลบินและความสามารถในการจับ-ปล่อยออกซิเจน (**ผลงานตีพิมพ์ที่ 3**), การศึกษาสารที่มีคุณสมบัติในการยับยั้งการเกิดเม็ดเลือดแดงรูปเคียว (**ผลงานตีพิมพ์ที่ 4**) (5) การศึกษาสารสีที่มีคุณสมบัติในการลดปริมาณ methemoglobin ในเม็ดเลือดแดง (**ผลงานตีพิมพ์ที่ 5**) และ (6) การศึกษาโครงสร้างและกลไกการทำงานของ อัลฟา-1-ไมโครโกลบูลิน (alpha-1-microglobulin) และการจับกับฮีม (heme) (**ผลงานตีพิมพ์ที่ 6**) โดย **ผลงานตีพิมพ์ที่ 1-5** เป็นศึกษาความสามารถของสารในระดับควอนตัม โดยใช้ความสัมพันธ์เชิงปริมาณระหว่างความสัมพันธ์ของโครงสร้างเคมีและการออกฤทธิ์ของสาร (quantitative structure–activity relationship, QSAR) เพื่อนำมาคัดกรองสารที่มีศักยภาพในการเกิดปฏิกิริยา **ผลงานตีพิมพ์ที่ 6** เป็นศึกษาการจับ

ของฮีม โดยวิธี tryptophan fluorescence quenching, UV-Vis spectrophotometry, circular dichroism spectrometer, surface plasmon resonance, electrophoretic migration shift, โครมาโตกราฟีแบบกรองผ่านเจล และ วิธีการจำลองแบบระดับโมเลกุล จากผลการทดลองพบว่า ตำแหน่งที่จับกับฮีมอยู่บริเวณ lipocalin และ ลูปที่1 และ 4 ของอัลฟา-1-ไมโครโกลบูลิน โดยปฏิกิริยามีเกี่ยวข้องกับโซ่ด้านข้างของกรดอะมิโนต่าง ๆ ประกอบด้วย cysteine ตำแหน่งที่ 34 lysine ตำแหน่งที่ 92, 118 และ 130 และ histidine ตำแหน่งที่ 123 ดังนั้น จากการศึกษาในครั้งนี้ คาดว่าสามารถให้ข้อมูลเชิงลึกเกี่ยวกับกลไกการทำงานและคุณสมบัติของฮีโมโกลบิน รวมทั้งความสัมพันธ์ของโครงสร้างสารทางเคมีและการทำงานของฮีโมโกลบิน นอกจากนี้ ความรู้ที่ได้จากการศึกษา สามารถนำไปวางแผนเพิ่มเติมเกี่ยวกับการสังเคราะห์สารโมเลกุลขนาดเล็กที่มีที่ฤทธิ์ หรือเกิดอันตรกิริยากับโปรตีนและเซลล์เป้าหมายอื่นๆ ตามที่ได้อธิบายไว้ในการศึกษาคุณสมบัติ และการทำงานของโปรตีนเรืองแสง (**ผลงานตีพิมพ์ที่ 7**), การศึกษาคุณสมบัติสารยับยั้ง P-glycoprotein (**ผลงานตีพิมพ์ที่ 8**), และสารยับยั้ง acetylcholinesterase (**ผลงานตีพิมพ์ที่ 9**), phosphodiesterase (**ผลงานตีพิมพ์ที่ 10**) ซึ่งความรู้และผลงานวิจัยได้สรุปเป็นบทความนิพนธ์ปริทรรศน์ 3 เรื่อง (**ผลงานตีพิมพ์ที่ 11-13**) และ หนังสือบทความทางวิชาการ 1 เรื่อง (**ผลงานตีพิมพ์ที่ 14**)

คำหลัก : ฮีโมโกลบิน, แบบจำลองเชิงโมเลกุล, ชีวสารสนเทศาตร์, เคมีสารสนเทศศาสตร์ และการทำเหมืองข้อมูล

Abstract

Project Code : RSA5780031

Project Title : Exploring the Structure-Function Relationship of Hemoglobins for Therapeutic Applications

Investigator : Associate Professor Dr. Chanin Nantasenamat

E-mail Address : chanin.nan@mahidol.edu

Project Period : 16 June 2014 – 16 June 2017

Blood transfusion is a life-saving medical procedure that provides donor blood to recipients who had loss blood from surgery or traumatic injury. Although immensely beneficial there are some inherent problems with donated blood that had burdened blood banking systems such as limited availability, short shelf-life, practical limitations in blood delivery, inadequate storage facilities, infectious risk and blood group incompatibilities. Such concerns are impeding factors that are limiting people from donating blood or receiving them from transfusion. The development of safe and effective artificial blood may resolve such issues. A common approach to the development of artificial blood is to use hemoglobin and make it work outside the red blood cells. Hemoglobin is in the reduced and non-toxic state when it is in the protective confinement of the red blood cell but becomes toxic and reactive upon intravascular hemolysis. Therefore, it is of great interest to seek out ways to remedy this shortcoming of hemoglobin. Structural studies of hemoglobin can resolve this issue and this is supported by the abundance of available X-ray crystallographic structures of hemoglobin. This study had explored the origins of (i) oxygen affinity modulatory properties allosteric effectors (**Publication 1**), (ii) splice switching activity of hemoglobin β-globin gene modulators (**Publication 2**), (iii) oxygen binding properties of a large set of hemoglobin variants (**Publication 3**), (iv) anti-sickling activity of several series of synthetic compounds (e.g. ethacrynic acid, bezyloxy acid, phenoxy acid, benzoic acid, proline salicylate, and alkanoic acid) (**Publication 4**), (v) methemoglobin reduction potential by electron mediators based on color dyes (**Publication 5**) as well as (vi) structural and biochemical characterization of two heme binding sites on α1-microglobulin (**Publication 6**). The first five studies employed quantum chemistry and machine learning to correlate structural features of the investigated compounds with their observed experimental biological properties via quantitative structure-activity relationship (QSAR) modeling. Molecular insights on the origins of these biological properties can be deduced from the obtained set of informative molecular descriptors that had been shown in all cases to be pertinent in differentiating potent compounds from their non-potent counterpart. In the sixth study, heme binding was investigated by tryptophan fluorescence quenching, UV–Vis spectrophotometry, circular dichroism, SPR, electrophoretic migration shift, gel filtration, catalase-like activity and molecular simulation. The results suggest that one heme-binding site is located in the lipocalin pocket and a second binding site between loops 1 and 4. Reactions with the hemes involve the side-groups of C34, K(92, 118, 130) and H123. It is expected that the knowledge gained from this study would help provide key mechanistic insights into the structure-function relationship of hemoglobin that is pertinent for its multi-faceted therapeutic applications. To the best of our knowledge, this research project explored for the first time the origins of several functional properties of hemoglobin that had for the most part been unexplored in spite of the abundant big data that had accumulated over the past decades. Looking ahead, it can be envisioned that the computational framework proposed in this project holds great utility for further exploration on the structure-function relationship of other functional properties of hemoglobin that have not yet been investigated herein. We are also planning on further exploring the experimental validation of computationally designed small molecule modulators against their respective target proteins and cells. Moreover, this computational framework had also been extended to other biological properties as demonstrated herein on studying and predicting functional properties of fluorescent proteins (**Publication 7**) as well as modulatory properties of P-glycoprotein inhibitors (**Publication 8**), acetylcholinesterase inhibitors (**Publication 9**) and phosphodiesterase (**Publication 10**). The accumulated best practice and knowledge had also been summarized in the form of 3 invited editorial/review articles (**Publications 11-13**) and 1 invited book chapter (**Publication 14**).

Keywords : hemoglobin; molecular modeling; bioinformatics; cheminformatics; data mining

Research Findings

Publication 1

Exploring the origins of structure-oxygen affinity relationship of human hemoglobin allosteric effector

1. Introduction

Several diseases (i.e. anemia, cancer, cardiovascular ailments, hemorrhages, ischemia and hemoglobinopathy H disease) are characterized by the insufficient supply of oxygen in peripheral tissues [1, 2, 3, 4, 5], a condition known as hypoxia. Therefore, the ability to increase the delivery of oxygen by red blood cells as to tackle the problem of hypoxia affords great therapeutic interest [6, 7].

It is generally accepted that a certain class of small-molecule known as allosteric effectors can modulate the oxygen binding property of hemoglobin (Hb) [8, 9, 10]. Particularly, these molecules bind preferentially to the larger central cavity of the T-state (when compared to the R-state) followed by stabilizing this conformation effectively shifting the R/T equilibrium and consequently lowering the oxygen affinity of Hb [11]. The endogenous allosteric effector of Hb, 2,3-bisphosphoglycerate (BPG), binds the allosteric cavity of Hb with a dissociation constant of 1.5×10^{-5} M. This is followed by a concomitant right-shift of the allosteric equilibrium that decreases its oxygen binding affinity as BPG stabilizes the deoxy T state via formation of intermolecular salt bridges between the two β-chains [11].

The search for novel and robust allosteric effectors has attracted much attention owing to its great therapeutic potential. Over the years, several structural classes of allosteric effectors of Hb had been discovered encompassing those based on the organic phosphates such as *myo*-inositol phosphates (i.e. myo-inositol hexakisphosphate, IHP; *myo*-inositol trispyrophosphate, ITPP) [12] as well as aromatic propionates such as the antilipidemic fibrate agents (i.e. clofibrate, CF; bezafibrate, BZF; BZF derivative, RSR-13; and BZF urea derivative, L35) [13]. The former class represents one of the first synthetic allosteric effectors to be studied in which IHP, a commonly available natural product, was found to bind 1,000 times more potently to the β-cleft of deoxygenated Hb [14]. Particularly, IHP displaces Hb-bound BPG consequently leading to lowered oxygen affinity resulting in increased and regulated release of oxygen to tissues [15]. The group of Nicolau and Lehn [16] introduced an IHP derivative, known as *myo*-inositol trispyrophosphate (ITPP), bearing 3 seven-membered cyclic pyrophosphate rings and the compound was shown to provide strong increase of oxygen release *in vitro* from both free Hb and red blood cells. Further studies demonstrated broad applicability of the compound for treating a wide range of disease for which tissues are in need of oxygen (i.e. ischemic heart disease, cardiovascular disease and tumor progression) [17, 18, 19]. The success of IHP and ITPP led the group of Nicolau and Lehn [12] to further extend their work towards the synthesis of *myo*-inositol derivatives encompassing inositol tetraphosphates (ITPs) and inositol bispyrophosphates (IBPPs) as allosteric effectors of human Hb.

Quantitative structure-activity/property relationship (QSAR/QSPR) is a computational methodology for discerning the inherent linear or non-linear relationship between a set of structural features of investigated molecules with their respective biological activity /chemical property [20, 21]. QSAR/QSPR had been successfully utilized to address a wide range of problems of biological [22, 23, 24, 25, 26, 27] and chemical [28, 29, 30, 31] importance. The essential steps in the construction of QSAR/QSPR models entail the following procedures: (i) compilation of data set of interest, (ii) optimize geometrical structures of molecules (if computing 3D features) (iii) compute molecular descriptors, (iv) select a subset of molecular descriptors either through chemical intuition or computational optimization, (v) divide data set into internal and external set and (vi) develop a predictive model using the internal set and evaluate predictivity on the external set.

Preliminary QSAR study by Hansch *et al.* [32] had touched upon modeling the allosteric interaction of alkylisonitriles (RN=C) with Hb by focusing on the hydrophobic properties of ligands responsible for such interactions. Therefore, there has not yet been any reported in-depth QSAR investigation for studying allosteric effectors of human hemoglobin for further utilization in remedying hypoxia-related diseases. To the best of our knowledge, this study represents the first QSAR model focused on unraveling the origins of allosteric effector activity on human Hb. This was achieved by compiling a data set of 27 *myo*-inositol derivatives based on tetrakisphosphates and bispyrophosphates reported by the group of Nicolau and Lehn [12]. Physicochemical features of investigated compounds, after feature selection, were described by a set of 5 molecular descriptors. QSAR models developed by several multivariate methods afforded good predictive performance as verified by internal and external validation.

2. Material and Methods

2.1. Data Set

A data set of 27 *myo*-inositol derivatives and their allosteric effector activity against human Hb were obtained from Koumbis *et al.* [12]. The allosteric activity was represented by the P_{50} value, which is a conventional measure of hemoglobin affinity for oxygen. An increase in the P_{50} value indicates a rightward shift of the standard curve thereby suggesting that a larger partial pressure is necessary to maintain 50% oxygen saturation thereby suggesting a decreased affinity. Conversely, a lower P_{50} leads to a leftward shift corresponding to higher oxygen affinity. As to achieve uniform distribution of data samples, the P_{50} values were subjected to data transformation by calculating its logarithmic values to the base of 10.

2.2. Geometry Optimization and Descriptor Calculation

Chemical structures of investigated compounds were drawn using ChemAxon Marvin version 6.2.1 [33] and their molecular geometries were optimized at the density functional theory (DFT) level using Becke's three-parameter Lee-Yang-Parr hybrid functional (B3LYP) in concomitant with the 6-311++G(d,p) basis set. Quantum chemical calculation was performed using Gaussian 09 [34] to derive a set of quantum chemical descriptors, which was obtained from the low energy conformer, comprising of the total energy of the molecule, highest occupied molecular orbital energy (HOMO), lowest unoccupied molecular orbital energy (LUMO), dipole moment (μ), electron affinity (EA), ionization potential (IP), energy difference of HOMO and LUMO states (HOMO-LUMO), Mulliken electronegativity (χ), Hardness (η), Softness (S), Electrophilicity (ω), Electrophilic index (ω_i), most negative atom in the molecule (Q_{neg}), most positive atom in the molecule (Q_{pos}) and the mean absolute atomic charge (Q_m).

Low energy conformers were subjected for further generation of an additional set of 3,224 molecular descriptors using DRAGON version 5.5 [35]. This descriptor set spanned 22 categories comprises of 48 constitutional descriptors, 119 topological descriptors, 47 walk and path counts, 33 connectivity indices, 47 information indices, 96 2D autocorrelation, 107 edge adjacency indices, 64 Burden eigenvalues, 21 topological charge indices, 44 eigenvalue-based indices, 41 randic molecular profiles, 74 geometrical descriptors, 150 RDF descriptors, 160 3D-MoRSE descriptors, 99 WHIM descriptors, 197 GETAWAY descriptors, 154 functional group counts, 120 atom-centered fragments, 14 charge descriptors, 29 molecular properties, 780 2D binary fingerprints and 780 2D frequency fingerprints.

2.3. Feature selection

Descriptors having constant value and variable pairs with correlation coefficient greater than 0.9 were subjected to removal using the Unsupervised Forward Selection algorithm [36]. Additional round of feature selection was performed using stepwise linear regression as calculated by SPSS Statistics 18.0 [37]. This led to the selection of important descriptors that will be subsequently used in correlating with hemoglobin allosteric activity of *myo*-inositol derivatives.

2.4 Data splitting

In order to obtain accurate and generalized QSAR models, the 27 compounds were divided into two parts comprising of internal set for constructing QSAR models using the leave-one-out cross-validation (LOO-CV) approach while the remaining subset of 7 compounds (i.e., **12d**, **17c**, **22, 24d**, **26a**, **26c**, **and 24b**) were used as the external set, which was sampled according to the Kennard-Stone algorithm [38] (Table 1).

2.5. Multivariate analysis

Physicochemical features of investigated *myo*-inositol derivatives as represented by selected molecular descriptors were correlated with their respective hemoglobin allosteric activity using multiple linear regression (MLR), artificial neural network (ANN) and support vector machine (SVM).

MLR is a classical multivariate approach that linearly correlates a set of independent variables (i.e. molecular descriptors) with the dependent variable of interest (i.e. allosteric activity).

ANN was performed using the back-propagation algorithm in which the residual error from prediction are propagated in a backward fashion from the output layer through the hidden layer and finally onto the input layer followed by readjustment of weights interconnecting the neurons. This process is carried out iteratively until convergence is reached.

SVM is a statistical learning approach proposed by Vapnik [39, 40] that utilizes kernel functions to transform data to higher dimension whereby SVM performs its learning and decision in a linear manner. This study employed the radial basis function (RBF) kernel for such data transformation.

All multivariate analysis were performed using Weka, version 3.4.2 [41]. Optimal parameters were determined empirically. Particularly, the number of nodes in the hidden layer, number of learning epochs, learning rate and momentum were subjected to optimization for ANN calculations whereas C and γ parameters were optimized for SVM classifier.

Y-scrambling is a widely used approach in order to ensure the robustness of the QSAR model [42]. This was performed by randomly shuffling the Y-variable while keeping the X-variable intact followed by computing the Y-scrambled model again. It is expected that the resulting models should have low R^2_{Tr} and Q^2_{CV}. Y-scrambling was performed for 100 runs using the R program [43]. Furthermore, the statistical metric r^2_m as proposed by Roy *et al.* [44] was also utilized to evaluate the robustness of QSAR models in which favorable models should afford $r^2_m > 0.5$ and that Δr^2_m should be < 0.2.

Principal component analysis (PCA) was performed using FactoMineR in R [45] as to visualize the chemical space of investigated compounds.

3. Results and Discussion

3.1. Chemical space of myo-inositol derivatives

Myo-inositol derivatives employed in this study are comprised of ITPs and IBPPs in which the former class constitutes substituted and unsubstituted inositol(1,3,4,6)P_4 (**12a-f**) and (±)-inositol(2,3,5,6)P4 (**17a-f**) along with (±)-inositol(1,2,3,4)P_4 (**22**) while the latter class constitutes inositol(1,6:3,4)BPP (**24a-f**) and (±)-inositol(2,3:5,6)BPP (**26a-e**). Chemical structures of investigated compounds are shown in Figure 1. These compounds were evaluated by Koumbis *et al.* [12] for their ability to shift the oxygen saturation curve to higher pO_2 values and subsequently assess the partial pressure of oxygen for half-saturation (P_{50}) of Hb when bound to the allosteric effectors. This study investigates the origins of allosteric effector properties by means of predictive QSAR modeling as summarized in Figure 2.

Visual representation of the overall distribution of data values for P_{50} and Lipinski's descriptors is shown as 3D bar plots in Figure 3. Generally, the affinity of compounds toward binding Hb as deduced from P_{50} values revealed the following trend for the four class of compounds: **12** > **17** > **24** > **26**. It can be seen from Figure 3A that ITPs (i.e. **12** and **17**) provided higher affinity than their IBPPs (i.e. **24** and **26**) counterpart.

Lipinski's descriptors comprising of molecular weight (MW), molecular lipophilicity (ALogP), number of hydrogen bond donors (nHDon) and the number of hydrogen bond acceptors (nHAcc) (Table 1) were analyzed in order to understand the general properties of these class of compounds. A notable characteristic distinguishing ITPs (**12** and **17**) from IBPPs (**24** and **26**) is the presence of higher number of hydrogen bond donors (Figure 3D) and acceptors (Figure 3E) in the former class. As hydrogen bond and electrostatic interaction are key binding mechanism for natural (2,3-BPG) [46] and synthetic [12] allosteric effectors thus ITPs were more effective than IBPPs plausibly due to the fact that these set of compounds are highly capable of forming hydrogen bonds than their IBBPs counterpart. Moreover, ITPs **12c** and **17c** followed by **12e** and **17e** were the most potent amongst the investigated compounds. These compounds also have higher ALogP (Figure 3B) and MW (Figure 3C) than the rest of the compounds thereby implying that these properties may influence allosteric activity of these sets of effectors.

Feature selection using UFS followed by stepwise linear regression identified five important descriptors consisting of nHDon, nHAcc, ALogP, topological polar surface area (TPSA) and Q_{pos} (Table 1). Interestingly, aside from TPSA and Q_{pos} the remaining three (i.e. nHDon, nHAcc and ALogP) of the selected descriptors were Lipinski's descriptors. TPSA refers to the surface area of oxygen, nitrogen, sulfur and attached hydrogen atoms while Q_{pos} pertained to the most positive atomic charge. Thus, the selected descriptors implied the importance of hydrogen bonds and electrostatic properties for the activity of allosteric effectors.

3.2. QSAR model of myo-inositol derivatives using MLR

In this study, the set of four important descriptors (Table 1) as selected by feature selection were used as independent variables while allosteric effector activity (i.e., $\log P_{50}$) was used as the dependent variable. The data set was separated into internal and external sets as to assess their internal and external predictive performances, respectively. Figure 4 displays the PCA plot of data points in the internal (red) and external (blue) sets in which the first two components explains 77.89% of the variance afforded by the 27 compounds.

MLR model was constructed using ridge parameter of R = 1.0E-8. The predictive performance of the constructed MLR models is shown in Table 2 and the MLR equation is shown below along with their respective statistical properties.

$$\log P_{50} = 0.0517(\text{nHDon}) + 0.0023(\text{TPSA}) + 0.0081(\text{ALogP}) + 1.3373(Q_{\text{pos}}) - 0.1697 \qquad (1)$$

n=20, R^2_{Tr} = 0.8859, Q^2_{CV} = 0.6306, Q^2_{Ext} = 0.8332, RMSE_{Tr} = 0.0775, RMES_{CV} = 0.1487, RMSE_{Ext} = 0.1015

As deduced from regression coefficient, the most important molecular descriptors were Q_{pos} > nHDon > ALogP > TPSA, which displayed corresponding values of 1.3373, 0.0517, 0.0081 and 0.0023, respectively. The MLR equation suggested that high potent allosteric effector in order to release oxygen from hemoglobin should have higher positively atomic charge, lipophilicity number of hydrogen bond donor and topological polar surface area. Plot of experimental versus predicted activities for investigated compounds is shown in Figure 5A.

3.3. QSAR model of myo-inositol derivatives using ANN and SVM

In addition to MLR which is linear machine learning approach, ANN and SVM which are nonlinear approach, was use for constructed more sophisticated QSAR model using the same dataset. Parameter optimization of ANN identified appropriate hidden node of 1, learning epoch of 1000 cycles, learning rate of 0.1 and momentum of 0.4 (Figures 6A-C). It can be observed from Table 2 that ANN model provided good predictive performance with the following parameters: R^2_{Tr} = 0.9235, Q^2_{CV} = 0.7484, Q^2_{Ext} = 0.8847, RMSE_{Tr} = 0.0642, RMES_{CV} = 0.1176 and RMSE_{Ext} = 0.0827. SVM model building was initiated by searching for optimal C and γ parameters. The search comprised of two-level including initial global grid search followed by a refined local grid search. Global grid search were identified optimal C and γ value as 2^3 and 2^{-1} respectively, while the refined local grid search identified optimal C and γ value of $2^{4.4}$ and $2^{-1.4}$, respectively (Figure 7 A-B). Furthermore, Table 2 shows the performance of SVM, which provided the following statistical parameters: R^2_{Tr} = 0.9876, Q^2_{CV} = 0.8722, Q^2_{Ext} = 0.9694, RMSE_{Tr} = 0.0298, RMES_{CV} = 0.0827 and RMSE_{Ext} = 0.0465. Plot of experimental versus predicted activities of the investigated compounds as predicted by ANN and SVM is shown in Figures 5B and 5C respectively.

As a result, it can be seen that all multivariate methods provided good performance in predicting the $\log P_{50}$ values of the investigated ITPs and IBBPs as hemoglobin allosteric effector. This is verified by both internal and external validations of the predictive QSAR models. Eriksson and Johansson [47] proposed a metric based on $R^2 - Q^2$ for describing the fraction of Y-data explained by accumulated chance correlations in which values greater than 0.2-0.3 suggests the risk of chance correlations, the presence of outliers or irrelevant descriptors in the data set, or the possibility of overfitting model. It can be seen that all multivariate methods afforded $R^2 - Q^2$ well below the aforementioned threshold where MLR, ANN and SVM had values of 0.2553, 0.1751 and 0.1155, respectively.

To further verify the validity of QSAR models, Y-scrambling experiments were performed in concomitant with the utilization of the r^2_m metric. It can be seen in Figure 8 that QSAR models developed using SVM afforded distinct difference in the distribution of data points of Y-scrambled models from the actual one. It is also observed that several Y-scrambled models afforded significantly large difference of $R^2 - Q^2$ suggesting its expected inadequacy in properly modeling the activity. Furthermore, it can also be observed that a few data points displayed higher Q^2 than their respective R^2. A closer analysis revealed that this model afforded the typically higher RMSE value for the CV set with respect to the training set while giving rise to highly negative Q values, which would in turn produce seemingly high Q^2 value, thereby suggesting that the scrambled models could not model the activity. Moreover, r^2_m values for MLR, ANN, and SVM were found to be 0.50, 0.68, and 0.84, respectively, whereas corresponding values for Δr^2_m were 0.01, 0.06, and 0.05, respectively. These results suggested that SVM and ANN models provided good predictive performance.

However, these two models are black-box models that are not readily interpretable. In spite of its lower predictive performance, the MLR model afforded essential insights on significant descriptors giving rise to the allosteric effector properties. Errors arising from these predictive models could be attributed to several factors including inherent experimental error as well as errors arising from the modeling process (i.e. limitation of learning method, sub-optimal descriptors from feature selection, sub-optimal learning parameters, applicability domain, etc.). In light of these potential factors, this study tries to address all of the aforementioned points in development of the QSAR model such as employing PCA analysis to visualize the applicability domain of internal and external sets, utilizing feature selection to select the most significant descriptors, performing parameter optimization as to obtain the best performing model as well as employing several machine learning methods.

Practical utilization of the QSAR models developed herein could be implemented by calculating molecular descriptors for new, unknown compounds followed by predicting their biological activity using the multivariate analysis methods. It should also be noted that to ensure reliable predictions, the applicability domain should also be determined by evaluating the molecular similarity (i.e. Tanimoto coefficient) of new compounds in relation to the constituting compounds from the model proposed herein. Furthermore, as more data becomes available, such information could potentially be used to update the QSAR model.

4. Conclusion

In summary, this study developed QSAR models for investigating the allosteric effector property of ITPs and IBBPs against human hemoglobin. Feature selection using UFS and stepwise linear regression method identified 4 important descriptors were selected from a total of 3,239 descriptors. Selected descriptors indicated the importance of hydrogen bonding capacity, lipohilicity and electrostatic properties, which were in agreement with the chemical space analysis performed herein. The QSAR approach presented herein could provide useful information on origins of allosteric effector properties that could further be used to guide the design of novel hemoglobin allosteric effectors.

References

1. Sun K, Xia Y. New insights into sickle cell disease: a disease of hypoxia. Curr Opin Hematol. 2013;20:215-21.
2. Rundqvist H, Johnson RS. Tumour oxygenation: implications for breast cancer prognosis. J Intern Med. 2013;274:105-12.
3. Lenihan CR, Taylor CT. The impact of hypoxia on cell death pathways. Biochem Soc Trans. 2013;41:657-63.
4. Voelkel NF, Mizuno S, Bogaard HJ. The role of hypoxia in pulmonary vascular diseases: a perspective. Am J Physiol Lung Cell Mol Physiol. 2013;304:L457-65.
5. Papassotiriou I, Kister J, Griffon N, Abraham DJ, Kanavakis E, Traeger-Synodinos J, Stamoulakatou A, Marden MC, Poyart C. Synthesized allosteric effectors of the hemoglobin molecule: a possible mechanism for improved erythrocyte oxygen release capability in hemoglobinopathy H disease. Exp Hematol. 1998;26:922-6.
6. Mairbaurl H, Weber RE. Oxygen transport by hemoglobin. Compr Physiol. 2012;2:1463-89.
7. Crawford JH, Chacko BK, Kevil CG, Patel RP. The red blood cell and vascular function in health and disease. Antioxid Redox Signal. 2004;6:992-9.
8. Lalezari I, Lalezari P, Poyart C, Marden M, Kister J, Bohn B, Fermi G, Perutz MF. New effectors of human hemoglobin: structure and function. Biochemistry. 1990;29:1515-23.
9. Perutz MF, Fermi G, Luisi B, Shaanan B, Liddington RC. Stereochemistry of cooperative mechanisms in hemoglobin. Acc Chem Res. 1987;20:309-21.
10. Safo MK, Bruno S. Allosteric Effectors of Hemoglobin: Past, Present and Future. Chemistry and Biochemistry of Oxygen Therapeutics: From Transfusion to Artificial Blood: John Wiley & Sons, Ltd; 2011. p. 285-300.
11. Arnone A. X-ray diffraction study of binding of 2,3-diphosphoglycerate to human deoxyhaemoglobin. Nature. 1972;237:146-9.
12. Koumbis AE, Duarte CD, Nicolau C, Lehn JM. Tetrakisphosphates and bispyrophosphates of myo-inositol derivatives as allosteric effectors of human hemoglobin: Synthesis, molecular recognition, and oxygen release. ChemMedChem. 2011;6:169-80.
13. Randad RS, Mahran MA, Mehanna AS, Abraham DJ. Allosteric modifiers of hemoglobin. 1. Design, synthesis, testing, and structure-allosteric activity relationship of novel hemoglobin oxygen affinity decreasing agents. J Med Chem. 1991;34:752-7.
14. Yonetani T, Park SI, Tsuneshige A, Imai K, Kanaori K. Global allostery model of hemoglobin. Modulation of O(2) affinity, cooperativity, and Bohr effect by heterotropic allosteric effectors. J Biol Chem. 2002;277:34508-20.
15. Teisseire BP, Ropars C, Vallez MO, Herigault RA, Nicolau C. Physiological effects of high-P50 erythrocyte transfusion on piglets. J Appl Physiol (1985). 1985;58:1810-7.
16. Fylaktakidou KC, Lehn JM, Greferath R, Nicolau C. Inositol tripyrophosphate: a new membrane permeant allosteric effector of haemoglobin. Bioorg Med Chem Lett. 2005;15:1605-8.
17. Kieda C, Greferath R, Crola da Silva C, Fylaktakidou KC, Lehn JM, Nicolau C. Suppression of hypoxia-induced HIF-1alpha and of angiogenesis in endothelial cells by myo-inositol trispyrophosphate-treated erythrocytes. Proc Natl Acad Sci USA. 2006;103:15576-81.
18. Biolo A, Greferath R, Siwik DA, Qin F, Valsky E, Fylaktakidou KC, Pothukanuri S, Duarte CD, Schwarz RP, Lehn JM, Nicolau C, Colucci WS. Enhanced exercise capacity in mice with severe heart failure treated with an allosteric effector of hemoglobin, myo-inositol trispyrophosphate. Proc Natl Acad Sci USA. 2009;106:1926-9.
19. Sihn G, Walter T, Klein JC, Queguiner I, Iwao H, Nicolau C, Lehn JM, Corvol P, Gasc JM. Anti-angiogenic properties of myo-inositol trispyrophosphate in ovo and growth reduction of implanted glioma. FEBS Lett. 2007;581:962-6.
20. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure-activity relationship. Excli J. 2009;8:74-88.

21. Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds. Exp Opin Drug Discov. 2010;5:633-54.
22. Mandi P, Nantasenamat C, Srungboonmee K, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR study of anti-prion activity of 2-aminothiazoles. Excli J. 2012;11:453-67.
23. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine. J Mol Graph Model. 2008;27:188-96.
24. Nantasenamat C, Piacham T, Tantimongcolwat T, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR model of the quorum-quenching *N*-acyl-homoserine lactone lactonase activity. J Biol Syst. 2008;16:279-93.
25. Thippakorn C, Suksrichavalit T, Nantasenamat C, Tantimongcolwat T, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Modeling the LPS neutralization activity of anti-endotoxins. Molecules. 2009;14:1869-88.
26. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul S, Prachayasittikul V. Predicting the free radical scavenging activity of curcumin derivatives. Chemometr Intell Lab Syst. 2011;109:207-16.
27. Worachartcheewan A, Nantasenamat C, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Modeling the activity of furin inhibitors using artificial neural network. Eur J Med Chem. 2009;44:1664-73.
28. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Quantitative structure-imprinting factor relationship of molecularly imprinted polymers. Biosens Bioelectron. 2007;22:3309-17.
29. Nantasenamat C, Isarankura-Na-Ayudhya C, Tansila N, Naenna T, Prachayasittikul V. Prediction of GFP spectral properties using artificial neural network. J Comput Chem. 2007;28:1275-89.
30. Nantasenamat C, Naenna T, Isarankura Na Ayudhya C, Prachayasittikul V. Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network. J Comput Aided Mol Des. 2005;19:509-24.
31. Nantasenamat C, Srungboonmee K, Jamsak S, Tansila N, Isarankura-Na-Ayudhya C, Prachayasittikul V. Quantitative structure–property relationship study of spectral properties of green fluorescent protein with support vector machine. Chemometr Intell Lab Syst. 2013;120:42-52.
32. Hansch C, Garg R, Kurup A, Mekapati SB. Allosteric interactions and QSAR: on the role of ligand hydrophobicity. Bioorg Med Chem. 2003;11:2075-84.
33. ChemAxon Ltd. MarvinSketch, Version 621 (2014) Budapest, Hungary.
34. Author. Gaussian 09. Revision A.1. Wallingford, Connecticut; 2009.
35. Talete srl. DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 55 (2007) Milano, Italy.
36. Whitley DC, Ford MG, Livingstone DJ. Unsupervised forward selection: a method for eliminating redundant variables. J Chem Inf Comput Sci. 2000;40:1160-8.
37. IBM Corporation. SPSS Statistics, Version 18 (2011) New York, USA.
38. Kennard RW, Stone LA. Computer aided design of experiments. Technometrics. 1969;11:137-48.
39. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20:273-97.
40. Vapnik VN. Statistical Learning Theory. New York: Wiley-Interscience; 1998.
41. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems): Morgan Kaufmann Publishers Inc.; 2005.
42. Rücker C, Rücker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model. 2007;47:2345-57.
43. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5:299-314.

44. Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN. Some case studies on application of "r_m^2" metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data. J Comput Chem. 2013;34:1071-82.
45. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Softw. 2008;25:1-18.
46. Nadolny C, Kempf I, Zundel G. Specific interactions of the allosteric effector 2,3-bisphosphoglycerate with human hemoglobin--a difference FTIR study. Biol Chem Hoppe Seyler. 1993;374:403-7.
47. Eriksson L, Johansson E. Multivariate design and modeling in QSAR. Chemometr Intell Lab Syst. 1996;34:1-19.

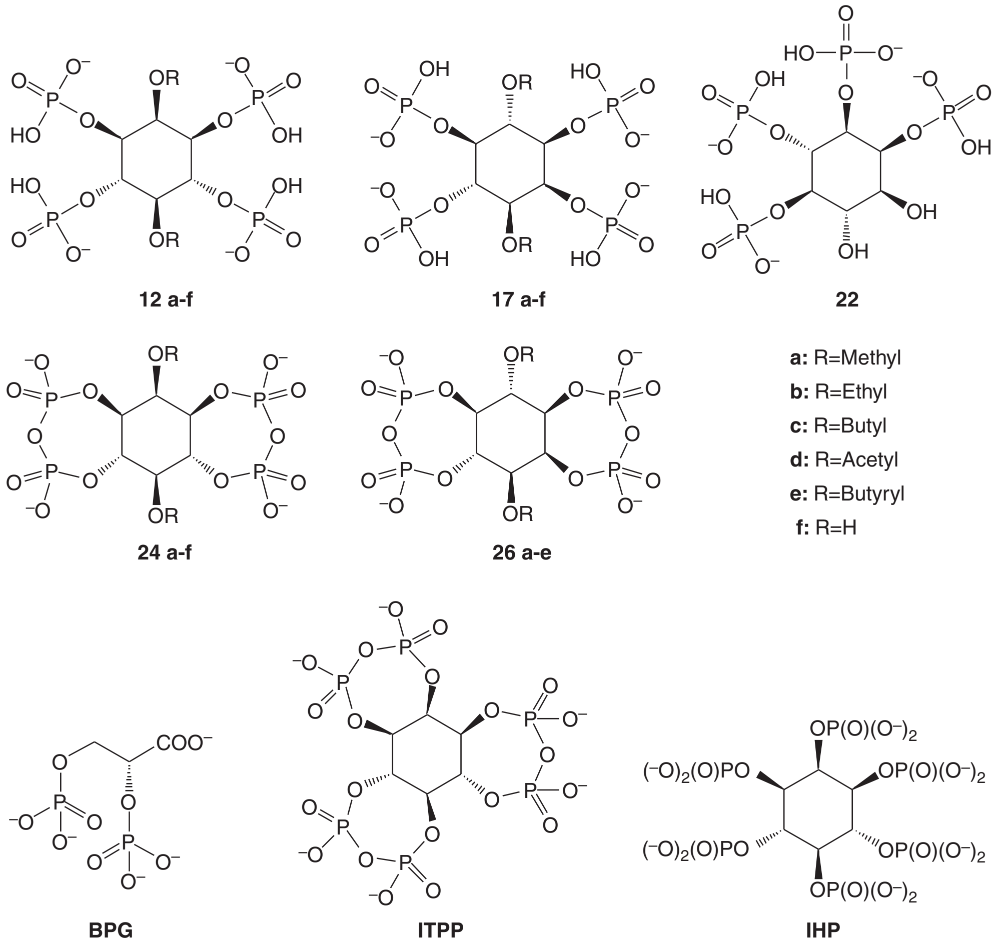


Figure 1. Chemical structures of ITPs (**12a-f**, **17a-f** and **22**) and IBBPs (**24a-f** and **26a-e**).

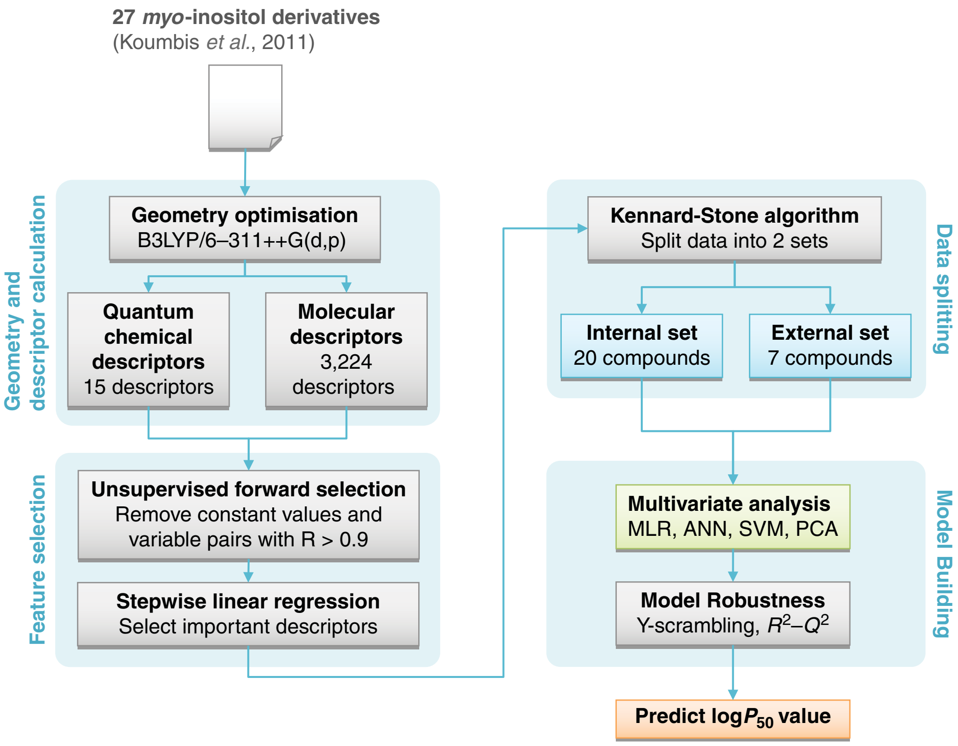


Figure 2. Schematic of the QSAR modeling workflow.

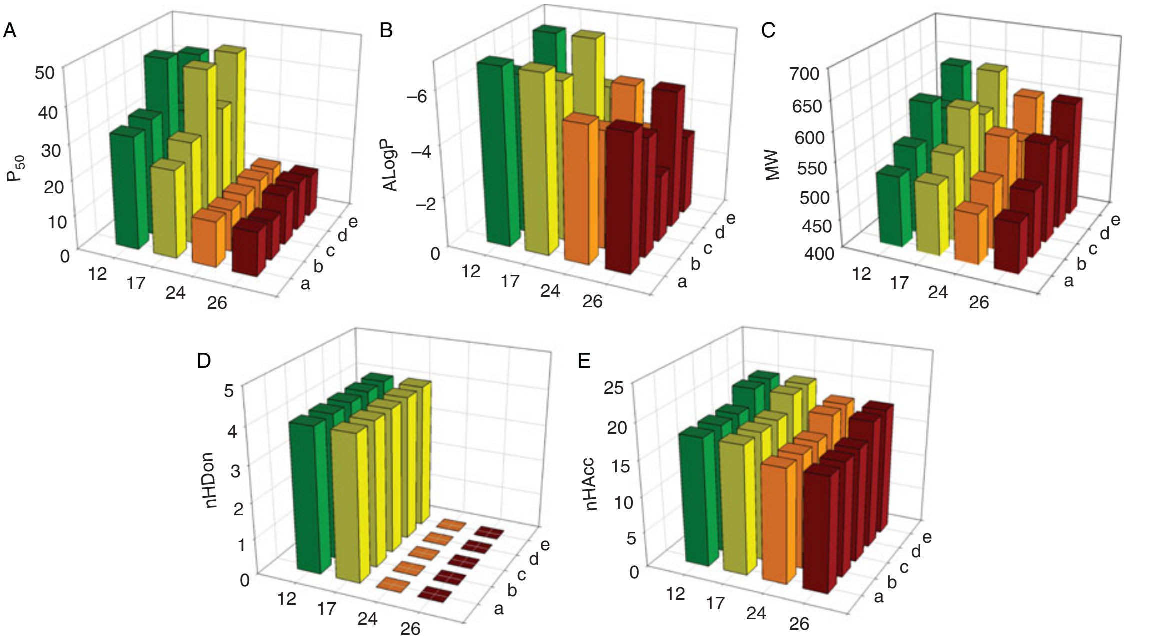


Figure 3. Three-dimensional bar plots showing overall distribution of data values for P_{50} (A) and Lipinski's descriptors (B-E).

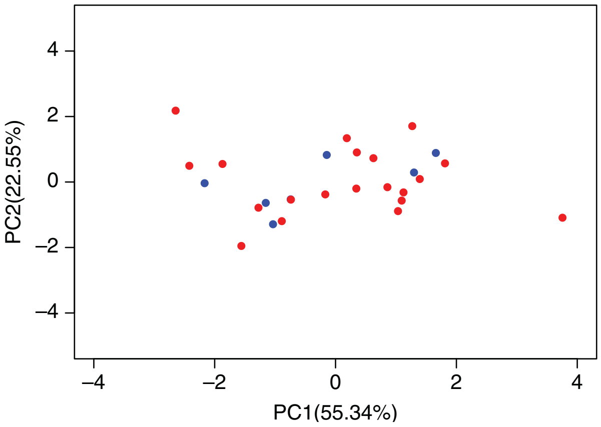


Figure 4. Chemical space spanned by internal and external sets as shown in blue and red colors, respectively.

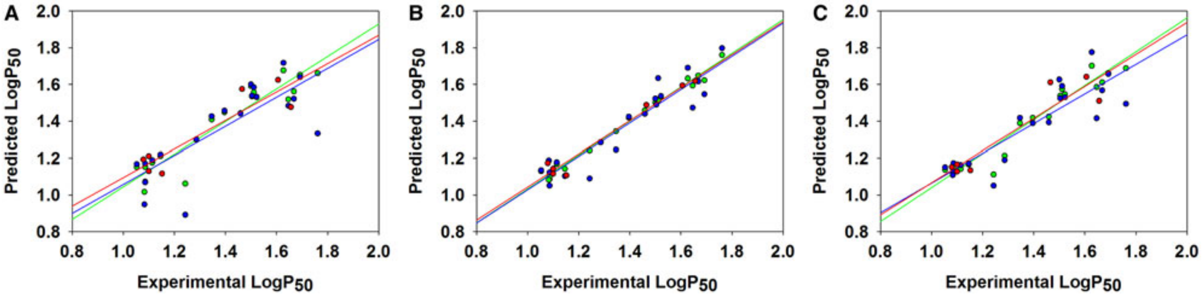


Figure 5. Plot of experimental versus predicted activities for investigated compounds using MLR (A), ANN (B) and SVM (C).

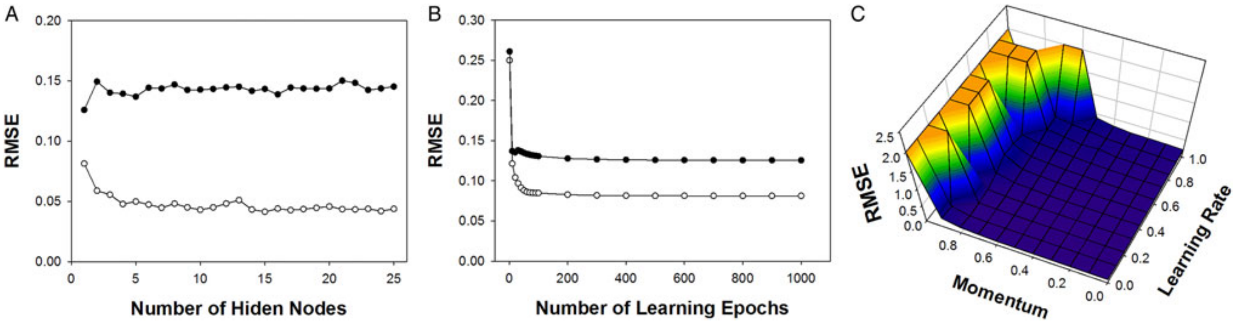


Figure 6. Optimization of ANN parameters consisting the number of nodes in the hidden layer (A), number of learning epochs (B) and the learning rate and momentum (C).

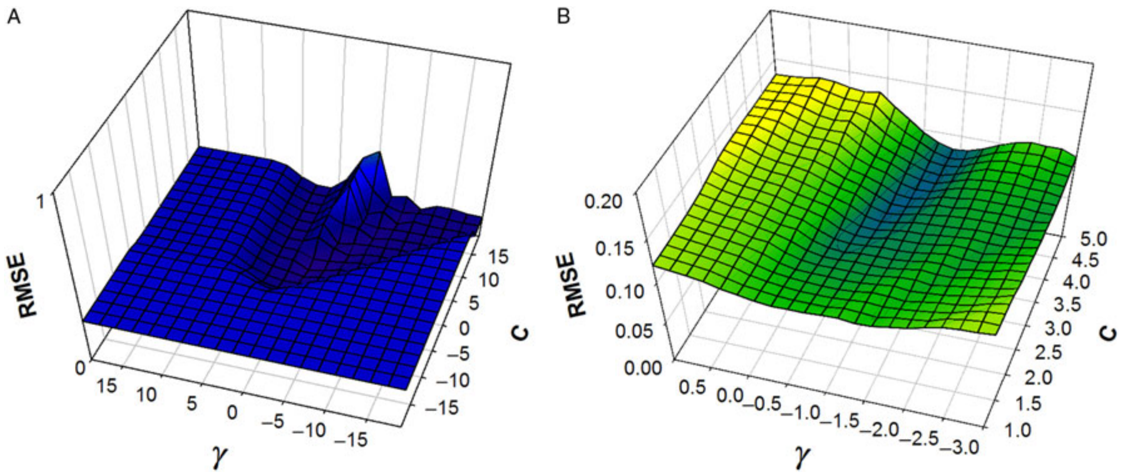


Figure 7. Optimization of SVM parameters by means of a global (A) and local (B) search.

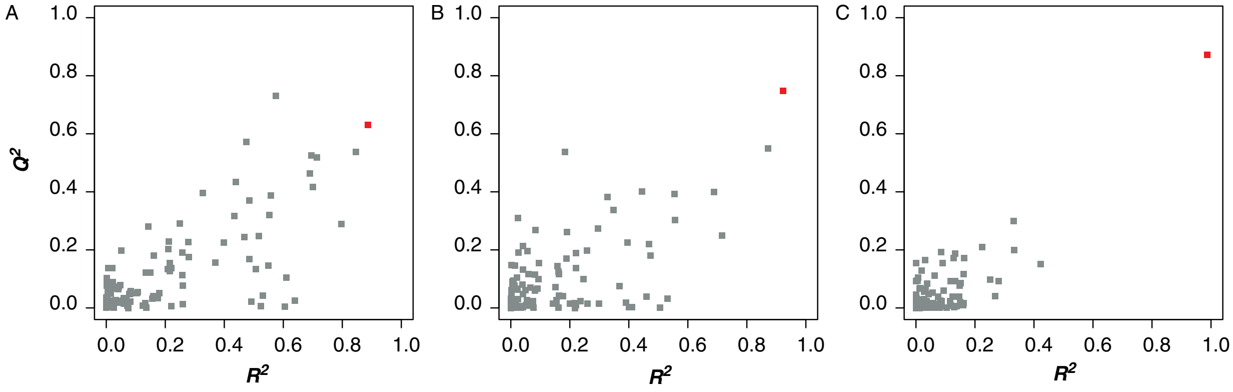


Figure 8. Plot of R_{Tr}^2 and Q_{CV}^2 from Y-scrambling experiments using MLR (A), ANN (B), and SVM (C). Gray and red points represent results from Y-scrambled and actual models, respectively.

Table 1. Data set of *myo*-inositol derivatives.

Compound	log P_{50}	nHDon	TPSA	ALogP	Q_{pos}
12a[a]	1.5038	4	336.06	−6.824	0.5939
12b[a]	1.5224	4	336.06	−6.126	0.5870
12c[a]	1.6684	4	336.06	−4.166	0.5982
12d[b]	1.4654	4	370.20	−6.882	0.5663
12e[a]	1.6274	4	370.20	−4.636	0.6294
12f[a]	1.6920	6	358.06	−7.641	0.5717
17a[a]	1.3962	4	336.06	−6.824	0.5289
17b[a]	1.4594	4	336.06	−6.126	0.5210
17c[b]	1.6561	4	336.06	−4.166	0.5343
17d[a]	1.4997	4	370.20	−6.882	0.5787
17e[a]	1.6464	4	370.20	−4.636	0.5109
17f[a]	1.5105	6	358.06	−7.641	0.5010
22[b]	1.6064	6	358.06	−7.641	0.5520
24a[a]	1.1139	0	273.60	−5.291	0.5765
24b[b]	1.0792	0	273.60	−4.593	0.5855
24c[a]	1.0864	0	273.60	−2.633	0.4830
24d[b]	1.1004	0	307.74	−5.349	0.5453
24e[a]	1.0864	0	307.74	−3.103	0.4879
24f[a]	1.2856	2	295.60	−6.108	0.5587
26a[b]	1.1004	0	273.60	−5.291	0.5419
26b[a]	1.0531	0	273.60	−4.593	0.5540
26c[b]	1.1523	0	273.60	−2.633	0.5160
26d[a]	1.1461	0	307.74	−5.349	0.5459
26e[a]	1.0828	0	307.74	−3.103	0.3866
BPG[a]	1.2430	0	204.59	−4.833	0.6046
IHP[a]	1.7612	0	493.38	−11.932	0.6081
ITPP[a]	1.3464	0	382.71	−7.629	0.5806

[a] Internal set.

[b] External set.

Table 2. Summary of the predictive performance of MLR, ANN and SVM models for predicting the Hb allosteric effector activity of myo-inositol derivatives.

Methods	R^2_{Tr}	Q^2_{CV}	Q^2_{Ext}	$\mathrm{RMSE}_{\mathrm{Tr}}$	$\mathrm{RMSE}_{\mathrm{CV}}$	$\mathrm{RMSE}_{\mathrm{Ext}}$
MLR	0.8859	0.6306	0.8332	0.0775	0.1487	0.1015
ANN	0.9235	0.7484	0.8847	0.0642	0.1176	0.0827
SVM	0.9876	0.8722	0.9694	0.0298	0.0827	0.0465

Publication 2

Exploring the origins of structure-oxygen affinity relationship of human hemoglobin allosteric effector

1. Introduction

Thalassemia is the most commonly inherited anemia, and it is widely distributed in the Mediterranean, the Middle-East, South-East Asia and sub-Saharan Africa. This condition can be linked to imbalanced α- and β-globin synthesis. The pathophysiology of β-thalassemia is associated with absent or reduced β-globin production, leading to an excess of α-globin polypeptides. This triggers a range of activities, including the formation of reactive oxygen species (ROS), release of hemin and free iron [1]. The clinical management of β-thalassemia largely depends on life-long red blood cell transfusions and iron chelation. Many alternative and experimental treatments have been developed, including promotion of fetal hemoglobin production, but to date, none of these treatments have reached widespread clinical use.

More than 200 alternative genetic lesions have been identified that affect the β-globin gene, resulting in a wide range of anemic severity. However, only approximately 10 of these lesions are responsible for 90 % of the cases. In particular, abnormal transcript splicing has been associated with β-thalassemia, clearly indicating that modulation of splicing can be a valuable approach for treatment. The use of splice switching oligonucleotides (SSO) has been promising [2]. Mutations in the second intron of the β-globin gene at positions 654 and 705 are clinically important and cause mis-splicing, which results in a defective protein. SSOs in antisense form to the 654 and 705 transcripts have been demonstrated to restore proper splicing in mouse models [3]. High-throughput screening of chemical libraries has also indicated that several compounds can be used to modulate such erroneous splicing.

Quantitative structure–activity relationship (QSAR) represents an important approach for elucidating the origin of biological activity for a set of compounds of interest as a function of their molecular descriptors [4,5]. The resulting QSAR models can reveal molecular features that are essential for active compounds and that can subsequently be used as therapeutic agents. We recently applied QSAR to understand the underlying physicochemical features defining Hb allosteric effector activity for a set of myo-inositols [6]. In this QSAR study, we examine the origin of the splice switching activity of Hb β-globin gene modulators via predictive QSAR modeling to draw conclusions regarding the most effective chemical structure that is useful in clinical settings.

Challenges in the development of QSAR/QSPR models can arise from the following: (i) selection of an appropriate subset of molecular descriptors from the many available descriptors, (ii) inability to interpret the descriptors, and (iii) the need to optimize chemical structures if three-dimensional descriptors are to be used. To address the above issues, QSAR models were developed using interpretable substructure fingerprints to quantitatively represent gene modulators. These descriptors were correlated to the splice switching activity of the hemoglobin β-globin gene using a wide array of machine learning methods, including rule-based, ensemble, non-linear classification and linear classification methods. Insights into important features governing the origin of splice switching activity as deduced from the constructed models could be used to further guide the design of novel ASOs with desired activity.

2. Materials and methods

2.1. Data set

A data set of small molecule modulators with splice switching activity against the Hb β-globin gene with a mutation in the second intron at position 654 (IVS2-654) was obtained from PubChem BioAssay (accession number AID 925) [7]. The data originated from a high-throughput screen of a chemical library of 64,405 compounds. Because 222 compounds were reported as having inconclusive activity, these compounds were therefore excluded. Compounds were treated with the QSAR curation workflow from Fourches et al. [8]. Briefly, the main steps are as follows: (i) removal of inorganics and mixtures, (ii) structural conversion and cleaning, (iii) normalization of specific chemotypes, (iv) removal of duplicates and (v) final manual checking. Thus, chemical compounds were curated using ChemAxon Standardizer with the following options: Strip Salts, Aromatize, Clean3D, Tautomerize, Neutralize, and Remove explicit hydrogens [9]. The resulting data set is composed of 39 and 60,647 active and inactive compounds. A representative subset of the chemical structures of these compounds (i.e., particularly the 39 active compounds) are shown in Fig. 1. The Open Babel software [10] was used to convert compounds from the SMILES notation to the SDF file format, which is suitable as an input for the PaDEL-Descriptor software.

2.2. Compound descriptors

Fingerprint descriptors provide descriptions of the constituting substructures inherently present in a molecule. There are essentially two versions of this descriptor, namely, the count and binary versions. Each bit in a string of fingerprint descriptors represents a distinct substructure [11]. In the count version, the numerical value, as the name implies, represents the frequency of that substructure present in a molecule, whereas in the binary version, values of 1 and 0 denote its presence and absence, respectively. These interpretable substructure fingerprints were calculated using the PaDEL-Descriptor software [12]. Thus, it can pinpoint substructures of a compound that are important for the activity of splice switching modulators [13].

2.3. Data filtering

Collinearity is a condition in which pairs of descriptors have a correlation with each other. It has a substantial negative impact on the computational analysis because correlated descriptors add more complexity to the model [14]. In addition, it also affects the interpretation of descriptors (i.e., substructure fingerprint count) because the resulting coefficient estimates (e.g., linear models) or feature usages (e.g., decision trees) are highly unstable. Moreover, the statistical assessment measures will be very sensitive to the predictive models and may inhibit the ability of prediction for new observations because correlated data create redundancy, which may overfit models. Overfitting is a condition when predictive models perfectly predict the training set. However, when new samples are introduced for prediction, the models perform ineffectively. In general, a Pearson's correlation coefficient of 0.7 is an indicator of high collinearity among predictors [15]. Thus, the cor function from the R package stats was used to calculate correlations among descriptors [16]. To obtain filtered descriptors with all pairwise correlations less than 0.7, the findCorrelation function from the R package caret with a cutoff at 70% was used [16]. The remaining descriptors that were used in the study are shown in Fig. 4.

2.4. Data pre-processing

Initially, the collected data set contained highly imbalanced data, in which 61,000 and 39 compounds were inactive and active compounds, respectively. To create a balanced data set, the undersampling approach was performed by applying K-means clustering on the inactive group of compounds. Prior to performing K-means clustering, PCA was utilized to reduced the dimension of the data set to obtained non-correlated variables, also known as principal component (PC) coefficients. These PC coefficients were used as inputs for performing K-means clustering to derive 39 clusters for a pool of inactive compounds [17]. A random point from each cluster was selected to represent a set of 39 inactive compounds. The resulting pre-processed data set is provided as supplementary data on figshare that is available at http://dx.doi.org/ 10.6084/m9.figshare.1609584.

2.5. Univariate analysis

Univariate analysis was conducted to investigate patterns, features and trends that were present in the substructure descriptors. It was performed by creating histogram plots using the R package ggplot2. The normality of each substructure's fingerprint for active and inactive compounds was assessed using the Shapiro-Wilk test using the shapiro.test function from the R package stats. The function pairwise.wilcox.test from the R package stats was used to perform a Mann–Whitney U test to measure the statistical significance of the investigated pairs (i.e., active and inactive compounds).

2.6. QSAR modeling

The decision tree algorithm J48 is Weka's implementation of the C4.5 algorithm that automatically generates classification rules in the form of a rule-based branching tree using the divide-and-conquer algorithm. It is considered to be one of the most trans-parent learning algorithms in which a series of readily understand-able if-then rules are formed. The construction involves two steps: growing and pruning. Growing starts from the root node, which branches out to form internal nodes that subsequently end up as leaf nodes. Internal nodes represent descriptors, branches describe descriptor values, and leaf nodes represent Y categorical classes (i.e., active and inactive). Once the trees are fully grown, the grown tree is pruned as a function of the predictive performance. The ad-vantage of pruning is that it reduces the complexity of the formed tree and reduces the chance of over fitting. The J48 function from the Rweka package was used to construct the QSAR models [18].

Random forest (RF) is an ensemble classifier that is composed of multiple decision tress. Similar to J48, classification starts at the root node, in which the data set at the node is split according to the value of descriptors that are selected such that the descriptors of different activities are predominantly moved to different branches. The classification is obtained by averaging the results of all trees by a majority vote from each tree. The RF classifier was generated using the R package randomForest with a total of 500 trees [19]. Two RF parameters were subjected to optimization including the ntree (i.e. the number of trees used for building the ensemble RF model) and mtry (i.e. number of descriptors to be sampled randomly as candidate features).

Support vector machine is a machine learning approach that can be used to perform both classification and regression, in which the kernel function is used to map the data into a high-dimensional feature space. The commonly used radial basis function kennel was used to construct a predictive model. The support vectors were fine-tuned with several parameters to obtain the optimal parameters, which are the width of the kernel function gamma and the error penalty parameters cost. The commonly used radial basis Gaussian kernel was selected along with tuned parameters ($C = 2$, $\gamma = 0.29$) to construct SVM machine learners. The train R package caret was used to fine-tune the model, and the support vector machine was constructed using the R package e1071 [20].

Artificial neural network (ANN) is a machine learning method that mimics the human brain, which is composed of networks of inter-connected neurons that function in relaying messages in the form of electrochemical signals. Brain cells are composed of dendrites, cell bodies and axons. A synapse is the connection between the axons of the nerve cell with the dendrites of an adjacent nerve cell. In a synapse, signals are transmitted from one cell to the next as neurotransmitters. Similar to the function of the human brain, the artificial neuron is interconnected in a feed-forward manner from the first through the last nodes, in which the connection between nodes of different layers is assigned as a weight, which can be expressed as a strength of the input data. The train function from the R package caret was used to construct the artificial neural network model while fine-tuning the parameters. The train function from the R package caret was also used to obtain optimal parameters (i.e., hidden layers = 7 and decay weight = 0.1), and the nnet function from the R package nnet was used to train the models [21].

Partial least squares discriminatory analysis (PLS-DA) is a linear classification method that seeks to categorize samples into groups (i.e., actives and inactives) based on the predictor characteristics. It is a robust (the parameters of the model do not change when samples are taken out) multivariate analysis that involves X and Y variables. It simultaneously projects the X into latent variables to correlate X predictors with Y responses. The extent of the influence that X has on the Y variable is revealed by the regression coefficient when PLS models are constructed for each responsive variable. The plsda function from the R package caret was used to construct predictive models [16].

2.7. Data splitting

The data set was randomly partitioned into two subsets (i.e., 80% as the internal set and 20% as the external set) using the sample_n function from the R package dplyr [22]. The internal set was used to train the model, while the external set was used to externally validate the performance of the predictive model. To avoid the bias that may arise from a single data split when training the model, predictive models were constructed from each of the 50 independent data splittings, and the mean and standard deviation values of statistical parameters were reported.

Ten-fold cross-validation (10-fold CV) and external testing were used to validate the robustness and reliability of the predictive models. Ten-fold CV was performed on the internal set, where the data set is separated into ten folds. Practically, one fold from the total of ten folds is left out as the testing set, while the remaining are used for training the models. This process is repeated iteratively until all the data samples had a chance to be left out as a testing test. External validation evaluates the stability of the predictive models by serving as unknown data previously not seen by the training model.

2.8. Validation of QSAR models

Model validation is essential in evaluating the results of empirical modeling. Several statistical parameters are used to assess the effective-ness and efficiency of constructed predictive models, including accuracy, sensitivity, specificity and Matthews correlation coefficient (MCC) on both dependent and independent test sets. These statistical parameters are commonly used in QSAR modelings [23]. Accuracy is the percentage of correctly classified instances relative to the total number of instances. Although it is commonly used to assess the predictability of predictive models, accuracy is not an optimal measure of model performance if the data are unbalanced. In contrast to accuracy, MCC is a mea-sure of assessment that is insensitive to unequal sizes of the classes and the costs of making certain errors. The Sen, sometimes considered as the true positive (TP) rate, is the proportion of true positives among all positively classified instances, whereas specificity is the proportion of true negatives (TN) among all negatively classified instances, thus defined as the false-positive rate.

These can be calculated using the equations described below:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100 \qquad (1)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100 \qquad (2)$$

$$Specificity = \frac{TN}{(TN + FP)} \times 100 \qquad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (4)$$

where TP is the instances of true positives, TN is the instances of true negatives, FP is the instances of false positives, and FN is the instance of false negatives. The range of MCC is from −1 to 1, in which a value of MCC = 1 indicates the best possible prediction, while MCC = −1 indicates the worst possible prediction. On the other hand, MCC = 0 suggests a random prediction scheme.

3. Results and discussion

The notion that the activity of compounds is governed by their physicochemical properties is a paradigm of QSAR. To gain understanding of the activity of splice switching modulators of the β-globin gene of hemoglobin, a set of easy-to-interpret substructure fingerprints was used to encode the compounds. The intercorrelation between predictors was removed using a Pearson's correlation coefficient threshold of 0.7 as an indicator of high collinearity [15]. To assess the robustness and reliability of the predictive model, the data set was split into two subsets: (i) internal set (i.e., used to train and fine tune the model) and an (ii) external set (i.e., used to validate the performance of the predictive models in the real-world setting). Feature importance selection was then performed using the built-in RF feature selector. To avoid random seeding when constructing the predictive models, 50 independent data splittings were performed, where each split was used to construct models. This was followed by computing the mean and standard deviation values for each statistical performance metric (i.e., accuracy, sensitivity, specificity and MCC) that was used in evaluating the predictive performance of the model. The general framework of the QSAR modelings is shown in Fig. 2.

3.1. Univariate analysis of Hb β-globin gene modulators

The normality of the data set (i.e., substructure fingerprint) was assessed using the Shapiro-Wilk test [24]. The results indicated that all descriptors afforded p values less than 0.05, suggesting that all descriptors exhibited a non-normal distribution. Thus, the Mann–Whitney U test was utilized to test whether two populations of activity (i.e., active or inactive) are equal or different from one another. As shown in Table 1, substructures SubFPC5, SubFPC26, SubFPC52, SubFPC100, SubFPC101, SubFPC135, SubFPC153, SubFPC179 and SubFPC180 displayed statistically significant differences with corresponding p values of 0.009, 0.046, 0.008, 0.000, 0.013, 0.003, 0.023, 0.003 and 0.034, respectively. The Lipinski's rule-of-5 indicated that nearly all compounds from the data set were drug-like as they had values in accordance with the following molecular properties: (i) molecular weight b 500 Da, (ii) LogP b 5, (iii) number of hydrogen bond donors b 5 and (iv) number of hydrogen bond acceptors b 10. Simple histograms of significant descriptors are shown in Fig. 3 to provide an overview of the relative distribution of the data values.

SubFPC5 is a count of alkene, which is a hydrocarbon that contains a carbon-carbon double bond, in the investigated compounds. The histogram plot of SubFPC5 shows that most of the inactive compounds have a count of 0, as shown in Fig. 3. In addition, it can be observed that SubFPC5 having a count of at least 1 is a general common feature for the active group compared to the inactive group, which have values of 0.267±0.495 (active) and 0.067±0.330, respectively (p b 0.05). However, it can be observed that active group had the same count with the inactive group with ranges of 2 (maximum) and 0 (minimum). Moreover, more than half of the active and the inactive groups have values of 0 counts.

SubFPC26 is a count of tertiary aliphatic amine, which is three organic substituents connected to a nitrogen atom. As shown in Fig. 1, the distribution of SubFPC26 shows a significant difference between the active and inactive groups, as was the case for SubFPC5, with values of 0.044±0.208 and 0.178±0.387, respectively (p b 0.05). Both of the ranges of active and inactive were not strikingly different ranges with the minimum count (0) to maximum count (1). Nevertheless, the Mann–Whitney U test revealed that the active and inactive groups are significantly different with a p value of 0.046. When comparing the active and inactive groups according to substructure count, the majority of the active group has a count of 0, whereas the majority of the inactive group has a count of 1, as shown in Fig. 3.

SubFPC52 is a count of imine, a nitrogen double bonded to carbon where a substituent is bonded to the nitrogen atom and two substituents are bonded to the carbon atom. The mean and standard deviation of SubFPC52 were 0.200±0.405 and 0.020±0.149 for active and inactive, respectively. The distribution range of SubFPC52 was similar to that of SubFPC5, where the majority of the inactive were higher than that of active and lower than that of active for substructure counts of 0 and 1, respectively. The Mann–Whitney test presented a p value of 0.008, indicating that the two groups were significantly different. The histogram plot of SubFPC52 in Fig. 3 shows that the frequency of in-active has more SubFPC52 than the active group and that active has more than inactive for counts of 0 and 1, respectively.

SubFPC100 is a count of secondary amide. The corresponding values of SubFPC100 are 0.111±0.318 and 0.533±548 for active and inactive, respectively. A simple statistical analysis revealed that the p value for the Mann-Whitney u test was (0.000). Note that the resulting p value is the smallest of all the significant descriptors. The majority of the active has a SubFPC100 count of 0, whereas the inactive groups were equally distributed for both count of 0 and count of 1, as shown in Fig. 3.

SubFPC101 is a count of tertiary amide in the investigated substructure. On average, inactive compounds had higher counts for SubFC101 when compared to active compounds with values of 0.244 ± 0.484 and 0.044±0.208 for inactive and active, respectively (p value b 0.05). The histogram clearly showed that SubFPC101 was negatively skewed to the left as the majority of the active and inactive compounds have a count of 0, as shown in Fig. 3. In general, the inactive compounds had a greater count number than the active compounds with maximum count ranges of 2 and 1, respectively.

SubFPC135 is a count of carboxyl derivatives. The histogram clearly showed that both types of compounds had a SubFPC135 close to zero. Notably, the inactive compounds had dynamic ranges of counts from 5 (maximum) to 0 (minimum), whereas the active compounds had a narrow range of counts from 2 (maximum) to 0 (minimum). It was found that the counts of SubFPC135 were higher in inactive compounds compared to the active compounds with values of 0.644 ± 1.151 and 0.111 ± 0.438, respectively (p b 0.05). The distribution of SubFPC135 for the inactive compounds was observed to be jagged relative to that of non-steroids, suggesting that inactive compounds tend to have SubFPC135 distributed in dynamic ranges.

SubFPC153 is a count of urethane, also known as ethyl carbamate. The histogram plot shows that all active compounds and a majority of in-active compounds lack SubFPC153 as a substructure, whereas a minority of inactive compounds contain 1 count. A simple statistical analysis revealed that average values were 0.000±0.000 and 0.111±0.318 for active and inactive compounds, respectively (p b 0.05). The histogram plot clearly showed that all active compounds have a count of 0 for SubFPC153, whereas the inactive compounds have a narrow range of counts of 1 at maximum and 0 at minimum, indicating that the inactive compounds may have at most 1 substructure of this type.

SubFPC179 is a count of hetero nitrogen basic hydrogen. The corresponding values for the SubFPC179 were 0.178 ± 0.387 and 0.000 ± 0.000 for active and inactive, respectively. The histogram plot clearly shows that inactives do not posses any of this sub-structure, whereas active had at least one SubFPC197, as shown in Fig. 3.

SubFPC180 is a count of hetero nitrogen basic no hydrogen. It was observed that the SubFPC180 were statistically different for active and inactive, with values of 0.067±0.252 and 0.267±0.539, respectively. It was found that the values of SubFPC180 are lower in active compounds compared to their inactive counterparts. As shown in Fig. 3, the distribution of the inactive was broader than that of active with ranges from 3 (maximum) to 0 (minimum). However, the majority of the active compounds do not possess SubFPC180 as a substructure. Nevertheless, the Mann–Whitney test revealed that the active and inactive groups are significantly different with a p value of 0.034.

3.2. QSAR modeling

Because compounds were characterized by substructure fingerprint descriptors, this allowed us to pinpoint the substructures that are important for modulating the splice switching activity of the β-globin gene chain. To avoid the inherent redundancy among the substructures, descriptors were filtered using a cutoff value of 0.70. The undersampling approach was applied to solve the class imbalance problem where the number of active and inactive compounds are significantly out of proportion. This approach had successfully been shown to be effective for handling the inherently imbalanced data derived from the PubChem database. [25].

As previously mentioned, the initial data set was split into an internal validation set and an external testing set, in which the former constituted 80% of the data set while the latter constituted the remaining 20% of the data set. QSAR models were developed with various machine learning methods, including rule-based models (e.g., J48), ensemble models (e.g., RF), non-linear classification models (e.g., ANN and SVM) and linear classification models (e.g., PLS-DA). To avoid the bias of a single data split, data splitting was performed for 50 iterations in which each split was used to construct a predictive model. The mean and standard deviation of the resulting predictive performance (e.g., accuracy, specificity, sensitivity and MCC) were computed as assessed by 10-fold CV and external sets.

Firstly, the rule-based J48 algorithm (i.e., Weka's implementation of the C4.5 algorithm) was used for identifying rules governing the relationship of independent variables (i.e., substructure fingerprint) with that of the dependent variable (i.e., activity). As shown in Table 2, the predictive performance of the training set and 10-fold CV set provided accuracies of 94.38±2.01 and 80.33±5.85, respectively. A closer examination of the predictive model revealed that the training set provided a better overall predictive performance than the 10-fold cross validation. Moreover, the external testing set afforded a moderate MCC value of 0.58±0.23. Secondly, RF is a popular ensemble technique in machine learning in which multiple decision trees are bagged to generate prediction and the predictions are averaged to provide the bagged model's prediction. The bagging of trees improves the predictive performance over a single tree by reducing the variance of the prediction.

As also shown in Table 2, the training set for RF was as high as 100.00±0.00% for accuracy, sensitivity and specificity and 1.00±0.00 for MCC. On the other hand, the 10-fold CV set was slightly lower with 89.50±13.45, 94.97±13.49, 84.29±22.27 and 0.80±0.25 for accuracy, sensitivity, specificity and MCC, respectively. Nevertheless, it can be observed that the predictive performance of the RF was considerably better compared with J48 In addition, it can be seen the highest performance for the external set was modeled by RF. Third, ANN is a powerful non-linear classification technique that works similar to a human brain. The outcome is modeled by the intermediates of predictor variables created through non-linear functions. The predictive performance of the ANN is high with 100.00±0.00 for accuracy, sensitivity and specificity and 1.00±0.00 for MCC. The 10-fold CV of the ANN is comparable to the RF with 85.48±15.96, 84.15±21.90, 86.16±24.93 and 0.74±0.29 for accuracy, sensitivity, specificity and MCC, respectively. As can be seen in Table 2, the MCC value for ANN was 0.74±0.20 which is slightly lower than of RF (0.75 ± 0.18) but superior to J48 (0.58 ± 0.23), SVM (0.67 ± 0.20) and PLS-DA (0.69 ± 0.22). Fourth, SVM is another non-linear modeling technique that is considered to be a powerful and highly flexible machine learner. It was originally developed for classification, and the predictive performance of the models is comparable to other machine learners (i.e., rule-based, ensemble and linear). As shown in Table 2, the assessment parameters for the SVM are relatively high, with accuracies of 96.47 ± 3.70 and 80.19 ± 15.62 for training and 10-fold CV, respectively. Finally, PLS-DA is a linear classification method in which predictors undergo dimension reduction with respect to the response. The predictive performance of PLS-DA is comparable to that of other machine learners, where the accuracy, sensitivity, specificity and MCC were 93.47±3.82, 93.47±5.93, 93.08±4.91 and 0.86±0.08, respectively, for the training set and 81.08 ± 14.97, 77.79 ± 27.74, 79.33 ± 29.61 and 0.65 ± 0.25, respectively, for the 10-fold CV. The predictive power of PLS-DA to accurately predict the activity of unknown compounds as assessed by the external set was good with MCC of 0.69±0.22. The overall predictive performances of the machine learning methods (i.e., J48, RF, ANN, SVM and PLS-DA) are highly comparable. However, based on the predictive power of both 10-fold CV and external sets, RF outperformed the other learning methods. This result indicates that the RF model was able to correlate unknown chemical structures with splice switching activity of the hemoglobin β-globin gene. Consequently, the RF model was chosen to represent the best QSAR model and further used in interpreting the feature importance governing the splice switching activity of β hemoglobin gene modulators.

3.3. Optimization of RF parameters

RF is increasingly used in QSAR modeling owing to its robust performance and built-in measure of feature importance. There are two RF parameters that can be optimized, which is comprised of ntree and mtry as previously mentioned in the Materials and Methods. Generally, according to the documentation of the randomForest R package, the default values of ntree and mtry are set at 500 and N (i.e. where N denote the total number of descriptors), respectively. Parameter optimization may lead to improvement in the resulting predictive performance as different data have their own unique characteristics (i.e. size, type and structure). As optimization of RF parameters is a highly time-consuming process owing to the relatively large size of ntree and mtry values coupled to the rather lengthy calculation of the k-fold CV scheme, therefore good default values could save considerable amount of computational cost. Thus, these type of benchmarking study had previously been employed for identifying optimal parameters of SVM as a function of signature fingerprints. [26].

Fig. 5 showed that mtry in the range of 1 and 10 turned out to be the best while mtry greater than 15 resulted in low accuracy. To get a better understanding on the best combination of mtry and ntree, the optimization step were repeated 50 times with independent training set. As shown in Fig. 6, the histogram suggests that the optimal range of mtry is no more than 5 while the ntree is no more than 1000. On the basis of the parameter optimization, the optimal range of mtry is between 1 to 5 while for ntree the optimal range is between 100 to 1000, which are the suggested good staring point in terms of trade-off between speed and performance. Nevertheless, it is worthy to note that using higher number of trees (i.e. greater than 1000) provided no benefit as it only adds to the complexity of the model, which possibly may give rise to overfitting while also prolonging the computational cost. In summary, the default values for the ntree and mtry parameters of 500 and N (i.e. where N denote the total number of descriptors), respectively, are within the recommended range of optimal values and are thus used for further analysis of the feature importance.

3.4. Importance of substructure fingerprints

The analysis of feature importance for each type of substructure fin-gerprint can provide a better understanding of β-globin modulators. Table 3 presents a list of substructure fingerprints and their descriptions that were utilized in the study. The efficient and effective built-in feature importance estimators of the RF method are utilized to identify informative features. Two measures, namely, mean decrease of Gini index (MDGI) and mean decrease of prediction accuracy, are generally available for ranking feature importance. Because the results of the MDGI measure are highly stable compared with the mean decrease of accuracy [27], the MDGI is adopted to rank features. To avoid the bias of random seed in evaluating feature importance, the average and standard deviation values of the MDGI on 50 runs of feature importance evaluations are used in the analysis.

The top 10 descriptors are SubFPC100, SubFPC1, SubFPC2, SubFPC101, SubFPC181, SubFPC135, SubFPC88, SubFPC5, SubFPC18 and SubFPC133, which can be defined as secondary amide, primary carbon, secondary carbon, tertiary amide, carboxyl derivative, carboxylic acid derivative, non-basic hetero nitrogen, alkyl aryl ether, alkene and nitrile, respectively.

Features with the top measure of Gini index are considered to be the most important. As shown in Fig. 7, the secondary amide ranked as the top feature. Note that compounds containing the amides hydroxycarbamide, the compound that works on reducing the imbalance between α- and β-globin gene [28], and isobutyramide, the compound that works on promoting the erythroid survival [29], are marketed as drugs for treating thalassemia and used as secondary and tertiary target levels, respectively. The analysis suggested that the amide functional group in the compounds is highly important in determining the activity of splice switching of the hemoglobin β-globin gene because it has the largest Gini index obtained from the built-in feature selector of RF. The second and third important features are primary carbon and secondary carbon, respectively. Note that carbon atoms, which are constituents of the basic forms of life, are important features. Organic compounds from both natural and synthetic sources have been major therapeutic agents for the treatment of various diseases. Our results suggested that the position of the carbon compound may play a role in determining the splice switching activity. The fourth most important feature is the tertiary amide. Note that amides are highly important in the determining the activity. This may be because the chemical properties of amides can change the conformation of the native DNA molecule [30], which may subsequently affect the transcription of DNA and translation of mRNA to protein. The fifth and sixth most important features are the carboxyl derivatives and carboxylic acid derivatives, respectively. The two substructures have a carbon atom attached to an oxygen atom through a double bond and an alcohol group. It has been shown that carboxylic acids can bind to DNA strands, as they have successfully been utilized as linkers to immobilize DNA with oligonucleotides [31]. This may be because helical DNA strands have a large number of hydrogen bond interactions, which may cause a hydrogen of DNA to interact with the oxygen of carboxylic acid derivatives, which has a high level of electronegativity. The seventh most important feature is the non-basic hetero nitrogen. Notably, the nitrogen atom and the nitrogen-containing substructure (i.e., amide and tertiary amide) are important in determining the activity of splice switching activity. In summary, it can be observed that the most important features come from the amide functional groups and the carboxylic acid groups. In summary, compounds containing amide functional groups and the carboxylic acid functional groups are important in deter-mining the activity of the splice switching, which can be used as a guide in the development of the novel splice switching compounds with high potency and selectivity.

4. Conclusion

Modulators of the Hb β-globin gene are important therapeutic agents for the treatment of thalassemia and other hemoglobinopathies. QSAR modeling was performed using the substructure fingerprint descriptors as an input to determine the substructure importance on the activity of modulators using the RF classifier, which provided excel-lent predictive ability with an accuracy of 89.50± 13.45, sensitivity of 94.97 ± 13.49, specificity of 84.29 ± 22.27 and MCC of 0.80 ± 0.25 for the internal validation set and an accuracy of 88.00±8.55, sensitivity of 87.89 ± 13.93, specificity of 87.51 ± 13.75 and MCC of 0.75 ± 0.18 for the external testing set. By utilizing the excellent built-in feature importance analysis parameters, three carbon-hetero bonds (carboxyl and derivatives) are shown to have significant weight in determining splice switching activity. Such insights can provide a better understanding of the origin of splice switching activity of the Hb β-globin gene and may be used as a general guideline for designing novel modulators.

References

[1] M.G. Olsson, M. Allhorn, L. Bülow, S.R. Hansson, D. Ley, M.L. Olsson, A. Schmidtchen, B. Åkerström, Pathological conditions involving extracellular hemoglobin: mo-lecular mechanisms, clinical significance, and novel therapeutic opportunities for α1-microglobulin, Antioxid. Redox Signal. 17 (2012) 813–846.

[2] R. Kole, T. Williams, L. Cohen, RNA modulation, repair and remodeling by splice switching oligonucleotides, Acta Biochim. Pol. 51 (2004) 373–378.

[3] H. Sierakowska, M.J. Sambade, S. Agrawal, R. Kole, Repair of thalassemic human beta-globin mRNA in mammalian cells by antisense oligonucleotides, Proc. Natl. Acad. Sci. USA 93 (1996) 12840–12844.

[4] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, A practical overview of quantitative structure–activity relationship, EXCLI J. 8 (2009) 74–88.

[5] C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Advances in computational methods to predict the biological activity of compounds, Expert Opin. Drug Discovery 5 (2010) 633–654.

[6] P. Mandi, W. Shoombuatong, C. Phanus-umporn, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, C. Nantasenamat, Exploring the origins of structure–oxygen affinity relationship of human haemoglobin allosteric effector, Mol. Simul. 41 (2015) 1283–1291.

[7] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, S.H. Bryant, PubChem: a public information system for analyzing bioactivities of small molecules, Nucleic Acids Res. 37 (2009) W623–W633.

[8] D. Fourches, E. Muratov, A. Tropsha, Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research, J. Chem. Inf. Model. 50 (2010) 1189–1204.

[9] C. Standardizer, Version 5.4. 4.1, ChemAxon, Budapest, Hungary, 2010.

[10] N.M. OLBoyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminf. 3 (2011) 33.

[11] J. Wikberg, M. Eklund, E. Willighagen, O. Spjuth, M. Lapins, O. Engkvist, J. Alvarsson, Introduction to pharmaceutical bioinformatics, Oakleaf Academic, 2010.

[12] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474.

[13] J. Klekota, F.P. Roth, Chemical substructures that enrich for biological activity, Bioinformatics 24 (2008) 2518–2525.

[14] M.T. Cronin, T.W. Schultz, Pitfalls in QSAR, J. Mol. Struct. 622 (2003) 39–51.

[15] G.D. Booth, M.J. Niccolucci, E.G. Schuster, Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation, Research Paper INT-470, United States Department of Agriculture, Forest Service, Ogden, USA, 1994.

[16] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. 28 (2008) 1–26.

[17] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a k-means clustering algorithm, J. R. Stat. Soc.: Ser. C: Appl. Stat. (1979) 100–108.

[18] K. Hornik, C. Buchta, A. Zeileis, Open-source machine learning: R meets Weka, Comput. Stat. 24 (2009) 225–232.

[19] A. Liaw, M. Wiener, Classification and regression by randomforest, R. News 2 (2002) 18–22.

[20] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: misc functions of the department of statistics (e1071), TU Wien, R Package Version 1.6-4, 2014.

[21] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, fourth ed. Springer, New York, 2002 (ISBN 0-387-95457-0).

[22] H. Wickham, R. Francois, dplyr: A grammar of data manipulation, R Package Version 0.4.1, 2015.

[23] S. Simeon, W. Shoombuatong, L. Preeyanon, V. Prachayasittikul, C. Nantasenamat, Predicting the oligomeric states of fluorescent proteins, PeerJ PrePrints 3 (2015) e1139.

[24] N.M. Razali, Y.B. Wah, Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests, J. Stat. Model. Anal. 2 (2011) 21–33.

[25] A.V. Zakharov, M.L. Peach, M. Sitzmann, M.C. Nicklaus, Qsar modeling of imbalanced high-throughput screening data in pubchem, J. Chem. Inf. Model. 54 (2014) 705–712.

[26] J. Alvarsson, M. Eklund, C. Andersson, L. Carlsson, O. Spjuth, J.E. Wikberg, Benchmarking study of parameter variation when using signature fingerprints together with support vector machines, J. Chem. Inf. Model. 54 (2014) 3211–3217.

[27] M.L. Calle, V. Urrea, Letter to the editor: stability of random forest importance measures, Brief. Bioinform. 12 (2011) 86–89.

[28] M. Bradai, M.T. Abad, S. Pissard, F. Lamraoui, L. Skopinski, M. de Montalembert, Hydroxyurea can eliminate transfusion requirements in children with severe β-thalassemia, Blood 102 (2003) 1529–1530.

[29] M.D. Cappellini, G. Graziadei, L. Ciceri, A. Comino, P. Bianchi, A. Porcella, G. Fiorelli, Oral isobutyramide therapy in patients with thalassemia intermedia: results of a phase II open study, Blood Cells Mol. Dis. 26 (2000) 105–111.

[30] K.S. Gates, T. Nooner, S. Dutta, Biologically relevant chemical reactions of N7-alkylguanine residues in DNA, Chem. Res. Toxicol. 17 (2004) 839–856.

[31] T. Heyduk, E. Heyduk, Molecular beacons for detecting DNA binding proteins, Nat. Biotechnol. 20 (2002) 171–176.

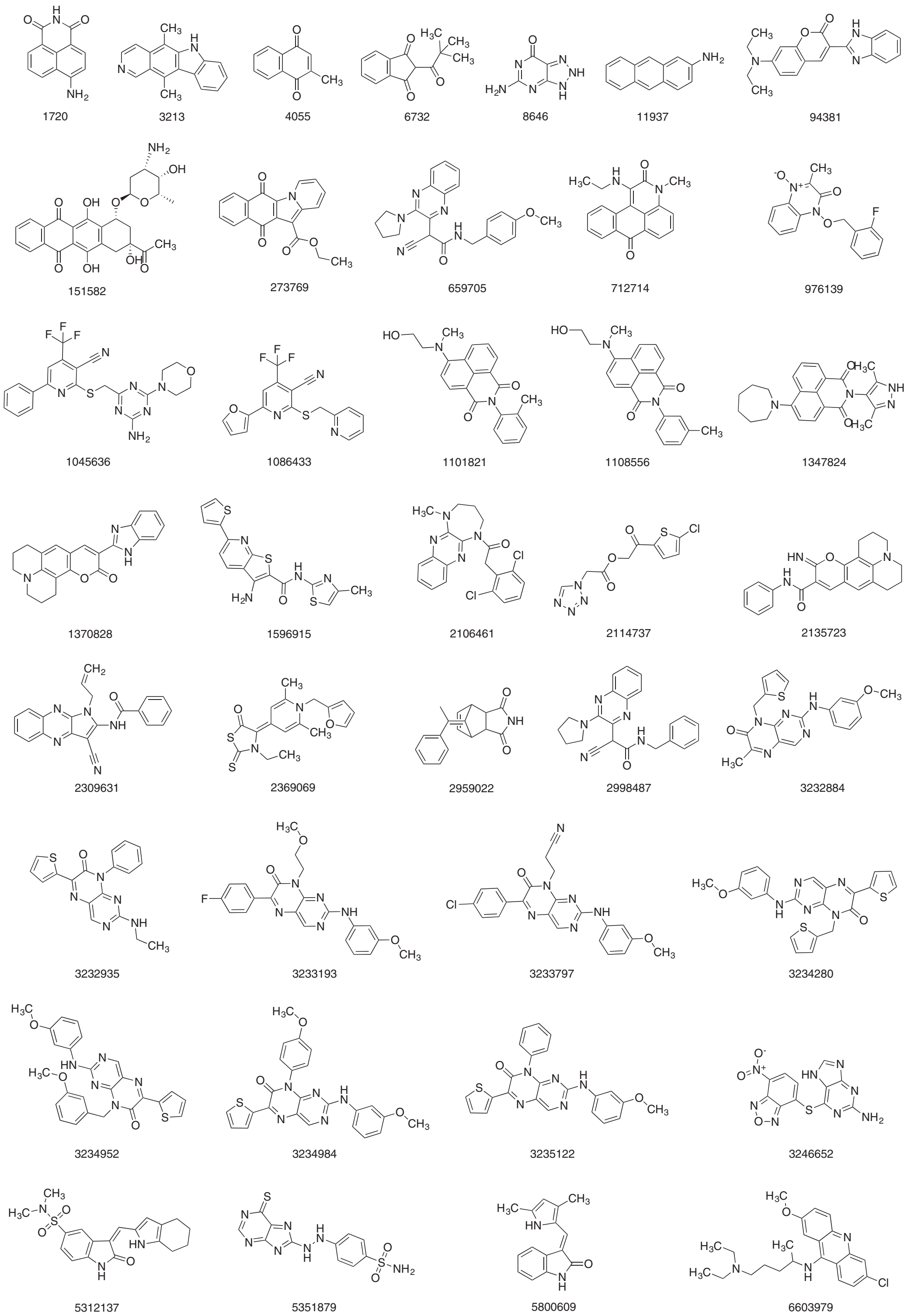


Figure 1. Chemical structures of the set of 39 active compounds with splice switching activity. PubChem CID numbers are shown beneath the chemical structures.

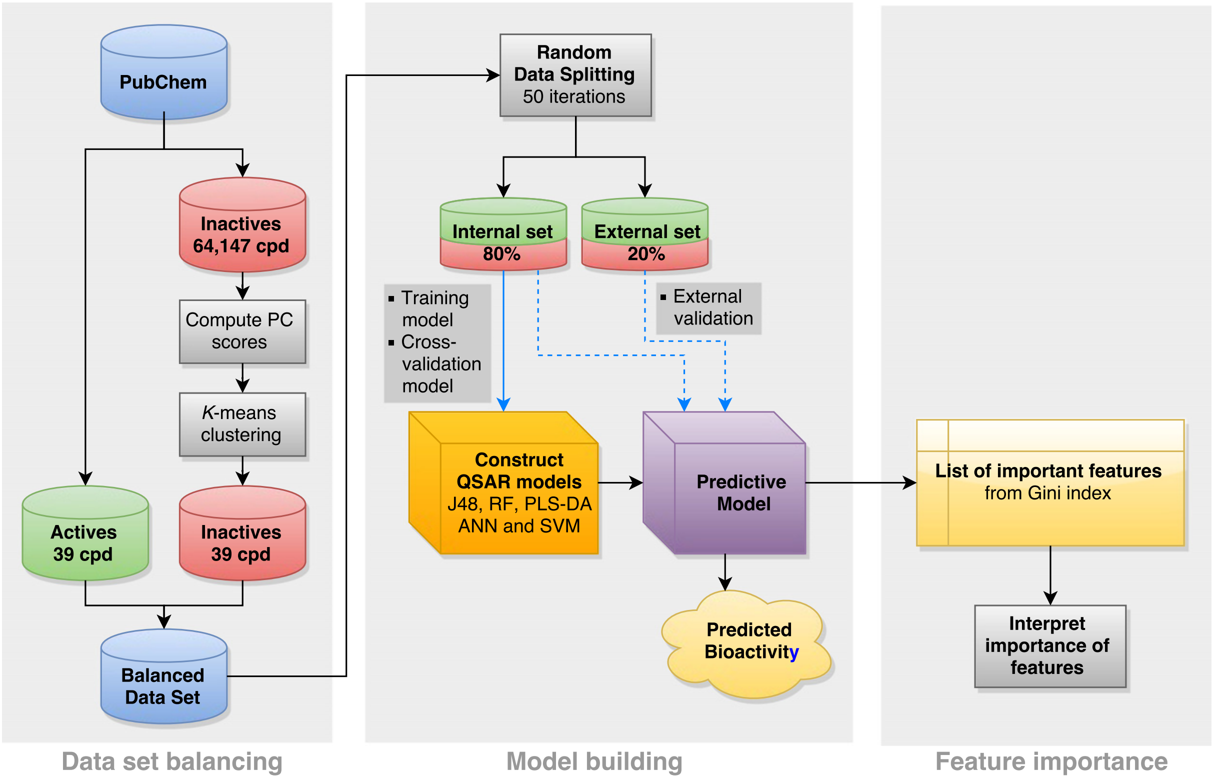


Figure 2. Workflow for the QSAR modeling of the hemoglobin β-globin gene modulators.

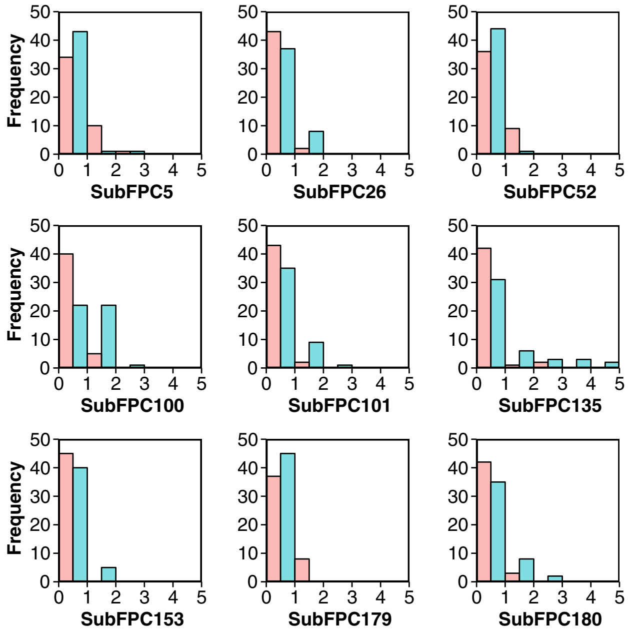


Figure 3. Histogram of substructure fingerprint counts of active (colored in pink) and inactive (colored in cyan) splice switching modulators.

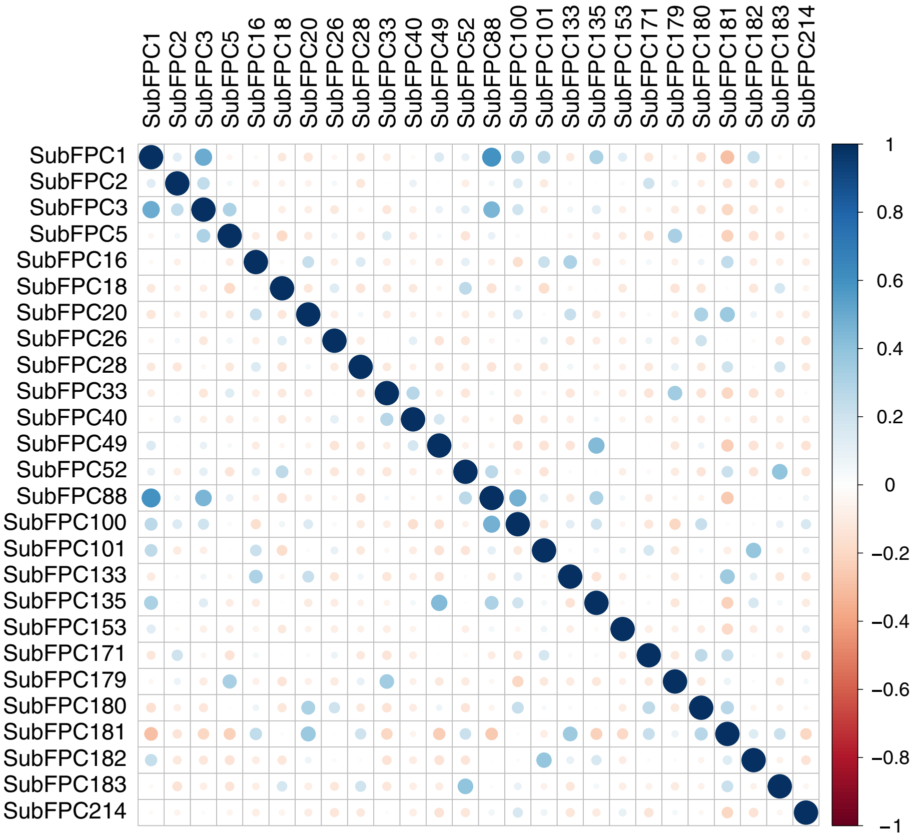


Figure 4. Intercorrelation matrix of the descriptors utilized for constructing the predictive models.

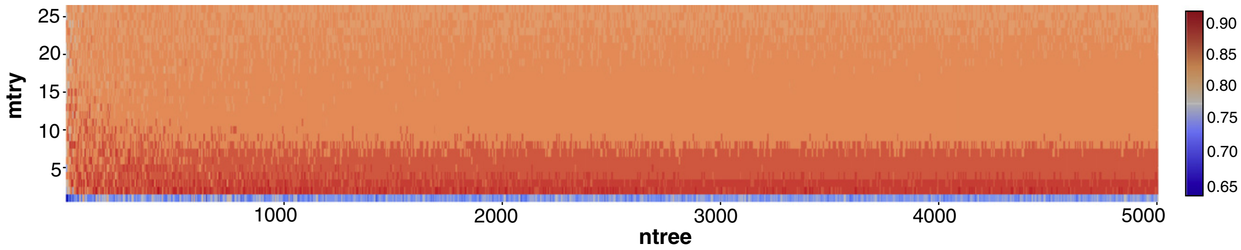


Figure 5. Heat map of fine tuning the RF parameters ntree and mtry as shown in the X and Y axes, respectively. Accuracy obtained from 10-fold CV is shown in the plot and color-coded according to their performance ranging from low (blue) to high (red) accuracy.

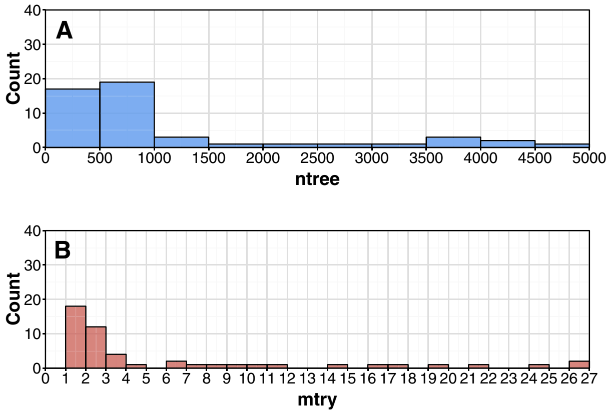


Figure 6. Histogram showing the binned distribution of optimal ntree (A) and mtry (B) values as obtained from model building from 50 independent data splits.

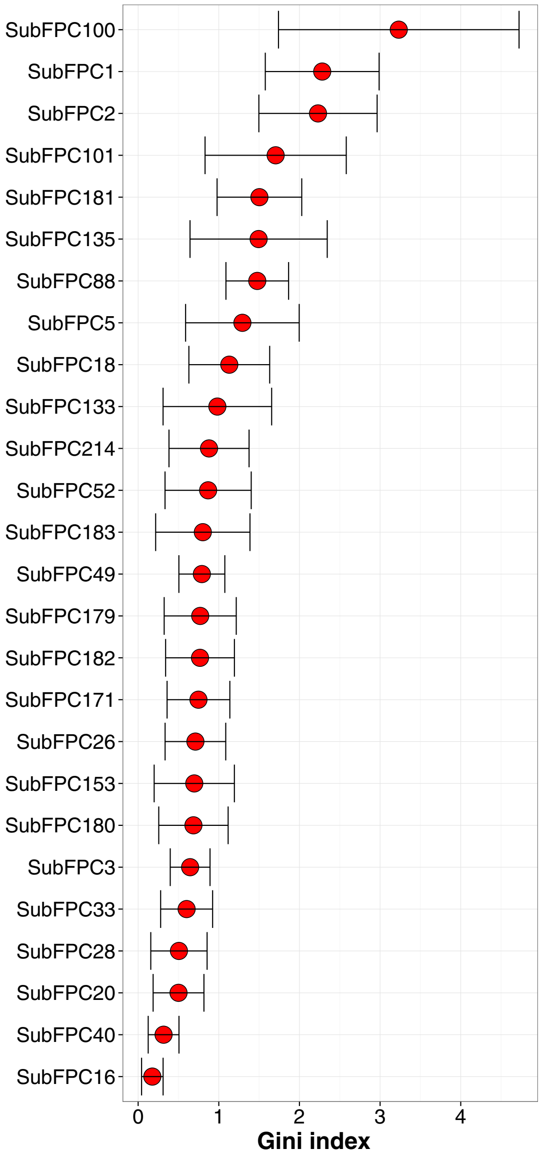


Figure 7. Importance of substructure fingerprints as a function of the Gini index. Features with the largest Gini index are deemed the most important.

Table 1. Summary of the mean and standard deviations of substructure fingerprints along with their P values derived from the statistical difference test using the Mann-Whitney U test.

	Active	Inactive	P value
SubFPC1	0.667 ± 0.929	1.533 ± 2.351	0.081
SubFPC2	0.756 ± 1.384	1.578 ± 3.258	0.223
SubFPC3	0.156 ± 0.638	0.267 ± 0.780	0.341
SubFPC5	0.267 ± 0.495	0.067 ± 0.330	0.009
SubFPC16	0.044 ± 0.208	0.044 ± 0.208	1.000
SubFPC18	0.289 ± 0.589	0.511 ± 0.944	0.378
SubFPC20	0.067 ± 0.252	0.156 ± 0.520	0.452
SubFPC26	0.044 ± 0.208	0.178 ± 0.387	0.046
SubFPC28	0.133 ± 0.344	0.022 ± 0.149	0.051
SubFPC33	0.178 ± 0.387	0.089 ± 0.288	0.220
SubFPC40	0.089 ± 0.288	0.022 ± 0.149	0.173
SubFPC49	0.267 ± 0.751	0.244 ± 0.529	0.509
SubFPC52	0.200 ± 0.405	0.022 ± 0.149	0.008
SubFPC88	1.067 ± 0.780	1.244 ± 1.836	0.614
SubFPC100	0.111 ± 0.318	0.533 ± 0.548	0.000
SubFPC101	0.044 ± 0.208	0.244 ± 0.484	0.013
SubFPC133	0.156 ± 0.367	0.044 ± 0.208	0.082
SubFPC135	0.111 ± 0.438	0.644 ± 1.151	0.003
SubFPC153	0.000 ± 0.000	0.111 ± 0.318	0.023
SubFPC171	0.156 ± 0.475	0.200 ± 0.457	0.416
SubFPC179	0.178 ± 0.387	0.000 ± 0.000	0.003
SubFPC180	0.067 ± 0.252	0.267 ± 0.539	0.034
SubFPC181	1.133 ± 1.217	0.844 ± 1.127	0.239
SubFPC182	0.089 ± 0.288	0.178 ± 0.442	0.327
SubFPC183	0.267 ± 0.618	0.067 ± 0.252	0.060
SubFPC214	0.067 ± 0.252	0.200 ± 0.405	0.065

Publication 3

Predicting the oxygen affinity of human hemoglobin

Abstract

Human hemoglobin (Hb) is instrumental in the transportation of oxygen (O_2) from lungs to tissues. In spite of several decades of investigations, the structural basis of O_2 binding has yet to be fully elucidated. Therefore, a comprehensive study on the physicochemical properties contributing to O_2 binding was performed herein on a large set of 326 non-redundant human Hb variants harboring a single point mutation on the a or b chains. Statistical and multivariate analyses were performed to gain insights into the origins of low and high O_2 binding affinities in human Hb. This study investigated the use of several classifiers (i.e. C5.0 decision tree, random forest, partial least squares discriminant analysis, artificial neural network and support vector machine) for predicting the O_2 affinity of Hb variants as a function of their sequence-based *z*-scale descriptors. Prediction performance of the constructed models were found to be robust as supported by accuracy, sensitivity and specificity greater than 80% and Matthews correlation coefficient greater than 0.6. Interpretation of the predictive models was also performed to gain useful insights on the underlying physicochemical properties giving rise to O_2 binding affinities, which may further guide structure-based design of novel human Hb variants with desired O_2 binding characteristics.

Publication 4

Origin of anti-sickling activity via QSAR modeling

Abstract

Sickle cell disease (SCD) is an autosomal recessive genetic disorder that has been recognized as a major public health problem by the WHO affecting 300,000 individuals worldwide. SCD arises from the A→T point mutation that causes the Glu6Val mutation in the hemoglobin β -globin gene thereby leading to sickle hemoglobin (HbS). At low oxygen tension, HbS are polymerized inside the red blood cells leading to gel or fiber formation thereby causing drastic decrease in the red cell deformability. Consequently, the complications of SCD leads to serious conditions such as anemia, microvascular occlusion, severe pain, stokes, renal dysfunction and infections. A lucrative therapeutic strategy to remedy complications of SCD affected patients is to employ anti-sickling agents for disrupting the HbS polymer. Therefore, this study aims to use quantitative structure-activity relationship (QSAR) modeling to elucidate the anti-sickling activity of 115 com- pounds. Briefly, the bioactivity of compounds were measured by a solubility assay described by Hofrichter et al. in which compounds were defined as active if their solubility ratios were greater than 1.06 and inactive if their solubility ratios were less than 1.06. Compounds were described by substructure descriptors and used in the construction of QSAR models via the random forest (RF) algorithm using rigorous validation. Good predictive performance was obtained as deduced by their accuracy (Ac), sensitivity (Sn), specificity (Sp) and Matthews correlation coefficient (MCC). Results indicated that Ac, Sn, Sp and MCC was in excess of 0.7 for the former three and greater than 0.5 for the latter statistical parameter. In addition, it was found that the top 5 important substructure descriptors for anti-sickling included conjugated double bond, arylchloride, Michael acceptor, alkene and vinylogous halide. Thus, this model is anticipated to be useful for guiding the design of robust compounds against the gelling activity of HbS.

Publication 5

QSAR modeling of methemoglobin reduction by electron mediators

Abstract

Hemoglobin vesicle (HbV) is an artificial oxygen carrier based on liposome-encapsulated hemoglobin which can potentially be used as an alternative source of blood. The advantages of HbV is that it can reduce the risk of infection and blood type mismatching while being easily excreted and has long shelf-life. However, the limitation of HbV lies in the auto-oxidation of ferrous Hb thereby leading to increased level of ferric methemoglobin (metHb). This condition leads to impaired oxygen transport and reduces the half-life of HbV. Previously, Kettisen *et al.* have found that electron mediator dyes can reduce metHb production in HbV. Therefore, this study aims to use quantitative structure-activity relationship (QSAR) modeling to elucidate the origin of electron mediation amongst the set of 15 dyes. These molecules were described by substructure fingerprints, molecular descriptors and quantum chemical descriptors and subsequently used in the construction of QSAR models using support vector machine, which afforded good predictive performance with accuracy, sensitivity, specificity and Matthew's correlation coefficient (MCC) of 99.00±3.16 %, 98.57±4.52 %, 100.00±0.00 % and 0.98±0.06, respectively, for leave-one-out cross validation as well as values of 98.00±6.32 %, 96.67±10.54 %, 100.00±0.00 % and 0.97±0.11, respectively, for external validation. It was found that important descriptors for electron mediation of metHb included HATS4m, Mor18m, R7e+, R7m+ and Mor12e, which corresponded to atomic mass and electronegativity of the molecule. Thus, this model is anticipated to be useful for guiding the design of robust compounds against the auto-oxidation of HbV.

Publication 6

Structural and biochemical characterization of two heme binding sites on α_1-microglobulin using site directed mutagenesis and molecular simulation

1. Introduction

The Lipocalin protein family includes approximately 50 members from bacteria, plants and animals (1-3). Although their structures are similar they have different and mostly non-related functions. A common feature is the folding of the single polypeptide into a β-barrel consisting of eight antiparallel β-strands with a closed bottom, an open end and a hydrophobic pocket which can carry small lipophilic ligands (4). One of the members of the Lipocalin protein family, D1-microglobulin (A1M)#, is a small (26 kDa), heterogeneously charged glycoprotein found in blood-plasma and interstitial fluid of all human tissues (5-7). A1M is expressed mainly in the liver and co-synthesized with bikunin, a proteinase inhibitor and component of extracellular matrix, from the AMBP-gene (D1-microglobulin-bikunin precursor protein) (8,9). After cleavage of the precursor protein the two mature proteins are secreted separately into the blood-stream (10). A large fraction of A1M forms complexes with other plasma proteins in blood. Approximately 50% of the protein is bound to immunoglobulin A (IgA) (11). Less abundant complexes are formed with albumin and prothrombin (12).

The physiological function of A1M has been suggested to be protection of cells and tissues against oxidative stress induced by extracellular hemoglobin and free radicals (7). This is supported by several recent papers, which demonstrate that A1M indeed can protect cell cultures and organ explants against oxidative damage (13-15), and therapeutic in vivo effects of the protein were shown in animal models of preeclampsia and hemoglobin-induced kidney damage (16,17). Mechanistically, the antioxidative protection is achieved by reductase and radical-binding activities of A1M, and it was shown that the C34 unpaired thiol group and the side-chains of K92, 118 and 130 of A1M are involved in the reductase and radical-binding activities (18-20). A1M has also been shown to bind heme≠ in vitro and in vivo (21-23) and was suggested to contribute to heme degradation by a mechanism still unknown, but which involves proteolytic activation of A1M (21). The heme binding of A1M was recently demonstrated to result in a trimeric complex with heme in a 1:2 stoichiometry (24). Furthermore, a heme binding site, which involves the C34 thiol group and H123 imidazole ring, has been proposed based on the crystal structure of A1M (25).

Free heme in plasma and extracellular fluid is the result of degradation of hemoglobin and other proteins carrying a heme prosthetic group. The heme-group presents a potential chemical threat to the organism by generation of reactive oxygen species (ROS), which are toxic and can cause severe damage to cells and tissues (26). We hypothesized that the above-mentioned side-chains of A1M are involved in the heme binding. To test this, we investigated wild type (Wt)-A1M and the previously constructed mutated forms of A1M: C34S- and K(3)T-A1M (27), the latter carrying Lys→Thr substitutions in positions 92, 118 and 130. We also constructed, prepared and investigated a new mutated form, H123S-A1M (Fig. 1A). The results suggest a two binding site-model where the three lysine side-chains participate in coordination of the first heme-group and the C34 and H123 side-chains coordinate the second heme-group.

2. Materials and Methods

2.1. Reagents and proteins

Heme (hemin; ferriprotoporphyrin IX chloride) was purchased from Porphyrin Products, Inc. (Logan, UT, U.S.A.). Stock solutions of heme were prepared by dissolving heme in DMSO to 10 mM and used within 10 h. Clarity Western ECL Substrate was from Bio-Rad Bio-Rad Laboratories (Hercules, CA, USA). Heme-agarose was from Sigma-Aldrich, Sweden. Sepharose CL-4B was from GE Healthcare, Sweden. Pierce BCA Protein Assay kit from Thermo Scientific, Sweden. AcroPrep Advance 96 Filter Plate 1.2µM supor® was from PALL Corporation (Port Washington, NY, USA), and Nunc TM 96-Well Microplates were purchased from Thermo Scientific.

2.2. Recombinant A1M

Wild-type (Wt) and mutated variants of A1M were expressed in Escherichia coli (E. coli) as described (27). Using site-directed mutagenesis a Cys → Ser substitution was introduced at amino acid position 34 in the C34S-A1M mutant, Lys → Thr substitutions at positions 92, 118, and 130 in the K(3)T-A1M mutant, and a His → Ser substitution at position 123 in the H123S-A1M mutant. The four forms of recombinant A1M, Wt-A1M, C34S-A1M, K(3)T-A1M and H123S-A1M, were purified and refolded as described (27) with the addition of ion-exchange chromatography and size exclusion purification steps as follows. The protein solution was applied to a column of DEAE-Sephadex A-50 (GE Healthcare, Uppsala, Sweden) equilibrated with the starting buffer (20 mM Tris-HCl, pH 8.0). A1M was eluted at a flow rate of 1 ml/min using a linear pH gradient consisting of 250 ml starting buffer and 250 ml elution buffer (20 mM Tris-HCl, 0.5 M NaCl, pH 8.0). Size-exclusion chromatography was run on a Superose 12 column obtained from GE Healthcare using Äkta purifier 10 system (GE Healtcare) run at a flow-rate of 1 ml/min. Wt-A1M without the N-terminal His8-tag was a generous gift from A1M-Pharma AB.

2.3. Secondary structure estimation by far-UV circular dichroism

The secondary structure of A1M variants was determined with a Jasco-J810 spectropolarimeter instrument using a 2mm quartz cuvette at continuous mode at speed 20nm/min, band width of 1nm and a 1nm resolution. Temperature was held at 22°C by a Peltier thermostat. The concentration of protein was 10µM in 20 mM Tris-HCl pH 8.0 + 0.15M NaCl. Each spectrum is a mean of five replicates. After subtraction of the buffer spectrum and recalculation into mean residue molar elipticity units, the CD spectra were analyzed using the CDPro package to determine different secondary structure content (28).

2.4. Spectrophotometric analysis of heme binding

Absorbance spectra were measured on a Beckman (Beckman Instruments, Fullerton, CA) DU 800 spectrophotometer using a scan rate of 1200 nm/min in the UV-VIS region between 250 and 700 nm at 22°C. The protein concentration was 10 µM in 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl, which was also used as blank. Heme was dissolved in DMSO to 10 mM and diluted in 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl to 3 mM. Volumes of this stock solution were then added to each protein solution, to a final concentration 20 µM. Blanks with equivalent concentration of DMSO was used for samples with heme. Protein solutions were scanned immediately after mixing with heme and after 1h or 24h.

2.5. Fluorescence spectroscopy

Fluorescence measurements were performed using a Jasco J-810 spectropolarimeter (equipped with a FMO-427 monochromator) with a 100 µl quartz cuvette (Hellma Precision Cell, Type no. 105.251-QS, light-path length 3mm in both excitation and emission modes) under nitrogen flow. Temperature was held at 22°C by a Peltier thermostat. Tryptophan fluorescence was measured by exciting at wavelength 295 nm. Fluorescence emission was detected from 310 to 400 nm with slits set at 5 nm bandpass. Emission spectra were recorded three times, averaged, and the peak at 346 nm was measured. A1M protein concentration was 1µM in 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl. At this concentration, the fluorescence signal of the protein was well resolved within the detector sensitivity (set at 900 V). The emission spectra of the protein solution (100 µl) without heme were recorded first and subsequently heme from a stock-solution was added (1µl at a time, up to 4 µl). Blank subtractions were made for Tris-HCl, pH 8.0, 0.15 M NaCl since the concentration of DMSO did not affect the emission spectra.

2.6. Induced circular dichroism

All samples containing heme were evaluated for the protein-induced chirality in the near-UV to visible CD range of 300-700 nm (referred to as visible CD). The visible CD measurements were performed on a Jasco J -815 Spectropolarimeter (JASCO Co., Japan) with the temperature maintained at 25± 0.5 oC. The spectra were recorded using a scan speed of 100 nm/min, bandwidth of 1.0 nm, and resolution of 0.2 nm, and accumulated in triplicate. A protein alone (Wt-A1M or mutant, respectively) was used as a blank. To accumulate the induced CD, A1M was used at an initial protein concentration of ~ 45µM, and the measurements were performed in a quartz cuvette with 1 cm pathlength. Heme content in the samples varied from ~ 4.5 µM to ~ 50 µM by using calculated amounts of the heme 2 mM stock solution in DMSO. The content of DMSO in the mixed samples did not exceed 4%. An ellipticity of induced CD spectra was expressed in millidegrees (mdeg).

2.7. Analytical size exclusion chromatography

Samples were size-fractionated on a Superose 12 10/300 fast protein liquid chromatography (FPLC) column (GE Healthcare, Uppsala, Sweden). A1M/heme molar ratio was 1:2. Heme was dissolved in DMSO to 10 mM and added from a 3 mM stock solution prepared as described above. 500 µl of sample was applied to the column after 1h of incubation with heme. The proteins were eluted with 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl and run at 0.5 ml/min. Fractions of 0.5 ml were collected.

2.8. Native PAGE and Western blot of A1M incubated with heme

A1M (5µM) in 20 mM Tris-HCl, pH 8.0 + 0.15 M NaCl was incubated for 30 min or 24h with 50, 10 and 0.1 µM heme. Samples were mixed with equal amounts of sample buffer for native PAGE, pH 6.8, and subjected to 12% CriterionTM TGXTM Precast Gels (Bio-Rad). The gels were either stained with Coomassie Brilliant Blue R-250 (BDH Chemicals, Ltd. Poole, UK) or transferred to polyvinylidene difluoride (PVDF) membranes using Trans-Blot Turbo system from Bio-Rad. The membranes were incubated in Clarity Western ECL Substrate and imaged with a digital imager (BioRad).

2.9. Binding of A1M to heme-agarose

Beads of heme-agarose and Sepharose were washed three times with an excess of 10mM Tris-HCl pH 8.0 + 0.125M NaCl, yielding a final 1:1 suspension in this buffer. The proteins were diluted in 10mM Tris-HCl pH 8.0 + 0.125M. Dilution series of all proteins were made to final concentrations of 10, 7.5, 5, 2.5, 1.25, and 0.625 µM. Seventy-five µl were then transferred to Nunc TM 96-Well Microplates in duplicates (one set for incubation with heme agarose and control Sepharose and one without beads). Twenty µl 50% heme-agarose or Sepharose were pipetted into the wells, using large-opening pipette-tips, and the plates were incubated in RT on a shaker for 30 minutes. The protein/beads mixtures were then transferred to AcroPrep Advance 96 Filter Plate and spun at 2000g for 2 minutes. Twenty-five µl of non-incubated samples and 25µl of each flow-through were transferred to a new Nunc TM 96-Well Microplate, 200µl of BCA reagent was added and the plates incubated at 37°C for 30 minutes. The absorbance was measured at wavelength 550 nm, using a Multilabel Counter (Victor™ 1420 Perkin Elmer Life and Analytical Sciences, Turku, Finland). Statistical analysis was performed using OriginPro 9.0 software (Microcal, Northampton, MA, USA).

2.10. Catalase-like activity assay

Monitoring of the Soret band intensity of heme equimolar samples with A1M mutants (C34S, K(3)T, H123S) after adding hydrogen peroxide was performed in comparison with Wt-A1M and free heme in Tris buffer according to the procedure described earlier (29) with minor modifications. Equimolar (L/P 1.0) heme/A1M samples were prepared by adding 20 µl of heme stock solution in DMSO (2.2 mM) to 1 ml of 45 µM solutions of A1M or each of the A1M mutants. After overnight (~20 hours) incubation at room temperature in dark, the samples were evaluated by UV/Vis measurements and diluted by buffer to adjust the Soret band intensity to ~ 0.6 AU. Heme sample in Tris buffer (with approximately the same absorbance intensity) was freshly prepared and used immediately. These UV/Vis spectra served as initial (zero time) baseline. After a 7 µl aliquot of 50 mM hydrogen peroxide was added to each sample, the time course of spectral changes were measured at 30s, 1 min, 2 min, 4 min, 6 min, 8 min and 10min.

2.11. Surface plasmon resonance (SPR)

SPR experiments were conducted on Biacore T200 (GE Healthcare, Piscataway, NJ). Anti-His mouse IgG1 monoclonal antibodies (R&D Systems, Minneapolis, MN) were immobilized on CM5 sensors by amine coupling, ~18,000 response units (RU). The tagged proteins were injected at a flow of 10 µl/min for 360 s. Prior to injection, a freshly prepared heme solution in DMSO was subjected to serial double dilutions using PBS buffer to create a range of eight heme concentrations, from 100 µM down to 0.625 µM.

Heme preparations were injected over captured A1M variants for 2 min with a flow 30
µl/min at 25°C, and the association was recorded, followed by the dissociation monitored during 10 min. Data were analyzed using the Biacore T200 evaluation software (GE Healthcare), subtracting the reference surface and buffer control signals from each curve. Data were globally fitted by simultaneous numerical integration to the association and dissociation parts of the interaction, using the heterogeneous ligand kinetic analysis models (T200 BIAevaluation software, version 1.0; Biacore AB, Uppsala, Sweden).

2.12. Molecular simulation of A1M-heme binding

A representative molecular structure of heme (accession number HEB) was obtained from PDBeChem (30) as a MOL file. Molecular volume was subsequently computed from the MOL file of heme b via the online tool from Molinspiration (31). Briefly, the method for computing the molecular volume is based on group contributions in which the sum of fragment contributions were fitted to actual three-dimensional molecular volume of a training set comprising twelve thousand molecules (32). The three-dimensional molecular structures of the training set were geometrically optimized using the semi-empirical Austin Model 1 (AM1) method.

In reconstructing the wild-type structure of A1M, residue 34 of the crystal structure (PDB ID: 3QKG) was mutated from serine to cysteine using PyMOL (33). This was performed whereby the rotamer was independent of the backbone such that the sulfhydryl moiety is pointed towards the imidazole moiety of H123. Molecular volume of the lipocalin pocket in the crystal structure of A1M was computed using CASTp (34).

A query on Protein Data Bank for structures having the lipocalin SCOP fold yielded 280 hits, of which 46 contained a bound heme. From this, 45 are either nitrophorin 1, 2 or 4 from Rhodnius prolixus (PDB ID: 1D2U, 1D3S, 1EUO, 1IKE, 1IKJ, 1KOI, 1NP1, 1NP4, 1PEE, 1PM1, 1SXU, 1SXW, 1SXX, 1SXY, 1SY0, 1SY1, 1SY2, 1SY3, 1T68, 1U0X, 1U17, 1U18, 1X8N, 1X8O, 1X8P, 1X8Q, 1YWA, 1YWB, 1YWC, 1YWD, 2A3F, 2ACP, 2AH7, 2AL0, 2ALL, 2AMM, 2ASN, 2AT0, 2EU7, 2GTF, 2HYS, 2NP1, 3C76, 3NP1, 4NP1) while the other was a nitrophorin-like protein from Arabidopsis thaliana (PDB ID: 3EMM). The former set constitutes amino acid length of 179-184 while the latter structure had 153 residues. Furthermore, owing to the fact that the former set of 45 structures spanned similar length, it was selected for further analysis. Structural alignment was performed using MultiProt (35).

The structure of wild-type A1M was docked to heme using HADDOCK (36). A second heme was subsequently docked to a putative heme binding site, which is formed by axial ligands comprising of C34 and H123 as proposed by Meining and Skerra (25) of the top-scoring model using PyMOL. The structure of A1M-heme complex was then refined using the energy-minimization in gas phase followed by molecular dynamics (MD) simulation on GROMACS, version 4.0 (37). During the refinement, the distance restraints of the side chain of C34 and H123 with the second heme were applied, and inappropriate bond lengths of any atom were fixed. The simulation was performed on the explicit-solvent periodic boundary conditions under the NPT condition at 300 K of temperature using the modified Berendsen thermostat (38) and 1 bar of pressure using Parrinello-Rahman barostat (39). GROMOS96-53A6 force field was applied for both protein and heme structures and the ionization states of amino acid residues were set according to the standard protocol (40). The SPC water was used as a solvent model. Bond lengths were constrained using LINCS algorithm that allows for a 2.0 fs-time step (41). A cut-off distance for the short-range neighbor list was set to 0.9 and 1.4 nm for the electrostatic and van der Waals interactions, respectively. Long-range electrostatic interactions are approximated using PME method (42). The same MD protocol was also applied for further analyzing dynamic properties of A1M-heme complex. For the data collection, atomic coordinates were recorded every 10 ps.

3. Results

3.1. Expression and characterization of A1M-mutants

The mutated side-group residues are high-lighted in Fig. 1A. Before measuring the heme binding of the mutated A1M-variants, purity and basic structural properties were analyzed by SDS-PAGE, far-UV CD spectroscopy and optical fluorescence (Fig. 2). As shown in Fig. 2A, no visible impurities could be detected. Similar apparent sizes were seen, as expected from the theoretical masses of the four variants (Wt-A1M: 22.64 kDa, C(34)S: 22.66, K(3)T: 22.56 and H(123)S: 22.59). Far-UV CD spectra of Wt-A1M and the mutated A1M-forms suggest a similar composition of secondary structure, i.e. mostly β-structure (Fig. 2B; Table I). The CD spectra of the four A1M-variants were also consistent with the X-ray crystallography-derived three-dimensional structure of Wt-A1M (25). Four tryptophan residues are found in A1M located at various positions (Fig. 1B) and tryptophan fluorescence spectra were recorded as an estimate of the overall conformation of the mutants (Fig. 2C). A higher intensity was obtained from the C(34)S-mutant, possibly as a result of an interaction between the closely located C34- and W36-residues, reducing the intensity of the fluorescence of W36 in Wt-A1M and the other mutants (Fig 1B). Apart from this, similar spectra were obtained, suggesting no major differences in the overall conformation of the four variants.

3.2. All A1M-variants bind heme

The binding of the heme-group to the A1M-variants was first investigated by quenching of the tryptophan fluorescence during heme-titration, using a protein:heme ratio of 0.4-20 (Fig. 3A,B). The results of this experiment show that heme was bound to all variants with a similar binding strength. Quenching of 50% of the tryptophan fluorescence of 1 μM A1M was achieved in the 0.8 μM range in all cases and approximately 70% of the tryptophan fluorescence was quenched by 2.5 μM heme. Furthermore, the binding of heme to the A1M variants was analyzed using heme-agarose titration (Fig. 4) and surface plasmon resonance

(SPR) (Table II). Both techniques confirm the binding of heme by all four A1M-variants, and indicate a slightly higher binding by Wt-A1M compared to the mutated variants. Heme-agarose titration with increasing amounts of the A1M variants (Fig. 4) yielded binding responses in the order Wt-A1M > C(34)S-A1M > H(123)S-A1M and K(3)T-A1M. Ovalbumin, a negative control protein, did not bind at all to heme-agarose and Wt-A1M did not bind to un-conjugated Sepharose. The SPR data analysis, within heme concentrations 0.625-100 μM (Fig. 5), resulted in dissociation constant (KD) values of 13.82 x10-6 (Wt-A1M), 11.81x10-6 (C34S), 9.65x10-6 [K(3)T] and 11.74x10-6 (H123S) (Table II), confirming the results of heme-agarose titration. Although these kinetic data fitted to a 1:1-binding model, SPR data obtained for an extended range of heme concentration (0.625-500 μM) fitted better to a 1:2 binding model, supporting the earlier proposed two binding site model (24). The latter should be interpreted with care, however, since low heme solubility and its increased aggregation at high concentrations greatly limit the reliability of precise determination of kinetic constants for the low affinity binding sites (43).

3.3. Oligomerization of A1M-heme complexes

The binding of heme was studied by gel-shift assay, using native PAGE, to analyze the electrophoretic mobility of the A1M-variants alone or in the presence of 0.1, 10 or 50 μM heme (Fig. 6A). A clear migration shift was seen of all four variants and the dose-dependence was similar. These results support the findings described above, i.e. binding of heme with similar strength by all A1M-variants. A similar migration shift was seen with Wt-A1M without the N-terminal His-tag, and no migration shift of the control proteins α1-acid glycoprotein and ovalbumin after heme-incubation, or of Wt-A1M in the presence of the carrier DMSO, could be seen (data not shown). To visualize the heme-group in the A1M-bands, the gels were analyzed by peroxidase-activity (ECL)-blotting (Fig 6B). All A1M-variants displayed heme-induced peroxidase-activity after incubation with heme. Three bands were seen with all variants, probably corresponding to the monomeric, dimeric and trimeric A1M-heme complexes previously reported (24). Interestingly, the peroxidase activity was much stronger in the dimeric and trimeric bands, when relating to the protein staining activity (Fig 6A vs B).

The sizes of the A1M-heme complexes were also investigated by Superose 12 gel-filtration (Fig. 7A). Incubation for 1 hour with heme (A1M:heme = 1:2) resulted in the appearance of larger forms besides the monomeric peak, suggesting an increased oligomerization of A1M in the presence of heme. This supports the results of the PAGE shown in Fig 6. The high molecular weight-forms were most pronounced in Wt-A1M and H(123)S-A1M, and less so in K(3)T- and C(34)S-A1M. Furthermore, a slight heme-induced shift of the monomeric peak towards higher molecular mass was seen in Wt-A1M.

3.4. UV-Vis absorbance of A1M-heme is dependent on C34, K92,118,1(32) and H132-residues

The heme binding was followed by UV-Vis absorbance spectrophotometry (Fig. 8). The heme binding could be confirmed, but the time-dependence of the binding was different for the various A1M-forms. A broad peak with a maximum (λmax) around 400 nm was seen immediately after mixing of 10 μM A1M + 20 μM heme for all variants (not shown). After

24 h, however, a red-shift of λmax towards a higher wavelength was seen for Wt-A1M, but not for any of the mutants (Fig. 8A; peak values of Soret-band shown in Table III). This red-shift was beginning after a few minutes and could be recorded at 30 min and onwards (not shown). Furthermore, zooming in on 500-700 nm wavelengths (Fig. 8B), the Wt-A1M-heme complex displayed a maximum at 540 nm, which was less pronounced in the mutants. The mutants also displayed an absorbance shoulder at 610 nm, this was most pronounced in the K(3)T-A1M-heme and H(123)S-heme complexes. The spectral differences could also be seen as a striking difference in color of the heme-complexed A1M-variants (Fig. 8C). At these concentrations (10 µM of both protein and heme), the Wt-A1M heme solution was red whereas the C(34)S-A1M heme complex was yellow and the K(3)T-A1M and H(123)S-A1M showed a similar yellow-brown color as free heme. The red-shifted Soret-band and the 540 nm-maximum is seen in ferrous (FeII) heme binding proteins (44). These results therefore suggest that A1M undergoes reducing reactions with the heme-group, involving the C34-residue and regulated by H123 and K92,118 and 130.

To minimize the contribution of unbound heme-groups to the spectra, the monomer peak fractions of the gel filtrated A1M-heme complexes were also analyzed by UV-Vis absorbance spectrometry. The peak wavelengths of the Soret-band (λmax) before and after gel filtration are shown in Table III. The red-shift of the Wt-A1M heme complex was also apparent after gel filtration, whereas the monomer fractions of three mutants showed a λmax below 400 nm. This suggests that the bound heme group is reduced by Wt-A1M but not the mutated forms.

3.5. Catalase-like activity

The catalase-like activity of A1M-heme complexes was measured by monitoring the Soret band after addition of H2O2 (Fig. 9). As evident from the figure, a significant shielding of the heme molecule was seen by all forms of A1M, as compared to heme alone. This further supports a binding of the heme group to A1M. However, plotting the peak value of each A1M-form as a function of time (Fig. 9D) demonstrates a less efficient shielding of the heme-molecule in C34S-A1M as compared to Wt-A1M and the other two mutants. This suggests that the C34 residue is essential for the heme coordination and/or redox activities of the bound heme-groups.

3.6. Induced visible CD support binding of two heme-groups

Neither heme nor A1M alone exhibits any CD activity in the visible range (not shown). However, when a small aliquot of heme solution is added to A1M, a CD activity in the heme absorbance region (390-415 nm) was induced (so called Cotton Effect) for each A1M variant (Fig. 10). These results support binding of heme to all A1M-variants. Titrations of heme were also consistent with the presence of two binding sites, as reported previously (24). At low heme concentrations (see bottom traces in Fig. 9, heme:A1M molar ratio 0.1), the induced CD proceeded to an equilibrium state relatively fast (< 40 min), suggesting binding to the primary binding site. As summarized in Table III, the peak wavelengths of the induced CD for Wt-A1M was 421 nm, and 417 nm, 397 nm and 416 nm for K(3)T-A1M, C34S-A1M and H123S-A1M, respectively, suggesting a different microenvironment of the heme molecule in the first binding site of each mutant. At higher heme concentrations (heme:A1M molar ratio 1.0; see upper traces in Fig. 10) the reaction was slower (> 5 hours to obtain equilibrium), consistent with binding at the second binding site at higher concentrations. The peak wavelengths (Table III) of the induced CD at heme:A1M ratio 1.0 were different from those shown for heme:A1M ratio 0.1, suggesting different heme environments and/or coordination at the primary and secondary binding sites.

3.7. Molecular simulation of A1M-heme binding

The lipocalin pocket, as identified by CASTp, essentially encompasses the inner cavity of the protein with a molecular volume of 2033 Å3. It was found that the inner lining of the pocket was composed of 41 residues. Apparently the molecular volume of heme, which is 538.9 Å3, could readily fit inside the lipocalin pocket (Fig. 11).

Before proceeding with the docking of heme to A1M, it is pertinent to explore the binding modality of other members of the lipocalin family that are known to bind heme. A total of 45 nitrophorin 1, 2 or 4 structures from Rhodnius prolixus were obtained from the PDB. These members of the lipocalin family had amino acid length in the range of 179-184 and their superimposition performed using MultiProt indicated high structural homology affording an RMSD value of 1.91 Å (Fig. 12). Analysis of these structures revealed that the bound heme was coordinated to axial H59 and ammonia ligands in nearly all cases.

Heme was docked to A1M using default parameters of HADDOCK without explicit definition of active and passive residues as they were assigned automatically. Active residues correspond to residues that are directly involved in the interaction whereas the passive residues denote residues that are in the vicinity of active residues. A typical docking simulation is comprised of the following steps: (i) rigid-body docking, (ii) scoring and filtering, (iii) semi-flexible refinement, (iv) water refinement, (v) scoring and analysis and finally (vi) clustering of docked structures. Docking results indicated that there were 193 structures in 6 clusters and that the top-performing cluster was the most populated with 142 structures affording a HADDOCK score of −86.7±7.9 and that the RMSD from the overall lowest-energy structure was 0.5±0.3 indicative of the close similarity of the docking conformation amongst members of this cluster. Thus, the lowest-energy structure of the A1M-heme complex was selected for further comparison with a representative structure from the aforementioned set of 45 structures from R. prolixus (PDB ID: 1X8P) obtained at an ultrahigh resolution of 0.85 Å. Superimposition of the two structures yielded an RMSD of 12.37 Å spanning the entire length of 164 residues in A1M (Fig. 13).

It can be seen that the bound heme in A1M leaves the putative outer heme binding site available. To investigate whether this site could accommodate another heme, a subsequent docking procedure was performed using PyMOL such that the iron atom of the second heme was coordinated to the SJ of C34 and NH of H123. After structural refinement using energy minimization and MD simulation, the final model was compared with the crystal structure of wild-type A1M (PDB ID: 3QKG; resolution of 2.3 Å) (25). A similar fold between the 3QKG structure and our heme-bound A1M model was observed with an RMSD of 1.29 Å (Fig. 14). The first heme binding site comprises residues from several strands of the E-barrel while the second heme binding site is formed by a few residues from the short D-helix at the open end and H123, which lies on loop 4, just opposite to the helix. Analyses of Ramachandran plot revealed that only 1.4 % of A1M residues were located in the disallowed region (data not shown) thereby indicating that appropriate stereochemical quality of the heme-bound A1M model was achieved.

To investigate fluctuations of the A1M structure upon heme binding and interactions of hemes with their pockets, MD simulations were performed for 30 ns on both heme-bound and heme-free A1M structures using explicit-solvent periodic boundary conditions. Stability of the protein and heme structures over the course of the simulation was observed by measuring the time evolution of root mean square deviation (RMSD) with respect to their initial structures (Fig. 15A). It can be seen that the simulated structures of A1M reached equilibrium prior to t=5 ns and remained stable throughout the simulation for both heme-bound and heme-free A1M models. RMSDs of the heme-bound A1M structure fluctuated around 0.2 nm whereas the heme-free A1M was slightly larger by | 0.1 nm, which implies that heme molecules were pertinent in stabilizing the A1M structure. The RMSD as a function of time for hemes in the A1M-heme simulation was also investigated. It was found that the second heme fluctuated to a much greater extent than the first heme in which RMSDs were 0.412r0.010 nm for the former and 0.141r0.003 nm for the latter (Fig.15B). Such differences may suggest either large flexibility of the binding pocket for the second heme or a reflection of the different binding strengths afforded by the two hemes.

To assess the conformational flexibility of A1M with respect to individual structural regions of the protein, root mean square fluctuation (RMSF) of the backbone as a function of the residue number was measured in both simulations from t= 5 to 30 ns (Fig. 15C). It can be seen that in the free A1M simulation, the four loops connecting neighboring E-strands (designated loop-1 to -4) at the open end of the eight-stranded E-barrel exhibited considerably high flexibility. This result is consistent with the experimental data in that those regions displayed high crystallographic B-factors, except for loop-4 of the 3QKG structure, which is responsible for the metal-binding site as well as exhibiting low temperature factor (25). As expected, presence of the second heme significantly decreased fluctuation of the binding site formed by the short α-helix of loop-1 (C34 to M40) and residues on loop-4 (S120 to G124). On the other hand, binding of the first heme did not alter the fluctuation of the lipocalin pocket in which a similar RMSF profile was observed in such region.

To investigate contributing factors that dominate the interaction of the A1M-heme complex, potential energies for interactions between each heme molecule and the protein were calculated as the sum of the electrostatic and van der Waals (vdW) interactions from t=5 to 30 ns (Fig. 15D). It can be seen that the first heme exhibited considerably large negative interaction energies with the protein, which is in contrast to the second heme that exhibited lower interactions. It should be noted that electrostatic energies were found to be major factors stabilizing the interaction between the first heme and A1M while vdW interactions contributed more dominantly in the second heme-bound A1M. These results suggest that the first heme binds the lipocalin pocket of A1M with more strength than the second heme, which may explain the higher degree of flexibility of the heme molecule when bound to the second binding pocket of A1M. The aforementioned evidences also suggests that binding of the first heme to the lipocalin pocket stabilizes the protein structure and predisposes A1M for binding of the second heme to the surface-exposed binding site.

4. Discussion

4.1. Side-groups of C34, H123, K92, K118 and K130 are involved in heme-binding

The investigation in this paper is based on site-directed mutagenesis, i.e. biochemical analysis of wild-type A1M and three mutated forms of the protein. In order to ascertain that functional differences were due to the amino acid substitutions rather than impurities or effects of incorrect folding, we first investigated biochemical properties of the recombinant A1M-species. The proteins appeared highly purified, and far-UV CD spectra of Wt-A1M and the mutated A1M-forms suggested a similar composition of secondary structure, i.e. mostly β-structure (Fig. 2B; Table I), consistent with the X-ray crystallography-derived three-dimensional structure of Wt-A1M (25). Tryptophan fluorescence spectra were also recorded as an estimate of the overall conformation of the mutants (Fig. 2C). The C(34)S-mutant displayed a higher fluorescence intensity than Wt-A1M and the other mutants. Sulfur-containing molecules are known fluorescence quenchers (45). The C34-residue and one of the four tryptophan residues, W36, are located closely together on the small helix on the rim of the lipocalin pocket (Fig. 1), hence there is a possibility of an interaction between these two residues reducing the intensity of the fluorescence of W36 in Wt-A1M and the other mutants. Apart from this, similar spectra were obtained, suggesting no major differences in the overall conformation of the four variants.

It was previously shown that heme binds specifically to A1M and that this feature is evolutionally conserved (21-23). Furthermore, it was shown that A1M forms a trimeric complex with heme in a 1:2 stoichiometry (24). In order to understand the mechanism of heme binding of A1M, the binding was studied by several different techniques, and to understand the structural requirements of the binding and possibly localize binding sites, the binding to mutated variants was also studied. In short, all methods showed that heme was bound to all variants and indicated a slightly higher binding by Wt-A1M compared to the mutated variants, whereas heme titration during CD-spectral analysis, UV-Vis absorbance spectrophotometry and the catalase-like activity assays revealed differences between the four A1M-variants suggesting involvement of the C34-, H123- and K(92,118,130)-side-groups.

The experimental results in this investigation were obtained using A1M of all four variants carrying an N-terminal His8-tag. We used the migration shift-PAGE assay to show that non-tagged Wt-A1M binds the heme-group with similar binding strength (43). Moreover, the results from SPR-analysis were obtained by studying A1M-molecules immobilized to the surface by binding via the His-tag. It was also shown previously that the His-tag does not contribute to, or interfere with, the reductase activities of A1M (19,20) and we therefore expect that any reductase activity of the A1M-variants in this investigation is not affected by the His-tag. Although this does not rule out that the His-tag may influence the heme binding, we therefore believe that the major conclusions of the work also are valid for non-tagged A1M.

4.2. Characterization of two heme-binding sites in A1M

Several different results suggest that each A1M molecule can bind two heme-groups simultaneously. A previous report used heme-titration, gel chromatography, and resonance Raman and EPR spectroscopy to demonstrate the formation of A1M:heme complexes with the stoichiometry 1:2 (24). In the present work, the differences in the heme-induced CD-spectra at low and high heme:A1M ratios are consistent with a subsequent filling of two binding-sites. Based on the spectral behavior of the four A1M-variants, we propose that one heme binding site is located in the lipocalin pocket, the other is located between loops 1 and 4 at the outer rim of the pocket, and that the sites are filled in that order by increasing heme concentrations. The existence of an inner binding site is supported experimentally by the almost complete quenching of the tryptophan fluorescence in spite of the fact that one of the tryptophan residues (W25) is located at the bottom of the pocket, and the dependence of the heme-induced visual CD spectra on the K(92, 118, 130) residues which line the interior pocket wall. The outer binding site is supported by the influence of the C34- and H123-residues on heme-titration effects, the less efficient shielding of the heme-group from H2O2-induced degradation in the C34S-mutant, and was also proposed by Meining and Skerra (25) based on its similarities with a group of heme binding proteins with a Cys-Pro dipeptide motif (46).

Our in silico modelling supports the possibility of simultaneous binding of two heme-groups to the proposed inner and outer binding site. The inner binding site is analogous to the heme binding sites of nitrophorins, heme binding members of the Lipocalin protein family. Dynamic analysis of the models also suggest a stronger binding of heme to the inner binding site, and/or a higher flexibility of the second binding site. This is consistent with a primary binding to the inner site and a secondary binding to the outer binding site. Interestingly, the inner binding site suggest close proximity between K118 and K130 side-groups and the non-pyrrolic groups of the heme-molecule (Fig. 14B), supporting the differences in heme-induced visual CD-spectra between Wt- and K(92,118,130)T-A1M. It has previously been shown that these three side-groups become covalently modified and cross-linked in vivo by yellow-brown, unidentified, size heterogeneous compounds with molecular masses between 122 and 282 atomic mass units (47). Based on this, it can be speculated that a reaction between the inner heme group and the A1M protein leads to degradation of the heme-group, yielding covalent attachment of degradation products to the lysyl side-chains. This hypothetical reaction may involve electron-transfer reactions of the iron atoms of both heme-groups as well as the thiol group of C34.

4.3. Possible electron transfer reaction between A1M and heme

The UV-Vis absorbance spectrum of Wt-A1M displayed the characteristic features of hemoglobin in its reduced form, i.e. a red-shifted Soret band and a peak at 540 nm, in contrast to the three mutated A1M forms (Fig. 9). A negative reduction potential of the C34 thiol group in combination with the three lysyl residues K92,118 and 130 was previously shown, i.e. A1M reduced methemoglobin, cytochrome c and free iron (19). It may therefore be speculated that Wt-A1M keeps the outer heme-group in its reduced, ferrous (Fe2+) form by a tentative reaction that involves coordination of the iron atom by the C34- and H123-residues, where the C34 thiol group also may serve as an electron-source. As proposed earlier, the K(92,118,130)-residues may regulate the electronegativity of the thiol group by creating a positive electrostatic microenvironment (7,19).

4.4. Physiological function of heme-binding to A1M

Several potential physiological functions of the heme binding by A1M are possible. The first, and perhaps most obvious, may be scavenging of free heme-groups to protect biomolecules from heme-induced toxic reactions. This function is supported by previous reports that A1M-binding of heme results in inhibition of cell-lysis, cell-cycle arrest and molecular damage otherwise induced by the free heme (13,48). Second, heme binding can be the first step in a series of reactions leading to heme-degradation. This also involves proteolytic cleavage of the C-terminal tetrapeptide Leu-Ile-Pro-Arg (pos 179-183) of A1M, as suggested in the report by Allhorn et al. (21). However, nothing is known about the detailed molecular mechanisms leading to the heme-degradation. The position of the C-terminus beyond amino acid residue Gly172 could not be determined in the published crystal structure of A1M (25) and it is therefore difficult to speculate on the mechanistic influence of the C-terminal tetrapeptide on binding and degradation of the heme-group. Thirdly, the heme-group, with the iron atom, may be employed as an enzymatic cofactor during the redox activities of A1M. Heme-groups are commonly employed as electron-active cofactors in peroxidases, reductases, dehydrogenases, etc, and it is therefore reasonable to imagine a role of the bound heme-groups in A1M in reduction, radical scavenging, or other antioxidative activities. This remains to be tested experimentally, however.

The concentration of A1M in blood plasma is 1-2 μM. Among heme binders in plasma, A1M is apparently not the strongest. Hemopexin has a much higher affinity. Albumin, which binds heme with a slightly higher affinity than A1M is much more abundant in blood. This suggests that the role of A1M is not primarily as a heme-scavenger in blood. A1M is secreted to blood from the liver, but rapidly transferred to the extravascular space (T1/2=2.5 min). The protein is also synthesized by most other epithelial cells and found both intracellularly and in the extracellular matrix, for example in skin and placenta. We therefore propose that the heme-reaction mechanisms of A1M are different from those of hemopexin and albumin, and of physiological importance in cells and extravascular fluids rather than in blood.

5. Conclusions

Several previous investigations have shown that the lipocalin A1M can bind to heme groups, and that this property constitutes part of its physiological antioxidative mechanisms (13,21,23). In this work we have investigated the structural requirements of the heme binding. The main conclusions of the present report are that two heme-groups can be accommodated simultaneously in the protein, and that the binding and reactions with the heme-groups are affected by the Cys34, Lys92, 118, 130, and His123 residues lining the lipocalin pocket.

6. References

1. Flower, D. R. (1996) The lipocalin protein family: structure and function. Biochem J 318 (Pt 1), 1-14
2. Åkerström, B., Flower, D. R., and Salier, J. P. (2000) Lipocalins: unity in diversity. Biochim Biophys Acta 1482, 1-8
3. Åkerström, B., Borregaard, N., Flower, D. R., and Salier, J. P. (eds). (2006) Lipocalins: An introduction, Landers Bioscience, Georgetown, Texas
4. Ganfornina, M. D., Sanchez, D., Greene, L. H., and Flower, D. R. (eds). (2006) The lipocalin protein family: Protein sequence, structure and relationship to the calycin superfamily, Landers Bioscience, Georgetown, Texas
5. Ekström, B., and Berggård, I. (1977) Human α1 -microglobulin . Purification procedure, chemical and physiochemical properties. J Biol Chem 252, 8048-8057
6. Åkerström, B., and Lögdberg, L. (1990) An intriguing member of the lipocalin protein family: a1 - microglobulin. Trends Biochem Sci 15, 240-243
7. Åkerström, B., and Gram, M. (2014) A1M, an extravascular tissue cleaning and housekeeping protein. Free Radic Biol Med 74, 274-282
8. Kaumeyer, J. F., Polazzi, J. O., and Kotick, M. P. (1986) The mRNA for a proteinase inhibitor related to the HI-30 domain of inter-alpha-trypsin inhibitor also encodes a1 -microglobulin (protein HC). Nucleic Acids Res 14, 7839-7850
9. Lindqvist, A., Bratt, T., Altieri, M., Kastern, W., and Åkerström, B. (1992) Rat alpha 1-microglobulin: co-expression in liver with the light chain of inter-alpha-trypsin inhibitor. Biochim Biophys Acta 1130, 63-67
10. Bratt, T., Olsson, H., Sjöberg, E. M., Jergil, B., and Åkerström, B. (1993) Cleavage of the alpha 1-microglobulin-bikunin precursor is localized to the Golgi apparatus of rat liver cells. Biochim Biophys Acta 1157, 147-154
11. Grubb, A., Mendez, E., Fernandez-Luna, J. L., Lopez, C., Mihaesco, E., and Vaerman, J. P. (1986) The molecular organization of the protein HC-IgA complex (HC-IgA). J Biol Chem 261, 14313-14320
12. Berggård, T., Thelin, N., Falkenberg, C., Enghild, J. J., and Åkerström, B. (1997) Prothrombin, albumin and immunoglobulin A form covalent complexes with alpha1-microglobulin in human plasma. Eur J Biochem 245, 676-683
13. Olsson, M. G., Olofsson, T., Tapper, H., and Åkerström, B. (2008) The lipocalin alpha1-microglobulin protects erythroid K562 cells against oxidative damage induced by heme and reactive oxygen species. Free Radic Res 42, 725-736
14. Olsson, M. G., Allhorn, M., Larsson, J., Cederlund, M., Lundqvist, K., Schmidtchen, A., Sörensen, O. E., Mörgelin, M., and Åkerström, B. (2011) Up-Regulation of A1M/a1 -microglobulin in Skin by Heme and Reactive Oxygen Species Gives Protection from Oxidative Damage. PLoS One 6, e27505
15. May, K., Rosenlöf, L., Olsson, M. G., Centlow, M., Mörgelin, M., Larsson, I., Cederlund, M., Rutardottir, S., Siegmund, W., Schneider, H., Åkerström, B., and Hansson, S. R. (2011) Perfusion of human placenta with hemoglobin introduces preeclampsia-like injuries that are prevented by alpha1-microglobulin. Placenta 32, 323-332
16. Wester-Rosenlöf, L., Casslen, V., Axelsson, J., Edström-Hägerwall, A., Gram, M., Holmqvist, M., Johansson, M. E., Larsson, I., Ley, D., Marsal, K., Mörgelin, M., Rippe, B., Rutardottir, S., Shohani, B., Åkerström, B., and Hansson, S. R. (2014) A1M/alpha1-microglobulin protects from heme-induced placental and renal damage in a pregnant sheep model of preeclampsia. PLoS One 9, e86353
17. Sverrisson, K., Axelsson, J., Rippe, A., Gram, M., Åkerström, B., Hansson, S. R., and Rippe, B. (2014) Extracellular fetal hemoglobin induces increases in glomerular permeability: inhibition with alpha1-microglobulin and tempol. Am J Physiol Renal Physiol 306, F442-448
18. Rutardottir, S., Nilsson, E. J., Pallon, J., Gram, M., and Åkerström, B. (2013) The cysteine 34 residue of A1M/alpha1-microglobulin is essential for protection of irradiated cell cultures and reduction of carbonyl groups. Free Radic Res 47, 541-550
19. Allhorn, M., Klapyta, A., and Åkerström, B. (2005) Redox properties of the lipocalin alpha1-microglobulin: reduction of cytochrome c, hemoglobin, and free iron. Free Radic Biol Med 38, 557-567

20. Åkerström, B., Maghzal, G. J., Winterbourn, C. C., and Kettle, A. J. (2007) The lipocalin alpha1-microglobulin has radical scavenging activity. J Biol Chem 282, 31493-31503

21. Allhorn, M., Berggård, T., Nordberg, J., Olsson, M. L., and Åkerström, B. (2002) Processing of the lipocalin α1-microglobulin by hemoglobin induces heme-binding and heme-degradation properties. Blood 99, 1894-1901

22. Allhorn, M., Lundqvist, K., Schmidtchen, A., and Åkerström, B. (2003) Heme-scavenging role of α1-microglobulin in chronic ulcers. J Invest Dermatol 121, 640-646

23. Larsson, J., Allhorn, M., and Åkerström, B. (2004) The lipocalin α1-microglobulin binds heme in different species. Arch Biochem Biophys 432, 196-204

24. Siebel, J. F., Kosinsky, R. L., Åkerström, B., and Knipp, M. (2012) Insertion of heme b into the structure of the Cys34-carbamidomethylated human lipocalin alpha(1)-microglobulin: formation of a [(heme)(2) (alpha(1)-Microglobulin)](3) complex. Chembiochem 13, 879-887

25. Meining, W., and Skerra, A. (2012) The crystal structure of human alpha(1)-microglobulin reveals a potential haem-binding site. Biochem J 445, 175-182

26. Finkel, T., and Holbrook, N. J. (2000) Oxidants, oxidative stress and the biology of ageing. Nature 408, 239-247

27. Kwasek, A., Osmark, P., Allhorn, M., Lindqvist, A., Åkerström, B., and Wasylewski, Z. (2007) Production of recombinant human alpha1-microglobulin and mutant forms involved in chromophore formation. Protein Expr Purif 53, 145-152

28. Lamar.colostate.edu/~/sreeram/CDPro/.

29. Karnaukhova, E., Krupnikova, S. S., Rajabi, M., and Alayash, A. I. (2012) Heme binding to human alpha-1 proteinase inhibitor. Biochim Biophys Acta 1820, 2020-2029

30. EMBL European Bioninformatics Institute, P. (2014).

31. Molinspiration, C. o. M. P. a. B. S. (2014).

32. Molinspiration, M. v. (2014).

33. W, D. (2002) PyMOL Release 0.99. Palo Alto, DeLano Scientific LLC.

34. Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., and Liang, J. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34, W116-118

35. Shatsky, M., Nussinov, R., and Wolfson, H. J. (2004) A method for simultaneous alignment of multiple protein structures. Proteins 56, 143-156

36. de Vries, S. J., van Dijk, M., and Bonvin, A. M. (2010) The HADDOCK web server for data-driven biomolecular docking. Nature protocols 5, 883-897

37. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005) GROMACS: fast, flexible, and free. J Comput Chem 26, 1701-1718

38. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. J. Chem. Phys 81, 3684

39. Parinello, M., and Rahman, A. (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. J. Appl. Phys 52, 7182

40. Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25, 1656-1676

41. Hess, B., Bekker, H., Berendsen, H. J., and Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. J Comput Chem 18, 1463-1472

42. Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald: An N log(N) method for Ewald sums in large systems J. Chem. Phys 98

43. Karnaukhova, E., Rutardottir, S., Rajabi, M., Wester Rosenlöf, L., Alayash, A. I., and Åkerström, B. (2014) Characterization of heme binding to recombinant alpha1-microglobulin. Frontiers in physiology 5, 465

44. Winterbourn, C. C. (1990) Oxidative reactions of hemoglobin. Methods Enzymol 186, 265-272

45. Kuznetsova, I. M., Yakusheva, T. A., and Turoverov, K. K. (1999) Contribution of separate tryptophan residues to intrinsic fluorescence of actin. Analysis of 3D structure. FEBS Lett 452, 205-210

46. Li, T., Bonkovsky, H. L., and Guo, J. T. (2011) Structural analysis of heme proteins: implications for design and prediction. BMC structural biology 11, 13

47. Berggård, T., Cohen, A., Persson, P., Lindqvist, A., Cedervall, T., Silow, M., Thogersen, I. B., Jönsson, J. A., Enghild, J. J., and Åkerström, B. (1999) a1 - microglobulin chromophores are located to three lysine residues semiburied in the lipocalin pocket and associated with a novel lipophilic compound. Protein Sci 8, 2611-2620

48. Olsson, M. G., Allhorn, M., Bülow, L., Hansson, S. R., Ley, D., Olsson, M. L., Schmidtchen, A., and Åkerström, B. (2012) Pathological Conditions Involving Extracellular Hemoglobin: Molecular Mechanisms, Clinical Significance, and Novel Therapeutic Opportunities for alpha(1)-Microglobulin. Antioxid Redox Signal 17, 813-846

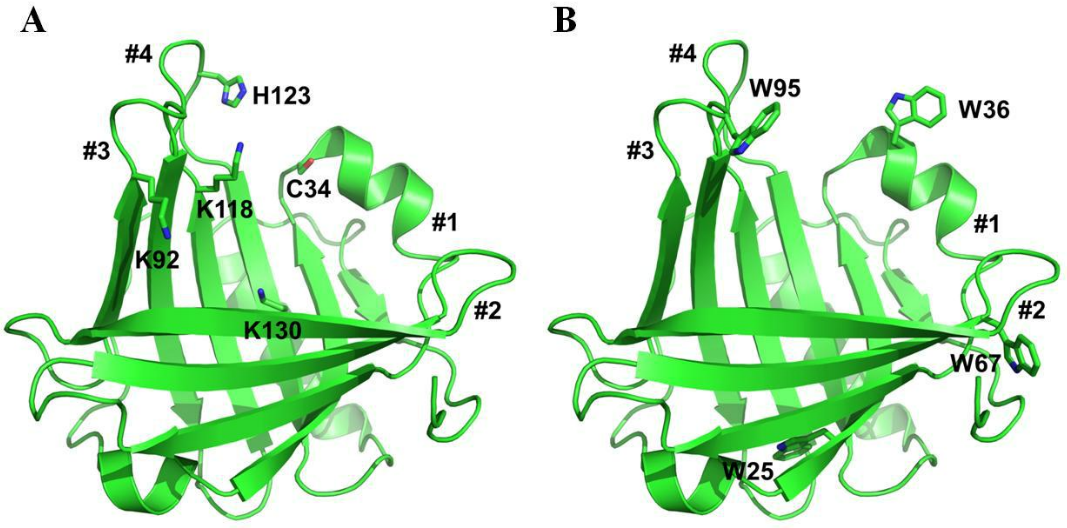


Figure 1. Three-dimensional structure of A1M. The illustration was generated using PyMOL (32) and coordinates from the crystal structure of human A1M (25). β-strands and α-helices are shown as green ribbons. The four loops at the open end of the lipocalin pocket are labeled #1 - #4. (A) Side-chains of C34, K92, K118, K130 and H123 are shown in green sticks. (B) Side-chains of W25, W34, W67 and W95 are shown in green sticks.

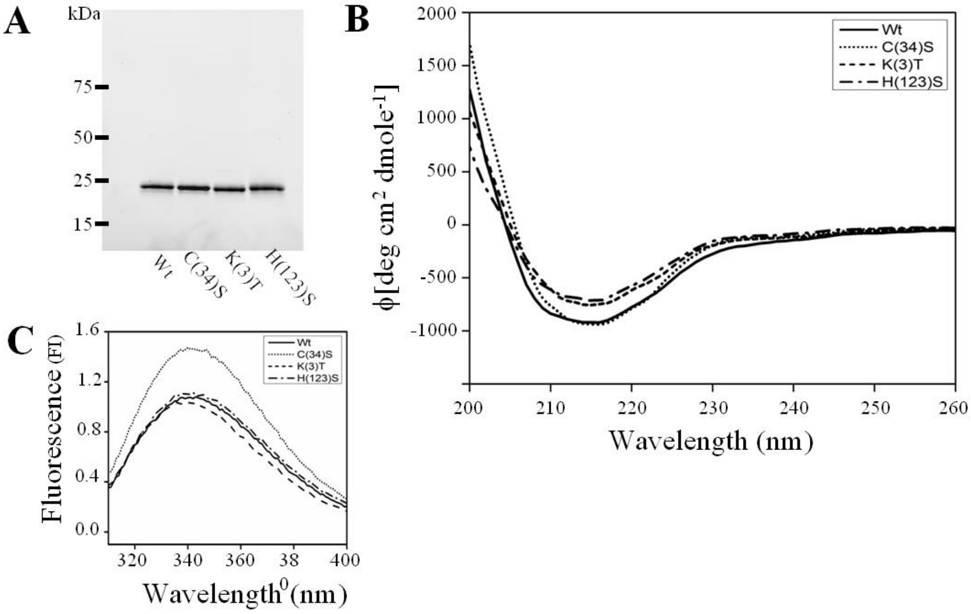


Figure 2.. Size, purity and spectroscopic properties of the four variants of recombinant A1M. (A) SDS-PAGE was performed in the presence of mercaptoethanol (T= 12 %). Approximately 2 µg of Wt-A1M, C(34)S-A1M, K(3)T-A1M, and H(123)S-A1M were applied to the gel and stained with Coommassie. (B) Far-UV CD spectra of A1M variants (10 µM in 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl). Similar spectra of Wt-A1M, C(24)S and K(3)-T-A1M were published in Kwasek et al. (27) and are included here for comparison of H(123)S-A1M. (C) Fluorescence spectra of A1M variants (1 µM) in 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl).

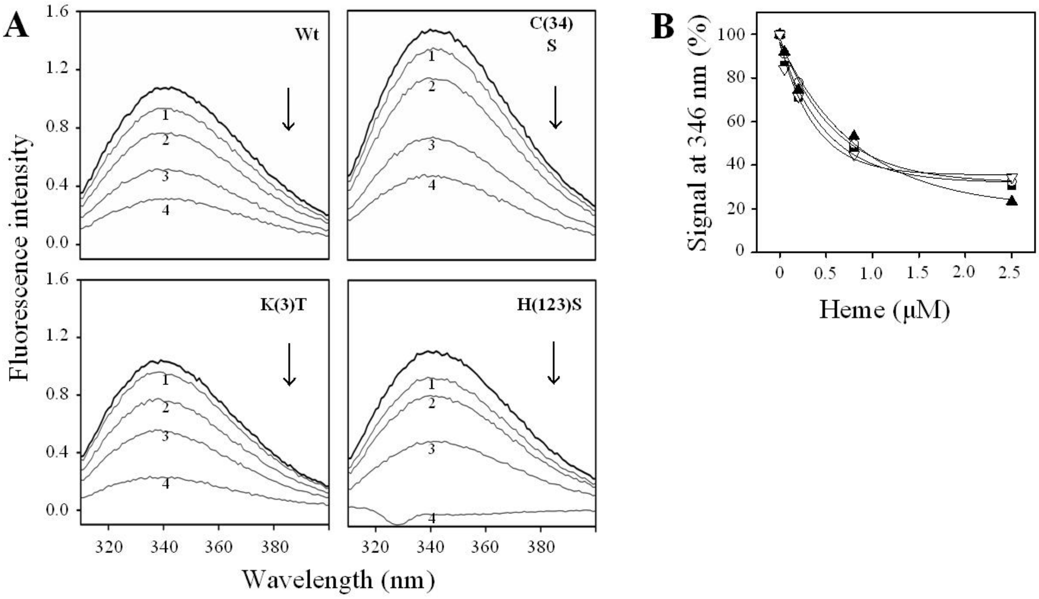


Figure 3. Heme-induced tryptophan fluorescence quenching in the four A1M-variants. (A) Fluorescence spectra of A1M-variants (1 µM) incubated with heme to a final concentration of; 1 = 0.05, 2 = 0.2, 3 = 0.8, 4 = 2.5 µM. (B) Normalized titration curve of A1M-variants with heme. The fluorescence intensity of each A1M-variant without heme-addition was set to 1. □:Wt, {: C(34)S,◁: K(3)T,▲: H(123)S

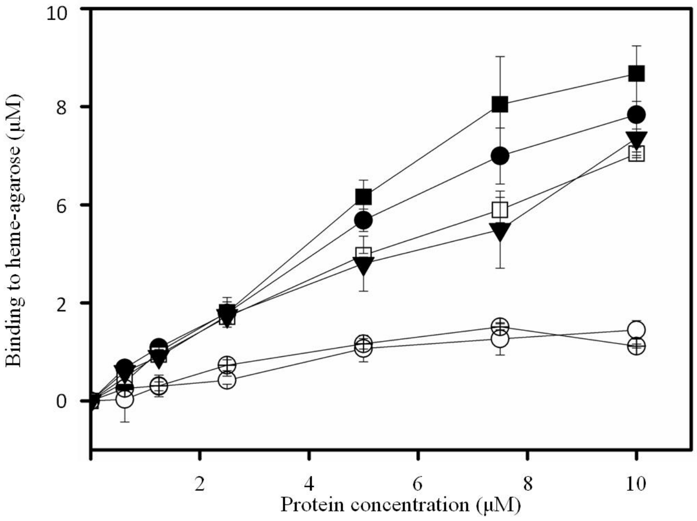


Figure 4. Binding of A1M-variants to heme- agarose. The proteins were diluted in 10mM Tris-HCl pH 8.0, 0.125M NaCl. Dilution series of all proteins (concentrations of 10, 7.5, 5, 2.5, 1.25, 0.625 µM) were incubated with heme agarose and unconjugated Sepharose for 30 min. Beads and unbound protein were separated by centrifugation through a filter plate. Non-incubated samples and each flow through, containing unbound protein, were analyzed with BCA Protein Assay Kit. Binding to unconjugated Sepharose was used as control. Binding to heme-agarose is shown as follows: „: Wt, z: C(34)S, □: K(3)T,▼: H(123)S, □: Ovalbumin. {:Wt-A1M to un-conjugated Sepharose. The mean of two replicates +/- SEM are shown.

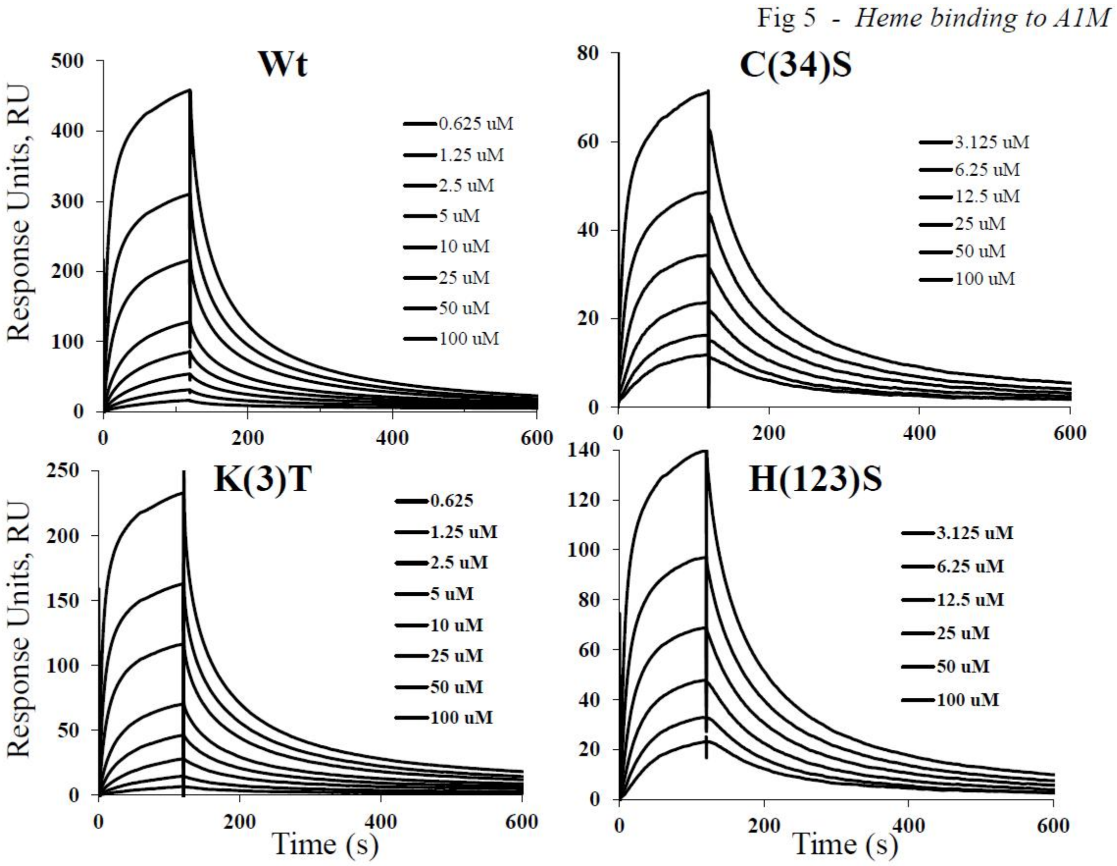


Figure 5. Binding of heme to A1M analyzed by surface plasmon resonance. Sensorgrams of the heme binding to wt-, C(34)S-, K(3)T- and H(123)S-A1M captured by the anti-His mouse IgG1 monoclonal antibody immobilized on CM5 sensor chip. Increasing signals were obtained using 0.625 – 100 μM heme.

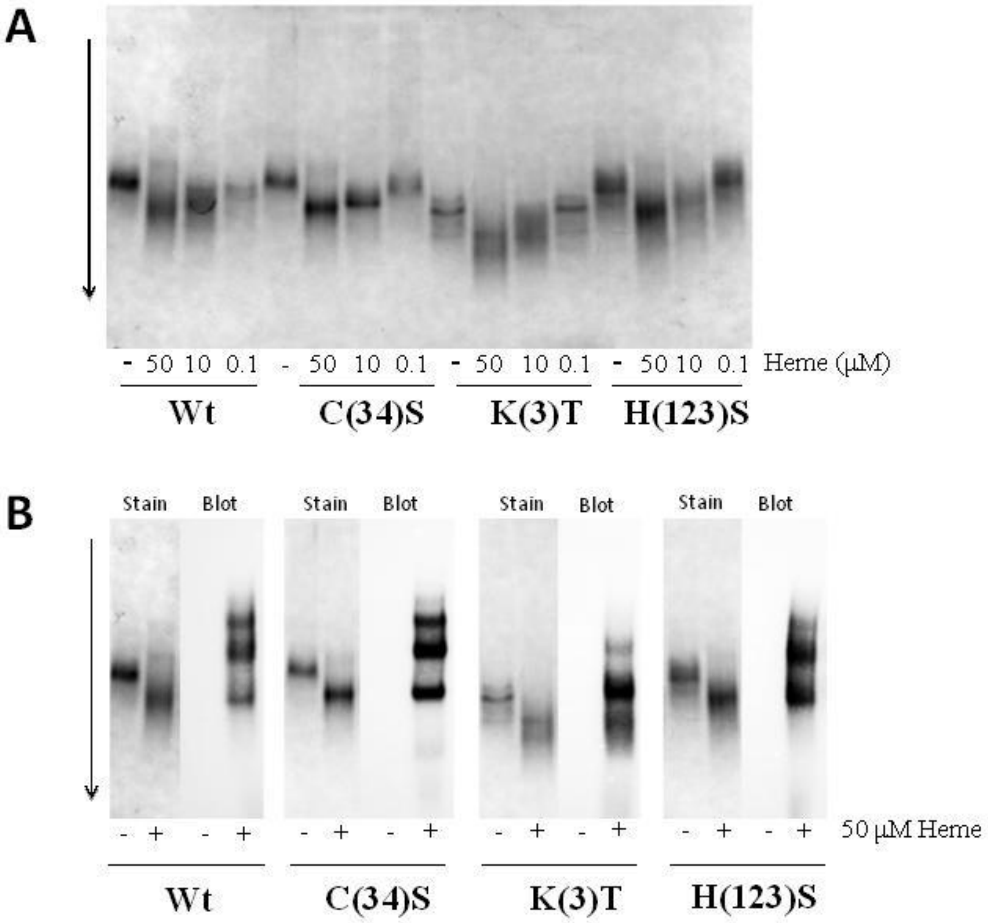

A
- 50 10 0.1 - 50 10 0.1 - 50 10 0.1 - 50 10 0.1 Heme (µM)
Wt
C(34)S
K(3)T
H(123)S
B
Stain Blot Stain Blot Stain Blot Stain Blot
- + - + - + - + - + - + - + - + 50 µM Heme
Wt
C(34)S
K(3)T
H(123)S

Figure 6. Heme-induced migration shift of A1M-variants. (A) Native PAGE of A1M (5 µM) in 20 mM Tris-HCl, pH 8.0, + 0.15 M NaCl incubated for 1h with 50, 10 or 0.1 µM heme. Samples were mixed with equal amounts of sample buffer for native PAGE, pH 6.8, and subjected to a 12% CriterionTM TGXTM Precast Gel and stained with Coommassie. (B) Peroxidase blotting of A1M-variants incubated for 4h without or with heme (50 µM), subjected to native PAGE in a 12% CriterionTM TGXTM Precast Gel and thereafter either stained with Coomassie (5 µg A1M), or transferred to polyvinylidene difluoride (PVDF) membranes (2 µg A1M). The membrane was incubated in Clarity Western ECL Substrate and imaged with a digital imager (BioRad).

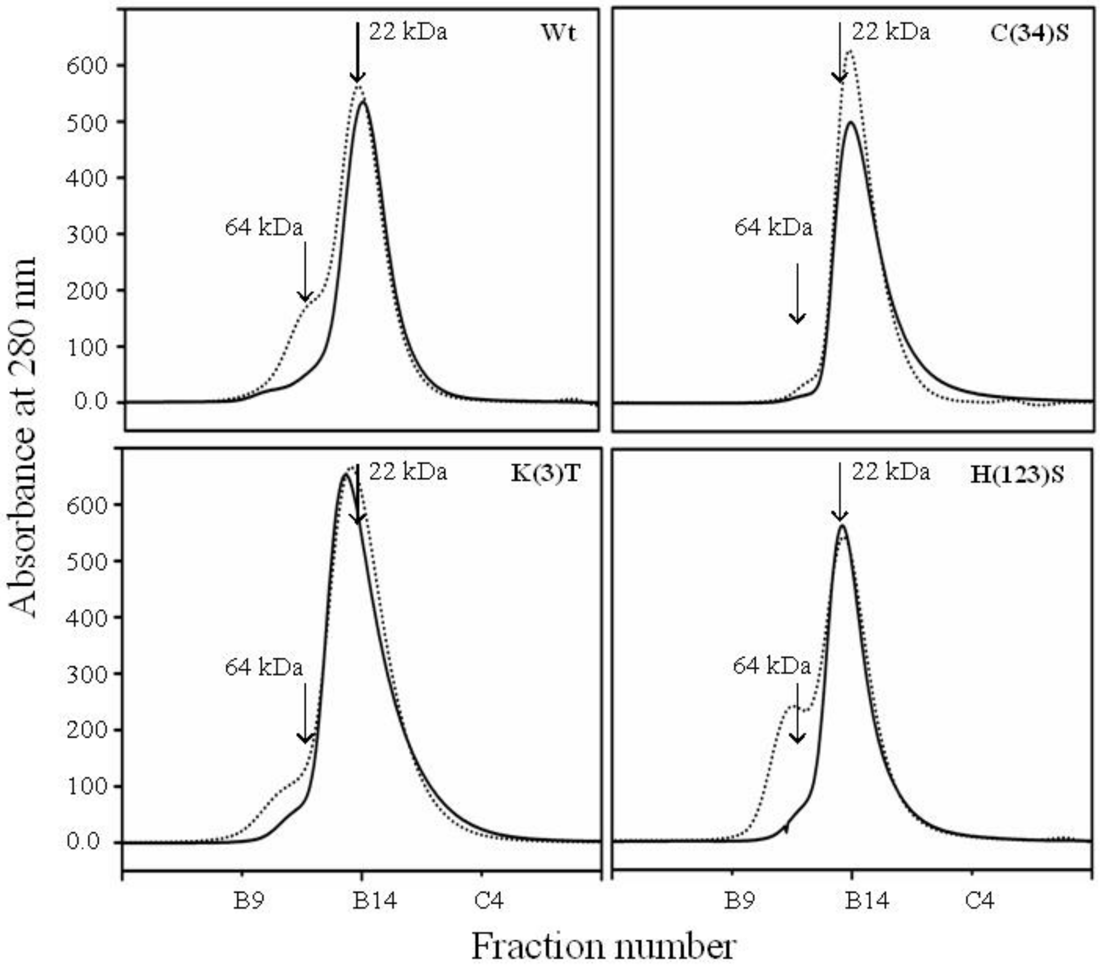


Figure 7. Size exclusion chromatography of A1M variants incubated with heme. A1M (0.5 ml of a 44 µM-solution) was applied on a Superose 12 FPLC column and eluted with 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl at 0.5 ml/min (unbroken line). A1M and heme in molar ratio of 1:2 were incubated for 1h and run on Superose 12 FPLC column as above (dotted line). The size calibration on the column was performed with Wt-A1M (22 kDa) and with human hemoglobin (64 kDa).

Figure 8 *Heme binding to A1M*

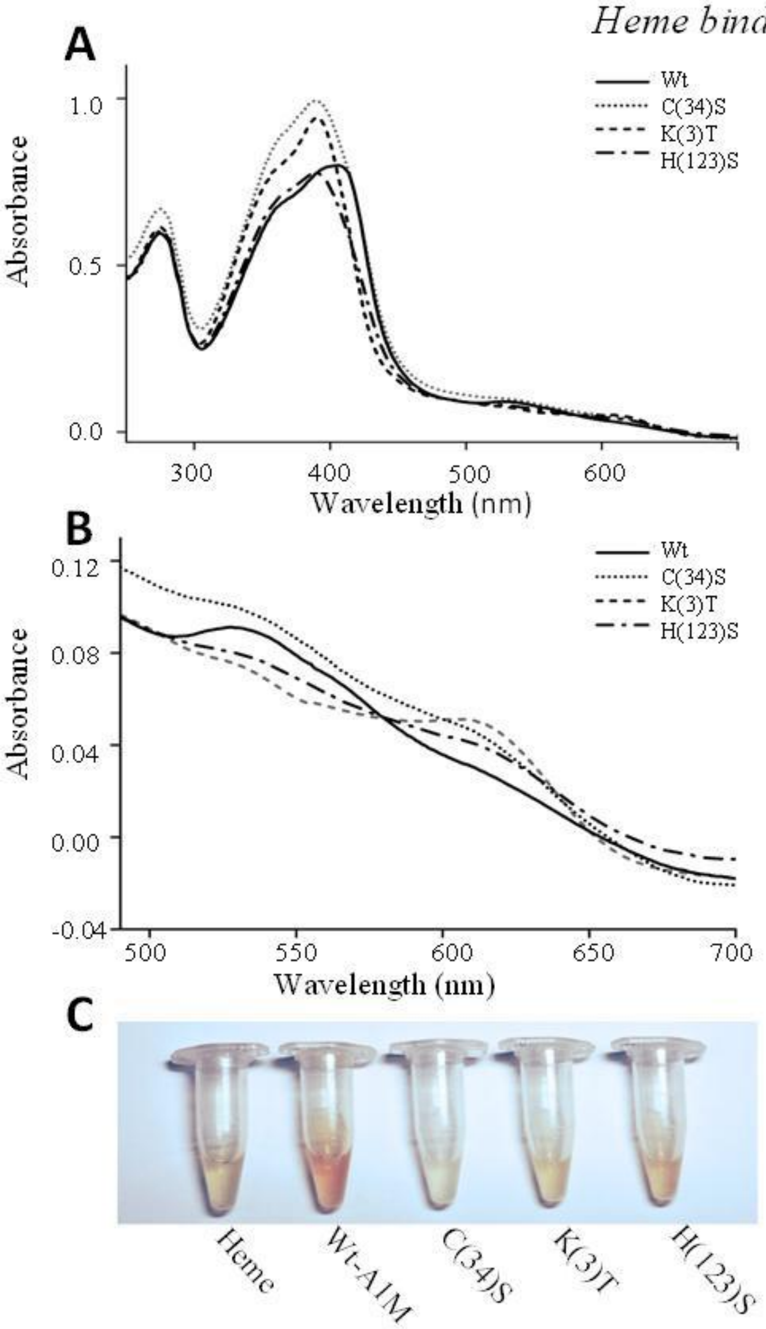


Figure 8. Absorbance spectrophotometry of A1M variants incubated with heme. (A) Absorbance spectrum of A1M variants (10 µM) in the presence of 20 µM heme in 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl after 24h of incubation. A similar absorbance spectrum of Wt-A1M was previously published in Karnaukhova et al. (43). (B). Close-up of the absorbance spectrum between 500-700 nm of A1M variants (10 µM) in the presence of 20 µM heme in 20 mM Tris-HCl, pH 8.0, 0.15 M NaCl after 24 h of incubation. (C) The A1M variants incubated with heme and heme only were also analyzed visually. Protein:heme ratio was 1:2 and the heme concentration 20 µM.

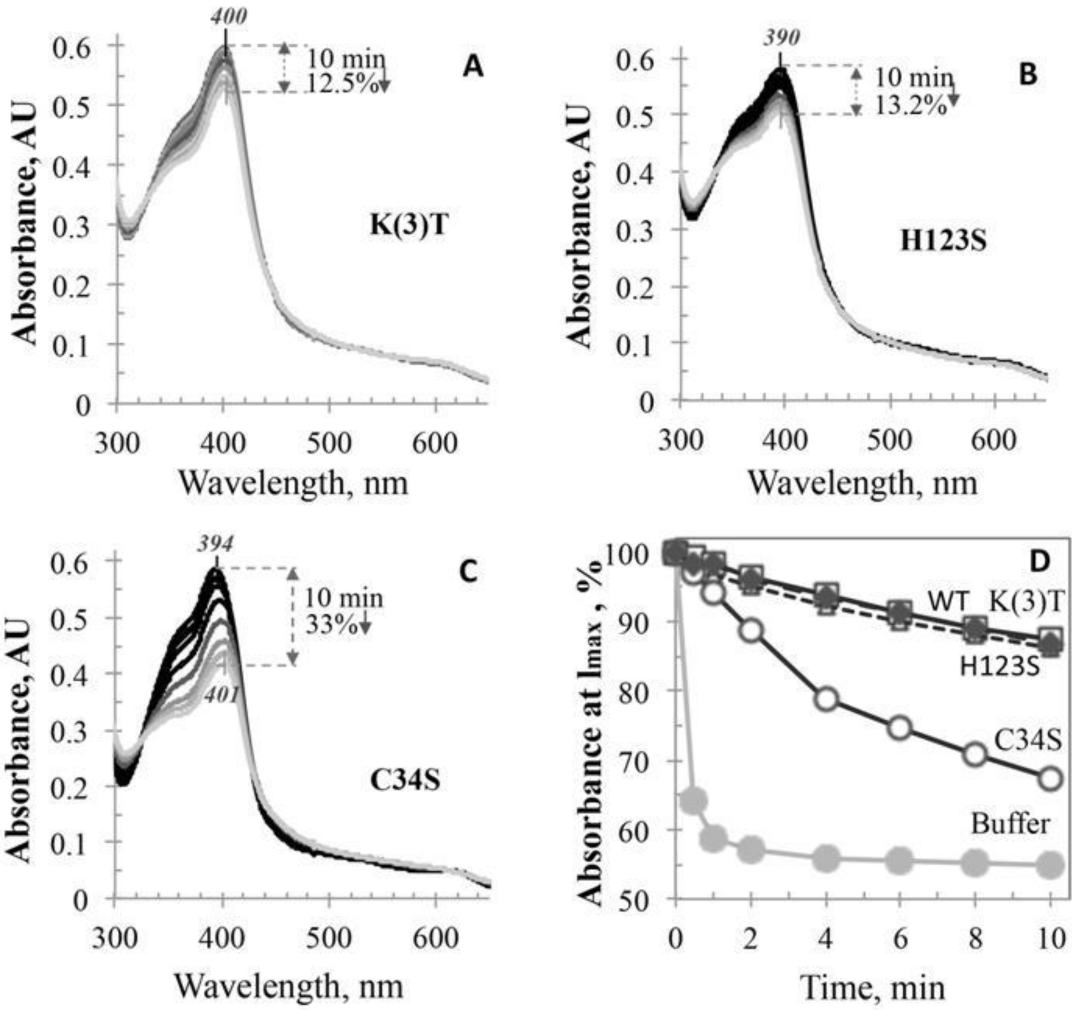


Figure 9. Time-course UV/Vis absorbance data for catalase-like activity of the heme complexes with A1M mutants: (A) K(3)T; (B) H123S, and (C) C34S. The upper trace of each plot shows initial absorbance spectrum of each heme/mutant sample (1:1) as recorded over 10 min with time points taken at 30 s, 1 min, 2 min, 4 min, 6 min, 8 min, and 10 min after adding a 7 µl aliquot of 50 mM H2O2 stock solution to 1 ml of the protein sample. Plot (D) shows the percentage of the remaining intensity of the Soret band of each mutant in comparison with 0 s.

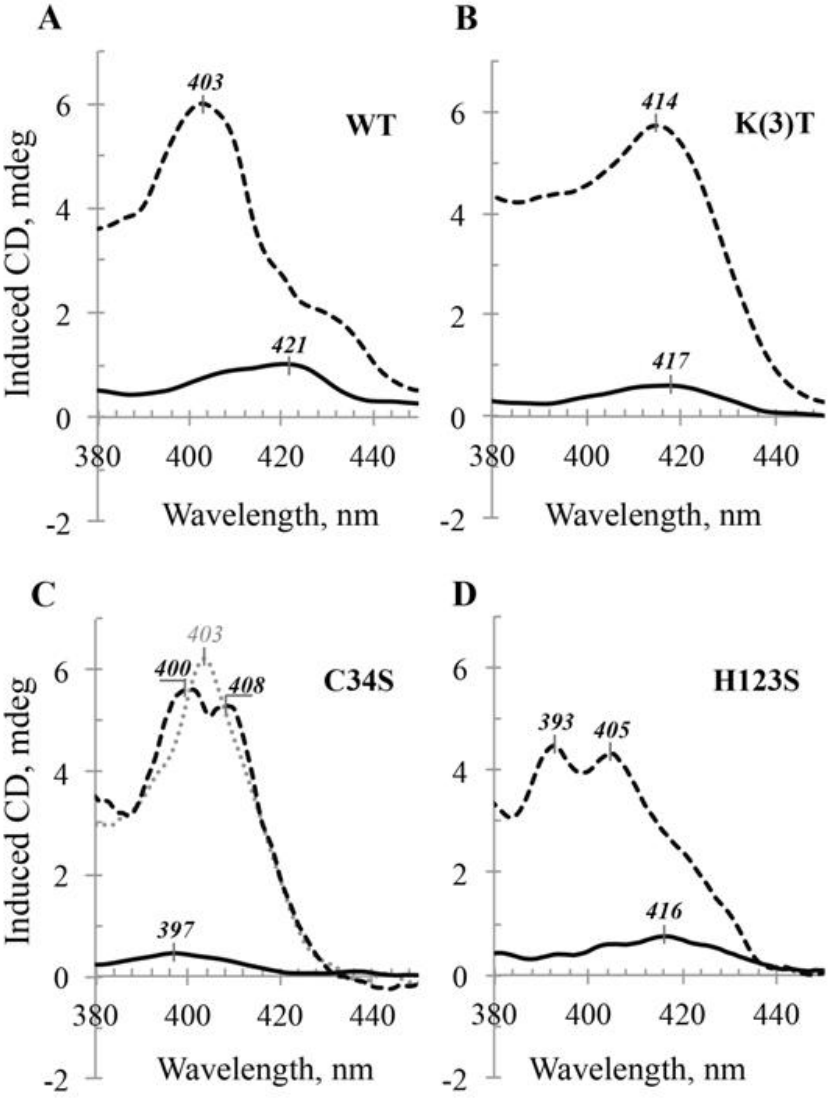


Figure 10. Induced CD data for heme complexes with A1M. Wt-A1M (A), and A1M mutants K(3)T (B), C34S (C), and H123S (D). Induced CD spectra are shown for heme:A1M ratio 0.1 (solid traces in A-D, the spectra were recorded at 1 h after adding heme, when no additional spectral changes were observed) and heme:A1M ratio 1.0 (broken traces, measured 20 h after adding heme, when no additional changes were observed). The broken trace in 9A was previously published in Karnaukhova et al (43). Gray dotted trace shown for C34S plot (C) corresponds to an intermediate state observed for this C34S mutant at L/P 1.0 at the time point 1.5 h after adding heme. Protein concentration in all samples was 45µM.

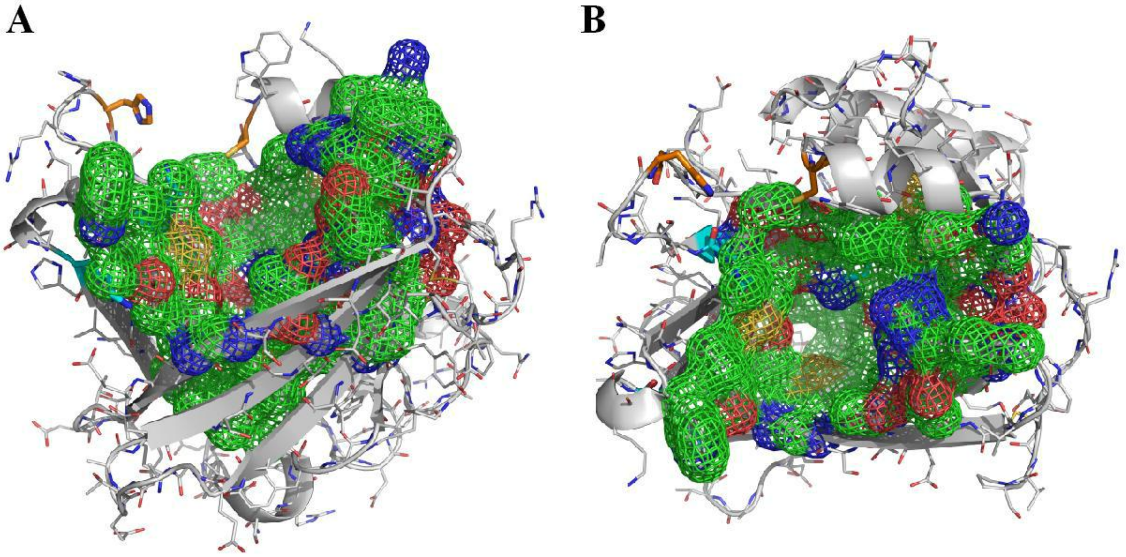


Figure 11. Identified pocket in the crystal structure of A1M. Residues lining the pocket are represented as meshes and color coded by their constituent carbon, nitrogen and oxygen atoms as green, blue and red, respectively. Residues C34 and H123 are shown in orange whereas residues K92, K118 and K130 in cyan. Secondary structure and remaining side chain carbon atoms are displayed in white. Structures are shown from the side (A) and top (B) views.

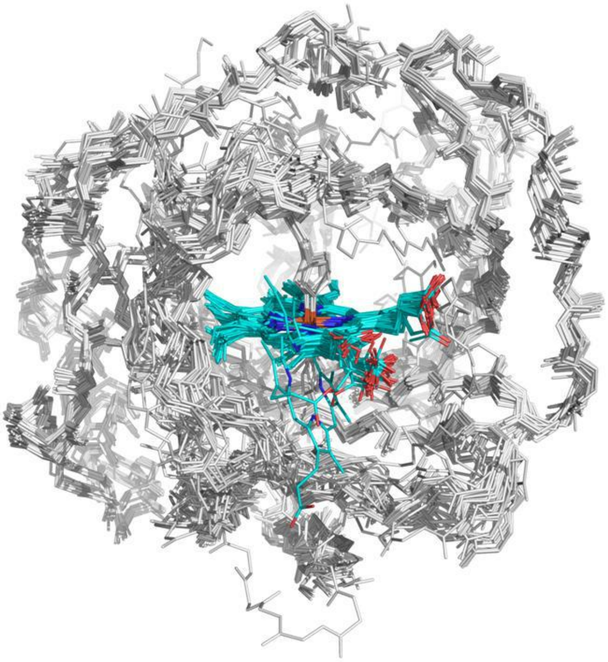


Figure 12. Structural superimposition of 45 heme-nitrophorin complexes from R. proxilus. Carbon backbone are color coded as cyan color for heme moiety and white color for the protein structure. Nitrogen and oxygen atoms are color coded as blue and red, respectively, while the iron atom of the heme moiety is shown in orange color.

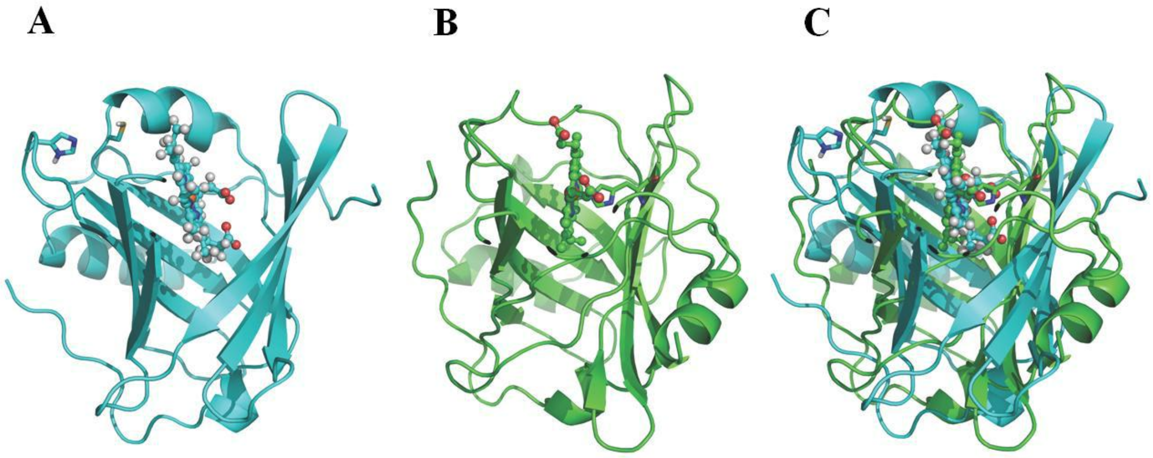


Figure 13. Structures of heme complexes with A1M (A), nitrophorin 4 from R. prolixus (B) and superimposition of both A1M and nitrophorin 4-heme complexes (C). Carbon atoms are shown in cyan and green colors in panel A and B, respectively, while oxygen and hydrogen atoms are shown in red and white, respectively.

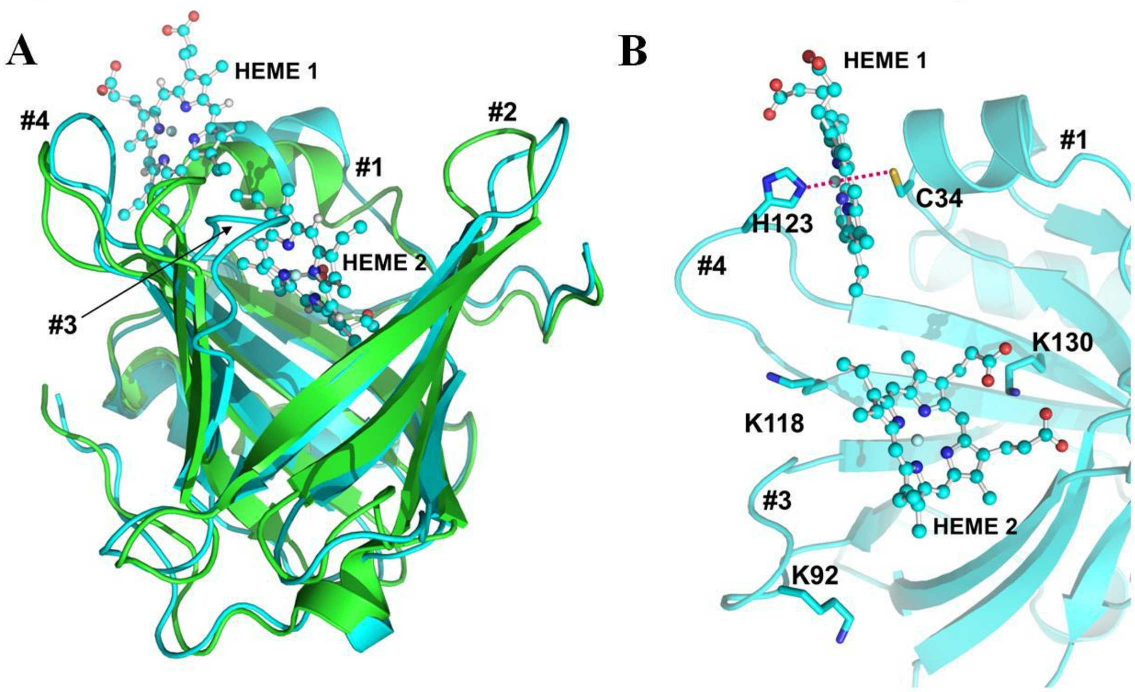


Figure 14. Modelling of binding of two heme-molecules to A1M. Superimposition between the structure of A1M complexed with 2 heme molecules and the 3QKG structure of apo-A1M (A), and a close view of two different heme binding sites of A1M (B), heme molecules are represented in a ball and stick model, coordinated residues are shown as sticks and magenta lines indicate the coordination. The four loops are labeled #1 - #4.

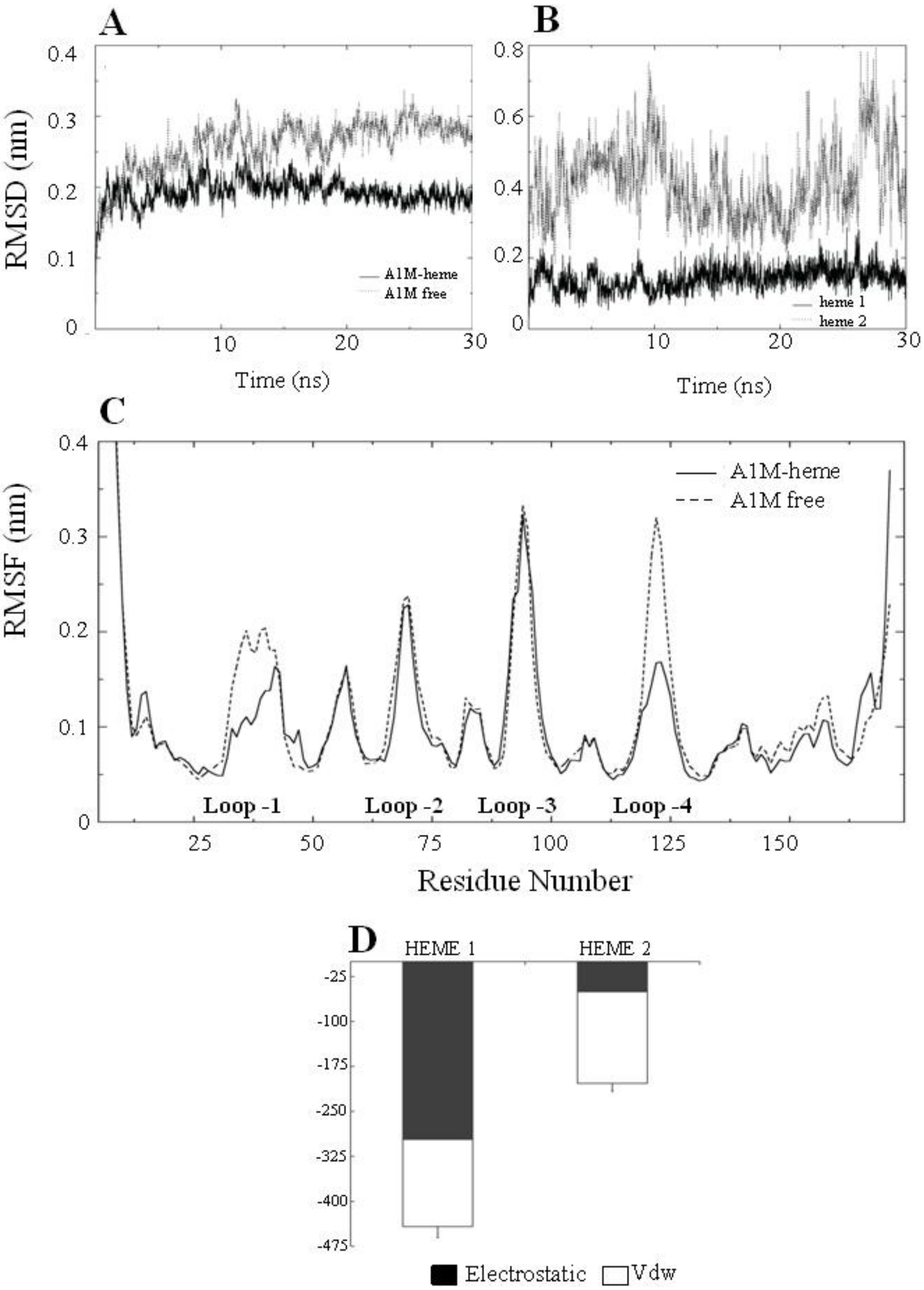


Figure 15. MD simulation data of A1M-heme complexes. (A) Time evolution of root mean square deviation (RMSD) with respect to the initial structure for the protein backbone atoms and (B) the heavy atoms of heme molecules. (C) Backbone root mean square fluctuation (RMSF) for each amino acid residue from MD simulations of heme-bound and heme-free structures of A1M. (D) Potential energies for interactions between A1M and each heme molecule. The electrostatic and vdW contributions are shown in black and white, respectively. Error bars represent the standard deviation in the sum of the interactions.

Outputs from this TRF funded research project

Publications

The knowledge and findings accumulated in this research project had been summarized in the form of 10 research articles, 3 review/editorial articles and 1 book chapter. Of the total of 14 publications, 11 have already been published while 3 manuscripts are in preparation and is anticipated to be submitted within this year.

Research articles

1. Mandi P, Shoombuatong W, Phanus-umporn C, Isarankura-Na-Ayudhya C, Prachayasittikul V, Bulow L, **Nantasenamat C***. Exploring the origins of structure-oxygen affinity relationship of human hemoglobin allosteric effector. ***Molecular Simulation*** 41 (2015) 1283-1291.
(2016 Impact Factor: 1.254)
2. Simeon S, Moller R, Almgren D, Li H, Phanus-umporn C, Prachayasittikul V, Bulow L, **Nantasenamat C***. Unraveling the origin of splice switching activity of hemoglobin-globin gene modulators via QSAR modeling. ***Chemometrics and Intelligent Laboratory Systems*** 151 (2016) 51-60.
(2016 Impact Factor: 2.303)
3. Simeon S, Phanum-umporn C, Shoombuatong W, Wikberg JES, Bulow L, **Nantasenamat C***. Predicting the oxygen affinity of human hemoglobin. ***Manuscript in Preparation*** (2017).
4. Phanus-umporn C, Anuwongcharoen N, Mandi P, Simeon S, Shoombuatong W, **Nantasenamat C***. Origin of anti-sickling activity via QSAR modeling. ***Manuscript in Preparation*** (2017).
5. Phanus-umporn C, Shoombuatong W, Choomwattana S, Anuwongcharoen N, Mandi P, Prachayasittikul V, **Nantasenamat C***. QSAR modeling of methemoglobin reduction by electron mediators. ***Manuscript in Preparation*** (2017).
6. Rutardottir S, Karnaukhova E, **Nantasenamat C**, Songtawee N, Prachayasittikul V, Rajabi M, Rosenlof L, Alayash A, Akerstrom B. Structural and biochemical characterization of two heme binding sites on α1-microglobulin using site directed mutagenesis and molecular simulation. ***Biochimica et Biophysica Acta - Proteins and Proteomics*** 1864 (2016) 29-41.
(2016 Impact Factor: 2.773)
7. Simeon S, Shoombuatong W, Anuwongcharoen N, Preeyanon L, Prachayasittikul V, Wikberg JES, **Nantasenamat C***. osFP: a web server for predicting the oligomeric states of fluorescent proteins. ***Journal of Cheminformatics*** 8 (2016) 72.
(2016 Impact Factor: 4.220)
8. Prachayasittikul V, Worachartcheewan A, Toropova AP, Toropov AA, Prachayasittikul V, **Nantasenamat C**. Large-scale classification of P-glycoprotein inhibitors using SMILES-based descriptors. ***SAR and QSAR in Environmental Research*** 28 (2017) 1-16.
(2016 Impact Factor: 1.642)

9. Simeon S, Anuwongcharoen N, Shoombuatong W, Malik AA, Prachayasittikul V, Wikberg JES, **Nantasenamat C***. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. ***PeerJ*** 4 (2016) e2322.
   (2016 Impact Factor: 2.177)
10. Rasti B, Schaduangrat N, Shahangian SS, **Nantasenamat C***. Exploring the origin of phosphodiesterase inhibition via proteochemometric modeling. ***RSC Advances*** 7 (2017) 28056-28068.
   (2016 Impact Factor: 3.108)

Review and Editorial articles

11. **Nantasenamat C***, Prachayasittikul V. Maximizing computational tools for successful drug discovery. ***Expert Opinion on Drug Discovery*** 10 (2015) 321-329.
   (2016 Impact Factor: 3.876)
12. Prachayasittikul V, Prathipati P, Pratiwi R, Phanus-umporn C, Malik AA, Schaduangrat N, Seenprachawong K, Wongchitrat P, Supokawej A, Prachayasittikul V, Wikberg JES, **Nantasenamat C***. Exploring the epigenetic drug discovery landscape. ***Expert Opinion in Drug Discovery*** 12 (2017) 345-362.
   (2016 Impact Factor: 3.876)
13. Shoombuatong W, Prathipati P, Prachayasittikul V, Schaduangrat N, Malik AA, Wanwimolruk S, Wikberg JES, Gleeson MP, Spjuth O, **Nantasenamat C***. Towards predicting the cytochrome P450 modulation: From QSAR to proteochemometric modeling. ***Current Drug Metabolism*** (2017) DOI: 10.2174/1389200218666170320121932.
   (2016 Impact Factor: 2.659)

Utilization of research results

Commercial:

The osFP web server (**Research Article 6**) has been made as a freely available software that however provides value-added service in helping protein engineers to investigate the oligomeric states of their proteins before proceeding to production phase.

Policy: -

Public:

- Conference Chairman of the 1st International Conference on Pharmaceutical Bioinformatics that was co-organized between Faculty of Medical Technology, Mahidol University and Uppsala University during January 24-26, 2016 at Centara Grand Mirage Beach Resort Pattaya, Thailand.

Academia:

This project had trained several undergraduate and graduate students as well as post-doctoral fellows and new lecturers. The knowledge and insights obtained from this research project had also been used as case studies in lectures as part of graduate level courses.

Others

Book chapter

14. Shoombuatong W, Prathipati P, Owasirikul W, Worachartcheewan A, Simeon S, Anuwongcharoen N, Wikberg JES, **Nantasenamat C***. Towards the revival of interpretable QSAR models. In: Roy K, ***Advances in QSAR modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences***. Challenges and Advances in Computational Chemistry and Physics (2017) pp. 3-55, Cham, Switzerland: Springer Nature. DOI: 10.1007/978-3-319-56850-8_1.

Talks

1. **Nantasenamat C***, Predicting the oxygen affinity of human hemoglobin, 15th International Symposium on Blood Substitutes and Oxygen Therapeutics (ISBS2015), June 21-24, 2015, Lund, Sweden
2. **Nantasenamat C***. Towards interactive and reproducible QSAR models for studying the origins of bioactivity. Symposium I: Systems Bioscience and Medicine. Systems Biosciences: Frontiers in integrative research, May 19, 2016, Institute of Molecular Biosciences, Mahidol University, Thailand.

Conference Poster Presentation

- Phanus-umporn C, Shoombuatong W, Choomwattana S, Anuwongcharoen N, Mandi P, Prachayasittikul V, **Nantasenamat C***. QSAR modeling of methemoglobin reduction by electron mediators. Poster Presentation. 1st International Conference on Pharmaceutical Bioinformatics, January 24-26, 2016, Centara Grand Mirage Beach Resort Pattaya, Thailand.

Molecular Simulation, 2015
Vol. 41, No. 15, 1283–1291, http://dx.doi.org/10.1080/08927022.2014.981180

Exploring the origins of structure–oxygen affinity relationship of human haemoglobin allosteric effector

Prasit Mandi[a,b], Watshara Shoombuatong[a], Chuleeporn Phanus-umporn[a,b], Chartchalerm Isarankura-Na-Ayudhya[b], Virapong Prachayasittikul[b], Leif Bülow[c] and Chanin Nantasenamat[a,b]*

[a]*Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand;* [b]*Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand;* [c]*Department of Pure and Applied Biochemistry, Chemical Center, Lund University, Lund, Sweden*

(*Received 4 August 2014; final version received 21 October 2014*)

A data set comprising 27 *myo*-inositol derivatives based on tetrakisphosphates and bispyrophosphates were used in the development of quantitative structure–activity relationship model for investigating its allosteric effector property against human haemoglobin (Hb). Three-dimensional structures of the investigated compounds were subjected to geometry optimisations at the density functional theory level. Physicochemical features of low-energy conformers were represented by quantum chemical and molecular descriptors. Feature selection by means of unsupervised forward selection and stepwise linear regression resulted in a set of four important descriptors. Multivariate analysis was performed using multiple linear regression (MLR), artificial neural network (ANN) and support vector machine (SVM). Robustness of the predictive performance of all methods was deduced from internal and external validation, which afforded Q^2_{CV} values of 0.6306, 0.7484 and 0.8722 using MLR, ANN and SVM, respectively, for the former and Q^2_{Ext} values of 0.8332, 0.8847 and 0.9694, respectively, for the latter. The predictive model is anticipated to be useful for further guiding the rational design of robust allosteric effectors of human Hb.

Keywords: haemoglobin; allosteric effectors; inositol; QSAR; data mining

1. Introduction

Several diseases (i.e. anaemia, cancer, cardiovascular ailments, haemorrhages, ischaemia and haemoglobinopathy H disease) are characterised by the insufficient supply of oxygen in peripheral tissues,[1–5] a condition known as hypoxia. Therefore, the ability to increase the delivery of oxygen by red blood cells to tackle the problem of hypoxia affords great therapeutic interest.[6,7]

It is generally accepted that a certain class of small-molecule known as allosteric effectors can modulate the oxygen-binding property of haemoglobin (Hb).[8–10] Particularly, these molecules bind preferentially to the larger central cavity of the T-state (when compared with the R-state) followed by stabilising this conformation effectively shifting the R/T equilibrium and consequently lowering the oxygen affinity of Hb.[11] The endogenous allosteric effector of Hb, 2,3-bisphosphoglycerate (BPG), binds the allosteric cavity of Hb with a dissociation constant of 1.5×10^{-5} M. This is followed by a concomitant right-shift of the allosteric equilibrium that decreases its oxygen-binding affinity as BPG stabilises the deoxy T state via formation of intermolecular salt bridges between the two β-chains.[11]

The search for novel and robust allosteric effectors has attracted much attention owing to its great therapeutic potential. Over the years, several structural classes of allosteric effectors of Hb had been discovered encompassing those based on the organic phosphates such as *myo*-inositol phosphates (i.e. *myo*-inositol hexakisphosphate, IHP; *myo*-inositol trispyrophosphate, ITPP) [12] as well as aromatic propionates such as the antilipidaemic fibrate agents (i.e. clofibrate, bezafibrate, BZF; BZF derivative, RSR-13; and BZF urea derivative, L35).[13] The former class represents one of the first synthetic allosteric effectors to be studied in which IHP, a commonly available natural product, was found to bind 1000 times more potently to the β-cleft of deoxygenated Hb.[14] Particularly, IHP displaces Hb-bound BPG consequently leading to lowered oxygen affinity resulting in increased and regulated release of oxygen to tissues.[15] The group of Fylaktakidou et al. [16] introduced an IHP derivative, known as ITPP, bearing three seven-membered cyclic pyrophosphate rings, and the compound was shown to strongly increase the oxygen release *in vitro* from both free Hb and red blood cells. Further studies demonstrated broad applicability of the compound for treating a wide range of disease for which tissues are in need of oxygen (i.e. ischaemic heart disease, cardiovascular disease and tumour progression).[17–19] The success of IHP and ITPP led the group of Koumbis et al. [12] to further extend their work towards the synthesis of *myo*-inositol derivatives encompassing inositol tetraphosphates (ITPs)

*Corresponding author. Email: chanin.nan@mahidol.ac.th and inositol bispyrophosphates (IBPPs) as allosteric effectors of human Hb.

Quantitative structure–activity/property relationship (QSAR/QSPR) is a computational methodology for discerning the inherent linear or non-linear relationship between a set of structural features of investigated molecules with their respective biological activity/chemical property.[20,21] QSAR/QSPR had been successfully utilised to address a wide range of problems of biological [22–27] and chemical [28–31] importance. The essential steps in the construction of QSAR/QSPR models entail the following procedures: (i) compilation of data set of interest, (ii) optimisation of geometrical structures of molecules (if computing 3D features), (iii) computation of molecular descriptors, (iv) selection of a subset of molecular descriptors either through chemical intuition or computational optimisation, (v) division of data set into internal and external sets and (vi) development of a predictive model using the internal set and evaluation of predictivity on the external set.

Preliminary QSAR study by Hansch et al. [32] had touched upon modelling the allosteric interaction of alkylisonitriles (RN=C) with Hb by focusing on the hydrophobic properties of ligands responsible for such interactions. Therefore, there has not yet been any reported in-depth QSAR investigation for studying the allosteric effectors of human Hb for further utilisation in remedying hypoxia-related diseases. To the best of our knowledge, this study represents the first QSAR model focused on unravelling the origins of allosteric effector activity on human Hb. This was achieved by compiling a data set of 27 *myo*-inositol derivatives based on tetrakisphosphates and bispyrophosphates reported by the group of Koumbis et al. [12]. Physicochemical features of investigated compounds, after feature selection, were described by a set of five molecular descriptors. QSAR models developed by several multivariate methods afforded good predictive performance as verified by internal and external validation.

2. Material and methods

2.1. Data set

A data set of 27 *myo*-inositol derivatives and their allosteric effector activity against human Hb were obtained from Koumbis et al. [12]. The allosteric activity was represented by the P_{50} value, which is a conventional measure of Hb affinity for oxygen. An increase in the P_{50} value indicates a rightward shift of the standard curve thereby suggesting that a larger partial pressure is necessary to maintain 50% oxygen saturation thereby suggesting a decreased affinity. Conversely, a lower P_{50} leads to a leftward shift corresponding to higher oxygen affinity. To achieve uniform distribution of data samples, the P_{50} values were subjected to data transformation by calculating its logarithmic values to the base of 10.

2.2. Geometry optimisation and descriptor calculation

Chemical structures of investigated compounds were drawn using ChemAxon Marvin version 6.2.1 [33] and their molecular geometries were optimised at the density functional theory level using Becke's three-parameter Lee–Yang–Parr hybrid functional (B3LYP) in concomitant with the 6-311 + + G(d,p) basis set. Quantum chemical calculation was performed using Gaussian 09 [34] to derive a set of quantum chemical descriptors, which was obtained from the low energy conformer, comprising the total energy of the molecule, highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, dipole moment (μ), electron affinity, ionisation potential, energy difference of HOMO and LUMO states (HOMO–LUMO), Mulliken electronegativity (χ), hardness (η), softness (S), electrophilicity (ω), electrophilic index (ω_i), most negative atom in the molecule (Q_{neg}), most positive atom in the molecule (Q_{pos}) and the mean absolute atomic charge (Q_m).

Low energy conformers were subjected for further generation of an additional set of 3224 molecular descriptors using DRAGON version 5.5 [35]. This descriptor set spanned 22 categories comprising 48 constitutional descriptors, 119 topological descriptors, 47 walk and path counts, 33 connectivity indices, 47 information indices, 96 2D autocorrelation, 107 edge adjacency indices, 64 Burden eigenvalues, 21 topological charge indices, 44 eigenvalue-based indices, 41 randic molecular profiles, 74 geometrical descriptors, 150 RDF descriptors, 160 3D-MoRSE descriptors, 99 WHIM descriptors, 197 GETAWAY descriptors, 154 functional group counts, 120 atom-centred fragments, 14 charge descriptors, 29 molecular properties, 780 2D binary fingerprints and 780 2D frequency fingerprints.

2.3. Feature selection

Descriptors having constant value and variable pairs with correlation coefficient greater than 0.9 were subjected to removal using the Unsupervised Forward Selection (UFS) algorithm.[36] Additional round of feature selection was performed using stepwise linear regression as calculated by SPSS Statistics 18.0 [37]. This led to the selection of important descriptors that will be subsequently used in correlating with Hb allosteric activity of *myo*-inositol derivatives.

2.4 Data splitting

In order to obtain accurate and generalised QSAR models, the 27 compounds were divided into two parts comprising internal set for constructing QSAR models using the leave-one-out cross-validation approach while the subset of the remaining seven compounds (i.e. **12d**, **17c**, **22**, **24d**, **26a**, **26c** and **24b**) was used as the external set, which was sampled according to the Kennard–Stone algorithm [38] (Table 1).

2.5. *Multivariate analysis*

Physicochemical features of investigated *myo*-inositol derivatives as represented by selected molecular descriptors were correlated with their respective Hb allosteric activity using multiple linear regression (MLR), artificial neural network (ANN) and support vector machine (SVM).

MLR is a classical multivariate approach that linearly correlates a set of independent variables (i.e. molecular descriptors) with the dependent variable of interest (i.e. allosteric activity).

ANN was performed using the back-propagation algorithm in which the residual error from prediction are propagated in a backward fashion from the output layer through the hidden layer and finally onto the input layer followed by readjustment of weights interconnecting the neurons. This process is carried out iteratively until convergence is reached.

SVM is a statistical learning approach proposed by Vapnik [39,40] that utilises kernel functions to transform data to higher dimension whereby SVM performs its learning and decision in a linear manner. This study employed the radial basis function kernel for such data transformation.

Table 1. Data set of *myo*-inositol derivatives.

Compound	log P_{50}	nHDon	TPSA	ALogP	Q_{pos}
12a[a]	1.5038	4	336.06	− 6.824	0.5939
12b[a]	1.5224	4	336.06	− 6.126	0.5870
12c[a]	1.6684	4	336.06	− 4.166	0.5982
12d[b]	1.4654	4	370.20	− 6.882	0.5663
12e[a]	1.6274	4	370.20	− 4.636	0.6294
12f[a]	1.6920	6	358.06	− 7.641	0.5717
17a[a]	1.3962	4	336.06	− 6.824	0.5289
17b[a]	1.4594	4	336.06	− 6.126	0.5210
17c[b]	1.6561	4	336.06	− 4.166	0.5343
17d[a]	1.4997	4	370.20	− 6.882	0.5787
17e[a]	1.6464	4	370.20	− 4.636	0.5109
17f[a]	1.5105	6	358.06	− 7.641	0.5010
22[b]	1.6064	6	358.06	− 7.641	0.5520
24a[a]	1.1139	0	273.60	− 5.291	0.5765
24b[b]	1.0792	0	273.60	− 4.593	0.5855
24c[a]	1.0864	0	273.60	− 2.633	0.4830
24d[b]	1.1004	0	307.74	− 5.349	0.5453
24e[a]	1.0864	0	307.74	− 3.103	0.4879
24f[a]	1.2856	2	295.60	− 6.108	0.5587
26a[b]	1.1004	0	273.60	− 5.291	0.5419
26b[a]	1.0531	0	273.60	− 4.593	0.5540
26c[b]	1.1523	0	273.60	− 2.633	0.5160
26d[a]	1.1461	0	307.74	− 5.349	0.5459
26e[a]	1.0828	0	307.74	− 3.103	0.3866
BPG[a]	1.2430	0	204.59	− 4.833	0.6046
IHP[a]	1.7612	0	493.38	− 11.932	0.6081
ITPP[a]	1.3464	0	382.71	− 7.629	0.5806

[a] Internal set.
[b] External set.

All multivariate analyses were performed using Weka, version 3.4.2.[41] Optimal parameters were determined empirically. Particularly, the number of nodes in the hidden layer, number of learning epochs, learning rate and momentum were subjected to optimisation for ANN calculations, whereas C and γ parameters were optimised for SVM classifier.

Y-scrambling is a widely used approach in order to ensure the robustness of the QSAR model.[42] This was performed by randomly shuffling the Y-variable while keeping the X-variable intact followed by computing the Y-scrambled model again. It is expected that the resulting models should have low R^2_{Tr} and Q^2_{CV}. Y-scrambling was performed for 100 runs using the R program.[43] Furthermore, the statistical metric r^2_{m} as proposed by Roy et al. [44] was also utilised to evaluate the robustness of QSAR models in which favourable models should afford $r^2_{\mathrm{m}} > 0.5$ and that Δr^2_{m} should be < 0.2.

Principal component analysis (PCA) was performed using FactoMineR in R [45] to visualise the chemical space of investigated compounds.

3. Results and discussion

3.1. *Chemical space of* myo-*inositol derivatives*

myo-Inositol derivatives employed in this study comprised ITPs and IBPPs in which the former class constitutes substituted and unsubstituted inositol(1,3,4,6)P_4 (**12a–f**) and (±)-inositol(2,3,5,6)P_4 (**17a–f**) along with (±)-inositol(1,2,3,4)P_4 (**22**) while the latter class constitutes inositol(1,6:3,4)BPP (**24a–f**) and (±)-inositol(2,3:5,6) BPP (**26a–e**). Chemical structures of investigated compounds are shown in Figure 1. These compounds were evaluated by Koumbis et al. [12] for their ability to shift the oxygen saturation curve to higher $p\mathrm{O}_2$ values and subsequently assess the partial pressure of oxygen for half-saturation (P_{50}) of Hb when bound to the allosteric effectors. This study investigates the origins of allosteric effector properties by means of predictive QSAR modelling as summarised in Figure 2.

Visual representation of the overall distribution of data values for P_{50} and Lipinski's descriptors is shown as 3D bar plots in Figure 3. In general, the affinity of compounds towards binding Hb as deduced from P_{50} values revealed the following trend for the four class of compounds: **12** > **17** > **24** > **26**. It can be seen from Figure 3(A) that ITPs (i.e. **12** and **17**) provided higher affinity than their IBPP (i.e. **24** and **26**) counterparts.

Lipinski's descriptors comprising molecular weight (MW), molecular lipophilicity (ALogP), number of hydrogen bond donors (nHDon) and the number of hydrogen bond acceptors (nHAcc) (Table 1) were

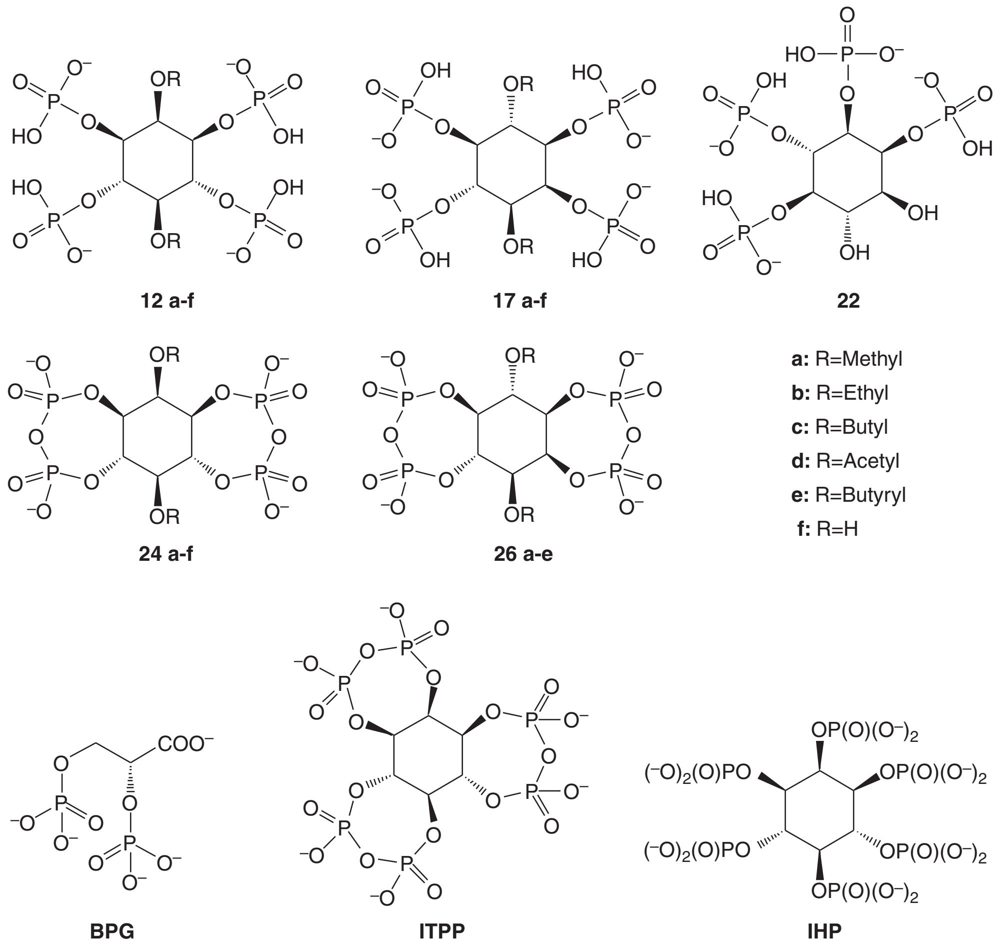


Figure 1. Chemical structures of ITPs (**12a–f**, **17a–f** and **22**) and IBBPs (**24a–f** and **26a–e**).

analysed in order to understand the general properties of these class of compounds. A notable characteristic distinguishing ITPs (**12** and **17**) from IBPPs (**24** and **26**) is the presence of higher nHDon (Figure 3(D)) and nHAcc (Figure 3(E)) in the former class. As hydrogen bond and electrostatic interaction are key binding mechanisms for natural (2,3-BPG) [46] and synthetic [12] allosteric effectors, ITPs were more effective than IBPPs plausibly due to the fact that these set of compounds are highly capable of forming hydrogen bonds than their IBBP counterparts. Moreover, ITPs **12c** and **17c** followed by **12e** and **17e** were the most potent amongst the investigated compounds. These compounds also have higher ALogP (Figure 3(B)) and MW (Figure 3(C)) than the rest of the compounds thereby implying that these properties may influence allosteric activity of these sets of effectors.

Feature selection using UFS followed by stepwise linear regression identified five important descriptors consisting of nHDon, nHAcc, ALogP, topological polar surface area (TPSA) and Q_{pos} (Table 1). Interestingly, aside from TPSA and Q_{pos}, the remaining three (i.e. nHDon, nHAcc and ALogP) of the selected descriptors were Lipinski's descriptors. TPSA refers to the surface area of oxygen, nitrogen, sulphur and attached hydrogen atoms while Q_{pos} pertained to the most positive atomic charge. Thus, the selected descriptors implied the importance of hydrogen bonds and electrostatic properties for the activity of allosteric effectors.

3.2. *QSAR model of* myo-*inositol derivatives using MLR*

In this study, the set of four important descriptors (Table 1) as selected by feature selection were used as independent variables while allosteric effector activity (i.e. log P_{50}) was used as the dependent variable. The data set was separated into internal and external sets to assess their internal and external predictive performances, respectively. Figure 4 displays the PCA plot of data points in the internal (red) and external (blue) sets in which the first two components

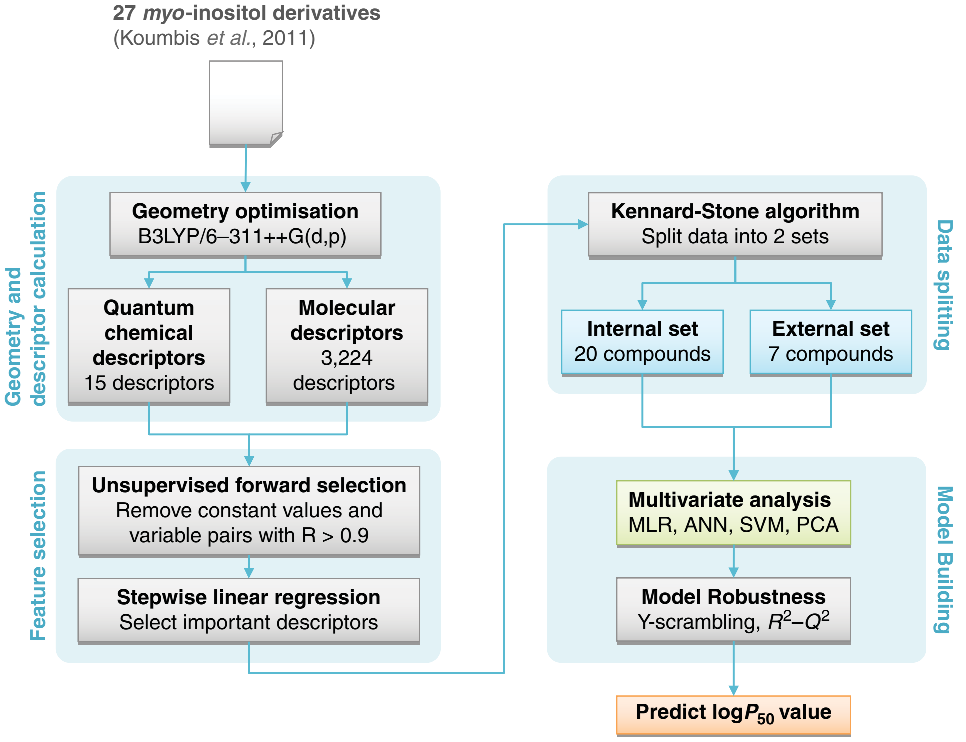


Figure 2. (Colour online) Schematic of the QSAR modelling workflow.

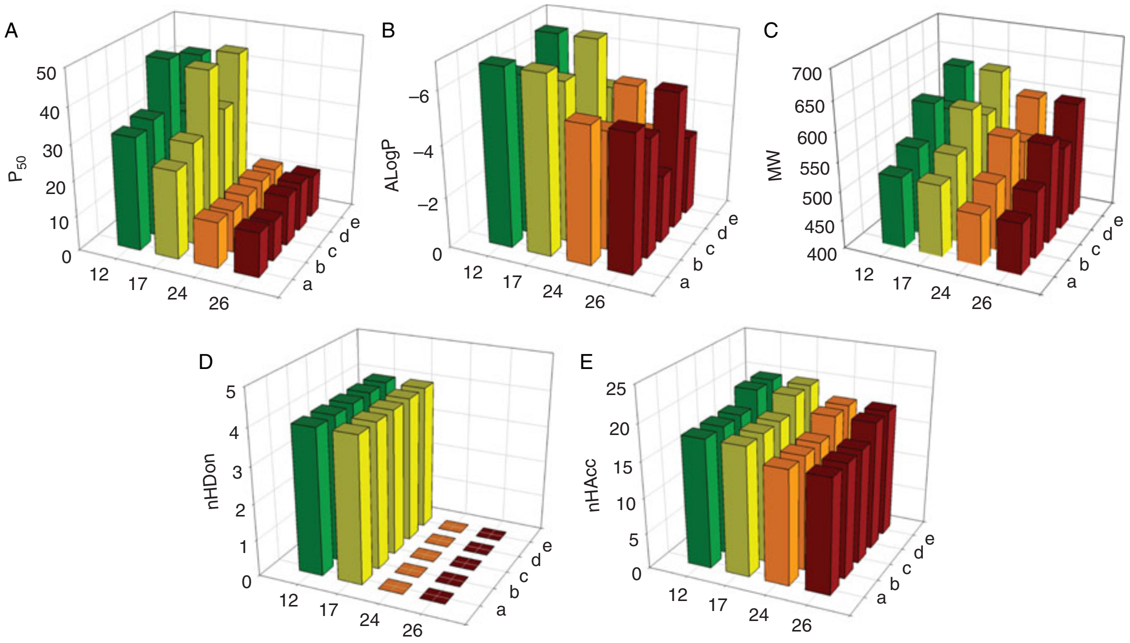


Figure 3. (Colour online) Three-dimensional bar plots showing overall distribution of data values for P_{50} (A) and Lipinski's descriptors (B–E).

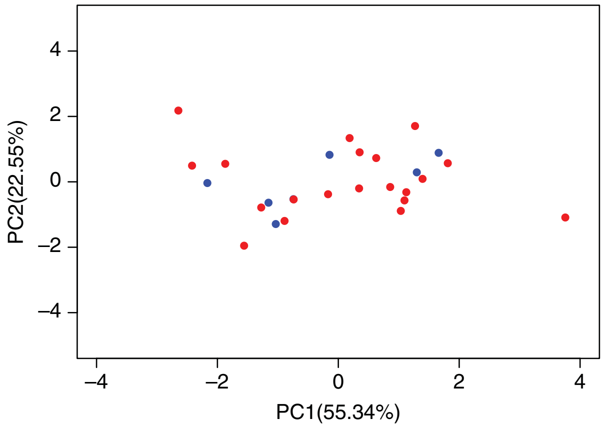


Figure 4. (Colour online) Chemical space spanned by internal and external sets as shown in blue and red colours, respectively.

explain 77.89% of the variance afforded by the 27 compounds.

MLR model was constructed using ridge parameter of $R = 1.0\text{E} - 8$. The predictive performance of the constructed MLR models is shown in Table 2 and the MLR equation is shown in the following along with their respective statistical properties.

$$\begin{aligned}\log P_{50} &= 0.0517(\text{nHDon}) + 0.0023(\text{TPSA})\\ &\quad + 0.0081(\text{ALogP}) + 1.3373(Q_{\text{pos}})\\ &\quad - 0.1697, \qquad (1)\end{aligned}$$

where $n = 20$, $R^2_{\text{Tr}} = 0.8859$, $Q^2_{\text{CV}} = 0.6306$, $Q^2_{\text{Ext}} = 0.8332$, $\text{RMSE}_{\text{Tr}} = 0.0775$, $\text{RMES}_{\text{CV}} = 0.1487$, $\text{RMSE}_{\text{Ext}} = 0.1015$.

As deduced from regression coefficient, the most important molecular descriptors were $Q_{\text{pos}} > \text{nHDon} > \text{ALogP} > \text{TPSA}$, which displayed corresponding values of 1.3373, 0.0517, 0.0081 and 0.0023, respectively. The MLR equation suggested that high potent allosteric effector in order to release oxygen from Hb should have higher positively atomic charge, lipophilicity number of hydrogen bond donor and topological polar surface area. Plot of experimental versus predicted activities for investigated compounds is shown in Figure 5(A).

Table 2. Summary of the predictive performance of MLR, ANN and SVM models for predicting the Hb allosteric effector activity of *myo*-inositol derivatives.

Methods	R^2_{Tr}	Q^2_{CV}	Q^2_{Ext}	RMSE_{Tr}	RMSE_{CV}	RMSE_{Ext}
MLR	0.8859	0.6306	0.8332	0.0775	0.1487	0.1015
ANN	0.9235	0.7484	0.8847	0.0642	0.1176	0.0827
SVM	0.9876	0.8722	0.9694	0.0298	0.0827	0.0465

3.3. *QSAR model of* myo-*inositol derivatives using ANN and SVM*

In addition to MLR which is a linear machine learning approach, ANN and SVM, which are nonlinear approaches, were used for constructing more sophisticated QSAR model using the same data set. Parameter optimisation of ANN identified appropriate hidden node of 1, learning epoch of 1000 cycles, learning rate of 0.1 and momentum of 0.4 (Figure 6(A)–(C)). It can be observed from Table 2 that ANN model provided good predictive performance with the following parameters: $R^2_{\text{Tr}} = 0.9235$, $Q^2_{\text{CV}} = 0.7484$, $Q^2_{\text{Ext}} = 0.8847$, $\text{RMSE}_{\text{Tr}} = 0.0642$, $\text{RMES}_{\text{CV}} = 0.1176$ and $\text{RMSE}_{\text{Ext}} = 0.0827$. SVM model building was initiated by searching for optimal C and γ parameters. The search comprised two-levels including initial global grid search followed by a refined local grid search. Global grid search identified optimal C and γ values as 2^3 and 2^{-1}, respectively, while the refined local grid search identified optimal C and γ values of $2^{4.4}$ and $2^{-1.4}$, respectively (Figure 7(A),(B)). Furthermore, Table 2 shows the performance of SVM, which provided the following statistical parameters: $R^2_{\text{Tr}} = 0.9876$, $Q^2_{\text{CV}} = 0.8722$, $Q^2_{\text{Ext}} = 0.9694$, $\text{RMSE}_{\text{Tr}} = 0.0298$, $\text{RMES}_{\text{CV}} = 0.0827$ and $\text{RMSE}_{\text{Ext}} = 0.0465$. Plot of experimental versus predicted activities of the investigated compounds as predicted by ANN and SVM is shown in Figure 5(B),(C), respectively.

As a result, it can be seen that all multivariate methods provided good performance in predicting the $\log P_{50}$ values of the investigated ITPs and IBBPs as Hb allosteric effector. This is verified by both internal and external validations of the predictive QSAR models. Eriksson and Johansson [47] proposed a metric based on $R^2 - Q^2$ for describing the fraction of Y-data explained by accumulated chance correlations in which values greater than 0.2–0.3 suggests the risk of chance correlations, the presence of outliers or irrelevant descriptors in the data set, or the possibility of overfitting model. It can be seen that all multivariate methods afforded $R^2 - Q^2$ well below the aforementioned threshold where MLR, ANN and SVM had values of 0.2553, 0.1751 and 0.1155, respectively.

To further verify the validity of QSAR models, Y-scrambling experiments were performed in concomitant with the utilisation of the r^2_{m} metric. It can be seen in Figure 8 that QSAR models developed using SVM afforded distinct difference in the distribution of data points of Y-scrambled models from the actual one. It is also observed that several Y-scrambled models afforded significantly large difference of $R^2 - Q^2$ suggesting its expected inadequacy in properly modelling the activity. Furthermore, it can also be observed that a few data points

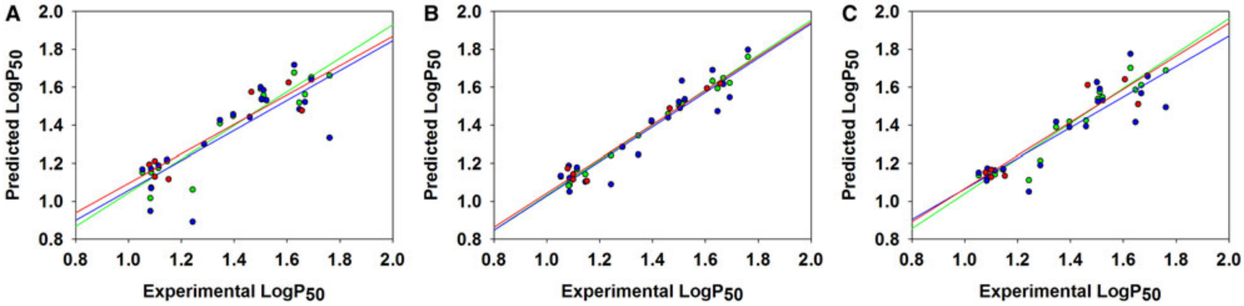


Figure 5. (Colour online) Plot of experimental versus predicted activities for investigated compounds using MLR (A), ANN (B) and SVM (C).

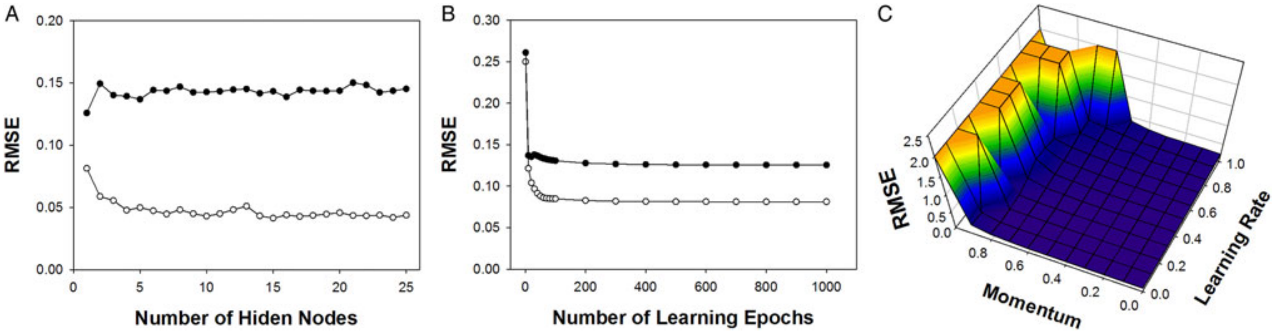


Figure 6. (Colour online) Optimisation of ANN parameters consisting of the number of nodes in the hidden layer (A), number of learning epochs (B) and the learning rate and momentum (C).

displayed higher Q^2 than their respective R^2. A closer analysis revealed that this model afforded the typically higher RMSE value for the CV set with respect to the training set while giving rise to highly negative Q values, which would in turn produce seemingly high Q^2 value, thereby suggesting that the scrambled models could not model the activity. Moreover, r_m^2 values for MLR, ANN and SVM were found to be 0.50, 0.68 and 0.84, respectively, whereas corresponding values for Δr_m^2 were 0.01, 0.06 and 0.05, respectively. These results suggested that SVM and ANN models provided good predictive performance. However, these two models are black-box models that are not readily interpretable. In spite of its lower predictive performance, the MLR model afforded essential insights on significant descriptors giving rise to the allosteric effector properties. Errors arising from these

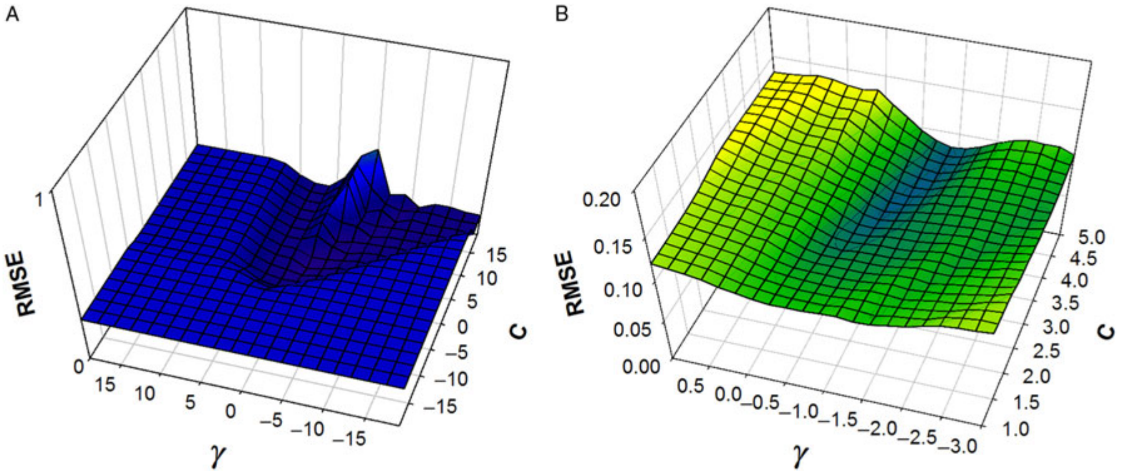


Figure 7. (Colour online) Optimisation of SVM parameters by means of a global (A) and local (B) search.

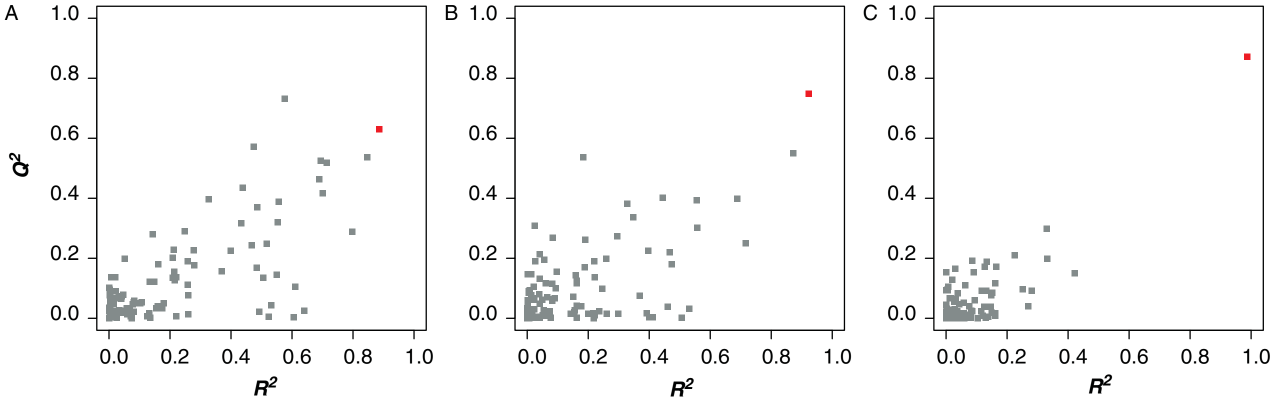


Figure 8. (Colour online) Plot of R^2_{Tr} and Q^2_{CV} from Y-scrambling experiments using MLR (A), ANN (B), and SVM (C). Grey and red points represent results from Y-scrambled and actual models, respectively.

predictive models could be attributed to several factors including inherent experimental error as well as errors arising from the modelling process (i.e. limitation of learning method, sub-optimal descriptors from feature selection, sub-optimal learning parameters, applicability domain, etc.). In light of these potential factors, this study tries to address all of the aforementioned points in development of the QSAR model such as employing PCA to visualise the applicability domain of internal and external sets, utilising feature selection to select the most significant descriptors, performing parameter optimisation to obtain the best performing model as well as employing several machine learning methods.

Practical utilisation of the QSAR models developed herein could be implemented by calculating molecular descriptors for new, unknown compounds followed by predicting their biological activity using the multivariate analysis methods. It should also be noted that to ensure reliable predictions, the applicability domain should also be determined by evaluating the molecular similarity (i.e. Tanimoto coefficient) of new compounds in relation to the constituting compounds from the model proposed herein. Furthermore, as more data become available, such information could potentially be used to update the QSAR model.

4. Conclusion

In summary, this study developed QSAR models for investigating the allosteric effector property of ITPs and IBBPs against human Hb. Feature selection using UFS and stepwise linear regression method identified four important descriptors, which were selected from a total of 3239 descriptors. Selected descriptors indicated the importance of hydrogen bonding capacity, lipophilicity and electrostatic properties, which were in agreement with the chemical space analysis performed herein. The QSAR approach presented herein could provide useful information on origins of allosteric effector properties that could further be used to guide the design of novel Hb allosteric effectors.

Funding

This research is supported by a Research Career Development Grant [grant number RSA5780031] from the Thailand Research Fund to C.N. Research assistantship from Mahidol University to P.M. is gratefully acknowledged. W.S. is grateful for the Mahidol University Postdoctoral Fellowship under the supervision of C.N. and V.P.

References

[1] Sun K, Xia Y. New insights into sickle cell disease: a disease of hypoxia. Curr Opin Hematol. 2013;20:215–221.
[2] Rundqvist H, Johnson RS. Tumour oxygenation: implications for breast cancer prognosis. J Int Med. 2013;274:105–112.
[3] Lenihan CR, Taylor CT. The impact of hypoxia on cell death pathways. Biochem Soc Trans. 2013;41:657–663.
[4] Voelkel NF, Mizuno S, Bogaard HJ. The role of hypoxia in pulmonary vascular diseases: a perspective. Am J Physiol Lung Cell Mol Physiol. 2013;304:L457–L465.
[5] Papassotiriou I, Kister J, Griffon N, Abraham DJ, Kanavakis E, Traeger-Synodinos J, Stamoulakatou A, Marden MC, Poyart C. Synthesized allosteric effectors of the hemoglobin molecule: a possible mechanism for improved erythrocyte oxygen release capability in hemoglobinopathy H disease. Exp Hematol. 1998;26: 922–926.
[6] Mairbaurl H, Weber RE. Oxygen transport by hemoglobin. Compr Physiol. 2012;2:1463–1489.
[7] Crawford JH, Chacko BK, Kevil CG, Patel RP. The red blood cell and vascular function in health and disease. Antioxid Redox Signal. 2004;6:992–999.
[8] Lalezari I, Lalezari P, Poyart C, Marden M, Kister J, Bohn B, Fermi G, Perutz MF. New effectors of human hemoglobin: structure and function. Biochemistry. 1990;29:1515–1523.
[9] Perutz MF, Fermi G, Luisi B, Shaanan B, Liddington RC. Stereochemistry of cooperative mechanisms in hemoglobin. Acc Chem Res. 1987;20:309–321.
[10] Safo MK, Bruno S. Allosteric effectors of hemoglobin: past, present and future. Chemistry and biochemistry of oxygen therapeutics:

from transfusion to artificial blood. West Sussex: John Wiley & Sons, Ltd.; 2011.

[11] Arnone A. X-ray diffraction study of binding of 2,3-diphosphoglycerate to human deoxyhaemoglobin. Nature. 1972;237: 146–149.

[12] Koumbis AE, Duarte CD, Nicolau C, Lehn JM. Tetrakisphosphates and bispyrophosphates of *myo*-inositol derivatives as allosteric effectors of human hemoglobin: synthesis, molecular recognition, and oxygen release. ChemMedChem. 2011;6:169–180.

[13] Randad RS, Mahran MA, Mehanna AS, Abraham DJ. Allosteric modifiers of hemoglobin. 1. Design, synthesis, testing, and structure–allosteric activity relationship of novel hemoglobin oxygen affinity decreasing agents. J Med Chem. 1991;34:752–757.

[14] Yonetani T, Park SI, Tsuneshige A, Imai K, Kanaori K. Global allostery model of hemoglobin. Modulation of O(2) affinity, cooperativity, and Bohr effect by heterotropic allosteric effectors. J Biol Chem. 2002;277:34508–34520.

[15] Teisseire BP, Ropars C, Vallez MO, Herigault RA, Nicolau C. Physiological effects of high-P_{50} erythrocyte transfusion on piglets. J Appl Physiol. 1985;8:1810–1817.

[16] Fylaktakidou KC, Lehn JM, Greferath R, Nicolau C. Inositol tripyrophosphate: a new membrane permeant allosteric effector of haemoglobin. Bioorg Med Chem Lett. 2005;15:1605–1608.

[17] Kieda C, Greferath R, Crola da Silva C, Fylaktakidou KC, Lehn JM, Nicolau C. Suppression of hypoxia-induced HIF-1alpha and of angiogenesis in endothelial cells by *myo*-inositol trispyrophosphate-treated erythrocytes. Proc Natl Acad Sci USA. 2006;103: 15576–15581.

[18] Biolo A, Greferath R, Siwik DA, Qin F, Valsky E, Fylaktakidou KC, Pothukanuri S, Duarte CD, Schwarz RP, Lehn JM, Nicolau C, Colucci WS. Enhanced exercise capacity in mice with severe heart failure treated with an allosteric effector of hemoglobin, *myo*-inositol trispyrophosphate. Proc Natl Acad Sci USA. 2009;106:1926–1929.

[19] Sihn G, Walter T, Klein JC, Queguiner I, Iwao H, Nicolau C, Lehn JM, Corvol P, Gasc JM. Anti-angiogenic properties of *myo*-inositol trispyrophosphate in ovo and growth reduction of implanted glioma. FEBS Lett. 2007;581:962–966.

[20] Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure–activity relationship. Excli J. 2009;8:74–88.

[21] Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds. Exp Opin Drug Discov. 2010;5:633–654.

[22] Mandi P, Nantasenamat C, Srungboonmee K, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR study of anti-prion activity of 2-aminothiazoles. Excli J. 2012;11:453–467.

[23] Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Prediction of bond dissociation enthalpy of antioxidant phenols by support vector machine. J Mol Graph Model. 2008;27: 188–196.

[24] Nantasenamat C, Piacham T, Tantimongcolwat T, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR model of the quorum-quenching *N*-acyl-homoserine lactone lactonase activity. J Biol Syst. 2008;16:279–293.

[25] Thippakorn C, Suksrichavalit T, Nantasenamat C, Tantimongcolwat T, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Modeling the LPS neutralization activity of anti-endotoxins. Molecules. 2009;14:1869–1888.

[26] Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul S, Prachayasittikul V. Predicting the free radical scavenging activity of curcumin derivatives. Chemometr Intell Lab Syst. 2011;109:207–216.

[27] Worachartcheewan A, Nantasenamat C, Naenna T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Modeling the activity of furin inhibitors using artificial neural network. Eur J Med Chem. 2009;44: 1664–1673.

[28] Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. Quantitative structure–imprinting factor relationship of molecularly imprinted polymers. Biosens Bioelectron. 2007;22: 3309–3317.

[29] Nantasenamat C, Isarankura-Na-Ayudhya C, Tansila N, Naenna T, Prachayasittikul V. Prediction of GFP spectral properties using artificial neural network. J Comput Chem. 2007;28:1275–1289.

[30] Nantasenamat C, Naenna T, Isarankura Na Ayudhya C, Prachayasittikul V. Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network. J Comput Aided Mol Des. 2005;19:509–524.

[31] Nantasenamat C, Srungboonmee K, Jamsak S, Tansila N, Isarankura-Na-Ayudhya C, Prachayasittikul V. Quantitative structure–property relationship study of spectral properties of green fluorescent protein with support vector machine. Chemometr Intell Lab Syst. 2013;120:42–52.

[32] Hansch C, Garg R, Kurup A, Mekapati SB. Allosteric interactions and QSAR: on the role of ligand hydrophobicity. Bioorg Med Chem. 2003;11:2075–2084.

[33] ChemAxon Ltd. MarvinSketch, Version 6.2.1. Budapest: ChemAxon Ltd; 2014.

[34] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA, Jr, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam NJ, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ. Gaussian 09. Revision A.1. Connecticut; 2009.

[35] Talete srl. DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.5. Milano; 2007.

[36] Whitley DC, Ford MG, Livingstone DJ. Unsupervised forward selection: a method for eliminating redundant variables. J Chem Inf Comput Sci. 2000;40:1160–1168.

[37] IBM Corporation. SPSS Statistics, Version 18. New York; 2011.

[38] Kennard RW, Stone LA. Computer aided design of experiments. Technometrics. 1969;11:137–148.

[39] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995; 20:273–297.

[40] Vapnik VN. Statistical learning theory. New York: Wiley-Interscience; 1998.

[41] Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 3rd ed. Massachusetts: Morgan Kaufmann Publishers Inc.; 2005.

[42] Rücker C, Rücker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model. 2007;47:2345–2357.

[43] Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5:299–314.

[44] Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN. Some case studies on application of "r_{m}^2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. J Comput Chem. 2013;34:1071–1082.

[45] Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Softw. 2008;25:1–18.

[46] Nadolny C, Kempf I, Zundel G. Specific interactions of the allosteric effector 2,3-bisphosphoglycerate with human hemoglobin – a difference FTIR study. Biol Chem Hoppe Seyler. 1993;374:403–407.

[47] Eriksson L, Johansson E. Multivariate design and modeling in QSAR. Chemometr Intell Lab Syst. 1996;34:1–19.

Chemometrics and Intelligent Laboratory Systems 151 (2016) 51–60

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Unraveling the origin of splice switching activity of hemoglobin β-globin gene modulators via QSAR modeling

Saw Simeon [a], Rickard Möller [b], Daniel Almgren [b], Hao Li [a], Chuleeporn Phanus-umporn [a], Virapong Prachayasittikul [c], Leif Bülow [b], Chanin Nantasenamat [a,*]

[a] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[b] Pure and Applied Biochemistry, Chemical Center, Lund University, Lund 221 00, Sweden
[c] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

ARTICLE INFO

Article history:
Received 5 September 2015
Received in revised form 21 November 2015
Accepted 2 December 2015
Available online 18 December 2015

Keywords:
Hemoglobin
Thalassemia
Quantitative structure–activity relationship
QSAR
Data mining

ABSTRACT

β-Thalassemia is a blood disease caused by a mutation in the second intron of the β-globin gene of hemoglobin that leads to abnormal hemoglobin production. Low molecular weight compounds have been proposed to modulate defective splicing by binding unwanted splicing sites, thereby restoring correct splicing. This study investigates the origin of this splice switching activity in a set of 39 active and 61,000 inactive compounds. The *K*-means algorithm was applied to the inactive compound points with 39 clusters, in which a point from each cluster was selected to create a balanced data set of 39 active and inactive compounds. To avoid random bias, predictive models (i.e., decision tree (DT), random forest (RF), artificial neural network (ANN), partial least squares discriminant analysis (PLS-DA) and support vector machine (SVM)) were constructed 50 times. The performances of the predictive models were statistically assessed in terms of accuracy, sensitivity, specificity and Matthews correlation coefficient (MCC). RF provided an accuracy of 89.50 ± 13.45, sensitivity of 94.97 ± 13.49, specificity of 84.29 ± 22.27, and MCC of 0.80 ± 0.25 for 10-fold CV, and it provided and accuracy of 88.00 ± 8.55, sensitivity of 87.89 ± 13.93, specificity of 87.51 ± 13.75, and MCC of 0.75 ± 0.18 for external testing. Taking advantage of the built-in feature selector of RF, a thorough analysis of feature importance was conducted. Newly identified fingerprint substructures, namely, three carbon-hetero bonds (i.e., secondary amide, tertiary amide, carboxyl derivative, carboxylic acid derivative and nitrile), carbon-carbon bonds (i.e., primary carbon, secondary carbon and alkene), aromatics (hetero N nonbasic) and carbon-hetero bond (alkyl aryl ether), may provide a better understanding of the structural variations governing the splice switching activity of the hemoglobin β-globin gene.

1. Introduction

Thalassemia is the most commonly inherited anemia, and it is widely distributed in the Mediterranean, the Middle-East, South-East Asia and sub-Saharan Africa. This condition can be linked to imbalanced α- and β-globin synthesis. The pathophysiology of β-thalassemia is associated with absent or reduced β-globin production, leading to an excess of α-globin polypeptides. This triggers a range of activities, including the formation of reactive oxygen species (ROS), release of hemin and free iron [1]. The clinical management of β-thalassemia largely depends on life-long red blood cell transfusions and iron chelation. Many alternative and experimental treatments have been developed, including promotion of fetal hemoglobin production, but to date, none of these treatments have reached widespread clinical use.

More than 200 alternative genetic lesions have been identified that affect the β-globin gene, resulting in a wide range of anemic severity. However, only approximately 10 of these lesions are responsible for 90 % of the cases. In particular, abnormal transcript splicing has been associated with β-thalassemia, clearly indicating that modulation of splicing can be a valuable approach for treatment. The use of splice switching oligonucleotides (SSO) has been promising [2]. Mutations in the second intron of the β-globin gene at positions 654 and 705 are clinically important and cause mis-splicing, which results in a defective protein. SSOs in antisense form to the 654 and 705 transcripts have been demonstrated to restore proper splicing in mouse models [3]. High-throughput screening of chemical libraries has also indicated that several compounds can be used to modulate such erroneous splicing.

Quantitative structure–activity relationship (QSAR) represents an important approach for elucidating the origin of biological activity for a set of compounds of interest as a function of their molecular descriptors [4,5]. The resulting QSAR models can reveal molecular features that are essential for active compounds and that can subsequently be used as therapeutic agents. We recently applied QSAR to understand the underlying physicochemical features defining Hb allosteric effector

[*] Corresponding author.
E-mail address: chanin.nan@mahidol.ac.th (C. Nantasenamat).

http://dx.doi.org/10.1016/j.chemolab.2015.12.002
0169-7439/© 2015 Elsevier B.V. All rights reserved.

activity for a set of *myo*-inositols [6]. In this QSAR study, we examine the origin of the splice switching activity of Hb β-globin gene modulators via predictive QSAR modeling to draw conclusions regarding the most effective chemical structure that is useful in clinical settings.

Challenges in the development of QSAR/QSPR models can arise from the following: (i) selection of an appropriate subset of molecular descriptors from the many available descriptors, (ii) inability to interpret the descriptors, and (iii) the need to optimize chemical structures if three-dimensional descriptors are to be used. To address the above issues, QSAR models were developed using interpretable substructure fingerprints to quantitatively represent gene modulators. These descriptors were correlated to the splice switching activity of the hemoglobin β-globin gene using a wide array of machine learning methods, including rule-based, ensemble, non-linear classification and linear classification methods. Insights into important features governing the origin of splice switching activity as deduced from the constructed models could be used to further guide the design of novel ASOs with desired activity.

2. Materials and methods

2.1. Data set

A data set of small molecule modulators with splice switching activity against the Hb β-globin gene with a mutation in the second intron at position 654 (IVS2-654) was obtained from PubChem BioAssay (accession number AID 925) [7]. The data originated from a high-throughput screen of a chemical library of 64,405 compounds. Because 222 compounds were reported as having inconclusive activity, these compounds were therefore excluded. Compounds were treated with the QSAR curation workflow from Fourches et al. [8]. Briefly, the main steps are as follows: (i) removal of inorganics and mixtures, (ii) structural conversion and cleaning, (iii) normalization of specific chemotypes, (iv) removal of duplicates and (v) final manual checking. Thus, chemical compounds were curated using ChemAxon Standardizer with the following options: *Strip Salts*, *Aromatize*, *Clean3D*, *Tautomerize*, *Neutralize*, and *Remove explicit hydrogens* [9]. The resulting data set is composed of 39 and 60,647 active and inactive compounds. A representative subset of the chemical structures of these compounds (i.e., particularly the 39 active compounds) are shown in Fig. 1. The Open Babel software [10] was used to convert compounds from the SMILES notation to the SDF file format, which is suitable as an input for the PaDEL-Descriptor software.

2.2. Compound descriptors

Fingerprint descriptors provide descriptions of the constituting substructures inherently present in a molecule. There are essentially two versions of this descriptor, namely, the count and binary versions. Each bit in a string of fingerprint descriptors represents a distinct substructure [11]. In the count version, the numerical value, as the name implies, represents the frequency of that substructure present in a molecule, whereas in the binary version, values of 1 and 0 denote its presence and absence, respectively. These interpretable substructure fingerprints were calculated using the PaDEL-Descriptor software [12]. Thus, it can pinpoint substructures of a compound that are important for the activity of splice switching modulators [13].

2.3. Data filtering

Collinearity is a condition in which pairs of descriptors have a correlation with each other. It has a substantial negative impact on the computational analysis because correlated descriptors add more complexity to the model [14]. In addition, it also affects the interpretation of descriptors (i.e., substructure fingerprint count) because the resulting coefficient estimates (e.g., linear models) or feature usages (e.g., decision trees) are highly unstable. Moreover, the statistical assessment measures will be very sensitive to the predictive models and may inhibit the ability of prediction for new observations because correlated data create redundancy, which may overfit models. Overfitting is a condition when predictive models perfectly predict the training set. However, when new samples are introduced for prediction, the models perform ineffectively. In general, a Pearson's correlation coefficient of 0.7 is an indicator of high collinearity among predictors [15]. Thus, the *cor* function from the R package *stats* was used to calculate correlations among descriptors [16]. To obtain filtered descriptors with all pairwise correlations less than 0.7, the *findCorrelation* function from the R package *caret* with a cutoff at 70% was used [16]. The remaining descriptors that were used in the study are shown in Fig. 4.

2.4. Data pre-processing

Initially, the collected data set contained highly imbalanced data, in which 61,000 and 39 compounds were inactive and active compounds, respectively. To create a balanced data set, the undersampling approach was performed by applying *K*-means clustering on the inactive group of compounds. Prior to performing *K*-means clustering, PCA was utilized to reduced the dimension of the data set to obtained non-correlated variables, also known as principal component (PC) coefficients. These PC coefficients were used as inputs for performing *K*-means clustering to derive 39 clusters for a pool of inactive compounds [17]. A random point from each cluster was selected to represent a set of 39 inactive compounds. The resulting pre-processed data set is provided as supplementary data on figshare that is available at http://dx.doi.org/10.6084/m9.figshare.1609584.

2.5. Univariate analysis

Univariate analysis was conducted to investigate patterns, features and trends that were present in the substructure descriptors. It was performed by creating histogram plots using the R package *ggplot2*. The normality of each substructure's fingerprint for active and inactive compounds was assessed using the Shapiro-Wilk test using the *shapiro.test* function from the R package *stats*. The function *pairwise.wilcox.test* from the R package *stats* was used to perform a Mann–Whitney U test to measure the statistical significance of the investigated pairs (i.e., active and inactive compounds).

2.6. QSAR modeling

The decision tree algorithm J48 is Weka's implementation of the C4.5 algorithm that automatically generates classification rules in the form of a rule-based branching tree using the divide-and-conquer algorithm. It is considered to be one of the most transparent learning algorithms in which a series of readily understandable if-then rules are formed. The construction involves two steps: growing and pruning. Growing starts from the root node, which branches out to form internal nodes that subsequently end up as leaf nodes. Internal nodes represent descriptors, branches describe descriptor values, and leaf nodes represent Y categorical classes (i.e., active and inactive). Once the trees are fully grown, the grown tree is pruned as a function of the predictive performance. The advantage of pruning is that it reduces the complexity of the formed tree and reduces the chance of over fitting. The *J48* function from the *Rweka* package was used to construct the QSAR models [18].

Random forest (RF) is an ensemble classifier that is composed of multiple decision tress. Similar to J48, classification starts at the root node, in which the data set at the node is split according to the value of descriptors that are selected such that the descriptors of different activities are predominantly moved to different branches. The classification is obtained by averaging the results of all trees by a majority vote from each tree. The RF classifier was generated using the R package

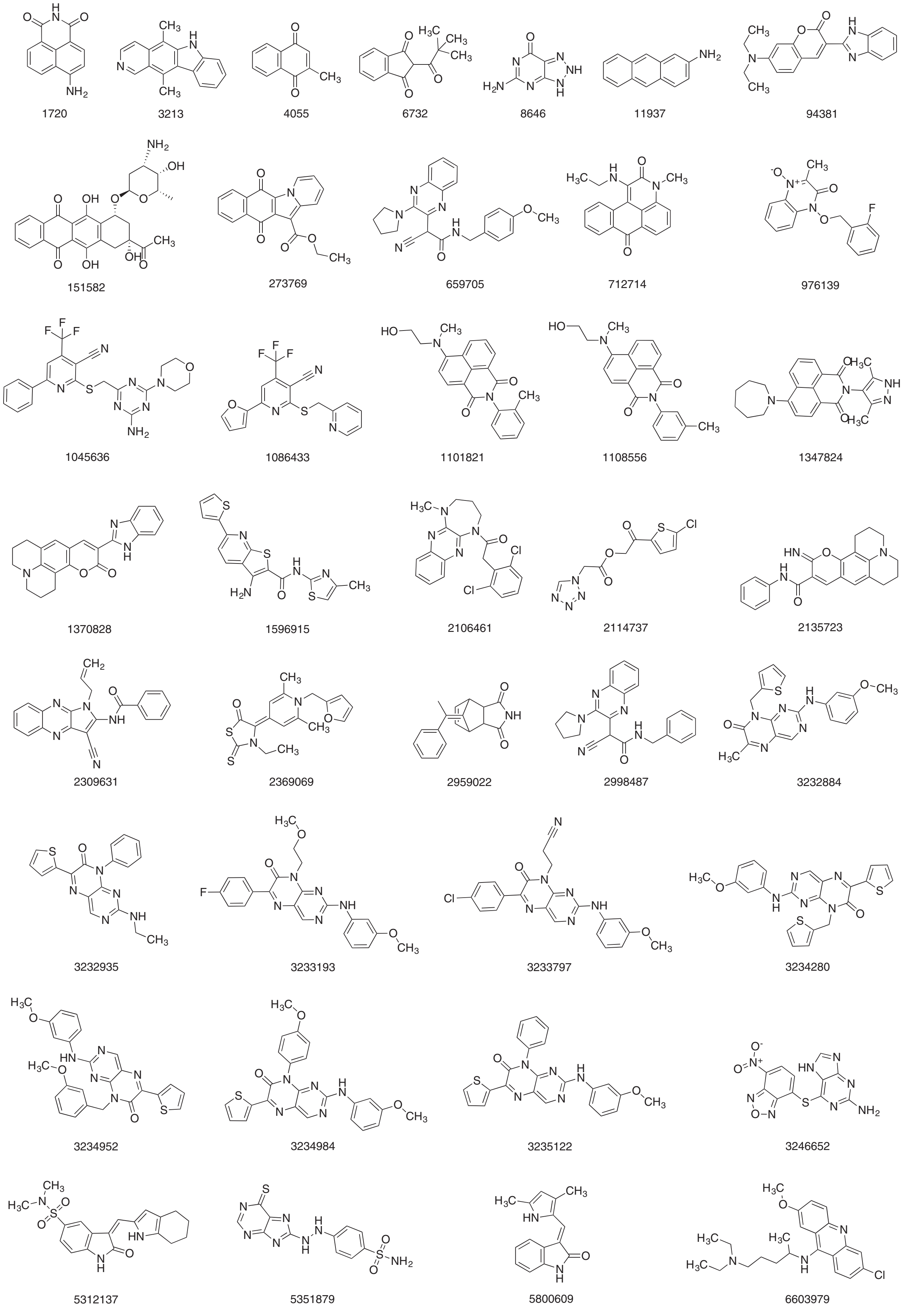


Fig. 1. Chemical structures of the set of 39 active compounds with splice switching activity. PubChem CID numbers are shown beneath the chemical structures.

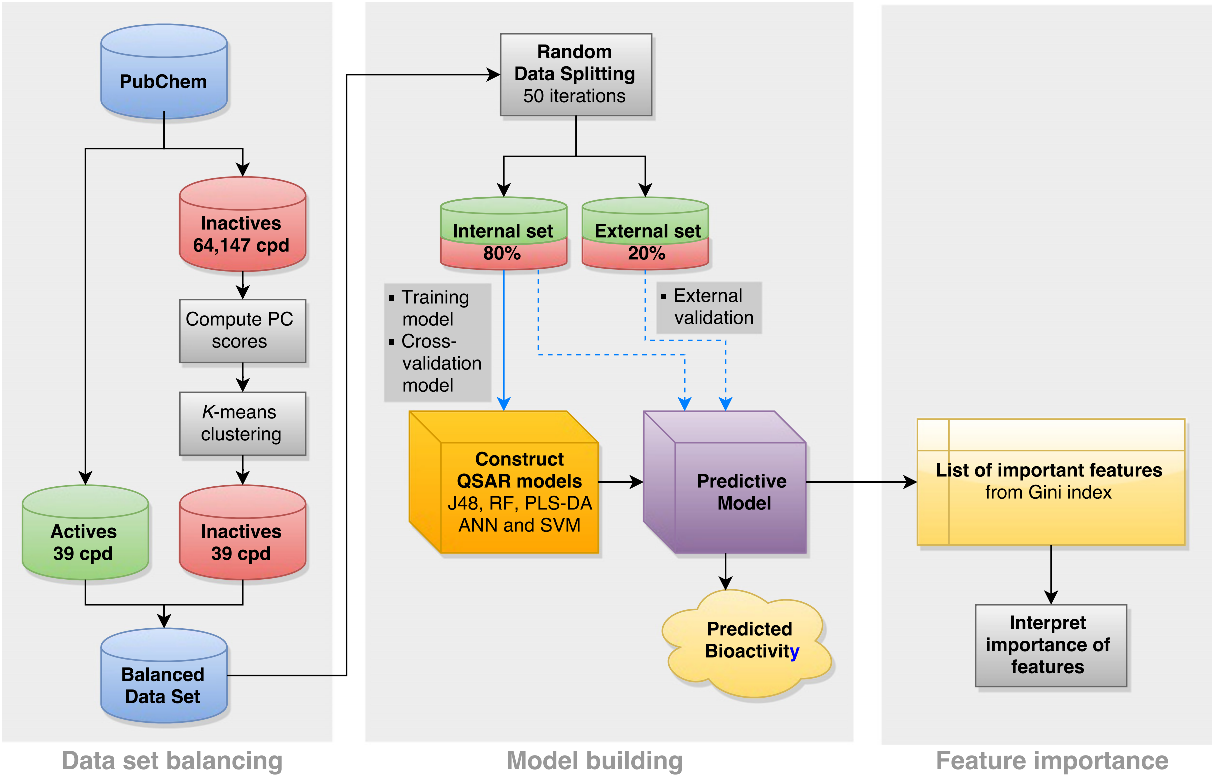


Fig. 2. Workflow for the QSAR modeling of the hemoglobin β-globin gene modulators.

randomForest with a total of 500 trees [19]. Two RF parameters were subjected to optimization including the ntree (i.e. the number of trees used for building the ensemble RF model) and mtry (i.e. number of descriptors to be sampled randomly as candidate features).

Support vector machine is a machine learning approach that can be used to perform both classification and regression, in which the kernel function is used to map the data into a high-dimensional feature space. The commonly used radial basis function kennel was used to construct a predictive model. The support vectors were fine-tuned with several parameters to obtain the optimal parameters, which are the width of the kernel function gamma and the error penalty parameters cost. The commonly used radial basis Gaussian kernel was selected along with tuned parameters ($C = 2$, $\gamma = 0.29$) to construct SVM machine learners. The *train* R package *caret* was used to fine-tune the model, and the support vector machine was constructed using the R package *e1071* [20].

Artificial neural network (ANN) is a machine learning method that mimics the human brain, which is composed of networks of interconnected neurons that function in relaying messages in the form of electrochemical signals. Brain cells are composed of dendrites, cell bodies and axons. A synapse is the connection between the axons of the nerve cell with the dendrites of an adjacent nerve cell. In a synapse, signals are transmitted from one cell to the next as neurotransmitters. Similar to the function of the human brain, the artificial neuron is interconnected in a feed-forward manner from the first through the last nodes, in which the connection between nodes of different layers is assigned as a weight, which can be expressed as a strength of the input data. The train function from the R package *caret* was used to construct the artificial neural network model while fine-tuning the parameters. The *train* function from the R package *caret* was also used to obtain optimal parameters (i.e., hidden layers = 7 and decay weight = 0.1), and the *nnet* function from the R package *nnet* was used to train the models [21].

Partial least squares discriminatory analysis (PLS-DA) is a linear classification method that seeks to categorize samples into groups (i.e., actives and inactives) based on the predictor characteristics. It is a robust (the parameters of the model do not change when samples are taken out) multivariate analysis that involves **X** and **Y** variables. It

Table 1
Summary of the mean and standard deviations of substructure fingerprints along with their *P* values derived from the statistical difference test using the Mann–Whitney *U* test.

	Active	Inactive	P value
SubFPC1	0.667 ± 0.929	1.533 ± 2.351	0.081
SubFPC2	0.756 ± 1.384	1.578 ± 3.258	0.223
SubFPC3	0.156 ± 0.638	0.267 ± 0.780	0.341
SubFPC5	0.267 ± 0.495	0.067 ± 0.330	0.009
SubFPC16	0.044 ± 0.208	0.044 ± 0.208	1.000
SubFPC18	0.289 ± 0.589	0.511 ± 0.944	0.378
SubFPC20	0.067 ± 0.252	0.156 ± 0.520	0.452
SubFPC26	0.044 ± 0.208	0.178 ± 0.387	0.046
SubFPC28	0.133 ± 0.344	0.022 ± 0.149	0.051
SubFPC33	0.178 ± 0.387	0.089 ± 0.288	0.220
SubFPC40	0.089 ± 0.288	0.022 ± 0.149	0.173
SubFPC49	0.267 ± 0.751	0.244 ± 0.529	0.509
SubFPC52	0.200 ± 0.405	0.022 ± 0.149	0.008
SubFPC88	1.067 ± 0.780	1.244 ± 1.836	0.614
SubFPC100	0.111 ± 0.318	0.533 ± 0.548	0.000
SubFPC101	0.044 ± 0.208	0.244 ± 0.484	0.013
SubFPC133	0.156 ± 0.367	0.044 ± 0.208	0.082
SubFPC135	0.111 ± 0.438	0.644 ± 1.151	0.003
SubFPC153	0.000 ± 0.000	0.111 ± 0.318	0.023
SubFPC171	0.156 ± 0.475	0.200 ± 0.457	0.416
SubFPC179	0.178 ± 0.387	0.000 ± 0.000	0.003
SubFPC180	0.067 ± 0.252	0.267 ± 0.539	0.034
SubFPC181	1.133 ± 1.217	0.844 ± 1.127	0.239
SubFPC182	0.089 ± 0.288	0.178 ± 0.442	0.327
SubFPC183	0.267 ± 0.618	0.067 ± 0.252	0.060
SubFPC214	0.067 ± 0.252	0.200 ± 0.405	0.065

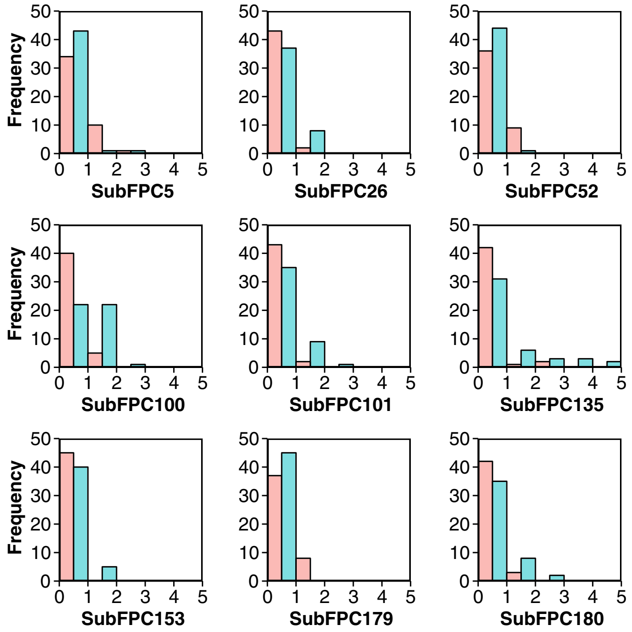


Fig. 3. Histogram of substructure fingerprint counts of active (colored in pink) and inactive (colored in cyan) splice switching modulators.

simultaneously projects the **X** into latent variables to correlate **X** predictors with **Y** responses. The extent of the influence that **X** has on the **Y** variable is revealed by the regression coefficient when PLS models are constructed for each responsive variable. The *plsda* function from the R package *caret* was used to construct predictive models [16].

2.7. Data splitting

The data set was randomly partitioned into two subsets (i.e., 80% as the internal set and 20% as the external set) using the *sample_n* function from the R package *dplyr* [22]. The internal set was used to train the model, while the external set was used to externally validate the performance of the predictive model. To avoid the bias that may arise from a single data split when training the model, predictive models were constructed from each of the 50 independent data splittings, and the mean and standard deviation values of statistical parameters were reported.

Ten-fold cross-validation (10-fold CV) and external testing were used to validate the robustness and reliability of the predictive models. Ten-fold CV was performed on the internal set, where the data set is separated into ten folds. Practically, one fold from the total of ten folds is left out as the testing set, while the remaining are used for training the models. This process is repeated iteratively until all the data samples had a chance to be left out as a testing test. External validation evaluates the stability of the predictive models by serving as unknown data previously not seen by the training model.

2.8. Validation of QSAR models

Model validation is essential in evaluating the results of empirical modeling. Several statistical parameters are used to assess the effectiveness and efficiency of constructed predictive models, including accuracy, sensitivity, specificity and Matthews correlation coefficient (MCC) on both dependent and independent test sets. These statistical parameters are commonly used in QSAR modelings [23]. Accuracy is the

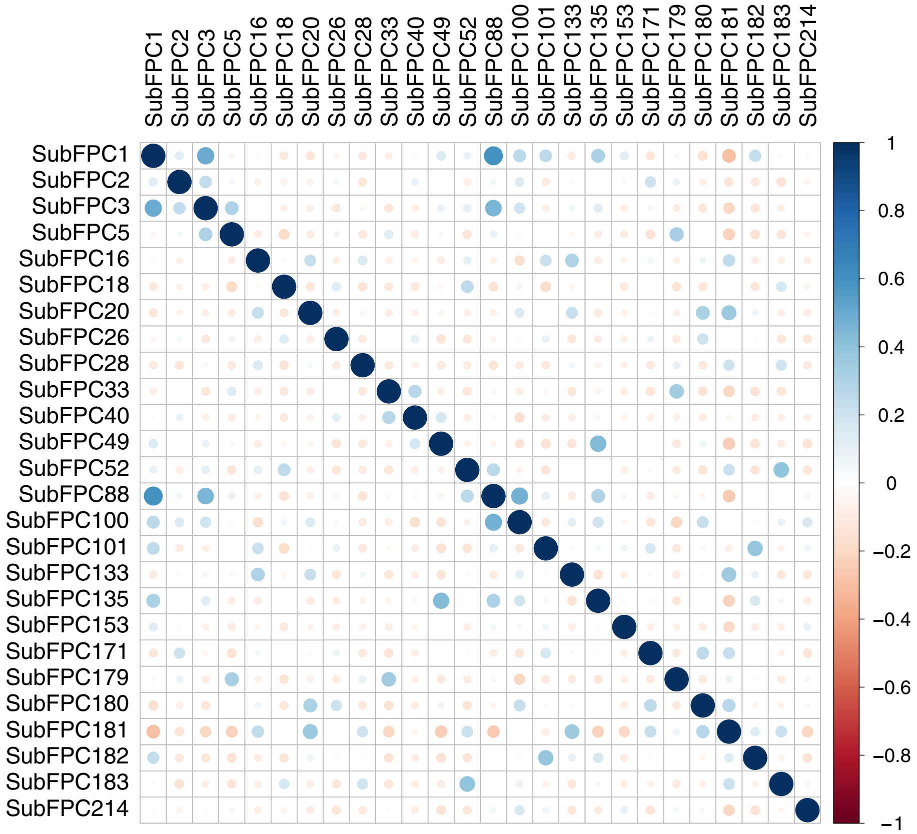


Fig. 4. Intercorrelation matrix of the descriptors utilized for constructing the predictive models.

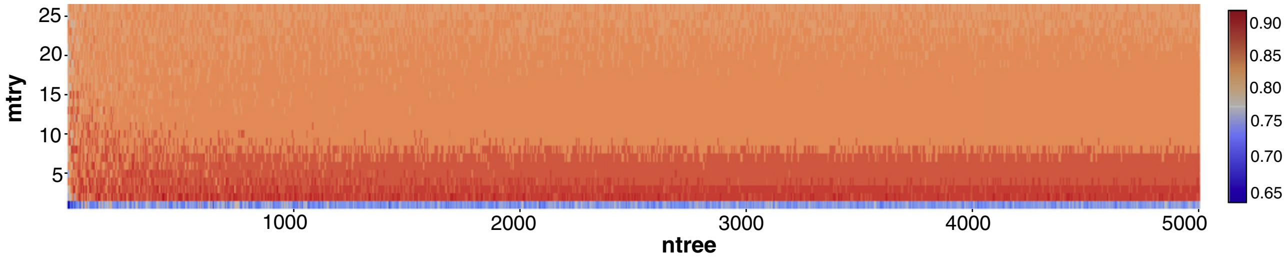


Fig. 5. Heat map of fine tuning the RF parameters ntree and mtry as shown in the X and Y axes, respectively. Accuracy obtained from 10-fold CV is shown in the plot and color-coded according to their performance ranging from low (blue) to high (red) accuracy.

percentage of correctly classified instances relative to the total number of instances. Although it is commonly used to assess the predictability of predictive models, accuracy is not an optimal measure of model performance if the data are unbalanced. In contrast to accuracy, MCC is a measure of assessment that is insensitive to unequal sizes of the classes and the costs of making certain errors. The Sen, sometimes considered as the true positive (TP) rate, is the proportion of true positives among all positively classified instances, whereas specificity is the proportion of true negatives (TN) among all negatively classified instances, thus defined as the false-positive rate.

These can be calculated using the equations described below:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100 \tag{1}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100 \tag{2}$$

$$Specificity = \frac{TN}{(TN + FP)} \times 100 \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

where TP is the instances of true positives, TN is the instances of true negatives, FP is the instances of false positives, and FN is the instance of false negatives. The range of MCC is from −1 to 1, in which a value of MCC = 1 indicates the best possible prediction, while MCC = −1 indicates the worst possible prediction. On the other hand, MCC = 0 suggests a random prediction scheme.

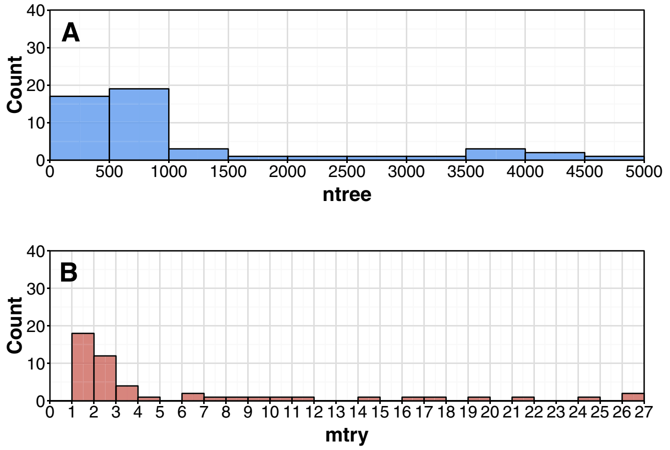


Fig. 6. Histogram showing the binned distribution of optimal ntree (A) and mtry (B) values as obtained from model building from 50 independent data splits.

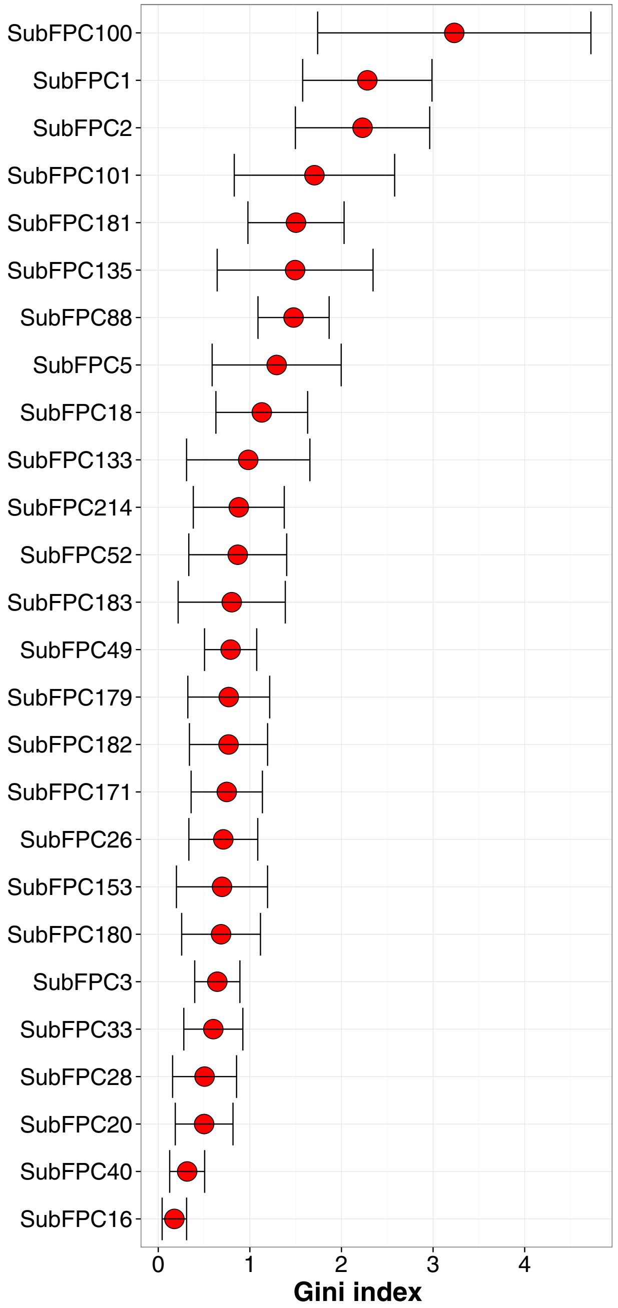


Fig. 7. Importance of substructure fingerprints as a function of the Gini index. Features with the largest Gini index are deemed the most important.

3. Results and discussion

The notion that the activity of compounds is governed by their physicochemical properties is a paradigm of QSAR. To gain understanding of the activity of splice switching modulators of the β-globin gene of hemoglobin, a set of easy-to-interpret substructure fingerprints was used to encode the compounds. The intercorrelation between predictors was removed using a Pearson's correlation coefficient threshold of 0.7 as an indicator of high collinearity [15]. To assess the robustness and reliability of the predictive model, the data set was split into two subsets: (i) internal set (i.e., used to train and fine tune the model) and an (ii) external set (i.e., used to validate the performance of the predictive models in the real-world setting). Feature importance selection was then performed using the built-in RF feature selector. To avoid random seeding when constructing the predictive models, 50 independent data splittings were performed, where each split was used to construct models. This was followed by computing the mean and standard deviation values for each statistical performance metric (i.e., accuracy, sensitivity, specificity and MCC) that was used in evaluating the predictive performance of the model. The general framework of the QSAR modelings is shown in Fig. 2.

3.1. Univariate analysis of Hb β-globin gene modulators

The normality of the data set (i.e., substructure fingerprint) was assessed using the Shapiro-Wilk test [24]. The results indicated that all descriptors afforded *p* values less than 0.05, suggesting that all descriptors exhibited a non-normal distribution. Thus, the Mann–Whitney *U* test was utilized to test whether two populations of activity (i.e., active or inactive) are equal or different from one another. As shown in Table 1, substructures SubFPC5, SubFPC26, SubFPC52, SubFPC100, SubFPC101, SubFPC135, SubFPC153, SubFPC179 and SubFPC180 displayed statistically significant differences with corresponding *p* values of 0.009, 0.046, 0.008, 0.000, 0.013, 0.003, 0.023, 0.003 and 0.034, respectively. The Lipinski's rule-of-5 indicated that nearly all compounds from the data set were drug-like as they had values in accordance with the following molecular properties: (i) molecular weight < 500 Da, (ii) LogP < 5, (iii) number of hydrogen bond donors < 5 and (iv) number of hydrogen bond acceptors < 10. Simple histograms of significant descriptors are shown in Fig. 3 to provide an overview of the relative distribution of the data values.

SubFPC5 is a count of alkene, which is a hydrocarbon that contains a carbon-carbon double bond, in the investigated compounds. The histogram plot of SubFPC5 shows that most of the inactive compounds have a count of 0, as shown in Fig. 3. In addition, it can be observed that SubFPC5 having a count of at least 1 is a general common feature for the active group compared to the inactive group, which have values of 0.267 ± 0.495 (active) and 0.067 ± 0.330, respectively ($p < 0.05$). However, it can be observed that active group had the same count with the inactive group with ranges of 2 (maximum) and 0 (minimum). Moreover, more than half of the active and the inactive groups have values of 0 counts.

SubFPC26 is a count of tertiary aliphatic amine, which is three organic substituents connected to a nitrogen atom. As shown in Fig. 1, the distribution of SubFPC26 shows a significant difference between the active and inactive groups, as was the case for SubFPC5, with values of 0.044 ± 0.208 and 0.178 ± 0.387, respectively ($p < 0.05$). Both of the ranges of active and inactive were not strikingly different ranges with the minimum count (0) to maximum count (1). Nevertheless, the Mann–Whitney U test revealed that the active and inactive groups are significantly different with a *p* value of 0.046. When comparing the active and inactive groups according to substructure count, the majority of the active group has a count of 0, whereas the majority of the inactive group has a count of 1, as shown in Fig. 3.

Table 2
Summary of the predictive performance as a function of different descriptor types and assessed by 10-fold CV and external validation. Predictive performance comparison of J48, RF, ANN, SVM and PLS-DA.

Predictive Models	Training Set			10-fold CV				External Validation			
	Accuracy	Sensitivity	Specificity	MCC	Accuracy	Sensitivity	Specificity	MCC	Accuracy	Sensitivity	Specificity	MCC
J48	94.38 ± 2.01	94.41 ± 3.36	94.61 ± 3.46	0.89 ± 0.04	80.33 ± 5.85	81.64 ± 6.83	79.05 ± 7.35	0.60 ± 0.12	78.13 ± 12.20	80.25 ± 18.50	76.49 ± 18.45	0.58 ± 0.23
RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	89.50 ± 13.45	94.97 ± 13.49	84.29 ± 22.27	0.80 ± 0.25	88.00 ± 8.55	87.89 ± 13.93	87.51 ± 13.75	0.75 ± 0.18
ANN	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	85.48 ± 15.96	84.15 ± 21.90	86.16 ± 24.93	0.74 ± 0.29	86.63 ± 10.34	83.40 ± 17.31	89.90 ± 12.91	0.74 ± 0.20
SVM	96.47 ± 3.70	94.98 ± 6.02	97.84 ± 3.62	0.93 ± 0.07	80.19 ± 15.62	72.17 ± 29.14	86.09 ± 26.85	0.66 ± 0.25	82.88 ± 10.24	85.23 ± 19.45	80.45 ± 18.94	0.67 ± 0.20
PLS-DA	93.47 ± 3.82	93.47 ± 5.93	93.08 ± 4.91	0.86 ± 0.08	81.08 ± 14.97	77.79 ± 27.74	79.33 ± 29.61	0.65 ± 0.25	84.25 ± 11.38	83.53 ± 15.87	85.74 ± 15.05	0.69 ± 0.22

SubFPC52 is a count of imine, a nitrogen double bonded to carbon where a substituent is bonded to the nitrogen atom and two substituents are bonded to the carbon atom. The mean and standard deviation of SubFPC52 were 0.200 ± 0.405 and 0.020 ± 0.149 for active and inactive, respectively. The distribution range of SubFPC52 was similar to that of SubFPC5, where the majority of the inactive were higher than that of active and lower than that of active for substructure counts of 0 and 1, respectively. The Mann–Whitney test presented a p value of 0.008, indicating that the two groups were significantly different. The histogram plot of SubFPC52 in Fig. 3 shows that the frequency of inactive has more SubFPC52 than the active group and that active has more than inactive for counts of 0 and 1, respectively.

SubFPC100 is a count of secondary amide. The corresponding values of SubFPC100 are 0.111 ± 0.318 and 0.533 ± 548 for active and inactive, respectively. A simple statistical analysis revealed that the p value for the Mann-Whitney u test was (0.000). Note that the resulting p value is the smallest of all the significant descriptors. The majority of the active has a SubFPC100 count of 0, whereas the inactive groups were equally distributed for both count of 0 and count of 1, as shown in Fig. 3.

SubFPC101 is a count of tertiary amide in the investigated substructure. On average, inactive compounds had higher counts for SubFC101 when compared to active compounds with values of 0.244 ± 0.484 and 0.044 ± 0.208 for inactive and active, respectively (p value < 0.05). The histogram clearly showed that SubFPC101 was negatively skewed to the left as the majority of the active and inactive compounds have a count of 0, as shown in Fig. 3. In general, the inactive compounds had a greater count number than the active compounds with maximum count ranges of 2 and 1, respectively.

SubFPC135 is a count of carboxyl derivatives. The histogram clearly showed that both types of compounds had a SubFPC135 close to zero. Notably, the inactive compounds had dynamic ranges of counts from 5 (maximum) to 0 (minimum), whereas the active compounds had a narrow range of counts from 2 (maximum) to 0 (minimum). It was found that the counts of SubFPC135 were higher in inactive compounds compared to the active compounds with values of 0.644 ± 1.151 and 0.111 ± 0.438, respectively ($p < 0.05$). The distribution of SubFPC135 for the inactive compounds was observed to be jagged relative to that of non-steroids, suggesting that inactive compounds tend to have SubFPC135 distributed in dynamic ranges.

SubFPC153 is a count of urethane, also known as ethyl carbamate. The histogram plot shows that all active compounds and a majority of inactive compounds lack SubFPC153 as a substructure, whereas a minority of inactive compounds contain 1 count. A simple statistical analysis revealed that average values were 0.000 ± 0.000 and 0.111 ± 0.318 for active and inactive compounds, respectively ($p < 0.05$). The histogram plot clearly showed that all active compounds have a count of 0 for SubFPC153, whereas the inactive compounds have a narrow range of counts of 1 at maximum and 0 at minimum, indicating that the inactive compounds may have at most 1 substructure of this type.

SubFPC179 is a count of hetero nitrogen basic hydrogen. The corresponding values for the SubFPC179 were 0.178 ± 0.387 and 0.000 ± 0.000 for active and inactive, respectively. The histogram plot clearly shows that inactives do not posses any of this substructure, whereas active had at least one SubFPC197, as shown in Fig. 3.

SubFPC180 is a count of hetero nitrogen basic no hydrogen. It was observed that the SubFPC180 were statistically different for active and inactive, with values of 0.067 ± 0.252 and 0.267 ± 0.539, respectively. It was found that the values of SubFPC180 are lower in active compounds compared to their inactive counterparts. As shown in Fig. 3, the distribution of the inactive was broader than that of active with ranges from 3 (maximum) to 0 (minimum). However, the majority of the active compounds do not possess SubFPC180 as a substructure. Nevertheless, the Mann–Whitney test revealed that the active and inactive groups are significantly different with a p value of 0.034.

Table 3
List of top substructure fingerprints and their descriptions.

Fingerprints	Description
SubFPC1	Primary carbon
SubFPC2	Secondary carbon
SubFPC3	Tertiary carbon
SubFPC5	Alkene
SubFPC16	Dialkylether
SubFPC18	Alkylarylether
SubFPC20	Alkylarylthioether
SubFPC26	Tertiary aliph amine
SubFPC28	Primary arom amine
SubFPC33	Tertiary mixed amine
SubFPC40	1,2-Aminoalcohol
SubFPC49	Ketone
SubFPC52	Imine
SubFPC88	Carboxylic acid derivative
SubFPC100	Secondary amide
SubFPC101	Tertiary amide
SubFPC133	Nitrile
SubFPC135	Carboxyl derivative
SubFPC153	Urethan
SubFPC171	Arylchloride
SubFPC179	Hetero N basic H
SubFPC180	Hetero N basic no H
SubFPC181	Hetero N nonbasic
SubFPC182	Hetero O
SubFPC183	Hetero S
SubFPC214	Sulfonic derivative

3.2. QSAR modeling

Because compounds were characterized by substructure fingerprint descriptors, this allowed us to pinpoint the substructures that are important for modulating the splice switching activity of the β-globin gene chain. To avoid the inherent redundancy among the substructures, descriptors were filtered using a cutoff value of 0.70. The undersampling approach was applied to solve the class imbalance problem where the number of active and inactive compounds are significantly out of proportion. This approach had successfully been shown to be effective for handling the inherently imbalanced data derived from the PubChem database. [25].

As previously mentioned, the initial data set was split into an internal validation set and an external testing set, in which the former constituted 80% of the data set while the latter constituted the remaining 20% of the data set. QSAR models were developed with various machine learning methods, including rule-based models (e.g., J48), ensemble models (e.g., RF), non-linear classification models (e.g., ANN and SVM) and linear classification models (e.g., PLS-DA). To avoid the bias of a single data split, data splitting was performed for 50 iterations in which each split was used to construct a predictive model. The mean and standard deviation of the resulting predictive performance (e.g., accuracy, specificity, sensitivity and MCC) were computed as assessed by 10-fold CV and external sets.

Firstly, the rule-based J48 algorithm (i.e., Weka's implementation of the C4.5 algorithm) was used for identifying rules governing the relationship of independent variables (i.e., substructure fingerprint) with that of the dependent variable (i.e., activity). As shown in Table 2, the predictive performance of the training set and 10-fold CV set provided accuracies of 94.38 ± 2.01 and 80.33 ± 5.85, respectively. A closer examination of the predictive model revealed that the training set provided a better overall predictive performance than the 10-fold cross validation. Moreover, the external testing set afforded a moderate MCC value of 0.58 ± 0.23. Secondly, RF is a popular ensemble technique in machine learning in which multiple decision trees are bagged to generate prediction and the predictions are averaged to provide the bagged model's prediction. The bagging of trees improves the predictive performance over a single tree by reducing the variance of the prediction.

As also shown in Table 2, the training set for RF was as high as 100.00 ± 0.00% for accuracy, sensitivity and specificity and 1.00 ± 0.00 for MCC. On the other hand, the 10-fold CV set was slightly lower with 89.50 ± 13.45, 94.97 ± 13.49, 84.29 ± 22.27 and 0.80 ± 0.25 for accuracy, sensitivity, specificity and MCC, respectively. Nevertheless, it can be observed that the predictive performance of the RF was considerably better compared with J48 In addition, it can be seen the highest performance for the external set was modeled by RF. Third, ANN is a powerful non-linear classification technique that works similar to a human brain. The outcome is modeled by the intermediates of predictor variables created through non-linear functions. The predictive performance of the ANN is high with 100.00 ± 0.00 for accuracy, sensitivity and specificity and 1.00 ± 0.00 for MCC. The 10-fold CV of the ANN is comparable to the RF with 85.48 ± 15.96, 84.15 ± 21.90, 86.16 ± 24.93 and 0.74 ± 0.29 for accuracy, sensitivity, specificity and MCC, respectively. As can be seen in Table 2, the MCC value for ANN was 0.74 ± 0.20 which is slightly lower than of RF (0.75 ± 0.18) but superior to J48 (0.58 ± 0.23), SVM (0.67 ± 0.20) and PLS-DA (0.69 ± 0.22). Fourth, SVM is another non-linear modeling technique that is considered to be a powerful and highly flexible machine learner. It was originally developed for classification, and the predictive performance of the models is comparable to other machine learners (i.e., rule-based, ensemble and linear). As shown in Table 2, the assessment parameters for the SVM are relatively high, with accuracies of 96.47 ± 3.70 and 80.19 ± 15.62 for training and 10-fold CV, respectively. Finally, PLS-DA is a linear classification method in which predictors undergo dimension reduction with respect to the response. The predictive performance of PLS-DA is comparable to that of other machine learners, where the accuracy, sensitivity, specificity and MCC were 93.47 ± 3.82, 93.47 ± 5.93, 93.08 ± 4.91 and 0.86 ± 0.08, respectively, for the training set and 81.08 ± 14.97, 77.79 ± 27.74, 79.33 ± 29.61 and 0.65 ± 0.25, respectively, for the 10-fold CV. The predictive power of PLS-DA to accurately predict the activity of unknown compounds as assessed by the external set was good with MCC of 0.69 ± 0.22. The overall predictive performances of the machine learning methods (i.e., J48, RF, ANN, SVM and PLS-DA) are highly comparable. However, based on the predictive power of both 10-fold CV and external sets, RF outperformed the other learning methods. This result indicates that the RF model was able to correlate unknown chemical structures with splice switching activity of the hemoglobin β-globin gene. Consequently, the RF model was chosen to represent the best QSAR model and further used in interpreting the feature importance governing the splice switching activity of β hemoglobin gene modulators.

3.3. Optimization of RF parameters

RF is increasingly used in QSAR modeling owing to its robust performance and built-in measure of feature importance. There are two RF parameters that can be optimized, which is comprised of ntree and mtry as previously mentioned in the Materials and Methods. Generally, according to the documentation of the *randomForest* R package, the default values of ntree and mtry are set at 500 and $\sqrt{N}$ (i.e. where N denote the total number of descriptors), respectively. Parameter optimization may lead to improvement in the resulting predictive performance as different data have their own unique characteristics (i.e. size, type and structure). As optimization of RF parameters is a highly time-consuming process owing to the relatively large size of ntree and mtry values coupled to the rather lengthy calculation of the *k*-fold CV scheme, therefore good default values could save considerable amount of computational cost. Thus, these type of benchmarking study had previously been employed for identifying optimal parameters of SVM as a function of signature fingerprints. [26].

Fig. 5 showed that mtry in the range of 1 and 10 turned out to be the best while mtry greater than 15 resulted in low accuracy. To get a better understanding on the best combination of mtry and ntree, the optimization step were repeated 50 times with independent training set. As shown in Fig. 6, the histogram suggests that the optimal range of mtry is no more than 5 while the ntree is no more than 1000. On the basis of the parameter optimization, the optimal range of mtry is between 1 to 5 while for ntree the optimal range is between 100 to 1000, which are the suggested good staring point in terms of trade-off between speed and performance. Nevertheless, it is worthy to note that using higher number of trees (i.e. greater than 1000) provided no benefit as it only adds to the complexity of the model, which possibly may give rise to overfitting while also prolonging the computational cost. In summary, the default values for the ntree and mtry parameters of 500 and $\sqrt{N}$ (i.e. where N denote the total number of descriptors), respectively, are within the recommended range of optimal values and are thus used for further analysis of the feature importance.

3.4. Importance of substructure fingerprints

The analysis of feature importance for each type of substructure fingerprint can provide a better understanding of β-globin modulators. Table 3 presents a list of substructure fingerprints and their descriptions that were utilized in the study. The efficient and effective built-in feature importance estimators of the RF method are utilized to identify informative features. Two measures, namely, mean decrease of Gini index (MDGI) and mean decrease of prediction accuracy, are generally available for ranking feature importance. Because the results of the MDGI measure are highly stable compared with the mean decrease of accuracy [27], the MDGI is adopted to rank features. To avoid the bias of random seed in evaluating feature importance, the average and standard deviation values of the MDGI on 50 runs of feature importance evaluations are used in the analysis.

The top 10 descriptors are SubFPC100, SubFPC1, SubFPC2, SubFPC101, SubFPC181, SubFPC135, SubFPC88, SubFPC5, SubFPC18 and SubFPC133, which can be defined as secondary amide, primary carbon, secondary carbon, tertiary amide, carboxyl derivative, carboxylic acid derivative, non-basic hetero nitrogen, alkyl aryl ether, alkene and nitrile, respectively.

Features with the top measure of Gini index are considered to be the most important. As shown in Fig. 7, the secondary amide ranked as the top feature. Note that compounds containing the amides hydroxycarbamide, the compound that works on reducing the imbalance between α- and β-globin gene [28], and isobutyramide, the compound that works on promoting the erythroid survival [29], are marketed as drugs for treating thalassemia and used as secondary and tertiary target levels, respectively. The analysis suggested that the amide functional group in the compounds is highly important in determining the activity of splice switching of the hemoglobin β-globin gene because it has the largest Gini index obtained from the built-in feature selector of RF. The second and third important features are primary carbon and secondary carbon, respectively. Note that carbon atoms, which are constituents of the basic forms of life, are important features. Organic compounds from both natural and synthetic sources have been major therapeutic agents for the treatment of various diseases. Our results suggested that the position of the carbon compound may play a role in determining the splice switching activity. The fourth most important feature is the tertiary amide. Note that amides are highly important in the determining the activity. This may be because the chemical properties of amides can change the conformation of the native DNA molecule [30], which may subsequently affect the transcription of DNA and translation of mRNA to protein. The fifth and sixth most important features are the carboxyl derivatives and carboxylic acid derivatives, respectively. The two substructures have a carbon atom attached to an oxygen atom through a double bond and an alcohol group. It has been shown that carboxylic acids can bind to DNA strands, as they have successfully been utilized as linkers to immobilize DNA with oligonucleotides [31]. This may be because helical DNA strands have a large number of hydrogen bond interactions, which may cause a hydrogen of DNA to interact with the oxygen of carboxylic acid derivatives, which has a high level of electronegativity. The seventh most important feature is the non-basic hetero nitrogen. Notably, the nitrogen atom and the nitrogen-containing substructure (i.e., amide and tertiary amide) are important in determining the activity of splice switching activity. In summary, it can be observed that the most important features come from the amide functional groups and the carboxylic acid groups. In summary, compounds containing amide functional groups and the carboxylic acid functional groups are important in determining the activity of the splice switching, which can be used as a guide in the development of the novel splice switching compounds with high potency and selectivity.

4. Conclusion

Modulators of the Hb β-globin gene are important therapeutic agents for the treatment of thalassemia and other hemoglobinopathies. QSAR modeling was performed using the substructure fingerprint descriptors as an input to determine the substructure importance on the activity of modulators using the RF classifier, which provided excellent predictive ability with an accuracy of 89.50 ± 13.45, sensitivity of 94.97 ± 13.49, specificity of 84.29 ± 22.27 and MCC of 0.80 ± 0.25 for the internal validation set and an accuracy of 88.00 ± 8.55, sensitivity of 87.89 ± 13.93, specificity of 87.51 ± 13.75 and MCC of 0.75 ± 0.18 for the external testing set. By utilizing the excellent built-in feature importance analysis parameters, three carbon-hetero bonds (carboxyl and derivatives) are shown to have significant weight in determining splice switching activity. Such insights can provide a better understanding of the origin of splice switching activity of the Hb β-globin gene and may be used as a general guideline for designing novel modulators.

Acknowledgments

This research is supported by a Research Career Development Grant (No. RSA5780031) to CN from the Thailand Research Fund. SS is a graduate student grateful for support from the Thailand Research Fund and Mahidol University. Partial support from the Office of Higher Education Commission and Mahidol University under the National Research Universities Initiative is also gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] M.G. Olsson, M. Allhorn, L. Bülow, S.R. Hansson, D. Ley, M.L. Olsson, A. Schmidtchen, B. Åkerström, Pathological conditions involving extracellular hemoglobin: molecular mechanisms, clinical significance, and novel therapeutic opportunities for α1-microglobulin, Antioxid. Redox Signal. 17 (2012) 813–846.
[2] R. Kole, T. Williams, L. Cohen, RNA modulation, repair and remodeling by splice switching oligonucleotides, Acta Biochim. Pol. 51 (2004) 373–378.
[3] H. Sierakowska, M.J. Sambade, S. Agrawal, R. Kole, Repair of thalassemic human beta-globin mRNA in mammalian cells by antisense oligonucleotides, Proc. Natl. Acad. Sci. USA 93 (1996) 12840–12844.
[4] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, A practical overview of quantitative structure–activity relationship, EXCLI J. 8 (2009) 74–88.
[5] C. Nantasenamat, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Advances in computational methods to predict the biological activity of compounds, Expert Opin. Drug Discovery 5 (2010) 633–654.
[6] P. Mandi, W. Shoombuatong, C. Phanus-umporn, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, C. Nantasenamat, Exploring the origins of structure–oxygen affinity relationship of human haemoglobin allosteric effector, Mol. Simul. 41 (2015) 1283–1291.
[7] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, S.H. Bryant, PubChem: a public information system for analyzing bioactivities of small molecules, Nucleic Acids Res. 37 (2009) W623–W633.
[8] D. Fourches, E. Muratov, A. Tropsha, Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research, J. Chem. Inf. Model. 50 (2010) 1189–1204.
[9] C. Standardizer, Version 5.4. 4.1, ChemAxon, Budapest, Hungary, 2010.
[10] N.M. OLBoyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminf. 3 (2011) 33.
[11] J. Wikberg, M. Eklund, E. Willighagen, O. Spjuth, M. Lapins, O. Engkvist, J. Alvarsson, Introduction to pharmaceutical bioinformatics, Oakleaf Academic, 2010.
[12] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474.
[13] J. Klekota, F.P. Roth, Chemical substructures that enrich for biological activity, Bioinformatics 24 (2008) 2518–2525.
[14] M.T. Cronin, T.W. Schultz, Pitfalls in QSAR, J. Mol. Struct. 622 (2003) 39–51.
[15] G.D. Booth, M.J. Niccolucci, E.G. Schuster, Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation, Research Paper INT-470, United States Department of Agriculture, Forest Service, Ogden, USA, 1994.
[16] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. 28 (2008) 1–26.
[17] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a k-means clustering algorithm, J. R. Stat. Soc.: Ser. C: Appl. Stat. (1979) 100–108.
[18] K. Hornik, C. Buchta, A. Zeileis, Open-source machine learning: R meets Weka, Comput. Stat. 24 (2009) 225–232.
[19] A. Liaw, M. Wiener, Classification and regression by randomforest, R. News 2 (2002) 18–22.
[20] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: misc functions of the department of statistics (e1071), TU Wien, R Package Version 1.6-4, 2014.
[21] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, fourth ed. Springer, New York, 2002 (ISBN 0-387-95457-0).
[22] H. Wickham, R. Francois, dplyr: A grammar of data manipulation, R Package Version 0.4.1, 2015.
[23] S. Simeon, W. Shoombuatong, L. Preeyanon, V. Prachayasittikul, C. Nantasenamat, Predicting the oligomeric states of fluorescent proteins, PeerJ PrePrints 3 (2015) e1139.
[24] N.M. Razali, Y.B. Wah, Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests, J. Stat. Model. Anal. 2 (2011) 21–33.
[25] A.V. Zakharov, M.L. Peach, M. Sitzmann, M.C. Nicklaus, Qsar modeling of imbalanced high-throughput screening data in pubchem, J. Chem. Inf. Model. 54 (2014) 705–712.
[26] J. Alvarsson, M. Eklund, C. Andersson, L. Carlsson, O. Spjuth, J.E. Wikberg, Benchmarking study of parameter variation when using signature fingerprints together with support vector machines, J. Chem. Inf. Model. 54 (2014) 3211–3217.
[27] M.L. Calle, V. Urrea, Letter to the editor: stability of random forest importance measures, Brief. Bioinform. 12 (2011) 86–89.
[28] M. Bradai, M.T. Abad, S. Pissard, F. Lamraoui, L. Skopinski, M. de Montalembert, Hydroxyurea can eliminate transfusion requirements in children with severe β-thalassemia, Blood 102 (2003) 1529–1530.
[29] M.D. Cappellini, G. Graziadei, L. Ciceri, A. Comino, P. Bianchi, A. Porcella, G. Fiorelli, Oral isobutyramide therapy in patients with thalassemia intermedia: results of a phase II open study, Blood Cells Mol. Dis. 26 (2000) 105–111.
[30] K.S. Gates, T. Nooner, S. Dutta, Biologically relevant chemical reactions of N7-alkylguanine residues in DNA, Chem. Res. Toxicol. 17 (2004) 839–856.
[31] T. Heyduk, E. Heyduk, Molecular beacons for detecting DNA binding proteins, Nat. Biotechnol. 20 (2002) 171–176.

Predicting the oxygen affinity of human hemoglobin

Saw Simeon[1], Chuleeporn Phanus-umporn[1], Watshara Shoombuatong[1], Jarl E. S. Wikberg[2], Leif Bülow[3], and Chanin Nantasenamat*[1]

[1]**Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand**
[2]**Department of Pharmaceutical Biosciences, Uppsala University, Uppsala 751 24, Sweden**
[3]**Pure and Applied Biochemistry, Chemical Center, Lund University, Lund 221 00, Sweden**

ABSTRACT

Human hemoglobin (Hb) is instrumental in the transportation of oxygen (O_2) from lungs to tissues. In spite of several decades of investigations, the structural basis of O_2 binding has yet to be fully elucidated. Therefore, a comprehensive study on the physicochemical properties contributing to O_2 binding was performed herein on a large set of 326 non-redundant human Hb variants harboring a single point mutation on the α or β chains. Statistical and multivariate analyses were performed to gain insights into the origins of low and high O_2 binding affinities in human Hb. This study investigated the use of several classifiers (i.e. C5.0 decision tree, random forest, partial least squares discriminant analysis, artificial neural network and support vector machine) for predicting the O_2 affinity of Hb variants as a function of their sequence-based *z*-scale descriptors. Prediction performance of the constructed models were found to be robust as supported by accuracy, sensitivity and specificity greater than 80% and Matthews correlation coefficient greater than 0.6. Interpretation of the predictive models was also performed to gain useful insights on the underlying physicochemical properties giving rise to O_2 binding affinities, which may further guide structure-based design of novel human Hb variants with desired O_2 binding characteristics.

Keywords: hemoglobin, oxygen binding, oxygen affinity, quantitative structure-activity relationship, data mining

INTRODUCTION

Human hemoglobin (Hb) plays a crucial role as an oxygen (O_2) transporter from lungs to tissues. Hb is a 64 kDa as tetrameric protein consisting of two α and two β chains organized symmetrically as a dimer of $\alpha\beta$ dimers (i.e. $\alpha_1\beta_1$ and $\alpha_2\beta_2$) where α and β chains are comprised of 141 and 146 residues, respectively. The individual Hb subunit adopts a globin fold where 6 of the 8 α-helices are arranged as a 3-on-3 α-helical sandwich in which helices A, E and F helices are stacked on top of B, G and H. Such globin fold contains a heme co-factor that is axially coordinated to the Nε_2 atom of proximal histidine at the F8 position (i.e. αHis87 and βHis92). Hb binds to gaseous ligands such as O_2, CO or NO, which is situated between the Fe(II) atom of heme and the Nε_2 atom of distal histidine at the E7 position (i.e. αHis58 and βHis63).

Hb has been used as a classical model of protein allostery, a concept that serves as the basis for understanding the structure-function relationship governing protein function. Such allostery can be explained by the two-state model of Monod et al. (1965) where cooperative O_2 binding arises from the conversion of Hb structure between the high affinity R (oxy-Hb) and low affinity T (deoxy-Hb) states. Although there exists equilibrium between these two structural states, the allosteric control of these states can be modulated by pH (i.e. Bohr effect), temperature and allosteric effectors (i.e. 2,3-bisphosphoglycerate). The cooperativity of O_2 binding, which can be measured by Hill coefficient, is a parameter that describes the allosteric property of Hb (Turner et al., 1992). It is worthy to note that there

*Corresponding author. E-mail: chanin.nan@mahidol.edu

Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

Origin of anti-sickling activity via QSAR modeling[†]

Chuleeporn Phanus-umporn,[a] Nuttapat Anuwongcharoen,[a] Prasit Mandi,[b] Saw Simeon,[a] Watshara Shoombuatong,[a] and Chanin Nantasenamat[a,*]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

Sickle cell disease (SCD) is an autosomal recessive genetic disorder that has been recognized as a major public health problem by the WHO affecting 300,000 individuals worldwide. SCD arises from the A→T point mutation that causes the Glu6Val mutation in the hemoglobin β-globin gene thereby leading to sickle hemoglobin (HbS). At low oxygen tension, HbS are polymerized inside the red blood cells leading to gel or fiber formation thereby causing drastic decrease in the red cell deformability. Consequently, the complications of SCD leads to serious conditions such as anemia, microvascular occlusion, severe pain, stokes, renal dysfunction and infections. A lucrative therapeutic strategy to remedy complications of SCD affected patients is to employ anti-sickling agents for disrupting the HbS polymer. Therefore, this study aims to use quantitative structure-activity relationship (QSAR) modeling to elucidate the anti-sickling activity of 115 compounds. Briefly, the bioactivity of compounds were measured by a solubility assay described by Hofrichter et al. in which compounds were defined as active if their solubility ratios were greater than 1.06 and inactive if their solubility ratios were less than 1.06. Compounds were described by substructure descriptors and used in the construction of QSAR models via the random forest (RF) algorithm using rigorous validation. Good predictive performance was obtained as deduced by their accuracy (Ac), sensitivity (Sn), specificity (Sp) and Matthews correlation coefficient (MCC). Results indicated that Ac, Sn, Sp and MCC was in excess of 0.7 for the former three and greater than 0.5 for the latter statistical parameter. In addition, it was found that the top 5 important substructure descriptors for anti-sickling included conjugated double bond, arylchloride, Michael acceptor, alkene and vinylogous halide. Thus, this model is anticipated to be useful for guiding the design of robust compounds against the gelling activity of HbS.

1 Introduction

Hemoglobin (Hb) is a tetrameric iron-containing protein located within red blood cells (RBCs) that plays an important role in the transport of oxygen from lungs to tissues. Hemoglobin A (HbA) is the most common adult form of hemoglobin that is made up of two α-chains and two β-chains constituting 141 and 146 amino acids, respectively, in length. Mutations of globin chain genes causes structural alteration and perturbation of the globin chain that eventually culminates in Hb-associated diseases as seen in HbA hemoglobin S (HbS), hemoglobin C (HbC) and hemoglobin E (HbE) as well as thalassemia (i.e. decreased globin chain production).

Sickle cell disease (SCD) is a syndrome characterized by the presence of intra-erythrocytic HbS, which is formed from the polymerization of deoxygenated HbS (deoxy-HbS)[2]. During the tense (T) conformation or deoxy-state, which follows the passage of RBCs in the microcirculation, the Hb molecule undergoes a conformational change. HbS arises from the substitution of the hydrophilic glutamic acid at the sixth position (Glu6) of the β-globin chain to the hydrophobic valine (Val6)[3]. Afterwards, the Val6 residue of β_2-globin chain interacts hydrophobically with Phe85 and Leu88 of the other hemoglobin molecules. This interaction constitutes the basis for polymerization. The same β_2-globin chain of the first Hb molecule also contains Glu121, which interacts with Gly16 of the β_2-globin chain of a third Hb molecule. In the meanwhile, His20 from the α_2-globin chain of the first Hb molecule interacts with Glu6 from the β_1-globin chain of the third Hb molecule. Another interaction involves Asp73 from the β_2-globin chain of the first Hb interacts with Thr4 from the β_2-globin chain of the forth Hb molecule. Furthermore, there is also an interaction between Glu121 from the β_1-globin chain of the first Hb molecule and proline from α_2-globin chain and

[a] *Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; E-mail: chanin.nan@mahidol.edu*

[b] *Department of Community Medical Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

[†] Electronic Supplementary Information (ESI) available: Curated data set used in this study. See DOI: 10.1039/b000000x/

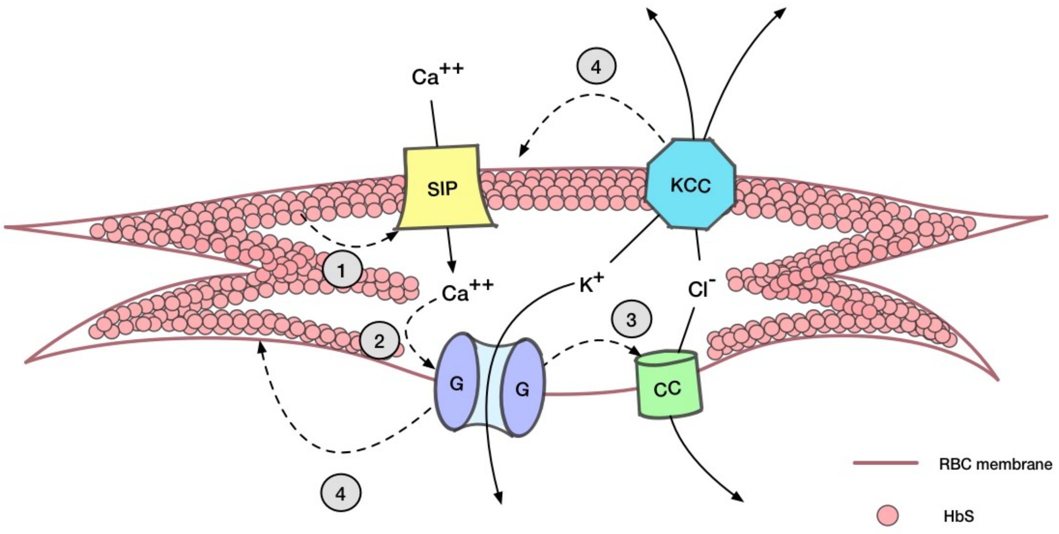


Fig. 1 Transport pathways contributing to sickle cell dehydration. KCC, indicates K:Cl cotransporter; SIP, sickling-induced pathway; G, Gardos pathway; and CC, Cl conductance pathway. Modified from [1]

His116 of β_2-globin chain of the fifth Hb molecule. This interaction model is supported by earlier reports on the nature of polymerization of sickle cells. These complex molecular interactions among the hemoglobin tetramers and with neighboring hemoglobin molecules play a vital role in the polymerization of the HbS cells [4]. Accordingly, this polymer forms and lengthens in helical fibers trigger cascade of cellular abnormalities, which participate in the overall pathophysiological mechanism [2,5]. Sickle cell dehydration is thought to result from deoxy-HbS polymers and dysregulation of cation homeostasis in sickle cell. Deoxy-HbS polymers activate a nonselective cation leak via calcium Ca^{2+} entry by sickling-induced (SIP) pathway. The increasing of cytoplasmic Ca^{2+} triggers activation of the Gardos (G) pathway, which mediates rapid K^+ efflux and water loss. After that, the K^+ efflux can be balanced by Cl^- exit via a chloride conductance (CC) pathway. This abnormality may thus facilitate a vicious spiral in which sickling and Gardos channel activation reinforce each other to dehydrate the cell [1] (Fig. 1).

Furthermore, Deoxy-HbS polymers becomes denatured and hemichromes concentrate at the internal side of the membrane together with protein cytoskeleton, in particular protein band 3. This process comes along with the loss of heme and with the liberation of Fe^{3+} which promotes the oxidizing microenvironment. All of these phenomena give rise to extra and intravascular hemolysis, sickle vasculopathy and vasoocclusive disease [5,6]. Therefore, the rational for our study was as follow the treatment of SCD is based on pathophysiology to inhibit the deoxy-HbS polymerization by binding to small molecules.

Quantitative structure-activity relationship (QSAR) represents an important approach for elucidating the origin of biological activity for a set of compounds of interest as a function of their molecular descriptors. The resulting QSAR models can reveal molecular features that are essential for the biological activity of potent compounds that can subsequently be used as therapeutic agents. Therefore, we applied QSAR to rationalize the underlying physicochemical features defining anti-sickling activity in several series of compounds reported by Abraham et al. In this study, we examined the utility of several sets of substructure fingerprint descriptors in modeling the anti-sickling activity. Important physicochemical features were then decoded from such predictive QSAR models as to discern the crucial chemical substructures influencing the anti-sickling activity.

2 Materials and Methods

A schematic summary of the QSAR modeling process performed in this study is provided in Figure 2.

Data collection

Compounds with anti-sickling activity will be compiled from the literature [7–13], which afforded an initial set of 133 compounds. Removal of redundant compounds resulted in a final set of 115 compounds. Compounds will be treated with the QSAR curation workflow as described by [14]. Chemaxon standardizers was utilized using the same protocol from our previous studies [15]. The anti-sickling activity will be represented by the solubility of HbS(drugs) / solubility of HbS (control). Solubility ratios were greater than 1.06 have been estimated as necessary for decreasing the clinical severity of sickle cell disease. Therefore, compounds will be classified into two types, which is comprised of 32 active compounds (solubility ratios ≥ 1.06) and 83 inactive compounds (solubility rations < 1.06).

Compound descriptors

Fingerprints descriptors provides descriptions of the constituting substructures inherently present in a molecule. These fingerprints will calculate using PaDel-Descriptor software. The software currently calculates 1875 descriptors (1444 1D, 2D

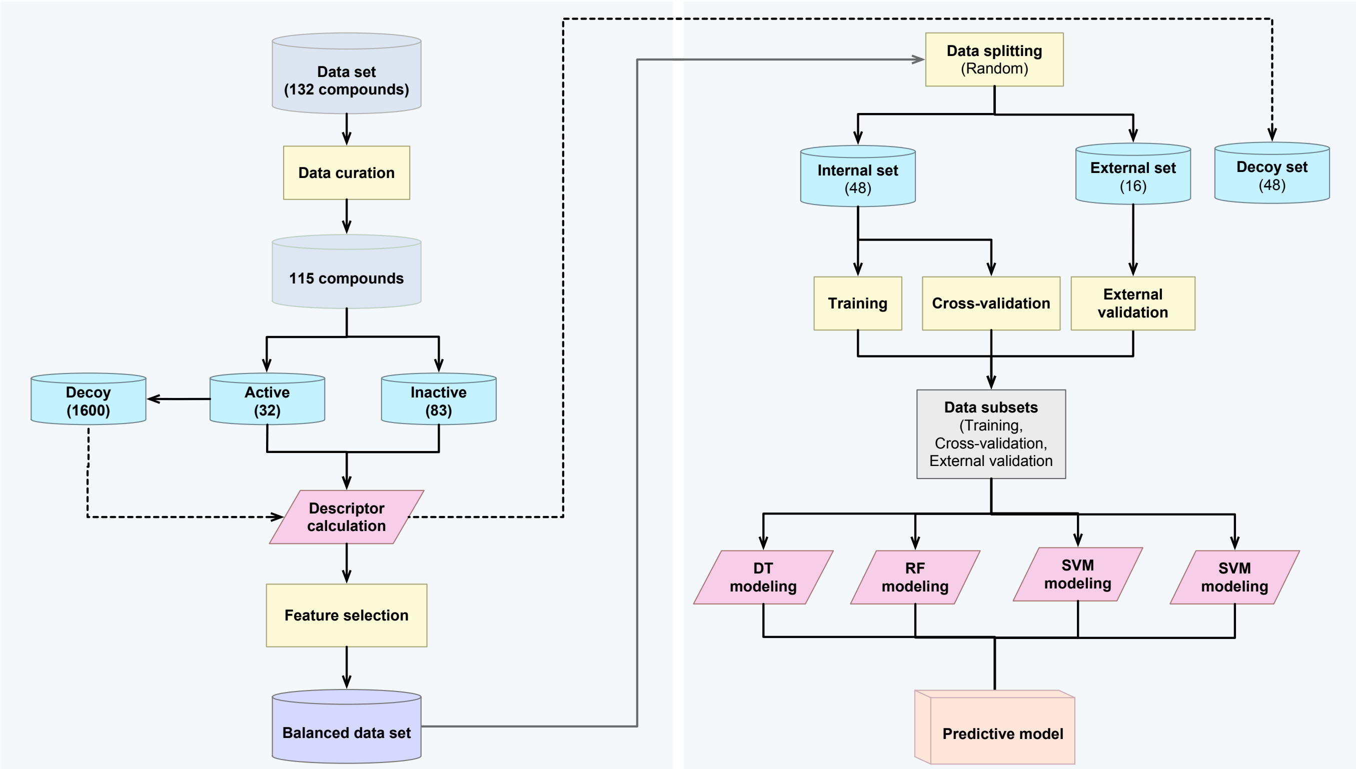


Fig. 2 Workflow of QSAR modeling for investigating anti-sickling activity.

descriptors and 431 3D descriptors) and 12 types of fingerprints (total 16092 bits). In this project, we performed only 12 types of fingerprints. Fingerprints are calculated mainly using The Chemistry Development Kit which include Fingerprinter, ExtendedFingerprinter, EStateFingerprinter, GraphOnlyFingerprinter, MACCSFingerprinter, PubchemFingerprinter, SubstructureFingerprinter, Substructure- FingerprintCount, KlekotaRothFingerprinter, KlekotaRothFingerprintCount, AtomPairs2D- Fingerprinter and AtomPairs2DFingerprintCount. The numerical value represents the fingerprints that present in a molecule and frequency of fingerprints that present in a molecule as fingerprint counting.

2.1 Data filtering

2.2 Data balancing

The collected data set contained imbalanced data, in which 32 active compounds and 83 inactive compounds. To create a balanced data set, the undersampling approach will be performed by the random process 100 times in the majority class (inactive compounds). This random process will be selected a set of 32 inactive compounds.

2.3 Data splitting

In order to obtain accurate and generalized of QSAR model, the 115 compounds will be divided into two parts comprising 80% as the internal set and 20% as the external set using random 100 times samplings. These processes will be performed 48 as the internal set and 16 as the external set for establishing QSAR model. The internal set will be used to train the model, while the external set will be used to externally validate the performance the predictive model. The 5-fold cross-validation (5-fold CV) and external set will be used to validate the robustness and reliability of predictive model. The 5-fold CV will be performed on the internal set. The data is divided into 5 groups. Practically, 1 group from 5 is left out act as the external validation, while the remaining are used for training the models. This process is repeated until all the data samples had a chance to be left out. External validation evaluates the stability of predictive model by act as unknown data.

2.4 QSAR modeling

Machine learning methods are used to construct classification models of anti-sicking agents as a function of their physicochemical properties of compounds. Models were generated using several machine learning methods as implemented under the R programming environment, version 3.3.2 Insert the appropriate in-text reference, as described below.

Decision tree (DT) is a supervised approach to classify samples into categories of activity of interest as rule-based branching tree using the divide and conquer algorithm. This algorithm is considered to be one of the most transparent learning in which a series of readily understandable. The construction involves two steps for growing and pruning. Growing starts from the root node, which branches out from internal node that subsequently end up as leaf nodes. Internal nodes represent fingerprint descriptors, branches describe the criteria of counting number value of fingerprint, and leaf nodes represent Y categorical classed (i.e., active and inactive). Once the tree is fully grown, the grown tree is pruned as a function of the predictive performance. The advantage of pruning is that it reduces the complexity of interpretation. C50 package will be used to construct the DT model.

Random forest (RF) is an ensemble classifier that produces multiple decision trees, using a randomly selected subset of training samples and variables. The classification starts at the root node, in which the data set at the node is split according to the value of descriptors that are selected such that the descriptors of different activities are predominantly moved to different branches. The classification is obtained by averaging the results of all trees by a majority vote from each tree. The RF classifier will be generated using the R package. The Gini index will be used to investigate the ability of a potential discriminative of each feature. The mean decrease of the Gini index (MDGI) will be considered for selecting important descriptors. The descriptors with the largest value of MDGI represented the most important feature as the descriptor contributes most to prediction performance.

Support vector machine (SVM) constructs a hyperplane or set of hyperplanes that separate different class from each other. SVM handle non linearity by using a kernel function to map the input variables into a very high-dimensional space in which a hyperplane can be used to do the separation. Intuitively, a good separate between classes is achieved by a hyperplane that has the largest distance to the nearest training observation of any class. The goal of SVM modeling is to find the optimal hyperplane that separates clusters in such a way that observation from one class are on one side of the plane and observations from other classes are on the other side of plane. The vectors near the hyperplane are the support vectors. Therefore, SVM analysis should produce a hyperplane that completely separate the observations into non-overlapping groups. SVM will be constructed using the R package e1071.

Artificial neural network (ANN) is a machine learning method that mimics the human brain, which is composed of networks of interconnected neurons that function in relaying messages in the form of electrochemical signals. Brain cells are composed of dendrites, cell bodies and axons. A synapse is the connection between the axons of the nerve cell with the dendrites of an adjacent nerve cell. In a synapse, signals are transmitted from one cell to the next as neurotransmitters. Similar to the function of the human brain, the artificial neuron is interconnected in a feed-forward manner from the first through the last nodes, in which the connection between nodes of different layers is assigned as a weight, which can be expressed as a strength of the input data. The train function from the R package caret will be used to construct the artificial neural network model while fine-tuning the parameters. The train function from the R package caret will be also used to obtain optimal parameters, and the nnet function from the R package nnet will be used to train the models.

2.5 Validation of QSAR model

Model validation is the most important step for evaluating the results of modeling. There are many statistical parameters that used to assess the effective and efficiency of constructed predictive models using several statistical metrics as described hereafter.

Accuracy (Ac) is the percentage of correctly classification that relative to the total number of the data. Although it is commonly used to assess the predictability of predictive models, Ac is not an optimal measure of unbalanced data. Therefore, Matthews correlation coefficient (MCC) is appropriated to measure due to insensitivity in unbalanced data. Sensitivity (Sn), sometimes considered as true positive (TP) rate, measures the proportion of positives that are correctly identified, whereas specificity (Sp) measures the proportion of negatives that are correctly identified. FP is false positives while FN is false negative.

These parameters can be calculated using the equation described below:

$$Ac = \frac{TP+TN}{(TP+TN+FP+FN)} \times 100 \tag{1}$$

$$Sn = \frac{TP}{(TP+FN)} \times 100 \tag{2}$$

$$Sp = \frac{TN}{(TN+FP)} \times 100 \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

where *TP*, *TN*, *FP* and *FN* represents the instances of true positives, true negatives, false positives and false negatives, respectively. The value of MCC ranges from –1 to 1 in which an MCC of 1 indicates the best possible prediction scenario while an MCC of –1 indicates the worst possible prediction. On the other hand, an MCC of 0 is indicative of random prediction.

3 Results and Discussion

The performance of the training set resulted in accuracy of ...

Fig. 3 Transport pathways contributing to sickle cell dehydration .

Fig. 4 Plot of experimental versus predicted pIC$_{50}$ values for models constructed with 12 different fingerprint descriptors. Shown are models built with CDK fingerprint (A), CDK extended fingerprint (B), E-State fingerprint (C), CDK graph only fingerprint (D), MACCS fingerprint (E), PubChem fingerprint (F), substructure fingerprint (G), substructure fingerprint count (H), Klekota-Roth fingerprint (I), Klekota-Roth fingerprint count (J), 2D atom pairs (K) and 2D atom pairs count (L).

Table 1 Performance summary of QSAR models for predicting anti-sickling agents.

Descriptor class	*N*	Training set				5–fold CV set				External set			
		Ac	*Sn*	*Sp*	*MCC*	*Ac*	*Sn*	*Sp*	*MCC*	*Ac*	*Sn*	*Sp*	*MCC*
CDK	885	99.81±0.67	99.92±0.56	99.72±1.15	1.00±0.01	78.50±5.98	77.37±7.02	80.52±6.70	0.57±0.12	81.06±10.77	80.50±12.84	84.26±12.83	0.63±0.22
CDK extended	892	99.88±0.49	99.88±0.68	99.88±0.68	1.00±0.01	78.83±5.72	78.44±6.55	79.77±6.11	0.58±0.11	80.19±9.36	80.75±10.95	82.50±11.69	0.62±0.18
CDK graph only	441	96.52±2.48	95.93±3.51	97.29±2.58	0.93±0.05	76.29±7.34	76.04±8.06	76.99±7.53	0.53±0.15	77.63±11.39	77.80±13.23	79.89±12.93	0.56±0.23
E-State	18	90.69±3.01	90.28±4.74	91.56±3.69	0.82±0.06	79.88±7.19	79.02±7.51	81.24±7.66	0.60±0.14	82.13±8.62	81.12±10.91	86.25±11.03	0.66±0.17
MACCS	103	97.23±2.02	98.17±2.50	96.53±3.30	0.95±0.04	76.52±5.41	76.68±6.38	76.82±5.50	0.53±0.11	79.19±9.31	80.47±11.31	80.32±10.88	0.60±0.19
PubChem	299	97.10±2.48	97.06±3.31	97.30±2.78	0.94±0.05	77.31±6.43	75.99±7.08	79.29±6.63	0.55±0.13	78.75±9.52	77.79±11.44	83.01±11.91	0.59±0.19
Substructure	38	92.75±3.14	95.30±4.18	90.78±3.93	0.86±0.06	79.10±6.87	79.71±7.79	78.94±6.92	0.58±0.14	81.56±8.86	82.96±11.35	82.81±10.97	0.64±0.18
Substructure count	45	95.58±2.80	98.52±2.43	93.15±4.02	0.91±0.05	80.48±5.46	81.85±6.99	79.67±5.22	0.61±0.11	82.38±8.99	84.82±11.32	83.27±11.29	0.66±0.17
Klekota-Roth	340	98.00±1.95	98.80±1.96	97.32±2.86	0.96±0.04	78.27±5.71	79.12±6.45	77.96±6.22	0.57±0.11	79.19±9.44	81.84±12.08	79.89±11.67	0.60±0.19
Klekota-Roth count	366	98.88±1.49	99.31±1.52	98.53±2.40	0.98±0.03	78.40±5.33	78.50±6.38	78.65±5.17	0.57±0.11	78.81±9.60	80.71±13.37	80.88±11.22	0.59±0.19
2D atom pairs	133	94.58±3.45	94.94±4.51	94.46±3.56	0.89±0.07	77.04±5.71	76.76±6.38	77.73±6.11	0.54±0.11	78.69±10.20	78.57±11.45	81.47±12.54	0.59±0.20
2D atom pairs count	167	99.19±1.02	99.96±0.40	98.48±1.95	0.98±0.02	77.79±6.39	78.47±7.05	77.52±6.65	0.56±0.13	77.63±10.28	79.17±12.19	78.96±11.92	0.57±0.20

Table 2 Performance difference of Training against 5-fold CV (Diff$_{Tr-CV}$) and external (Diff$_{Tr-Ext}$) sets.

Descriptor class	*N*	Diff$_{Tr-CV}$				Diff$_{Tr-Ext}$			
		Ac	*Sn*	*Sp*	*MCC*	*Ac*	*Sn*	*Sp*	*MCC*
CDK	885	21.31	22.55	19.2	0.43	18.75	15.66	15.46	0.37
CDK extended	892	21.05	21.44	20.11	0.42	19.69	17.38	17.38	0.38
CDK graph only	441	20.23	19.89	20.3	0.4	18.89	16.04	17.4	0.37
E-State	18	10.81	11.26	10.32	0.22	8.56	4.03	5.31	0.16
MACCS	103	20.71	21.49	19.71	0.42	18.04	17.85	16.21	0.35
PubChem	299	19.79	21.07	18.01	0.39	18.35	14.05	14.29	0.35
Substructure	38	13.65	15.59	11.84	0.28	11.19	12.49	7.97	0.22
Substructure count	45	15.1	16.67	13.48	0.3	13.2	15.25	9.88	0.25
Klekota-Roth	340	19.73	19.68	19.36	0.39	18.81	18.91	17.43	0.36
Klekota-Roth count	366	20.48	20.81	19.88	0.41	20.07	18.43	17.65	0.39
2D atom pairs	133	17.54	18.18	16.73	0.35	15.89	13.47	12.99	0.3
2D atom pairs count	167	21.4	21.49	20.96	0.42	21.56	21	19.52	0.41

4 Conclusion

...

Acknowledgments

...

References

1 C. H. Joiner, *Blood*, 2008, **111**, 3918–3919.

2 C. Madigan and P. Malik, *Expert Rev Mol Med*, 2006, **8**, 1–23.

3 E. M. Novelli and M. T. Gladwin, *Chest*, 2016, **149**, 1082–1093.

4 I. O. Nurain, C. O. Bewaji, J. S. Johnson, R. D. Davenport and Y. Zhang, *Molecular Pharmaceutics*, 2016.

5 M.-H. OdiÃĺvre, E. Verger, A. C. Silva-Pinto and J. Elion, *The Indian Journal of Medical Research*, 2011, **134**, 532–537.

6 N. C. e. Fernando Ferreira Costa, *Sickle Cell Anemia: From Basic Science to Clinical Practice*, Springer International Publishing, 1st edn, 2016.

7 M. O. Fatope and D. J. Abraham, *Journal of Medicinal Chemistry*, 1987, **30**, 1973–1977.

8 P. E. Kennedy, F. L. Williams and D. J. Abraham, *J Med Chem*, 1984, **27**, 103–5.

9 D. J. Abraham, A. S. Mehanna and F. L. Williams, *J Med Chem*, 1982, **25**, 1015–7.

10 D. J. Abraham, M. Mokotoff, L. Sheh and J. E. Simmons, *J Med Chem*, 1983, **26**, 549–54.

11 D. J. Abraham, D. M. Gazze, P. E. Kennedy and M. Mokotoff, *Journal of Medicinal Chemistry*, 1984, **27**, 1549–1559.

12 D. J. Abraham, P. E. Kennedy, A. S. Mehanna, D. C. Patwa and F. L. Williams, *J Med Chem*, 1984, **27**, 967–78.

13 D. J. Abraham, A. S. Mehanna, F. S. Williams, J. Cragoe, E. J. and J. Woltersdorf, O. W., *J Med Chem*, 1989, **32**, 2460–7.

14 D. Fourches, E. Muratov and A. Tropsha, *Journal of chemical information and modeling*, 2010, **50**, 1189–1204.

15 S. Simeon, R. Möller, D. Almgren, H. Li, C. Phanus-umporn, V. Prachayasittikul, L. Bülow and C. Nantasenamat, *Chemometrics and Intelligent Laboratory Systems*, 2016, **151**, 51–60.

QSAR modeling of methemoglobin reduction by electron mediators

Chuleeporn Phanus-umporn[a], Watshara Shoombuatong[a],
Saowapak Choomwattana[a], Nuttapat Anuwongcharoen[a,b], Prasit Mandi[a],
Virapong Prachayasittikul[b], Chanin Nantasenamat[a,*],

[a] *Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*
[b] *Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

[*]Corresponding author: chanin.nan@mahidol.ac.th

Hemoglobin vesicle (HbV) is an artificial oxygen carrier based on liposome-encapsulated hemoglobin which can potentially be used as an alternative source of blood. The advantages of HbV is that it can reduce the risk of infection and blood type mismatching while being easily excreted and has long shelf-life. However, the limitation of HbV lies in the auto-oxidation of ferrous Hb thereby leading to increased level of ferric methemoglobin (metHb). This condition leads to impaired oxygen transport and reduces the half-life of HbV. Previously, Kettisen *et al.* have found that electron mediator dyes can reduce metHb production in HbV. Therefore, this study aims to use quantitative structure-activity relationship (QSAR) modeling to elucidate the origin of electron mediation amongst the set of 15 dyes. These molecules were described by substructure fingerprints, molecular descriptors and quantum chemical descriptors and subsequently used in the construction of QSAR models using support vector machine, which afforded good predictive performance with accuracy, sensitivity, specificity and Matthew's correlation coefficient (MCC) of 99.00±3.16 %, 98.57±4.52 %, 100.00±0.00 % and 0.98±0.06, respectively, for leave-one-out cross validation as well as values of 98.00±6.32 %, 96.67±10.54 %, 100.00±0.00 % and 0.97±0.11, respectively, for external validation. It was found that important descriptors for electron mediation of metHb included HATS4m, Mor18m, R7e+, R7m+ and Mor12e, which corresponded to atomic mass and electronegativity of the molecule. Thus, this model is anticipated to be useful for guiding the design of robust compounds against the auto-oxidation of HbV.

Keywords: hemoglobin vesicles, auto-oxidation, electron mediators, methemoglobin, quantitative structure-activity relationship; QSAR

Biochimica et Biophysica Acta 1864 (2016) 29–41

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Structural and biochemical characterization of two heme binding sites on α_1-microglobulin using site directed mutagenesis and molecular simulation

Sigurbjörg Rutardottir [a], Elena Karnaukhova [b], Chanin Nantasenamat [c,d], Napat Songtawee [c], Virapong Prachayasittikul [d], Mohsen Rajabi [b], Lena Wester Rosenlöf [a], Abdu I. Alayash [b], Bo Åkerström [a,*]

[a] Division of Infection Medicine, Lund University, Lund, Sweden
[b] Laboratory of Biochemistry and Vascular Biology, Division of Hematology Research and Review, Center for Biologics Evaluation and Research, Food and Drug Administration, MD, USA
[c] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand
[d] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form 12 September 2015
Accepted 7 October 2015
Available online 21 October 2015

Keywords:
α_1-Microglobulin
Heme
Site-directed mutagenesis
Molecular simulation

ABSTRACT

Background: α_1-Microglobulin (A1M) is a reductase and radical scavenger involved in physiological protection against oxidative damage. These functions were previously shown to be dependent upon cysteinyl-, C34, and lysyl side-chains, K(92, 118,130). A1M binds heme and the crystal structure suggests that C34 and H123 participate in a heme binding site. We have investigated the involvement of these five residues in the interactions with heme.
Methods: Four A1M-variants were expressed: with cysteine to serine substitution in position 34, lysine to threonine substitutions in positions (92, 118, 130), histidine to serine substitution in position 123 and a wt without mutations. Heme binding was investigated by tryptophan fluorescence quenching, UV–Vis spectrophotometry, circular dichroism, SPR, electrophoretic migration shift, gel filtration, catalase-like activity and molecular simulation.
Results: All A1M-variants bound to heme. Mutations in C34, H123 or K(92, 118, 130) resulted in significant absorbance changes, CD spectral changes, and catalase-like activity, suggesting involvement of these side-groups in coordination of the heme-iron. Molecular simulation support a model with two heme-binding sites in A1M involving the mutated residues. Binding of the first heme induces allosteric stabilization of the structure predisposing for a better fit of the second heme.
Conclusions: The results suggest that one heme-binding site is located in the lipocalin pocket and a second binding site between loops 1 and 4. Reactions with the hemes involve the side-groups of C34, K(92, 118, 130) and H123.
General significance: The model provides a structural basis for the functional activities of A1M: heme binding activity of A1M.

1. Introduction

The Lipocalin protein family includes approximately 50 members from bacteria, plants and animals [1–3]. Although their structures are similar they have different and mostly non-related functions. A common feature is the folding of the single polypeptide into a β-barrel consisting of eight antiparallel β-strands with a closed bottom, an open end and a hydrophobic pocket which can carry small lipophilic ligands [4]. One of the members of the Lipocalin protein family, α_1-microglobulin (A1M), is a small (26 kDa), heterogeneously charged glycoprotein found in blood-plasma and interstitial fluid of all human tissues [5–7]. A1M is expressed mainly in the liver and co-synthesized with bikunin, a proteinase inhibitor and component of extracellular matrix, from the *AMBP*-gene (α_1-microglobulin-bikunin precursor protein) [8,9]. After cleavage of the precursor protein the two mature proteins are secreted separately into the blood-stream [10]. A large fraction of A1M forms complexes with other plasma proteins in blood. Approximately 50% of the protein is bound to immunoglobulin A (IgA) [11]. Less abundant complexes are formed with albumin and prothrombin [12].

The physiological function of A1M has been suggested to be protection of cells and tissues against oxidative stress induced by extracellular hemoglobin and free radicals [7]. This is supported by several recent papers, which demonstrate that A1M indeed can protect cell cultures and organ explants against oxidative damage [13–15], and therapeutic *in vivo* effects of the protein were shown in animal models of preeclampsia and hemoglobin-induced kidney damage [16,17]. Mechanistically, the antioxidative protection is achieved by reductase and radical-

Abbreviations: A1M, α_1-microglobulin; AMBP, A1M-bikunin precursor protein; IgA, immunoglobulin A; ROS, reactive oxygen species; FPLC, fast protein liquid chromatography; SPR, surface plasmon resonance; MD, molecular dynamics; CD, circular dichroism; RMSD, root mean square deviation; RMSF, root mean square fluctuation.

* Corresponding author at: Division of Infection Medicine, Lund University, Sölvegatan 19, 22184 Lund, Sweden.
E-mail address: sigurbjorg.rutardottir@med.lu.se (B. Åkerström).

http://dx.doi.org/10.1016/j.bbapap.2015.10.002
1570-9639/© 2015 Elsevier B.V. All rights reserved.

binding activities of A1M, and it was shown that the C34 unpaired thiol group and the side-chains of K92, 118 and 130 of A1M are involved in the reductase and radical-binding activities [18–20]. A1M has also been shown to bind heme[1] *in vitro* and *in vivo* [21–23] and was suggested to contribute to heme degradation by a mechanism still unknown, but which involves proteolytic activation of A1M [21]. The heme binding of A1M was recently demonstrated to result in a trimeric complex with heme in a 1:2 stoichiometry [24]. Furthermore, a heme binding site, which involves the C34 thiol group and H123 imidazole ring, has been proposed based on the crystal structure of A1M [25].

Free heme in plasma and extracellular fluid is the result of degradation of hemoglobin and other proteins carrying a heme prosthetic group. The heme-group presents a potential chemical threat to the organism by generation of reactive oxygen species (ROS), which are toxic and can cause severe damage to cells and tissues [26]. We hypothesized that the above-mentioned side-chains of A1M are involved in the heme binding. To test this, we investigated wild type (Wt)-A1M and the previously constructed mutated forms of A1M: C34S- and K [3]T-A1M [27], the latter carrying Lys → Thr substitutions in positions 92, 118 and 130. We also constructed, prepared and investigated a new mutated form, H123S-A1M (Fig. 1A). The results suggest a two binding site-model where the three lysine side-chains participate in coordination of the first heme-group and the C34 and H123 side-chains coordinate the second heme-group.

2. Materials and methods

2.1. Reagents and proteins

Heme (hemin; ferriprotoporphyrin IX chloride) was purchased from Porphyrin Products, Inc. (Logan, UT, U.S.A.). Stock solutions of heme were prepared by dissolving heme in DMSO to 10 mM and used within 10 h. Clarity Western ECL Substrate was from Bio-Rad Bio-Rad Laboratories (Hercules, CA, USA). Heme-agarose was from Sigma-Aldrich, Sweden. Sepharose CL-4B was from GE Healthcare, Sweden. Pierce BCA Protein Assay kit from Thermo Scientific, Sweden. AcroPrep Advance 96 Filter Plate 1.2 μM supor® was from PALL Corporation (Port Washington, NY, USA), and Nunc TM 96-Well Microplates were purchased from Thermo Scientific.

2.2. Recombinant A1M

Wild-type (Wt) and mutated variants of A1M were expressed in *Escherichia coli* (*E. coli*) as described [27]. Using site-directed mutagenesis a Cys → Ser substitution was introduced at amino acid position 34 in the C34S-A1M mutant, Lys → Thr substitutions at positions 92, 118, and 130 in the K [3]T-A1M mutant, and a His → Ser substitution at position 123 in the H123S-A1M mutant. The four forms of recombinant A1M, Wt-A1M, C34S-A1M, K [3]T-A1M and H123S-A1M, were purified and refolded as described [27] with the addition of ion-exchange chromatography and size exclusion purification steps as follows. The protein solution was applied to a column of DEAE-Sephadex A-50 (GE Healthcare, Uppsala, Sweden) equilibrated with the starting buffer (20 mM Tris–HCl, pH 8.0). A1M was eluted at a flow rate of 1 ml/min using a linear pH gradient consisting of 250 ml starting buffer and 250 ml elution buffer (20 mM Tris–HCl, 0.5 M NaCl, pH 8.0). Size-exclusion chromatography was run on a Superose 12 column obtained from GE Healthcare using Äkta purifier 10 system (GE Healthcare) run at a flow-rate of 1 ml/min. Wt-A1M without the N-terminal His$_8$-tag was a generous gift from A1M-Pharma AB.

2.3. Secondary structure estimation by far-UV circular dichroism

The secondary structure of A1M variants was determined with a Jasco-J810 spectropolarimeter instrument using a 2 mm quartz cuvette at continuous mode at speed 20 nm/min, band width of 1 nm and a 1 nm resolution. Temperature was held at 22 °C by a Peltier thermostat. The concentration of protein was 10 μM in 20 mM Tris–HCl pH 8.0 + 0.15 M NaCl. Each spectrum is a mean of five replicates. After subtraction of the buffer spectrum and recalculation into mean residue molar elipticity units, the CD spectra were analyzed using the CDPro package to determine different secondary structure content [28].

2.4. Spectrophotometric analysis of heme binding

Absorbance spectra were measured on a Beckman (Beckman Instruments, Fullerton, CA) DU 800 spectrophotometer using a scan rate of 1200 nm/min in the UV–VIS region between 250 and 700 nm at 22 °C. The protein concentration was 10 μM in 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl, which was also used as blank. Heme was dissolved in DMSO to 10 mM and diluted in 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl to 3 mM. Volumes of this stock solution were then added to each protein solution, to a final concentration 20 μM. Blanks with equivalent concentration of DMSO were used for samples with heme. Protein solutions were scanned immediately after mixing with heme and after 1 h or 24 h.

2.5. Fluorescence spectroscopy

Fluorescence measurements were performed using a Jasco J-810 spectropolarimeter (equipped with a FMO-427 monochromator) with a 100 μl quartz cuvette (Hellma Precision Cell, Type no. 105.251-QS, light-path length 3 mm in both excitation and emission modes) under nitrogen flow. Temperature was held at 22 °C by a Peltier thermostat. Tryptophan fluorescence was measured by exciting at wavelength 295 nm. Fluorescence emission was detected from 310 to 400 nm with slits set at 5 nm bandpass. Emission spectra were recorded three times, averaged, and the peak at 346 nm was measured. A1M protein concentration was 1 μM in 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl. At this concentration, the fluorescence signal of the protein was well resolved within the detector sensitivity (set at 900 V). The emission spectra of the protein solution (100 μl) without heme were recorded first and subsequently heme from a stock-solution was added (1 μl at a time, up to 4 μl). Blank subtractions were made for Tris–HCl, pH 8.0, 0.15 M NaCl since the concentration of DMSO did not affect the emission spectra.

2.6. Induced circular dichroism

All samples containing heme were evaluated for the protein-induced chirality in the near-UV to visible CD range of 300–700 nm (referred to as visible CD). The visible CD measurements were performed on a Jasco J-815 Spectropolarimeter (JASCO Co., Japan) with the temperature maintained at 25 ± 0.5 °C. The spectra were recorded using a scan speed of 100 nm/min, bandwidth of 1.0 nm, and resolution of 0.2 nm, and accumulated in triplicate. A protein alone (Wt-A1M or mutant, respectively) was used as a blank. To accumulate the induced CD, A1M was used at an initial protein concentration of ~45 μM, and the measurements were performed in a quartz cuvette with 1 cm pathlength. Heme content in the samples varied from ~4.5 μM to ~50 μM by using calculated amounts of the heme 2 mM stock solution in DMSO. The content of DMSO in the mixed samples did not exceed 4%. An ellipticity of induced CD spectra was expressed in millidegrees (mdeg).

2.7. Analytical size exclusion chromatography

Samples were size-fractionated on a Superose 12 10/300 fast protein liquid chromatography (FPLC) column (GE Healthcare, Uppsala,

[1] The term “heme” is used in this article to denote both the ferrous and ferric forms.

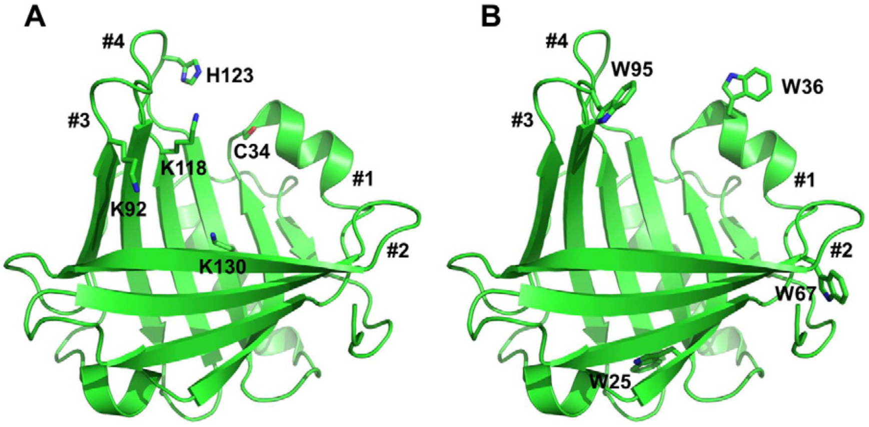


Fig. 1. Three-dimensional structure of A1M. The illustration was generated using PyMOL [32] and coordinates from the crystal structure of human A1M [25]. β-strands and α-helices are shown as green ribbons. The four loops at the open end of the lipocalin pocket are labeled #1–#4. (A) Side-chains of C34, K92, K118, K130 and H123 are shown in green sticks. (B) Side-chains of W25, W34, W67 and W95 are shown in green sticks.

Sweden). A1M/heme molar ratio was 1:2. Heme was dissolved in DMSO to 10 mM and added from a 3 mM stock solution prepared as described above. 500 μl of sample was applied to the column after 1 h of incubation with heme. The proteins were eluted with 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl and run at 0.5 ml/min. Fractions of 0.5 ml were collected.

2.8. Native PAGE and Western blot of A1M incubated with heme

A1M (5 μM) in 20 mM Tris–HCl, pH 8.0 + 0.15 M NaCl was incubated for 30 min or 24 h with 50, 10 and 0.1 μM heme. Samples were mixed with equal amounts of sample buffer for native PAGE, pH 6.8, and subjected to 12% Criterion™ TGX™ Precast Gels (Bio-Rad). The gels were either stained with Coomassie Brilliant Blue R-250 (BDH Chemicals, Ltd. Poole, UK) or transferred to polyvinylidene difluoride (PVDF) membranes using Trans-Blot Turbo system from Bio-Rad. The membranes were incubated in Clarity Western ECL Substrate and imaged with a digital imager (BioRad).

2.9. Binding of A1M to heme-agarose

Beads of heme-agarose and Sepharose were washed three times with an excess of 10 mM Tris–HCl pH 8.0 + 0.125 M NaCl, yielding a final 1:1 suspension in this buffer. The proteins were diluted in 10 mM Tris–HCl pH 8.0 + 0.125 M. Dilution series of all proteins were made to final concentrations of 10, 7.5, 5, 2.5, 1.25, and 0.625 μM. Seventy-five μl were then transferred to Nunc TM 96-Well Microplates in duplicates (one set for incubation with heme agarose and control Sepharose and one without beads). Twenty μl 50% heme-agarose or Sepharose were pipetted into the wells, using large-opening pipette-tips, and the plates were incubated in RT on a shaker for 30 min. The protein/beads mixtures were then transferred to AcroPrep Advance 96 Filter Plate and spun at 2000 g for 2 min. Twenty-five μl of non-incubated samples and 25 μl of each flow-through were transferred to a new Nunc TM 96-Well Microplate, 200 μl of BCA reagent was added and the plates incubated at 37 °C for 30 min. The absorbance was measured at wavelength 550 nm, using a Multilabel Counter (Victor™ 1420 Perkin Elmer Life and

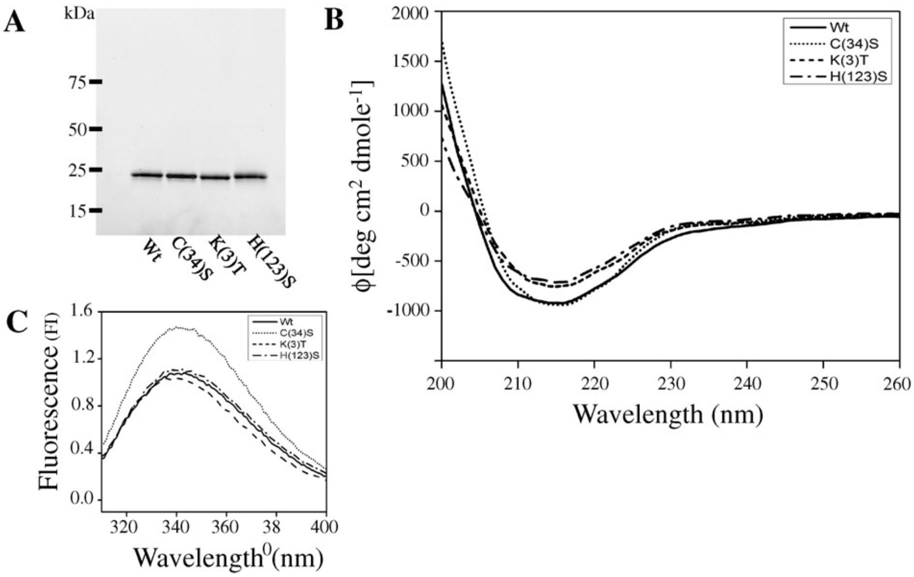


Fig. 2. Size, purity and spectroscopic properties of the four variants of recombinant A1M. (A) SDS-PAGE was performed in the presence of mercaptoethanol (T = 12%). Approximately 2 μg of Wt-A1M, C [34]S-A1M, K [3]T-A1M, and H(123)S-A1M were applied to the gel and stained with Coommassie. (B) Far-UV CD spectra of A1M variants (10 μM in 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl). Similar spectra of Wt-A1M, C [24]S and K [3]-T-A1M were published in Kwasek et al. [27] and are included here for comparison of H(123)S-A1M. (C) Fluorescence spectra of A1M variants (1 μM) in 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl).

Table 1
Secondary structure prediction of recombinant A1M-variants.

	α-Helix (%)	β-Sheet (%)	Turns (%)	Unordered (%)
Wt	4.3	43.3	22.2	30.0
C [34]S	2.0	48.6	18.6	29.6
K [3]T	1.2	44.9	20.9	32.3
H(123)S	4.2	47.2	19.5	28.9

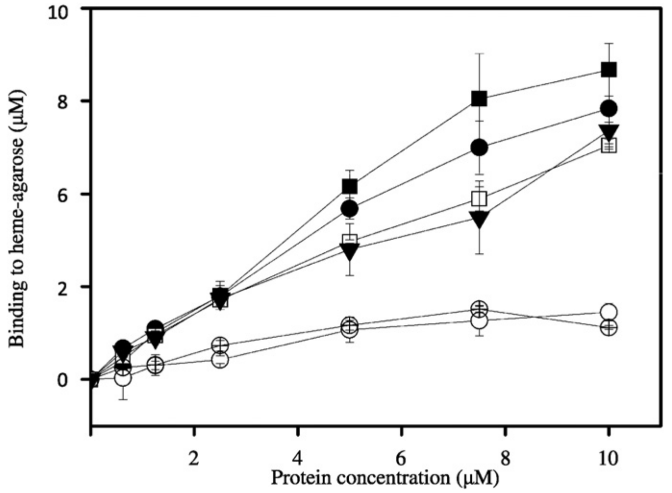


Fig. 4. Binding of A1M-variants to heme-agarose. The proteins were diluted in 10 mM Tris–HCl pH 8.0, 0.125 M NaCl. Dilution series of all proteins (concentrations of 10, 7.5, 5, 2.5, 1.25, 0.625 μM) were incubated with heme agarose and unconjugated Sepharose for 30 min. Beads and unbound protein were separated by centrifugation through a filter plate. Non-incubated samples and each flow through, containing unbound protein, were analyzed with BCA Protein Assay Kit. Binding to unconjugated Sepharose was used as control. Binding to heme-agarose is shown as follows: ■: Wt, ●: C [34]S, □: K [3]T, ▼: H(123)S, ⊗: Ovalbumin. ○: Wt-A1M to un-conjugated Sepharose. The mean of two replicates +/− SEM are shown.

Analytical Sciences, Turku, Finland). Statistical analysis was performed using OriginPro 9.0 software (Microcal, Northampton, MA, USA).

2.10. Catalase-like activity assay

Monitoring of the Soret band intensity of heme equimolar samples with A1M mutants (C34S, K(3)T, H123S) after adding hydrogen peroxide was performed in comparison with Wt-A1M and free heme in Tris buffer according to the procedure described earlier [29] with minor modifications. Equimolar (L/P 1.0) heme/A1M samples were prepared by adding 20 μl of heme stock solution in DMSO (2.2 mM) to 1 ml of 45 μM solutions of A1M or each of the A1M mutants. After overnight (~20 h) incubation at room temperature in dark, the samples were evaluated by UV/Vis measurements and diluted by buffer to adjust the Soret band intensity to ~0.6 AU. Heme sample in Tris buffer (with approximately the same absorbance intensity) was freshly prepared and used immediately. These UV/Vis spectra served as initial (zero time) baseline. After a 7 μl aliquot of 50 mM hydrogen peroxide was added to each sample, the time course of spectral changes were measured at 30 s, 1 min, 2 min, 4 min, 6 min, 8 min and 10 min.

2.11. Surface plasmon resonance (SPR)

SPR experiments were conducted on Biacore T200 (GE Healthcare, Piscataway, NJ). Anti-His mouse IgG1 monoclonal antibodies (R&D Systems, Minneapolis, MN) were immobilized on CM5 sensors by amine coupling, ~18,000 response units (RU). The tagged proteins were injected at a flow of 10 μl/min for 360 s. Prior to injection, a freshly prepared heme solution in DMSO was subjected to serial double dilutions using PBS buffer to create a range of eight heme concentrations, from 100 μM down to 0.625 μM. Heme preparations were injected over captured A1M variants for 2 min with a flow 30 μl/min at 25 °C, and the association was recorded, followed by the dissociation monitored during 10 min. Data were analyzed using the Biacore T200 evaluation software (GE Healthcare), subtracting the reference surface and buffer control signals from each curve. Data were globally fitted by simultaneous numerical integration to the association and dissociation parts of the interaction, using the heterogeneous ligand kinetic analysis models (T200 BIAevaluation software, version 1.0; Biacore AB, Uppsala, Sweden).

2.12. Molecular simulation of A1M-heme binding

A representative molecular structure of heme (accession number HEB) was obtained from PDBeChem [30] as a MOL file. Molecular volume was subsequently computed from the MOL file of heme b *via* the online tool from Molinspiration [31]. Briefly, the method for computing

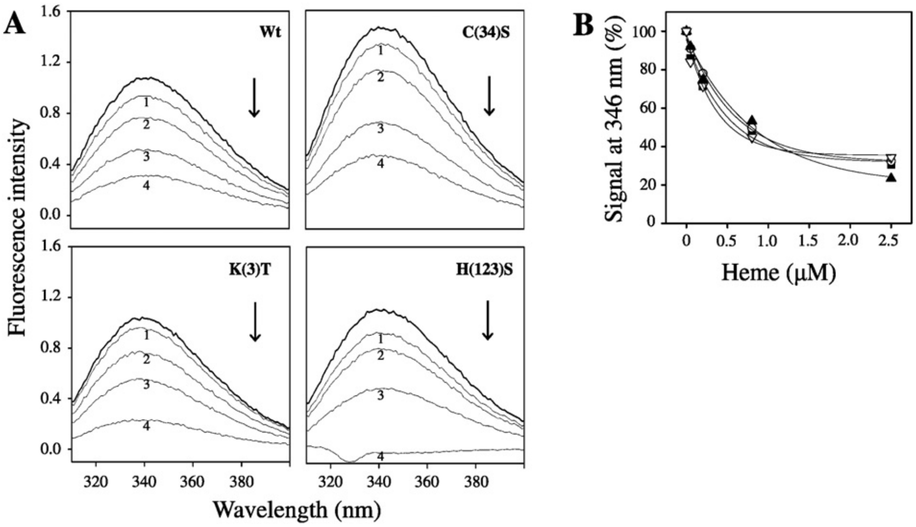


Fig. 3. Heme-induced tryptophan fluorescence quenching in the four A1M-variants. (A) Fluorescence spectra of A1M-variants (1 μM) incubated with heme to a final concentration of; 1 = 0.05, 2 = 0.2, 3 = 0.8, 4 = 2.5 μM. (B) Normalized titration curve of A1M-variants with heme. The fluorescence intensity of each A1M-variant without heme-addition was set to 100%. The data were fitted to lines with the equations $y = 31.9 + 64^{*}e^{-0.0020^{*}x}$ (wt), $y = 31.9 + 66^{*}e^{-0.0017^{*}x}$ (C34S), $y = 20.2 + 76^{*}e^{-0.0012^{*}x}$ (K3T), and $y = 35.5 + 61^{*}e^{-0.0026^{*}x}$ (H123S). n:Wt, ¡: C(34)S,s: K(3)T,▲: H(123)S.

Table 2
Kinetic rate constant and dissociation constants determined by SPR for the heme interactions with A1M variants.

A1M	k_a (M^{-1} s^{-1})	k_d (s^{-1})	K_D (M)[a]
Wt[b]	568.25	7.69×10^{-3}	$13.55 \times 10^{-6} \pm 0.4$
C34S	557.40	6.21×10^{-3}	$11.17 \times 10^{-6} \pm 0.9$
K [3]T	655.55	5.89×10^{-3}	$9.03 \times 10^{-6} \pm 0.9$
H123S	608.15	6.79×10^{-3}	$11.20 \times 10^{-6} \pm 0.8$

[a] All the kinetics parameter were analyzed using the Biaevaluation software, version 4.1.1. The data were fitted using 1:1 L binding model represented by equation (A + B AB).

[b] Previously published in Karnaukhova et al. ref. 43.

the molecular volume is based on group contributions in which the sum of fragment contributions were fitted to actual three-dimensional molecular volume of a training set comprising twelve thousand molecules [32]. The three-dimensional molecular structures of the training set were geometrically optimized using the semi-empirical Austin Model 1 (AM1) method.

In reconstructing the wild-type structure of A1M, residue 34 of the crystal structure (PDB ID: 3QKG) was mutated from serine to cysteine using PyMOL [33]. This was performed whereby the rotamer was independent of the backbone such that the sulfhydryl moiety is pointed towards the imidazole moiety of H123. Molecular volume of the lipocalin pocket in the crystal structure of A1M was computed using CASTp [34].

A query on Protein Data Bank for structures having the lipocalin SCOP fold yielded 280 hits, of which 46 contained a bound heme. From this, 45 are either nitrophorin 1, 2 or 4 from *Rhodnius prolixus* (PDB ID: 1D2U, 1D3S, 1EUO, 1IKE, 1IKJ, 1KOI, 1NP1, 1NP4, 1PEE, 1PM1, 1SXU, 1SXW, 1SXX, 1SXY, 1SY0, 1SY1, 1SY2, 1SY3, 1T68, 1U0X, 1U17, 1U18, 1X8N, 1X8O, 1X8P, 1X8Q, 1YWA, 1YWB, 1YWC, 1YWD, 2A3F, 2ACP, 2AH7, 2AL0, 2ALL, 2AMM, 2ASN, 2AT0, 2EU7, 2GTF, 2HYS, 2NP1, 3C76, 3NP1, 4NP1) while the other was a nitrophorin-like protein from *Arabidopsis thaliana* (PDB ID: 3EMM). The former set constitutes amino acid length of 179–184 while the latter structure had 153 residues. Furthermore, owing to the fact that the former set of 45 structures spanned similar length, it was selected for further analysis. Structural alignment was performed using MultiProt [35].

The structure of wild-type A1M was docked to heme using HADDOCK [36]. A second heme was subsequently docked to a putative heme binding site, which is formed by axial ligands comprising of C34 and H123 as proposed by Meining and Skerra [25] of the top-scoring model using PyMOL. The structure of A1M-heme complex was then refined using the energy-minimization in gas phase followed by molecular dynamics (MD) simulation on GROMACS, version 4.0 [37]. During the refinement, the distance restraints of the side chain of C34 and H123 with the second heme were applied, and inappropriate bond lengths of any atom were fixed. The simulation was performed on the explicit-solvent periodic boundary conditions under the NPT condition at 300 K of temperature using the modified Berendsen thermostat [38] and 1 bar of pressure using Parrinello-Rahman barostat [39]. GROMOS96-53A6 force field was applied for both protein and heme structures and the ionization states of amino acid residues were set according to the standard protocol [40]. The SPC water was used as a solvent model. Bond lengths were constrained using LINCS algorithm that allows for a 2.0 fs-time step [41]. A cut-off distance for the short-range neighbor list was set to 0.9 and 1.4 nm for the electrostatic and van der Waals interactions, respectively. Long-range electrostatic interactions are approximated using PME method [42].

The same MD protocol was also applied for further analyzing dynamic properties of A1M-heme complex. For the data collection, atomic coordinates were recorded every 10 ps.

3. Results

3.1. Expression and characterization of A1M-mutants

The mutated side-group residues are high-lighted in Fig. 1A. Before measuring the heme binding of the mutated A1M-variants, purity and

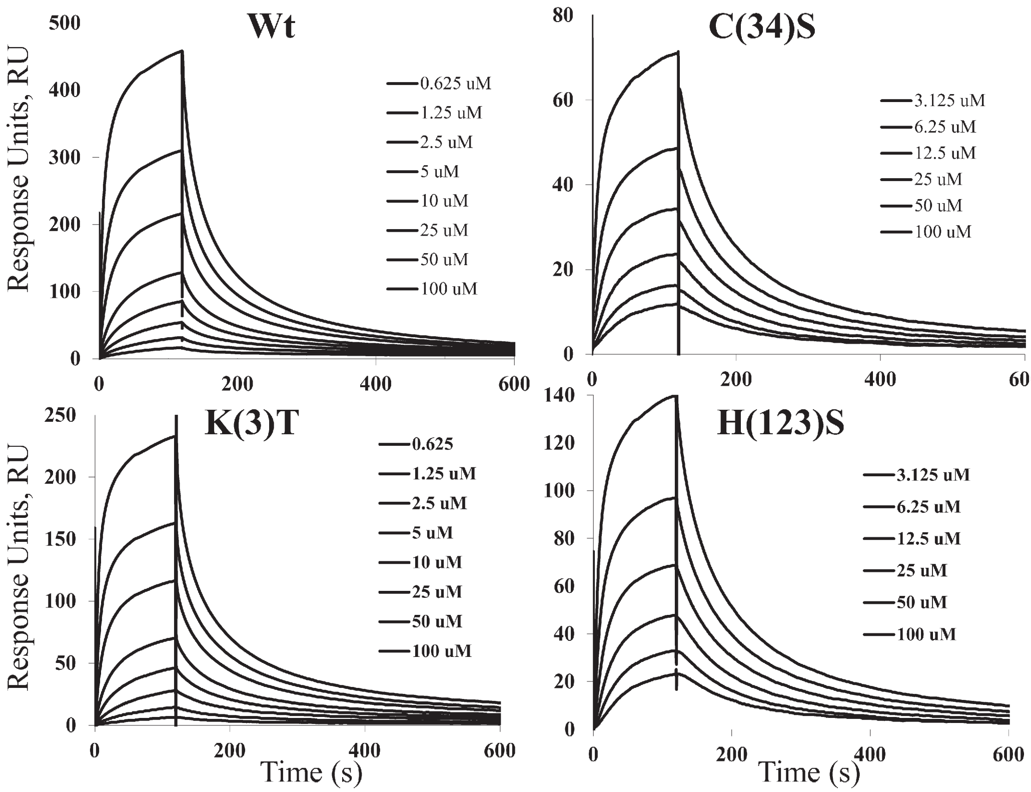


Fig. 5. Binding of heme to A1M analyzed by surface plasmon resonance. Sensorgrams of the heme binding to wt-, C [34]S-, K [3]T- and H(123)S-A1M captured by the anti-His mouse IgG1 monoclonal antibody immobilized on CM5 sensor chip. Increasing signals were obtained using 0.625–100 μM heme.

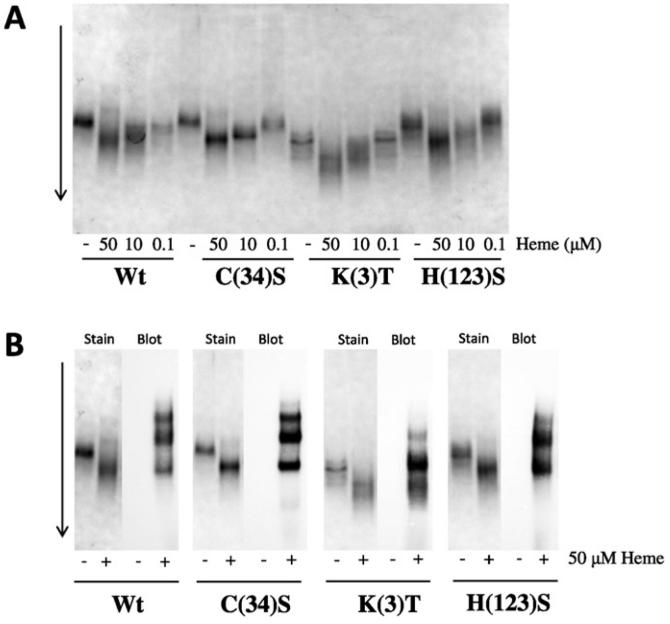

A
- 50 10 0.1 - 50 10 0.1 - 50 10 0.1 - 50 10 0.1 Heme (μM)
Wt
C(34)S
K(3)T
H(123)S
B
Stain Blot Stain Blot Stain Blot Stain Blot
- + - + - + - + - + - + - + - + 50 μM Heme
Wt
C(34)S
K(3)T
H(123)S

Fig. 6. Heme-induced migration shift of A1M-variants. (A) Native PAGE of A1M (5 μM) in 20 mM Tris–HCl, pH 8.0, + 0.15 M NaCl incubated for 1 h with 50, 10 or 0.1 μM heme. Samples were mixed with equal amounts of sample buffer for native PAGE, pH 6.8, and subjected to a 12% Criterion™ TGX™ Precast Gel and stained with Coommassie. (B) Peroxidase blotting of A1M-variants incubated for 4 h without or with heme (50 μM), subjected to native PAGE in a 12% Criterion™ TGX™ Precast Gel and thereafter either stained with Coomassie (5 μg A1M), or transferred to polyvinylidene difluoride (PVDF) membranes (2 μg A1M). The membrane was incubated in Clarity Western ECL Substrate and imaged with a digital imager (BioRad).

basic structural properties were analyzed by SDS-PAGE, far-UV CD spectroscopy and optical fluorescence (Fig. 2). As shown in Fig. 2A, no visible impurities could be detected. Similar apparent sizes were seen, as expected from the theoretical masses of the four variants (Wt-A1M: 22.64 kDa, C(34)S: 22.66, K [3]T: 22.56 and H(123)S: 22.59). Far-UV CD spectra of Wt-A1M and the mutated A1M-forms suggest a similar composition of secondary structure, *i.e.* mostly β-structure (Fig. 2B; Table 1). The CD spectra of the four A1M-variants were also consistent with the X-ray crystallography-derived three-dimensional structure of Wt-A1M [25]. Four tryptophan residues are found in A1M located at various positions (Fig. 1B) and tryptophan fluorescence spectra were recorded as an estimate of the overall conformation of the mutants (Fig. 2C). A higher intensity was obtained from the C [34]S-mutant, possibly as a result of an interaction between the closely located C34- and W36-residues, reducing the intensity of the fluorescence of W36 in Wt-A1M and the other mutants (Fig. 1B). Apart from this, similar spectra were obtained, suggesting no major differences in the overall conformation of the four variants.

3.2. All A1M-variants bind heme

The binding of the heme-group to the A1M-variants was first investigated by quenching of the tryptophan fluorescence during heme-titration, using a protein:heme ratio of 0.4–20 (Fig. 3A,B). The results of this experiment show that heme was bound to all variants with a similar binding strength. Quenching of 50% of the tryptophan fluorescence of 1 μM A1M was achieved in the 0.8 μM range in all cases and approximately 70% of the tryptophan fluorescence was quenched by 2.5 μM heme. Furthermore, the binding of heme to the A1M variants was analyzed using heme-agarose titration (Fig. 4) and surface plasmon resonance (SPR) (Table 2). Both techniques confirm the binding of heme by all four A1M-variants, and indicate a slightly higher binding by Wt-A1M compared to the mutated variants. Heme-agarose titration with increasing amounts of the A1M variants (Fig. 4) yielded binding responses in the order Wt-A1M > C [34]S-A1M > H(123)S-A1M and K [3]T-A1M. Ovalbumin, a negative control protein, did not bind at all to heme-agarose and Wt-A1M did not bind to un-conjugated Sepharose.

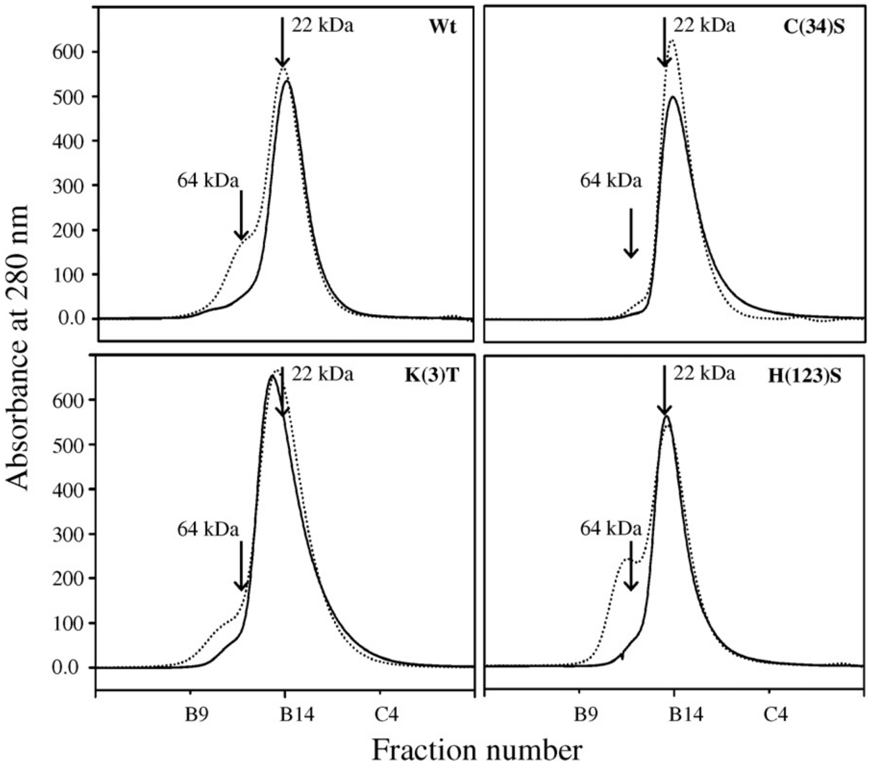


Fig. 7. Size exclusion chromatography of A1M variants incubated with heme. A1M (0.5 ml of a 44 μM-solution) was applied on a Superose 12 FPLC column and eluted with 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl at 0.5 ml/min (unbroken line). A1M and heme in molar ratio of 1:2 were incubated for 1 h and run on Superose 12 FPLC column as above (dotted line). The size calibration on the column was performed with Wt-A1M (22 kDa) and with human hemoglobin (64 kDa).

The SPR data analysis, within heme concentrations 0.625–100 μM (Fig. 5), resulted in dissociation constant (K_D) values of 13.82 × 10^{-6} (Wt-A1M), 11.81 × 10^{-6} (C34S), 9.65 × 10^{-6} [K [3]T] and 11.74 ×10^{-6} (H123S) (Table 2), confirming the results of heme-agarose titration. Although these kinetic data fitted to a 1:1-binding model, SPR data obtained for an extended range of heme concentration (0.625–500 μM) fitted better to a 1:2 binding model, supporting the earlier proposed two binding site model [24]. The latter should be interpreted with care, however, since low heme solubility and its increased aggregation at high concentrations greatly limit the reliability of precise determination of kinetic constants for the low affinity binding sites [43].

3.3. Oligomerization of A1M-heme complexes

The binding of heme was studied by gel-shift assay, using native PAGE, to analyze the electrophoretic mobility of the A1M-variants alone or in the presence of 0.1, 10 or 50 μM heme (Fig. 6A). A clear migration shift was seen of all four variants and the dose-dependence was similar. These results support the findings described above, *i.e.* binding of heme with similar strength by all A1M-variants. A similar migration shift was seen with Wt-A1M without the N-terminal His-tag, and no migration shift of the control proteins α_1-acid glycoprotein and ovalbumin after heme-incubation, or of Wt-A1M in the presence of the carrier DMSO, could be seen (data not shown). To visualize the heme-group in the A1M-bands, the gels were analyzed by peroxidase-activity (ECL)-blotting (Fig. 6B). All A1M-variants displayed heme-induced peroxidase-activity after incubation with heme. Three bands were seen with all variants, probably corresponding to the monomeric, dimeric and trimeric A1M-heme complexes previously reported [24]. Interestingly, the peroxidase activity was much stronger in the dimeric and trimeric bands, when relating to the protein staining activity (Fig. 6A *vs* B).

The sizes of the A1M-heme complexes were also investigated by Superose 12 gel-filtration (Fig. 7A). Incubation for 1 h with heme (A1M:heme = 1:2) resulted in the appearance of larger forms besides the monomeric peak, suggesting an increased oligomerization of A1M in the presence of heme. This supports the results of the PAGE shown in Fig. 6. The high molecular weight-forms were most pronounced in Wt-A1M and H(123)S-A1M, and less so in K [3]T- and C [34]S-A1M. Furthermore, a slight heme-induced shift of the monomeric peak towards higher molecular mass was seen in Wt-A1M.

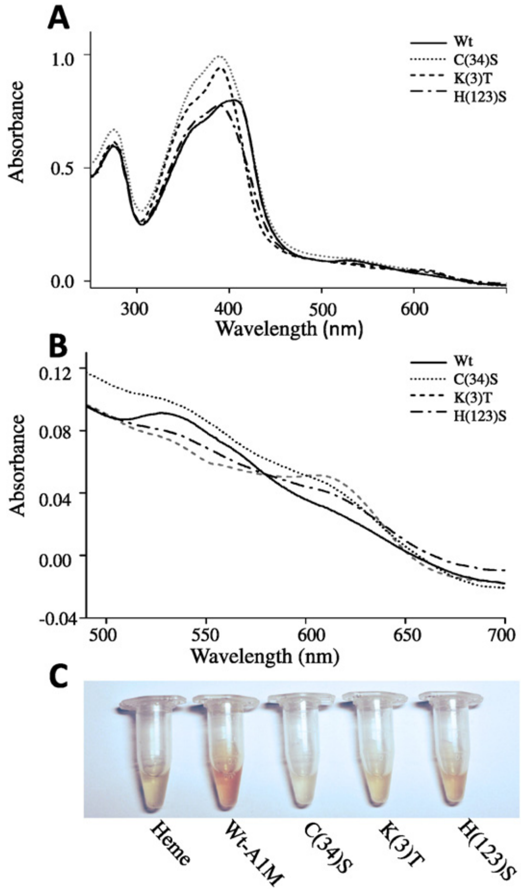


Fig. 8. Absorbance spectrophotometry of A1M variants incubated with heme. (A) Absorbance spectrum of A1M variants (10 μM) in the presence of 20 μM heme in 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl after 24 h of incubation. A similar absorbance spectrum of Wt-A1M was previously published in Karnaukhova et al. [43]. (B). Close-up of the absorbance spectrum between 500 and 700 nm of A1M variants (10 μM) in the presence of 20 μM heme in 20 mM Tris–HCl, pH 8.0, 0.15 M NaCl after 24 h of incubation. (C) The A1M variants incubated with heme and heme only were also analyzed visually. Protein:heme ratio was 1:2 and the heme concentration 20 μM.

3.4. UV–Vis absorbance of A1M-heme is dependent on C34, K92, 118, 1(32) and H132-residues

The heme binding was followed by UV–Vis absorbance spectrophotometry (Fig. 8). The heme binding could be confirmed, but the time-dependence of the binding was different for the various A1M-forms. A broad peak with a maximum (λ_{max}) around 400 nm was seen immediately after mixing of 10 μM A1M + 20 μM heme for all variants (not shown). After 24 h, however, a red-shift of λ_{max} towards a higher wavelength was seen for Wt-A1M, but not for any of the mutants (Fig. 8A; peak values of Soret-band shown in Table 3). This red-shift was beginning after a few minutes and could be recorded at 30 min and onwards (not shown). Furthermore, zooming in on 500–700 nm wavelengths (Fig. 8B), the Wt-A1M-heme complex displayed a maximum at 540 nm, which was less pronounced in the mutants. The mutants also displayed an absorbance shoulder at 610 nm, this was most pronounced in the K [3]T-A1M-heme and H(123)S-heme complexes. The spectral differences could also be seen as a striking difference in color of the heme-complexed A1M-variants (Fig. 8C). At these concentrations (10 μM of both protein and heme), the Wt-A1M heme solution was red whereas the C [34]S-A1M heme complex was yellow and the K [3]T-A1M and H(123)S-A1M showed a similar yellow-brown color as free heme. The red-shifted Soret-band and the 540 nm-maximum are seen in ferrous (FeII) heme binding proteins [44]. These results therefore suggest that A1M undergoes reducing reactions with the heme-group, involving the C34-residue and regulated by H123 and K92, 118 and 130.

To minimize the contribution of unbound heme-groups to the spectra, the monomer peak fractions of the gel filtrated A1M-heme complexes were also analyzed by UV–Vis absorbance spectrometry. The peak wavelengths of the Soret-band (λ_{max}) before and after gel filtration are shown in Table 3. The red-shift of the Wt-A1M heme

Table 3
Summary of the UV/Vis absorbance spectra and induced CD for heme complexes with A1M variants at low (heme:A1M ratio 0.1) and high (heme:A1M ratio 1.0) relative heme content.

	Heme, μM	Heme:A1M ratio	Absorbance Soretλmax, nm		Induced CD, nm
				Monomer[a] gelfiltration	
Wt	4.5	0.1	422		421
	45	1.0	414	414	403
C (3)S	4.5	0.1	397		397
	45	1.0	394	392	400, 408
K (3)T	4.5	0.1	421		417
	45	1.0	399	399	414
H(123)S	4.5	0.1	421		416
	45	1.0	390	391	393, 405

[a] The heme:A1M ratio was 2.0 before application to gel filtration column.

complex was also apparent after gel filtration, whereas the monomer fractions of three mutants showed a λ_{max} below 400 nm. This suggests that the bound heme group is reduced by Wt-A1M but not the mutated forms.

3.5. Catalase-like activity

The catalase-like activity of A1M-heme complexes was measured by monitoring the Soret band after addition of H_2O_2 (Fig. 9). As evident from the figure, a significant shielding of the heme molecule was seen by all forms of A1M, as compared to heme alone. This further supports a binding of the heme group to A1M. However, plotting the peak value of each A1M-form as a function of time (Fig. 9D) demonstrates a less efficient shielding of the heme-molecule in C34S-A1M as compared to Wt-A1M and the other two mutants. This suggests that the C34 residue is essential for the heme coordination and/or redox activities of the bound heme-groups.

3.6. Induced visible CD support binding of two heme-groups

Neither heme nor A1M alone exhibits any CD activity in the visible range (not shown). However, when a small aliquot of heme solution is added to A1M, a CD activity in the heme absorbance region (390–415 nm) was induced (so called Cotton Effect) for each A1M variant (Fig. 10). These results support binding of heme to all A1M-variants. Titrations of heme were also consistent with the presence of two binding sites, as reported previously [24]. At low heme concentrations (see bottom traces in Fig. 9, heme:A1M molar ratio 0.1), the induced CD proceeded to an equilibrium state relatively fast (<40 min), suggesting binding to the primary binding site. As summarized in Table 3, the peak wavelengths of the induced CD for Wt-A1M was 421 nm, and 417 nm, 397 nm and 416 nm for K [3]T-A1M, C34S-A1M and H123S-A1M, respectively, suggesting a different microenvironment of the heme molecule in the first binding site of each mutant. At higher heme concentrations (heme:A1M molar ratio 1.0; see upper traces in Fig. 10) the reaction was slower (>5 h to obtain equilibrium), consistent with binding at the second binding site at higher concentrations. The peak wavelengths (Table 3) of the induced CD at heme:A1M ratio 1.0 were different from those shown for heme:A1M ratio 0.1, suggesting different heme environments and/or coordination at the primary and secondary binding sites.

3.7. Molecular simulation of A1M-heme binding

The lipocalin pocket, as identified by CASTp, essentially encompasses the inner cavity of the protein with a molecular volume of 2033 Å^3. It was found that the inner lining of the pocket was composed of 41 residues. Apparently the molecular volume of heme, which is 538.9 Å^3, could readily fit inside the lipocalin pocket (Fig. 11).

Before proceeding with the docking of heme to A1M, it is pertinent to explore the binding modality of other members of the lipocalin family that are known to bind heme. A total of 45 nitrophorin 1, 2 or 4 structures from *Rhodnius prolixus* were obtained from the PDB. These members of the lipocalin family had amino acid length in the range of 179–184 and their superimposition performed using MultiProt indicated high structural homology affording an RMSD value of 1.91 Å (Fig. 12). Analysis of these structures revealed that the bound heme was coordinated to axial H59 and ammonia ligands in nearly all cases.

Heme was docked to A1M using default parameters of HADDOCK without explicit definition of active and passive residues as they were

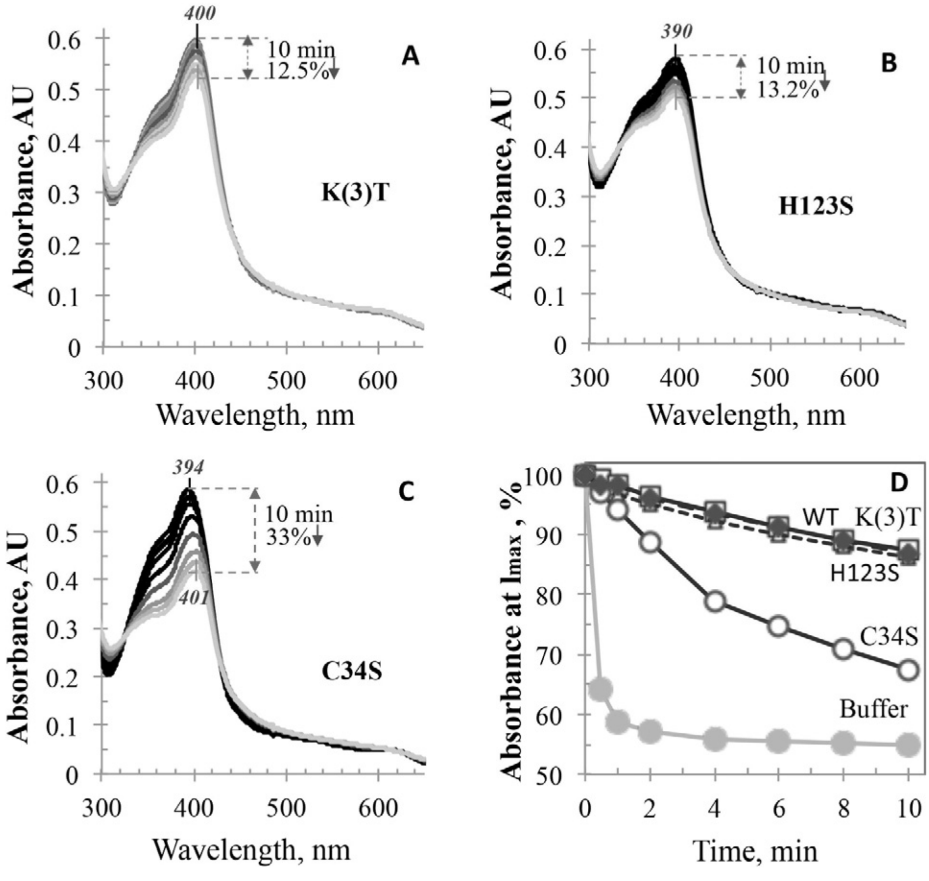


Fig. 9. Time-course UV/Vis absorbance data for catalase-like activity of the heme complexes with A1M mutants: (A) K [3]T; (B) H123S, and (C) C34S. The upper trace of each plot shows initial absorbance spectrum of each heme/mutant sample (1:1) as recorded over 10 min with time points taken at 30 s, 1 min, 2 min, 4 min, 6 min, 8 min, and 10 min after adding a 7 μl aliquot of 50 mM H_2O_2 stock solution to 1 ml of the protein sample. Plot (D) shows the percentage of the remaining intensity of the Soret band of each mutant in comparison with 0 s.

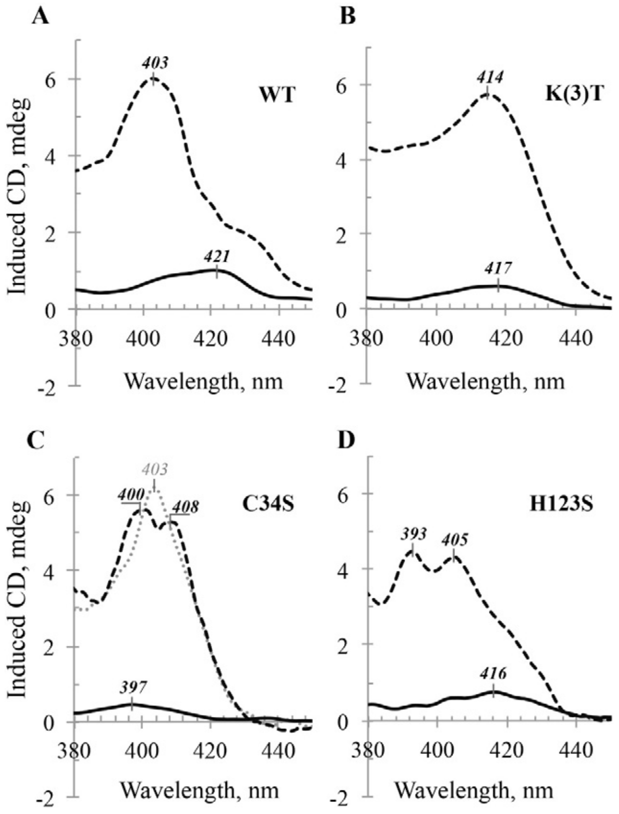


Fig. 10. Induced CD data for heme complexes with A1M. Wt-A1M (A), and A1M mutants K [3]T (B), C34S (C), and H123S (D). Induced CD spectra are shown for heme:A1M ratio 0.1 (solid traces in A-D, the spectra were recorded at 1 h after adding heme, when no additional spectral changes were observed) and heme:A1M ratio 1.0 (broken traces, measured 20 h after adding heme, when no additional changes were observed). The broken trace in 9A was previously published in Karnaukhova et al. [43]. Gray dotted trace shown for C34S plot (C) corresponds to an intermediate state observed for this C34S mutant at L/P 1.0 at the time point 1.5 h after adding heme. Protein concentration in all samples was 45 μM.

assigned automatically. Active residues correspond to residues that are directly involved in the interaction whereas the passive residues denote residues that are in the vicinity of active residues. A typical docking simulation is comprised of the following steps: (i) rigid-body docking, (ii) scoring and filtering, (iii) semi-flexible refinement, (iv) water refinement, (v) scoring and analysis and finally (vi) clustering of docked structures. Docking results indicated that there were 193 structures in 6 clusters and that the top-performing cluster was the most populated with 142 structures affording a HADDOCK score of -86.7 ± 7.9 and that the RMSD from the overall lowest-energy structure was 0.5 ± 0.3 indicative of the close similarity of the docking conformation among members of this cluster. Thus, the lowest-energy structure of the A1M-heme complex was selected for further comparison with a representative structure from the aforementioned set of 45 structures from *R. prolixus* (PDB ID: 1X8P) obtained at an ultrahigh resolution of 0.85 Å. Superimposition of the two structures yielded an RMSD of 12.37 Å spanning the entire length of 164 residues in A1M (Fig. 13).

It can be seen that the bound heme in A1M leaves the putative outer heme binding site available. To investigate whether this site could accommodate another heme, a subsequent docking procedure was performed using PyMOL such that the iron atom of the second heme was coordinated to the Sγ of C34 and Nε of H123. After structural refinement using energy minimization and MD simulation, the final model was compared with the crystal structure of wild-type A1M (PDB ID: 3QKG; resolution of 2.3 Å) [25]. A similar fold between the 3QKG structure and our heme-bound A1M model was observed with an RMSD of 1.29 Å (Fig. 14). The first heme binding site comprises residues from several strands of the β-barrel while the second heme binding site is formed by a few residues from the short α-helix at the open end and H123, which lies on loop 4, just opposite to the helix. Analyses of Ramachandran plot revealed that only 1.4% of A1M residues were located in the disallowed region (data not shown) thereby indicating that appropriate stereochemical quality of the heme-bound A1M model was achieved.

To investigate fluctuations of the A1M structure upon heme binding and interactions of hemes with their pockets, MD simulations were performed for 30 ns on both heme-bound and heme-free A1M structures using explicit-solvent periodic boundary conditions. Stability of the protein and heme structures over the course of the simulation was observed by measuring the time evolution of root mean square deviation (RMSD) with respect to their initial structures (Fig. 15A). It can be seen that the simulated structures of A1M reached equilibrium prior to $t = 5$ ns and remained stable throughout the simulation for both heme-bound and heme-free A1M models. RMSDs of the heme-bound A1M structure fluctuated around 0.2 nm whereas the heme-free A1M was slightly larger by ≈0.1 nm, which implies that heme molecules were pertinent in stabilizing the A1M structure. The RMSD as a function of time for hemes in the A1M-heme simulation was also investigated. It was found that the second heme fluctuated to a much greater extent than the first heme in which RMSDs were 0.412 ± 0.010 nm for the former and 0.141 ± 0.003 nm for the latter (Fig.15B). Such differences may suggest either large flexibility of the binding pocket for the second heme or a reflection of the different binding strengths afforded by the two hemes.

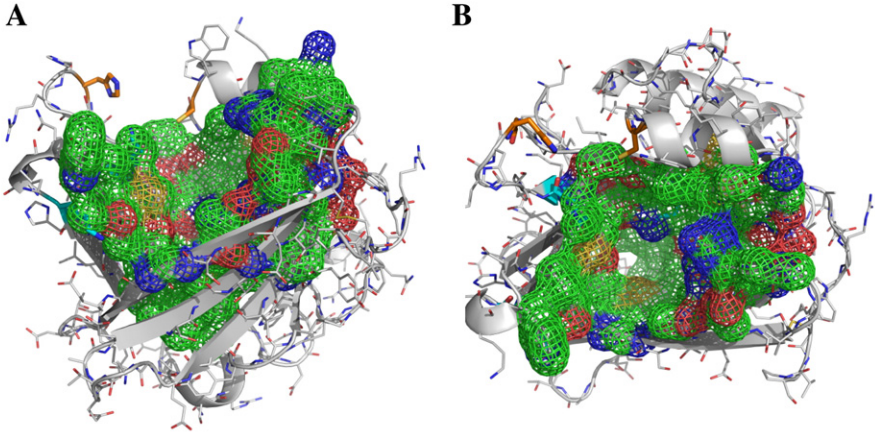


Fig. 11. Identified pocket in the crystal structure of A1M. Residues lining the pocket are represented as meshes and color coded by their constituent carbon, nitrogen and oxygen atoms as green, blue and red, respectively. Residues C34 and H123 are shown in orange whereas residues K92, K118 and K130 in cyan. Secondary structure and remaining side chain carbon atoms are displayed in white. Structures are shown from the side (A) and top (B) views.

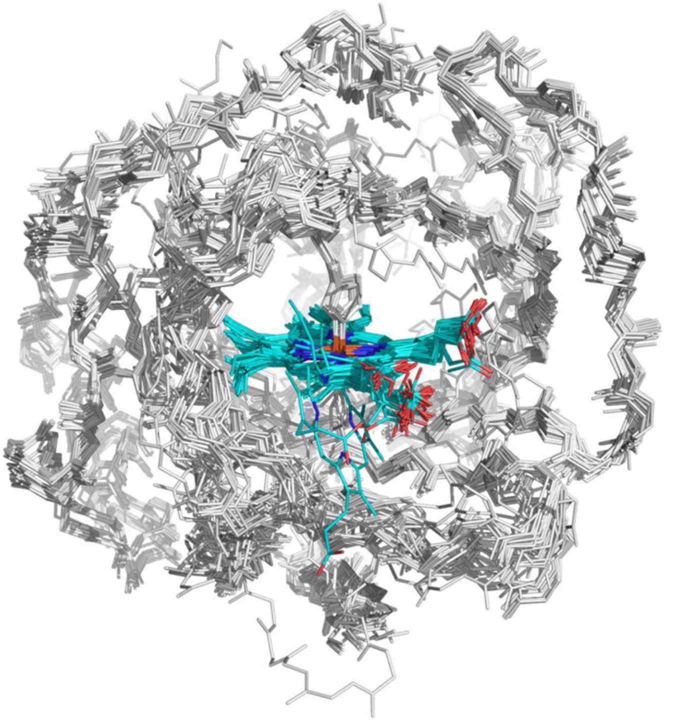


Fig. 12. Structural superimposition of 45 heme-nitrophorin complexes from *R. proxilus*. Carbon backbone are color coded as cyan color for heme moiety and white color for the protein structure. Nitrogen and oxygen atoms are color coded as blue and red, respectively, while the iron atom of the heme moiety is shown in orange color.

To assess the conformational flexibility of A1M with respect to individual structural regions of the protein, root mean square fluctuation (RMSF) of the backbone as a function of the residue number was measured in both simulations from $t = 5$ to 30 ns (Fig. 15C). It can be seen that in the free A1M simulation, the four loops connecting neighboring β-strands (designated loop-1 to −4) at the open end of the eight-stranded β-barrel exhibited considerably high flexibility. This result is consistent with the experimental data in that those regions displayed high crystallographic B-factors, except for loop-4 of the 3QKG structure, which is responsible for the metal-binding site as well as exhibiting low temperature factor [25]. As expected, presence of the second heme significantly decreased fluctuation of the binding site formed by the short α-helix of loop-1 (C34 to M40) and residues on loop-4 (S120 to G124). On the other hand, binding of the first heme did not alter the fluctuation of the lipocalin pocket in which a similar RMSF profile was observed in such region.

To investigate contributing factors that dominate the interaction of the A1M-heme complex, potential energies for interactions between each heme molecule and the protein were calculated as the sum of the electrostatic and van der Waals (vdW) interactions from $t = 5$ to 30 ns (Fig. 15D). It can be seen that the first heme exhibited considerably large negative interaction energies with the protein, which is in contrast to the second heme that exhibited lower interactions. It should be noted that electrostatic energies were found to be major factors stabilizing the interaction between the first heme and A1M while vdW interactions contributed more dominantly in the second heme-bound A1M. These results suggest that the first heme binds the lipocalin pocket of A1M with more strength than the second heme, which may explain the higher degree of flexibility of the heme molecule when bound to the second binding pocket of A1M. The aforementioned evidences also suggest that binding of the first heme to the lipocalin pocket stabilizes the protein structure and predisposes A1M for binding of the second heme to the surface-exposed binding site.

4. Discussion

4.1. Side-groups of C34, H123, K92, K118 and K130 are involved in heme-binding

The investigation in this paper is based on site-directed mutagenesis, *i.e.* biochemical analysis of wild-type A1M and three mutated forms of the protein. In order to ascertain that functional differences were due to the amino acid substitutions rather than impurities or effects of incorrect folding, we first investigated biochemical properties of the recombinant A1M-species. The proteins appeared highly purified, and far-UV CD spectra of Wt-A1M and the mutated A1M-forms suggested a similar composition of secondary structure, *i.e.* mostly β-structure (Fig. 2B; Table 1), consistent with the X-ray crystallography-derived three-dimensional structure of Wt-A1M [25]. Tryptophan fluorescence spectra were also recorded as an estimate of the overall conformation of the mutants (Fig. 2C). The C [34]S-mutant displayed a higher fluorescence intensity than Wt-A1M and the other mutants. Sulfur-containing molecules are known fluorescence quenchers [45]. The C34-residue and one of the four tryptophan residues, W36, are located closely together on the small helix on the rim of the lipocalin pocket (Fig. 1), hence there is a possibility of an interaction between these two residues reducing the intensity of the fluorescence of W36 in Wt-A1M and the other mutants. Apart from this, similar spectra were obtained, suggesting no major differences in the overall conformation of the four variants.

It was previously shown that heme binds specifically to A1M and that this feature is evolutionally conserved [21–23]. Furthermore, it was shown that A1M forms a trimeric complex with heme in a 1:2 stoichiometry [24]. In order to understand the mechanism of heme binding of A1M, the binding was studied by several different techniques, and to understand the structural requirements of the binding and possibly localize binding sites, the binding to mutated variants was also studied. In short, all methods showed that heme was bound to all variants and

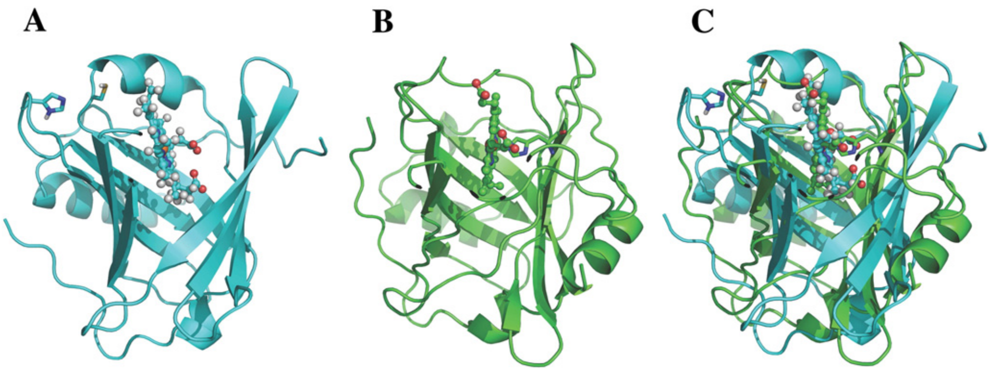


Fig. 13. Structures of heme complexes with A1M (A), nitrophorin 4 from *R. prolixus* (B) and superimposition of both A1M and nitrophorin 4-heme complexes (C). Carbon atoms are shown in cyan and green colors in panel A and B, respectively, while oxygen and hydrogen atoms are shown in red and white, respectively.

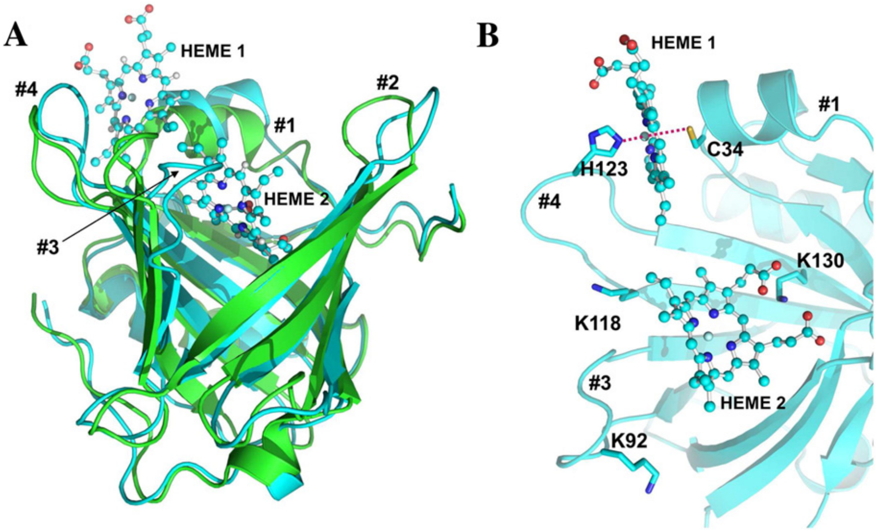


Fig. 14. Modeling of binding of two heme-molecules to A1M. Superimposition between the structure of A1M complexed with 2 heme molecules and the 3QKG structure of apo-A1M (A), and a close view of two different heme binding sites of A1M (B), heme molecules are represented in a ball and stick model, coordinated residues are shown as sticks and magenta lines indicate the coordination. The four loops are labeled #1–#4.

indicated a slightly higher binding by Wt-A1M compared to the mutated variants, whereas heme titration during CD-spectral analysis, UV–Vis absorbance spectrophotometry and the catalase-like activity assays revealed differences between the four A1M-variants suggesting involvement of the C34-, H123- and K(92,118,130)-side-groups.

The experimental results in this investigation were obtained using A1M of all four variants carrying an N-terminal His$_8$-tag. We used the migration shift-PAGE assay to show that non-tagged Wt-A1M binds the heme-group with similar binding strength [43]. Moreover, the results from SPR-analysis were obtained by studying A1M-molecules immobilized to the surface by binding *via* the His-tag. It was also shown previously that the His-tag does not contribute to, or interfere with, the reductase activities of A1M [19,20] and we therefore expect that any reductase activity of the A1M-variants in this investigation is not affected by the His-tag. Although this does not rule out that the His-tag may influence the heme binding, we therefore believe that the major conclusions of the work also are valid for non-tagged A1M.

4.2. Characterization of two heme-binding sites in A1M

Several different results suggest that each A1M molecule can bind two heme-groups simultaneously. A previous report used heme-titration, gel chromatography, and resonance Raman and EPR spectroscopy to demonstrate the formation of A1M:heme complexes with the stoichiometry 1:2 [24]. In the present work, the differences in the heme-induced CD-spectra at low and high heme:A1M ratios are consistent with a subsequent filling of two binding-sites. Based on the spectral behavior of the four A1M-variants, we propose that one heme binding site is located in the lipocalin pocket, the other is located between loops 1 and 4 at the outer rim of the pocket, and that the sites are filled in that order by increasing heme concentrations. The existence of an inner binding site is supported experimentally by the almost complete quenching of the tryptophan fluorescence in spite of the fact that one of the tryptophan residues (W25) is located at the bottom of the pocket, and the dependence of the heme-induced visual CD spectra on the K(92, 118, 130) residues which line the interior pocket wall. The outer binding site is supported by the influence of the C34- and H123-residues on heme-titration effects, the less efficient shielding of the heme-group from H_2O_2-induced degradation in the C34S-mutant, and was also proposed by Meining and Skerra [25] based on its similarities with a group of heme binding proteins with a Cys-Pro dipeptide motif [46].

Our *in silico* modeling supports the possibility of simultaneous binding of two heme-groups to the proposed inner and outer binding site. The inner binding site is analogous to the heme binding sites of nitrophorins, heme binding members of the Lipocalin protein family. Dynamic analysis of the models also suggests a stronger binding of heme to the inner binding site, and/or a higher flexibility of the second binding site. This is consistent with a primary binding to the inner site and a secondary binding to the outer binding site. Interestingly, the inner binding site suggests close proximity between K118 and K130 side-groups and the non-pyrrolic groups of the heme-molecule (Fig. 14B), supporting the differences in heme-induced visual CD-spectra between Wt- and K(92,118,130)T-A1M. It has previously been shown that these three side-groups become covalently modified and cross-linked *in vivo* by yellow-brown, unidentified, size heterogeneous compounds with molecular masses between 122 and 282 amu [47]. Based on this, it can be speculated that a reaction between the inner heme group and the A1M protein leads to degradation of the heme-group, yielding covalent attachment of degradation products to the lysyl side-chains. This hypothetical reaction may involve electron-transfer reactions of the iron atoms of both heme-groups as well as the thiol group of C34.

4.3. Possible electron transfer reaction between A1M and heme

The UV–Vis absorbance spectrum of Wt-A1M displayed the characteristic features of hemoglobin in its reduced form, *i.e.* a red-shifted Soret band and a peak at 540 nm, in contrast to the three mutated A1M forms (Fig. 9). A negative reduction potential of the C34 thiol group in combination with the three lysyl residues K92,118 and 130 was previously shown, *i.e.* A1M reduced methemoglobin, cytochrome c and free iron [19]. It may therefore be speculated that Wt-A1M keeps the outer heme-group in its reduced, ferrous (Fe2+) form by a tentative reaction that involves coordination of the iron atom by the C34- and H123-residues, where the C34 thiol group also may serve as an electron-source. As proposed earlier, the K(92,118,130)-residues may regulate the electronegativity of the thiol group by creating a positive electrostatic microenvironment [7,19].

4.4. Physiological function of heme-binding to A1M

Several potential physiological functions of the heme binding by A1M are possible. The first, and perhaps most obvious, may be

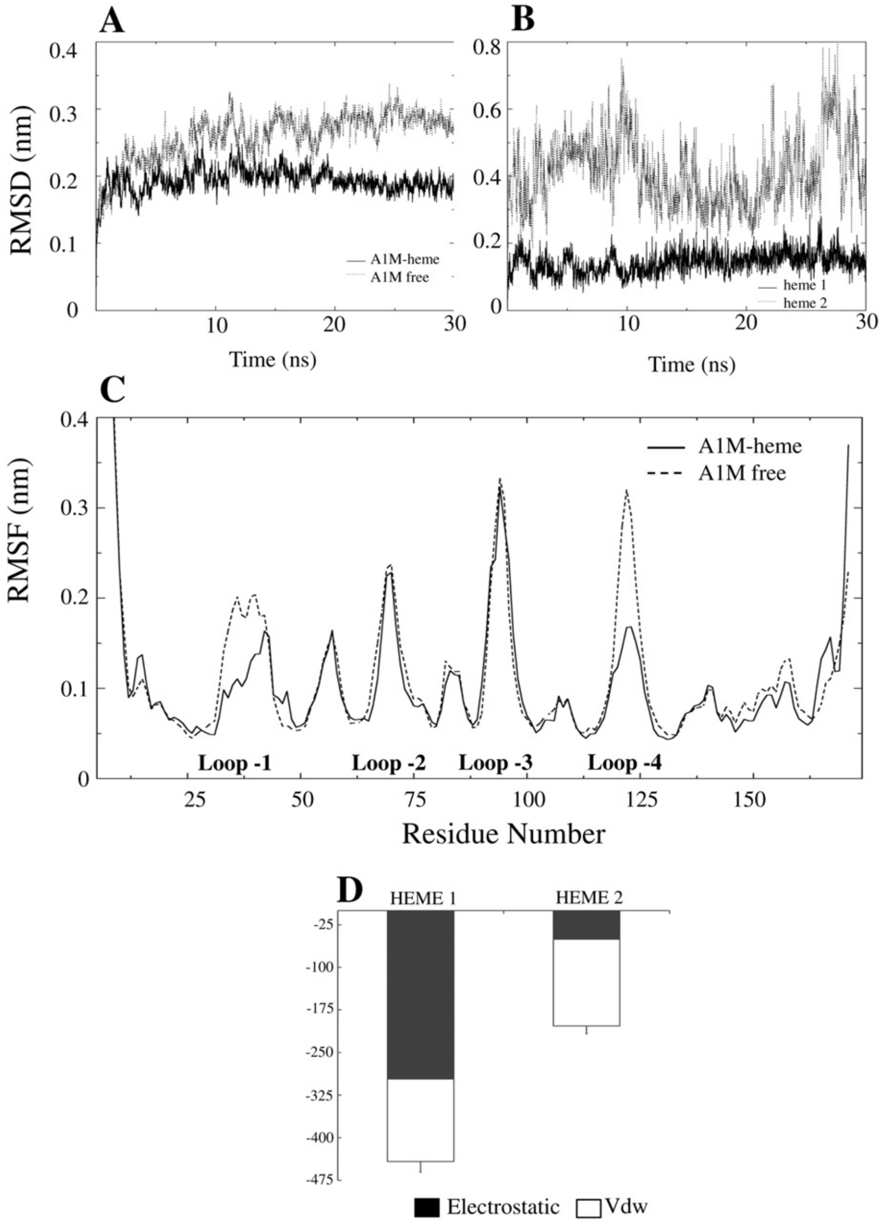


Fig. 15. MD simulation data of A1M-heme complexes. (A) Time evolution of root mean square deviation (RMSD) with respect to the initial structure for the protein backbone atoms and (B) the heavy atoms of heme molecules. (C) Backbone root mean square fluctuation (RMSF) for each amino acid residue from MD simulations of heme-bound and heme-free structures of A1M. (D) Potential energies for interactions between A1M and each heme molecule. The electrostatic and vdW contributions are shown in black and white, respectively. Error bars represent the standard deviation in the sum of the interactions.

scavenging of free heme-groups to protect biomolecules from heme-induced toxic reactions. This function is supported by previous reports that A1M-binding of heme results in inhibition of cell-lysis, cell-cycle arrest and molecular damage otherwise induced by the free heme [13,48]. Second, heme binding can be the first step in a series of reactions leading to *heme-degradation*. This also involves proteolytic cleavage of the C-terminal tetrapeptide Leu-Ile-Pro-Arg (pos 179–183) of A1M, as suggested in the report by Allhorn et al. [21]. However, nothing is known about the detailed molecular mechanisms leading to the heme-degradation. The position of the C-terminus beyond amino acid residue Gly172 could not be determined in the published crystal structure of A1M [25] and it is therefore difficult to speculate on the mechanistic influence of the C-terminal tetrapeptide on binding and degradation of the heme-group. Thirdly, the heme-group, with the iron atom, may be employed as an *enzymatic cofactor* during the redox activities of A1M. Heme-groups are commonly employed as electron-active cofactors in peroxidases, reductases, dehydrogenases, *etc*., and it is therefore reasonable to imagine a role of the bound heme-groups in A1M in reduction, radical scavenging, or other antioxidative activities. This remains to be tested experimentally, however.

The concentration of A1M in blood plasma is 1–2 μM. Among heme binders in plasma, A1M is apparently not the strongest. Hemopexin has a much higher affinity. Albumin, which binds heme with a slightly higher affinity than A1M is much more abundant in blood. This suggests that the role of A1M is not primarily as a heme-scavenger in blood. A1M is secreted to blood from the liver, but rapidly transferred to the extravascular space ($T1/2 = 2.5$ min). The protein is also synthesized by most other epithelial cells and found both intracellularly and in the extracellular matrix, for example in skin and placenta. We therefore propose that the heme-reaction mechanisms of A1M are different from those of hemopexin and albumin, and of physiological importance in cells and extravascular fluids rather than in blood.

5. Conclusions

Several previous investigations have shown that the lipocalin A1M can bind to heme groups, and that this property constitutes part of its physiological antioxidative mechanisms [13,21,23]. In this work we have investigated the structural requirements of the heme binding. The main conclusions of the present report are that two heme-groups can be accommodated simultaneously in the protein, and that the binding and reactions with the heme-groups are affected by the Cys34, Lys92, 118, 130, and His123 residues lining the lipocalin pocket.

Conflicts of interest

Bo Åkerström is a founder and share-holder of A1M Pharma AB which holds patent rights on medical uses of A1M based on its heme-binding properties.

Acknowledgments

SR, LWR and BÅ are indebted to Maria E Johansson for invaluable technical assistance. This work was supported by the Swedish Medical Research Council (VR), governmental ALF research grants to Lund University and Lund University Hospital, the Royal Physiographic Society in Lund, the Foundations of Greta and Johan Kock, Alfred Österlund, the Blood and Defense Network, Lund University, the Crafoord Foundation (20081029) and A1M Pharma AB. CN is supported by a Research Career Development Grant (No. RSA5780031) from the Thailand Research Fund. AIA acknowledges the support from the National Institutes of Health (NIH) under grants HL110900 and the U.S. Food and Drug Administration (MODSCI Grants).

References

[1] D.R. Flower, The lipocalin protein family: structure and function, Biochem. J. 318 (Pt 1) (1996) 1–14.

[2] B. Åkerström, D.R. Flower, J.P. Salier, Lipocalins: unity in diversity, Biochim. Biophys. Acta 1482 (2000) 1–8.

[3] B. Åkerström, N. Borregaard, D.R. Flower, J.P. Salier (Eds.), *Lipocalins: An Introduction*, Landers Bioscience, Georgetown, Texas, 2006.

[4] M.D. Ganfornina, D. Sanchez, L.H. Greene, D.R. Flower (Eds.), *The Lipocalin Protein Family: Protein Sequence, Structure and Relationship to the Calycin Superfamily*, Landers Bioscience, Georgetown, Texas, 2006.

[5] B. Ekström, I. Berggård, Human α1-microglobulin. Purification procedure, chemical and physiochemical properties, J. Biol. Chem. 252 (1977) 8048–8057.

[6] B. Åkerström, L. Lögdberg, An intriguing member of the lipocalin protein family: a_1-microglobulin, Trends Biochem. Sci. 15 (1990) 240–243.

[7] B. Åkerström, M. Gram, A1M, an extravascular tissue cleaning and housekeeping protein, Free Radic. Biol. Med. 74 (2014) 274–282.

[8] J.F. Kaumeyer, J.O. Polazzi, M.P. Kotick, The mRNA for a proteinase inhibitor related to the HI-30 domain of inter-alpha-trypsin inhibitor also encodes a_1-microglobulin (protein HC), Nucleic Acids Res. 14 (1986) 7839–7850.

[9] A. Lindqvist, T. Bratt, M. Altieri, W. Kastern, B. Åkerström, Rat alpha 1-microglobulin: co-expression in liver with the light chain of inter-alpha-trypsin inhibitor, Biochim. Biophys. Acta 1130 (1992) 63–67.

[10] T. Bratt, H. Olsson, E.M. Sjöberg, B. Jergil, B. Åkerström, Cleavage of the alpha 1-microglobulin-bikunin precursor is localized to the Golgi apparatus of rat liver cells, Biochim. Biophys. Acta 1157 (1993) 147–154.

[11] A. Grubb, E. Mendez, J.L. Fernandez-Luna, C. Lopez, E. Mihaesco, J.P. Vaerman, The molecular organization of the protein HC-IgA complex (HC-IgA), J. Biol. Chem. 261 (1986) 14313–14320.

[12] T. Berggård, N. Thelin, C. Falkenberg, J.J. Enghild, B. Åkerström, Prothrombin, albumin and immunoglobulin A form covalent complexes with alpha1-microglobulin in human plasma, Eur. J. Biochem. 245 (1997) 676–683.

[13] M.G. Olsson, T. Olofsson, H. Tapper, B. Åkerström, The lipocalin alpha1-microglobulin protects erythroid K562 cells against oxidative damage induced by heme and reactive oxygen species, Free Radic. Res. 42 (2008) 725–736.

[14] M.G. Olsson, M. Allhorn, J. Larsson, M. Cederlund, K. Lundqvist, A. Schmidtchen, O.E. Sörensen, M. Mörgelin, B. Åkerström, Up-regulation of A1M/a_1-microglobulin in skin by heme and reactive oxygen species gives protection from oxidative damage, PLoS One 6 (2011), e27505.

[15] K. May, L. Rosenlöf, M.G. Olsson, M. Centlow, M. Mörgelin, I. Larsson, M. Cederlund, S. Rutardottir, W. Siegmund, H. Schneider, B. Åkerström, S.R. Hansson, Perfusion of human placenta with hemoglobin introduces preeclampsia-like injuries that are prevented by alpha1-microglobulin, Placenta 32 (2011) 323–332.

[16] L. Wester-Rosenlöf, V. Casslen, J. Axelsson, A. Edström-Hägerwall, M. Gram, M. Holmqvist, M.E. Johansson, I. Larsson, D. Ley, K. Marsal, M. Mörgelin, B. Rippe, S. Rutardottir, B. Shohani, B. Åkerström, S.R. Hansson, A1M/alpha1-microglobulin protects from heme-induced placental and renal damage in a pregnant sheep model of preeclampsia, PLoS One 9 (2014), e86353.

[17] K. Sverrisson, J. Axelsson, A. Rippe, M. Gram, B. Åkerström, S.R. Hansson, B. Rippe, Extracellular fetal hemoglobin induces increases in glomerular permeability: inhibition with alpha1-microglobulin and tempol, Am. J. Physiol. Ren. Physiol. 306 (2014) F442–F448.

[18] S. Rutardottir, E.J. Nilsson, J. Pallon, M. Gram, B. Åkerström, The cysteine 34 residue of A1M/alpha1-microglobulin is essential for protection of irradiated cell cultures and reduction of carbonyl groups, Free Radic. Res. 47 (2013) 541–550.

[19] M. Allhorn, A. Klapyta, B. Åkerström, Redox properties of the lipocalin alpha1-microglobulin: reduction of cytochrome c, hemoglobin, and free iron, Free Radic. Biol. Med. 38 (2005) 557–567.

[20] B. Åkerström, G.J. Maghzal, C.C. Winterbourn, A.J. Kettle, The lipocalin alpha1-microglobulin has radical scavenging activity, J. Biol. Chem. 282 (2007) 31493–31503.

[21] M. Allhorn, T. Berggård, J. Nordberg, M.L. Olsson, B. Åkerström, Processing of the lipocalin α1-microglobulin by hemoglobin induces heme-binding and heme-degradation properties, Blood 99 (2002) 1894–1901.

[22] M. Allhorn, K. Lundqvist, A. Schmidtchen, B. Åkerström, Heme-scavenging role of α1-microglobulin in chronic ulcers, J. Investig. Dermatol. 121 (2003) 640–646.

[23] J. Larsson, M. Allhorn, B. Åkerström, The lipocalin α1-microglobulin binds heme in different species, Arch. Biochem. Biophys. 432 (2004) 196–204.

[24] J.F. Siebel, R.L. Kosinsky, B. Åkerström, M. Knipp, Insertion of heme b into the structure of the Cys34-carbamidomethylated human lipocalin alpha(1)-microglobulin: formation of a [(heme)(2) (alpha(1)-microglobulin)](3) complex, Chembiochem 13 (2012) 879–887.

[25] W. Meining, A. Skerra, The crystal structure of human alpha(1)-microglobulin reveals a potential haem-binding site, Biochem. J. 445 (2012) 175–182.

[26] T. Finkel, N.J. Holbrook, Oxidants, oxidative stress and the biology of ageing, Nature 408 (2000) 239–247.

[27] A. Kwasek, P. Osmark, M. Allhorn, A. Lindqvist, B. Åkerström, Z. Wasylewski, Production of recombinant human alpha1-microglobulin and mutant forms involved in chromophore formation, Protein Expr. Purif. 53 (2007) 145–152.

[28] Lamar.colostate.edu/~/sreeram/CDPro/.

[29] E. Karnaukhova, S.S. Krupnikova, M. Rajabi, A.I. Alayash, Heme binding to human alpha-1 proteinase inhibitor, Biochim. Biophys. Acta 1820 (2012) 2020–2029.

[30] EMBL, European Bioninformatics Institute, P, 2014.

[31] Molinspiration, C. o. M. P. a. B. S. (2014).

[32] Molinspiration, M. v. (2014).

[33] W, D, PyMOL Release 0.99, DeLano Scientific LLC, Palo Alto, 2002.

[34] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, J. Liang, CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues, Nucleic Acids Res. 34 (2006) W116–W118.

[35] M. Shatsky, R. Nussinov, H.J. Wolfson, A method for simultaneous alignment of multiple protein structures, Proteins 56 (2004) 143–156.

[36] S.J. de Vries, M. van Dijk, A.M. Bonvin, The HADDOCK web server for data-driven biomolecular docking, Nat. Protoc. 5 (2010) 883–897.

[37] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J. Berendsen, GROMACS: fast, flexible, and free, J. Comput. Chem. 26 (2005) 1701–1718.

[38] H.J.C. Berendsen, J.P.M. Postma, W.F. Van Gunsteren, A. Dinola, J.R. Haak, Molecular dynamics with coupling to an external bath, J. Chem. Phys 81 (1984) 3684.

[39] M. Parinello, A. Rahman, Polymorphic transitions in single crystals: a new molecular dynamics method, J. Appl. Phys. 52 (1981) 7182.

[40] C. Oostenbrink, A. Villa, A.E. Mark, W.F. van Gunsteren, A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6, J. Comput. Chem. 25 (2004) 1656–1676.

[41] B. Hess, H. Bekker, H.J. Berendsen, J.G.E.M. Fraaije, LINCS: a linear constraint solver for molecular simulations, J. Comput. Chem. 18 (1997) 1463–1472.

[42] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: an N log(N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993).

[43] E. Karnaukhova, S. Rutardottir, M. Rajabi, L. Wester Rosenlöf, A.I. Alayash, B. Åkerström, Characterization of heme binding to recombinant alpha1-microglobulin, Front. Physiol. 5 (2014) 465.

[44] C.C. Winterbourn, Oxidative reactions of hemoglobin, Methods Enzymol. 186 (1990) 265–272.

[45] I.M. Kuznetsova, T.A. Yakusheva, K.K. Turoverov, Contribution of separate tryptophan residues to intrinsic fluorescence of actin. Analysis of 3D structure, FEBS Lett. 452 (1999) 205–210.

[46] T. Li, H.L. Bonkovsky, J.T. Guo, Structural analysis of heme proteins: implications for design and prediction, BMC Struct. Biol. 11 (2011) 13.

[47] T. Berggård, A. Cohen, P. Persson, A. Lindqvist, T. Cedervall, M. Silow, I.B. Thogersen, J.A. Jönsson, J.J. Enghild, B. Åkerström, a_1-microglobulin chromophores are located to three lysine residues semiburied in the lipocalin pocket and associated with a novel lipophilic compound, Protein Sci. 8 (1999) 2611–2620.

[48] M.G. Olsson, M. Allhorn, L. Bülow, S.R. Hansson, D. Ley, M.L. Olsson, A. Schmidtchen, B. Åkerström, Pathological conditions involving extracellular hemoglobin: molecular mechanisms, clinical significance, and novel therapeutic opportunities for alpha(1)-microglobulin, Antioxid. Redox Signal. 17 (2012) 813–846.

Simeon *et al. J Cheminform (2016) 8:72*
DOI 10.1186/s13321-016-0185-8

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

osFP: a web server for predicting the oligomeric states of fluorescent proteins

Saw Simeon[1], Watshara Shoombuatong[1], Nuttapat Anuwongcharoen[1], Likit Preeyanon[2], Virapong Prachayasittikul[2], Jarl E. S. Wikberg[3] and Chanin Nantasenamat[1*]

Abstract

Background: Currently, monomeric fluorescent proteins (FP) are ideal markers for protein tagging. The prediction of oligomeric states is helpful for enhancing live biomedical imaging. Computational prediction of FP oligomeric states can accelerate the effort of protein engineering efforts of creating monomeric FPs. To the best of our knowledge, this study represents the first computational model for predicting and analyzing FP oligomerization directly from the amino acid sequence.

Results: After data curation, an exhaustive data set consisting of 397 non-redundant FP oligomeric states was compiled from the literature. Results from benchmarking of the protein descriptors revealed that the model built with amino acid composition descriptors was the top performing model with accuracy, sensitivity and specificity in excess of 80% and MCC greater than 0.6 for all three data subsets (e.g. training, tenfold cross-validation and external sets). The model provided insights on the important residues governing the oligomerization of FP. To maximize the benefit of the generated predictive model, it was implemented as a web server under the R programming environment.

Conclusion: osFP affords a user-friendly interface that can be used to predict the oligomeric state of FP using the protein sequence. The advantage of osFP is that it is platform-independent meaning that it can be accessed via a web browser on any operating system and device. osFP is freely accessible at http://codes.bio/osfp/ while the source code and data set is provided on GitHub at https://github.com/chaninn/osFP/.

Keywords: Fluorescent protein, FP, Green fluorescent protein, GFP, Oligomeric state, Data mining, Web server

Background

Many coral fluorescent proteins (FP) are observed in anthozoans and because of the fact that their tertiary structures are homologous to the *Aequorea victoria* jellyfish, they are often referred to as green fluorescent protein (GFP)-like. These FPs represent an important class of bioluminescent proteins because of their immense utility for biomedical imaging in the life sciences. Such popularity lies in the diversity of their spectral colors and their independence from co-factors owing to the autocatalytic post-translational modifications of the three or four amino acid precursors of the chromophore. Although the inherently weak dimerization of *Aequorea* GFP does not hinder its usage as a protein tag, the obligate tetrameric structure of DsRed has greatly impeded its utilization as a genetically encoded fusion tag because of possible perturbations to the tagged protein. Although oligomeric FPs in corals can serve as "sunscreen" to prevent coral bleaching, steric conflicts and stoichiometric clashes can occur when DsRed is tagged to oligomeric protein of interest (e.g. actin, tubulin, connexin or histone) [1].

Despite being the essential tagging tool for live biomedical imaging, the oligomerization of FPs hinders their utilization, problems have been reported, such as abnormal localizations, perturbing normal functions, interfering with signaling cascades, and preventing normal oligomerization fusion products within specific organelles. Shcherbo et al. [2] stressed that Katushaka, the dimeric far-red mutants of FPs from the sea anemone

*Correspondence: chanin.nan@mahidol.ac.th
[1] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
Full list of author information is available at the end of the article

Entacmaea quadricolor, can form abnormal localization in Phoenix eco cells. Mizuno et al. [3] demonstrated that aggregation of DsRed disturbs normal function of calmodulin in the cytosol. Zacharias [4] stressed that oligomerization of FPs interfered with target protein signaling cascades when using them as tagging probes for fluorescence resonance energy transfer (FRET). Lauf et al. [5] showed that tetrameric DsRed tagged with connexins led to problems because DsRed crosslinked between different connexins, negatively affected the connexin function. Jain [6] reported that in the secretory pathway of endocrine cells, EGFP oligomerized through the disulphide-linkage of Cys49 and Cys71. Typically, there are two ways to overcome oligomerization problems: modify the FPs to monomeric states through rational and/or random mutagenesis or look for natural monomeric FPs from other organisms. In regards to the former, Zacharias [4] converted the weak dimeric *A. victoria* FP to the monomeric form by scrutinizing the crystal structure of GFP and replacing hydrophobic residues with polar charged amino acids. As for the latter approach, Shagin et al. [7] screened for FPs from hydrozoan species from the ocean and observed that one in six copepoda FPs were monomeric.

Quantitative structure-property relationship (QSPR) represents an important paradigm that allows the prediction of biological and chemical properties of interests as a function of their physicochemical properties through the use of machine learning methods [8–10]. Garian [11] proposed one of the first study for predicting protein oligomerization using decision tree (DT) in which primary sequences of proteins from the SWISS-PROT database (Release 34) were classified as homodimers or non-homodimers. Afterwards, several computational models based on different machine learning algorithms were then reported such as support vector machines (SVM) [12–14], function of degree of disagreement (FDOD) [15], *k*-NN algorithm [16] and probability approaches [17]. Details of existing methods for predicting protein oligomerization properties are provided in Table 1. Although several predictive models have been reported for predicting protein oligomerization, however no computational studies exists for specifically analyzing and investigating FPs.

To the best of our knowledge, this study proposes the first computational model for predicting the oligomeric states directly from the protein sequence of a large set of FPs compiled from the literature. It is also worthy to note that the sample size of this study is comparatively smaller than the aforementioned studies on protein oligomerization but such disparity is limited by the currently available experimental data on FP oligomerization. Current machine learning methods that are being used to construct predictive models for predicting the protein oligomeric state ranges from simple and interpretable approach (e.g. DT) to more sophisticated (e.g. NN, SVM, etc.) approaches. As this study aims for simple and interpretable predictive models, the DT approach was employed for classifying FPs as being either monomeric or oligomeric. In spite of its simplicity, most of the predictive models built as a function of various classes of protein descriptors afforded robust performance as deduced by the statistical parameters. The best model was further developed as the osFP web server that is freely available at http://codes.bio/osfp/. As to encourage further developments and extension of the predictive model and web server, the source code, complete data set and example files are provided on the repository page of osFP on GitHub at https://github.com/chaninn/osFP/.

Methods

Data sets

A data set consisting of 409 FPs along with their oligomeric states were compiled from the primary literature and is available on the osFP repository page on GitHub at https://github.com/chaninn/osFP/. Monomeric FPs are ideal tools for fluorescent tagging in biomedical imaging whereas oligomeric FPs hinder their usage as tagging labels because of their tendencies to aggregate. Therefore, we aimed to classify the FPs as being either in the monomeric or oligomeric states.

Redundant sequences in the training or testing data may lead to overestimation of the model in which the learning method could only reproduce its own input data rather than being able to interpolate and extrapolate [18]. Without considering the homology relatedness, the predictive performance will be inflated. In fact, many of the FPs were obtained via site-directed mutagenesis from just a few wild-type sequences. For example, the Ala206Lys mutation in green fluorescent protein from *Aequorea victoria* caused the FP to change from oligomeric (weak dimer) to the monomeric state. Thus, it is important to consider homology reduction for the sequence-based classification performed herein. Redundancy reduction of the sequence was performed using the CD–HIT algorithm [19] as implemented in the *cdhitHR* function of the *BioSeqClass* R package [20]. Threshold values of 0.95, 0.99 and 1.00 (i.e. corresponding to 95%, 99% and 100%, respectively) was set using the *identity* argument to produce a reduced subset consisting of 136, 261 and 397 FPs, respectively.

The data set was randomly divided into two subsets consisting of an internal set (80%) and an external set (20%) in which the former set was used to constructing predictive models as full training and tenfold cross-validation (tenfold CV) while samples in the latter set was

Table 1 Summary of existing studies for predicting oligomeric states from protein sequences

Data set	Method	Internal set size	External set size	Sequence features	Source
SWISS-PROT (release 34)	DT	1639	N/A	PCP	[11]
	SVM	1639	N/A	AAC, AC	[14]
	FDOD	1639	N/A	QSO	[15]
SWISS-PROT (release 34) after removing similar protein sequence	SVM	1568	N/A	QSO	[13]
	SVM	1568	N/A	AAC, DPC, AACD	[21]
	k-NN	1568	N/A	QSO	[16]
	SVM	1568	1283	PseAAC	[12]
SWISS-PROT (release 40)	DA	3174	332	PseAAC	[22]
	SVM	3174	N/A	FS, MSE	[23]
	NN	3174	332	PseAAC	[24]
UniProtKB (release 15.6)	Probability	5495	N/A	AAC, DPC	[17]
	Fuzzy *k*-NN	5495	N/A	PseAAC	[25]
SWISS-PROT (release 55.3)	OET-*k*-NN	6702	N/A	FunD, PsePSSM	[26]
	DWT_DT	6702	N/A	PseAAC, PCP	[27]
FP data set	DT	318	79	AAC, DPC, TPC AC, CTD, Ctriad QSO, PseAAC	This study

DT decision tree, *DWT_DT* discrete wavelength transform and decision tree, *FDOD* function of degree of disagreement, *DA* discriminatory analysis, *SVM* support vector machine, *NN* neural network, *k-NN k*-nearest neighbors, *Fuzzy k-NN* Fuzzy *k*-nearest neighbors, *OET-k-NN* optimized evidence-theoretic *k*-NN algorithm, *AAC* amino acid composition, *AACD* amino acid composition distribution, *AC* autocorrelation descriptors derived from several physicochemical properties including Geary, Moreau-Broto and Moran, *APseAAC* amphiphilic pseudo-amino acid composition, *CTD* composition, transition and distribution, *Ctriad* conjoint triad descriptors, *DPC* dipeptide composition, *DWT_DT* discrete wavelet transform and decision tree, *FDOD* function of degree of disagreement, *FS* the factor scores, *FunD* functional domain composition, *MSE* multi-scale energy, *PCP* physicochemical properties, *PseAAC* pseudo amino acid composition, *PsePSSM* pseudo position-specific score matrix, *TPC* tripeptide composition, *QSO* quasi-sequence-order descriptors

predicted using the aforementioned trained model. This data splitting was performed for 100 iterations followed by computing the mean values for the performance metrics.

Protein descriptor calculation

Several classes of amino acid descriptors consisting of 8420 amino acid/dipeptide/tripeptide composition (AAC/DPC/TPC) descriptors (i.e. 20 amino acid composition, 400 dipeptide composition and 8000 tripeptide composition descriptors, 720 autocorrelation (AC) descriptors (i.e. 240 normalized Moreau-Broto autocorrelation, 240 Moran autocorrelation and 240 Geary autocorrelation descriptors), 147 composition, transition and distribution (CTD) descriptors, 343 conjoint triad (Ctriad) descriptors, 160 quasi-sequence-order (QSO) (i.e. 60 sequence-order-coupling number and 100 quasi-sequence-order descriptors) and 130 pseudo-amino acid composition (PseAAC) descriptors (i.e. 50 pseudo-amino acid composition and 80 amphiphilic pseudo-amino acid composition) were used to represent features of the amino acid sequence.

The amino acid/peptide composition descriptors were computed using *extractAAC*, *extractDC* and *extractTC* functions from the *protr* R package. Amino acid composition is the proportion of all twenty naturally occurring amino acids, dipeptide composition constitutes 400 possible sequence of dipeptides and tripeptide composition encompasses 8000 possible sequence of tripeptides. These three sets of descriptors can be defined by the following equations:

$$f(r) = \frac{N_r}{N} \quad r = 1, 2, \ldots, 20. \tag{1}$$

where N_r is the number of amino acid type r and N is the length of the sequence.

$$f(r, s) = \frac{N_{rs}}{N - 1} \quad r, s = 1, 2, \ldots, 20. \tag{2}$$

where N_{rs} is the number of dipeptides represented by amino acid types r and s.

$$f(r, s, t) = \frac{N_{rst}}{N - 2} \quad r, s, t = 1, 2, \ldots, 20 \tag{3}$$

where N_{rst} is the number of tripeptides represented by amino acid types r, s and t.

AC descriptors were computed using *extractMoreauBroto*, *extractMoran* and *extractGeary* functions from the *protr* R package [28] to obtain the normalized Moreau-Broto autocorrelation, Moran autocorrelation and Geary autocorrelation descriptors, respectively. The AC descriptors are defined based on the distribution of the physicochemical properties of amino acid, which can be derived from the AAindex database. The normalized Moreau-Broto autocorrelation is defined by 8 physicochemical properties consisting of normalized average hydrophobicity scales, average flexibility indices, polarizability parameter, free energy of solution in water, residue accessible surface area in tripeptide, residue volume, steric parameter, relative mutability with respective AAindex database ID of CIDH920105, BHAR880101, CHAM820101, CHAM820102, CHOC760101, BIGC6 70101, CHAM810101 and DAYM780201, respectively.

Moreau-Broto autocorrelation descriptor is summarized below:

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad d = 1, 2, \ldots, 30 \tag{4}$$

where d is the lag of the autocorrelation while P_i and P_{i+d} are properties of the amino acids at positions i and $i + d$, respectively.

Moran autocorrelation descriptors can be defined as follows:

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \bar{P}')(P_{i+d} - \bar{P}')}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \bar{P}')^2} \quad d = 1, 2, \ldots, 30 \tag{5}$$

where d, P_i and P_{i+d} are as defined above while $\bar{P}'$ is the considered property P along the sequence:

$$\bar{P}' = \frac{\sum_{i=1}^{N} P_i}{N} \tag{6}$$

where d, P, P_i and P_{i+d} are as described above.

Geary autocorrelation descriptors are defined as follows:

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \bar{P}')^2} \quad d = 1, 2, \ldots, 30 \tag{7}$$

where d, P, P_i and P_{i+d} are as mentioned previously.

CTD descriptors were computed using *extractCTDC*, *extractCTDT* and *extractCTDD* functions from the *protr* R package to produce the composition, transition and distribution descriptors, respectively. Briefly, the amino acid were categorized according to their properties (i.e. hydrophobicity, normalized van dar Waals volume, polarity, polarizability, charge, secondary structure and solvent accessibility) which are denoted as sub-class 1, sub-class 2 and sub-class 3. The composition descriptors is the global percentage for each class of each sequence. Depending on the sub-class, the sequence were encoded using the following equation:

$$C_r = \frac{n_r}{n} \quad r = 1, 2, 3 \tag{8}$$

where n_r is the number of amino acid type r in the encoded sequence and N is the length of the sequence. Transition is the percent frequency of a transition from one category to another, which can be calculated as follows:

$$T_{rs} = \frac{n_{rs} + n_{sr}}{N - 1} \quad rs = \text{‘12’}, \text{‘13’}, \text{‘23’} \tag{9}$$

where n_{rs} and n_{sr} are the numbers of dipeptide encoded as “rs” and “sr”, respectively, in the sequence and N is the length of the sequence. Distribution descriptor describes the chain length in which the first residue as well as 25, 50, 75 and 100% of amino acids reside for a specified encoded class.

Ctriad descriptors were obtained using the *extractCTriad* function from the *protr* R package. The conjoint triad descriptors are abstracts descriptors of protein pairs based on the categories of amino acid. The twenty natural amino acids were catogerizied based on their dipoles and volumes of the side chains because electrostatic and hydrophobic interactions, respectively, play an important part in protein–protein interaction. These two parameters were calculated via density-functional theory method B3LYP/6-31G and molecular modeling approach. The amino acids are then further categorized into seven classes based on their dipoles and values of their respective side chains. Triads can be defined as a unit of any three continuous amino acids, considering the properties of sandwiched amino acid and its vicinal amino acids.

QSO descriptors were computed using *extractSOCN* and *extractQSO* functions from the *protr* R package. The Quasi-Sequence-Order descriptors based from the distance matrix between twenty amino acids as proposed by [29].

PseAAC descriptors were calculated using *extractPAAC* and *extractAPAAC* functions of the *protr* R package. PseAAC can also be called type 1 pseudo-amino acid composition as they are based on the original hydrophobicity values, hydrophilicity and side chain masses, which can be summarized as follows:

$$H_1(i) = \frac{H_1^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^o(i)}{\sqrt{\frac{\sum_{i=1}^{20}[H_1^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^o(i)]^2}{20}}} \tag{10}$$

where $H_1^o(i)$, $H_2^o(i)$ and $M^o(i)$ $(i = 1, 2, 3, \ldots, 20)$ represents the hydrophobicity values, the hydrophilicity values and the original side chain masses of the 20 naturally occurring amino acids.

The APseAAC, also known as type 2 pseudo-amino acid composition, is defined by the following equation:

$$\begin{aligned} H^1_{i,j} &= H_1(i)H_1(j) \\ H^2_{i,j} &= H_2(i)H_2(j) \end{aligned} \tag{11}$$

where $H_1(i)$ and $H_2(j)$ represents hydrophobicity and hydrophilicity, respectively.

From these qualities, sequence order factors can be defined as follows:

$$\begin{aligned} \tau_1 &= \frac{1}{N-1}\sum_{i=1}^{N-1} H^1_{i,i+1} \\ \tau_2 &= \frac{1}{N-1}\sum_{i=1}^{N-1} H^2_{i,i+1} \\ \tau_3 &= \frac{1}{N-2}\sum_{i=1}^{N-2} H^1_{i,i+2} \\ \tau_4 &= \frac{1}{N-2}\sum_{i=1}^{N-2} H^2_{i,i+2} \\ &\ldots \\ \tau_{2\lambda-1} &= \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda} H^1_{i,i+\lambda} \\ \tau_{2\lambda} &= \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda} H^2_{i,i+\lambda} \end{aligned} \tag{12}$$

A set of APseAAC descriptors can be defined as:

$$P_c = \frac{f_c}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda}\tau_j} \quad (1 < c < 20) \tag{13}$$

$$P_c = \frac{w\tau_u}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda}\tau_j} \quad (21 < u < 20 + 2\lambda) \tag{14}$$

where w is the weighting factor and is taken as $w = 0.5$.

Thus, six descriptor classes consisting of AAC/DPC/TPC, AC, CTD, Ctriad, QSO and PseAAC were benchmark for its ability to predict the oligomeric states of FP.

Feature selection

Intercorrelation (or collinearity) is a condition where pairs of descriptors have a major correlation with each others. It has negative impact on the analysis as highly correlated predictors add more complexity to the model than information they provide. In addition, one of the key principle in the analysis of high dimensional data, which is also known as the *curse of dimensionality* that tempts practitioners to fall into a trap in which the inclusion of a higher number of features will yield higher performance for the predictive model. Indeed, adding additional features that are truly associated with the outcome (e.g. oligomerization) is expected to improve the predictive model. On the other hand, the addition of noise features that are not truly relevant to the outcome is expected to deteriorate the model thereby leading to a reduction of the model performance. This is because the incorporation of noise features tends to increase the risk of overfitting. As low collinearity is favorable for retaining a non-redundant set of descriptors and as there is no strict criteria on the removal threshold, therefore typically high threshold value for the correlation coefficient are employed. Cronin and Schultz et al. [30] pointed out that there seems to be no consensus on the threshold criterion for the correlation coefficient as acceptable values ranged from less than 0.4 to 0.9. Thus, the *cor* function from the *caret* R package [31] was used to calculate correlations between descriptors. Subsequently, collinear descriptors were removed using an arbitrary threshold of 0.7 for the Pearson's correlation coefficient as implemented by the *findCorrelation* function from the *caret* R package. Such threshold value is deemed to be a stringent value for exclusion of descriptors displaying mild intercorrelation with one another whereas a high threshold value of 0.9, for instance, would allow fewer removal of descriptors while allowing descriptor pairs with mild intercorrelation to be included in the model.

Multivariate analysis

A decision tree (DT) algorithm was utilized for constructing a computational model to predict FP oligomeric states. Because the DT method affords interpretable rules for estimating feature importance pertaining to FP oligomeric states, it is helpful in revealing the different characteristics between monomeric and oligomeric states. The construction of a DT model requires the following: (i) all samples in the internal set belong to a single class; (ii) the tree depth is close to maximum; and (iii) the number of classes in the terminal node is less than the minimum number of classes of the parent nodes. In general, the root node is a variable with the highest information gain, whereas the other internal nodes provide the second and subsequent highest information gain thereafter. Machine learning models were built in the R statistical programming language using the *J48* function from the *RWeka* R package.

Statistical assessment of predictive model

For any empirical learning method, statistical assessment of the model robustness is an important process. Four measurements were used to evaluate the prediction performances of the proposed model: accuracy (Ac), sensitivity (Sn), specificity (Sp) and Matthews' correlation coefficient (MCC). These parameters are defined as follows:

$$Ac = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100 \tag{15}$$

$$Sn = \frac{TP}{(TP + FN)} \times 100 \tag{16}$$

$$Sp = \frac{TN}{(TN + FP)} \times 100 \tag{17}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{18}$$

where TP is the instances of true positives, TN is the instances of true negatives, FP is the instances of false positives and FN is the instance of false negatives. In this study, a tenfold CV procedure was used to confirm the reliability and robustness of the QSPR models using the training set. Additionally, external validation set was used to assess the generalizability to our proposed model for predicting unknown samples. It should be noted that the range of MCC is from −1 to 1 in which a value of 1 indicates the best possible prediction while −1 indicates the worst possible prediction. On the other hand, a value of 0 suggests the occurrence of random prediction.

Development of the osFP webserver

The osFP web server was developed using the web application framework known as *Shiny* under the R statistical programming language. Technically, the Shiny web application framework is comprised of two components: (i) ui.R (i.e. the user interface script) and (ii) server.R (i.e. the server script). The user interface script is responsible for producing the layout of the web application that users can see and interact (i.e. entering the input sequence of FP in FASTA format for calculation submission) while the server script performs the calculations and generates the output (i.e. prediction results of the oligomeric state). As it is computationally intensive to compute 8420 descriptors (i.e. 20 amino acid, 400 dipeptide and 8000 tripeptide composition descriptors), only the top 20 important features as revealed by the DT model were used in the production environment (i.e. the osFP web server). As such, this required slight adaptation to the descriptor calculation functions from the *protr* R package, particularly by computing only specific descriptors from the list of the top 20 important features instead of the default total number of 8420 descriptors for the three descriptor classes.

osFP is hosted on a Ubuntu Linux server via the the cloud infrastructure provider, DigitalOcean. The benefits of hosting on the cloud is many: (i) low start-up cost (i.e. no need for costly investments on hardware, no maintenance cost and no need for server administrator), (ii) scalable resources (i.e. when the need for more RAM or storage arises the server can be upgraded) (iii) operating systems are pre-installed and available in several Linux distributions (i.e. no need for lengthy installation of the operating system as a working server takes under a minute to be provisioned), (iv) full access and control of the server (i.e. freedom to install and configure softwares) and (v) the whole server can be backed up as an image.

The provisioned web server used to host osFP is based on Ubuntu version 14.10. Firstly, the R base software and associated packages (i.e. shiny, shinythemes, shinyjs, protr, seqinr, RWeka and markdown, which are used on the osFP web server) were installed via the apt-get package handling utility in the command line. Secondly, the RStudio Shiny Server, which is available at https://www.rstudio.com/products/shiny/download-server/, was installed. At default, the directory for housing Shiny applications is set to /srv/shiny-server/ while the Shiny application would typically run at port number 3838, therefore the base URL will look something like http://192.168.1.1:3838/ where 192.168.1.1 represents the IP address while the full URL would look something like http://192.168.1.1:3838/osfp/. There is a workaround to hiding the port number but one needs to configure the Shiny configuration file (i.e. available at /etc/shiny-server/shiny-server.conf) and/or the Apache server settings (i.e. available at /etc/apache2/sites-available/). There is an excellent step-by-step tutorial provided by the DigitalOcean user community on installing and configuring the Shiny server [32]. Although, the Shiny server supports the use of databases such as MySQL, however the simplicity and moderate size of the data set employed herein makes satisfactory use of the CSV file format to store, retrieve and analyze the data via functionalities of the R environment.

Results and discussion

Predicting FP oligomeric states

Figure 1 illustrates the flowchart of the workflow used to predict and analyze the oligomerization of FPs. In the study, six classes of protein features were benchmark to provide a better picture on which protein features can be considered sufficient to provide insights on the oligomerization of FPs. To avoid the possibility of obtaining prediction results that may arise from chance correlation from a single calculation, the multivariate analysis was performed for 100 independent iterations where each run involves random data splitting to an internal and external sets consisting of 80 and 20%, respectively.

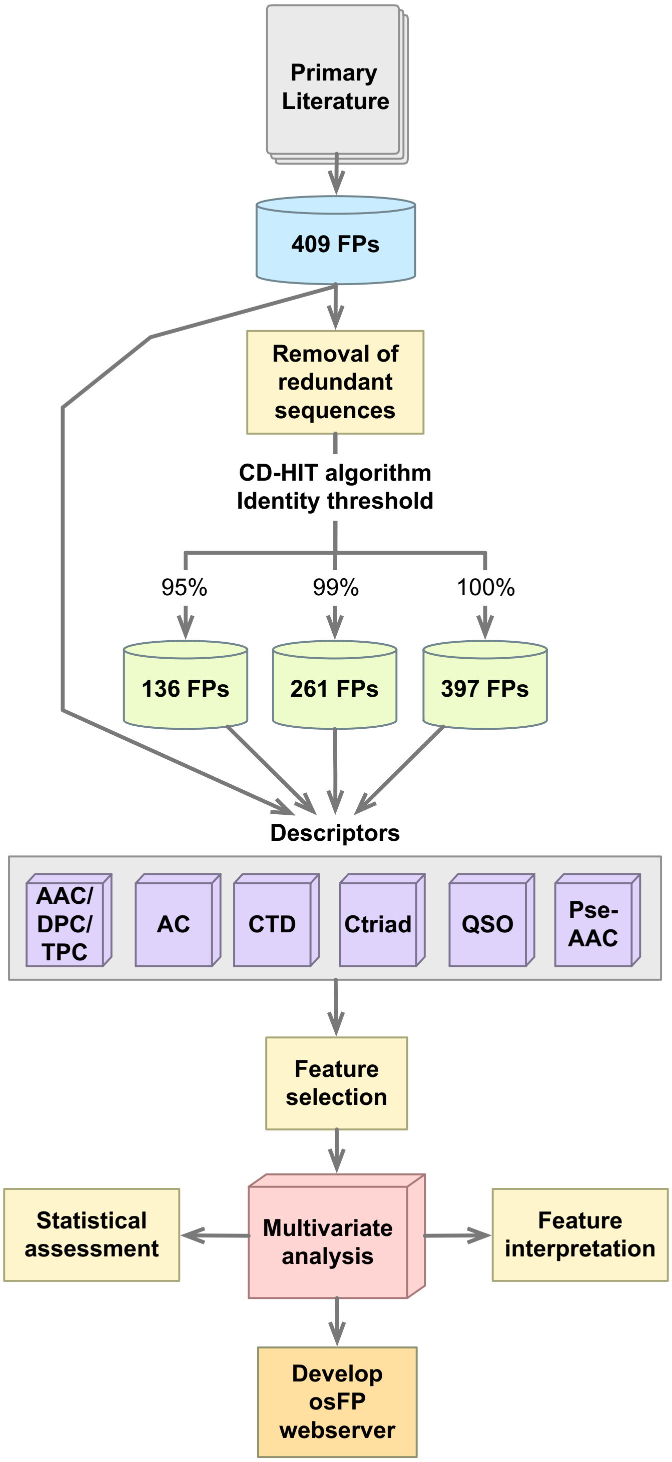


Fig. 1 Workflow of QSPR modeling for predicting oligomeric states of FP

Judging from the performances, although different descriptor sets may capture different aspects of amino acids however all descriptor sets afforded similar level of performance. This may indicate that all descriptor sets capture the oligomerization space well and can be used as a features to train predictive QSPR models as assessed via Ac, Sn, Sp and MCC. However, the best performing method was AAC/DPC/TPC descriptors whereas AC descriptors ranked last, indicating that the amino acid composition descriptors are capable of capturing information on the oligomerization of FP. The predictive performance of the six classes of protein descriptors were further discussed in the following paragraphs.

The internal set was used to construct a predictive model by means of the J48 algorithm as to discriminate FPs to either monomers or oligomers. The predictive model was fine tuned using tenfold CV as to prevent overtraining on the internal set and then tested on an external set in order to assess its ability to accurately predict unknown samples. Table 2 provides the mean performance comparison amongst the various types of protein descriptors as assessed by the training set, the tenfold CV set and the external set.

It was observed that models built with AAC/DPC/TPC outperformed the others with Ac, Sn, Sp and MCC of 83.01 ± 2.04%, 83.26 ± 2.19%, 82.95 ± 2.27% and 0.66 ± 0.04, respectively, for the tenfold CV set. Furthermore, AAC/DPC/TPC also afforded the best performance on the external set with Ac, Sn, Sp and MCC of 83.26 ± 3.58%, 83.77 ± 5.14%, 83.37 ± 4.45% and 0.67 ± 0.07, respectively. On the other hand, the autocorrelation descriptors afforded the lowest performance with Ac, Sn, Sp and MCC of 78.48 ± 4.76%, 78.65 ± 5.66%, 78.89 ± 5.62% and 0.57±0.10, respectively. The predictive models built using CTD, Ctriad, QSO and PseAAC provided moderate performance with MCC of 0.60 ± 0.10, 0.62 ± 0.10, 0.63 ± 0.08 and 0.63 ± 0.09, respectively.

When looking into the relationship between protein sequence features and oligomerization, the phylogenetic relationships between sequences in the data sets should be taken into account. By not considering homologous relatedness amongst the FP samples, a problem in which FPs are the products of site-directed mutagenesis from a few wild-type sequences may arise. On the other hand, one site mutation may convert oligomeric FP to the monomeric state. For instance, the Ala206Lys mutation could convert a weakly oligomeric GFP to the monomeric form. Therefore, homologous reduction with the threshold of 100, 99 and 95% were considered.

Table 2 shows the results of the predictive models for the data set that removes all identical homologous sequence via the use of an identity threshold of 100% (non-redundant data set), it can be seen that the top performing tenfold CV set was built using amino acid composition, which afforded the highest performance with Ac, Sn, Sp and MCC of 83.07 ± 2.04%, 83.26 ± 2.19%, 82.95 ± 2.27% and 0.66 ± 0.04, respectively, whereas the performance of tenfold CV of autocorrelation descriptors (i.e. normalized Moreau-Broto autocorrelation, Moran autocorrelation and Geary autocorrelation) was 78.49 ± 4.76%, 78.36 ± 2.36%, 78.67 ± 2.30% and 0.57 ± 0.04 for Ac, Sn, Sp and MCC, respectively. However, the J48 model built with CTD, Conjoint, QSO and PseAAC were comparable with Ac, Sn, Sp and MCC in the ranges of 80.15–81.13, 79.66–81.19, 79.82–81.14 and 0.60–0.62, respectively.

The performance of models with homologous sequence reduction set at 99% is shown in Table 3. Again, J48 model built using AAC/DPC/TPC descriptors was the top performing model as assessed via tenfold CV with Ac, Sn, Sp and MCC with 79.40 ± 2.75%, 80.78 ± 1.86%, 77.98 ± 3.20% and 0.59 ± 0.06, respectively, when compared to other J48 models built with different descriptor sets. As for the external set, the J48 model with the lowest performance was made with AC descriptors having Ac, Sn, Sp and MCC of 72.73 ± 6.11%, 74.89 ± 6.72%, 71.43 ± 7.57% and 0.46 ± 0.12, respectively. Nevertheless, it can be observed that J48 models built with different descriptor set performance well as assessed via tenfold CV set and external set.

For the performance of the sequence homologous reduction at 95%, it can be seen that the top performing model of the tenfold CV set resulted in Ac, Sn, Sp and MCC of 72.13 ± 4.18%, 79.83 ± 3.66%, 61.03 ± 5.34% and 0.42 ± 0.09, respectively, was from the model built using amino acid composition. On the other hand, the other models built using different descriptors were comparable as shown in Table 4. As for the external set, again, model built with amino acid composition outperform others with Ac, Sn, Sp and MCC of 72.89 ± 7.08%, 79.85 ± 6.92%, 64.16 ± 11.20% and 0.43 ± 0.15, respectively.

Identifying informative features

Investigating feature importance of each type of protein descriptor can provide insights into FP oligomerization. Herein, the efficient built-in feature importance selector of the DT algorithm was used. In the DT algorithm, the estimation of feature importance is calculated from the feature usage based on information gain. The feature with the highest usage score is the most important feature because it maximizes the prediction performance. Since amino acid composition provided the highest performance, it was selected as an input to explore important features for discriminating the oligomers from the monomers.

Figure 2 demonstrates the top ten informative descriptors with the following feature usage: RMY (95.96 ± 7.29), LI (44.56 ± 18.72), MVS (34.12 ± 15.28), ML (28.82 ± 14.14), YS (24.79 ± 12.90), KLE (21.38 ± 12.45), SF (19.28 ± 11.69), NR (17.01 ± 10.67), HY (14.92 ± 9.99),

Table 2 Summary of performance of QSAR models for predicting the oligomeric state of FPs (100% homologous sequence reduction) using the J48 algorithm

Descriptors	Training set				Tenfold CV set				External set			
	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
AAC/DPC/TPC	97.40 ± 0.77	97.72 ± 1.23	97.14 ± 1.31	0.95 ± 0.02	83.07 ± 2.04	83.26 ± 2.19	82.95 ± 2.27	0.66 ± 0.04	83.26 ± 3.58	83.77 ± 5.14	83.37 ± 4.45	0.67 ± 0.07
AC	98.70 ± 0.58	98.54 ± 0.93	98.86 ± 0.75	0.97 ± 0.01	78.36 ± 2.35	78.36 ± 2.36	78.67 ± 2.30	0.57 ± 0.04	78.49 ± 4.76	78.65 ± 5.66	78.89 ± 5.62	0.57 ± 0.10
CTD	97.58 ± 0.82	98.05 ± 0.99	97.20 ± 1.33	0.95 ± 0.02	80.28 ± 2.37	79.66 ± 2.98	80.88 ± 2.33	0.60 ± 0.05	80.40 ± 4.92	80.37 ± 6.63	81.01 ± 5.34	0.61 ± 0.10
Ctriad	95.46 ± 1.08	96.20 ± 1.51	94.85 ± 1.87	0.91 ± 0.02	80.38 ± 2.01	81.07 ± 2.43	79.82 ± 2.08	0.61 ± 0.04	81.06 ± 4.83	81.80 ± 5.79	80.98 ± 5.55	0.62 ± 0.10
QSO	98.35 ± 0.63	98.35 ± 0.83	98.36 ± 0.93	0.97 ± 0.01	80.15 ± 1.91	80.11 ± 2.25	80.27 ± 2.16	0.60 ± 0.04	81.42 ± 4.00	81.86 ± 5.21	81.58 ± 5.00	0.63 ± 0.08
PseAAC	98.51 ± 0.62	98.63 ± 0.85	98.42 ± 1.05	0.97 ± 0.01	81.13 ± 1.89	81.19 ± 2.24	81.14 ± 2.09	0.62 ± 0.04	81.40 ± 4.66	81.13 ± 5.48	82.19 ± 5.54	0.63 ± 0.09

Table 3 Summary of performance of QSAR models for predicting the oligomeric state of FPs (99% homologous sequence reduction) using the J48 algorithm

Descriptors	Training set				Tenfold CV set				External set			
	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
AAC/DPC/TPC	98.22 ± 0.71	98.73 ± 0.89	97.69 ± 1.20	0.97 ± 0.01	79.40 ± 2.75	80.78 ± 1.86	77.98 ± 3.20	0.59 ± 0.06	80.78 ± 5.72	82.12 ± 6.28	80.12 ± 7.50	0.62 ± 0.12
AC	98.22 ± 0.71	98.73 ± 0.89	97.69 ± 1.20	0.96 ± 0.01	72.88 ± 3.46	74.52 ± 3.65	71.20 ± 3.74	0.46 ± 0.07	72.73 ± 6.11	74.89 ± 6.72	71.43 ± 7.57	0.46 ± 0.12
CTD	97.66 ± 0.90	98.06 ± 1.17	97.29 ± 1.46	0.95 ± 0.02	74.40 ± 3.03	75.01 ± 3.32	73.92 ± 3.42	0.49 ± 0.06	74.89 ± 5.79	75.76 ± 6.64	74.83 ± 6.95	0.50 ± 0.12
Ctriad	95.25 ± 1.87	96.62 ± 1.57	94.00 ± 3.64	0.91 ± 0.04	73.66 ± 2.84	75.78 ± 3.08	71.54 ± 3.15	0.47 ± 0.06	74.22 ± 6.02	77.26 ± 7.06	72.19 ± 7.21	0.49 ± 0.12
QSO	98.50 ± 0.64	98.63 ± 1.00	98.39 ± 1.15	0.97 ± 0.01	75.78 ± 2.71	77.37 ± 3.08	74.11 ± 2.74	0.51 ± 0.05	76.71 ± 5.27	78.92 ± 6.19	75.19 ± 6.35	0.54 ± 0.11
PseAAC	98.17 ± 0.74	98.38 ± 1.08	97.98 ± 1.36	0.96 ± 0.01	74.35 ± 3.00	75.87 ± 2.83	72.81 ± 3.67	0.49 ± 0.06	74.71 ± 5.74	77.17 ± 6.60	72.97 ± 6.93	0.50 ± 0.12

Table 4 Summary of performance of QSAR models for predicting the oligomeric state of FPs (95% homologous sequence reduction) using the J48 algorithm

Descriptors	Training set				Tenfold CV set				External set			
	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
AAC/DPC/TPC	97.54 ± 1.19	99.25 ± 0.90	94.85 ± 2.50	0.95 ± 0.03	72.13 ± 4.18	79.83 ± 3.66	61.03 ± 5.34	0.42 ± 0.09	72.89 ± 7.08	79.85 ± 6.92	64.16 ± 11.20	0.43 ± 0.15
AC	98.35 ± 0.87	99.31 ± 0.92	96.81 ± 1.95	0.97 ± 0.02	70.71 ± 4.45	77.73 ± 3.63	59.80 ± 6.10	0.38 ± 0.09	70.30 ± 8.55	77.40 ± 7.91	60.99 ± 13.19	0.38 ± 0.18
CTD	97.97 ± 1.06	98.33 ± 1.40	97.50 ± 1.95	0.96 ± 0.02	69.40 ± 4.95	75.24 ± 4.39	60.62 ± 6.33	0.39 ± 0.10	70.18 ± 7.79	75.54 ± 7.39	63.17 ± 12.39	0.38 ± 0.17
Ctriad	96.62 ± 1.33	98.07 ± 1.52	94.35 ± 2.89	0.93 ± 0.03	68.64 ± 5.99	76.28 ± 4.49	57.20 ± 8.12	0.34 ± 0.12	71.26 ± 8.36	78.04 ± 7.20	62.51 ± 12.24	0.40 ± 0.17
QSO	98.10 ± 1.08	98.55 ± 1.25	97.42 ± 2.33	0.96 ± 0.02	68.98 ± 4.21	76.15 ± 3.45	57.59 ± 5.63	0.34 ± 0.09	69.93 ± 6.90	77.19 ± 5.75	60.30 ± 11.15	0.37 ± 0.14
PseAAC	98.24 ± 0.92	98.38 ± 1.30	98.07 ± 1.71	0.96 ± 0.02	69.39 ± 4.97	76.34 ± 3.98	58.20 ± 6.67	0.35 ± 0.10	69.67 ± 8.03	76.92 ± 7.01	59.53 ± 10.36	0.36 ± 0.17

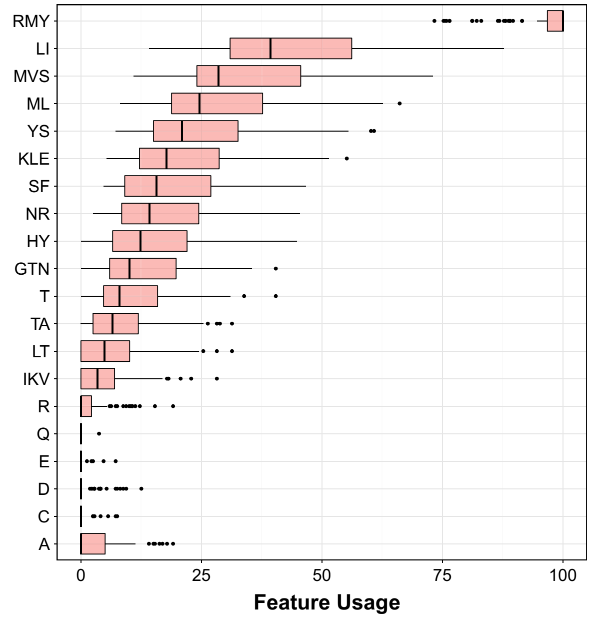


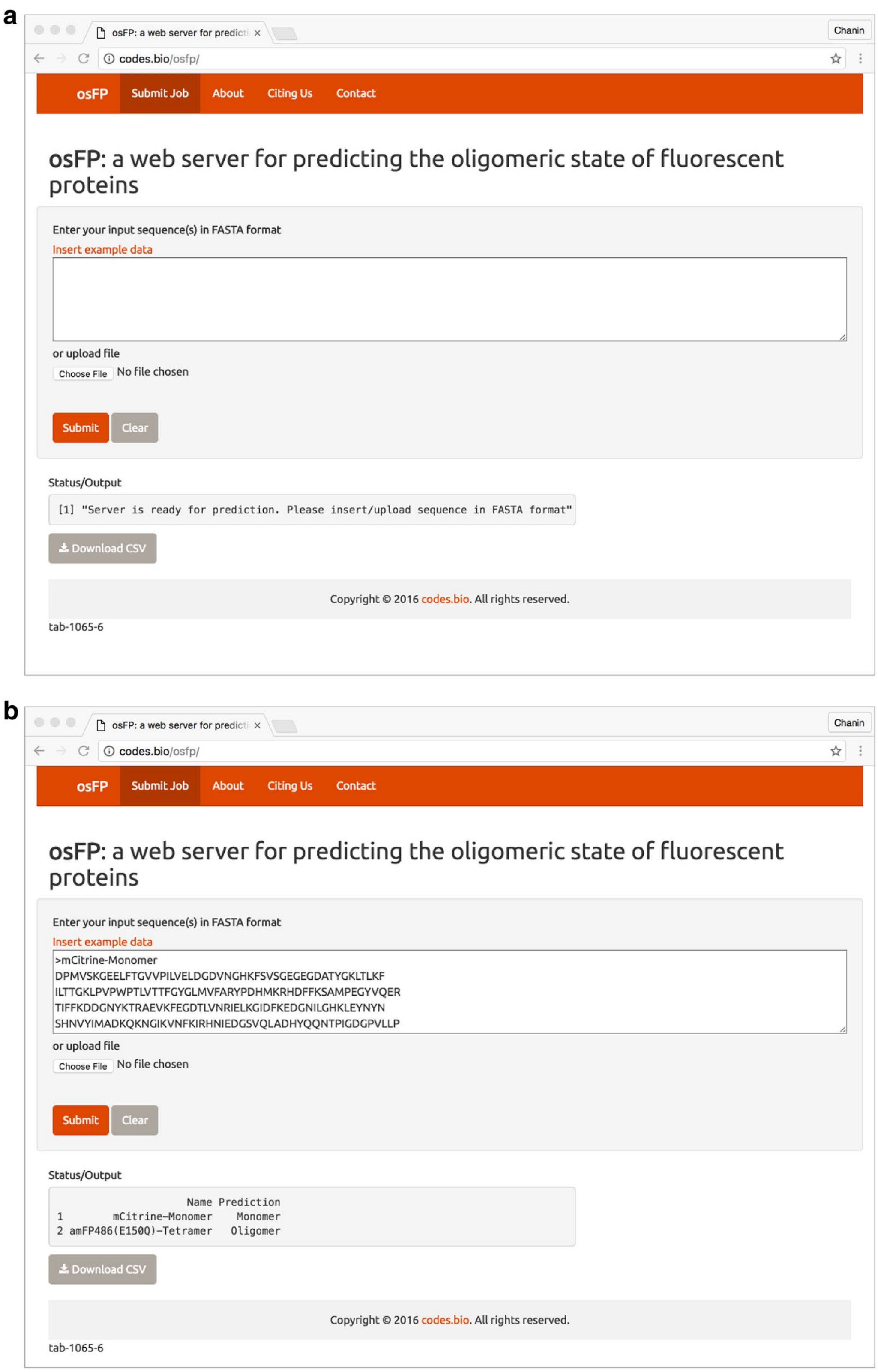
Fig. 2 Box plot of the feature usage from the predictive model of FP oligomerization. Features with the highest usage is deemed to be the most important

GTN (13.24 ± 9.58) and T (11.05 ± 3.34). Notably, the top informative descriptors was the tripeptide RMY, which is comprised of the positively-charged Arg, the hydrophobic Met as well as the aromatic/hydrophobic Tyr. The second most important feature was the dipeptide LI, which are hydrophobic amino acids. Subsequent features from the top ten informative descriptors were also primarily hydrophobic in nature. This finding is corroborated by the experimental findings of Yarbrough et al. [33] in which the crystal structure of *Discosoma sp.* DsRed indicated that the oligomeric interfaces of subunits A and B consisted mostly of hydrophobic interaction along with a few hydrogen bonds and salt bridges. In a similar manner, the first discovered photoconvertible Kaede from *Trachyphyllia geoffroyi* displayed dominant hydrophobic interactions between the oligomeric interface at the A and C subunits [34]. Additionally, *Heteractis crispa* HcRed, the commercially available dimeric FP from Clontech, was converted to a dimer from a tetramer via the replacement of the hydrophobic Leu at position 123 to the aromatic His residue thereby perturbing the tetrameric hydrophobic interface. These findings reiterated that FP oligomerization are stabilized by several hydrophobic contact. Thus, hydrophobic residues at the interface were substituted with polar residues in attempt to create monomeric FPs [33–36]. Along with hydrophobic contacts, several other interactions including the formation of coordination bonds, ionic interactions, van der Waals' contacts, electrostatic interactions, hydrogen bondings and π-π stackings may mediate FP oligomerization at the oligomeric interface.

osFP web server

To maximize the utility of the predictive model of FP oligomerization, a web server was developed using the Shiny package under the R programming environment. The utilization of Shiny boasts several benefits. The first advantage is the seamless integration of the web server with the aforementioned predictive model that was also built in R. The second benefit is that there is no requirement for developers to have an extensive knowledge of web development (i.e. although it may be useful). Thirdly, Shiny is platform-independent and can launch locally from any R environment (console R, RGui, RStudio, etc.) on any operating system whether Windows, Mac or Linux. Alternatively, users could also setup a remote server with installed instances of R and Shiny such that only a web browser is required to gain access to the application. As users can run their own instance of osFP, they can choose to customize the code to their own needs, run the application offline as well as ensuring strict privacy of the input data (i.e. it should be noted that the osFP web server does not cache or store the input data submitted by users). Most importantly, the fourth reason is that Shiny facilitates rapid development and deployment of web applications, which is especially beneficial for the scientific community as predictive models can be readily deployed as a web server, which is accessible to a wider group of users instead of confined to those with a background in computer science.

The web server user interface accepts the input sequence data of FP in FASTA format and relays such information to the server script in which a predictive model is constructed and applied for classifying the input sequence(s) as being either monomeric or oligomeric. A screenshot of the osFP webserver is shown in Fig. 3. Under the hood, two R scripts are primarily responsible for driving the osFP web server along with the auxiliary role of the markdown files (e.g. about.md, cite.md and contact.md), which stores the content text that appears on the website. Firstly, the ui.R script performs as implied by its file name that is to house the user interface elements such as the website name, the navigation bar (i.e. links to the markdown file to display the respective constituent text appearing on the about, cite and contact tabs), the input text box, the file upload button, the *Insert example data* link, the submit button and the *Status/Output* text box. It should be noted that the website theme of the osFP web server is based on the shinythemes package



a

osFP: a web server for predicting the oligomeric state of fluorescent proteins

Enter your input sequence(s) in FASTA format

Insert example data

or upload file

No file chosen

Status/Output

```
[1] "Server is ready for prediction. Please insert/upload sequence in FASTA format"
```

Copyright © 2016 codes.bio. All rights reserved.

tab-1065-6

b

osFP: a web server for predicting the oligomeric state of fluorescent proteins

Enter your input sequence(s) in FASTA format

Insert example data

```
>mCitrine-Monomer
DPMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKF
ILTTGKLPVPWPTLVTTFGYGLMVFARYPDHMKRHDFFKSAMPEGYVQER
TIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYN
SHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLP
```

or upload file

No file chosen

Status/Output

```
                    Name Prediction
1        mCitrine-Monomer    Monomer
2 amFP486(E150Q)-Tetramer   Oligomer
```

Copyright © 2016 codes.bio. All rights reserved.

tab-1065-6

Fig. 3 Screenshot of the osFP web server. Shown are the web server before (**a**) and after (**b**) prediction

in R, which at default makes use of the themes provided by Bootswatch (https://bootswatch.com/). These themes are written in Bootstrap (i.e. a HTML, CSS, and JS framework) that enable websites to be responsive and mobile-friendly (i.e. compresses the website width to fit onto a smart phone or tablet or expands the website width to fit the screen of a desktop or laptop monitor).

Secondly, the server.R script processes the data and builds the model as summarized by the following pseudocode:

```
1. Import R packages
2. Define function for computing amino acid based
   descriptors
3. Model building
   • Accepts input FASTA sequence data from the text
     box or uploaded file
   • Process the sequence data by computing the amino
     acid based descriptors for both the training and
     input data sets
   • Combine descriptors and constructs the DT model
     using C4.5 algorithm
   • Applies the constructed model to predict the oligo-
     meric states of the input sequence data
4. Outputs the prediction results in an output text box
   on the webpage
5. Makes prediction results available for download as
   a CSV file.
```

The procedure for using the osFP web server is summarized below.

Step 1. Before starting the prediction, users should wait until the gray box that is found under the *Status/Output* heading shows the following text *Server is ready for prediction. Please insert/upload sequence in FASTA format*.

Step 2. Once the aforementioned message appears, users can enter their query sequence into the Input box or upload their sequence file by clicking on the *Choose file* button (i.e. found below the *Enter your input sequence(s) in FASTA format* heading). Finally, click on the *Submit* button to initiate the prediction process.

At the onset, users may also want to try out the functionality of the osFP web server via the use of an example input data by clicking on the *Insert example data* link. This calls upon the *updateTextInput* function from the Shiny package so as to insert the example FASTA data stored in the *fastaexample* variable into the input text box. Similarly, users can initiate the prediction process by clicking on the *Submit* button.

Step 3. The prediction results are automatically displayed in a gray box below the *Status/Output* heading. Users can also download the prediction results as a CSV file by clicking on the *Download CSV button*.

Conclusion

This study represents the attempt in the development of a computational model for predicting and analyzing FP oligomerization from protein sequences using six classes of sequence descriptors consisting of AAC/DPC/TPC, AC, CTD, Ctriad, QSO and PseAAC. Findings indicated that the DT algorithm utilizing AAC/DPC/TPC (i.e. amino acid/peptide composition) outperformed the other descriptor class. Identification of informative features as obtained from the feature usage scores of DT revealed that the oligomeric interface are predominantly occupied by hydrophobic residues with a few electrostatic residues engaging in salt bridges. The results presented herein provide a glimpse on the important residues at the oligomeric interface that may be useful for guiding the rational design of monomeric forms of FP. To benefit the scientific community the predictive model was deployed as the osFP web server as well as providing the source codes and data sets on GitHub as to encourage further extension or adaptation of the web server. It is worthy to note that as new experimental data becomes available on the oligomeric states of FPs, the predictive model proposed herein could be continually updated by these growing data as to augment the model's coverage and accuracy.

Authors' contributions

CN conceived and designed the study. SS compiled the data, constructed the predictive models and designed the first draft of the web server. CN extended the web server by simplifying the user interface and optimizing the code performance. SS, WS, NA, LP, VP, JESW and CN analyzed the data. SS and CN drafted the manuscript. All authors read and approved the final manuscript.

Author details

[1] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand. [2] Department of Community Medical Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand. [3] Department of Pharmaceutical Biosciences, Uppsala University, 751 24 Uppsala, Sweden.

Acknowledgements

This research is supported by the Swedish Research Links program (No. C0610701) from the Swedish Research Council. WS and CN are grateful for the New Scholar Research Grant (No. MRG5980220) and the Research Career Development Grant (No. RSA5780031), respectively, from the Thailand Research Fund. The internet browser icon used in the graphical abstract is designed by Freepik from Flaticon.

Competing interests

The authors declare that they have no competing interests.

Received: 7 September 2016 Accepted: 29 November 2016
Published online: 20 December 2016

References

1. Baird GS, Zacharias DA, Tsien RY (2000) Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral. Proc Nal Acad Sci USA 97(22):11984–11989
2. Shcherbo D, Merzlyak EM, Chepurnykh TV, Fradkov AF, Ermakova GV, Solovieva EA, Lukyanov KA, Bogdanova EA, Zaraisky AG, Lukyanov S (2007) Bright far-red fluorescent protein for whole-body imaging. Nat Methods 4(9):741–746
3. Mizuno H, Sawano A, Eli P, Hama H, Miyawaki A (2001) Red fluorescent protein from Discosoma as a fusion tag and a partner for fluorescence resonance energy transfer. Biochemistry 40(8):2502–2510
4. Zacharias DA (2002) Sticky caveats in an otherwise glowing report: oligomerizing fluorescent proteins and their use in cell biology. Sci Signal 2002(131):23
5. Lauf U, Lopez P, Falk MM (2001) Expression of fluorescently tagged connexins: a novel approach to rescue function of oligomeric DsRed-tagged proteins. FEBS Lett 498(1):11–15
6. Jain R, Joyce P, Molinete M, Halban P, Gorr S (2001) Oligomerization of green fluorescent protein in the secretory pathway of endocrine cells. Biochem J 360:645–649
7. Shagin DA, Barsova EV, Yanushevich YG, Fradkov AF, Lukyanov KA, Labas YA, Semenova TN, Ugalde JA, Meyers A, Nunez JM (2004) GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity. Mol Biol Evol 21(5):841–850
8. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V (2009) A practical overview of quantitative structure-activity relationship. EXCLI J 8(7):74–88
9. Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V (2010) Advances in computational methods to predict the biological activity of compounds. Exp Opin Drug Discov 5(7):633–654. doi:10.1517/17460441.2010.492827
10. Nantasenamat C, Prachayasittikul V (2015) Maximizing computational tools for successful drug discovery. Exp Opin Drug Discov 10(4):321–329
11. Garian R (2001) Prediction of quaternary structure from primary structure. Bioinformatics 17(6):551–556
12. Qiu J-D, Suo S-B, Sun X-Y, Shi S-P, Liang R-P (2011) OligoPred: a web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into chou's pseudo amino acid composition. J Mol Graph Model 30:129–134
13. Song J, Tang H (2005) Support vector machines for classification of homo-oligomeric proteins by incorporating subsequence distributions. J Mol Struct (Thoechem) 722(1):97–101
14. Zhang S-W, Pan Q, Zhang H-C, Zhang Y-L, Wang H-Y (2003) Classification of protein quaternary structure with support vector machine. Bioinformatics 19(18):2390–2396
15. Song J, Tang H (2004) Accurate classification of homodimeric vs other homooligomeric proteins using a new measure of information discrepancy. J Chem Inf Comput Sci 44(4):1324–1327
16. Song J (2007) Prediction of homo-oligomeric proteins based on nearest neighbour algorithm. Comput Biol Med 37(12):1759–1764
17. Carugo O (2007) A structural proteomics filter: prediction of the quaternary structural type of hetero-oligomeric proteins on the basis of their sequences. J Appl Crystallogr 40(6):986–989
18. Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution, pp 783–791
19. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659
20. Hong L (2015) BioSeqClass: Classification for biological sequences. R package version 1.24.0
21. Shi J, Pan Q, Zhang S, Cheng Y (2005) Classification of protein homo-oligomers using amino acid composition distribution. Shengwu Wuli Xuebao 22(1):49–56
22. Chou K-C, Cai Y-D (2003) Predicting protein quaternary structure by pseudo amino acid composition. Proteins Struct Funct Bioinf 53(2):282–289
23. Zhang S-W, Chen W, Zhao C-H, Cheng Y-M, Pan Q (2007) Predicting protein quaternary structure with multi-scale energy of amino acid factor solution scores and their combination. In: Zhang DY (ed) Medical Biometrics: First International Conference, ICMB 2008, Hong Kong, China, January 4–5, 2008, Proceedings. Springer, Berlin, pp 65–72
24. Xiao X, Lin W-Z (2009) Application of protein grey incidence degree measure to predict protein quaternary structural types. Amino Acids 37(4):741–749
25. Xiao X, Wang P, Chou K-C (2011) Quat-2L: a web-server for predicting protein quaternary structural attributes. Mol Divers 15(1):149–155
26. Shen H-B, Chou K-C (2009) QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J Proteome Res 8(3):1577–1584
27. Sun X-Y, Shi S-P, Qiu J-D, Suo S-B, Huang S-Y, Liang R-P (2012) Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. Mol BioSyst 8(12):3178–3184
28. Xiao N, Xu Q, Cao D (2015) protr: Generating various numerical representation schemes of protein sequence. R package, version 1.1-1
29. Chou K-C (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278(2):477–483
30. Cronin MTD, Schultz TW (2003) Pitfalls in QSAR. J Mol Struct (Theochem) 622(12):39–51
31. Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(5):1–26
32. Attali D (2016) How to set up Shiny server on Ubuntu 14.04. https://www.digitalocean.com/community/tutorials/how-to-set-up-shiny-server-on-ubuntu-14-04. Accessed 9 Nov
33. Yarbrough D, Wachter RM, Kallio K, Matz MV, Remington SJ (2001) Refined crystal structure of DsRed, a red fluorescent protein from coral, at 2.0-Å resolution. Proc Natl Acad Sci USA 98(2):462–467
34. Hayashi I, Mizuno H, Tong KI, Furuta T, Tanaka F, Yoshimura M, Miyawaki A, Ikura M (2007) Crystallographic evidence for water-assisted photo-induced peptide cleavage in the stony coral fluorescent protein Kaede. J Mol Biol 372(4):918–926
35. Campbell RE, Tour O, Palmer AE, Steinbach PA, Baird GS, Zacharias DA, Tsien RY (2002) A monomeric red fluorescent protein. Proc Natl Acad Sci USA 99(12):7877–7882
36. Wilmann PG, Petersen J, Pettikiriarachchi A, Buckle AM, Smith SC, Olsen S, Perugini MA, Devenish RJ, Prescott M, Rossjohn J (2005) The 2.1 Å crystal structure of the far-red fluorescent protein HcRed: inherent conformational flexibility of the chromophore. J Mol Biol 349(1):223–237

SAR AND QSAR IN ENVIRONMENTAL RESEARCH, 2017
VOL. 28, NO. 1, 1–16
http://dx.doi.org/10.1080/1062936X.2016.1264468

Large-scale classification of P-glycoprotein inhibitors using SMILES-based descriptors

V. Prachayasittikul[a], A. Worachartcheewan[a,b,c], A. P. Toropova[d], A. A. Toropov[d], N. Schaduangrat[a], V. Prachayasittikul[e] and C. Nantasenamat[a]

[a]Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand; [b]Department of Community Medical Technology, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand; [c]Department of Clinical Chemistry, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand; [d]IRCCS, Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy; [e]Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand

ABSTRACT

P-glycoprotein (Pgp) inhibition has been considered as an effective strategy towards combating multidrug-resistant cancers. Owing to the substrate promiscuity of Pgp, the classification of its interacting ligands is not an easy task and is an ongoing issue of debate. Chemical structures can be represented by the simplified molecular input line entry system (SMILES) in the form of linear string of symbols. In this study, the SMILES notations of 2254 Pgp inhibitors including 1341 active, and 913 inactive compounds were used for the construction of a SMILE-based classification model using CORrelation And Logic (CORAL) software. The model provided an acceptable predictive performance as observed from statistical parameters consisting of accuracy, sensitivity and specificity that afforded values greater than 70% and MCC value greater than 0.6 for training, calibration and validation sets. In addition, the CORAL method highlighted chemical features that may contribute to increased and decreased Pgp inhibitory activities. This study highlights the potential of CORAL software for rapid screening of prospective compounds from a large chemical space and provides information that could aid in the design and development of potential Pgp inhibitors.

ARTICLE HISTORY
Received 7 June 2016
Accepted 21 November 2016

KEYWORDS
P-glycoprotein inhibitors; CORAL software; classification model; SMILES; multidrug resistance; pharmacokinetic

Introduction

Considerable attention has been given to P-glycoprotein (Pgp) transporter due to its clinical impacts on multidrug resistance and pharmacokinetic profiles of substrate drugs [1]. Human Pgp is a protein belonging to the ATP binding cassette (ABC) family [2], which is encoded by multidrug resistance genes (i.e. MDR1) [2]. Pgp is a 170 kD membrane-bound polypeptide [3] comprising 1280 amino acids [4]. Pgp functions as an efflux pump to extrude a wide range of structurally diverse hydrophobic substances out of the cell [5,6]. Due to its expression in many physical barriers [2] and pharmacokinetic-related [7] organs, it plays a role in limiting cellular uptake, distribution, excretion and toxicity of many xenobiotics and toxic substances

CONTACT A. Worachartcheewan ✉ apilak.woa@mahidol.ac.th
Supplemental data for this article can be accessed here

[5]. In addition, it influences the pharmacokinetic or ADMET (A = absorption, D = distribution, M = metabolism, E = excretion and T = toxicity) profiles of its substrate drugs [7,8]. Pgp is considered as a contributing factor of multidrug resistance [5,7,9] on account of many anticancer drugs being substrates of Pgp [10]. Furthermore, Pgp overexpression is found in many types [11] and various stages [7] of cancer cells. The overexpression increases efflux activity, thereby impairing the delivery of anticancer agents from reaching their target sites [5,7,9]. Although structurally unrelated, the broad specificity of Pgp allows a wide range of anticancer drugs to be recognized and extruded out of cancer cells [5,9]. This phenomenon leads to a simultaneous resistance to a number of structurally and functionally unrelated anticancer agents called multidrug resistance [7]. Hence, an inhibition of this efflux pump is one of the strategies geared towards improving the pharmacokinetics of drugs as well as combating multidrug resistance [5]. In this regard, the development of novel Pgp inhibitors for therapeutic applications is an active research area gaining much attention [1,12]. Currently, many classes of Pgp inhibitors have been developed from diverse types of compounds and natural products [4,13–15]. However, many aspects of Pgp and its interacting compounds need to be elucidate for the development of Pgp inhibitors which have a desired treatment outcome [16].

Pgp is one of the most studied transporters, due to its promiscuity [17,18]. The presence of multiple binding sites together with its broad specific recognition allows non-specific and simultaneous binding of hydrophobic compounds [6]. In addition, many available experimental assays use different criteria to classify Pgp-interacting compounds, which leads to conflicting reports of their endpoints [19,20]. The classification of Pgp ligands is not straightforward and the issue is still under debate [21]. In this regard, many computational methods have been employed in an attempt to understand this transporter, such as quantitative structure–activity relationship (QSAR) [22–26], classification structure–property relationship (CSPR) [27–34], molecular docking [22,23,35] and homology modelling [36].

The molecular structure of a compound can be represented by simplified molecular input line entry system (SMILES) notations [37–39]. SMILES is a chemical language designed for a human/machine interface [39]. In a computational aspect, SMILES can be interpreted in a fast and compact manner, thereby significantly saving time and space [39]. To date, SMILES is believed to be the best compromise between the human and machine aspect of chemical notation. In addition, its ability to facilitate information processing beyond the conventional methods has been well documented [37,39–42]. The use of SMILES for the development of QSAR/CSPR models can help avoid general problems of using molecular descriptors, such as the selection of an appropriate subset of informative descriptors from a large available set of descriptors and the interpretation of important descriptors obtained from the constructed models [40–43]. Moreover, the use of the SMILES notation can greatly conserve time, as geometrical optimization is not required for descriptor calculations [43–45]. CORrelation And Logic (CORAL) software (http://www.insilico.eu/coral) is a computational tool for conformation-independent QSAR/CSPR analysis. Herein, the SMILES notations are used instead of molecular descriptors, which need to be calculated from optimized structures [38,43]. The CORAL software has been successfully employed for predicting diverse types of compounds and biological activities, including anticancer [43,46–48], antiviral [49,50], antimalarial [51,52] and toxicity of compounds [53,54].

To date, the classification model constructed by CORAL software has not been reported for Pgp inhibitors. In this study, a CORAL software based on Monte Carlo technique was employed to construct a classification model from 2254 compounds (1341 active, 913 inactive).

In addition, the SMILES attributes influencing the Pgp inhibitory effects of the compounds were highlighted to provide a simple and rapid screening of potential Pgp inhibitors.

Materials and methods

Data set

A set of compounds in the form of SMILES notations together with their endpoints (i.e. active or inactive) were retrieved from the admetSAR database [55]. The Pgp inhibitors in the database were obtained from the original studies in which the details regarding experimental assays and classification criteria are provided in the supplementary information. According to their heterogeneity, uniform SMILES notations were created from the original SMILES using ChemAxon Marvin Sketch version 6.3.1 command line [28]. Subsequently, all newly generated SMILES along with their endpoints were combined and sorted in a single Excel worksheet.

Lipinski's rule of five is a set of *in silico* guidelines used to prioritize compounds with high oral absorption. The rule of five has been used to reduce attrition rate arising due to poor pharmacokinetic properties. However, recent research has revealed that strictly following the rule of five may lead to a loss of opportunity in finding potential compounds acting on difficult targets [56]. Moreover, exceptions to the rule of five have been demonstrated in many recently approved drugs [56], especially natural products [57] and peptides [56]. The chemical spaces beyond the rule of five are noted for a high degree of diversity, complexity and novelty; however, these spaces have not yet been explored [56]. Therefore, the current trend in drug discovery has moved towards exploring beyond the chemical spaces of the rule of five [58,59]. Current research suggested that the oral druggable space has been extended from the rule of five, in which the limited MW has been extended from ≤500 Da to ≤1000 Da [56]. In addition, the compounds in the expanded rule of five spaces are suggested to interact with the Pgp efflux (either as substrates or inhibitors) [56]. However, 1000 Da is used as the upper limit due to the ability of the drugs to move or distribute within the body, since large drugs with a molecular size greater than 1000 Da are not readily diffused across physiological membranes or barriers [60]. In regards to the above context, compounds with a MW ≥ 1000 Da were discarded from this study. Furthermore, redundant compounds and overlapping classified compounds which were categorized as belonging to more than one class were identified and excluded from this study. This produced a final set of 1341 active and 913 inactive compounds that were used for model construction as summarized in the schematic workflow shown in Figure 1. Briefly, the 2254 available compounds were randomly split into training (≈50%), calibration (≈25%), and validation (≈25%) sets. The training set was used for constructing the model correlating to binary endpoints (i.e. active = 1, and inactive = 0) with their respective molecular descriptors. Next, the calibration set was used to avoid overtraining, in which the optimization process is stopped when the correlation coefficient between optimal descriptor and endpoint (Figure 2(b)) decreases. Lastly, the external validation set was used for the final evaluation of the predictive potential of the model.

CORAL method

The SMILES attributes represented their chemical elements as symbols in the structure such as cycles and branching, and were employed for model development using CORAL software.

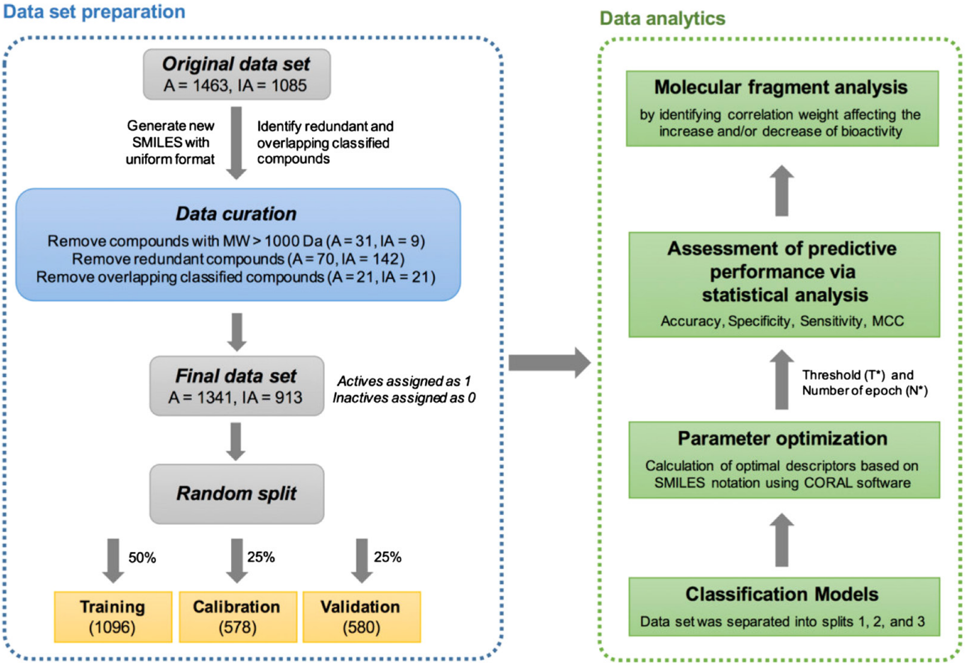

Data set preparation
Original data set
A = 1463, IA = 1085
Generate new SMILES with uniform format
Identify redundant and overlapping classified compounds
Data curation
Remove compounds with MW > 1000 Da (A = 31, IA = 9)
Remove redundant compounds (A = 70, IA = 142)
Remove overlapping classified compounds (A = 21, IA = 21)
Final data set
A = 1341, IA = 913
Actives assigned as 1
Inactives assigned as 0
Random split
50%
25%
25%
Training (1096)
Calibration (578)
Validation (580)
Data analytics
Molecular fragment analysis
by identifying correlation weight affecting the increase and/or decrease of bioactivity
Assessment of predictive performance via statistical analysis
Accuracy, Specificity, Sensitivity, MCC
Threshold (T*) and Number of epoch (N*)
Parameter optimization
Calculation of optimal descriptors based on SMILES notation using CORAL software
Classification Models
Data set was separated into splits 1, 2, and 3

Figure 1. Schematic workflow of this study.

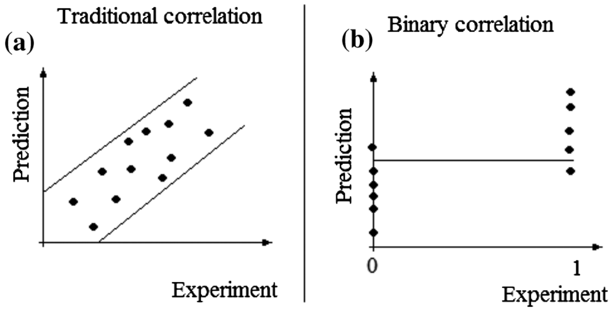


Figure 2. Classification model based on binary correlations.

CORAL software is a free program available at http://www.insilico.eu/coral. It is a computational tool for constructing regression [43,61,62] and classification [48,63] models based on Monte Carlo methods in which SMILES of the compounds together with their endpoints (i.e. active or inactive) are used as input data. Initially, all compounds were randomly divided into three data sets: training (50%), calibration (25%) and validation (25%) sets. The first two subsets served as the internal validation, while the latter set served as the external validation and, as such, can be used to evaluate the interpolation and extrapolation capability of the predictive model, respectively.

The classification models were calculated by CORAL software using the same algorithms. Regression models were calculated based on true correlations, whereas that of classification models were based on pseudo correlations (Supplementary Figure S1, available via the

Supplementary Content tab on the article's online page). The data set for the classification models was assigned as one of two possible activity values (1 = active or 0 = inactive).

The optimal descriptors based on SMILES notations were calculated according to the following Equation (1)

$$DCW(Threshold, N_{\mathrm{epoch}}) = \sum CW(S_k) + \sum CW(SS_k) \tag{1}$$

where S and SS are one-atom and two-atom fragments of SMILES. The atom of SMILES is defined as one symbol or two symbols which cannot be examined separately (e.g. Cl, Br, etc.). The $CW(S_k)$ and $CW(SS_k)$ attributes are correlation weights that give maximal correlation between the optimal descriptor and the binary endpoint for a training set. The threshold is a parameter for discriminating SMILES fragments into two classes: (i) rare, if the number of SMILES that contains this fragment in the training set is less than the threshold, and (ii) non-rare, if this number is larger than the threshold. The epoch of the Monte Carlo optimization is one cycle of modification of all correlation weights. The numerical data on the correlation weights were obtained via the Monte Carlo method. The correlation weights give maximal values of the correlation coefficient between the endpoint (0, 1) and $DCW\,(Threshold, N_{epoch})$. The pseudo regression (binary) model is calculated by the least squares method according to the following Equation (2)

$$Model = C_0 + C_1 \times DCW(Threshold, N_{\mathrm{epoch}}) \tag{2}$$

For an active compound, the true positive (TP) value is assigned if the model > 0.5, whereas a false negative (FN) value is assigned if the model ≤ 0.5. For an inactive compound, a true negative (TN) value is assigned for the model ≤ 0.5, whereas a false positive (FP) value is assigned if the model > 0.5.

Statistical assessment of predictive models

A set of statistical parameters for evaluating the predictive performance of the constructed models included accuracy, sensitivity, specificity and Matthews Correlation Coefficient (MCC), as shown in Equations (3)–(6) [63,64].

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{3}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{4}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{5}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives or over-predictions and FN is the number of false negatives or missed predictions. MCC coefficient values of −1, 0 and +1 indicate total disagreement between the prediction and observation, no better than random prediction and a perfect prediction, respectively. In addition, Y-randomization text [64] has been utilized to validate the predictive potential of the suggested models.

Results and discussion

CORAL classification

A set of 2254 compounds (1341 active and 913 inactive) were randomly divided into training (1096), calibration (578) and validation (580) sets. Initially, parameter optimization was performed to obtain preferable threshold and number of epochs (N_{epoch}) [43]. The data set was split into three splits (split 1, split 2 and split 3) and was calculated with N_{epoch} of 20 and threshold values in the range of 1–3. In addition, a start value of D_{start} = 0.5, precision = 0.1 and the number of probes (optimization) = 3 was used. For all splits, the preferable threshold (T*) and number of epochs (*N**) of the Monte Carlo optimization were kept at *T** = 1 and *N** = 3, respectively. The models for splits 1, 2 and 3 were obtained as follows:

Split 1

$$Inhibitory\ activity = -0.3296(\pm 0.0009) + 0.007760(\pm 0.000007) * DCW(1,3) \quad (7)$$

Split 2

$$Inhibitory\ activity = -0.3341(\pm 0.0009) + .008308(\pm .000007) * DCW(1,5) \quad (8)$$

Split 3

$$Inhibitory\ activity = -0.3227(\pm 0.0009) + 0.008667(\pm 0.000007) * DCW(1,3) \quad (9)$$

Classification models described by Equations (7)–(9) define a compound as active if its inhibitory activity > 0.5 and as inactive if its inhibitor activity ≤ 0.5. In other words, the classification into the active or inactive category (i.e. according to the calculated inhibitory activity with Equations (7), (8) or (9)) is defined by the formula:

$$Class = \begin{cases} Inhibitoractivity > 0.5, class = 1(active) \\ inhibiroractivity \leq 0.5, class = 0(inactive) \end{cases} \quad (10)$$

Detailed results of the classification model for external validation sets of split 1–3 are provided in Supplementary Tables S2–S4 (available online). In addition, the statistical characteristics of the classification model for three random splits along with three runs of the Monte Carlo optimization are displayed in Table 1. The results showed a high accuracy and sensitivity of greater than 80% and specificity of greater than 70% in splits 1–3 for training, calibration and validation sets.

It should be noted that the classification model is based on special binary correlations (Figure 2(b)) where some traditional criteria cannot be used. Figure 2(a) gives a graphical illustration for this point. In the case of traditional correlation, if $r^2 \approx 0.4$ then the regression model is deemed to be poor. However, in the case of binary correlation with $r^2 \approx 0.4$, the specificity, sensitivity, accuracy and MCC may be considered as satisfactory or even excellent

Table 1. Statistical characteristics of the classification models of inhibitor activity for three random splits in three runs of the Monte Carlo method optimization.

Split	*Run*	*Statistical characteristics*	*Training set*	*Calibration set*	*Validation set*
1		**Total**	**1096**	**578**	**580**
	1	TP	561	307	304
		TN	335	177	174
		FP	112	53	62
		FN	88	41	40
		Sensitivity	0.8644	0.8822	0.8837
		Specificity	0.7494	0.7696	0.7373
		Accuracy	0.8175	0.8374	0.8241
		MCC	0.6197	0.6582	0.6322
		Y-randomization*	0.5175	0.5591	–
	2	TP	566	307	303
		TN	332	177	167
		FP	115	53	69
		FN	83	41	41
		Sensitivity	0.8721	0.8822	0.8808
		Specificity	0.7427	0.7696	0.7076
		Accuracy	0.8193	0.8374	0.8103
		MCC	0.6229	0.6582	0.6027
		Y-randomization	0.5179	0.5678	–
	3	TP	560	307	304
		TN	334	178	168
		FP	113	52	68
		FN	89	41	40
		Sensitivity	0.8629	0.8822	0.8837
		Specificity	0.7472	0.7739	0.7119
		Accuracy	0.8157	0.8391	0.8138
		MCC	0.6159	0.6620	0.6101
		Y-randomization	0.5158	0.5668	–
2		***Total***	**1159**	**536**	**559**
	1	TP	627	271	287
		TN	371	167	173
		FP	97	55	50
		FN	64	43	49
		Sensitivity	0.9074	0.8631	0.8542
		Specificity	0.7927	0.7523	0.7758
		Accuracy	0.8611	0.8172	0.8229
		MCC	0.7094	0.6209	0.6304
		Y-randomization	0.5951	0.5236	–
	2	TP	625	269	286
		TN	368	169	171
		FP	100	53	52
		FN	66	45	50
		Sensitivity	0.9045	0.8567	0.8512
		Specificity	0.7863	0.7613	0.7668
		Accuracy	0.8568	0.8172	0.8175
		MCC	0.7003	0.6215	0.6190
		Y-randomization	0.5869	0.5229	–
	3	TP	629	271	289
		TN	363	167	171
		FP	105	55	52
		FN	62	43	47
		Sensitivity	0.9103	0.8631	0.8601
		Specificity	0.7756	0.7523	0.7668
		Accuracy	0.8559	0.8172	0.8229
		MCC	0.6984	0.6209	0.6294
		Y-randomization	0.5866	0.5225	–
3		***Total***	**1134**	**555**	**565**
	1	TP	621	268	305
		TN	335	195	162

(Continued)

Table 1. (*Continued*).

Split	*Run*	*Statistical characteristics*	*Training set*	*Calibration set*	*Validation set*
		FP	89	54	58
		FN	69	38	40
		Sensitivity	0.9000	0.8758	0.8841
		Specificity	0.7995	0.7831	0.7364
		Accuracy	0.8607	0.8342	0.8265
		MCC	0.7057	0.6641	0.6312
		Y-randomization	0.5801	0.5529	–
	2	TP	624	269	301
		TN	354	191	161
		FP	90	58	59
		FN	66	37	44
		Sensitivity	0.9043	0.8791	0.8725
		Specificity	0.7973	0.7671	0.7318
		Accuracy	0.8624	0.8288	0.8177
		MCC	0.7092	0.6532	0.6128
		Y-randomization	0.5851	0.5420	–
	3	TP	625	272	299
		TN	352	193	162
		FP	92	56	58
		FN	65	34	46
		Sensitivity	0.9058	0.8889	0.8667
		Specificity	0.7928	0.7751	0.7364
		Accuracy	0.8616	0.8378	0.8159
		MCC	0.7071	0.6717	0.6097
		Y-randomization	0.5812	0.5614	–

*A model has predictive potential if the Y-randomization test is larger than 0.5 [64].

(i.e. the majority of substances shown in the diagram are classified correctly, with only one false positive and one false negative).

Furthermore, chemical elements involving increased and decreased activities together with their correlation weights and frequencies are shown in the supplementary information (Table S1, available online). The correlation weight values indicated the class of the compound as follows: (i) if SMILES attribute is characterized by a stable positive value, the attribute can be examined as a promoter for active Pgp inhibitory activity of the compound; (ii) if SMILES attribute is characterized by a stable negative value, the attribute can be examined as a promoter for inactive activity of the compound; (iii) attribute with an unclear role (attributes which are characterized by both positive and negative values); and (iv) blocked attributes (absent in the training set). The possible promoters that are involved in increased and decreased activities are presented in Table 2. All technical details related to three splits examined in this work are available on request.

Our results indicated that a large carbon skeleton, branched structure, biphenyl ring, naphthalene ring, and a ring containing nitrogen (N) atom as key influential features for increased inhibitory activity. In contrast, the presence of sulphur (S), oxygen (O)-containing ring, O atom together with double bond and five rings were noted for decreased activity. Unfortunately, some part of SMILES attribute have no physical interpretation (e.g. [...(.......; [...-.......; \...3.......; etc.). In addition, in order to infer the correct interpretation, the frequency of the attribute obtained from training and test sets should be taken into account.

Challenging issues

Despite being one of the most studied proteins, many aspects of this drug transporter are still not fully understood and are currently being debated [21,35]. Great challenges arise

Table 2. List of SMILES attributes which can be interpreted as promoters of activity or inactivity of compounds.

Role	*SMILES attribute*	*Interpretation*
Promoter of activity increase	C...........	Carbon (large carbon skeleton)
	C...(.......	branching
	C...1.......	Presence ring
	C...2.......	Presence of two rings (biphenyl)
	c...2.......	Presence of two rings (naphthalene)
	N...1.......	Presence of ring with nitrogen
Promoter of activity decrease	N...........	Presence of nitrogen
	O...1.......	Presence of oxygen in ring
	O...=.......	Presence of oxygen and double bond
	S...........	Presence of sulphur
	c...5.......	Presence of five rings

mainly due to the lack of a human crystallographic structure, the highly flexible and dynamic binding nature of the transporter, the presence of multiple substrate binding sites and the polyspecificity of substrate and inhibitor recognition [65,66]. In addition, the mechanisms of transport and inhibition of this protein are still not clearly understood [35].

Pgp is a dynamic efflux pump. A transport cycle of the Pgp is strongly involved with a switch between two conformations of the pump, that is, inward-facing (a conformation that prevents extrusion of substrate) and outward-facing (a conformation that allows extrusion of substrate from the cell) conformations [6,67]. Briefly, a lipophilic substrate is passively diffused through the lipid membrane to reach its binding site. The binding of substrate induces ATP binding, ATP hydrolysis and dimerization of ATP domains, giving rise to conformational changes of the transporter into a form preferable for substrate extrusion. After the substrate is extruded, another ATP binds and ATP hydrolysis occurs to reset the pump into the inward-facing form [6,67]. The dynamic nature of Pgp and the key role of ATP hydrolysis in its transport function have been highlighted by molecular dynamics. Molecular dynamic studies indicated that the conformational changes are driven by ATP hydrolysis, which occurs after ATP binding and upon substrate extrusion [68–72]. The dynamic nature of Pgp means that it is a research area with continual progress; however, the lack of a crystallographic structure of human Pgp is a limitation that hinders the reliability of studies [65].

Pgp inhibition is one of the strategies for achieving therapeutic activity [1,5]. However, the inhibition of Pgp is dynamic, and can be achieved by alteration of many steps of the Pgp transport cycle. Typically, the efflux function of Pgp can be inhibited by three main mechanisms: (i) blocking the drug binding site (either competitively or non-competitively/allosterically) [73], (ii) alteration of ATP hydrolysis [74] and (iii) alteration of lipid membrane integrity [75]. Competitive inhibitors are also Pgp substrates which compete with substrate drugs for transport by interacting at drug binding sites [12]. In contrast, non-competitive inhibitors bind at different sites (modulatory sites), thereby preventing translocation and dissociation of the substrate [35,76]. In addition, pharmaceutical excipients (i.e. surfactants) which are capable of interacting with lipid membranes exhibit an indirect Pgp inhibitory effect via an alteration of lipid membrane integrity [12]. Various mechanisms of inhibition, along with different conditions and judging criteria of available experimental assays, lead to a conflict in reports and unclear classification of Pgp inhibitors. In this regard, extensive studies are essential for the development of novel Pgp inhibitors with desired activity.

Important features required for Pgp inhibition

Besides the complex mechanisms of transport and inhibition, the ambiguous classification of Pgp inhibitors is due to its broad-specific nature. Furthermore, a compound's lipophilicity is the only requirement for Pgp recognition and interaction in the initial step of its transport cycle [12]. Moreover, the substrate-binding sites are large and flexible enough to be occupied by more than one substrate simultaneously [6,9,67]. This simple requirement, together with the presence of multiple highly flexible substrate-binding sites, gives rise to the promiscuity of Pgp [6,9,67]. In addition, polyspecific recognition of Pgp is a factor hindering the rational design of inhibitors, and therefore is an area of great interest [65]. To our knowledge, the majority of Pgp-interacting compounds share common structural and physicochemical properties (i.e. lipophilicity, molecular size, flexibility, bulkiness and aromaticity) [20,31,77,78]. However, the definite essential structural features and properties required for recognition, transport and inhibition of Pgp have not yet been clearly understood [20].

Structure–activity relationships (SAR) and *in silico* studies have suggested essential chemical properties required for potent Pgp inhibitors [79] such as lipophilicity [80–86], molecular size [31,87] and bulkiness [31,88], aromaticity [88–90] and hydrogen bond acceptor [84,85,89–92]. Hydrogen bond interaction was noted to play important roles in substrate recognition, transport and release of Pgp ligands [93]. In addition, hydrophobic [23], π-stacking [23,94] and electrostatic [84,95,96] interactions are also noted for their roles. High lipophilicity of a compound is essential for partitioning through the lipid membrane in order to access the Pgp binding sites. In addition, the presence of hydrogen bond acceptors and/or hydrogen bond donors, tertiary N atom, aromatic ring and multiple hydrophobic regions within the molecules is required for H-bonding, electrostatic, π-stacking and hydrophobic interactions, respectively [79].

According to our active promoters (Table 2), the role of each promoter is discussed with regard to the key chemical properties. The presence of a large carbon skeleton, branched structure and rings may indicate a large molecular size and bulkiness. In addition, molecular weight (MW) is a good representative of the molecular size of a compound [78]. MW has been noted for its large influence on Pgp transport [97] and inhibition [31]. Furthermore, the bulkiness of the compound is highly correlated to MW because a larger molecule contains more rings and branched structures [31]. Moreover, the presence of a bulky aromatic system has been designated with a strong inhibitory effect [88]. In addition, planar aromatic rings (i.e. biphenyl and naphthalene rings) containing π-bonds may facilitate the formation of π–π stacking interactions with Pgp. Lipophilicity is a key physicochemical property relating to membrane permeability [98]. Interaction with lipid membrane is considered as a primary criteria for substrate recognition by Pgp [12] and multiple hydrophobic features are required for hydrophobic interactions [79]. In this regard, a complicated ring system may govern the access and recognition by Pgp and/or may play a role in the alteration of lipid membrane integrity as it affects molecular size, shape and lipophilicity of the compound. In addition, N atom is considered to be a hallmark atom of Pgp inhibitors [99], and its role as a hydrogen bond acceptor for H-bonding interaction has been also highlighted [89]. Therefore, the presence of nitrogen-containing ring structures may improve the capability of the compound to form H-bonds as well as electrostatic interactions. Furthermore, the presence of a tertiary amine may be essential for strengthening the binding via electrostatic interactions with Pgp [100].

Predictive models are only beneficial if they can be used for reliably predicting the activity of all compounds of interest. However, this is not always the case, as models are only applicable for a limited number of compounds that are structurally similar to those that have been used to train the model. Thus, it is essential to assess whether compounds belonging to various data subsets are within the applicability domain or not. Therefore, Supplementary Tables S2–S4 (available online) describe for each split whether compounds from the training, calibration and validation sets fall in the applicability domain or not. The criterion used for defining the applicability domain via the so-called defects of SMILES has been described previously in the literature [101].

There are two categories of QSAR models: (i) the regression models, which aim to predict numerical values of an endpoint for different substances; and (ii) the classification models, which aim to predict, in the form of Yes/No, various abilities, such as toxicity [102], fungicide activity [103], anticancer activity [104], etc. The CORAL software enables the possibility to build up both mentioned categories of the QSAR models [48,64].

Conclusions

The CORAL method was successfully employed in the construction of a SMILES-based classification model for classifying 2254 compounds as Pgp inhibitors or non-inhibitors with acceptable predictive performances. The SMILES-based classification model provided distinct aspects from the conventional methods in which the modelling can be performed without the need of fastidious processes (i.e. structural geometrical optimization, descriptor calculation and descriptor selection). Furthermore, the CORAL method allows the discovery of important chemical features (as represented by SMILES attributes) that may contribute to increased and decreased Pgp inhibitory activities of the compounds. Finally, this study highlights, for the first time, the potential usage of CORAL software for classifying Pgp inhibitors from a large available library. The simplicity of the CORAL software could benefit the prompt screening of potential compounds, as it can minimize the chemical space into a manageable size for further investigation and development.

Funding

This work was supported by the Annual Budget Grant (2556-2558 B.E.) from Mahidol University; the Mahidol University Talent Management Program under Grant number TM60/2557; the Research Career Development Grant from the Thailand Research Fund under Grant number RSA5780031; the Office of the Higher Education Commission and Mahidol University under the National Research Universities Initiative (2558 B.E.); and the EU project PROSIL funded under the LIFE program (project LIFE12 ENV/IT/000154).

ORCID

V. Prachayasittikul http://orcid.org/0000-0001-6338-3721
A. Worachartcheewan http://orcid.org/0000-0003-3021-3632
A. P. Toropova http://orcid.org/0000-0002-4194-9963
A. A. Toropov http://orcid.org/0000-0001-6864-6340
N. Schaduangrat http://orcid.org/0000-0002-0842-8277
V. Prachayasittikul http://orcid.org/0000-0001-7942-1083
C. Nantasenamat http://orcid.org/0000-0003-1040-663X

References

[1] V. Prachayasittikul and V. Prachayasittikul, *P-glycoprotein transporter in drug development*, EXCLI J. 15 (2016), pp. 113–118.

[2] O. Fardel, E. Kolasa, and M. Le Vee, *Environmental chemicals as substrates, inhibitors or inducers of drug transporters: Implication for toxicokinetics, toxicity and pharmacokinetics*, Expert Opin. Drug Metab. Toxicol. 8 (2012), pp. 29–46.

[3] R.L. Juliano and V. Ling, *A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants*, Biochim. Biophys. Acta 445 (1976), pp. 152–162.

[4] T. Bansal, M. Jaggi, R.K. Khar, and S. Talegaonkar, *Emerging significance of flavonoids as P-glycoprotein inhibitors in cancer chemotherapy*, J. Pharm. Pharm. Sci. 12 (2009), pp. 46–78.

[5] M.L. Amin, *P-glycoprotein inhibition for optimal drug delivery*, Drug Target Insights 7 (2013), pp. 27–34.

[6] S.V. Ambudkar, I.W. Kim, and Z.E. Sauna, *The power of the pump: Mechanisms of action of P-glycoprotein (ABCB1)*, Eur. J. Pharm. Sci. 27 (2006), pp. 392–400.

[7] R. Krishna and L.D. Mayer, *Multidrug resistance (MDR) in cancer. Mechanisms, reversal using modulators of MDR and the role of MDR modulators in influencing the pharmacokinetics of anticancer drugs*, Eur. J. Pharm. Sci. 11 (2000), pp. 265–283.

[8] J.H. Lin and M. Yamazaki, *Role of P-glycoprotein in pharmacokinetics: Clinical implications*, Clin. Pharmacokinet. 42 (2003), pp. 59–98.

[9] M. Hennessy and J.P. Spiers, *A primer on the mechanics of P-glycoprotein the multidrug transporter*, Pharmacol. Res. 55 (2007), pp. 1–15.

[10] F.J. Sharom, *The P-glycoprotein multidrug transporter*, Essays Biochem. 50 (2011), pp. 161–178.

[11] D. Drach, S. Zhao, J. Drach, and M. Andreeff, *Low incidence of MDR1 expression in acute promyelocytic leukaemia*, Br. J. Haematol. 90 (1995), pp. 369–374.

[12] K.M.R. Srivalli and P.K. Lakshmi, *Overview of P-glycoprotein inhibitors: A rational outlook*, Braz. J. Pharm. Sci. 48 (2012), pp. 353–367.

[13] D. Balayssac, N. Authier, A. Cayre, and F. Coudore, *Does inhibition of P-glycoprotein lead to drug-drug interactions?*, Toxicol. Lett. 156 (2005), pp. 319–329.

[14] D. Lopez and S. Martinez-Luis, *Marine natural products with P-glycoprotein inhibitor properties*, Mar. Drugs 12 (2014), pp. 525–546.

[15] H.M. Abdallah, A.M. Al-Abd, R.S. El-Dine, and A.M. El-Halawany, *P-glycoprotein inhibitors of natural origin as potential tumor chemo-sensitizers: A review*, J. Adv. Res. 6 (2015), pp. 45–62.

[16] G. Szakács, J.K. Paterson, J.A. Ludwig, C. Booth-Genthe, and M.M. Gottesman, *Targeting multidrug resistance in cancer*, Nat. Rev. Drug Discov. 5 (2006), pp. 219–234.

[17] M.M. Gottesman, I. Pastan, and S.V. Ambudkar, *P-glycoprotein and multidrug resistance*, Curr. Opin. Genet. Dev. 6 (1996), pp. 610–617.

[18] M.M. Gottesman, T. Fojo, and S.E. Bates, *Multidrug resistance in cancer: Role of ATP-dependent transporters*, Nat. Rev. Cancer 2 (2002), pp. 48–58.

[19] F.J. Sharom, *The P-glycoprotein efflux pump: How does it transport drugs?*, J. Membr. Biol. 160 (1997), pp. 161–175.

[20] A. Seelig, *A general pattern for substrate recognition by P-glycoprotein*, Eur. J. Biochem. 251 (1998), pp. 252–261.

[21] Y.H. Wang, Y. Li, S.L. Yang, and L. Yang, *Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach*, J. Chem. Inf. Model. 45 (2005), pp. 750–757.

[22] P.H. Palestro, L. Gavernet, G.L. Estiu, and L.E. Bruno Blanch, *Docking applied to the prediction of the affinity of compounds to P-glycoprotein*, BioMed. Res. Int. 2014 (2014), pp. 358–425.

[23] M. Ghandadi, A. Shayanfar, M. Hamzeh-Mivehroud, and A. Jouyban, *Quantitative structure activity relationship and docking studies of imidazole-based derivatives as P-glycoprotein inhibitors*, Med. Chem. Res. 23 (2014), pp. 4700–4712.

[24] J. Shen, Y. Cui, J. Gu, Y. Li, and L. Li, *A genetic algorithm-back propagation artificial neural network model to quantify the affinity of flavonoids toward P-glycoprotein*, Comb. Chem. High Throughput Screen. 17 (2014), pp. 162–172.

[25] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, and V. Prachayasittikul, *A practical overview of quantitative structure-activity relationship*, EXCLI J. 8 (2009), pp. 74–88.

[26] C. Nantasenamat, C. Isarankura-Na-Ayudhya, and V. Prachayasittikul, *Advances in computational methods to predict the biological activity of compounds*, Expert Opin. Drug Discov. 5 (2010), pp. 633–654.

[27] J. Levatić, J. Ćurak, M. Kralj, T. Šmuc, M. Osmak, and F. Supek, *Accurate models for P-gp drug recognition induced from a cancer cell line cytotoxicity screen*, J. Med. Chem. 56 (2013), pp. 5691–5708.

[28] F. Klepsch, P. Vasanthanathan, and G.F. Ecker, *Ligand and structure-based classification models for prediction of P-glycoprotein inhibitors*, J. Chem. Inf. Model. 54 (2014), pp. 218–229.

[29] Z. Wang, Y. Chen, H. Liang, A. Bender, R.C. Glen, and A. Yan, *P-glycoprotein substrate models using support vector machines based on a comprehensive data set*, J. Chem. Inf. Model. 51 (2011), pp. 1447–1456.

[30] J.E. Penzotti, M.L. Lamb, E. Evensen, and P.D.J. Grootenhuis, *A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein*, J. Med. Chem. 45 (2002), pp. 1737–1740.

[31] L. Chen, Y. Li, Q. Zhao, H. Peng, and T. Hou, *ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques*, Mol. Pharm. 8 (2011), pp. 889–900.

[32] D. Li, L. Chen, Y. Li, S. Tian, H. Sun, and T. Hou, *ADMET evaluation in drug discovery. 13. Development of in silico prediction models for P-glycoprotein substrates*, Mol. Pharm. 11 (2014), pp. 716–726.

[33] M. Adenot and R. Lahana, *Blood-brain barrier permeation models: Discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 239–248.

[34] V. Prachayasittikul, A. Worachartcheewan, W. Shoombuatong, V. Prachayasittikul, and C. Nantasenamat, *Classification of P-glycoprotein-interacting compounds using machine learning methods*, EXCLI J. 14 (2015), pp. 958–970.

[35] M. Zeino, M.E.M. Saeed, O. Kadioglu, and T. Efferth, *The ability of molecular docking to unravel the controversy and challenges related to P-glycoprotein – A well-known, yet poorly understood drug transporter*, Invest. New Drugs 32 (2014), pp. 618–625.

[36] H. Yamaguchi, Y. Kidachi, K. Kamiie, T. Noshita, and H. Umetsu, *Homology modeling and structural analysis of human P-glycoprotein*, Bioinformation 8 (2012), pp. 1066–1074.

[37] A.P. Toropova, A.A. Toropov, and E. Benfenati, *A quasi-QSPR modelling for the photocatalytic decolourization rate constants and cellular viability (CV%) of nanoparticles by CORAL*, SAR QSAR Environ. Res. 26 (2015), pp. 29–40.

[38] A.A. Toropov, A.P. Toropova, and E. Benfenati, *Additive SMILES-based carcinogenicity models: Probabilistic principles in the search for robust predictions*, Int. J. Mol. Sci. 10 (2009), pp. 3106–3127.

[39] D. Weininger, *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*, J. Chem. Inf. Comput. Sci. 28 (1988), pp. 31–36.

[40] P.G. Achary, *Simplified molecular input line entry system-based optimal descriptors: QSAR modelling for voltage-gated potassium channel subunit Kv7.2*, SAR QSAR Environ Res 25 (2014), pp. 73–90.

[41] P.G.R. Achary, *QSPR modelling of dielectric constants of π-conjugated organic compounds by means of the CORAL software*, SAR QSAR Environ. Res. 25 (2014), pp. 507–526.

[42] S. Begum and P.G.R. Achary, *Simplified molecular input line entry system-based: QSAR modelling for MAP kinase-interacting protein kinase (MNK1)*, SAR QSAR Environ. Res. 26 (2015), pp. 343–361.

[43] A. Worachartcheewan, P. Mandi, V. Prachayasittikul, A.P. Toropova, A.A. Toropov, and C. Nantasenamat, *Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors*, Chemometr. Intell. Lab. Syst. 138 (2014), pp. 120–126.

[44] J.B. Veselinovic, G.M. Nikolic, N.V. Trutic, J.V. Zivkovic, and A.M. Veselinovic, *Monte Carlo QSAR models for predicting organophosphate inhibition of acetycholinesterase*, SAR QSAR Environ. Res. 26 (2015), pp. 449–460.

[45] M.A. Turabekova, B.F. Rasulev, F.N. Dzhakhangirov, A.A. Toropov, D. Leszczynska, and J. Leszczynski, *Aconitum and delphinium diterpenoid alkaloids of local anesthetic activity: Comparative QSAR*

analysis based on GA-MLRA/PLS and optimal descriptors approach, J. Environ. Sci. Health C, Environ. Carcinog. Ecotoxicol. Rev. 32 (2014), pp. 213–238.

[46] Q. Li, X. Ding, H. Si, and H. Gao, *QSAR model based on SMILES of inhibitory rate of 2, 3-diarylpropenoic acids on AKR1C3*, Chemometr. Intell. Lab. Syst. 139 (2014), pp. 132–138.

[47] L.M.A. Mullen, P.R. Duchowicz, and E.A. Castro, *QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents*, Chemometr. Intell. Lab. Syst. 107 (2011), pp. 269–275.

[48] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, D. Leszczynska, and J. Leszczynski, *CORAL: Classification model for predictions of anti-sarcoma activity*, Curr. Top. Med. Chem. 12 (2012), pp. 2741–2744.

[49] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, and V. Prachayasittikul, *QSAR study of H1N1 neuraminidase inhibitors from influenza a virus*, Lett. Drug Des. Discov. 11 (2014), pp. 420–427.

[50] A.P. Toropova, A.A. Toropov, J.B. Veselinović, F.N. Miljković, and A.M. Veselinović, *QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method*, Eur. J. Med. Chem. 77 (2014), pp. 298–305.

[51] V.H. Masand, A.A. Toropov, A.P. Toropova, and D.T. Mahajan, *QSAR models for anti-malarial activity of 4-aminoquinolines*, Curr. Comput. Aided Drug Des. 10 (2014), pp. 75–82.

[52] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, and E.A. Castro, *QSAR on aryl-piperazine derivatives with activity on malaria*, Chemometr. Intell. Lab. Syst. 110 (2012), pp. 81–88.

[53] A.P. Toropova, A.A. Toropov, A. Lombardo, A. Roncaglioni, E. Benfenati, and G. Gini, *CORAL: QSAR models for acute toxicity in fathead minnow (*Pimephales promelas*)*, J. Comput. Chem. 33 (2012), pp. 1218–1223.

[54] A.P. Toropova, A.A. Toropov, E. Benfenati, D. Leszczynska, and J. Leszczynski, *QSAR model as a random event: A case of rat toxicity*, Bioorg. Med. Chem. 23 (2015), pp. 1223–1230.

[55] F. Cheng, W. Li, Y. Zhou, J. Shen, Z. Wu, G. Liu, P.W. Lee, and Y. Tang, *admetSAR: A comprehensive source and free tool for assessment of chemical ADMET properties*, J. Chem. Inf. Model. 52 (2012), pp. 3099–3105.

[56] B.C. Doak, B. Over, F. Giordanetto, and J. Kihlberg, *Oral druggable space beyond the rule of 5: Insights from drugs and clinical candidates*, Chem. Biol. 21 (2014), pp. 1115–1142.

[57] J.L. Medina-Franco, *Chapter 21 - Discovery and development of lead compounds from natural sources using computational approaches, in Evidence-Based Validation of Herbal Medicine*, P.K Mukherjee, Ed., Elsevier, Boston, MA, 2015, pp. 455–475.

[58] S. Kesavan and L.A. Marcaurelle, *Translational synthetic chemistry*, Nat. Chem. Biol. 9 (2013), pp. 210–213.

[59] F. Giordanetto and J. Kihlberg, *Macrocyclic drugs and clinical candidates: What can medicinal chemists learn from their properties?*, J. Med. Chem. 57 (2014), pp. 278–295.

[60] B.G. Katzung, *Basic & Clinical Pharmacology*, Mcgraw Hill Companies, UK, 2003.

[61] A.A. Toropov, A.P. Toropova, A. Lombardo, A. Roncaglioni, N. De Brita, G. Stella, and E. Benfenati, *CORAL: The prediction of biodegradation of organic compounds with optimal SMILES-based descriptors*, Cent. Eur. J. Chem. 10 (2012), pp. 1042–1048.

[62] J. García, P.R. Duchowicz, M.F. Rozas, J.A. Caram, M.V. Mirífico, F.M. Fernández, and E.A. Castro, *A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases*, J. Mol. Graph. Model. 31 (2011), pp. 10–19.

[63] A.A. Toropov, A.P. Toropova, B.F. Rasulev, E. Benfenati, G. Gini, D. Leszczynska, and J. Leszczynski, *CORAL: Binary classifications (active/inactive) for liver-related adverse effects of drugs*, Curr. Drug Saf. 7 (2012), pp. 257–261.

[64] P.K. Ojha and K. Roy, *Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection*, Chemometr. Intell. Lab. 109 (2011), pp. 146–161.

[65] R. Prajapati, U. Singh, A. Patil, K.S. Khomane, P. Bagul, A.K. Bansal, and A.T. Sangamwar, *In silico model for P-glycoprotein substrate prediction: Insights from molecular dynamics and in vitro studies*, J. Comput. Aided Mol. Des. 27 (2013), pp. 347–363.

[66] M.A. Demel, O. Krämer, P. Ettmayer, E.E.J. Haaksma, and G.F. Ecker, *Predicting ligand interactions with ABC transporters in ADME*, Chem. Biodivers. 6 (2009), pp. 1960–1969.

[67] S.G. Aller, J. Yu, A. Ward, Y. Weng, S. Chittaboina, R. Zhuo, P.M. Harrell, Y.T. Trinh, Q. Zhang, I.L. Urbatsch, and G. Chang, *Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding*, Science 323 (2009), pp. 1718–1722.

[68] J.M. Damas, A.S.F. Oliveira, A.M. Baptista, and C.M. Soares, *Structural consequences of ATP hydrolysis on the ABC transporter NBD dimer: Molecular dynamics studies of HlyB*, Protein Sci. 20 (2011), pp. 1220–1230.

[69] A.S. Oliveira, A.M. Baptista, and C.M. Soares, *Insights into the molecular mechanism of an ABC transporter: Conformational changes in the NBD dimer of MJ0796*, J. Phys. Chem. B. B 114 (2010), pp. 5486–5496.

[70] J.P. Becker, G. Depret, F. Van Bambeke, P.M. Tulkens, and M. Prévost, *Molecular models of human P-glycoprotein in two different catalytic states*, BMC Struct. Biol. 9 (2009).

[71] M. Scian, M. Acchione, M. Li, and W.M. Atkins, *Reaction dynamics of ATP hydrolysis catalyzed by P-glycoprotein*, Biochemistry 53 (2014), pp. 991–1000.

[72] J.G. Wise, *Catalytic transitions in the human MDR1 P-glycoprotein drug binding sites*, Biochemistry 51 (2012), pp. 5125–5141.

[73] M.V.S. Varma, Y. Ashokraj, C.S. Dey, and R. Panchagnula, *P-glycoprotein inhibitors and their screening: A perspective from bioavailability enhancement*, Pharmacol. Res. 48 (2003), pp. 347–359.

[74] A.B. Shapiro and V. Ling, *Effect of quercetin in Hoechst 33342 transport by purified and reconstituted P-glycoprotein*, Biochem. Pharmacol. 53 (1997), pp. 587–596.

[75] S. Drori, G.D. Eytan, and Y.G. Assaraf, *Potentiation of anticancer-drug cytotoxicity by multidrug-resistance chemosensitizers involves alterations in membrane fluidity leading to increased membrane permeability*, Eur. J. Biochem. 228 (1995), pp. 1020–1029.

[76] N. Maki, P. Hafkemeyer, and S. Dey, *Allosteric modulation of human P-glycoprotein. Inhibition of transport by preventing substrate translocation and dissociation*, J. Biol. Chem. 278 (2003), pp. 18132–18139.

[77] F. Broccatelli, *QSAR models for P-glycoprotein transport based on a highly consistent data set*, J. Chem. Inf. Model. 52 (2012), pp. 2462–2470.

[78] V. Prachayasittikul, P. Mandi, S. Prachayasittikul, V. Prachayasittikul, and C. Nantasenamat, *Exploring the chemical space of P-glycoprotein interacting compounds*, Mini Rev. Med. Chem. 16 (2016).

[79] H. Liu, Z. Ma, and B. Wu, *Structure-activity relationships and in silico models of P-glycoprotein (ABCB1) inhibitors*, Xenobiotica 43 (2013), pp. 1018–1026.

[80] P. Chiba, G. Ecker, D. Schmid, J. Drach, B. Tell, S. Goldenberg, and V. Gekeler, *Structural requirements for activity of propafenone-type modulators in P-glycoprotein-mediated multidrug resistance*, Mol. Pharmacol. 49 (1996), pp. 1122–1130.

[81] P. Chiba, M. Hitzler, E. Richter, M. Huber, C. Tmej, E. Giovagnoni, and G. Ecker, *Studies on propafenone-type modulators of multidrug resistance III: Variations on the nitrogen*, Quant. Struct.-Act. Relat. 16 (1997), pp. 361–366.

[82] F. Klepsch, P. Chiba, and G.F. Ecker, *Exhaustive sampling of docking poses reveals binding hypotheses for propafenone type inhibitors of p-glycoprotein*, PLoS Comput. Biol. 7 (2011), e1002036.

[83] I. Pajeva and M. Wiese, *Molecular modeling of phenothiazines and related drugs as multidrug resistance modifiers: A comparative molecular field analysis study*, J. Med. Chem. 41 (1998), pp. 1815–1826.

[84] H. Müller, I.K. Pajeva, C. Globisch, and M. Wiese, *Functional assay and structure-activity relationships of new third-generation P-glycoprotein inhibitors*, Bioorg. Med. Chem. 16 (2008), pp. 2448–2462.

[85] I.M. Tsakovska, *QSAR and 3D-QSAR of phenothiazine type multidrug resistance modulators in P388/ADR cells*, Bioorg. Med. Chem. 11 (2003), pp. 2889–2899.

[86] A. Ramu and N. Ramu, *Reversal of multidrug resistance by phenothiazines and structurally related compounds*, Cancer Chemother. Pharmacol. 30 (1992), pp. 165–173.

[87] I. Jabeen, K. Pleban, U. Rinner, P. Chiba, and G.F. Ecker, *Structure-activity relationships, ligand efficiency, and lipophilic efficiency profiles of benzophenone-type inhibitors of the multidrug transporter P-glycoprotein*, J. Med. Chem. 55 (2012), pp. 3261–3273.

[88] C. Globisch, I.K. Pajeva, and M. Wiese, *Structure-activity relationships of a series of tariquidar analogs as multidrug resistance modulators*, Bioorg. Med. Chem. 14 (2006), pp. 1588–1598.

[89] F. Broccatelli, E. Carosati, A. Neri, M. Frosini, L. Goracci, T.I. Oprea, and G. Cruciani, *A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields*, J. Med. Chem. 54 (2011), pp. 1740–1751.

[90] S. Ekins, R.B. Kim, B.F. Leake, A.H. Dantzig, E.G. Schuetz, L.B. Lan, K. Yasuda, R.L. Shepard, M.A. Winter, J.D. Schuetz, J.H. Wikel, and S.A. Wrighton, *Three-dimensional quantitative structure-activity relationships of inhibitors of P-glycoprotein*, Mol. Pharmacol. 61 (2002), pp. 964–973.

[91] Z. Parveen, G. Brunhofer, I. Jabeen, T. Erker, P. Chiba, and G.F. Ecker, *Synthesis, biological evaluation and 3D-QSAR studies of new chalcone derivatives as inhibitors of human P-glycoprotein*, Bioorg. Med. Chem. 22 (2014), pp. 2311–2319.

[92] T. Langer, M. Eder, R.D. Hoffmann, P. Chiba, and G.F. Ecker, *Lead identification for modulators of multidrug resistance based on in silico screening with a pharmacophoric feature model*, Arch. Pharm. 337 (2004), pp. 317–327.

[93] A. Seelig, E. Landwojtowicz, H. Fischer, and X. Li Blatter, Towards P-glycoprotein structure–activity relationships, in *Drug Bioavailability: Estimation of solubility, Permeability, Absorption and Bioavailability*, H. van de Waterbeemd, H. Lennernäs and P. Artursson, eds., Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG, 2004, pp. 461–492.

[94] A.B. Pawagi, J. Wang, M. Silverman, R.A.F. Reithmeier, and C.M. Deber, *Transmembrane aromatic amino acid distribution in P-glycoprotein: A functional role in broad substrate specificity*, J. Mol. Biol. 235 (1994), pp. 554–564.

[95] A. Seelig and X. Li, *Blatter, and F. Wohnsland, Substrate recognition by P-glycoprotein and the multidrug resistance- associated protein MRP1: A comparison*, Int. J Clin. Pharmacol. Ther. 38 (2000), pp. 111–121.

[96] G. Chang and C.B. Roth, *Structure of MsbA from* E. coli: *A homolog of the multidrug resistance ATP binding cassette (ABC) transporters*, Science 293 (2001), pp. 1793–1800.

[97] M.P. Gleeson, *Generation of a set of simple, interpretable ADMET rules of thumb*, J. Med. Chem. 51 (2008), pp. 817–834.

[98] H. van de Waterbeemd and E. Gifford, *ADMET in silico modelling: Towards prediction paradise?*, Nat. Rev. Drug Discov. 2 (2003), pp. 192–204.

[99] M. Huber, D. Schmid, G. Ecker, and P. Chiba, *The importance of a nitrogen atom in modulators of multidrug resistance*, Mol. Pharmacol. 56 (1999), pp. 791–796.

[100] K.H. Kim, *3D-QSAR Analysis of 2,4,5- and 2,3,4,5-substituted imidazoles as potent and nontoxic modulators of P-glycoprotein mediated MDR*, Bioorg. Med. Chem. 9 (2001), pp. 1517–1523.

[101] M. Gobbi, M. Beeg, M.A. Toropova, A.A. Toropov, and M. Salmona, *Monte Carlo method for predicting of cardiac toxicity: hERG blocker compounds*, Toxicol. Lett. 250–251 (2016), pp. 42–46.

[102] D. Fourches, J.C. Barnes, N.C. Day, P. Bradley, J.Z. Reed, and A. Tropsha, *Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species*, Chem. Res. Toxicol. 23 (2010), pp. 171–183.

[103] A. Speck-Planche, V.V. Kleandrova, and J.A. Rojas-Vargas, *QSAR model toward the rational design of new agrochemical fungicides with a defined resistance risk using substructural descriptors*, Mol. Divers. 15 (2011), pp. 901–909.

[104] A. Speck-Planche, V.V. Kleandrova, F. Luan, and M. Cordeiro, *Rational drug design for anti-cancer chemotherapy: Multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents*, Bioorg. Med. Chem. 20 (2012), pp. 4848–4855.

Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking

Saw Simeon[1,*], Nuttapat Anuwongcharoen[1,*], Watshara Shoombuatong[1], Aijaz Ahmad Malik[1], Virapong Prachayasittikul[2], Jarl E.S. Wikberg[3] and Chanin Nantasenamat[1]

[1] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand

[2] Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand

[3] Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

[*] These authors contributed equally to this work.

ABSTRACT

Alzheimer's disease (AD) is a chronic neurodegenerative disease which leads to the gradual loss of neuronal cells. Several hypotheses for AD exists (e.g., cholinergic, amyloid, tau hypotheses, etc.). As per the cholinergic hypothesis, the deficiency of choline is responsible for AD; therefore, the inhibition of AChE is a lucrative therapeutic strategy for the treatment of AD. Acetylcholinesterase (AChE) is an enzyme that catalyzes the breakdown of the neurotransmitter acetylcholine that is essential for cognition and memory. A large non-redundant data set of 2,570 compounds with reported IC_{50} values against AChE was obtained from ChEMBL and employed in quantitative structure-activity relationship (QSAR) study so as to gain insights on their origin of bioactivity. AChE inhibitors were described by a set of 12 fingerprint descriptors and predictive models were constructed from 100 different data splits using random forest. Generated models afforded R^2, Q^2_{CV} and Q^2_{Ext} values in ranges of 0.66–0.93, 0.55–0.79 and 0.56–0.81 for the training set, 10-fold cross-validated set and external set, respectively. The best model built using the substructure count was selected according to the OECD guidelines and it afforded R^2, Q^2_{CV} and Q^2_{Ext} values of 0.92 ± 0.01, 0.78 ± 0.06 and 0.78 ± 0.05, respectively. Furthermore, Y-scrambling was applied to evaluate the possibility of chance correlation of the predictive model. Subsequently, a thorough analysis of the substructure fingerprint count was conducted to provide informative insights on the inhibitory activity of AChE inhibitors. Moreover, Kennard–Stone sampling of the actives were applied to select 30 diverse compounds for further molecular docking studies in order to gain structural insights on the origin of AChE inhibition. Site-moiety mapping of compounds from the diversity set revealed three binding anchors encompassing both hydrogen bonding and van der Waals interaction. Molecular docking revealed that compounds **13**, **5** and **28** exhibited the lowest binding energies of −12.2, −12.0 and −12.0 kcal/mol, respectively, against human AChE, which is modulated by hydrogen bonding, π–π stacking and hydrophobic interaction inside the binding pocket. These information may be used as guidelines for the design of novel and robust AChE inhibitors.

Submitted 9 May 2016
Accepted 13 July 2016
Published 9 August 2016

Corresponding author
Chanin Nantasenamat,
chanin.nan@mahidol.ac.th

Academic editor
Walter de Azevedo Jr

Additional Information and Declarations can be found on page 25

DOI **10.7717/peerj.2322**

How to cite this article Simeon et al. (2016), Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. **PeerJ 4:e2322; DOI 10.7717/peerj.2322**

Subjects Bioinformatics, Computational Biology, Neuroscience, Drugs and Devices, Pharmacology
Keywords Acetylcholinesterase, Acetylcholinesterase inhibitor, Alzheimer's disease, Dementia, Neurodegenerative disease, Quantitative structure-activity relationship, QSAR, Data mining

INTRODUCTION

Neurodegenerative diseases is caused by the progressive loss of neural cells thereby leading to nervous system dysfunction (*Beal, 1995*; *Kuca et al., 2016*). In particular, Alzheimer's disease (AD) is a debilitating illness that is expected to triple by the year 2050 (*Brookmeyer et al., 2007*). AD is characterized by gradual cognitive impairment, memory loss and decline in speech, behavioral abnormality and eventually death. The pathological changes in AD are mainly attributed to the dramatic loss of neurons in many areas of the central nervous system accompanied by a great reduction in the levels of neurotransmitters. Acetylcholine (ACh) is a neurotransmitter possessing important cognitive and muscular functions. Particularly, in the peripheral nervous system, ACh is found at the neuromuscular junction where it is involved in muscle contraction while in the central nervous system, it is involved in cognitive functions such as thought, learning and memory.

Acetylcholinesterases (AChE) is an enzyme that catalyzes the breakdown of ACh to choline and acetic acid (*Quinn, 1987*). Thus, a promising therapeutic approach is to maintain the level of ACh by inhibiting the enzyme that is responsible for its breakdown. The structure of AChE is comprised of four main subsites consisting of anionic subsite, esteratic site, oxyanion hole and the acyl pocket (*Bourne, Taylor & Marchot, 1995*). The anionic site is involved in the binding of the positive quaternary amine of ACh (*Ordentlich et al., 1993*). The substrate interacts with the 14 aromatic residues that forms the active site. Of these 14 aromatic residues, Trp84 is important for the enzyme activity because when it is replaced by alanine, the activity of the enzyme decreased by 3,000-fold (*Tougu, 2001*). The esteratic site contains the catalytic triad consisting of Ser203, His447 and Glu334 (i.e., resembling that of chymotrypsin and other serine proteases) that hydrolyzes ACh to acetate and choline (*Harel et al., 1993*). The mechanism of the hydrolysis starts from the carboxyl ester leads to the formation of an acyl-enzyme and choline. Finally, the acyl-enzyme undergoes nucleophilic attack by water molecules thereby regenerating the enzyme (*Tougu, 2001*). The oxyanion hole consisting of Gly121, Gly122 and Ala204 contribute hydrogen bond donors to help stabilize the tetrahedral intermediate of ACh form during catalysis. The acyl pocket consisting of Phe295 and Phe297 are gatekeepers that limit the dimension of substrates that can enter the active site.

AChE inhibitors form one of the most actively investigated classes of compounds having been labeled as a potential agent for the treatment of AD by inhibiting AChE from hydrolyzing ACh, thereby leading to increases in the level of ACh (*Birks, 2006*). Generally, AChE inhibitors can be classified into reversible and irreversible inhibitors. Reversible inhibitor bind to the enzyme at allosteric sites as to reduce the activity of the enzyme whether or not the enzyme has already bind the substrate or not. Tacrine is a reversible AChE inhibitor that was synthesized nearly five decades ago and in 1993 it has become the first drug to be marketed for the treatment of AD with approval from the US. Food and

Drug Administration (*Racchi et al., 2004*). On the other hand, irreversible inhibitors such as metrifonate (*Morris et al., 1998*) bind to the target enzymes and dissociates very slowly from the enzyme via either covalent or non-covalent interactions (*Kitz & Wilson, 1962*).

Quantitative structure–activity relationship (QSAR) is a paradigm that enables the prediction of biological activities for compounds of interest as a function of their descriptors through the use of statistical or machine learning methods (*Nantasenamat et al., 2009*). Aside from the ability to predict the activity, QSAR models have been instrumental in enabling understanding on the origin of these biological activities by means of interpreting the descriptors used in building such models.

Historically, the first QSAR investigation of AChE inhibitors was reported by *Mundy et al. (1978)* almost 40 years ago in which the log(1/LD_{50} for a series of twelve substituted 0,0-dimethyl 0-(*p*-nitrophenyl) phosphorothioates and 0-analogs was predicted as a function of the octanol/water partition coefficient. Analysis of the literature of QSAR studies of AChE revealed that much of the early studies are classical QSAR models (i.e., Hansch and Free-Wilson approach) that are based on small congeneric compound set and primarily aimed at predicting AChE inhibition as to investigate the toxic effect of pesticides of various chemotypes belonging to either organophosphates (*Mager, 1983*; *Aaviksaar, 1990*) or carbamates (*Su & Lien, 1980*; *Goldblum, Yoshimoto & Hansch, 1981*; *Walters & Hopfinger, 1986*). Recent QSAR studies are based on the use of large and heterogeneous data sets comprising of structurally diverse chemotypes. This include the study from *Yan & Wang (2012)* where they predicted AChE inhibition for a large set of 404 compounds using multiple linear regression and support vector machine. Furthermore, *Lee & Barron (2016)* performed a 3D-QSAR investigation on a large set of 341 compounds comprising of organophosphates and carbamates. Moreover, *Veselinović et al. (2015)* compiled a set of 278 organophosphates for which they developed QSAR models for predicting AChE inhibition using SMILES-based descriptors.

Research in this field had experienced two distinct transitions when viewed from biological and computational viewpoints. Biologically, early QSAR studies treat AChE as a biomarker of toxicity from pesticides while investigations from later years had shifted the focus by viewing AChE as a therapeutic target for the treatment of AD. In regards to the latter point, viewpoint on targeting AChE as a single target for treating AD is starting to be replaced by the multi-target concept in which the treatment for AD can be approached by a panel of key targets (*Fang et al., 2015*; *Huang et al., 2011*). Computationally, early studies are predominantly based on simple 2D-QSAR (*Mundy et al., 1978*; *Su & Lien, 1980*) while later years started to use more sophisticated approach for understanding AChE inhibition encompassing 3D-QSAR (*Deb et al., 2012*; *Lee & Barron, 2016*; *Prado-Prado et al., 2012*), molecular dynamics (*Shen et al., 2002*), molecular docking (*Lu et al., 2011*; *Deb et al., 2012*; *Giacoppo et al., 2015*), pharmacophore modeling (*Lu et al., 2011*; *Gupta & Mohan, 2014*) and statistical molecular design (*Andersson et al., 2014*; *Prado-Prado, Escobar & Garcia-Mera, 2013*).

Herein, we propose the first large-scale QSAR investigation for predicting AChE inhibition, which to the best of our knowledge represents the largest collection of 2,570 non-redundant compounds. QSAR models were built using interpretable learning methods (e.g., random forest) and descriptors (i.e., molecular fingerprints) as to unravel the underlying AChE inhibitory activity, which was performed in accordance with guidelines of the Organisation for Economic Cooperation and Development (OECD). Molecular docking was also performed on a chemically diverse set of compounds selected from active AChE inhibitors. Together, the ligand and structure-based approach employed in this study is anticipated to be useful in the design and development of robust AChE inhibitors.

MATERIALS AND METHODS

A summary of the workflow of this study is provided in Fig. 1. Briefly, this included a large-scale QSAR model for predicting and analyzing the AChE inhibition, which was performed in accordance with the OECD guidelines as follows: (i) a data set with a defined endpoint; (ii) an unambiguous learning algorithm; (iii) a defined applicability domain of the QSAR model; (iv) using appropriate measures of goodness-of-fit, robustness and predictivity; (v) a mechanistic interpretation of the QSAR model. Furthermore, molecular docking was also performed on a chemically diverse data set as to elucidate the underlying binding mechanism. To facilitate the reproducibility of the research work performed herein, the data set (Data S1) and codes (Data S2) used to perform the multivariate analysis are provided as Supplemental Information.

Data set

A data set of inhibitors against human AChE (Target ID CHEMBL220) were compiled from the ChEMBL 20 database (*Gaulton et al., 2012*) that is comprised of a total number of 9,242 bioactivity data points from 5,049 compounds. SMILES notations of the compounds were curated with ChemAxon's Standardizer (*ChemAxon Kft., 2015*) using the same parameter settings as described in our previous study (*Simeon et al., 2016*). The initial data set was assembled from several bioactivity measurement units including (in order of decreasing data size) IC_{50}, K_i, % activity, % inhibition, MIC, EC_{50}, etc. IC_{50} was selected for further investigation as they constituted the largest subset with 4,910 compounds. A closer look revealed that 1,301 compounds had no reported IC_{50} values or had lesser/greater than signs which were subjected to removal thereby leaving 3,609 compounds. Only data points having nM as the bioactivity unit were selected for further study, which produced 3,596 compounds. Furthermore, redundant compounds having different bioactivity values were kept if the standard deviation of IC_{50} was less than 2 and this resulted in 2,571 compounds. Moreover, some compounds were found to have no SMILES notation associated with it and were thus removed. A final data set comprising of 2,570 compounds was obtained.

Description of inhibitors

AChE inhibitors were encoded by a vector of fingerprint descriptors accounting for its molecular constituents. Prior to calculating descriptors, salts were removed and tautomers were standardized using the built-in function of the PaDEL-Descriptor software (*Yap, 2011*).

Although fingerprint descriptors are able to capture the feature space of chemical compounds, their ability to be used as descriptors for bioactivity modeling can vary. In

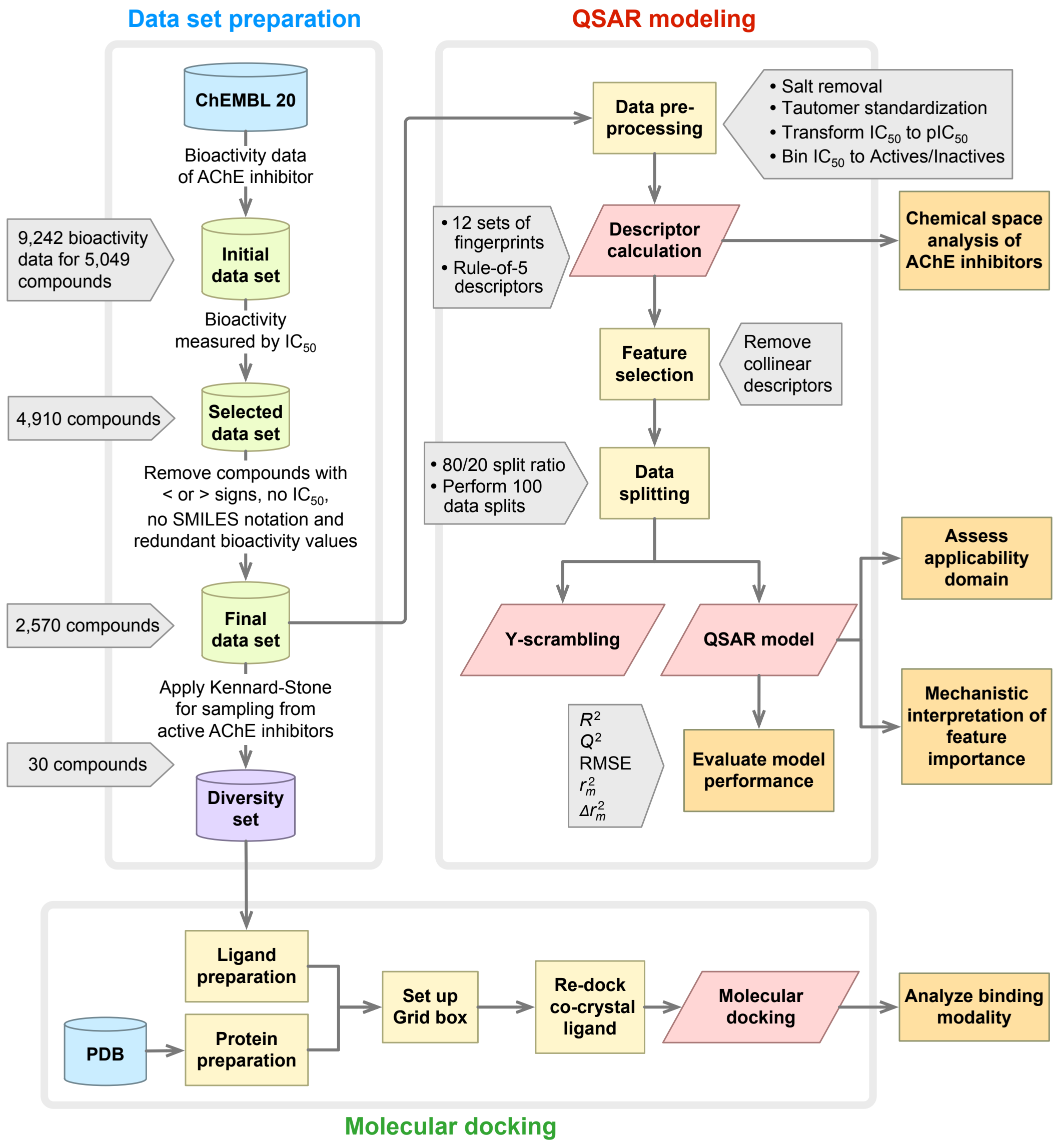

Data set preparation
QSAR modeling
ChEMBL 20
Bioactivity data of AChE inhibitor
9,242 bioactivity data for 5,049 compounds
Initial data set
Bioactivity measured by IC_{50}
4,910 compounds
Selected data set
Remove compounds with < or > signs, no IC_{50}, no SMILES notation and redundant bioactivity values
2,570 compounds
Final data set
Apply Kennard-Stone for sampling from active AChE inhibitors
30 compounds
Diversity set
Data pre-processing
• Salt removal
• Tautomer standardization
• Transform IC_{50} to pIC_{50}
• Bin IC_{50} to Actives/Inactives
• 12 sets of fingerprints
• Rule-of-5 descriptors
Descriptor calculation
Chemical space analysis of AChE inhibitors
Feature selection
Remove collinear descriptors
• 80/20 split ratio
• Perform 100 data splits
Data splitting
Assess applicability domain
Y-scrambling
QSAR model
Mechanistic interpretation of feature importance
R^2
Q^2
RMSE
r_m^2
Δr_m^2
Evaluate model performance
Ligand preparation
PDB
Protein preparation
Set up Grid box
Re-dock co-crystal ligand
Molecular docking
Analyze binding modality
Molecular docking

Figure 1 Workflow of QSAR modeling and molecular docking for investigating AChE inhibitory activity.

fact, performance differences existing amongst the different fingerprint type has been the subject of several investigations into its utilization for bioactivity modeling. *Riniker & Landrum (2013)* benchmarked and assessed the performance of predictive models constructed from 2D fingerprint descriptors obtained from RDKit.

In this study, the suitability of 12 different fingerprint descriptors for predicting the bioactivity of AChE inhibitors was investigated. Table 1 summarizes the employed fingerprints along with their corresponding size, description and reference.

Additionally, the four molecular descriptors that are used to define the Lipinski's rule-of-five comprising of molecular weight (MW), logarithm of the octanol/water partition coefficient (ALogP), number of hydrogen bond donor (nHBDon) and number of hydrogen bond acceptor (nHBAcc) were also computed by the PaDEL-Descriptor software.

Table 1 **Summary of 12 sets of fingerprint descriptors employed in this study.**

No.	Fingerprint	Number	Description	Reference
1	CDK	1024	Fingerprint of length 1024 and search depth of 8	*Steinbeck et al. (2003)*
2	CDK extended	1024	Extends the fingerprint with additional bits describing ring features	*Steinbeck et al. (2003)*
3	CDK graph only	1024	A special version that considers only the connectivity and not bond order	*Steinbeck et al. (2003)*
4	E-state	79	Electrotopological state atom types	*Hall & Kier (1995)*
5	MACCS	166	Binary representation of chemical features defined by MACCS keys	*Durant et al. (2002)*
6	PubChem	881	Binary representation of substructures defined by PubChem	*NCBI (2009)*
7	Substructure	307	Presence of SMARTS patterns for functional groups	*Laggner (2005)*
8	Substructure count	307	Count of SMARTS patterns for functional groups	*Laggner (2005)*
9	Klekota–Roth	4860	Presence of chemical substructures	*Klekota & Roth (2008)*
10	Klekota–Roth count	4860	Count of chemical substructures	*Klekota & Roth (2008)*
11	2D atom pairs	780	Presence of atom pairs at various topological distances	*Carhart, Smith & Venkataraghavan (1985)*
12	2D atom pairs count	780	Count of atom pairs at various topological distances	*Carhart, Smith & Venkataraghavan (1985)*

Feature selection

Collinearity is a condition where descriptor pairs are known to have intercorrelation, which not only add complexity to the model but could potentially give rise to bias. To remedy this, the *cor* function from the R package *stats* was used to find the pairwise correlation among descriptors, and descriptors in a pair with a Pearson's correlation coefficient greater than the threshold of 0.7 was filtered out using the *findCorrelation* function from the R package *caret* to obtain a smaller subset of descriptors (*Kuhn, 2008*).

Data splitting

To avoid the possibility of bias that may arise from a single data split when building predictive models (*Puzyn et al., 2011*), predictive models were constructed from 100 independent data splits and the mean and standard deviation values of statistical parameters were reported. The data set was split into internal and external sets in which the former comprises 80% whereas the latter constitutes 20% of the initial data set. The *sample* function from the R *base* package was used to split the data.

Multivariate analysis

Supervised learning is to learn a model from labeled training data which can be used to make prediction about unseen or future data (*James et al., 2013*). This study constructs regression models, which affords the prediction of the continuous response variable (i.e., pIC_{50}) as a function of predictors (i.e., fingerprint descriptors).

Random forest (RF) is an ensemble classifier that is composed of several decision trees (*Breiman, 2001*). Briefly, the main idea behind RF is that instead of building a deep decision tree with an ever-growing number of nodes, which may be at risk for overfitting and overtraining of the data, rather multiple trees are generated as to minimize the variance instead of maximizing the accuracy. As such, the results will be more noisier when compared to a well-trained decision tree, yet these results are usually reliable and robust. The *ranger* function from the R package *ranger*, which is a fast implementation of the RF algorithm that was used for constructing the models (*Wright & Ziegler, 2015*).

Validation of QSAR models

Model validation is an important process, which should be performed to ensure that a fitted model can accurately predict responses for future or unknown subjects. Two statistical parameters were used to evaluate the performance of the QSAR models consisting of Pearson's correlation coefficient (r) and root mean squared error (RMSE). The r value is a commonly used metric to represent the degree of relationship between two variables of interest. It can range from -1 to $+1$ in which negative values are indicative of negative correlation between two variables and vice versa. RMSE is a commonly used parameter to assess the relative error of the predictive model. The predictive performance of the QSAR models was verified by 10-fold cross-validation, external validation and Y-scrambling test.

The 10-fold cross-validation technique does not used the entire data set to build predictive model. Instead, it splits the data into training and testing data set by allowing model that is built with training data set us allow to assess the performance of the model on the testing data set. By performing repeats of the 10-fold validation, the average accuracies can be used to truly assess the performance of the predictive model.

Y-scrambling test was used to ensure the robustness of the predictive model not only to rule out the possibility of chance correlations but also to assess the statistical significance of R^2 and Q^2, ensuring the generalizability of QSAR model. The true Y-dependent variable (i.e., pIC_{50}) was randomly scrambled and the statistical assessment parameters are recalculated. Performance of the Y-scrambling test can be deduced from the regression line of the plot of R^2 versus Q^2. Intercept values for R^2 and Q^2 as denoted by iR^2 and iQ^2, respectively, were calculated. Negative iQ^2 is indicative of an acceptable QSAR model and that there is no chance correlation from the real model (*Eriksson et al., 2003*). Furthermore, r_m^2 and Δr_m^2 metrics as introduced by *Roy et al. (2013)* were used to verify the robustness of the proposed QSAR model in which an acceptable QSAR model should give $r_m^2 > 0.5$ and $\Delta r_m^2 < 0.2$.

Applicability domain analysis

The applicability domain (AD) estimates the likelihood of reliable prediction for compounds on the basis of how similar they are to compounds used to build the model. Thus, compounds falling outside the AD may lead to unreliable predictions. The most common approach for determining AD is described by *Gramatica (2007)* and *Tropsha, Gramatica & Gombar (2003)*, which is to compute the leverage values for each compound. The leverage value allows one to identify whether new compounds will lie within or outside the domain. Leverage values for all compounds are calculated via adjustment of X to give the hat matrix H:

$$H = X(X^T X)^{-1} X^T \tag{1}$$

where X is a two-dimensional matrix comprising of n compounds and m descriptors while X^T is the transpose of X. Meanwhile, the leverage value of the ith compound (h_i) is the ith diagonal element of H:

$$h_i = x_i^T (X^T X)^{-1} x_i \tag{2}$$

where x_i is the descriptor row-vector of the ith compound. The warning leverage h^* is calculated by:

$$h^* = 3(p+1)/n. \tag{3}$$

Practically, the leverage value along with the William's plot is often used to assess the AD of QSAR models. The William's plot is constructed by depicting the standardized residuals versus the leverage value for each compound's h_i. If the ith compound has $h_i > h^*$ then it means that the ith compound exerts a great influence on the QSAR model and may be excluded from the AD. In spite of this, it does not appear to be an outlier because its standardized residual may be small.

Molecular docking

The co-crystal structure of human AChE with donepezil (PDB ID: 4EY7) was retrieved from the Protein Data Bank and initially prepared by removing alternative side chains and water molecules. The protein was prepared via the rebuilding of bonds and the addition of missing hydrogen atoms. Subsequently, the protein was cleaned by merging the atomic charge and removing lone pair atoms, non-polar hydrogen atoms and non-standard amino acid residues. Grid box was set up to provide coverage of the active site of human AChE with a dimension of 40 × 30 × 40 Å (X, Y and Z axes of −13.987, −41.668 and 27.109, respectively). Molecular docking was consequently performed with AutoDock Vina (*Trott & Olson, 2010*) using default parameters. The docking protocol was validated in order to ensure its reliability for subsequent analysis of the studied compounds. This was performed by extracting the co-crystal ligand, donepezil, from the PDB file and re-docked to the co-crystal human AChE protein. The root mean squared deviation (RMSD) of the atomic position between the original orientation of the co-crystal ligand and the re-docked ligand is computed and is deemed acceptable if the RMSD value is less than or equal to 2.0 Å.

A set of 30 representative and chemically diverse compounds, which will be referred hereafter as the diversity set, were extracted from the full set of active AChE inhibitors (i.e., IC_{50} <1 μM) using the Kennard–Stone algorithm (*Kennard & Stone, 1969*). These compounds were used as ligands for molecular docking against the human AChE. The binding energy (kcal/mol) of AChE inhibitors were calculated according to the built-in scoring function of Autodock Vina and conformers providing the lowest binding energy were selected for further analysis of the binding mode. Furthermore, key-interacting residues and their moiety preferences were analyzed using LigPlot+ (*Wallace, Laskowski & Thornton, 1995*), Maestro (*Schrödinger, 2015b*) and the SiMMap web server (*Bollback, 2006*). Finally, three-dimensional structure of protein–ligand interaction was created and visualized using Pymol (*Schrödinger, 2015a*).

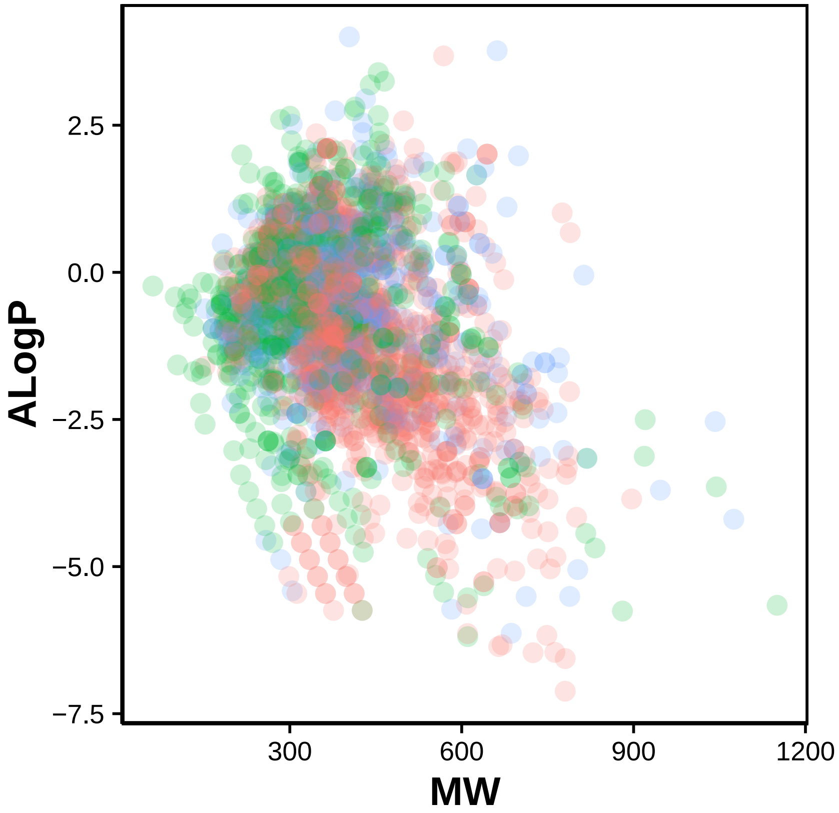


Figure 2 **Chemical space of AChE inhibitors.** Actives, inactives and intermediates are shown in green, red and blue colors, respectively.

RESULTS AND DISCUSSION

Chemical space of AChE inhibitors

Navigation of the chemical space of AChE inhibitors was performed to gain insights into the structure–activity relationship by analyzing the Lipinski's rule-of-five descriptors. Chemical space analysis may provide important knowledge on the general character of compounds governing inhibitory properties of compounds. Exploratory data analysis was performed using the Lipinski's rule-of-five descriptors comprising of MW, ALogP, nHBDon and nHBAcc. MW represents the molecular size of a compound that is commonly used because of it can be easily interpreted and calculated as well as appropriate size of a compound is important for its passage via lipid membrane. ALogP is a widely used parameter for determining the lipophilicity of a compound and used in calculating the membrane penetration and permeability of compounds. nHBDon and nHBAcc describe the number of hydrogen bond donors and hydrogen bond acceptors, respectively, which is used to measuring hydrogen bonding capacity. Visualization of the chemical space of ALogP as a function of MW is shown in Fig. 2, as to investigate the chemical space of AChE inhibitors. A dense distribution of inhibitors was observed within the space of MW starting from approximately 300–600 Da and within the space of ALogP ranging from approximately −2.5 to 2.5. In addition, the box plot of the Lipinski's descriptors is

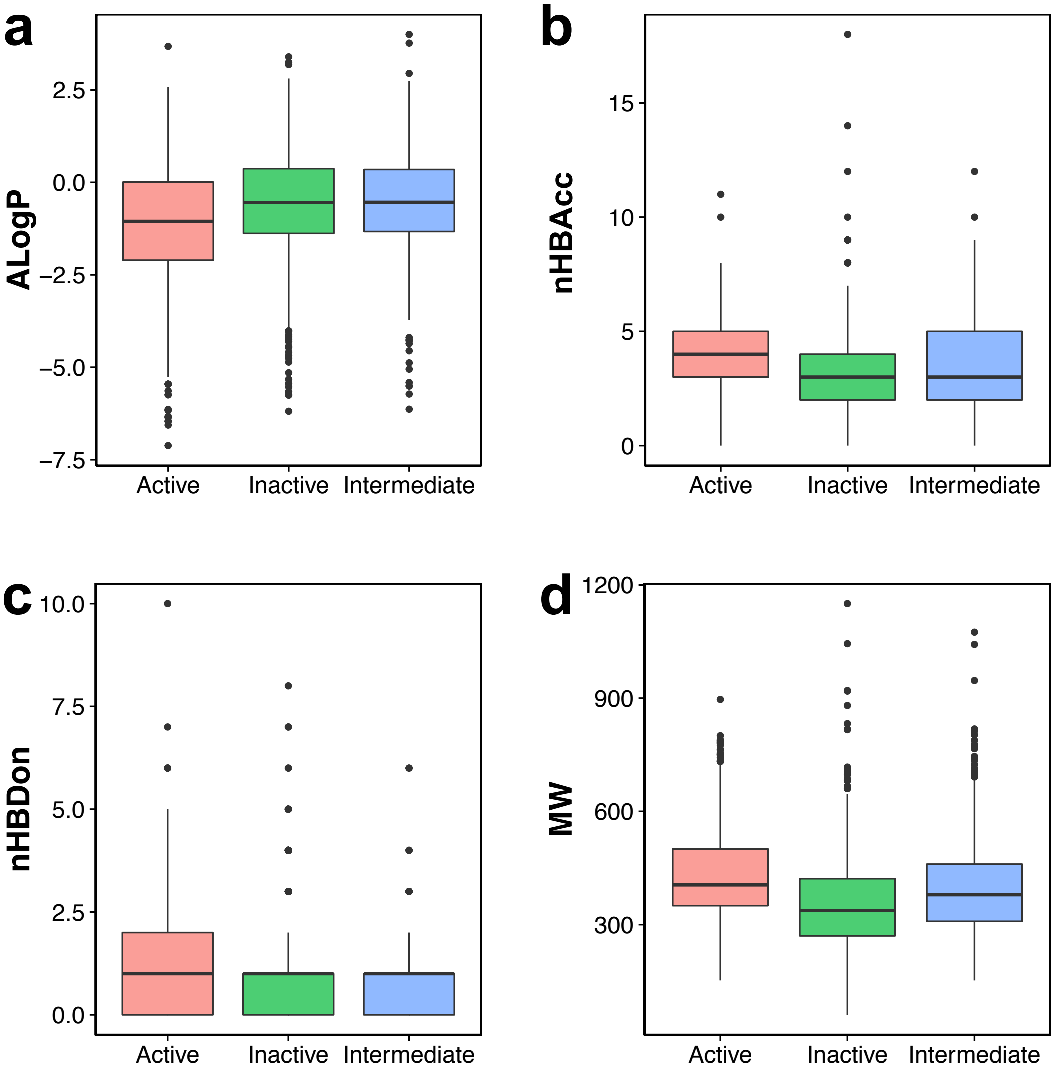


Figure 3 Box plot of AChE inhibitors using Lipinski's rule-of-five descriptors.

shown in Fig. 3. Compounds with negative ALogP values approximately of closer to 0.0 can be found in inactive inhibitors whereas most of the active inhibitors tend to possess approximately lower values in average of ALogP values.

Visual representation of the overall distribution of data values of Lipinski's descriptors is shown as box plots in Fig. 3 in which the ALogP, nHBAcc, nHBDon and MW are shown in Figs. 3A, 3B, 3C and 3D, respectively. Analysis of the box plots revealed that there were no differences amongst the three bioactivity classes for nHBAcc and nHBAcc as deduced from the boundaries of the boxes (i.e., representing the first and third quartiles). ALogP and MW were found to display differences amongst the bioactivity classes. Particularly, ALogP values for actives were the lowest while negligible differences were observed for the other two classes. Furthermore, MW for actives were the largest amongst the three bioactivity classes, which is followed by the intermediates while inactives were smallest.

QSAR model for predicting AChE inhibitory activity

A data set comprising of 2,570 compounds were used for construction of QSAR models. Particularly, twelve sets of fingerprint descriptors were benchmarked in order to find the best performing set. Prior to modeling, feature selection was applied to remove collinear descriptors. Each of the twelve models were then built using a data split ratio of 80/20 in which 80% of the data set was used as the internal set and 20% as the external set. This procedure was iteratively performed in which each of the 100 independent data splits were used for model construction and the performance results given in Table 2 are the mean and standard deviation values derived from these runs.

It can be observed that all twelve models are capable of capturing the inhibitory activity space of AChE inhibitors as they provided R^2 and Q^2 (i.e., both 10-fold CV and external sets) greater than the threshold values proposed by *Golbraikh & Tropsha (2002)* of 0.6 and 0.5, respectively, which is indicative of robust model performance. The possibility of chance correlation can be assessed from the R^2–Q^2 margin as described by *Eriksson & Johansson (1996)* in which values <0.2–0.3 are indicative of predictive and reliable models while values >0.2–0.3 suggests possible chance correlation or the presence of outliers in the data set. Furthermore, observation of the Q^2_{CV}–Q^2_{Ext} margin revealed that the difference was negligible with values in the range of 0 and 0.01.

Generally, it can be seen that models with larger descriptor size, namely CDK and CDK extended, afforded the best performance with Q^2_{CV} of 0.79 ± 0.07 and 0.79 ± 0.06, respectively, and Q^2_{Ext} of 0.80 ± 0.04 and 0.81 ± 0.04, respectively. The opposite also holds true as the model with the least number of descriptors were also found to perform the worst amongst the other fingerprints with Q^2_{CV} of 0.55 ± 0.09 and Q^2_{Ext} of 0.56 ± 0.05. In a nutshell, the model performance in order of decreasing value is as follows: CDK extended > CDK > MACCS ≈ Substructure count ≈ Klekota–Rota count > PubChem > Klekota–Roth ≈ 2D atom pairs count > CDK graph only > Substructure > 2D atom pairs > E-state.

The best performing model is not necessarily the best choice considering the fact that the descriptor size for the best models were quite high and is consequently prone to overfitting. It was found that the substructure count provided reasonably good predictive performance (i.e., Q^2_{CV} and Q^2_{Ext} of 0.78 ± 0.06 and 0.78 ± 0.05, respectively) with the advantage of making use of a small set of 26 descriptors. Therefore, this fingerprint was selected for further interpretation of the feature importance.

To further check the reliability and validity of the selected model, Y-scrambling test was performed for 100 iterations. Table 3 demonstrates that QSAR models built using substructure count has a low Q^2 (−0.0013), which rules out the possibility of chance correlation. Furthermore, model afforded an r^2_m value of 0.61 ± 0.06 thereby revealing its robustness. It is observed in Table 3 that the value of Δr^2_m is greater that 0.2 but also close to 0.20.

As shown in Fig. 4, it can also be seen that scatter plots of experimental versus predicted pIC_{50} of panels A, C and F displayed narrower variance of the data points than the other methods as assessed via 10-fold cross-validation and external set.

Table 2 **Performance summary of QSAR models for predicting pIC_{50}.**

Descriptor class	N	Training set		10–fold CV set		External set		$R^2 - Q^2_{\mathrm{CV}}$	$R^2 - Q^2_{\mathrm{Ext}}$
		R^2	RMSE_{Tr}	Q^2_{CV}	$\mathrm{RMSE}_{\mathrm{CV}}$	Q^2_{Ext}	$\mathrm{RMSE}_{\mathrm{Ext}}$		
CDK	960	0.93 ± 0.01	0.44 ± 0.04	0.79 ± 0.07	0.76 ± 0.15	0.80 ± 0.04	0.73 ± 0.09	0.14	0.13
CDK extended	948	0.94 ± 0.01	0.42 ± 0.03	0.79 ± 0.06	0.76 ± 0.12	0.81 ± 0.04	0.72 ± 0.08	0.15	0.13
CDK graph only	198	0.87 ± 0.01	0.61 ± 0.03	0.72 ± 0.06	0.87 ± 0.13	0.72 ± 0.05	0.87 ± 0.09	0.15	0.15
E-State	21	0.66 ± 0.03	1.00 ± 0.05	0.55 ± 0.09	1.11 ± 0.13	0.56 ± 0.05	1.10 ± 0.08	0.11	0.10
MACCS	77	0.89 ± 0.01	0.56 ± 0.03	0.77 ± 0.07	0.81 ± 0.15	0.77 ± 0.04	0.80 ± 0.09	0.12	0.12
PubChem	103	0.90 ± 0.01	0.55 ± 0.03	0.76 ± 0.05	0.80 ± 0.11	0.78 ± 0.03	0.79 ± 0.08	0.14	0.12
Substructure	30	0.75 ± 0.01	0.85 ± 0.03	0.64 ± 0.06	1.00 ± 0.13	0.65 ± 0.05	0.98 ± 0.08	0.11	0.10
Substructure count	26	0.92 ± 0.01	0.50 ± 0.02	0.78 ± 0.06	0.77 ± 0.14	0.78 ± 0.05	0.77 ± 0.10	0.14	0.14
Klekota–Roth	111	0.89 ± 0.01	0.59 ± 0.03	0.74 ± 0.07	0.85 ± 0.14	0.76 ± 0.05	0.81 ± 0.10	0.15	0.13
Klekota–Roth count	72	0.91 ± 0.01	0.52 ± 0.03	0.78 ± 0.06	0.79 ± 0.14	0.78 ± 0.05	0.77 ± 0.11	0.13	0.13
2D atom pairs	42	0.75 ± 0.03	0.85 ± 0.06	0.61 ± 0.08	1.03 ± 0.15	0.60 ± 0.06	1.05 ± 0.12	0.14	0.15
2D atom pairs count	38	0.92 ± 0.01	0.51 ± 0.02	0.74 ± 0.07	0.84 ± 0.15	0.76 ± 0.05	0.82 ± 0.10	0.18	0.16

Table 3 **Performance summary of QSAR models assessed using additional statistical metrics.**

Descriptor class	*N*	Training set		10-fold CV set		External set		iR^2	iQ^2
		r_m^2	Δr_m^2	r_m^2	Δr_m^2	r_m^2	Δr_m^2		
CDK	960	0.82 ± 0.02	0.07 ± 0.01	0.62 ± 0.09	0.20 ± 0.06	0.64 ± 0.05	0.19 ± 0.03	0.0003	−0.0003
CDK extended	948	0.83 ± 0.01	0.07 ± 0.01	0.62 ± 0.08	0.20 ± 0.05	0.65 ± 0.05	0.18 ± 0.03	0.0006	−0.0005
CDK graph only	198	0.70 ± 0.02	0.14 ± 0.01	0.51 ± 0.07	0.27 ± 0.05	0.53 ± 0.05	0.26 ± 0.03	0.0007	−0.0006
E-State	21	0.35 ± 0.03	0.38 ± 0.02	0.27 ± 0.07	0.40 ± 0.04	0.28 ± 0.05	0.41 ± 0.03	0.0011	−0.0009
MACCS	77	0.73 ± 0.01	0.12 ± 0.01	0.57 ± 0.09	0.23 ± 0.05	0.58 ± 0.05	0.23 ± 0.03	0.0005	−0.0004
PubChem	103	0.74 ± 0.02	0.12 ± 0.01	0.57 ± 0.07	0.23 ± 0.04	0.59 ± 0.05	0.22 ± 0.03	0.0006	−0.0005
Substructure	30	0.50 ± 0.02	0.28 ± 0.01	0.39 ± 0.07	0.34 ± 0.05	0.41 ± 0.05	0.33 ± 0.03	0.0033	–0.0027
Substructure count	26	0.77 ± 0.01	0.10 ± 0.01	0.60 ± 0.08	0.22 ± 0.05	0.61 ± 0.06	0.21 ± 0.04	0.0015	−0.0013
Klekota–Roth	111	0.71 ± 0.02	0.14 ± 0.01	0.54 ± 0.08	0.25 ± 0.05	0.56 ± 0.06	0.24 ± 0.03	0.0006	–0.0004
Klekota–Roth count	72	0.76 ± 0.02	0.11 ± 0.01	0.60 ± 0.08	0.22 ± 0.05	0.61 ± 0.07	0.21 ± 0.04	0.0006	−0.0005
2D atom pairs	42	0.49 ± 0.03	0.28 ± 0.02	0.35 ± 0.08	0.36 ± 0.04	0.35 ± 0.06	0.36 ± 0.03	0.0010	–0.0008
2D atom pairs count	38	0.75 ± 0.01	0.10 ± 0.01	0.52 ± 0.05	0.26 ± 0.05	0.54 ± 0.06	0.25 ± 0.04	0.0006	−0.0005

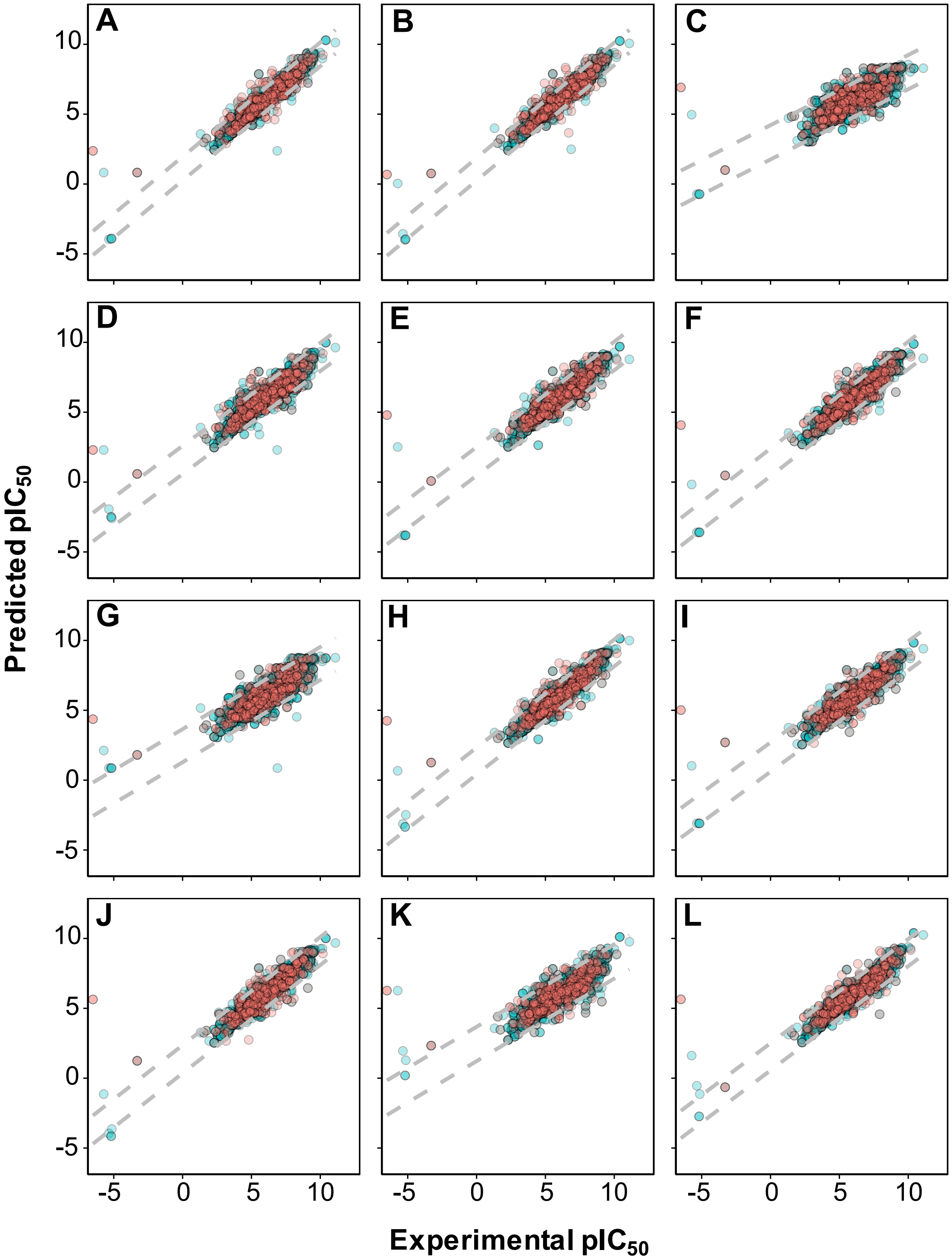


Figure 4 Plot of experimental versus predicted pIC$_{50}$ values for models constructed with 12 different fingerprint descriptors. Shown are models built with CDK fingerprint (A), CDK extended fingerprint (B), E-State fingerprint (C), CDK graph only fingerprint (D), MACCS fingerprint (E), PubChem fingerprint (F), substructure fingerprint (G), substructure fingerprint count (H), Klekota–Roth fingerprint (I), Klekota–Roth fingerprint count (J), 2D atom pairs (K) and 2D atom pairs count (L).

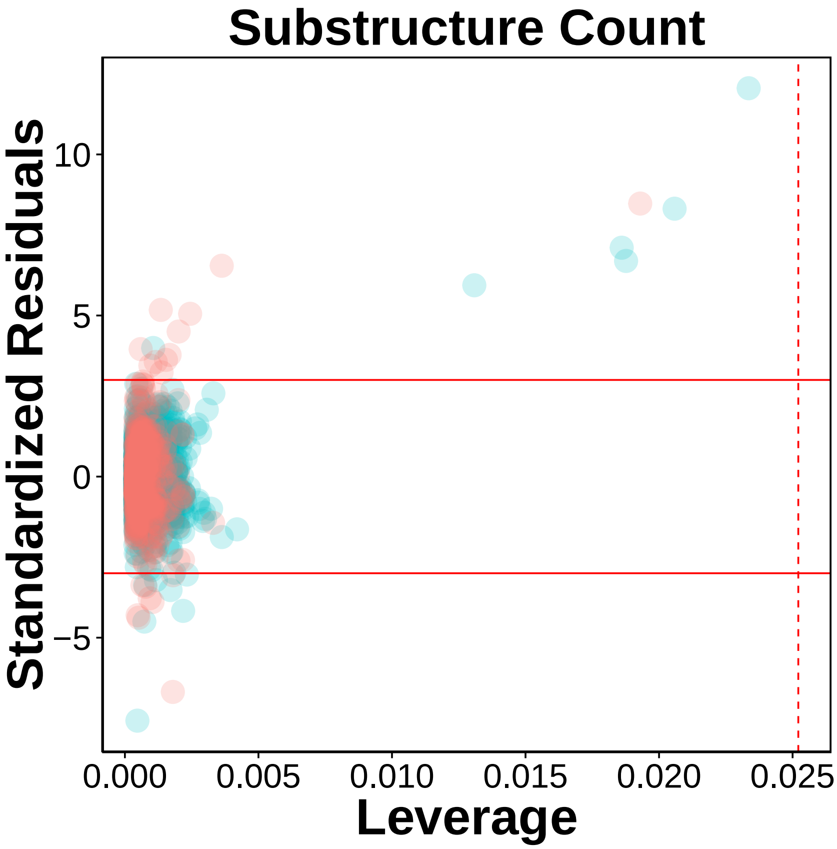


Figure 5 Evaluation of applicability domain using William plots for QSAR model built with substructure fingerprint count. Compounds in the internal and external sets are shown by blue and red dots, respectively. The solid and dashed lines correspond to the ±3 standardized residual and the warning leverage value ($h^* = 0.025$), respectively.

Applicability domain

The AD of the proposed QSAR model was defined as provided by the Williams plot shown in Fig. 5. The employed data set consisting of 2,570 compounds was randomly split to two separate subset in which the first subset constituting 80% of the data set was used as an internal set while the second subset constituting the remaining 20% were used as an external set. Compounds representing the internal set (blue dots) and external set (red dots) are shown in the Williams plot and it can be clearly seen that almost all of the 2,570 compounds were located within the boundaries of applicability domain, which indicated that our proposed QSAR model had a well-defined AD.

As can be seen in Fig. 5, very few compounds indeed fall outside the ±3 standardized residual range. This consisted of six compounds (**997**, **1829**, **62**, **1096**, **13**, **677**) from the internal set and 11 compounds (**2116**, **2120**, **2388**, **2117**, **2392**, **2323**, **2423**, **2424**, **2507**, **2219**, **2422**) from the external set that had standardized residual higher than 3. On the other hand, 7 compounds (**1567**, **576**, **1644**, **1098**, **2022**, **1447** and **322**) from the internal set and 8 compounds (**2486**, **2353**, **2130**, **2553**, **2389**, **2072**, **2103**, **2125**) from the external set had standardized residual lower than −3. The corresponding chemical structures are provided in Table S1.

Mechanistic interpretation of feature importance

Feature importance analysis help reveal features that are important toward bioactivity. There are essentially two parameters for evaluating the relative importance of features used in models using the RF algorithm: (i) accuracy and (ii) Gini index (i.e., variance of the responses). The latter was selected as a metric for ranking important features (i.e., mean decrease of the Gini index) for predicting the pIC_{50} of AChE inhibitors (Fig. 6). Table 4 lists the substructure fingerprints along with their respective descriptions.

As can be seen in Fig. 6, the top ranking feature is secondary carbon (SubFPC2), which is a carbon atom with two carbon neighbors. In the context of drug design, such central carbon atom may be more difficult to be accessed and metabolized by cytochrome P450 (*Uetrecht & Trager, 2007*) and therefore are more metabolically stable.

The second most important feature is the rotatable bond (SubFPC302). Based on the rule of three for defining lead-like compounds, a compound may have a lead-like characters if it does have rotatable bonds of no more than 3. On the other hand, *Veber et al. (2002)* noted that the the upper limit of a orally bioavailable drugs is of seven rotatable bonds. Nevertheless, it has been found that number of rotatable bonds provide better discrimination between compounds that are orally active and those that are not. *Kryger, Silman & Sussman (1999)* claimed that E2020 (i.e., also known as donepezil and marketed as Aricept) needs at least two rotatable bonds on each side of the piperidine in which two aromatic moieties of E2020 interact with Trp86 and Trp286 (human AChE numbering), suggesting that links between aromatic systems of the inhibitor against its AChE counterparts are essential to yield high affinity.

The third important substructure is the aromatic ring (SubFPC274). Findings from X-ray crystallographic study showed that in the binding site of the co-crystal structure of AChE with tacrine, the aromatic ring of acridine engages in a π–π stacking interaction with the indole of Trp86 (human AChE numbering), thereby indicating the importance of the aromatic ring for AChE inhibition (*Chen et al. 2012*).

The fourth important feature is C ONS bond (SubFPC295), which is defined as the presence of any carbon connected with either oxygen, nitrogen or sulfur atom in a molecule. These atoms are considered as high electron density atoms, which exerted from higher electronegativity comparing with a carbon atom. Unequal sharing of electron pair making covalent bond contribute polarity and afford dipole moment to a molecule, which able to generate a dipole–dipole attraction such as hydrogen bond between two polar molecules. Furthermore, increasing of polarity and presenting of hydrogen bond improve water solubility, which is essential characteristic of a drug.

The fifth important substructure is secondary mixed amine (SubFPC32). The importance of the moiety was demonstrated in the work of *Bembenek et al. (2008)* in which a structure-based approach was used to reveal that in the catalytic triad, Trp86 interacts with the quaternary amine of ACh through a cation–π interaction. Furthermore, in the

Table 4 **List of top substructure fingerprints and their corresponding description.**

Fingerprints	Description
SubFPC1	Primary carbon
SubFPC2	Secondary carbon
SubFPC3	Tertiary carbon
SubFPC5	Alkene
SubFPC18	Alkylarylether
SubFPC23	Amine
SubFPC26	Tertiary aliphatic amine
SubFPC28	Primary aromatic amine
SubFPC32	Secondary mixed amine
SubFPC35	Ammonium
SubFPC49	Ketone
SubFPC88	Carboxylic acid derivative
SubFPC100	Secondary amide
SubFPC135	Carbonyl derivative
SubFPC137	Vinylogous ester
SubFPC143	Carbonic acid diester
SubFPC153	Urethan
SubFPC171	Arylchloride
SubFPC180	Hetero N basic H
SubFPC181	Hetero N nonbasic
SubFPC182	Hetero O
SubFPC184	Heteroaromatic ring
SubFPC274	Aromatic ring
SubFPC275	Heterocyclic ring
SubFPC276	Epoxide
SubFPC287	Conjugated double bond
SubFPC295	C ONS bond
SubFPC296	Charged
SubFPC298	Cation
SubFPC300	1,3-Tautomerizable
SubFPC301	1,5-Tautomerizable
SubFPC302	Rotatable bond
SubFPC303	Michael acceptor
SubFPC307	Chiral center specified

'anionic' site Trp286 appears to attract the amine moiety via cation and/or hydrophobic interactions.

The sixth, seventh and eighth important substructures are heterocyclic rings (SubFPC275), tertiary carbon (SubFPC3) and primary carbon (SubFPC1). Heterocycles are of high relevance in the design of AChE inhibitors as it allows π–π stacking interaction with key amino acid residues in the binding site of AChE. It is observed that the binding site of the AChE are highly hydrophobic in nature. Particularly, the Trp286 (human AChE numbering) which is a part of the peripheral anionic site of the AChE is involved in the

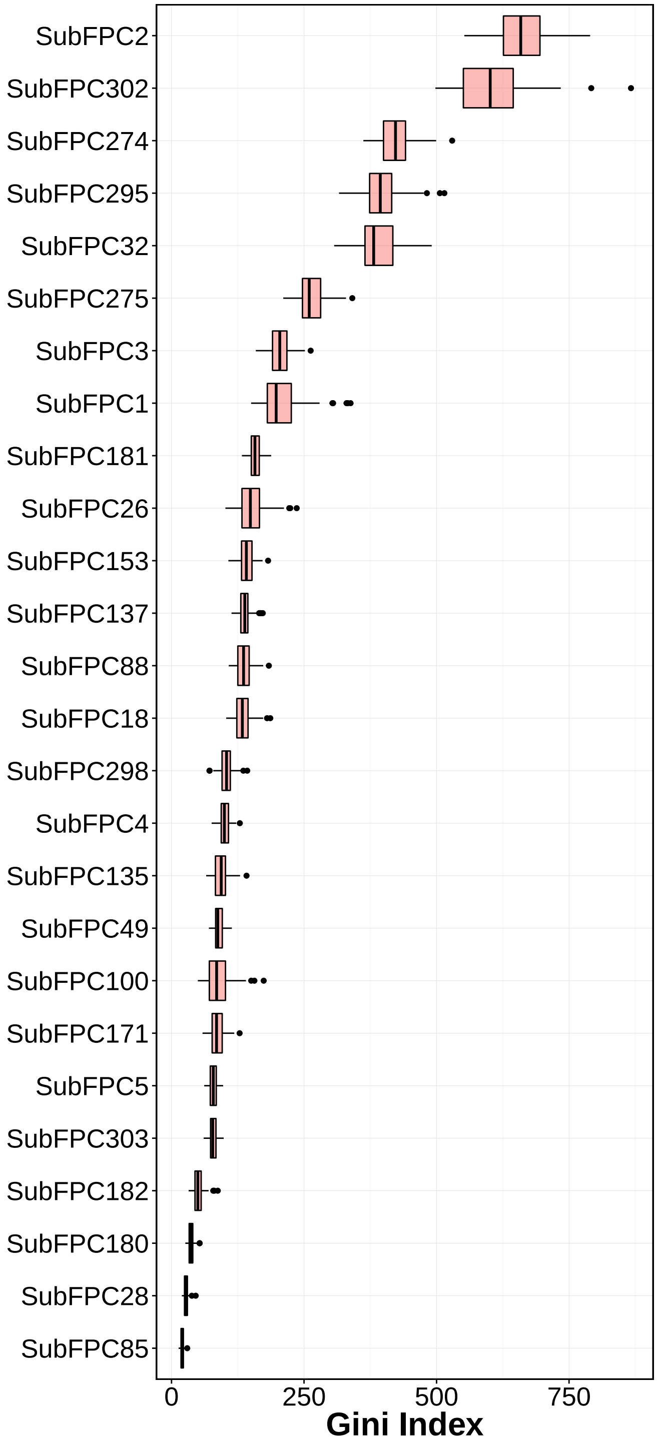


Figure 6 Plot of feature importance as exemplified by the Gini index.

π–π interaction with heterocycles of AChE inhibitors (*Lu et al., 2011*). The aforementioned explanation made for the secondary carbon is also applicable for the tertiary carbon in which the higher number of carbon neighbors would also confer high stability against cytochrome P450 metabolism.

The ninth and tenth important substructures are Hetero N nonbasic (SubFPC181) and tertiary aliphatic amine (SubFPC26) in which the former is defined as aromatic nitrogen. The presence of the nitrogen atom, which is a cationic moiety deemed to interact with aromatic residues through π-cation interaction as observed in E2020 against *Torpedo californica* AChE (tcAChE). The charge of nitrogen atom located in piperidine ring provide π-cation interaction with the side chain of Phe337 (Phe330 in tcAChE numbering) (*Guo et al., 2004*). Since the active site gorge of AChE is comprised of several aromatic residues known as the aromatic patch, the addition of cationic moiety could possibly increase the binding affinity when the ligand is arranged in a suitable conformation with respect to the aromatic side chain of residues in the active site.

It is worthy to note that substructures pertaining to the covalent inhibitors, carbamates and organophosphates, were not found in the top ten important features. A manual inspection of the 307 substructure fingerprints revealed that there were none specifically describing the characteristic feature of carbamate and organophosphates. However, there were quite a few substructures that resembled partial features of the carbamate (e.g., carboxylic acid (SubFPC84), carboxylic ester (SubFPC85), "NOS methylen ester and similar" (SubFPC65), etc.) as well as substructures resembling partial features of organophosphates (e.g., phosphonic monoester (SubFPC230), phosphonic diester (SubFPC231), phosphonic monoamide (SubFPC232), phosphonic esteramide (SubFPC234), phosphonic acid derivative (SubFPC235), etc.). In spite of the presence of these descriptors, important features obtained from predictive model as revealed by the Gini index did not contain these features in the top ten. A possible explanation for such observation could be that there were a few carbamates (123 compounds) and organophosphates (18 compounds) present in the data set and its relative importance may have been masked by other features that represented the other structural class. Thus, it seems very lucrative for a future large-scale study to be performed focusing on these two compound classes.

MOLECULAR DOCKING OF AChE INHIBITORS

To gain a further understanding on the non-covalent interaction between AChE and their inhibitors, a chemically diverse set of 30 representative compounds was extracted from active AChE inhibitors (i.e., having IC_{50} <1 μM) using the Kennard–Stone algorithm and subjected to an investigation on its binding modality against the active site of AChE. Figure 7 shows the distribution of the selected subset of compounds in the context of the full set of actives, which was found to provide a full coverage of the original chemical space. The chemical structures of these compounds are shown in Fig. 8.

The active site of this enzyme is buried inside a narrow gorge of 20 Å deep, which permits multiple enzyme-substrate interaction thereby facilitating the formation of the transition state of ACh (*Silman & Sussman, 2008*; *Zhou, Wang & Zhang, 2010*; *Cheung et*

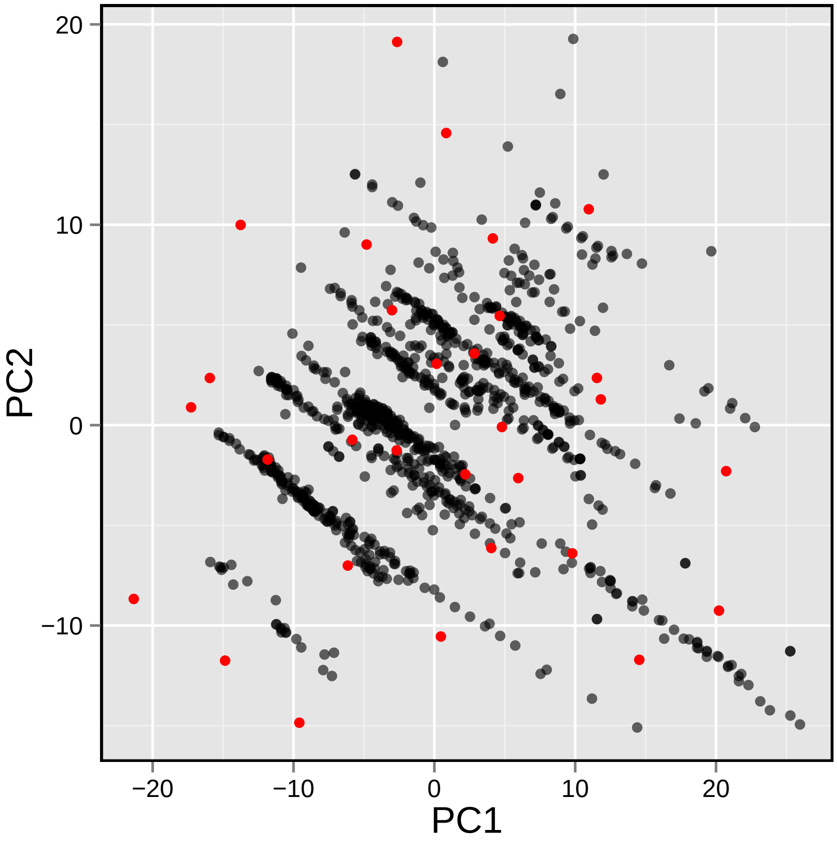


Figure 7 **Plot showing the distribution of active AChE inhibitors (gray circles) and the diversity set (red circles) selected for molecular docking.**

al., 2012). The entry of the active site gorge is lined up by peripheral anionic site (PAS), which is composed of Tyr72, Tyr124, Trp286 and Try341. The function of PAS is to trap the substrate via π-cation interaction and proceed through the constriction residues Tyr124 and Phe338 and onto the catalytic site (*Silman & Sussman, 2008*; *Dvir et al., 2010*). As a serine hydrolase, AChE contains Ser203, Glu334 and His447 in the catalytic triad that catalyzed the acylation and deacylation of ACh. The catalytic triad is surrounded by the catalytic anionic site (CAS) (i.e., contains Trp86, Glu202 and Tyr337), oxyanion hole (i.e., comprising of Glu121, Glu122 and Ala204) and the acyl pocket (i.e., comprised of Phe295 and Phe297). These sites interact with ACh and positions it in a suitable orientation for interacting with the catalytic triad as well as providing proton transfer that is essential for nucleophilic substitution during the catalytic reaction (*Zhou, Wang & Zhang, 2010*). As a result of the acylation process, the proton transfer from Ser203 to His447 induces the oxygen atom of Ser203 to engage in the nucleophilic attack of the carbonyl group of ACh, which consequently breaks down the choline moiety and forms a covalent acylenzyme complex between Ser203 and the acetyl group. This complex consequently proceeds with deacylation, which follows a similar mechanism with the acylation stage. The protonation of Glu202 provides a water molecule and the proton transfer from His447 to the water molecule leads to the nucleophilic attack against the

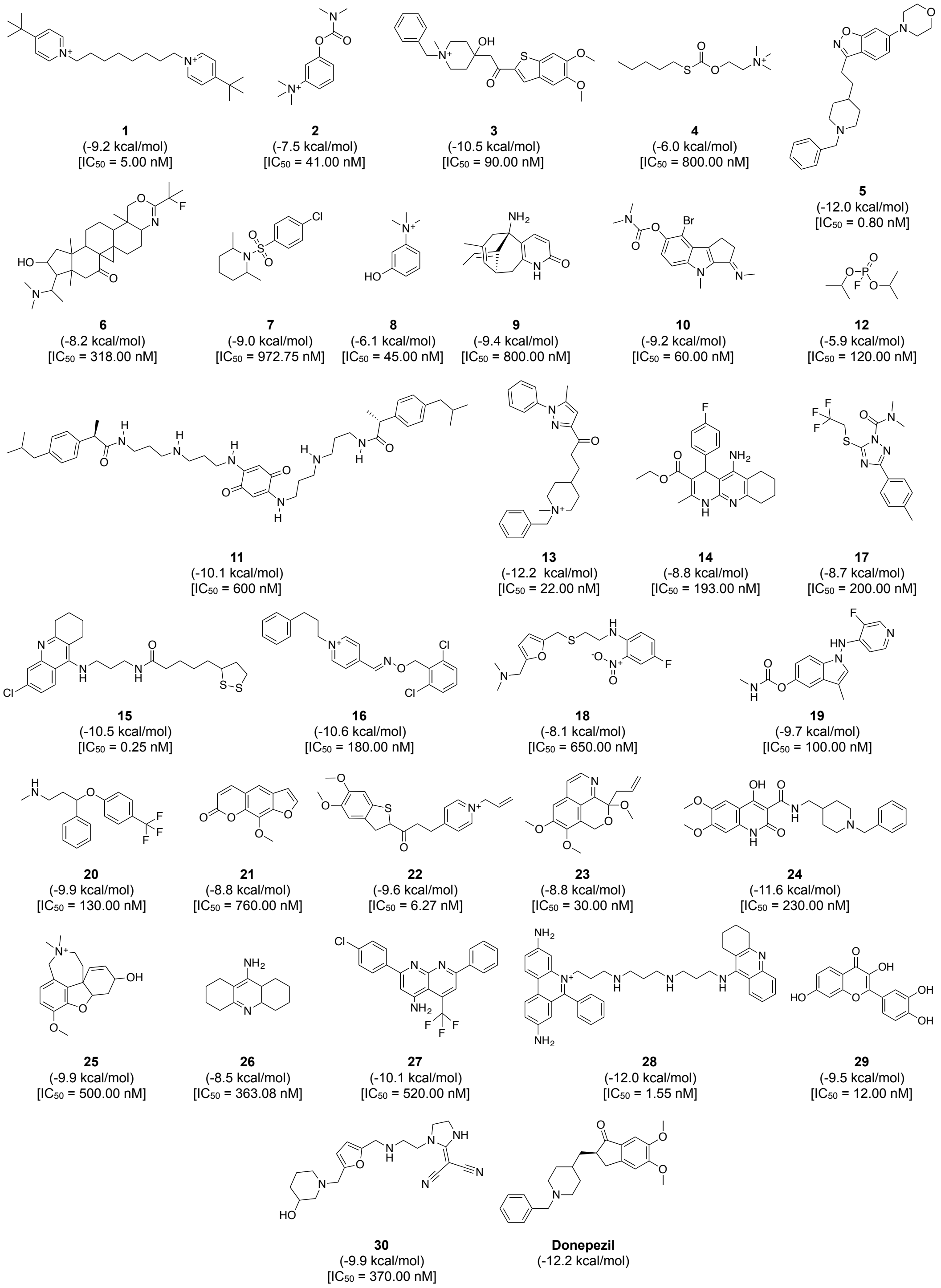


Figure 8 Chemical structures, binding energy and bioactivity of the diversity set consisting of 30 representative compounds from active AChE inhibitors.

acetyl group of the complex. Finally, this results in the breaking down of the complex thereby restoring wild-type AChE and causing the release of acetic acid from the active site (*Zhou, Wang & Zhang, 2010*). In the case of organophosphate and carbamate poisoning, the active site of AChE is phosphorylated or carbamylated, which is no longer capable to hydrolyze the ACh substrate. Organophosphates target and phosphorylates Ser203 in the same manner as that of ACh. However, the phosphoryl moiety is highly stable and leads to the irreversible inhibition of AChE. The mechanism of carbamate inhibition is virtually identical to organophosphate poisoning with the exception that the carbamylated serine moiety is less stable and is therefore able to be regenerated to the active enzyme form. Such understanding on the catalytic mechanism of substrate and inhibitory mechanism of AChE provides useful insights for the development of therapeutic agents targeting AChE (*Fukuto, 1990*).

Prior to carrying out the molecular docking calculations, the docking protocol was validated by re-docking the co-crystal ligand and protein. It was found that the re-docked ligand exhibited negligible deviation from the co-crystal conformation with an RMSD value of 0.963 Å, which was deemed to be suitable for further molecular docking investigation and its subsequent interpretation. Consequently, the binding modality was analyzed in order to gain understanding on the contribution of key residues in interacting with the investigated set of 30 compounds. This was performed using the SiMMap web server, which revealed three major binding anchors: Hbond1, vdW1 and vdW2 along with their site-moiety preferences. The first anchor site involves hydrogen bond interaction between Tyr124 (i.e., an important residue in the PAS that is spatially located as a bottleneck between the peripheral region and the catalytic site of AChE) and the following ligand moieties: secondary amide, secondary amine, nitrogen moiety in aromatic ring, ketone and ester. Such interaction can be observed in the co-crystal structure of huperzine A with human AChE (*Cheung et al., 2012*). Interestingly, analysis of the important features from QSAR models also revealed the importance of "C ONS bond," "secondary mixed amine," "heterocyclic" and "hetero N non-basic" as they were found to be in the top ten important substructures and is therefore crucial for forming hydrogen bonds. Furthermore, the other anchor sites involve van der Waals interaction in which members of the first van der Waals interaction site (vdW1) are comprised of Tyr124, Phe338 and Tyr341, which has a preference to interact with heterocyclic, aromatic, phenol and other non-polar moieties from representative inhibitors. The second van der Waals' interaction site (vdW2) is consisted of Trp86 and Gly121 with preference for the following ligand moieties: aromatic ring, heterocyclic ring, aliphatic moiety with alkene linkage and phenol moiety. These residues contain either bulky aromatic ring or non-polar moiety as their side chain to provide the van der Waal's surface contact against non-polar moiety from ligands. Notably, aromatic and heterocyclic substructures were also observed in the top ten important substructures for predicting the inhibitory activity of human AChE inhibitors.

Furthermore, analysis of the binding energy from the 30 representative compounds revealed that compounds **13**, **5** and **28** exhibited the lowest binding energy of −12.2, −12.0 and −12.0 kcal/mol, respectively, when interacting with the human AChE binding site, which is comparable to donepezil (−12.2 kcal/mol) as indicated in Fig. 8. Key interacting

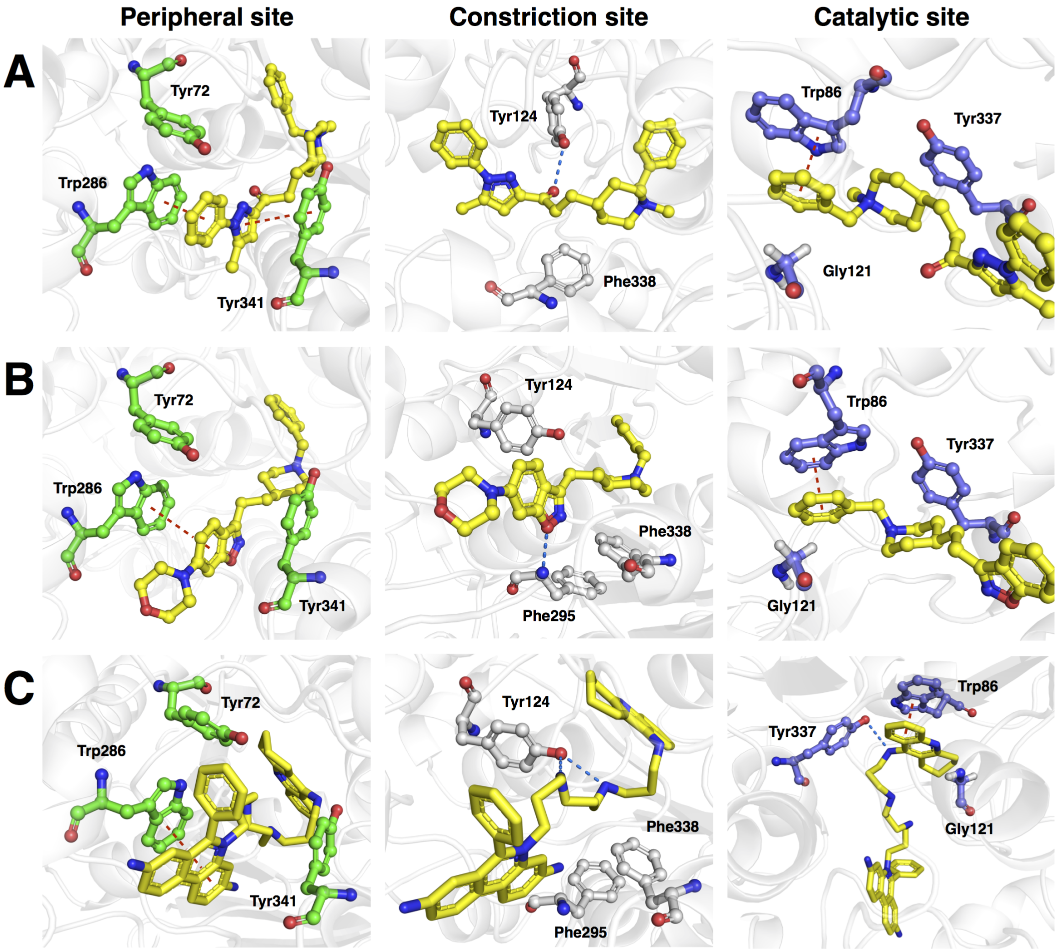


Figure 9 Molecular docking of the top three representative compounds against AChE. The binding between three subsites of AChE with **13**, **5** and **28** are shown in (A, B) and (C), respectively. The representative compounds are highlighted as yellow sticks while residues participating in the peripheral site, constriction site and catalytic site are labeled as green, white and blue sticks, respectively. Hydrogen bonds are shown by blue dash line while the red dash line represents the π–π stacking interaction between ligand and residues in the active site of AChE.

residues and their moiety preference was deduced from the protein–ligand interaction diagram generated by LigPlot+ (*Wallace, Laskowski & Thornton, 1995*) in combination with Maestro (*Schrödinger, 2015b*) and their three-dimensional structure was visualized by PyMOL (*Schrödinger, 2015a*).

Figure 9A revealed three major interaction subsites for **13** inside the binding pocket of human AChE. The first interaction site (Fig. 9A) is formed by residues from the PAS at the gorge opening consisting of Trp286 and Tyr341 both of which engages in π–π stacking interaction where the terminal benzene attached with pyrazole moiety of the ligand interacts with the former residue while the pyrazole moiety of **13** provide contact with the phenol moiety of the latter residue, which are deemed to increase the binding fitness against the active site of human AChE. The second interaction subsite (Fig. 9A) is dominated by Tyr124 and Phe338 in which the former interacts with the ketone moiety of **13** by means of a hydrogen bond, which tends to increase the binding affinity of this compound. The side chain of Phe338 is involved in hydrophobic interaction with carbon atoms from the piperidine ring and the central aliphatic chain of **13**. It can be noted that these residues are members of constriction site, which arranged as bottleneck of active site. The third interaction subsite (Fig. 9A) is moderated by π–π stacking between Trp86 and the terminal benzene with an attached piperidine moiety as well as hydrophobic interaction between Tyr337 and the piperidine moiety of **13** in which both residues belonged to the CAS. In addition, Gly121 of oxyanion hole also interact with terminal benzene of the ligand through hydrophobic contact increasing binding affinity against catalytic site of AChE.

Analysis of the binding modality of compound **5** revealed interactions with all subsites of the AChE active site gorge as illustrated in Fig. 9B. PAS was the first subsite dominated by π-π stacking between Trp286 and the benzisoxazole moiety of compound **5**, which is essential for stabilizing the binding affinity of the ligand against entry into the gorge. The second interaction site was observed at the constriction site in which the piperidine moiety makes contact with Phe338 via hydrophobic interaction thereby increasing the binding fitness against the bottleneck region of the active site. Similar hydrophobic interaction was also observed in the binding pocket in which Tyr337 from CAS interacts with the piperidine moiety and Gly121 of the oxyanion hole interacts with the terminal benzene of compound **5**. The π-π interaction between the terminal benzene and Trp86 from the CAS was deemed to increase the binding fitness with the catalytic site of AChE. In addition, hydrogen bond interaction is facilitated by the nitrogen atom from Phe295 at the acyl pocket to the oxygen atom from the benzisoxazole moiety. Notably, all sites from the active site gorge are snugly bound by compound **5**, which is deemed to exhibit strong intermolecular interaction with human AChE.

The binding energy of compound **28** was similar to that of compound **5** as indicated in Fig. 8. These compounds possessed several aromatic rings at both terminal, which are favourable for interacting with aromatic residues lining up the surface of the gorge and these are known as the aromatic patch. The 5,7-dihydrophenanthridine moiety facilitates π–π stacking with the side chain of Trp286 from PAS at the gorge opening. Meanwhile, this moiety also engages in hydrophobic interaction with Phe295, which tends to increase the binding fitness for the acyl pocket. Aside from the former moiety, 1,2,3,4-tetrahydroacridine at the opposite terminal provides π–π interaction with Trp86 and hydrophobic contact with Tyr337 where both of which are members of CAS in the catalytic site. Furthermore, long aliphatic chain linking the two aromatic moieties provide hydrophobic contact with several aromatic residues in the aromatic patch consisting of Tyr72, Tyr124, Trp286, Tyr337, Phe338 and Tyr341 (i.e., these residues are the members of PAS, CAS and constriction site of the gorge). Moreover, this chain contain several nitrogen atoms, which can act as hydrogen bond donor to Tyr124 and Tyr337 from PAS and CAS, respectively. This would tighten the binding between compound **28** and the active site gorge. The binding modality of this compound was shown in Fig. 9C.

It should be noted that compounds exhibiting strong binding fitness against AChE are those that interact with residues from both PAS and CAS at the entry and inner pocket of the gorge, respectively, as dual-binding site inhibitor through either π–π stacking or

π-cation interaction together with hydrophobic contact. These compounds competes with the natural substrate in interacting with these residues. For non-covalent inhibitors, the aromatic moiety is preferred for occupying the interaction sites while hydrophobic moieties are preferred for making contact with the aromatic residues surrounding the catalytic site. Hydrogen bond donors such as secondary amine and heterocyclic ring can be employed for interacting with the oxygen atom on the side chain of Tyr residues. Interestingly, this finding is corroborated by the feature importance results obtained from the QSAR model as shown in Fig. 6 in which the aromatic moiety, C ONS bond, secondary mixed amine, heterocyclic ring and the hetero N non-basic moiety were found amongst the top ten important substructures that are essential for the bioactivity of AChE inhibitor.

CONCLUSION

In conclusion, twelve sets of fingerprint descriptors were used for constructing QSAR models and their performances were comparatively evaluated. It was observed that several fingerprint descriptors afforded good performance for the constructed models indicating that they could capture the feature space of AChE inhibitors. By taking advantage of the built-in feature importance estimator from RF known as the Gini index, the following important features that are critical for AChE inhibition were identified: secondary carbon (SubFPC2), rotatable bond (SubFPC302), aromatic (SubFPC274), C ONS bond (SubFPC295), secondary mixed amine (SubFPC32) and heterocyclic (SubFPC275). Results from molecular docking also support the aforementioned findings from the QSAR models in which the aromatic, heteroaromatic and heterocyclic rings were preferable moieties for interacting with the hydrophobic pocket of AChE. It is anticipated that the knowledge gained from this study could be used as general guidelines for the design of novel AChE inhibitors.

ACKNOWLEDGEMENTS

The authors would like to thank the peer reviewers and academic editor for providing comments that helped to improve this manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This research is supported by a Research Career Development Grant (No. RSA5780031) and a New Scholar Grant (No. MRG5980220) from the Thailand Research Fund to CN and WS, respectively, as well as the Swedish Research Links program (No. C0610701) from the Swedish Research Council to CN and JESW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Thailand Research Fund: RSA5780031, MRG5980220.
Swedish Research Council: C0610701.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Saw Simeon, Nuttapat Anuwongcharoen, Watshara Shoombuatong and Aijaz Ahmad Malik performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Virapong Prachayasittikul and Jarl E.S. Wikberg analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.
- Chanin Nantasenamat conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

The data set has been supplied as Supplemental Dataset.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.2322#supplemental-information.

REFERENCES

Aaviksaar A. 1990. QSAR in reactions of organophosphorus inhibitors with acetylcholinesterase. *Phosphorus, Sulfur, and Silicon and the Related Elements* **51(1–4)**:47–50 DOI 10.1080/10426509008040679.

Andersson CD, Hillgren JM, Lindgren C, Qian W, Akfur C, Berg L, Ekström F, Linusson A. 2014. Benefits of statistical molecular design, covariance analysis, and reference models in QSAR: a case study on acetylcholinesterase. *Journal of Computer-Aided Molecular Design* **29(3)**:199–215 DOI 10.1007/s10822-014-9808-1.

Beal MF. 1995. Aging, energy, and oxidative stress in neurodegenerative diseases. *Annals of Neurology* **38(3)**:357–366 DOI 10.1002/ana.410380304.

Bembenek SD, Keith JM, Letavic MA, Apodaca R, Barbier AJ, Dvorak L, Aluisio L, Miller KL, Lovenberg TW, Carruthers NI. 2008. Lead identification of acetylcholinesterase inhibitors-histamine H3 receptor antagonists from molecular modeling. *Bioorganic & Medicinal Chemistry* **16(6)**:2968–2973 DOI 10.1016/j.bmc.2007.12.048.

Birks J. 2006. Cholinesterase inhibitors for Alzheimer's disease. *Cochrane Database of Systematic Reviews* (**1**):CD005593.

Bollback JP. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7**:88 DOI 10.1186/1471-2105-7-88.

Bourne Y, Taylor P, Marchot P. 1995. Acetylcholinesterase inhibition by fasciculin: crystal structure of the complex. *Cell* **83(3)**:503–512 DOI 10.1016/0092-8674(95)90128-0.

Breiman L. 2001. Random forests. *Machine Learning* **45(1)**:5–32 DOI 10.1023/A:1010933404324.

Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. 2007. Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia* **3(3)**:186–191 DOI 10.1016/j.jalz.2007.04.381.

Carhart RE, Smith DH, Venkataraghavan R. 1985. Atom pairs as molecular features in structure–activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **25(2)**:64–73.

ChemAxon Kft. 2015. Standardizer. Version 15.9.14.0.

Chen Y, Fang L, Peng S, Liao H, Lehmann J, Zhang Y. 2012. Discovery of a novel acetylcholinesterase inhibitor by structure-based virtual screening techniques. *Bioorganic & Medicinal Chemistry Letters* **22(9)**:3181–3187 DOI 10.1016/j.bmcl.2012.03.046.

Cheung J, Rudolph MJ, Burshteyn F, Cassidy MS, Gary EN, Love J, Franklin MC, Height JJ. 2012. Structures of human acetylcholinesterase in complex with pharmacologically important ligands. *Journal of Medicinal Chemistry* **55(22)**:10282–10286 DOI 10.1021/jm300871x.

Deb PK, Sharma A, Piplani P, Akkinepally RR. 2012. Molecular docking and receptor-specific 3D-QSAR studies of acetylcholinesterase inhibitors. *Molecular Diversity* **16(4)**:803–823 DOI 10.1007/s11030-012-9394-x.

Durant JL, Leland BA, Henry DR, Nourse JG. 2002. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences* **42(6)**:1273–1280 DOI 10.1021/ci010132r.

Dvir H, Silman I, Harel M, Rosenberry TL, Sussman JL. 2010. Acetylcholinesterase: from 3D structure to function. *Chemico–Biological Interactions* **187(1–3)**:10–22 DOI 10.1016/j.cbi.2010.01.042.

Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental Health Perspectives* **111(10)**:1361–1375 DOI 10.1289/ehp.5758.

Eriksson L, Johansson E. 1996. Multivariate design and modeling in QSAR. *Chemometrics and Intelligent Laboratory Systems* **34(1)**:1–19 DOI 10.1016/0169-7439(96)00023-8.

Fang J, Li Y, Liu R, Pang X, Li C, Yang R, He Y, Lian W, Liu A-L, Du G-H. 2015. Discovery of multitarget-directed ligands against Alzheimer's disease through systematic prediction of chemical–protein interactions. *Journal of Chemical Information and Modeling* **55(1)**:149–164 DOI 10.1021/ci500574n.

Fukuto TR. 1990. Mechanism of action of organophosphorus and carbamate insecticides. *Environmental Health Perspectives* **87**:245–254 DOI 10.1289/ehp.9087245.

Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40(D1)**:D1100–D1107 DOI 10.1093/nar/gkr777.

Giacoppo JO, CC França T, Kuča K, Cunha EFD, Abagyan R, Mancini DT, Ramalho TC. 2015. Molecular modeling and *in vitro* reactivation study between the oxime BI-6 and acetylcholinesterase inhibited by different nerve agents. *Journal of Biomolecular Structure and Dynamics* **33(9)**:2048–2058 DOI 10.1080/07391102.2014.989408.

Golbraikh A, Tropsha A. 2002. Beware of q^2! *Journal of Molecular Graphics and Modelling* **20(4)**:269–276 DOI 10.1016/S1093-3263(01)00123-1.

Goldblum A, Yoshimoto M, Hansch C. 1981. Quantitative structure–activity relationship of phenyl N-methylcarbamate inhibition of acetylcholinesterase. *Journal of Agricultural and Food Chemistry* **29(2)**:277–288 DOI 10.1021/jf00104a017.

Gramatica P. 2007. Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science* **26(5)**:694–701 DOI 10.1002/qsar.200610151.

Guo J, Hurley MM, Wright JB, Lushington GH. 2004. A docking score function for estimating ligand–protein interactions: application to acetylcholinesterase inhibition. *Journal of Medicinal Chemistry* **47(22)**:5492–5500 DOI 10.1021/jm049695v.

Gupta S, Mohan CG. 2014. Dual binding site and selective acetylcholinesterase inhibitors derived from integrated pharmacophore models and sequential virtual screening. *BioMed Research International* **2014**:291214 DOI 10.1155/2014/291214.

Hall LH, Kier LB. 1995. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *Journal of Chemical Information and Computer Sciences* **35(6)**:1039–1045 DOI 10.1021/ci00028a014.

Harel M, Schalk I, Ehret-Sabatier L, Bouet F, Goeldner M, Hirth C, Axelsen P, Silman I, Sussman J. 1993. Quaternary ligand binding to aromatic residues in the active-site gorge of acetylcholinesterase. *Proceedings of the National Academy of Sciences of the United States of America* **90(19)**:9031–9035 DOI 10.1073/pnas.90.19.9031.

Huang W, Tang L, Shi Y, Huang S, Xu L, Sheng R, Wu P, Li J, Zhou N, Hu Y. 2011. Searching for the multi-target-directed ligands against Alzheimer's disease: discovery of quinoxaline-based hybrid compounds with AChE, H3R and BACE 1 inhibitory activities. *Bioorganic & Medicinal Chemistry* **19(23)**:7158–7167 DOI 10.1016/j.bmc.2011.09.061.

James G, Witten D, Hastie T, Tibshirani R. 2013. *An introduction to statistical learning: with applications in R*. New York: Springer.

Kennard RW, Stone LA. 1969. Computer aided design of experiments. *Technometrics* **11(1)**:137–148 DOI 10.1080/00401706.1969.10490666.

Kitz R, Wilson IB. 1962. Esters of methanesulfonic acid as irreversible inhibitors of acetylcholinesterase. *Journal of Biological Chemistry* **237(10)**:3245–3249.

Klekota J, Roth FP. 2008. Chemical substructures that enrich for biological activity. *Bioinformatics* **24(21)**:2518–2525 DOI 10.1093/bioinformatics/btn479.

Kryger G, Silman I, Sussman JL. 1999. Structure of acetylcholinesterase complexed with E2020 (Aricept®): implications for the design of new anti-Alzheimer drugs. *Structure* **7**(**3**):297–307 DOI 10.1016/S0969-2126(99)80040-9.

Kuca K, Soukup O, Maresova P, Korabecny J, Nepovimova E, Klimova B, Honegr J, Ramalho TC, França TC. 2016. Current approaches against Alzheimer's disease in clinical trials. *Journal of the Brazilian Chemical Society* **27**(**4**):641–649 DOI 10.5935/0103-5053.20160048.

Kuhn M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**(**5**):1–26.

Laggner C. 2005. SMARTS patterns for functional group classification. Inte:Ligand Software-Entwicklungs und Consulting GmbH. *Available at https://github.com/openbabel/openbabel/blob/master/data/SMARTS_InteLigand.txt*.

Lee S, Barron MG. 2016. A mechanism-based 3D-QSAR approach for classification and prediction of acetylcholinesterase inhibitory potency of organophosphate and carbamate analogs. *Journal of Computer-Aided Molecular Design* **30**(**4**):347–363 DOI 10.1007/s10822-016-9910-7.

Lu S-H, Wu JW, Liu H-L, Zhao J-H, Liu K-T, Chuang C-K, Lin H-Y, Tsai W-B, Ho Y. 2011. The discovery of potential acetylcholinesterase inhibitors: a combination of pharmacophore modeling, virtual screening, and molecular docking studies. *Journal of Biomedical Science* **18**(**8**):Article 22 DOI 10.1186/1423-0127-18-22.

Mager P. 1983. QSAR applied to aging of phosphylated acetylcholinesterase. *Pharmazie* **38**(**4**):271–272.

Morris JC, Cyrus PA, Orazem J, Mas J, Bieber F, Ruzicka BB, Gulanski B. 1998. Metrifonate benefits cognitive, behavioral, and global function in patients with Alzheimer's disease. *Neurology* **50**(**5**):1222–1230 DOI 10.1212/WNL.50.5.1222.

Mundy R, Bowman M, Farmer J, Haley T. 1978. Quantitative structure activity study of a series of substituted 0,0-dimethyl 0-(p-nitrophenyl) phosphorothioates and 0-analogs. *Archives of Toxicology* **41**(**2**):111–123 DOI 10.1007/BF00302523.

Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. 2009. A practical overview of quantitative structure–activity relationship. *EXCLI Journal* **8**(**7**):74–88 DOI 10.17877/DE290R-690.

NCBI. 2009. PubChem Substructure Fingerprint. Version 1.3. *Available at ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt*.

Ordentlich A, Barak D, Kronman C, Flashner Y, Leitner M, Segall Y, Ariel N, Cohen S, Velan B, Shafferman A. 1993. Dissection of the human acetylcholinesterase active center determinants of substrate specificity. Identification of residues constituting the anionic site, the hydrophobic site, and the acyl pocket. *Journal of Biological Chemistry* **268**(**23**):17083–17095.

Prado-Prado FJ, Escobar M, Garcia-Mera X. 2013. Review of bioinformatics and theoretical studies of acetylcholinesterase inhibitors. *Current Bioinformatics* **8**(**4**):496–510 DOI 10.2174/1574893611308040012.

Prado-Prado F, Garcia-Mera X, Escobar M, Alonso N, Caamano O, Yanez M, Gonzalez-Diaz H. 2012. 3D MI-DRAGON: new model for the reconstruction of US FDA drug-target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. *Current Topics in Medicinal Chemistry* **12**(**16**):1843–1865 DOI 10.2174/156802612803989228.

Puzyn T, Mostrag-Szlichtyng A, Gajewicz A, Skrzyński M, Worth AP. 2011. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Structural Chemistry* **22**(**4**):795–804 DOI 10.1007/s11224-011-9757-4.

Quinn DM. 1987. Acetylcholinesterase: enzyme structure, reaction dynamics, and virtual transition states. *Chemical Reviews* **87**(**5**):955–979 DOI 10.1021/cr00081a005.

Racchi M, Mazzucchelli M, Porrello E, Lanni C, Govoni S. 2004. Acetylcholinesterase inhibitors: novel activities of old molecules. *Pharmacological Research* **50**(**4**):441–451 DOI 10.1016/j.phrs.2003.12.027.

Riniker S, Landrum GA. 2013. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics* **5**:Article 26 DOI 10.1186/1758-2946-5-1.

Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN. 2013. Some case studies on application of "r_m^2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. *Journal of Computational Chemistry* **34**(**12**):1071–1082 DOI 10.1002/jcc.23231.

Schrödinger LLC. 2015a. PyMOL Molecular Graphics System. Version 1.7.6.3. *Available at https://www.schrodinger.com/pymol/.*

Schrödinger LLC. 2015b. Maestro. Version 10.4. *Available at https://www.schrodinger.com/Maestro/.*

Shen T, Tai K, Henchman RH, McCammon JA. 2002. Molecular dynamics of acetylcholinesterase. *Accounts of Chemical Research* **35**(**6**):332–340 DOI 10.1021/ar010025i.

Silman I, Sussman JL. 2008. Acetylcholinesterase: how is structure related to function? Proceedings of the IX international meeting on cholinesterases, *Chemico-Biological Interactions* **175**(**1–3**):3–10 DOI 10.1016/j.cbi.2008.05.035.

Simeon S, Möller R, Almgren D, Li H, Phanus-umporn C, Prachayasittikul V, Bülow L, Nantasenamat C. 2016. Unraveling the origin of splice switching activity of hemoglobin β-globin gene modulators via QSAR modeling. *Chemometrics and Intelligent Laboratory Systems* **151**:51–60 DOI 10.1016/j.chemolab.2015.12.002.

Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. 2003. The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences* **43**(**2**):493–500 DOI 10.1021/ci025584y.

Su C, Lien E. 1980. QSAR of acetylcholinesterase inhibitors: a reexamination of the role of charge-transfer. *Research Communications in Chemical Pathology and Pharmacology* **29**(**3**):403–415.

Tougu V. 2001. Acetylcholinesterase: mechanism of catalysis and inhibition. *Current Medicinal Chemistry—Central Nervous System Agents* **1**(**2**):155–170 DOI 10.2174/1568015013358536.

Tropsha A, Gramatica P, Gombar VK. 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science* **22(1)**:69–77 DOI 10.1002/qsar.200390007.

Trott O, Olson AJ. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **31(2)**:455–461 DOI 10.1002/jcc.21334.

Uetrecht J, Trager W. 2007. *Drug metabolism: chemical and enzymatic aspects.* Boca Raton: CRC Press.

Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD. 2002. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* **45(12)**:2615–2623 DOI 10.1021/jm020017n.

Veselinović J, Nikolić G, Trutić N, Živković J, Veselinović A. 2015. Monte Carlo QSAR models for predicting organophosphate inhibition of acetycholinesterase. *SAR and QSAR in Environmental Research* **26(6)**:449–460 DOI 10.1080/1062936X.2015.1049665.

Wallace AC, Laskowski RA, Thornton JM. 1995. LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Engineering* **8(2)**:127–134 DOI 10.1093/protein/8.2.127.

Walters D, Hopfinger A. 1986. Case studies of the application of molecular shape analysis to elucidate drug action. *Journal of Molecular Structure* **134(3–4)**:317–323 DOI 10.1016/0166-1280(86)80004-5.

Wright MN, Ziegler A. 2015. ranger: a fast implementation of random forests for high dimensional data in C++ and R. ArXiv preprint. arXiv:1508.04409.

Yan A, Wang K. 2012. Quantitative structure and bioactivity relationship study on human acetylcholinesterase inhibitors. *Bioorganic & Medicinal Chemistry Letters* **22(9)**:3336–3342 DOI 10.1016/j.bmcl.2012.02.108.

Yap CW. 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **32(7)**:1466–1474 DOI 10.1002/jcc.21707.

Zhou Y, Wang S, Zhang Y. 2010. Catalytic reaction mechanism of acetylcholinesterase determined by Born–Oppenheimer ab initio QM/MM molecular dynamics simulations. *Journal of Physical Chemistry B* **114(26)**:8817–8825 DOI 10.1021/jp104258d.

RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2017, 7, 28056

Exploring the origin of phosphodiesterase inhibition *via* proteochemometric modeling

Behnam Rasti, *[a] Nalini Schaduangrat, [b] S. Shirin Shahangian [c] and Chanin Nantasenamat *[b]

The phosphodiesterase (PDE) superfamily, including all PDE families and subfamilies, are often implicated in diverse physiological disorders thereby making their selective inhibition of great necessity. Of the PDE4 family, the subfamilies of PDE4B and PDE4D have attracted attention due to their role in highly critical disorders such as asthma, acrodysostosis, cognition disorder and schizophrenia. Owing to their different levels of involvement in related disorders and within different subcellular compartments, the development of specific subfamily-selective compounds seems pertinent. Since achieving selectivity can be facilitated by considering the information of both compound and protein, thereby calling for proteochemometrics (PCM) to investigate the interaction space and selectivity of different chemical compounds towards different PDE4 isoforms. Several internal and external data sets were applied to validate the predictivity of the PCM model for interpolating on internal compounds as well as extrapolating on newly designed compounds. The *Y*-scrambling approach was applied to evaluate the possibility of chance correlation. Excellent values of 0.9973, 0.9037 and 0.9742 were observed for the training (R^2), internal cross-validation (Q^2) and external validation set ($Q_{\text{ext}}{}^2$), respectively. Practical utilization of this information was demonstrated *via* the design of a few novel compounds whereby structural changes to the compound can exert effects on the selectivity against both PDE4B and PDE4D. Our model provided knowledge on the structural features of compounds in order to discriminate the binding of PDE4B and PDE4D, which is valuable for the promising design of selective inhibitors.

Received 24th February 2017
Accepted 16th May 2017

DOI: 10.1039/c7ra02332d

rsc.li/rsc-advances

Introduction

Phosphodiesterases (PDEs) (EC 3.1.4.17) catalyze the production of 5′-AMP and 5′-GMP *via* the degradation of cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP), respectively.[1] These secondary messengers are crucial in physiological processes such as cell growth, apoptosis, immune responses, reproduction, inflammatory responses, *etc.*[2,3] Consequently, a lot of major disorders like obesity, diabetes, heart failure, arthritis, chronic obstructive pulmonary disease (COPD) and cancer can be engaged with PDE deficiency.[4–6] The PDE superfamily is comprised of eleven families. Among them, the PDE families are divided into cAMP-specific families (PDE4, PDE7, and PDE8), cGMP-specific families (PDE5, PDE6, and PDE9) as well as PDEs families with dual specificity (PDE1, PDE2, PDE3, PDE10, and PDE11).[5,6] Since each family posses unique tissue distribution, substrate specificity and functional properties, their inhibition can result in different biological outcomes.[7–14] In addition, PDE4 isoforms are mainly expressed in inflammatory and immune cells and are known to be involved in disorders such as asthma and COPD. Therefore, developing PDE4-selective inhibitors for such disorders has been of great interest to pharmaceutical companies. Two advanced inhibitors named cilomilast and roflumilast are in their phase III clinical trials.[5,15–18] Furthermore, due to promising clinical advancements in developing PDE4 inhibitors, there is also a great appeal growing towards developing specific inhibitors against PDE4 subfamilies including PDE4B and PDE4D.

Previous investigations have shown that selective inhibition of PDE4D is associated with a reduction of inflammation and improvement of cognition, while PDE4B-selective inhibitors seem to be potent therapeutics for allergic inflammation and asthma.[19–27] It has also been shown that people with mutations or single-nucleotide polymorphisms (SNPs) in their PDE4D are engaged with acrodysostosis[25] and ischaemic stroke[26] whereas PDE4B SNPs coupled with low levels of PDE4B expression are associated with schizophrenia.[27] Other studies with knockout mice proposed that the development of PDE4B-selective inhibitors for use in asthma and other inflammation-related diseases can cause fewer side effects as compared to classical PDE4 inhibitors.[28] To sum up, since the PDE4 subfamily are involved

[a] *Department of Microbiology, Faculty of Basic Sciences, Lahijan Branch, Islamic Azad University (IAU), Lahijan, Guilan, Iran. E-mail: rasti@liau.ac.ir*
[b] *Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand. E-mail: chanin.nan@mahidol.edu*
[c] *Department of Biology, Faculty of Sciences, University of Guilan, Rasht 41938-33697, Iran*

in different signalosomes within different subcellular compartments, the development of specific therapeutics and the design of novel isoform-selective inhibitors seem highly necessary. Therefore, in order to investigate the selectivity, the interaction space of multiple compounds across multiple proteins should be analyzed.

The proteochemometrics (PCM) approach fulfills this goal by correlating both compound and protein descriptions with biological activities.[29] Due to the crucial physiological impact resulted upon selective inhibition of different PDE4 subclasses, as well as owing to the fact that there have been no PCM studies regarding the PDE family as yet, therefore, here for the first time we applied the PCM approach to study the interaction space of PDE4 subfamilies and their inhibitors. In addition, thus far PCM have been successfully applied to investigate protein families such as G protein-coupled receptors,[30,31] proteases,[32,33] kinases,[34–36] antibodies,[37] cytochrome P450 [38,39] and carbonic anhydrase.[40,41] While the majority of these researches have used sequence based descriptors for describing proteins, whereby the latter has shown a positive impact on molecular interaction field (MIF) based descriptors called GRid-INdependent Descriptors (GRIND) on modeling and on the significance of using *z* scales-GRINDs combinatorial descriptors.[41] Hence, in the present study, we developed a unified PCM model with the combination of *z* scales and MIF-based GRIND to investigate the interaction space and the selectivity between two subfamilies of PDE4 (PDE4B and PDE4D) and a series of their inhibitors. This approach provides the ability to find differential structural features that can be taken into consideration for designing compounds with better selectivity towards PDE4B and PDE4D. The flowchart for PCM modeling of the compound–protein interaction space is represented in Fig. 1.

Materials and methods

Data set

The bioactivities of compounds were obtained from the BindingDB,[42,43] which is a publicly available database containing nearly 20 000 experimentally determined bioactivities of compound–protein complexes. There are 983 and 853 biological activities deposited in the BindingDB for PDE4B and PDE4D, respectively. Some of the data belongs to organisms other than humans and some do not have valid compound IDs. In some cases the reported activity has neither the type of interest nor an exact measured value (*e.g.* IC_{50} > 1000). There are also some cases in which the activity of an inhibitor is reported for one of the PDE4 isoforms. After initial filtration of the data set according to the points mentioned above, a set of 71 compounds with inhibitory data available for both PDE4B and PDE4D was selected based on the following extra filtration steps: (i) the difference between their inhibitory powers (*i.e.* IC_{50} of nanomolar potency) for the two isoforms of PDE was less than or equal to 2-fold. In other words, we made sure that the chosen compounds were not selective inhibitors in the case of isoforms PDE4B and PDE4D, (ii) compounds with more than one value reported for their IC_{50} were removed from the data set. Inhibitory activities of compounds were converted to pIC_{50} ($-\log IC_{50} \times 10^{-9}$) and the values are available in a CSV file provided on GitHub at https://github.com/chaninn/PDE4/.

Protein descriptors

Structures of human PDE4B (PDB ID: 3O0J) and PDE4D (PDB ID: 2PW3) were obtained from the Protein Data Bank (Fig. 10). Available PDB structures for PDE4B and PDE4D represent the catalytic domain of the proteins. Since GRIND descriptors are extracted from the 3D structures of proteins, the presence of the missing residues in the PDB files can negatively affect the process of descriptor calculation. Therefore, we chose 3O0J and 2PW3 structures as there were no missing residues reported in their PDB files. In addition, since 2PW3 represents the structure of PDE in complex with its native substrate, we used 2PW3 as the reference structure to extract those residues that make up the enzyme's cavity and have the potential to interact with the compounds. From the center of the cAMP, a cutoff of 10 Å was used and those residues falling within the applied cutoff were considered as compound interacting residues. The sequence of 3O0J was then aligned using Clustal Omega web server, version 1.2.4 [44] and cavity amino acids were identified in correspondent positions for the PDE4B isoform (Fig. 2). In accordance with our previous work,[41] albeit slight changes, we modified the ALMOND algorithm and introduced an ALMOND-like algorithm with the ability to calculate the GRIND descriptors for complex structures such as protein cavity. Briefly, our algorithm works by following 3 steps: (i) calculating MIFs using the program GRID[45] and finding the points with favorable interaction energies, (ii) reducing the points to those showing the best interaction energies, using the genetic algorithm.[46,47] Both the intensity of a point and the mutual node–node distances between the selected points are considered in the filtering process, (iii) for each MIF pairs of interaction, energies are multiplied and the greatest product is kept for each internode distance. This way a set of the chosen node is converted to a set of descriptors.

DRY (–CH3), O (carbonyl oxygen), N1 (amide nitrogen) and TIP probes were used for computing the MIFs. These probes are representatives for hydrophobic interactions, hydrogen bond acceptors, hydrogen bond donors and molecular shape, respectively. A grid spacing of 0.5 Å was applied while the center and size of the GRID box were adjusted to cover all compound interacting residues. The following parameters were applied in the case of the genetic algorithm: (i) population size of 200, (ii) maximum generation of 200, (iii) two-point crossovers with the rate of 0.8, and (iv) mutations with the rate of 0.01. Regarding each MIF, 2000 nodes were extracted and the smoothing window of 0.2 Å was used, resulting in 224 descriptors for each auto/cross MIF–MIF multiplication. Since we applied four types of probes (resulting in a final number of 10 auto/cross MIF–MIF combinations), the total number of 2240 descriptors for each protein isoform were calculated. Descriptors showing the same value for PDE4B and PDE4D were removed, resulting in the final number of 2200 descriptors for each protein isoform.

z-Scales were calculated for non-conserved compound interacting residues (8 residues). These residues were encoded

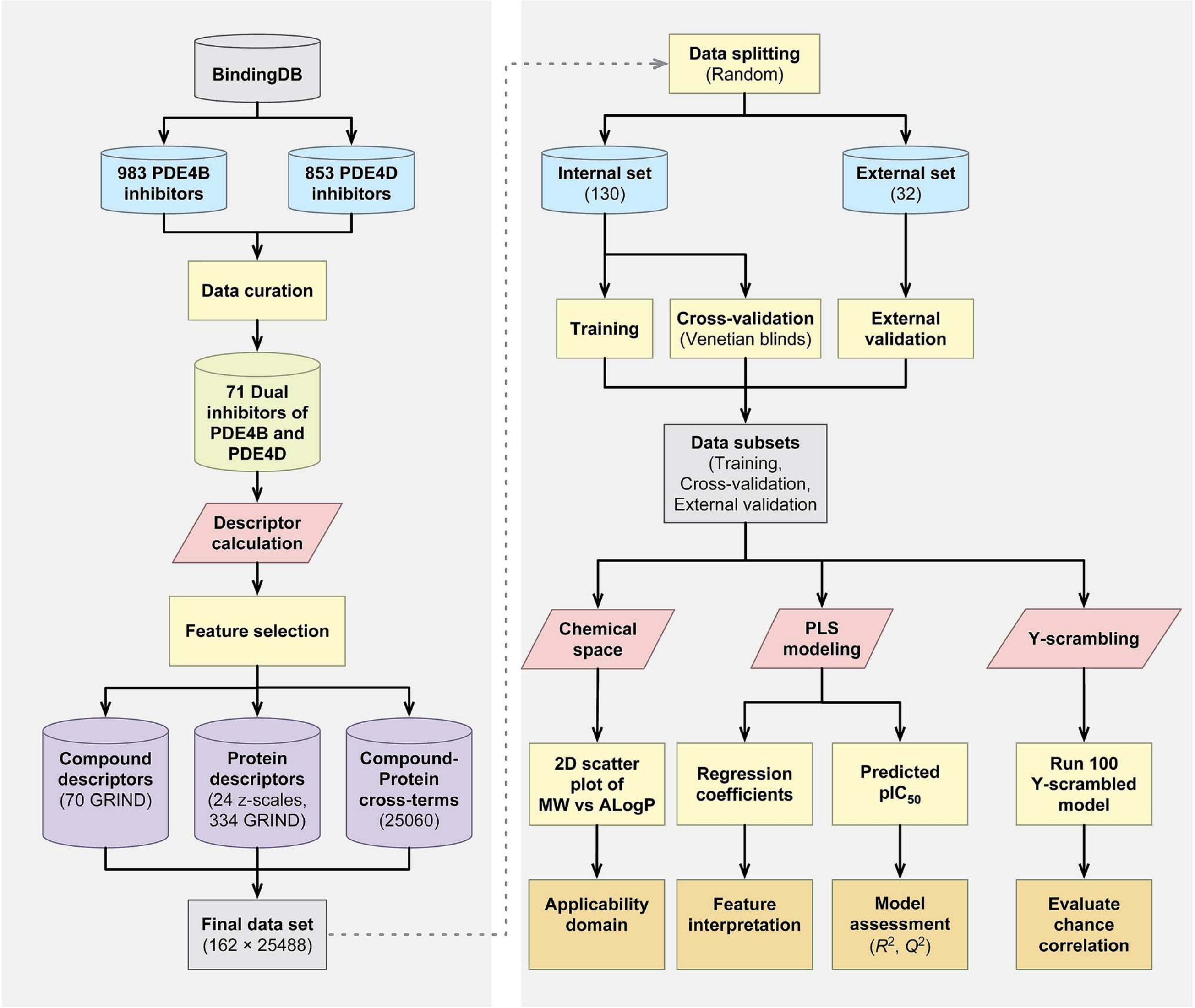

BindingDB
983 PDE4B inhibitors
853 PDE4D inhibitors
Data curation
71 Dual inhibitors of PDE4B and PDE4D
Descriptor calculation
Feature selection
Compound descriptors (70 GRIND)
Protein descriptors (24 z-scales, 334 GRIND)
Compound-Protein cross-terms (25060)
Final data set (162 × 25488)
Data splitting (Random)
Internal set (130)
External set (32)
Training
Cross-validation (Venetian blinds)
External validation
Data subsets (Training, Cross-validation, External validation
Chemical space
PLS modeling
Y-scrambling
2D scatter plot of MW vs ALogP
Regression coefficients
Predicted pIC$_{50}$
Run 100 Y-scrambled model
Applicability domain
Feature interpretation
Model assessment (R^2, Q^2)
Evaluate chance correlation

Fig. 1 Flowchart summarizing the PCM modeling of PDE4 inhibitors.

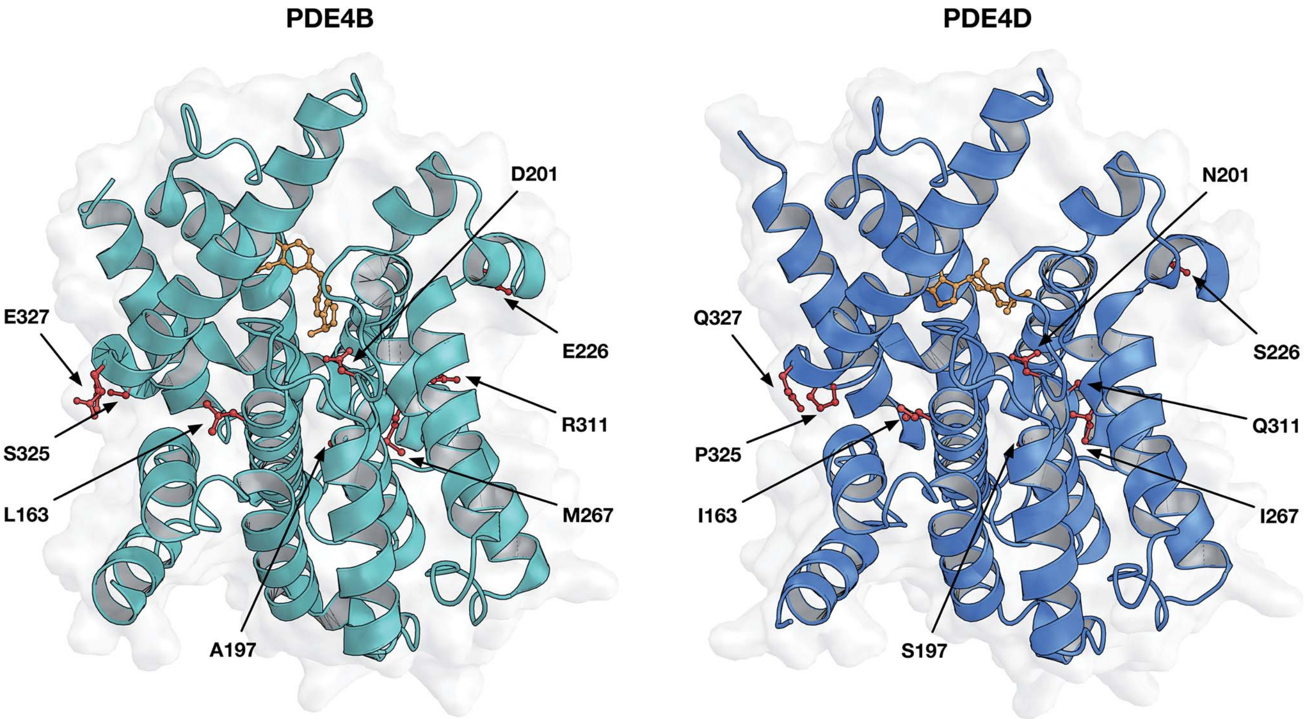
by three *z*-scale descriptors (z_1, z_2 and z_3), as derived by Sandberg *et al.*,[48] representing hydrophobicity, size/polarizability and polarity, respectively. Using these three *z*-scale descriptors we reached a total number of 8 × 3 = 24 descriptors for each protein.

Compound descriptors

Structures were obtained from ZINC website.[49,50] All geometries were optimized using SYBYL version 7.3 [51] and the Tripos force field was applied with a distance-dependent dielectric and the Powell conjugate gradient algorithm convergence criterion of 0.01 kcal mol^{-1} Å^{-1}. To calculate the partial atomic charges, the Gasteiger–Huckel method was used. GRIND 3D descriptors were calculated using the same algorithm as the one applied for calculating the descriptors of PDEs. The following modifications were considered when compared to proteins, (i) the GRID box was adjusted in a way that the whole compound was placed within, (ii) since the structure of compounds are less complex when compared to that of proteins, the number of extracted nodes and smoothing window were 100 and 0.4 Å for each MIF, respectively. As there are ten descriptor types (*i.e.* DRY–DRY, O–O, N1–N1, *etc.*) and each descriptor type has 62 distance iterations (starting from 0.4 and ending at 0.4 × 62 = 24.8), therefore upon applying the mentioned parameters, a final set of 620 descriptors was obtained for compounds. Descriptors showing the same value for all compounds were removed thereby resulting in a final number of 495 descriptors for each inhibitor.

Feature selection

Feature selection was applied to select the best fitted GRIND descriptors. GA-PLS consists of three basic steps: (i) generation of the initial chromosomes. Each chromosome contains different genes representing the presence/absence of a variable, (ii) calculation of Q^2 parameter to evaluate the fitness of each chromosome, (iii) reproduction in which processes such as crossing-over and mutation were carried out. Steps (ii) and (iii)



```
       80         90         100        110        120        130        140        150        160
PDE4D        TEQE DVLAKELEDV NKWGLHVFRI AELSGNRPLT VIMHTIFQER DLLKTFKIPV DTLITYLMTL EDHYHADVAY HNNIHAADVV
PDE4B        --NE DHLAKELEDL NKWGLNIFNV AGYSHNRPLT CIMYAIFQER DLLKTFRISS DTFITYMMTL EDHYHSDVAY HNSLHAADVA
               :* * *******: *****::*.: *  * *****  **::***** ******:*   **:***:*** *****:**** **.:*****.

       170        180        190        200        210        220        230        240        250
PDE4D  QSTHVLLSTP ALEAVFTDLE ILAAIFASAI HNVDHPGVSN QFLINTNSEL ALMYNDSSVL ENHHLAVGFK LLQEENCDIF QNLTKKQRQS
PDE4B  QSTHVLLSTP ALDAVFTDLE ILAAIFAAAI HDVDHPGVSN QFLINTNSEL ALMYNDESVL ENHHLAVGFK LLQEEHCDIF MNLTKKQRQT
       ********** **:******* *******:** *:******** ********** ******.*** ********** *****:****  ********:

       260        270        280        290        300        310        320        330        340
PDE4D  LRKMVIDIVL ATDMSKHMNL LADLKTMVET KKVTSSGVLL LDNYSDRIQV LQNMVHCADL SNPTKPLQLY RQWTDRIMEE FFRQGDRERE
PDE4B  LRKMVIDMVL ATDMSKHMSL LADLKTMVET KKVTSSGVLL LDNYTDRIQV LRNMVHCADL SNPTKSLELY RQWTDRIMEE FFQQGDKERE
       *******:** ********.* ********** ********** ****:***** *:******** ***** *:** ********** **:***:***

       350        360        370        380        390        400        410
PDE4D  RGMEISPMCD KHNASVEKSQ VGFIDYIVHP LWETWADLVH PDAQDILDTL EDNREWYQST IPQ
PDE4B  RGMEISPMCD KHTASVEKSQ VGFIDYIVHP LWETWADLVQ PDAQDILDTL EDNRNWYQSM I--
       ********** **.******* ********** *********: ********** ****:****  *
```

Fig. 2 Protein structures (top panel) and sequence alignment (bottom panel) of investigated phosphodiesterase isoforms PDE4B (PDB ID: 3O0J) and PDE4D (PDB ID: 2PW3). Environmental residues (indicated by red colored residues in the top panel and red text in the bottom panel) surrounding the cAMP substrate (colored orange) were selected for further descriptor generation.

continues until the designated number of generations was achieved.[47,52] PLS Toolbox 3.5 [53] was used and the genetic algorithm with default parameters was applied on both the compound and protein descriptors. The final numbers of 70 and 334 GRINDs were selected by the genetic algorithm for compounds and proteins, respectively.

Compound–protein cross-terms

In the present study, cross-terms are simply mathematical products of the compound descriptors with those of the proteins, representing the interaction space between compounds and proteins. Since compound and protein descriptors are 70 and 358 (24 *z*-scales + 334 GRINDs) respectively, the total number of cross-terms for each compound–protein complex is 25 060 (70 × 358). All descriptors were mean-centered and scaled to unit variance before cross-terms calculation. To prevent small blocks of descriptors being masked by large ones, we used block-scaling.

Multivariate modeling

Descriptors of the compounds, protein cavity residues and their cross-terms were correlated to the pIC_{50}s using partial least squares (PLS) regression. Briefly, PLS correlates the *X* matrix of predictors with the *Y* response variables by simultaneously projecting them to the PLS components and finding linear relationships between them. PLS modeling was performed using PLS Toolbox 3.5. The optimal number of latent variables to use for the construction of the PCM model was selected according to the method of Haaland and Thomas,[54] which resulted in the use of 14 latent variables.

For a PCM model consisting of P protein descriptors, L compound descriptors and $C \times P$ cross-terms, the regression equation is expressed as follows:

$$\mathrm{pIC}_{50} = \overline{\mathrm{pIC}_{50}} + \sum_{c=1}^{C} C_c {D_c}^C + \sum_{p=1}^{P} C_p {D_p}^P + \sum_{c=1,p=1}^{C\times P} C_{cp} {D_c}^C {D_p}^P \quad (1)$$

where $\overline{\mathrm{pIC}_{50}}$ represents the average pIC_{50}; D_c and D_p represent compound descriptors and protein descriptors, respectively; C_c, C_p and C_{cp} are regression coefficients of compound descriptors, protein descriptors and compound–protein cross-terms, respectively.

Model validation

To assess the ability of the model for predicting the IC_{50} of a new compound, data splitting was applied to select 20% as an external set while using 80% as the internal set. This study makes use of two types of external set that are termed as follows: (i) external-compounds and (ii) external-complexes. In the former, 12 compound–protein complexes were randomly excluded from the modeling process while in the latter 10 compounds (and its associated bioactivity against the two isoforms, which results in 20 bioactivity data points or compound–protein complexes) were excluded as the external set. Since information related to these complexes has not been seen in the PLS model, it is assumed that they could not have had an influence on the PLS model. Thus, we applied these two external sets consisting of 32 compound–protein complexes (making up the 20% subset) to evaluate the extrapolation capability of the model whereas the internal set consisting of 130 compound–protein complexes (constituting the 80% subset) was evaluated for its intrapolation capability. Thus, the internal set was used as both a training set as well as subjected to Venetian blinds cross-validation (VB-CV). Furthermore, the predictive model trained using the internal set was applied on the external sets as to evaluate the general assessment of the model's predictive power in regards to its bioactivity and selectivity toward new compounds.

Y-Scrambling was applied to test the robustness of the models. In this approach the variable Y is randomly shuffled, and a new model with scrambled data is generated to ensure the robustness of the PCM models and to rule out the possibility of chance correlations. Therefore, we built 100 new models using randomly shuffled variable Y. The R^2 and Q^2 values of scrambled and unscrambled models were plotted *versus* correlation coefficients between original and scrambled Y values. Regression line was conducted and the intercepts for R^2 and Q^2 (R^2 intercept and Q^2 intercept) were calculated.

Previous studies have revealed that in order to ensure the robustness of the models and to rule out the possibility of chance correlations, the R^2 intercept and the Q^2 intercept should not exceed 0.3 and 0.05, respectively.[55] Along with these validations, we randomly selected from BindingDB 10 new compounds with different selectivity for PDE4B and PDE4D and used this set as an external set to validate the power of the model for selectivity prediction of new inhibitors. Since information related to these compounds have not been used in the PLS model, this way we could make sure of the reliability of our model for predicting the selectivity of newly designed inhibitors, which we term the external-compound set.

Applicability domain analysis

Applicability domain (AD) was applied to estimate the likelihood of reliable prediction for the investigated compounds. The uncertainty of predictions refers to the number of compounds falling outside the AD. The most popular method for determining AD was described by Gramatica *et al.*[56] and Tropsha *et al.*,[57] which encompasses the computation of leverage values for each investigated compound. Using the leverage value we can identify whether a new compound falls within or outside the domain. Leverage values are calculated *via* adjustment of X as to yield the hat matrix H:

$$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \quad (2)$$

where X is a two-dimensional matrix made of n compounds and m descriptors whereas X^{T} is the transpose of X. Meantime, the leverage value of the i^{th} compound (h_i) is the i^{th} diagonal element of H:

$$h_i = x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i \quad (3)$$

where x_i is the descriptor row-vector of the i^{th} compound. The warning leverage h^* is calculated by:

$$h^* = 3(p + 1)/n \quad (4)$$

The leverage value along with the William's plot is usually applied to assess the AD of QSAR models. The William's plot is built by depicting the standardized residuals *versus* the leverage value for each compound's h_i. If the i^{th} compound has $h_i > h^*$ then it means that the i^{th} compound applies a great impact on the QSAR model and may be excluded from the AD.

Contribution of compound properties to protein selectivity

Since the significance of a compound descriptor to protein selectivity can be obtained from coefficients of cross-terms involving descriptors of interest, we applied the following equation to assess the contribution of a GRIND descriptor to the selectivity for a particular protein isoform.

$$\Delta y_{cp} = \frac{\mathrm{d}y}{\mathrm{d}D_c} = C_c + \sum_{p=1}^{P} C_{cp} D_p \quad (5)$$

where y_{cp} represents the change in selectivity of the c^{th} descriptor of compounds for a particular protein isoform p with descriptors D_1, $D_2 \ldots D_p$. $\mathrm{C_c}$ and C_{cp} denotes regression coefficients while $D_c(P + 1 \leq c \leq P + C)$ and $D_p(1 \leq p \leq P)$ denotes descriptors for compounds and protein isoforms, respectively.

Results and discussion

Non-specific inhibitors may inhibit several isoforms of phosphodiesterase, resulting in toxic side effects. Regarding the inhibition of PDE4D isoform in particular, with available developed inhibitors of PDE4, there are findings which suggest that inhibition of PDE4D is associated with the dose-limiting gastrointestinal side effects, while PDE4B seems to play major roles in activation of the T-cell receptor.[5,28,58–60] These investigations and studies similar to them, can justify the rational for the development of selective inhibitors for PDE4B and PDE4D. While the former can act as potential therapeutics for allergic inflammation and asthma,[20] the latter can reduce inflammation and improve cognition.[61] Since PDE4B and PDE4D are subclasses of PDE isoform 4, the sequences of their catalytic domains are substantially conserved and indeed their structures are highly similar. Therefore, designing compounds with the ability of discriminating one subclass from the other is a significantly challenging task. Considering that the properties of proteins are exploited in addition to the features of compounds in PCM modeling, PCM models would be able to catch differences in patterns of chemical interactions with regard to different compound–protein complexes. Furthermore, in the case of isoforms and their subclasses (*e.g.* PDE4) which show high sequence/structural similarities, PCM can catch chemical space differences, even those raised by the slightest sequence/structural dissimilarities. Therefore, using the PCM approach, we could capture some structural features that can be considered while designing compounds with higher selective tendencies towards a specific subclass of PDE4.

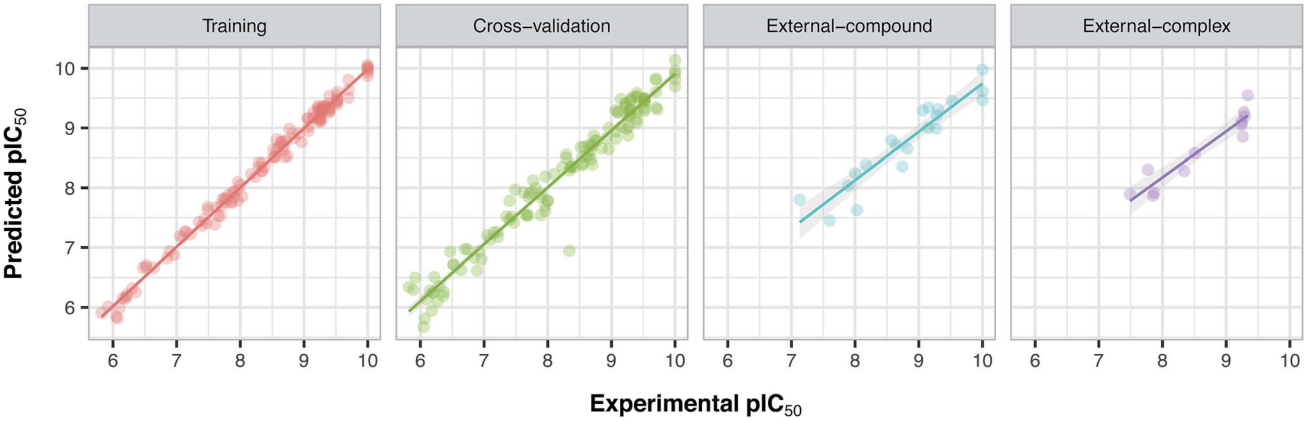


Fig. 3 Plots of experimental *versus* predicted pIC_{50} values for the training, cross-validation and external (external-compound and external-complex) sets.

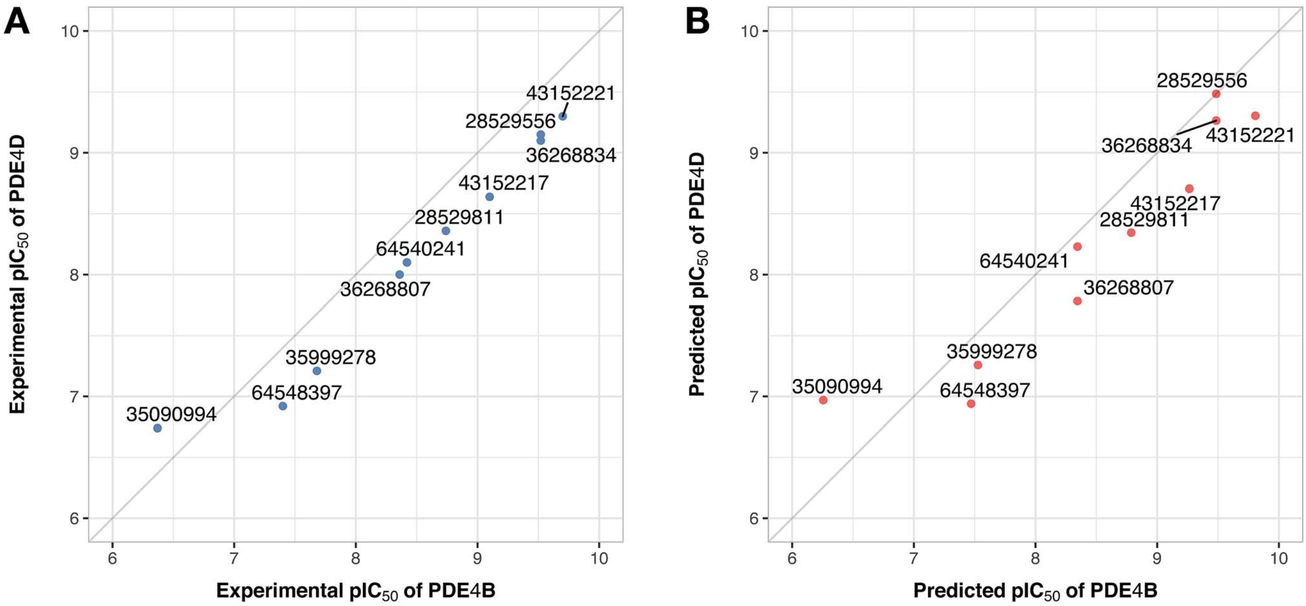


Fig. 4 Plots of the experimental (A) and predicted (B) pIC_{50} values of 10 compounds from the external-compound data set for PDE4B *versus* PDE4D. Compounds are denoted by filled circles and correspondingly labeled by their ZINC ID.

PCM modeling and assessment of the model validity

Since the positive impact of GRIND descriptors on modeling and the significance of using *z*-scale-GRIND combinatorial descriptors have been already confirmed,[41] we applied a combination of *z*-scale descriptors and GRIND descriptors in our modeling. Prior to modeling, feature selection was carried out with regard to GRIND descriptors using genetic algorithm in order to find the best fitted structural descriptors. Subsequently, 70 and 334 features were selected for compounds and proteins, respectively. In addition to GRIND descriptors, 3 *z*-scale descriptors were calculated in the case of proteins, resulting in a total number of 358 descriptors per protein isoform. *z*-Scale are the result of principal component analysis (PCA) performed on physicochemical properties of 87 natural/artificial amino acids. The first three PCs, called z_1, z_2 and z_3, are the representatives of the largest variations of physicochemical properties. Correlation between the descriptor matrix (consisting of compound descriptors, protein descriptors and their cross-terms) and biological activities (pIC_{50}s) were made using PLS.

Prior to modeling, a set of 12 randomly selected complexes were left out. We used this set in addition to 10 new compounds as an external set to validate the predictivity power of the model for activity and selectivity of new compounds. The resulting PCM model passed all the internal and external validation tests. Excellent values were observed for R^2 (0.99), Q^2 (0.92) and Q^2 pred (0.97), which are the indices for training, cross-validation and external cross-validation, respectively. Fig. 3 shows the experimental pIC_{50} values for the 32 complexes plotted *versus* their predicted values, revealing the highly effective predictivity power of the model. Moreover, 10 new compounds with activity against both isoforms were also randomly selected from the structure IDs deposited in the BindingDB (chemical structures along with their pIC_{50} values for PDE4B and PDE4D are provided as Data S4, Data S5 and Data S6, respectively). Prior to descriptor calculation, compounds geometries were optimized using SYBYL 7.3 (see the Methods for details). Comparing Fig. 4A with Fig. 4B reveals that not only is our model excellently able to predict the most active (43152221, 28529556 and 36268834) and the least active (35090994, 64548397 and 35999278) compounds, but is also capable of predicting the selectivity trend of new compounds towards PDE4B and PDE4D (with the exception of compound 28529556) most marvelously. Finally, the results of the *Y*-scrambling test (R^2 intercept of 0.05 and Q^2 intercept of −0.04) was able to confirm the robustness of the model (Fig. 5).

Applicability domain

Fig. 6 illustrates the applicability domain (AD) of the PCM model as defined by the Williams plot. The entire data set was split into two sets consisting of internal (80%) and external (20%) subsets, as described earlier in the Materials and methods section. As can be seen in the Williams plot, nearly all compounds are located within the boundaries of the applicability domain thereby suggesting a well-defined AD for the proposed PCM model.

Model interpretation and analyzing the contribution of compound properties for protein isoform selectivity

It has been shown by several X-ray crystallography investigations[5,14,62] that aromatic–aromatic interaction plays a major role

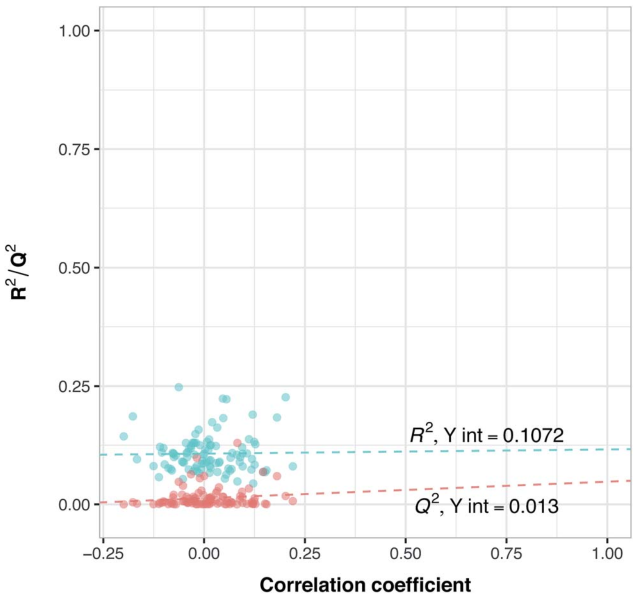


Fig. 5 *Y*-Scrambling plot of pIC_{50} for the PCM model. The *Y*-axis represents R^2 (blue circles) and Q^2 (red circles) coefficients for the original model and 100 models built based on randomly scrambled response data. The *X*-axis designates the correlation coefficient between the original and permuted response data.

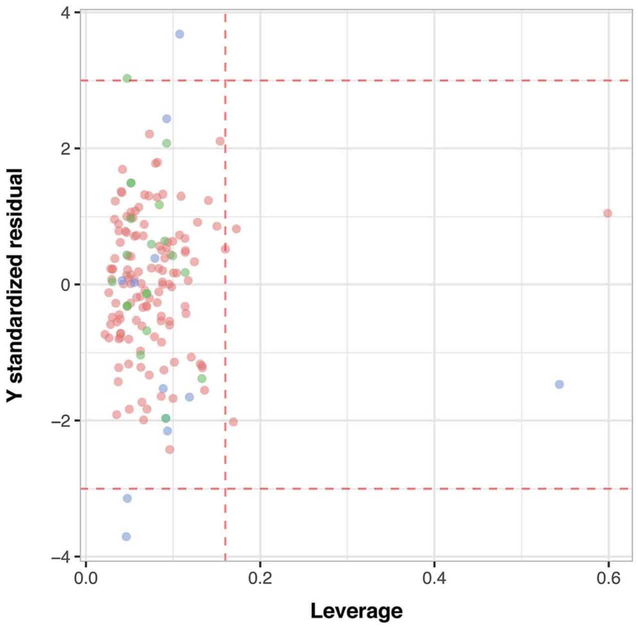


Fig. 6 Analysis of the applicability domain as evaluated *via* the Williams plot.

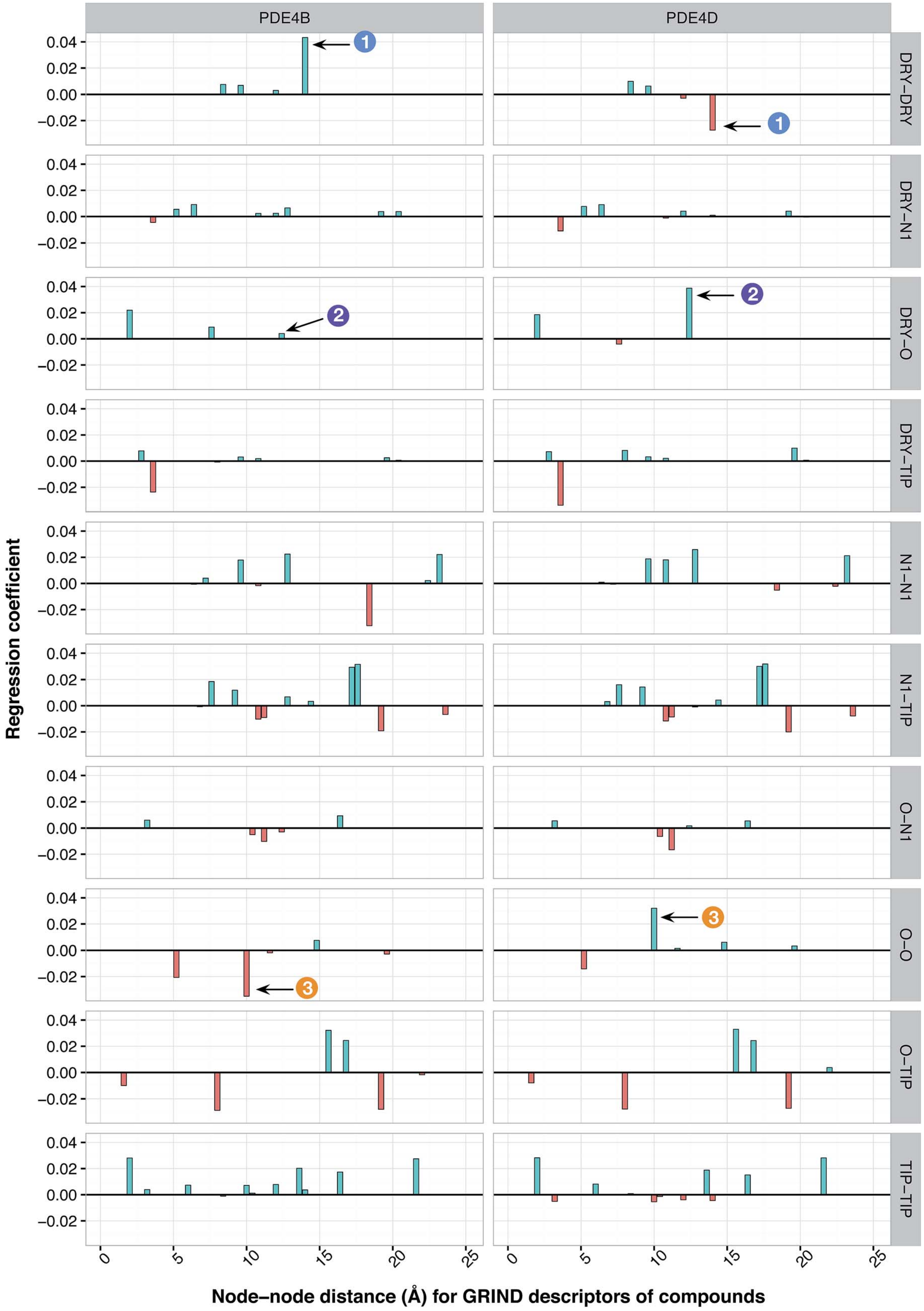


Fig. 7 Plot of important features showing the contribution of compound descriptors toward PDE inhibition. *Y*-Axes indicate the regression coefficients of descriptors. The interval within each sub-plot represents the node–node distance range of 0 and 24.8 Å for each respective GRIND descriptor. Descriptors that significantly discriminate between PDE4B and PDE4D are indicated by the numbered arrows.

in the average affinity of compounds toward the PDE4 isoforms. This is to be expected since nearly all available inhibitors are engaged in interacting with more than one aromatic ring and there is a considerable distribution of residues with aromatic side chains (*e.g.* Phe, Tyr and Trp) inside the protein cavity as well. Contribution of compound descriptors to the protein selectivity was assessed based on the measurements. Since Δy_{cp} can be considered as a regression coefficient of a GRIND descriptors for a specific protein isoform, comparing values for different protein isoforms can assess the impact of that specific compound property to their selectivity for a specific protein isoform. Fig. 7 illustrates the values for structural descriptors of compounds (selected by genetic algorithm) with regards to PDE4B and PDE4D. The *X*-axis tick labels represent a distance range from 0 Å to 24.8 Å. Some structural descriptors, showing highly discriminative behavior towards different isoforms of PDE4, are indicated in the figure and will be discussed in details.

DRY–DRY descriptor at distance of 14 Å

According to the unified PLS model, the DRY–DRY descriptor shows a significantly positive coefficient at the distance of 14 Å with regard to isoform PDE4B, in particular (Fig. 7, arrow 1). Comparing the chemical space of the enzymes cavities revealed the position where A197 in PDE4B is substituted by Ser in PDE4D. A closer look further revealed the presence of a hydrophobic residue (position 163) in the distance of almost 14 Å from position 197 of both cavities (Fig. 8A). Seeing that, Ser cannot be involved in hydrophobic interactions, this finding is highly compatible with the PLS coefficient of the DRY–DRY descriptor which is only positive for isoform PDE4B. Based on the PLS coefficient for the DRY–DRY descriptor and the structural evidence, our finding suggests that compounds having dual hydrophobic moieties with a distance of nearly 14 Å in their structures might show higher selectivity towards PDE4B. Fig. 8A clearly highlights the important role of hydrophobic–hydrophobic interactions toward selectivity.

O–O descriptor at distance of 10 Å

It seems that mutations in positions 197 and 201 are significantly critical for compound–protein selectivity, as they are captured twice by our PLS model. As clearly shown in Fig. 7 (arrow 3), a highly positive coefficient for the O–O descriptor exists at a distance of 10 Å with regard to PDE4D, while the correspondent descriptor is significantly negative for PDE4B. In addition, structural inspection of the enzymes cavities revealed that there is a spatial distance of 10 Å between the side chains of positions 197 and 201 (Fig. 8B). While these positions are occupied by Ser and Asn (both having side chains capable of accepting hydrogen bond) in PDE4D, the S197A substitution renders the O–O descriptor unfavorable in the case of PDE4B. This finding suggests that compounds with dual hydrogen donor moieties, located at a distance of 10 Å from each other, can possibly fit better in the cavity of PDE4D rather than in the cavity of PDE4B. Out of the plenty amino acids involved in the cavities of PDE4B and PDE4D, only 8 positions are not conserved. This definitely strengthens the necessity of applying these differences for the modeling process, as a few of them captured by our model has given a rise to valuable information.

DRY–O descriptor at distance of 12.4 Å

Inspecting the PLS coefficients revealed the presence of a significantly positive coefficient for DRY–O descriptor at a distance of 12.4 Å with regard to PDE4D, while the correspondent descriptor is hardly considerable for PDE4B (Fig. 7, arrow 2). The structural analysis showed that the spatial distance between positions 201 and 267 of the enzyme cavity (12.7 Å) is very close to that of the highly positive DRY–O descriptor (Fig. 8C). Based on the positive PLS coefficient and the similar distance between positions 201 and 267, it seems

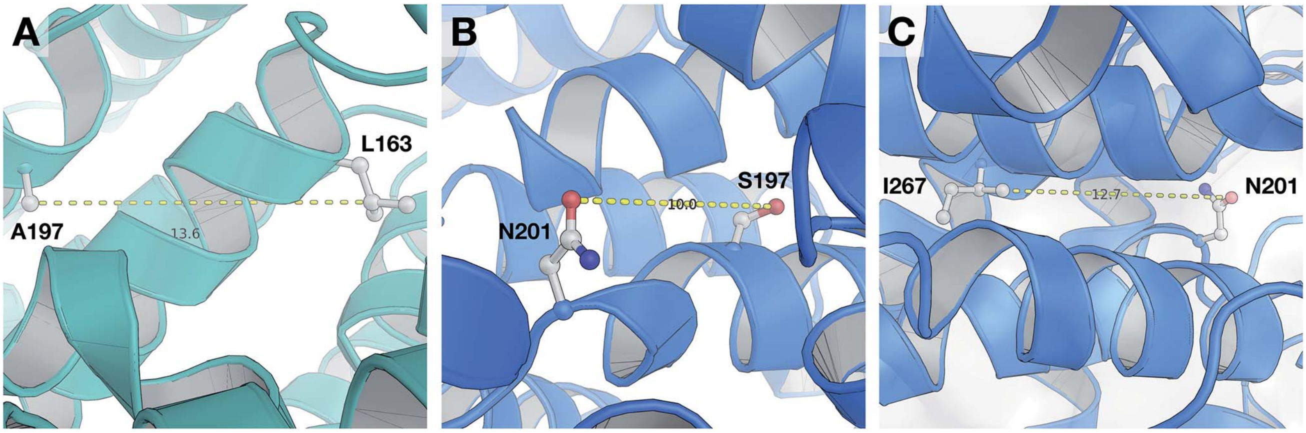


Fig. 8 Close-up structural representation of the cavity of PDE4B (A) and PDE4D (B and C). Panel (A) illustrates that two hydrophobic residues (L163 and A197) from PDE4B are located at a distance of 13.6 Å from each other, which corresponds to the DRY–DRY descriptor at a distance of 14 Å. Panel (B) shows that the two H-bond acceptor moieties (S197 and N201) from PDE4D are located at a distance of 10 Å from each other, which coincides with the O–O descriptor at a distance of 10 Å. Panel (C) reveals that an H-bond acceptor moiety (N201) and a hydrophobic residue (I267) are located at a distance of 12.7 Å from each other, which is in agreement with the DRY–O descriptor at distance of 12.4 Å.

that the I267M substitution (*i.e.* Ile is much more hydrophobic than Met) in PDE4B causes the inefficiency of the mentioned DRY–O descriptor in this isoform. Therefore, it is expected that inhibitors with dual hydrophobic/hydrogen donor moieties, placed in a distance close to that of the DRY–O descriptor, show better selectivity for isoform PDE4D.

Applicability of the model and the design of new inhibitors

To confirm the applicability of the PLS model and to show the impact of the differential descriptors on the selectivity, a few compounds were designed by modifying the studied inhibitors. The following criteria were considered for the selection of template compounds: (i) the selected compounds have exactly the same values of IC_{50} for PDE4B and PDE4D. This would make it easier to compare the changes in the predicted pIC_{50} values. (ii) The selected compounds lack the structural descriptor whose role is being studied. In this respect, the impact of a descriptor on the change of the pIC_{50} could be easily investigated by producing the desired descriptor values *via* modification of the compound structure. According to the mentioned criteria, inhibitors with ZINC IDs of 38431730, 26735077 and 36268795 were selected as template structures for applying structural modifications, which corresponded to DRY–DRY, DRY–O and O–O descriptors, respectively. As is illustrated in Fig. 9, the following structural modifications were applied on each compound in order to create the descriptors being investigated here: (i) in order to make a DRY–DRY descriptor at a distance of 14 Å for 38431730, we substituted the carboxyl group by two methyl moieties (38431730′), (ii) the DRY–O descriptor at a distance of 12.4 Å was created for 26735077 by converting the methyl group of the hydrocarbon chain to hydroxyl moiety (26735077′), (iii) the final modification was performed on 36268795 in order to provide this compound with the O–O descriptor at a distance of 10 Å. To do so, the methyl group linked to the five-membered ring was replaced by a hydroxyl group

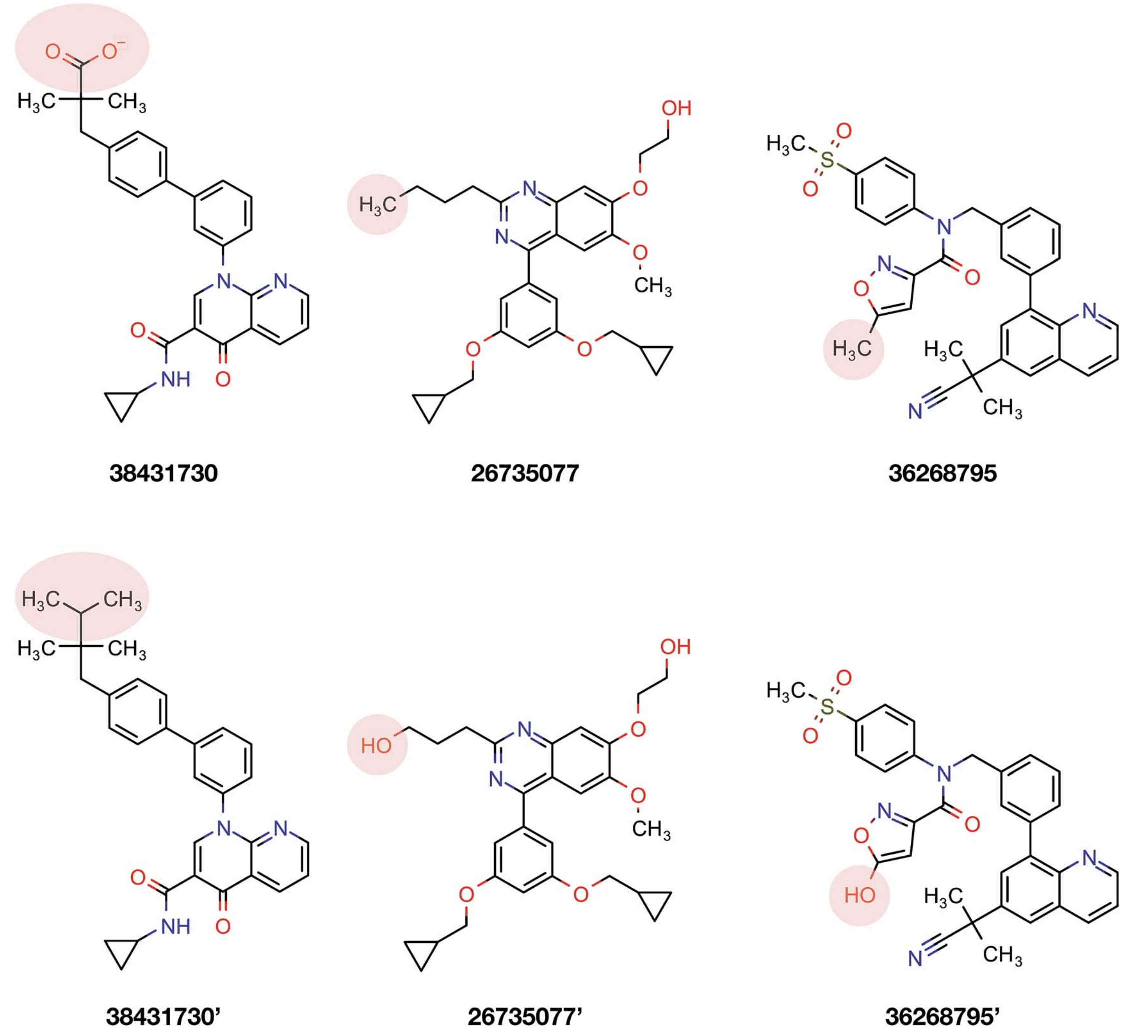


Fig. 9 Modification strategies utilized to investigate the effects of different functional moieties on the selectivity of compounds. Three template compounds used in the PCM modeling (first column consisting of ZINC ID 38431730, 26735077 and 36268795) as well as their derivatives (second column consisting of 38431730′, 26735077′ and 36268795′) that were used to generate the discussed DRY–DRY, O–O and DRY–O descriptors, respectively.

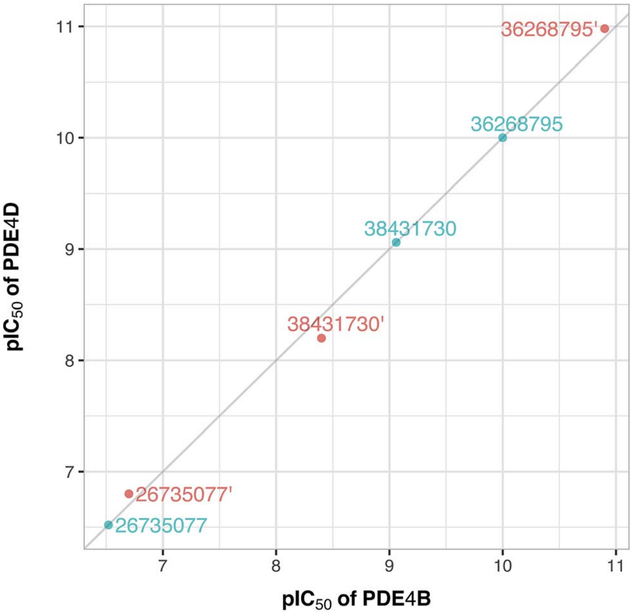


Fig. 10 Plot of the experimental/predicted pIC$_{50}$ values of template (cyan color) and designed (red color) compounds for PDE4B *versus* PDE4D. Experimental pIC$_{50}$ values are shown for template compounds while predicted pIC$_{50}$ values are shown for designed compounds. It should be noted that the former had identical pIC$_{50}$ values for both isoforms.

(36268795′). The designed compounds were subsequently put in the test set in order to predict their new biological activity.

Fig. 10 shows the predicted pIC$_{50}$ for each compound against PDE4B and PDE4D. Comparing the predicted pIC$_{50}$ values with the experimental ones reveals that the selectivity ratios of modified inhibitors have been altered according to our expectations. As shown, the selectivity ratio of 38431730, whose carboxylate moiety was substituted by two methyl groups as to generate the DRY–DRY descriptor of 38431730′, has been changed in favor of the PDE4B. In the case of 26735077 and 36268795 however, the derivatives 26735077′ and 36268795′ were in favor of increasing the selectivity ratios towards the PDE4D, as expected. Taken together, the results obtained by the compound design are indicative of the reliability of the proposed PCM model. Moreover, all three descriptors appears to be critical for the selective inhibition of PDE4B over PDE4D and *vice versa*.

Conformational difference of PDE4B and PDE4D

According to data collected from the crystal structure of PDE4B, it has been suggested that regions of the enzyme corresponding to the cavity's residues are highly flexible.[63,64] Particularly, results provided by crystallographic studies of PDE4 isoforms in complex with their inhibitors have revealed that despite the same observed pattern of compound–residue interactions, there exists a significant conformational difference. Interestingly, the average *B*-factor of the PDE4B structure was found to be higher than that of the PDE4D structure. Particularly, it can be seen that there exists a higher degree of conformational flexibility in the cavity of PDE4B when compared to that of PDE4D as shown in Fig. 11. Conformational variations reported for PDE4B and PDE4D can be attributed to available mutations in their sequences. For instance, it is known that T436 (PDE4B)/N362 (PDE4D) mutation causes variation in conformational properties of cavity residues followed by changes in the pattern of H-bonding interactions.[62] It has also been shown that the

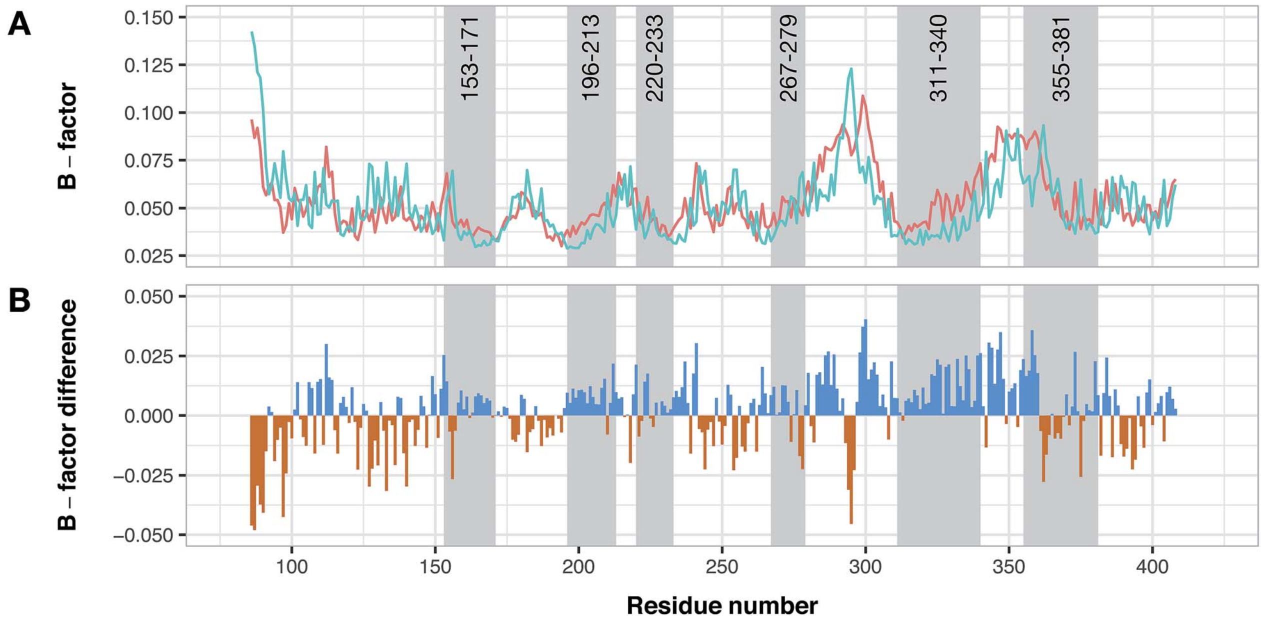


Fig. 11 Conformational analysis of PDE4B and PDE4D as assessed by *B*-factor values derived from their X-ray crystallographic structure. Panel (A) shows the normalized *B*-factor values of each residues of PDE4B and PDE4D while panel (B) shows the difference in the *B*-factor values in which blue and red bars correspond to flexible residues in PDE4B and PDE4D, respectively. Residue numbers are based on PDE4D (PDB ID: 2PW3) while gray shaded areas correspond to cavity residues.

active site of PDE4B is more hydrophobic in nature than that of PDE4D thereby suggesting that attention to the hydrophobic pocket can lead to effective subtype selective inhibitors.[62] This coincides with the results presented herein, which also suggests the crucial roles played by hydrophobic interactions along with some characteristic H-bond patterns in attaining the subtype selectivity. A close look at designed compounds proposed herein revealed that the substitution of a polar group with a hydrophobic one (*e.g.* hydroxyl/methyl) can alter the subtype selectivity in favor of PDE4B and *vice versa* when a hydrophobic moiety is replaced by a polar one (*e.g.* methyl/hydroxyl) (Fig. 9 and 10).

Conclusion

Due to the involvement of PDE4 subfamilies in major clinical disorders such as asthma, acrodysostosis, cognition disorder and schizophrenia, the necessity for the discovery of novel compounds with strong inhibitory properties towards these subfamilies is substantially growing. However, since the engagement of these subfamilies differs with regard to different disorders and subcellular compartments, developing new inhibitors without considering their selectivity properties can lead to undesirable side effects. Particularly, in cases similar to what is being presented here, when structural and sequence similarities of proteins are significant, the discovery of new selective inhibitors could be much more challenging. One way to overcome this issue is by applying the PCM approach in which the model takes in to account protein-based information, which in turn provides the opportunity of investigating the compound–protein interaction space in greater detail and thereby supporting the discovery of new selective inhibitors. Therefore, we investigated the interaction space and the selectivity properties that govern the inhibition of PDE4B and PDE4D by applying the PCM approach. Furthermore, utilizing the combination of *z*-scales and MIF-based GRIND descriptors, we found that specific structural features such as the presence of dual hydrophobic moieties at certain distances are crucial for selectivity properties. Aside from the critical role of hydrophobic–hydrophobic interactions in selectivity, other types of interactions such as hydrogen bonds with characteristic patterns were also found to be play an important role in governing the selectivity of compounds towards a specific receptor. The ability of the presented PCM model for capturing all of these discriminative structural details can be attributed to the consideration of the protein's structural/sequence differences in the model. In this manner, we were able to discover that substitutions such as A197S and I267M can influence the interaction space of investigated PDE4 and compounds. Finally, we believe that our findings can be taken into consideration for designing compounds with better selectivity towards either PDE4B or PDE4D.

Acknowledgements

This research is supported by a Research Career Development Grant (No. RSA5780031) and the Swedish Research Links program (No. C0610701) from the Swedish Research Council to CN.

References

1 V. C. Manganiello, F. Ahmad, Y. H. Choi, Y. Tang, R. Lindh, E. Zmuda-Trezbiatowska, H. Walz, H. Liu, H. L. Stenson and E. Degerman, in *Phosphodiesterase and Intracellular Signaling*, Mie University Press, Mie, Japan, 2007, pp. 101–125.

2 E. J. Tsai and D. A. Kass, *Pharmacol. Ther.*, 2009, **122**, 216–238.

3 S. H. Francis, M. A. Blount and J. D. Corbin, *Physiol. Rev.*, 2011, **91**, 651–690.

4 A. T. Bender and J. A. Beavo, *Pharmacol. Rev.*, 2006, **58**, 488–520.

5 D. H. Maurice, H. Ke, F. Ahmad, Y. Wang, J. Chung and V. C. Manganiello, *Nat. Rev. Drug Discovery*, 2014, **13**, 290–314.

6 F. Ahmad, T. Murata, K. Shimizu, E. Degerman, D. Maurice and V. Manganiello, *Oral Dis.*, 2015, **21**, 25–50.

7 Y. H. Jeon, Y. S. Heo, C. M. Kim, Y. L. Hyun, T. G. Lee, S. Ro and J. M. Cho, *Cell. Mol. Life Sci.*, 2005, **62**, 1198–1220.

8 K. Omori and J. Kotera, *Circ. Res.*, 2007, **100**, 309–327.

9 M. Zaccolo and M. A. Movsesian, *Circ. Res.*, 2007, **100**, 1569–1578.

10 H. Ke and H. Wang, *Curr. Top. Med. Chem.*, 2007, **7**, 391–403.

11 J. Hou, J. Xu, M. Liu, R. Zhao, H. B. Luo and H. Ke, *PLoS One*, 2011, **6**, e18092.

12 F. Meng, J. Hou, Y. X. Shao, P. Y. Wu, M. Huang, X. Zhu, Y. Cai, Z. Li, J. Xu, P. Liu, H. B. Luo, Y. Wan and H. Ke, *J. Med. Chem.*, 2012, **55**, 8549–8558.

13 S. K. Chen, P. Zhao, Y. X. Shao, Z. Li, C. Zhang, P. Liu, X. He, H. B. Luo and X. Hu, *Bioorg. Med. Chem. Lett.*, 2012, **22**, 3261–3264.

14 Z. Li, Y. H. Cai, Y. K. Cheng, X. Lu, Y. X. Shao, X. Li, M. Liu, P. Liu and H. B. Luo, *J. Chem. Inf. Model.*, 2013, **53**, 972–981.

15 T. J. Torphy, *Am. J. Respir. Crit. Care Med.*, 1998, **157**, 351–370.

16 K. F. Rabe, *Br. J. Pharmacol.*, 2011, **163**, 53–67.

17 H. Tenor, A. Hatzelmann, R. Beume, G. Lahu, K. Zech and T. D. Bethke, *Handb. Exp. Pharmacol.*, 2011, 85–119.

18 L. M. Fabbri, P. M. Calverley, J. L. Izquierdo-Alonso, D. S. Bundschuh, M. Brose, F. J. Martinez, K. F. Rabe and M2-127 and M2-128 study groups, *Lancet*, 2009, **374**, 695–703.

19 H. Wachtel, *Neuropharmacology*, 1983, **22**, 267–272.

20 M. D. Houslay, P. Schafer and K. Y. Zhang, *Drug Discovery Today*, 2005, **10**, 1503–1519.

21 M. P. Kelly and N. J. Brandon, *Prog. Brain Res.*, 2009, **179**, 67–73.

22 S. J. Clapcote, T. V. Lipina, J. K. Millar, S. Mackie, S. Christie, F. Ogawa, J. P. Lerch, K. Trimble, M. Uchiyama, Y. Sakuraba, H. Kaneda, T. Shiroishi, M. D. Houslay, R. M. Henkelman, J. G. Sled, Y. Gondo, D. J. Porteous and J. C. Roder, *Neuron*, 2007, **54**, 387–402.

23 J. K. Millar, S. Mackie, S. J. Clapcote, H. Murdoch, B. S. Pickard, S. Christie, W. J. Muir, D. H. Blackwood,

J. C. Roder, M. D. Houslay and D. J. Porteous, *J. Physiol.*, 2007, **584**, 401–405.

24 Z. DeMarch, C. Giampa, S. Patassini, G. Bernardi and F. R. Fusco, *Neurobiol. Dis.*, 2008, **30**, 375–387.

25 H. Lee, J. M. Graham, D. L. Rimoin, R. S. Lachman, P. Krejci, S. W. Tompson, S. F. Nelson, D. Krakow and D. H. Cohn, *Am. J. Hum. Genet.*, 2012, **90**, 746–751.

26 S. Gretarsdottir, G. Thorleifsson, S. T. Reynisdottir, A. Manolescu, S. Jonsdottir, T. Jonsdottir, T. Gudmundsdottir, S. M. Bjarnadottir, O. B. Einarsson, H. M. Gudjonsdottir, M. Hawkins, G. Gudmundsson, H. Gudmundsdottir, H. Andrason, A. S. Gudmundsdottir, M. Sigurdardottir, T. T. Chou, J. Nahmias, S. Goss, S. Sveinbjornsdottir, E. M. Valdimarsson, F. Jakobsson, U. Agnarsson, V. Gudnason, G. Thorgeirsson, J. Fingerle, M. Gurney, D. Gudbjartsson, M. L. Frigge, A. Kong, K. Stefansson and J. R. Gulcher, *Nat. Genet.*, 2003, **35**, 131–138.

27 S. H. Fatemi, D. P. King, T. J. Reutiman, T. D. Folsom, J. A. Laurence, S. Lee, Y. T. Fan, S. A. Paciga, M. Conti and F. S. Menniti, *Schizophr. Res.*, 2008, **101**, 36–49.

28 A. Robichaud, P. B. Stamatiou, S. L. Jin, N. Lachance, D. MacDonald, F. Laliberté, S. Liu, Z. Huang, M. Conti and C. C. Chan, *J. Clin. Invest.*, 2002, **110**, 1045–1052.

29 P. Prusis, R. Muceniece, P. Andersson, C. Post, T. Lundstedt and J. E. S. Wikberg, *Biochim. Biophys. Acta*, 2001, **1544**, 350–357.

30 M. Lapinsh, P. Prusis, T. Lundstedt and J. E. S. Wikberg, *Mol. Pharmacol.*, 2002, **61**, 1465–1475.

31 M. Lapinsh, P. Prusis, S. Uhlen and J. E. S. Wikberg, *Bioinformatics*, 2005, **21**, 4289–4296.

32 P. Prusis, M. Lapins, S. Yahorava, R. Petrovska, P. Niyomrattanakit, G. Katzenmeier and J. E. S. Wikberg, *Bioorg. Med. Chem.*, 2008, **16**, 9369–9377.

33 M. Lapins, M. Eklund, O. Spjuth, P. Prusis and J. E. Wikberg, *BMC Bioinf.*, 2008, **9**, 181.

34 M. Lapins and J. E. Wikberg, *BMC Bioinf.*, 2010, **11**, 339.

35 M. Fernandez, S. Ahmad and A. Sarai, *J. Chem. Inf. Model.*, 2010, **50**, 1179–1188.

36 V. Subramanian, P. Prusis, L. O. Pietilä, H. Xhaard and G. Wohlfahrt, *J. Chem. Inf. Model.*, 2013, **53**, 3021–3030.

37 I. Mandrika, P. Prusis, S. Yahorava, M. Shikhagaie and J. E. Wikberg, *Protein Eng., Des. Sel.*, 2007, **20**, 301–307.

38 A. Kontijevskis, J. Komorowski and J. E. Wikberg, *J. Chem. Inf. Model.*, 2008, **48**, 1840–1850.

39 S. Simeon, O. Spjuth, M. Lapins, S. Nabu, N. Anuwongcharoen, V. Prachayasittikul, J. E. S. Wikberg and C. Nantasenamat, *PeerJ*, 2016, **4**, e1979.

40 B. Rasti, M. H. Karimi-Jafari and J. B. Ghasemi, *Chem. Biol. Drug Des.*, 2016, **88**, 341–353.

41 B. Rasti, M. Namazi, M. H. Karimi-Jafari and J. B. Ghasemi, *Mol. Inf.*, 2017, **36**, 1600102.

42 X. Chen, M. Liu and M. K. Gilson, *Comb. Chem. High Throughput Screening*, 2001, **4**, 719–725.

43 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, 198–201.

44 M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins, *Bioinformatics*, 2007, **23**, 2947–2948.

45 Molecular Discovery Ltd., GRID, 1999, Oxford, UK.

46 D. Beasley, D. R. Bull and R. R. Martin, *Univ. Comput.*, 1993, **15**, 58–69.

47 D. Rogers and A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 854–866.

48 M. Sandberg, L. Eriksson, J. Jonsson, M. Sjostrom and S. Wold, *J. Med. Chem.*, 1998, **41**, 2481–2491.

49 J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.

50 J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J. Chem. Inf. Model.*, 2012, **52**, 1757–1768.

51 Tripos Associates, *SYBYL Molecular Modeling Software, version 7.3*, St. Louis, MO, 2006.

52 T. J. Hou, J. M. Wang, N. Liao and X. J. Xu, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 775–781.

53 *PLS Toolbox, version 3.5*, Eigenvector Research, Inc., Manson, WA, 2005.

54 D. M. Haaland and E. V. Thomas, *Anal. Chem.*, 1988, **60**, 1193–1202.

55 L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell and P. Gramatica, *Environ. Health Perspect.*, 2003, **111**, 1361–1375.

56 P. Gramatica, *QSAR Comb. Sci.*, 2007, **26**, 694–701.

57 A. Tropsha, P. Gramatica and V. K. Gombar, *QSAR Comb. Sci.*, 2003, **22**, 69–77.

58 S. L. Jin and M. Conti, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 7628–7633.

59 M. Ariga, B. Neitzert, S. Nakae, G. Mottin, C. Bertrand, M. P. Pruniaux, S. L. Jin and M. Conti, *J. Immunol.*, 2004, **173**, 7531–7538.

60 G. Hansen, S. Jin, D. T. Umetsu and M. Conti, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 6751–6756.

61 A. B. Burgin, O. T. Magnusson, J. Singh, P. Witte, B. L. Staker, J. M. Bjornsson, M. Thorsteinsdottir, S. Hrafnsdottir, T. Hagen, A. S. Kiselyov, L. J. Stewart and M. E. Gurney, *Nat. Biotechnol.*, 2010, **28**, 63–70.

62 P. Srivani, D. Usharani, E. D. Jemmis and G. N. Sastry, *Curr. Pharm. Des.*, 2008, **14**, 3854–3872.

63 H. Wang, M. S. Peng, Y. Chen, J. Geng, H. Robinson, M. D. Houslay, J. Cai and H. Ke, *Biochem. J.*, 2007, **408**, 193–201.

64 P. Cedervall, A. Aulabaugh, K. F. Geoghegan, T. J. McLellan and J. Pandit, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E1414–E1422.

EXPERT OPINION

1. Introduction
2. Expanding the chemical space
3. Big data in drug discovery
4. Virtual screening
5. Systems-based drug discovery
6. Interoperability and future of computational tools
7. Conclusion
8. Expert opinion

informa
healthcare

Maximizing computational tools for successful drug discovery

Chanin Nantasenamat* & Virapong Prachayasittikul[†]

*[†]Mahidol University, Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Bangkok, Thailand and *Mahidol University, Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Bangkok, Thailand*

Drug discovery is an iterative cycle of identifying promising hits followed by lead optimization via bioisosteric replacements. In the search for compounds affording good bioactivity, equal importance should also be placed on achieving those with favorable pharmacokinetic properties. Thus, the balance and realization of both key properties is an intricate problem that requires great caution. In this editorial, the authors explore the available computational tools in the context of the extant of big data that has borne out via advents of the Omics revolution. As such, the selection of appropriate computational tools for analyzing the vast number of chemical libraries, target proteins and interactomes is the first step toward maximizing the chance for success. However, in order to realize this, it is also necessary to have a solid foundation on the big concepts of drug discovery as well as knowing which tools are available in order to give drug discovery scientists the best opportunity.

Keywords: bioinformatics, cheminformatics, chemogenomics, computational tools, data mining, drug discovery

Expert Opin. Drug Discov. (2015) ***10****(4):321-329*

1. Introduction

Drug discovery entails the screening of large chemical libraries for promising hits, translating such hits to leads, enhancing them via lead optimization until a drug candidate is obtained for further validation by clinical trials [1]. It is an enduring process that can take 15 years and cost more than $1 billion to develop a new drug starting from its conception up until its release in the market. Computational tools have been instrumental at various stages of drug discovery and continue to be indispensible in the never-ending quests for life-saving drugs.

Doman *et al.* [2] employed computational tools in identifying 365 promising compounds from their virtual screen where 127 of which displayed good inhibitory effect against their target, tyrosine phosphatase-1B, while a high-throughput screening of 400,000 compounds yielded 81 compounds with good inhibitory activity. In this comparative analysis, it was shown that computational tools were able to afford a significantly higher hit rate of 35% over that of 0.021% from high-throughput screening. Lists of approved drug that are a result of computational drug design effort include angiotensin-converting enzyme inhibitor captopril, carbonic anhydrase inhibitor dorzolamide, fibrinogen antagonist tirofiban as well as three HIV drugs, including saquinavir, ritonavir and indinavir [3].

In spite of the great abundance of tools available for drug discovery [4,5], we are left in a situation of deciding what tools are potentially available for performing a certain task and which one to use. Thus, it is the ambition of this editorial to take a glimpse at what tools are available and how we can maximize our productivity by presenting the available toolbox for drug discovery. Selecting the appropriate computational tools is dependent on the goal of the drug discovery project that

10.1517/17460441.2015.1016497 © 2015 Informa UK, Ltd. ISSN 1746-0441, e-ISSN 1746-045X

may, for example, be any of the following: understanding the SAR of a compound series, synthesizing/optimizing novel leads, understanding the binding modality for compounds of interest, screening for novel compounds with desired bioactivity, identifying potential toxicity in compounds, designing multi-target inhibitors, predicting the side effects of drugs, etc. Thus, this editorial will briefly touch upon several key steps in the drug discovery process and their associated tools in the context of trying to cover most (if not all) of these essential topics in drug discovery. Particularly, available tools at various levels of drug discovery (i.e., fragments, ligands, protein and systems level) will be briefly discussed and a schematic of the connectivity of these concepts is provided in (**Figure 1**). Here, we also confer known or potential opportunities and problems that researchers may need to be aware of in their drug discovery endeavors.

2. Expanding the chemical space

As drug discovery is primarily a data-driven process that is reliant on producing medicinal chemistry data on the potential bioactivity of compounds and their target proteins of interests, it is therefore customary to first start with the discussion of data in drug discovery. One of the initial phases would be to acquire a large chemical library either by using existing public or private chemical databases or by chemical synthesis. The former may be a cost-effective path as large chemical libraries encompassing a wide range of molecules are immediately available for performing queries or available for carrying out high-throughput virtual screening. In spite of the convenience of using pre-existing databases, the ambitious task of expanding our known chemical space can be attributed to a large extent to high-throughput chemical synthesis (i.e., combinatorial chemistry, dynamic combinatorial chemistry, target-oriented synthesis and diversity-oriented synthesis) [6]. Despite the seemingly large chemical library produced by these efforts, we are still far from reaching the estimated chemical space of > 10^{60} molecules. The first step to scratching the surface for expanding our coverage of the chemical space was carried out by the research group of Reymond in which 166 billion molecules containing up to 17 atoms (i.e., C, N, O, S and halogens) were computationally enumerated [7]. Another approach for increasing the sampled chemical space is to employ molecular fragments where some of which are privileged structures that can bind a wide range of receptor sub-pockets and act as good chemical starting points [8]. Once promising fragments are identified from experimental or virtual screen, they can be computationally grown such that a central scaffold and linker will subsequently join the fragments. In situations where an active compound afford poor pharmacokinetic properties, medicinal chemists may perform scaffold hopping as to enhance the polarity, toxicity as well as flexibility profiles of the central scaffold while keeping the functional moieties at the periphery intact [9].

3. Big data in drug discovery

The advent of the omics era and advances in the synthesis of large chemical library have inevitably render the task of analyzing the generated big data a great challenge. Traditionally, data from bioactivity assays had been hidden inside the primary literature within the confinement of hard or soft copies of the research article. Curation of these bioactivity data is available in public databases such as PubChem [10], ChEMBL [11] and BindingDB [12] that correspondingly contain binding data/protein targets/small molecules of 200,000,000/8,000/1,900,000, 12,843,338/10,579/1,411,7-86 and 1,058,945/6,997/453,657, respectively. It is anticipated that further progress in biocuration will lead to the availability of ever more bioactivity data. And it is expected that text mining would be instrumental in automating or semi-automating the extraction of pertinent data from the overwhelming primary literature. In spite of the available big data in drug discovery, a great challenge that researchers are facing is the inherent error that may be present in the underlying data repositories that they depend on [13]. Such errors may arise from several factors (i.e., human error in data reporting, software bugs producing erroneous data, stereochemistry errors in chemical structures, errors in macromolecular sequences, etc.) and have far-reaching impact that goes beyond just being simple data errors but may skew the reliability of several tools, repositories and studies that rely on such erroneous data sources [14]. Several key players (whether it be the authors, publishers, readers and users of such data sources) can together help to resolve such issues by not taking the data for granted but should be cautious as well as scrutinizing the validity and quality of data resources.

In spite of growing progress in finding cures for several diseases, the lack of treatment and bioactivity data still persists for rare and neglected diseases [15]. Open innovation addresses the issue of finding cure for neglected diseases as well as tackling the issue of declining productivity and attrition rate of the pharmaceutical industry [16]. Particularly, such initiative is being carried out by pharmaceutical company whereby high-throughput screening data is made publicly available thereby allowing the pooling of resources in an open environment [17]. For example, several leading institutes are sharing their screening data for neglected tropical diseases on ChEMBL (https://www.ebi.ac.uk/chemblntd). Along a similar route, Bayer HealthCare initiated a crowdsourcing project called Grants4Targets that aims to foster collaboration between researchers in academia and pharmaceutical industry [18].

4. Virtual screening

In addition to experimental screening in drug discovery, virtual screening had firmly established itself as a crucial component in the drug discovery pipeline. Methodologically,

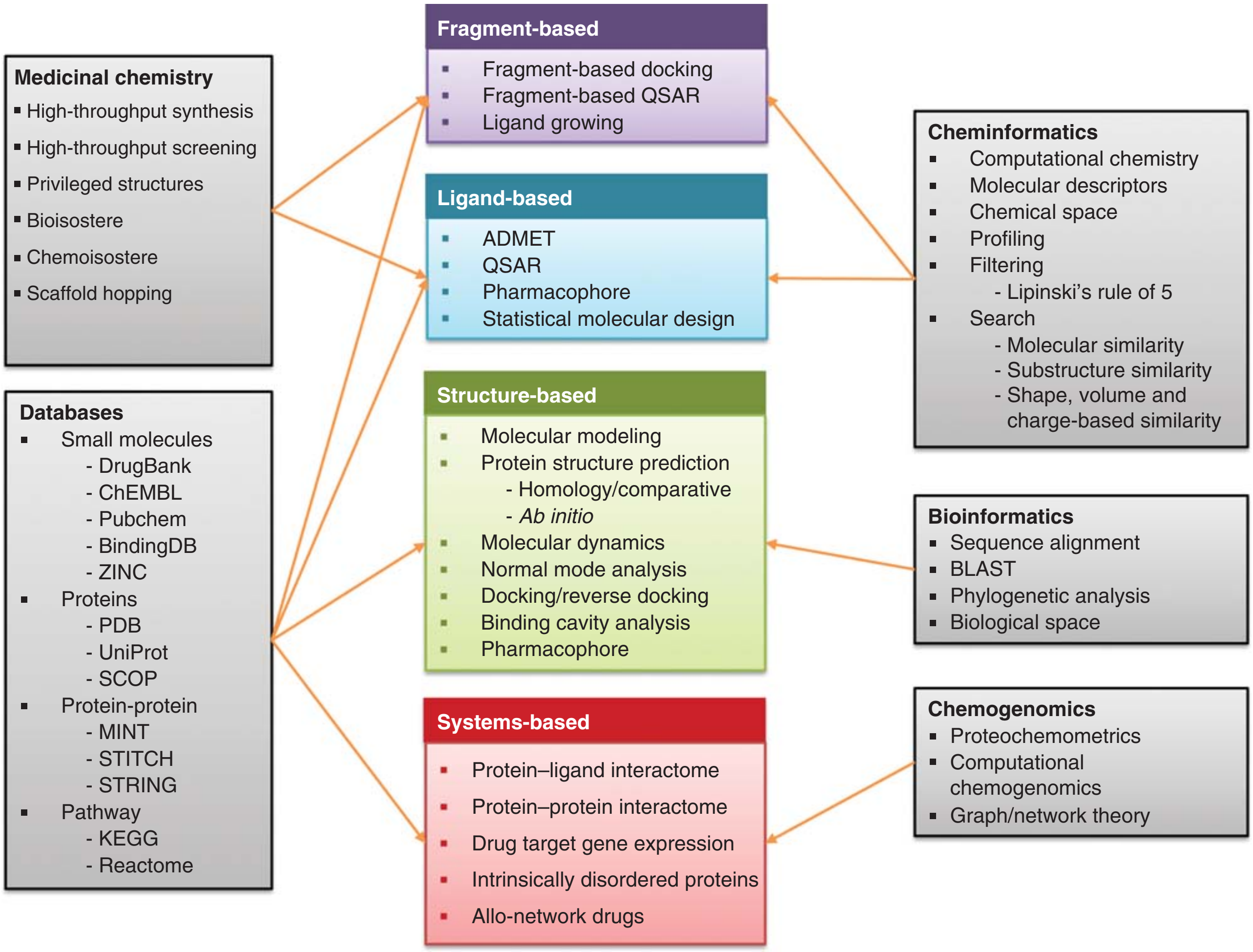


Figure 1. Schematic overview of computational drug discovery.

virtual screening can be classified into two major types: i) ligand-based virtual screening; and ii) structure-based (or target-based) virtual screening.

Ligand-based virtual screening, as the name implies, entails the use of ligands for virtual screening and is essentially based on the assumption that compounds with similar chemical structures have the tendency to share similar biology activities. Fundamentally, ligands can be described by their molecular fingerprints and, for example, the fingerprints of known actives can be used to extrapolate and seek for other molecules sharing similar properties (i.e., by means of Tanimoto coefficient or other distance-based approaches). Another popular class of ligand-based virtual screening approach is quantitative structure–activity relationship (QSAR) that requires the construction of computational models to learn from bioactivity data afforded by each ligand [19-21]. This approach involves the computation of molecular descriptors (i.e., constitutional, topological, charge, autocorrelations, etc.) for quantitatively describing the ligands and the subsequent use of machine learning methods to correlate such descriptors with their associated bioactivity. Notable examples on the utilization of QSAR in drug discovery generally fall into the following categories: prediction of ADMET properties [22,23], drug-likeness profiling [24-26] as well as predicting inhibition of target proteins [27], cancer cells [28] and pathogens [29].

Analogously, structure-based virtual screening makes use of the protein structure for investigating characteristics of the binding pocket, *de novo* design of ligands inside the binding pocket and performing protein-ligand docking. For the latter point, molecular dynamics has been shown to improve the performance of virtual screening against known crystal structures as it can generate snapshots for subsequent sampling of relevant protein conformation [30]. Along this line, researchers can also dock an ensemble of structures from the same protein against a library of ligands, dock several proteins to the same ligand (an approach called reverse or inverse docking) as well as cross-docking several ligands against an ensemble of proteins [31,32]. Novel ligands can also be tailor-made to the protein target of interest by docking fragments to the binding pocket followed by linking together the fragments to form a ligand. A notable example of such tool is AutoGrow developed by Durrant *et al.* [33] and a recent update [34] employs the rule of click chemistry in order to improve the synthesizability of produced ligands. A similar but ligand-based tool called LigMerge [35] can generate novel ligands from known actives by superimposing the largest substructures and systematically mixes and matches substituents at each atom. The aforementioned approaches are only applicable if the protein structure of interest exists. Therefore, in situations where the protein structure is not experimentally available from databases such as the PDB, researchers could also predict the structures either through *ab initio* or template-based homology or comparative modeling approaches [36].

It is noteworthy that pharmacophore modeling can be applied for both ligand and structure-based approaches by either aligning several ligand structures or by docking ligands to target proteins followed by analyzing the resulting docked structures. Pharmacophore model produces a spatial map of key chemical functionality that is required for ligands to potently bind the target protein of interest. Analogously, shape and charge similarity search can also be applied by both ligand and structure-based approaches whereby known active ligands and binding pockets are used as queries for potential binders from large chemical libraries. Notable tools for performing shape and charge similarity search are ROCS and EON, respectively, both available from OpenEye (http://www.eyesopen.com/). Although both are commercial software but an academic license is also available for researchers from academia. Recently, a new tool called Screen3D has recently been reported by ChemAxon for performing shape similarity search [37].

Central to these investigations is acquiring the actual compounds once they have been virtually identified to afford promising inhibitory effect against the target protein of interest. ZINC [38] addresses this issue as it provides access to 35 million compounds that are also commercially available. ChemSpider [39] is another equally important repository that aside from providing access to information on 32 million compounds not only links to chemical vendors but also links to patents, physicochemical information and other relevant repositories.

Tanrikulu *et al.* [40] suggests that virtual screening is not a standalone approach but rather is evenly matched and complementary to high-throughput screening. It was suggested that no single approach is superior to one another and that the complementary nature of ligand and structure-based approaches can be fused to increase enrichment rates. Furthermore, virtual screening was classified into four major classes (i.e., classical, parallel, iterative and integrated) as functions of how wet and virtual screening are connected. An excellent example on the merger of ligand and structure-based approach is the SABRE [41] program that combines shape similarity search of known active ligands followed by analyzing the generated consensus molecular shape-pattern within the confinement of the binding pocket.

Aside from the promises of virtual screening, there are also inherent flaws and potential pitfalls that practitioners should be aware of. Scior *et al.* [42] summarize this issue into four major categories: erroneous assumptions and expectations, data design and content, conformational sampling and ligand/target flexibility and choice of software. An additional caution for applying models derived from virtual screening is to consider its applicability domain as such model will only be valid for compounds sharing similar chemical space as those used to build the model [43]. Of particular note is the importance of being well versed with the strengths and limitations of selected computational tools and this could be achieved by first understanding the algorithmic details as well as its technicalities (i.e., file format, default *versus* custom parameters, interoperability, etc.).

5. Systems-based drug discovery

The aforementioned sections covered approaches for designing inhibitors against a single target. In recent years, there is a paradigm shift from the reductionist view of 'one drug, one target' to the holistic and systems approach of 'multi-drug, multi-target' [44]. In systems-based drug discovery, healthy and disease states are viewed as being modulated by biological networks of proteins and its regulators. It is worthy to note that the redundancy and resiliency in biological networks that are commonly found in complex diseases (i.e., cancer, diabetes and cardiovascular diseases) is the reason behind the observation that modulating a single target in the biological pathway(s) may still give rise to the disease. This is corroborated by a computational study on drug-target network performed by Yildirim *et al.* [45] suggested that drugs can act on multiple targets and that these drug targets are in turn linked to several diseases.

Once an undesirable property of the single-target paradigm (sometimes referred to as off-target binding, binding promiscuity or polypharmacology), multi-target drugs have found a new therapeutic route in its rebranding as drug repositioning (also called drug repurposing) in which existing FDA-approved drugs are reused for treating another disease [46]. The advantage of this approach is that the pharmacokinetic profiles of these drugs are implied to be safe as it had already gone through clinical trials. Chemogenomics is a discipline that employs chemical probes to characterize proteome functions in either forward or reverse manner depending on whether the process proceeds from phenotype to target or from target to phenotype [47]. The rather sparse and limited availability of bioactivity data for compounds against multi-targets as compared to those available for single targets in concomitant with the time-consuming and costly nature of experimental assays calls for computational methods (i.e., proteochemometrics, graph/network theory and reverse docking) that are capable of mapping such large-scale compound–protein interaction spectrum by discerning linkages between molecular and target similarity. Proteochemometrics is instrumental in discerning the structure–activity relationship for a series of compound against a series of proteins and it has been successfully applied on a wide range of protein families such as G-protein-coupled receptors, proteases, kinases and cytochrome P450s among others [48]. Graph/network theory sees the biological network of proteins and their regulators as graphs in which nodes are interconnected via vertices [49]. Reverse or inverse docking modifies the original concept of docking a series of ligands against a single-target protein to the docking of a series of proteins to a target compound [50]. Owing to the availability of huge volumes of data in medicinal chemistry, there is growing interest to harness the knowledge from these data via large-scale structure–activity relationships. Aiming to tackle this gap is the CARLSBAD database [51] that houses nearly 1,500,000 unique bioactivities against > 400,000 compounds. Similarly, AstraZeneca developed the SAR Connect [52] for mining the big data in drug discovery compiled from ChEMBL, GOSTAR and AstraZeneca's IBIS. Along this line is a semantic web explorer PharmaTrek [53] that allows users to perform complex query on multi-target polypharmacology in an intuitive manner. Thus, proteochemometrics as well as reverse docking could readily be used in drug repositioning FDA-approved drugs against a multitude of new diseases [50,54].

Conventional approaches in drug discovery had primarily focused on targeting orthosteric sites of target proteins in attempts to inhibit their function. However, there are other equally important therapeutic approaches that should be kept in mind such as targeting allosteric sites, intrinsically disordered proteins and protein–protein interactions as well as modulating gene expression [55]. The benefits of allosteric drugs is its greater specificity, reduced side effects and lower toxicity as well as its usage flexibility in which it can be used to target large protein–protein interaction for which small molecules may not be applicable as well as used to increase activity of its target and tackle drug-resistant targets [56]. AlloSteric Database [57] had recently updated its repository to encompass 1286 allosteric proteins and 22,008 allosteric modulators. Intrinsically disordered proteins represent lucrative drug targets owing to its importance and overrepresentation in signaling and major disease pathway [58]. Protein interaction network are known to moderate cellular signaling via interaction between proteins and pathologically significant protein–protein interaction therefore serve as promising drug targets [59]. Modulation of gene expression may be an alternative therapeutic route by indirectly tackling a target protein by controlling its amount rather than influencing its activity, which may be attractive in cases when modulating the activity is a formidable challenge (i.e., highly drug-resistant target or targets that lack a suitable binding site) [60,61].

6. Interoperability and future of computational tools

The lack of consistency and interoperability of computational tools presents a major bottleneck in drug discovery. As such, the input and output data for various tools are often of heterogeneous formats and are thus not amenable for easy and widespread adoption and implementation. Efforts to tackle this issue is a pressing need that had started to gain momentum in recent years. For example, Galaxy [62] is a web-based platform implemented in Python that supports interactive genomic analysis. Moreover, Cinfony [63] establishes itself as a central platform of cheminformatics toolkits by housing them all in one common interface as a Python module. Similarly, Biskit [64] and CSB [65] are Python modules that integrate several structural bioinformatics and computational structural biology tools under a unified platform. Furthermore, data from biological databases (i.e., BioModels, ChEMBL, KEGG and UniProt) can be programmatically accessed via a common Python framework called BioServices [66]. In the context of ligand-based drug discovery, the development of QSAR-ML [67] makes QSAR data sets open and interoperable as it embed the information of chemical structures, descriptors, software implementations and response values inside an XML-based file. The need to integrate data available from various sources and public databases can be performed by semantic web technology that would thereby allow interoperability to be achieved as implemented in a pioneering framework called Chem2Bio2RDF [68]. Ongoing efforts on achieving interoperability via establishment of open standards and open source software have been carried out by the Blue Obelisk [69] and (Open PHArmacological Concepts Triple Store) [70] consortiums.

Computational tools have once been confined to specific operating systems and issues of incompatibility may have prevented widespread usage. In recent years, tools are increasingly becoming platform-independent making its transition from desktop-based to cloud-based applications. Furthermore, researchers from small laboratories can now have access to large cloud-based supercomputer for their drug discovery projects via cloud computing services such as Amazon Web Services and Google Cloud Platform. One common problem in the life science is reproducible research and this issue is being felt also in drug discovery not only to reproduce research but also to validate and compare one's own work with previously published computational models. There are several best practices reviewed in the literature for performing reproducible computational research [71,72] and it is envisioned that one day there would be a repository for storing computational models such that future research would not have to, so to say, 'reinvent the wheel' in order to use a previously published method to tackle a new research problem. This has started to be realized by workflow tools such as Taverna [73] and iPython [74] as well as the utilization of so-called open laboratory notebook to keep track of research protocols as well as performing them on-the-fly using the latest information [75].

7. Conclusion

In spite of extensive effort by industry and academia to develop new drugs, there are still several diseases that are in need of therapeutic agents and have yet to be developed. The end of the post-genomic era had marked the beginning of several new chapters in life science encompassing transcriptomics, proteomics and metabolomics that are in concomitant with advents in high-throughput synthesis and screening. This had resulted in big data for drug discovery and making sense of these data is no easy task. Nevertheless, it is apparent that data-driven computational tools provide high hopes that many of the diseases under investigation can be brought under control.

8. Expert opinion

It is apparent that advents in science and technology will continue to produce overwhelming volume of data for drug discovery. Thus, the ability to unravel the inherently hidden knowledge from the big data of drug discovery would take us a step closer toward developing a new drug. Thus, computational tools play integral roles in hit identification and lead optimization campaigns. In spite of its immense utility, computational tools should not be blindly taken as definitive black box answers but should be viewed as decision-supporting tools for aiding human practitioners who will ultimately determine the fate of the drug discovery program. With this in mind, it is pertinent to have a broad perspective on the big picture of drug discovery and awareness of the available tools for performing designated tasks. Selection of appropriate computational tools is highly dependent on the available data at hand and on a case-by-case basis. For example, if available data are primarily based on known active ligands in the absence of the protein crystal structure then employing ligand-based virtual screening approaches may be preferable unless a suitable homologous protein can be identified that can serve as a template for protein structure prediction thereby opening up the possibility to use structure-based approaches. Emphasis should not only be placed on obtaining strong binders or those affording good pharmacological activity but equal importance should also be made on obtaining compounds with good pharmacokinetic profiles and low drug adverse effects, which are essential factors that can help steer away from late-stage failure in drug development. It is also important to be well versed with the ins and outs of employed tools as to dodge from potential pitfalls and maximize chances of a project's success. Focus should not be placed on solely one computational method but instead should harness the power of several (if not all) approaches (i.e., ligand, structure and systems level) as each provides complementary and different view of the same problem. The paradigm shift from single to multi-target approach in concomitant with the need to handle the increasingly high-dimensional data is the perfect storm calling for its reliance on systems-based approaches. In future drug discovery endeavors, it is necessary to develop a more streamlined infrastructure that better supports interoperability among various databases and tools. Semantic web technology has started to make this a reality and continued progress is called for to link the ever-expanding data in the life sciences and drug discovery. Instrumental in this expansion is the utilization of open notebook and workflow tools that can contribute to reproducible and interactive research.

Declaration of interest

The authors acknowledge support from the Research Career Development Grant (number RSA5780031) from the Thailand Research Fund to C Nantasenamat. Partial support from Office of Higher Education Commission and Mahidol University under the National Research Universities Initiative is also gratefully acknowledged. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Hughes JP, Rees S, Kalindjian SB, et al. Principles of early drug discovery. Brit J Pharmacol 2011;162(6):1239-49

•• **A concise introduction to the topic of drug discovery.**

2. Doman TN, McGovern SL, Witherbee BJ, et al. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 2002;45(11):2213-21

3. Sliwoski G, Kothiwale S, Meiler J, et al. Computational methods in drug discovery. Pharmacol Rev 2013;66(1):334-95

4. Villoutreix BO, Lagorce D, Labbe CM, et al. One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. Drug Discov Today 2013;18(21-22):1081-9

5. Swiss Institute of Bioinformatics. Click2Drug: directory of computer-aided drug design tools. Available from: http://www.click2drug.org/ [Last accessed 18 July 2014]

6. Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. Nature 2004;432(7019):855-61

7. Ruddigkeit L, van Deursen R, Blum LC, et al. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model 2012;52(11):2864-75

• **Interesting work on computationally enumerating an exhaustively large set of 166 billion compounds of 17 atoms or less.**

8. Erlanson DA, McDowell RS, O'Brien T. Fragment-based drug discovery. J Med Chem 2004;47(14):3463-82

9. Bohm HJ, Flohr A, Stahl M. Scaffold hopping. Drug Discov Today Technol 2004;1(3):217-24

10. Wang Y, Xiao J, Suzek TO, et al. PubChem's BioAssay database. Nucleic Acids Res 2012;40 (Database issue):D400-12

11. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(Database issue):D1100-7

12. Liu T, Lin Y, Wen X, et al. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 2007;35(Database issue):D198-201

13. Williams AJ, Ekins S. A quality alert and call for improved curation of public chemistry databases. Drug Discov Today 2011;16(17-18):747-50

• **Suggestions for improving the quality of public chemistry databases are provided.**

14. Williams AJ, Ekins S, Tkachenko V. Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discov Today 2012;17(13-14):685-701

• **Another interesting article on resolving the quality issue of public chemistry databases.**

15. Trouiller P, Olliaro P, Torreele E, et al. Drug development for neglected diseases: a deficient market and a public-health policy failure. Lancet 2002;359(9324):2188-94

16. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov 2011;10(6):428-38

17. Judd DB. Open innovation in drug discovery research comes of age. Drug Discov Today 2013;18(7-8):315-17

18. Dorsch H, Jurock AE, Schoepe S, et al. Grants4Targets: an open innovation initiative to foster drug discovery collaborations. Nat Rev Drug Discov 2015;14(1):74-6

19. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, et al. A practical overview of quantitative structure-activity relationship. Excli J 2009;874-88

•• **Reviews the principles and application of QSAR.**

20. Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds. Exp Opin Drug Discov 2010;5(7):633-54

21. Nantasenamat C, Worachartcheewan A, Jamsak S, et al. AutoWeka: toward an automated data mining software for QSAR and QSPR studies. Methods Mol Biol 2015;1260:119-47

• **Provides a concise tutorial on quickly getting started on making QSAR models with an automated data mining software.**

22. Chen L, Li Y, Yu H, et al. Computational models for predicting substrates or inhibitors of P-glycoprotein. Drug Discov Today 2012;17(7-8):343-51

23. Li D, Chen L, Li Y, et al. ADMET evaluation in drug discovery. 13. Development of in silico prediction models for P-glycoprotein substrates. Mol Pharm 2014;11(3):716-26

24. Shen M, Tian S, Li Y, et al. Drug-likeness analysis of traditional Chinese medicines: 1. property distributions of drug-like compounds, non-drug-like compounds and natural compounds from traditional Chinese medicines. J Cheminform 2012;4(1):31

25. Bickerton GR, Paolini GV, Besnard J, et al. Quantifying the chemical beauty of drugs. Nat Chem 2012;4(2):90-8

• **A new drug-likeness measure called the quantitative estimate of drug-likeness (QED).**

26. Lipinski CA, Lombardo F, Dominy BW, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001;46(1-3):3-26

• **The classical Lipinski's rule-of-5 for filtering drug-like compounds.**

27. Worachartcheewan A, Mandi P, Prachayasittikul V, et al. Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. Chemometr Intell Lab Syst 2014;138:120-6

28. Pingaew R, Worachartcheewan A, Nantasenamat C, et al. Synthesis, cytotoxicity and QSAR study of N-tosyl-1,2,3,4-tetrahydroisoquinoline derivatives. Arch Pharm Res 2013;36(9):1066-77

29. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, et al. Predicting antimicrobial activities of benzimidazole derivatives. Med Chem Res 2013;22(11):5418-30

30. Nichols SE, Baron R, Ivetac A, et al. Predictive power of molecular dynamics

receptor structures in virtual screening. J Chem Inf Model 2011;51(6):1439-46

31. Tian S, Sun H, Pan P, et al. Assessing an ensemble docking-based virtual screening strategy for kinase targets by considering protein flexibility. J Chem Inf Model 2014;54(10):2664-79

32. Dixit A, Verkhivker GM. Integrating ligand-based and protein-centric virtual screening of kinase inhibitors using ensembles of multiple protein kinase genes and conformations. J Chem Inf Model 2012;52(10):2501-15

33. Durrant JD, Amaro RE, McCammon JA. AutoGrow: a novel algorithm for protein inhibitor design. Chem Biol Drug Des 2009;73(2):168-78

34. Durrant JD, Lindert S, McCammon JA. AutoGrow 3.0: an improved algorithm for chemically tractable, semi-automated protein inhibitor design. J Mol Graph Model 2013;44:104-12

35. Lindert S, Durrant JD, McCammon JA. LigMerge: a fast algorithm to generate models of novel potential ligands from sets of known binders. Chem Biol Drug Des 2012;80(3):358-65

36. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol 2008;18(3):342-8

37. Kalaszi A, Szisz D, Imre G, et al. Screen3D: a novel fully flexible high-throughput shape-similarity search method. J Chem Inf Model 2014;54(4):1036-49

38. Irwin JJ, Shoichet BK. ZINC - a free database of commercially available compounds for virtual screening. J Chem Inf Model 2005;45(1):177-82

39. Pence HE, Williams A. ChemSpider: an online chemical information resource. J Chem Educ 2010;87(11):1123-4

40. Tanrikulu Y, Kruger B, Proschak E. The holistic integration of virtual screening in drug discovery. Drug Discov Today 2013;18(7-8):358-64

•• **Proposes the classification of virtual screening and discusses about how its integration with wet screening can increase enrichment rates.**

41. Wei NN, Hamza A. SABRE: ligand/structure-based virtual screening approach using consensus molecular-shape pattern recognition. J Chem Inf Model 2014;54(1):338-46

42. Scior T, Bender A, Tresadern G, et al. Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 2012;52(4):867-81

•• **A must read for all researchers in drug design to gain understanding on the potential pitfalls that lurks in virtual screening and how to avoid them.**

43. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. J Mol Graph Model 2008;26(8):1315-26

• **Provides a good coverage on the importance of applicability domain in virtual screening.**

44. Medina-Franco JL, Giulianotti MA, Welmaker GS, et al. Shifting from the single to the multitarget paradigm in drug discovery. Drug Discov Today 2013;18(9-10):495-501

45. Yildirim MA, Goh KI, Cusick ME, et al. Drug-target network. Nat Biotechnol 2007;25(10):1119-26

• **An early study showing the multi-target nature of drugs as elucidated from a drug-target network.**

46. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004;3(8):673-83

• **A review of on drug repositioning, which is an important concept that can drastically save costs in drug development by reusing drugs against a different target.**

47. Bredel M, Jacoby E. Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat Rev Genet 2004;5(4):262-75

• **Overview of chemogenomics and how it can be used to quickly find putative interacting compounds for orphan receptors or find putative receptors for compounds.**

48. van Westen GJ, Wegner JK, Ijzerman AP, et al. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. Med Chem Comm 2011;2(1):16-30

•• **Reviews the principles and applications of proteochemometric modeling.**

49. Pavlopoulos GA, Secrier M, Moschopoulos CN, et al. Using graph theory to analyze biological networks. BioData Min 2011;4:10

50. Kharkar PS, Warrier S, Gaud RS. Reverse docking: a powerful tool for drug repositioning and drug rescue. Future Med Chem 2014;6(3):333-42

51. Mathias SL, Hines-Kay J, Yang JJ, et al. The CARLSBAD database: a confederated database of chemical bioactivities. Database (Oxford) 2013;2013:bat044

52. Eriksson M, Nilsson I, Kogej T, et al. SARConnect: a tool to interrogate the connectivity between proteins, chemical structures and activity Data. Mol Inform 2012;31(8):555-68

53. Carrascosa MC, Massaguer OL, Mestres J. PharmaTrek: a semantic web explorer for OpeniInnovation in multitarget drug discovery. Mol Inform 2012;31(8):537-41

54. Dakshanamurthy S, Issa NT, Assefnia S, et al. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem 2012;55(15):6832-48

•• **A study utilizing proteochemometrics modeling for drug repositioning.**

55. Pei J, Yin N, Ma X, et al. Systems biology brings new dimensions for structure-based drug design. J Am Chem Soc 2014;136(33):11556-65

56. Lu S, Li S, Zhang J. Harnessing allostery: a novel approach to drug discovery. Med Res Rev 2014;34(6):1242-85

57. Huang Z, Mou L, Shen Q, et al. ASD v2.0: updated content and novel features focusing on allosteric regulation. Nucleic Acids Res 2014;42(Database issue):D510-16

58. Uversky VN. Intrinsically disordered proteins and novel strategies for drug discovery. Expert Opin Drug Discov 2012;7(6):475-88

59. Milroy LG, Grossmann TN, Hennig S, et al. Modulators of protein-protein interactions. Chem Rev 2014;114(9):4695-748

60. Gashaw I, Ellinghaus P, Sommer A, et al. What makes a good drug target? Drug Discov Today 2011;16(23-24):1037-43

61. Iorio F, Rittman T, Ge H, et al. Transcriptional data: a new gateway to drug repositioning? Drug Discov Today 2013;18(7-8):350-7

62. Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 2005;15(10):1451-5

63. O'Boyle NM, Hutchison GR. Cinfony - combining open source cheminformatics

toolkits behind a common interface. Chem Cent J 2008;2:24

- **A useful cheminformatics framework implemented on Python that integrates several tools under one common interface.**

64. Grunberg R, Nilges M, Leckner J. Biskit - a software platform for structural bioinformatics. Bioinformatics 2007;23(6):769-70

65. Kalev I, Mechelke M, Kopec KO, et al. CSB: a Python framework for structural bioinformatics. Bioinformatics 2012;28(22):2996-7

66. Cokelaer T, Pultz D, Harder LM, et al. BioServices: a common Python package to access biological web services programmatically. Bioinformatics 2013;29(24):3241-2

67. Spjuth O, Willighagen EL, Guha R, et al. Towards interoperable and reproducible QSAR analyses: exchange of datasets. J Cheminform 2010;2(1):5

68. Chen B, Dong X, Jiao D, et al. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 2010;11:255

69. O'Boyle NM, Guha R, Willighagen EL, et al. Open data, open source and open standards in chemistry: the Blue Obelisk five years on. J Cheminform 2011;3(1):37

70. Williams AJ, Harland L, Groth P, et al. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today 2012;17(21-22):1188-98

71. Preeyanon L, Pyrkosz AB, Brown CT. Reproducible bioinformatics research for biologists. Implementing reproducible research: Chapman and Hall/CRC;Boca Raton, Florida. 2014

72. Sandve GK, Nekrutenko A, Taylor J, et al. Ten simple rules for reproducible computational research. PLoS Comput Biol 2013;9(10):e1003285

- **Guidelines for performing reproducible computational research is provided.**

73. Wolstencroft K, Haines R, Fellows D, et al. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic Acids Res 2013;41(Web Server issue):W557-61

74. Pérez F, Granger BE. IPython: a system for interactive scientific computing. Comput Sci Eng 2007;9(3):21-9

75. Voegele C, Bouchereau B, Robinot N, et al. A universal open-source electronic laboratory notebook. Bioinformatics 2013;29(13):1710-12

Affiliation

Chanin Nantasenamat*[1] &
Virapong Prachayasittikul[†2]
[†,*]Authors for correspondence
[1]Mahidol University, Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, 10700 Bangkok, Thailand
E-mail: chanin.nan@mahidol.ac.th
[2]Mahidol University, Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, 10700 Bangkok, Thailand
E-mail: virapong.pra@mahidol.ac.th

EXPERT OPINION ON DRUG DISCOVERY, 2017
VOL. 12, NO. 4, 345–362
http://dx.doi.org/10.1080/17460441.2017.1295954

REVIEW

Exploring the epigenetic drug discovery landscape

Veda Prachayasittikul [a*], Philip Prathipati [b*], Reny Pratiwi [a], Chuleeporn Phanus-umporn [a], Aijaz Ahmad Malik [a], Nalini Schaduangrat [a], Kanokwan Seenprachawong [c], Prapimpun Wongchitrat [d], Aungkura Supokawej [c], Virapong Prachayasittikul [e], Jarl E. S. Wikberg [f] and Chanin Nantasenamat [a]

[a]Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand; [b]National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan; [c]Department of Clinical Microscopy, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand; [d]Center for Research and Innovation, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand; [e]Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand; [f]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

ABSTRACT

Introduction: Epigenetic modification has been implicated in a wide range of diseases and the ability to modulate such systems is a lucrative therapeutic strategy in drug discovery.

Areas covered: This article focuses on the concepts and drug discovery aspects of epigenomics. This is achieved by providing a survey of the following concepts: (i) factors influencing epigenetics, (ii) diseases arising from epigenetics, (iii) epigenetic enzymes as druggable targets along with coverage of existing FDA-approved drugs and pharmacological agents, and (iv) drug repurposing/repositioning as a means for rapid discovery of pharmacological agents targeting epigenetics.

Expert opinion: Despite significant interests in targeting epigenetic modifiers as a therapeutic route, certain classes of target proteins are heavily studied while some are less characterized. Thus, such orphan target proteins are not yet druggable with limited report of active modulators. Current research points towards a great future with novel drugs directed to the many complex multifactorial diseases of humans, which are still often poorly understood and difficult to treat.

ARTICLE HISTORY
Received 30 October 2016
Accepted 13 February 2017

KEYWORDS
Epigenomics; epigenetics; drug discovery; drugs; bioinformatics; cheminformatics; chemogenomics; proteochemometrics

1. Introduction

Nature modulates the expression or activity of proteins and genes using three basic mechanisms: mutation/copy number variation/translocation collectively called genetic events [1], modulation via natural substrates (both small and large molecules) [2], and epigenetic regulation, primarily via posttranslational modifications (PTMs) such as acetylation and methylation of histones and DNA [3]. Epigenetics is a term coined by Conrad Hal Waddington to describe ‘the branch of biology which studies the causal interaction between genes and their products, which bring the phenotype into being’ [4]. Particularly, Waddington’s developmental landscape diagram metaphorically describes how an environmental stimulus may exert its influence on the inheritance of acquired characteristics [5,6]. This landmark event and other milestones in the history of epigenetics are summarized in Figure 1. Thus, epigenetic modification is pertinent for the regulation of gene expression and differentiation and as heritable changes in gene activity or cellular phenotype without alteration of the DNA sequence [3]. Such epigenetic modification encompasses three major epigenetic markers (e.g. writers, readers, and erasers) acting on three major substrates (e.g. DNA, histones, and noncoding RNAs). It is worthy to note that various components of the epigenetic machinery are highly interconnected and influenced by various factors (e.g. gender, nutrition, environmental and chemical factors, social and economic status, aging and stress) as summarized in Figure 2.

DNA methylation directly affects the genomic DNA and is accomplished by DNA methyltransferases (DNMTs), which add methyl donor groups from the S-adenosylmethionine (SAM) to the 5′ position of the cytosine ring within 5′-cytosine-phosphate-guanosine-3′ (CpG) sites [7,8]. DNA methylation plays an important role in genomic imprinting (i.e. for conveying parent to offspring gene control), suppression of retrotransposons, maintenance of genome stability, X-chromosome inactivation, as well as other types of gene regulation [9,10]. An equally important mechanism aside from DNA methylation is DNA demethylation [11], which is the removal of the methyl group making such a process necessary for the reprogramming of genes.

Histone associates with DNA to form a complex known as nucleosome [12]. The nucleosome is composed of pairs of histone proteins H2A, H2B, H3, and H4 that are assembled in an octameric core with 147 base pairs of DNA wrapped around it. Repeating nucleosomes are linked and stabilized by histone H1 that is folded up to form the chromatin structure [13]. A myriad of histone PTMs exists for the epigenetic regulation of gene expression and these can be classified on the basis of their catalytic mechanisms and the substrates that they read, write, or erase [14]. Mechanisms employed by PTMs and currently explored in drug discovery research are illustrated in Figure 3. A large proportion of the epigenetic target space is PTM enzymes involved in reading,

CONTACT Chanin Nantasenamat ✉ chanin.nan@mahidol.edu

*These authors contributed equally to this work.

Article highlights

- Epigenetic events act as a mediator between the external/environmental factors (cause) and the manifested genetic landscape regulation (effect). Therefore, lifestyle choices and environmental factors affect gene expression thereby influencing the etiology of various diseases.
- In a nutshell, epigenetic events allow a reversible and robust switching on and off of gene expression via the attachment and processing of chemical tags on DNA and histones thereby regulating the euchromatin state that is required for active transcription.
- Dysregulation of targets mediating epigenetic events can give rise to pathological deterioration associated with cardiovascular diseases, neurological disorders, metabolic disorders, and cancer development. Epigenomic targets thus represents an attractive avenue for pharmacological intervention by small molecules.
- Epigenetic drugs have primarily been studied for their use in treating cancer, however research on their use in Alzheimer's, Asthma and a myriad of other CNS, CVS, inflammatory/immune diseases is steadily rising.
- Great interest has been invested in the discovery of inhibitors or modulators against several key epigenetic drug targets. However, current epigenetic modulators lack isoform selectivity and produce global epigenetic changes. Advances in structural/functional genomics and chemical biology have provided insights into the isoform selectivity requirements and sequence-specific epigenetic chemical probes.
- Drug repositioning or drug repurposing represents an interesting route for the rapid discovery of novel drugs targeting the epigenetic machinery and for expanding the therapeutic indication profiles of existing epigenetic drugs.

This box summarizes key points contained in the article.

writing, and erasing the epigenetic marks [15]. While methylation and acetylation of histone residues are well-known writers of epigenetic marks, there are however a number of other modifications that can be classified into those that neutralize the positive charge of lysine and arginine (e.g. acetylation, butyrylation, citrullination, crotonylation, and propionylation) and those that retain (e.g. methylation) or add one or more negative charges (e.g. succinylation, malonylation, ADP ribosylation, and phosphorylation) [16]. The chromatin structure is highly dependent on the integrity of the nucleosome complex, which is mediated by the highly basic histone N-terminal tail that protrudes from the nucleosome and makes contact with adjacent nucleosomes [7]. Charge neutralization or gain in negative charge via PTM of these tails would not only affect internucleosomal interactions but also the nucleosomal-DNA complex structure and thus affect the overall chromatin structure and the expression levels of the corresponding genes [16].

In addition to the aforementioned histone and DNA modifiers, the epigenetic machinery is also composed of noncoding RNAs (ncRNAs). ncRNAs are functional RNA molecules that do not encode proteins. ncRNAs can bind DNA and alter its conformation, and in effect regulate gene expression, mRNA stability at the posttranscriptional level. Although important in its own rights, this review will not delve deeper into the topic as emphasis will be placed on DNA and histone modifications [17].

2. Factors influencing epigenetics

The epigenome situates itself at the interface between the genome and the environment [18] as summarized in Figure 2. Susceptibility to epigenetic alterations is life-time dependent [19] where gametogenesis and early embryogenesis are considered to be critical periods with high genome plasticity [18]. Epigenetic memory conveys inheritance from generation to generation [19], meaning that memory can be transferred across generations [18] without the need of re-exposure to the same epigenetically driven factors [19]. The effect of altered epigenetic markers in early life stages is not only displayed as acute adaptive responses, but can also gradually manifest as adult-onset diseases [18] upon secondary triggers (e.g. aging and hormonal changes) [20]. The plasticity and reversibility of the epigenome render the hosts susceptible to reprogramming when exposed to external factors [19]. Thus, studies toward understanding the role of external factors in epigenetic alterations could at least partially demystify the rather complex etiology of multifactorial diseases, and may provide a promising strategy for the prevention and treatment of many complex multifactorially driven diseases [21].

3. Drugging epigenetic targets

Owing to its broad involvement in a wide range of diseases, epigenetics have received great attention for the discovery of novel therapeutic agents. Figure 3 summarizes the various enzymes that mediate covalent modifications of DNA and histones as a collective mechanism for epigenetic control of gene transcription. The intense efforts in the field have already bore fruits as several FDA-approved drugs targeting the epigenetic machinery have been developed. Moreover, efforts on drug repositioning or repurposing has also led to the discovery of novel indications for existing FDA-approved drugs in the field. In this review, a few drug repositioning case studies are discussed in section 5 under the heading 'Drug repurposing opportunities for epigenetic drug discovery.' Chemical structures of these drugs are shown in Figure 4 while protein structures of epigenetic modifiers as determined by X-ray crystallography are illustrated in Figure 5.

Table 1 illustrates the available bioactivity data for epigenetic drug targets from the recent Release 22 of the ChEMBL database [22]. The availability of bioactivity data can be taken as a relative measure of their druggability and importance for drug discovery research. Table 2 summarizes the available crystal structures and bioactivity data of epigenetic protein-ligand complexes. Resources such as this, as well as other manually curated bioinformatics database of epigenetic enzymes (e.g. dbEM) [23] and chemogenomics databases of epigenetic protein–ligand interactions (e.g. HEMD) [23,24]) are great starting points for ligand and structure-based drug design efforts [25]. It is implied that some epigenetic enzymes are more popular among scientists than others, and as a result some may be more heavily studied while others may remain as orphan targets in need of attention. Computational chemogenomics and proteochemometrics [26,27] are promising approaches for suggesting potential ligands for orphan target proteins on the basis of the molecular similarity concept in which similar ligands are implied to bind to similar target proteins, as well as *vice versa* where potential target proteins could be suggested for a ligand of interest.

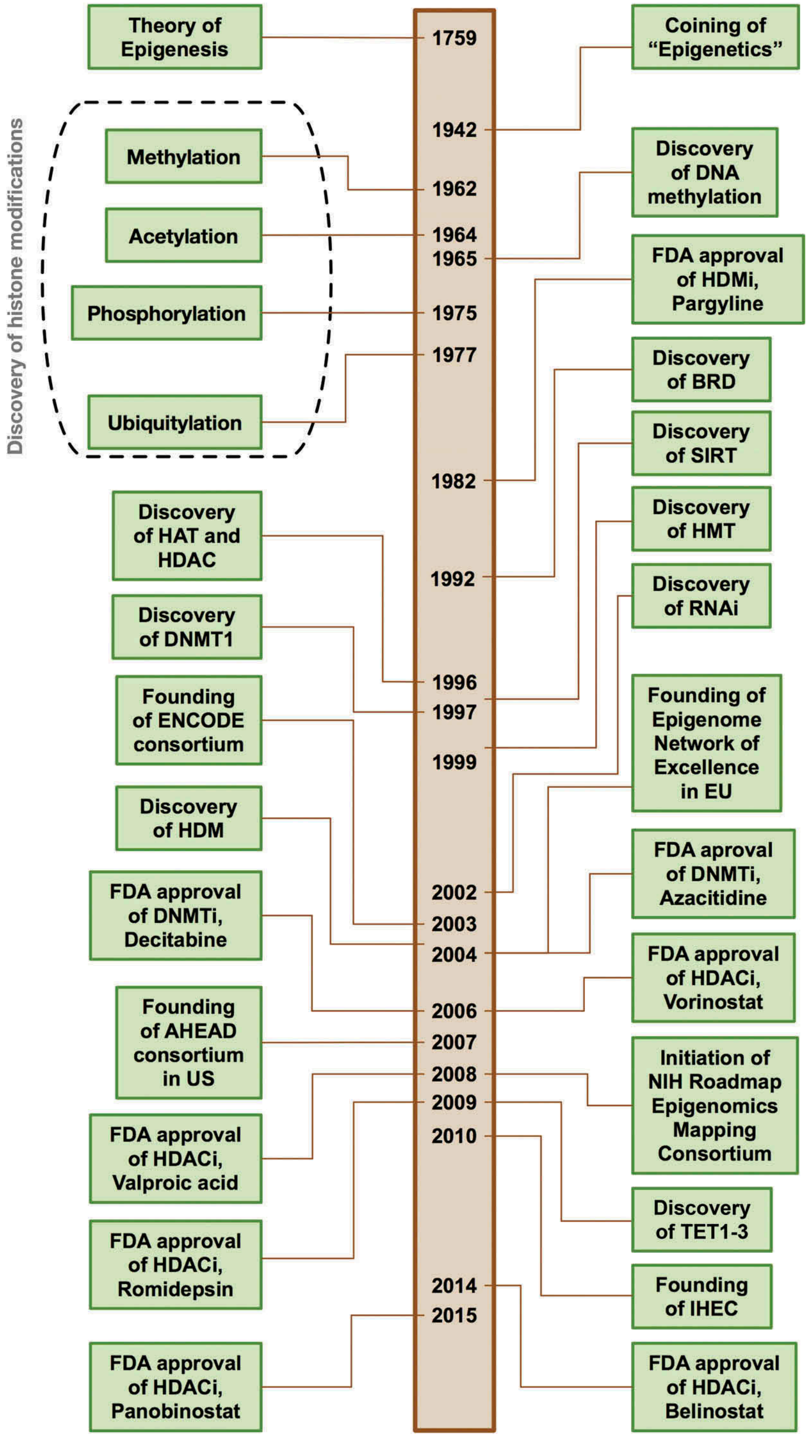

Theory of Epigenesis
1759
Coining of "Epigenetics"
Discovery of histone modifications
Methylation
Acetylation
Phosphorylation
Ubiquitylation
1942
Discovery of DNA methylation
1962
1964
1965
FDA approval of HDMi, Pargyline
1975
1977
Discovery of BRD
Discovery of SIRT
1982
Discovery of HAT and HDAC
Discovery of HMT
1992
Discovery of DNMT1
Discovery of RNAi
1996
Founding of ENCODE consortium
1997
Founding of Epigenome Network of Excellence in EU
1999
Discovery of HDM
FDA aproval of DNMTi, Azacitidine
2002
FDA approval of DNMTi, Decitabine
2003
2004
FDA approval of HDACi, Vorinostat
2006
Founding of AHEAD consortium in US
2007
2008
Initiation of NIH Roadmap Epigenomics Mapping Consortium
2009
2010
FDA approval of HDACi, Valproic acid
Discovery of TET1-3
FDA approval of HDACi, Romidepsin
2014
Founding of IHEC
2015
FDA approval of HDACi, Panobinostat
FDA approval of HDACi, Belinostat

Figure 1. Timeline of milestones in the history of epigenetics.

3.1. DNA methyltransferase inhibitors (DNMTi)

DNA methylation is controlled by the family of DNMTs comprising of DNMT1, DNMT2, DNMT3A, DNMT3B, and DNMT3 L [28]. The DNMTs share a common catalytic domain referred to as the AdoMet-dependent MTase fold [29]. Within this family, DNMT1 maintains methylation patterns after DNA replication while DNMT3A and DNMT3B together with its regulatory factor DNMT3L regulates *de novo* DNA methylation during the early embryonic development of mammals. Finally, DNMT2 is involved in cytoplasmic RNA methylation [30–33]. Alterations of DNA methylation (e.g. hypomethylation and hypermethylation) have been found to be correlated with cancers, genetic disorders, neurological and autoimmune diseases. Hence, DNMTs have gained prominence as drug targets and as such several small-molecule inhibitors targeting the DNMT family have been reported [34,35]. Two nucleoside-based DNMT inhibitors belonging to the cytidine chemotype, 5-aza-cytidine (Azacitidine or Vidaza) and 5-aza-2′-deoxycytidine (Decitabine or Dacogen), have been approved by the FDA for the

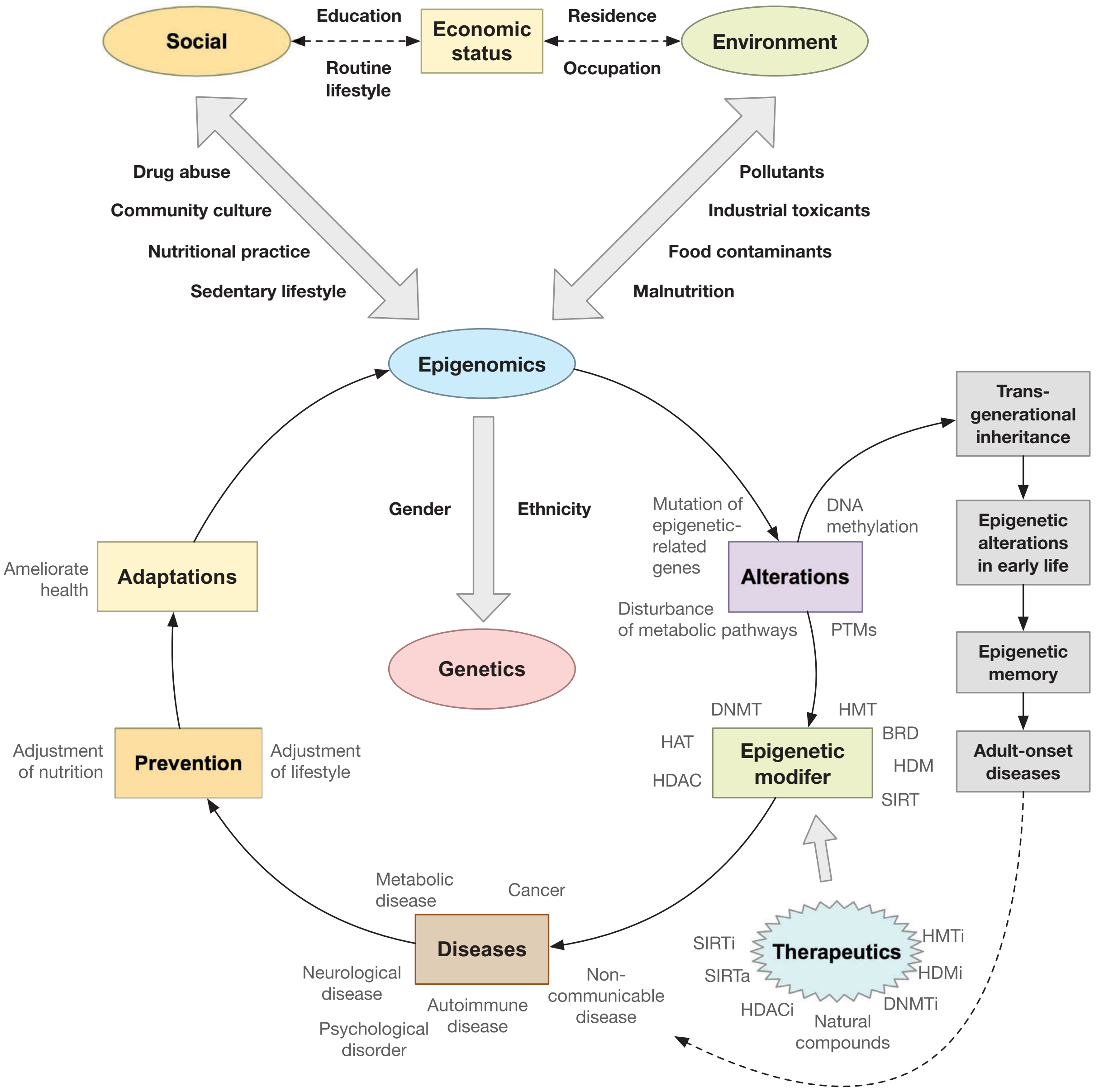


Figure 2. Epigenetics acts as a mediator between genetic and external factors in the development of diseases and its adaptations highlight its potential in therapy and disease prevention. Economic status may be considered to be a key factor that influences both social and environmental factors. An individual's income may govern their chance of toxic substance exposure as it governs their area of residence, education opportunity, occupation, workplace and nutritional practice.

treatment of myelodysplastic syndrome. These inhibitors intercalate between DNA base pairs and suppress the methylation of CpG islands that are generally enriched in transcriptionally relevant regulatory sequences also known as the promoter regions of genes [34,35]. However, DNMTi cytidine analogs are chemically unstable and have a highly promiscuous target association profile. Given these concerns, this calls for the need of more specific and selective DNMT inhibitors [36].

In recent years, the number of compounds tested as DNMT inhibitors have increased as reflected in the public databases. As summarized in Table 1, a query from the ChEMBL database [22,37] revealed that there were 502 compounds tested for DNMT1 (CHEMBL1993), 62 compounds tested for DNMT3A (CHEMBL1992), and 68 compounds for DNMT3B (CHEMBL6095). A number of potent non-nucleoside analogs targeting DNMT have been explored, including SGI-110, procainamide, epigallocatechin 3-gallate (EGCG), RG108, and hydralazine. SGI-110 (also known as guadecitabine or S110) is a CpG dinucleotide derivative of 5-aza-deoxycitidine (5-aza-dC or decitabine) [36]. Particularly, it is an oligonucleotide consisting of decitabine linked to the endogenous nucleoside deoxyguanosine via a phosphodiester bond. It is considered to be an efficient prodrug of decitabine [36] as the dinucleotide configuration provides protection against drug degradation by cytidine deaminase while maintaining the effect of its active metabolite, decitabine [38,39]. Thus, SGI-110 is considered to be a potent inhibitor of DNA methylation [38]. To date, SGI-110 is undergoing a phase III clinical trial for myelodysplastic syndrome and acute myeloid leukemia and a phase II clinical trial for hepatocellular carcinoma (http://clinicaltrials.gov). In addition to the identification of non-nucleoside analogs for

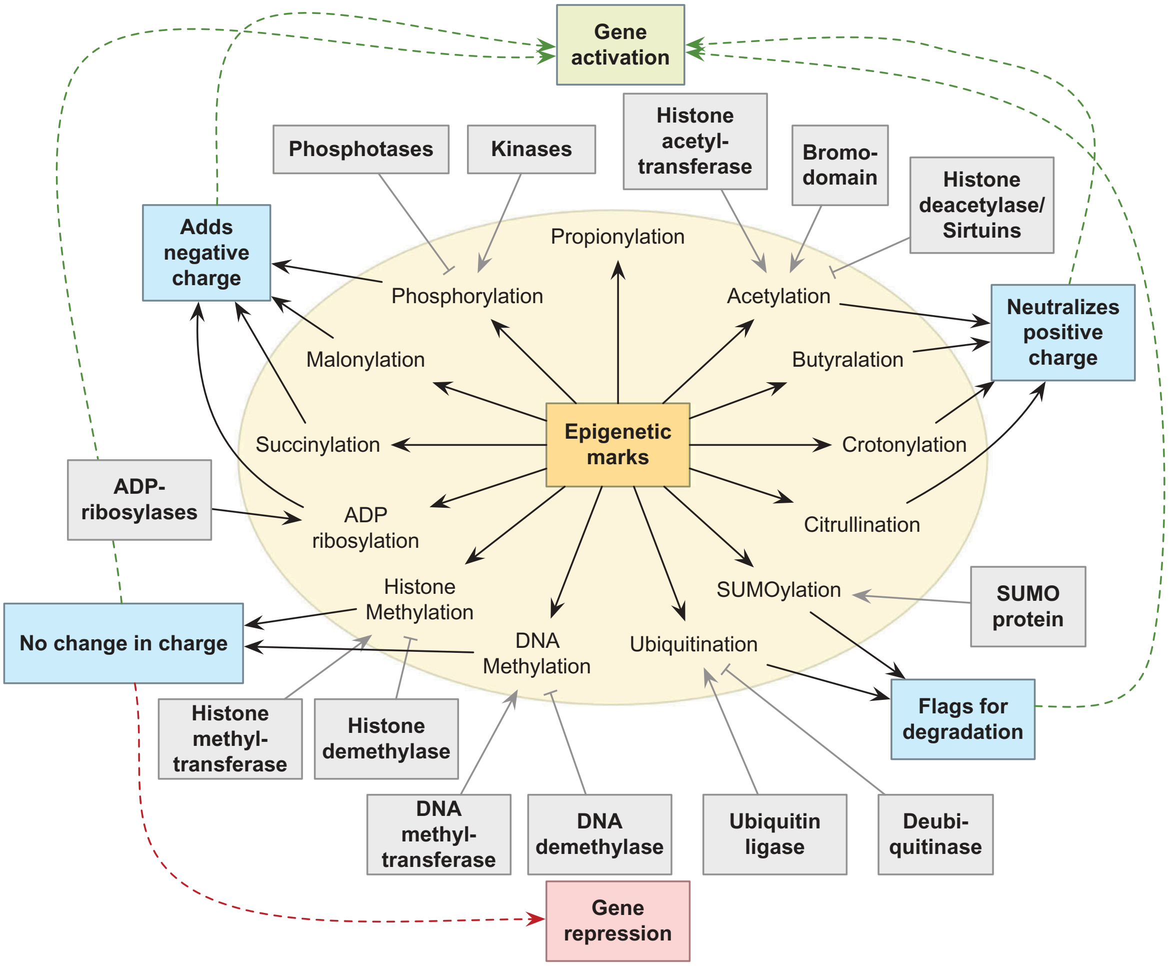


Figure 3. Illustration of the comprehensive spectrum of various DNA and PTM-mediated epigenetic marks. Effects of these marks on the charge and activation/repression of gene expression via an open chromatin structure is shown schematically. The open chromatin structure is characterized by the loss of complementary interactions between negatively-charged DNA and positively-charged residues in histone tails. The neutralization of positive charges or gain of negative charges leads to disruption of complementary interactions and results in an open chromatin structure and facilitates activation of gene expression.

DNMT targets, several strategies have been proposed including the development of allosteric inhibitors, SAM analogs for DNMT, DNA substrate competitors, combining two DNMT substrates SAM and cytosine/deoxycytidine in a single structure, and molecules for disruption of protein–protein interactions [36].

3.2. Histone acetyltransferase inhibitors (HATi)

Histone acetyltransferases (HATs) were first identified as regulators of tumor suppressors and were implicated in several diseases, including cancer progression, viral infection, and certain respiratory disorders [40]. Three naturally occurring small molecules have been described as HAT inhibitors: curcumin, garcinol, and anacardic acid [41]. Curcumin is an EP300- and CREBBP-specific inhibitor capable of repressing EP300-mediated p53 acetylation *in vivo* [42]. Its antitumor activities in a wide variety of cancers included, respectively, the downregulation and upregulation of CCND1 (cyclin D1) and CASP8 (caspase-8), as well as the inhibition of constitutive nuclear factor-κB (NF-κB) activation [43]. Garcinol and anacardic acid are both EP300 and KAT2B HAT inhibitors. Although, garcinol exhibits a much better cell permeability than anacardic acid, both may improve cancer therapy. Whereas, garcinol has been shown to induce apoptosis in HeLa cells while anacardic acid can sensitize cancer cells to ionizing radiation. A few other small molecules have been described as HAT inhibitors, but to date only a series of isothiazolones affecting EP300 and KAT2B activity were found to inhibit the growth of colon and ovarian cancer cells.

3.3. Histone deacetylase inhibitors (HDACi)

To date, 18 HDACs have been identified in mammals and categorized into four structurally and phylogenetically distinct classes, namely class I, IIA, IIB, and III. Class I is homologous to yeast Rpd3 deacetylase, IIA and IIB are homologous to yeast Hda1 deacetylase, and III is homologous to yeast Sir2. Interestingly, HDAC11 shows homology to enzymes of both classes I and II but is classified as a class IV enzyme. Class I and II HDACs as well as HDAC11 are zinc-dependent hydrolases whereas class III sirtuins are NAD-dependent enzymes. These enzymes are implicated in a wide variety of biological processes, such as apoptosis, differentiation, proliferation, and senescence [44]. Referring to

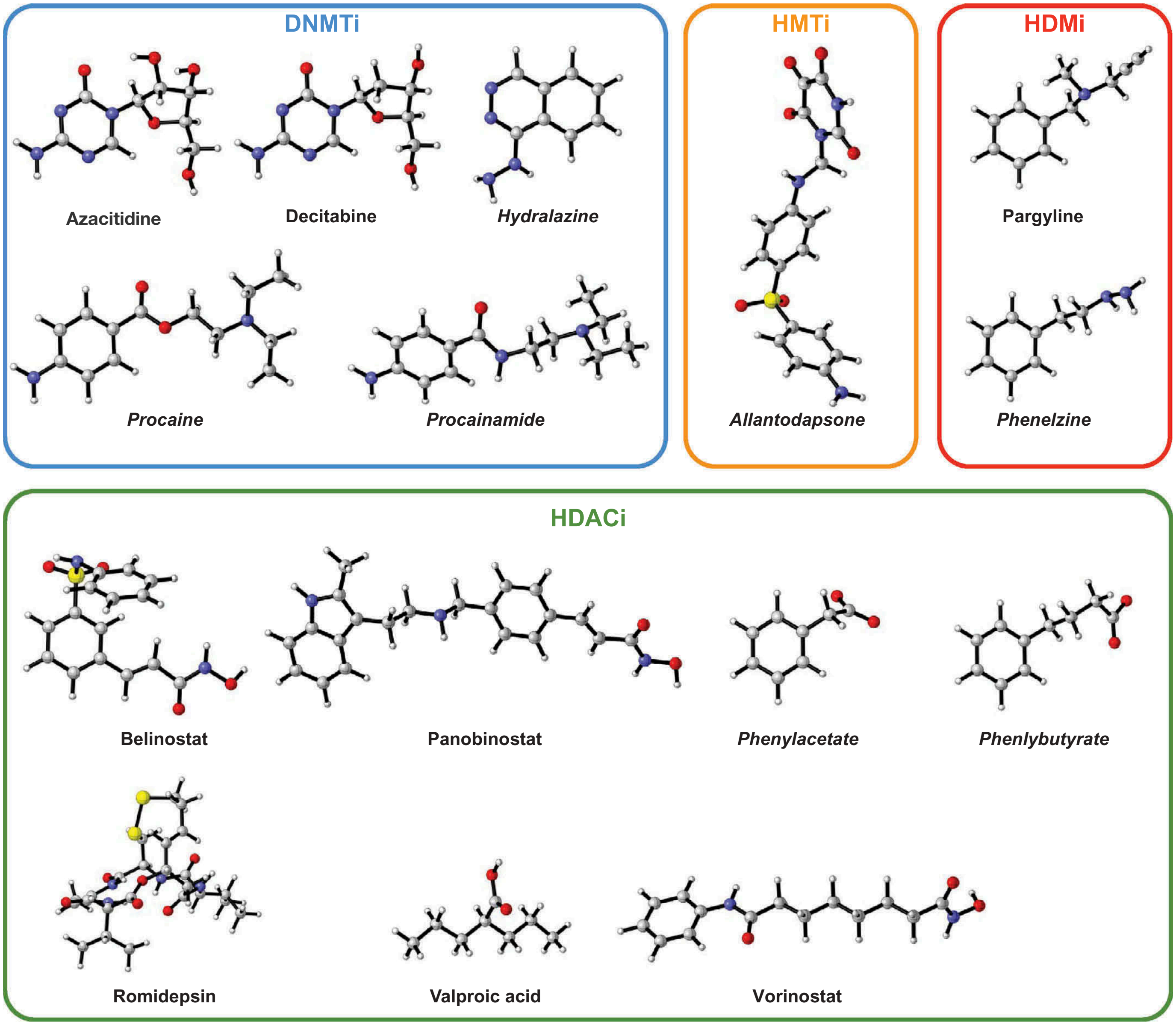


Figure 4. Chemical structures of FDA-approved drugs targeting the epigenetic machinery. Drugs that are FDA-approved for epigenetic targets are indicated by bold text whereas those that were approved for other indications and repurposed for epigenetic targets are highlighted as bold italic text. Epigenetic targets consisted of DNA methyltransferase inhibitors (DNMTi), histone methyltransferase inhibitors (HMTi), histone demethylase inhibitors (HDMi) and histone deacetylase inhibitors (HDACi).

Table 1, it is interesting to note that much of the current efforts have been directed toward HDAC1, HDAC6, HDAC8, and SIRT1 and SIRT2 with reported number of compounds showing bioactivity amounting to 3822, 2117, 1371, 1240, and 1260, respectively, while other HDACs have accumulated less than 1000 compounds.

The essential ligand-based pharmacophoric requirements for HDAC can be summarized as follows: (i) a capping group that interacts with residues at the active site entrance, (ii) a Zn-binding group (ZBG) that coordinates with the catalytic metal atom within the active site, and (iii) a linker group that binds with hydrophobic tunnel residues and positions the ZBG and the capping group for interaction in the active site [45]. Several HDACi chemotypes have been developed consisting of short-chain fatty acids (e.g. sodium butyrate, phenylbutyrate, pivanex, and valproic acid), cyclic tetrapeptides and natural compounds as well as the newer and more selective classes consisting of hydroxamic acids (e.g. vorinostat, belinostat, panobinostat, and dacinostat), benzamides (e.g. entinostat and mocetinostat), and bicyclic depsipeptide (e.g. romidepsin) [46].

Particularly, the majority of compounds under clinical trials are hydroxamic acid analogs [47]. The clinical success of hydroxamic analogs has been demonstrated first for the FDA-approved drug vorinostat [48]. HDAC inhibitory activity of compounds in this class can be attributed to the crucial polar hydroxamic group that interacts with the Zn-binding protein or chelates the Zn ion located at the catalytic site of the enzyme pocket, thereby leading to the inhibition of deacetylation [48]. Romidepsin, a cyclic tetrapeptide, is a naturally derived FDA-approved drug that blocks HDAC activity via the reduction of thiol released from the cell

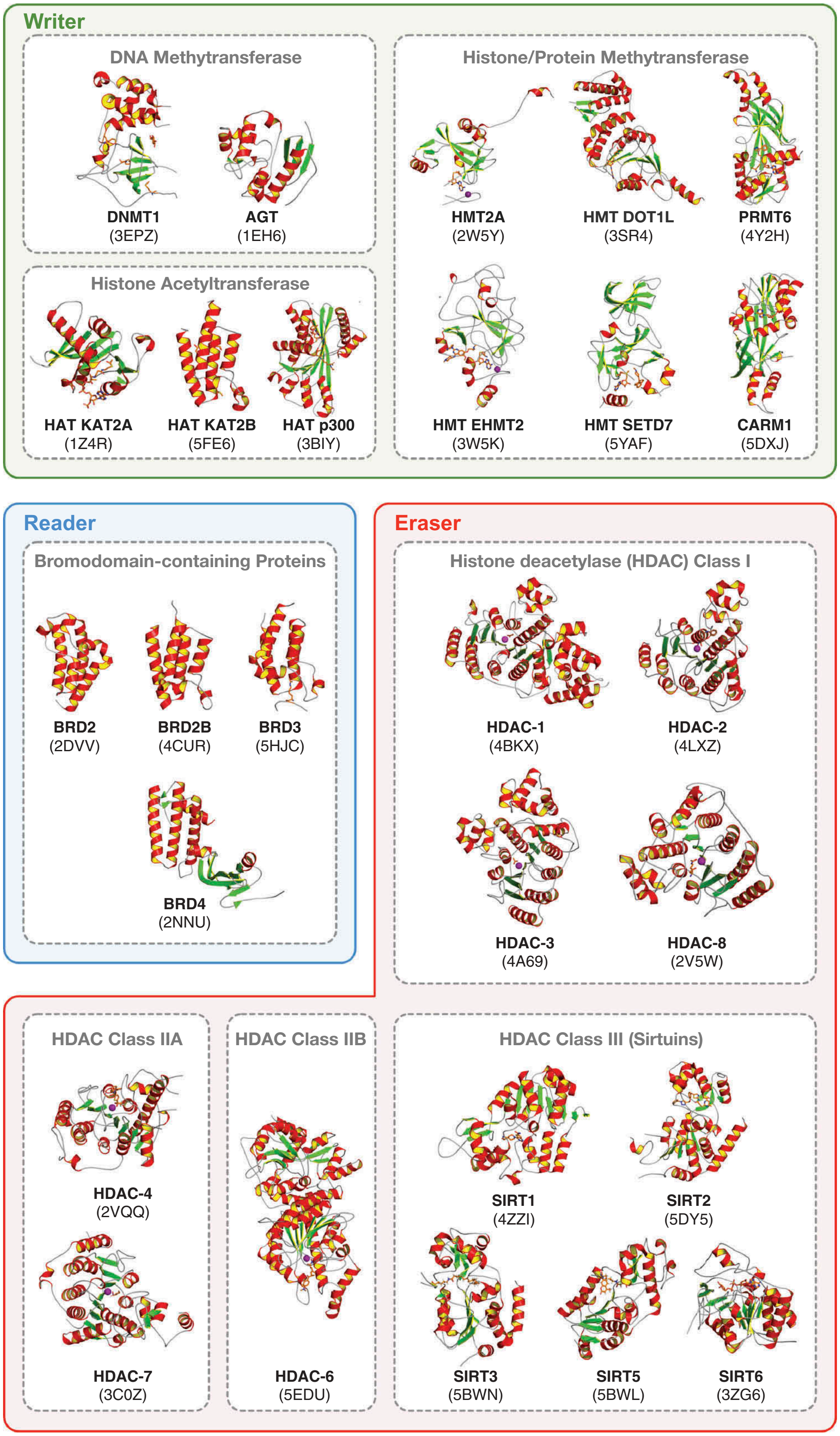


Figure 5. Protein structures of epigenetic drug targets. Proteins are classified into the three classes of epigenetic modifiers consisting of writers, readers and erasers. Alpha-helices are represented as a red ribbon (inner face shown in yellow), beta-strands are shown in green, loops are in gray and Zn ions are depicted as a purple sphere. Each structure is labeled by its acronym followed by the PDB ID in parenthesis on the subsequent line. Protein names and their acronyms are listed as follows: DNA (cytosine-5)-Methyltransferase-1, DNMT1; O-6-Alkylguanine-DNA Alkyltransferase, AGT; Histone-Lysine *N*-Methyltransferase 2A, HMT2A; Histone-Lysine *N*-Methyltransferase, H3 Lysine-79 specific (H3K79) DOT1L, HMT DOT1L Histone-Lysine *N*-Methyltransferase EHMT2, HMT EHMT2; Histone-Lysine *N*-Methyltransferase SETD7, HMT SETD7; Protein Arginine Methyltransferase PRMT6, PRMT6; Histone-Arginine Methyltransferase CARM1, CARM1; Histone Acetyltransferase KAT2A, HAT KAT2A; Histone Acetyltransferase KAT2B, HAT KAT2B; and Histone Acetyltransferase p300, HAT p300; Bromodomain-containing Protein-2, BRD2; Bromodomain containing Protein-2B, BRD2B; Bromodomain containing Protein-3, BRD3; Bromodomain containing Protein-4, BRD4; Histone Deacetylases 1 to 8, HDAC1 to HDAC8; and Sirtuins 1 to 6, SIRT1 to SIRT6. Full color available online.

Table 1. Summary of bioactivity data for compounds targeting the three classes of epigenetic modifiers.

CHEMBL ID	Target Name	UniProt	Number of compounds	Number of end points
Writer				
CHEMBL2864	6-O-methylguanine-DNA methyltransferase	P16455	167	387
CHEMBL1993	DNA (cytosine-5)-methyltransferase 1	P26358	502	597
CHEMBL3784	Histone acetyltransferase p300	Q09472	268	406
CHEMBL5500	Histone acetyltransferase PCAF	Q92831	393	500
CHEMBL5501	Histone acetyltransferase GCN5	Q92830	13857	14188
CHEMBL2189110	Histone-lysine N-methyltransferase EZH2	Q15910	338	589
CHEMBL6032	Histone-lysine N-methyltransferase, H3 lysine-9 specific 3	Q96KQ7	90127	92115
CHEMBL1293299	Histone-lysine N-methyltransferase MLL	Q03164	17174	17203
CHEMBL1795117	Histone-lysine N-methyltransferase, H3 lysine-79 specific	Q8TEK3	121	185
CHEMBL5523	Histone-lysine N-methyltransferase SETD7	Q8WTS6	166	188
CHEMBL5406	Histone-arginine methyltransferase CARM1	Q86X55	144	202
CHEMBL5524	Protein-arginine N-methyltransferase 1	Q99873	329	544
CHEMBL2093861	Menin/Histone-lysine N-methyltransferase MLL	O00255, Q03164	44110	47950
Reader				
CHEMBL1293289	Bromodomain-containing protein 2	P25440	237	346
CHEMBL1741220	Bromodomain adjacent to zinc finger domain protein 2B	Q9UIF8	55612	55612
CHEMBL1795186	Bromodomain-containing protein 3	Q15059	190	285
CHEMBL1163125	Bromodomain-containing protein 4	O60885	873	1646
Eraser				
CHEMBL325	Histone deacetylase 1	Q13547	3822	5453
CHEMBL1937	Histone deacetylase 2	Q92769	933	1334
CHEMBL1829	Histone deacetylase 3	O15379	862	1124
CHEMBL3524	Histone deacetylase 4	P56524	912	1274
CHEMBL2563	Histone deacetylase 5	Q9UQL6	333	416
CHEMBL1865	Histone deacetylase 6	Q9UBN7	2117	2969
CHEMBL2716	Histone deacetylase 7	Q8WUI4	413	524
CHEMBL3192	Histone deacetylase 8	Q9BY41	1371	1664
CHEMBL4145	Histone deacetylase 9	Q9UKV0	234	295
CHEMBL5103	Histone deacetylase 10	Q969S8	285	359
CHEMBL3310	Histone deacetylase 11	Q96DB2	245	302
CHEMBL4506	Sirtuin 1	Q96EB6	1240	2133
CHEMBL4462	Sirtuin 2	Q8IXJ6	1260	1921
CHEMBL4461	Sirtuin 3	Q9NTG7	373	467
CHEMBL2163183	Sirtuin 5	Q9NXA8	277	317

Briefly, writers, readers and erasers adds, reads and removes epigenetic marks such as methyl and acetyl. Data was obtained from ChEMBL database (version 22)[22] where only those with a compound count exceeding 100 are shown.

through the formation of a disulfide bond [47]. Thiol is essential for the interaction with the Zn-dependent pocket of HDACs and therefore the decreased availability of thiol leads to HDAC inhibition [47]. Following the discoveries of vorinostat and romidepsin, analogs of clinically potent second-generation inhibitors have been developed to improve their specificity and toxic profiles [48]. Emphasis has been paid on compounds belonging to the classes of hydroxamic acids (e.g. panobinostat, givinostat, and belinostat) and benzamides (e.g. entinostat and mocetinostat) [48]. Compounds in this generation exhibit improved profiles (i.e. improved efficacy, pharmacodynamic and pharmacokinetic properties with decreased toxicity). However, their mechanisms of action are the same as the clinically used ones and their ability to produce more effective clinical outcomes has yet to be seen [48]. Some promise can be found in classes I (e.g. RG2833, PCI-34051) and II (e.g. trifluoromethylooxadiazole (TFMO)) HDACi that are currently under preclinical development [48]. Of particular note is that adamantane and noradamantane are crucial scaffolds where compounds possessing these moieties exhibited HDAC inhibitory effects in the picomolar range [48]. Furthermore, natural compounds (e.g. diallyl disulfide, resveratrol, and spiruchostatin A) and other scaffolds (e.g. thioesters, epoxides, and electrophilic ketones) were also reported as HDACi [49].

Based on the analysis of existing HDACi and active compounds [50], several important issues could be considered in the design of HDAC inhibitors. First, as illustrated in Figure 6(a), isoform selective inhibition of HDACs could achieve beneficial efficacy profiles. With the exception of a few reports from the Bradner lab [51] and a few other groups, very few structure–activity relationship (SAR) studies have been reported aiming to improve the isoform selectivity. Second, the majority of HDACi have hydroxamic acid as a Zn-binding group. Due to concerns regarding the toxicity of the hydroxamic acid substructure and the general nature of Zn-chelating fragments, the identification of alternative Zn-binding groups or non-chelating fragments complementary to residues of the Zn-containing pocket are highly desirable [52]. Third, most computational studies of HDAC enzymes have not discussed or adequately compared the ionization states of HDAC enzymes and bound ligands. This limits the insights gained from most pharmacophore modeling studies that have only characterized metal groups as hydrogen bond acceptor/donor groups but not in terms of ionizable features, which are commonly seen in most metal chelators [50]. This has in turn critically limited the utility of pharmacophore models for virtual screening endeavors in the identification of HDACi with novel Zn-chelating fragments. Fourth, since HDACi is predominantly a metal chelator, the creation of a more effective scoring function that can effectively deal with molecular recognition events (i.e. the coordinate covalent bond formation) is needed. In particular, such scoring functions are required to advance HDACi design and development while generally advancing docking efforts against metal-containing proteins [50].

Table 2. Summary of X-ray crystal structures of epigenetic drug targets and their bound active ligand.

PDB ID	Target name	UniProt ID	Compound name	PDBeChem code	Bioactivity type	Bioactivity value (nM)
Writer						
4NVQ	Histone-lysine N-methyltransferase, H3 lysine-9 specific 3 (EHMT2)	Q96KQ7	A-366	2OD	IC_{50}	3.3
3QOX	Histone-lysine N-methyltransferase, H3 lysine-79 specific (DOT1L)	Q8TEK3	S-adenosyl-L-homocysteine	SAH	K_i	270
Reader						
4UYG	Bromodomain- containing protein 2 (BRD2)	P25440	I-BET726 (GSK1324726A)	73B	K_D	4.4
4NRB	Bromodomain adjacent to zinc finger domain protein 2B (BRD2B)	Q9UIF8	N01197	2LX	IC_{50}	38000
3LQJ	Histone-lysine N-methyltransferase 2A (MLL)	Q03164	H3(1–9)K4me3 peptide	-	K_D	4300
2YEL	Bromodomain- containing protein 4 (BRD4)	O60885	GW841819X	WSH	IC_{50}	15.5
Eraser						
3MAX	Histone deacetylase 2 (HDAC2)	Q92769	N-(4-aminobiphenyl-3-yl)benzamide	LLX	IC_{50}	27
2VQO	Histone deacetylase 4 (HDAC4)	P56524	2,2,2-trifluoro-1-(5-{3-phenyl-5H,6H,7H,8H-imidazo[1,2-a] pyrazine-7-carbonyl}thiophen-2-yl)ethane-1,1-diol	TFG	IC_{50}	317
3ZNR	Histone deacetylase 7 (HDAC7)	Q8WUI4	TMP269	NU9	K_i	36
2V5X	Histone deacetylase 8 (HDAC8)	Q9BY41	(2 R)-N~8~-hydroxy-2-{[(5-methoxy-2-methyl-1H-indol-3-yl) acetyl]amino}-N~1~-[2-(2-phenyl-1H-indol-3-yl)ethyl] octanediamide	V5X	IC_{50}	100

Only druggable targets discussed in Table 1 were processed. The bioactivity data were extracted from the BindingMOAD [55].

3.4. Sirtuin inhibitors (SIRTi) and modulators

Development of sirtuin (SIRT) modulators is an ongoing research where most compounds are still under preclinical investigation. Among all human SIRTs, the discovery of modulators has been driven toward SIRT1 and SIRT2. Specific inhibitors against SIRT1 have been suggested for cancer treatment [47]. SIRTi can be classified by their scaffolds as β-naphthols (e.g. sirtinol, splitomicin, salermide, and cambinol), indoles (e.g. EX-527 and oxyindole) and ureas (e.g. suramin and tenovin) [53]. In addition, other types such as chalcone and 1,4-dihydropyridine have been reported to inhibit SIRTs [53]. Great attention has been given to SIRT1 activators for conveying neuroprotection [47]. In addition, phenol derivatives such as resveratrol, quercetin, and piceatannol have been reported as SIRT1 activators [47]. Of note, resveratrol and its synthetic derivatives (e.g. SRT1720 and SRT2183) are promising compounds undergoing clinical trials [47]. These resveratrol-based compounds have been suggested to act as allosteric enzyme activators [47]. Furthermore, a different mechanism of SIRT1 activation has been reported for isonicotinamide whereby it interacts competitively with an endogenous SIRT1 inhibitor (e.g. nicotinamide) in order to promote deacetylation [47].

3.5. Histone demethylase inhibitors (HDMi)

In humans, the demethylation of N-methyl lysine residue is catalyzed by two distinct subfamilies of demethylases (KDMs), the flavin-dependent KDM1 subfamily and the 2-oxoglutarate (2OG)-dependent JmjC subfamily, both of which employ oxidative mechanisms [54]. Modulation of the histone methylation status is proposed to be important in epigenetic regulation and has substantial medicinal potential for the treatment of diseases including cancer and genetic disorders. Demethylases of the LSD1/KDM1 family share some sequence and structural similarities to amine oxidases and monoamine oxidase. Consequently, inhibitors of monoamine oxidases (MAOi) such as pargyline, phenelzine, and tranylcypromine can also inhibit the HDM KDM1A (Figure 4) [56]. Increasing the arsenal of inhibitors against the many HDMs involved in cancer will be a major challenge in the coming years. Furthermore, studies on the selective inhibition of the catalytic domain from both human KDM1/LSD and JmjC families of KDMs are progressing rapidly. Although these studies are at a relatively early stage, the signs suggest that with sufficient medicinal chemistry efforts, it will be possible to make highly potent and selective inhibitors against the catalytic domains from both families of human KDMs. To date, most KDM1 and JmjC KDM inhibition efforts have been focused on the extension of known inhibitors for other family members (i.e. mechanism-based inhibition of KDM1s and active site iron chelators for the JmjC KDMs). It is likely that the extension of those methods (i.e. by competing with histone substrate binding interactions) will lead to highly selective inhibitors of the catalytic domains [57].

3.6. Histone methyltransferase inhibitors (HMTi)

Histone/protein methyltransferases (HMTs/PMTs) catalyze the transfer of methyl groups from SAM to the side chains of lysine or arginine on the target protein. PMTs can be classified into lysine and arginine methyltransferases (PKMTs and PRMTs, respectively) [58]. All PKMTs contain the conserved catalytic 'SET' (Su(var)3–9, Enhancer-of-zeste, and Trothorax) domain whereby cofactors and substrates bind, with the exception of DOT1L [59,60]. The binding pocket of SAM and

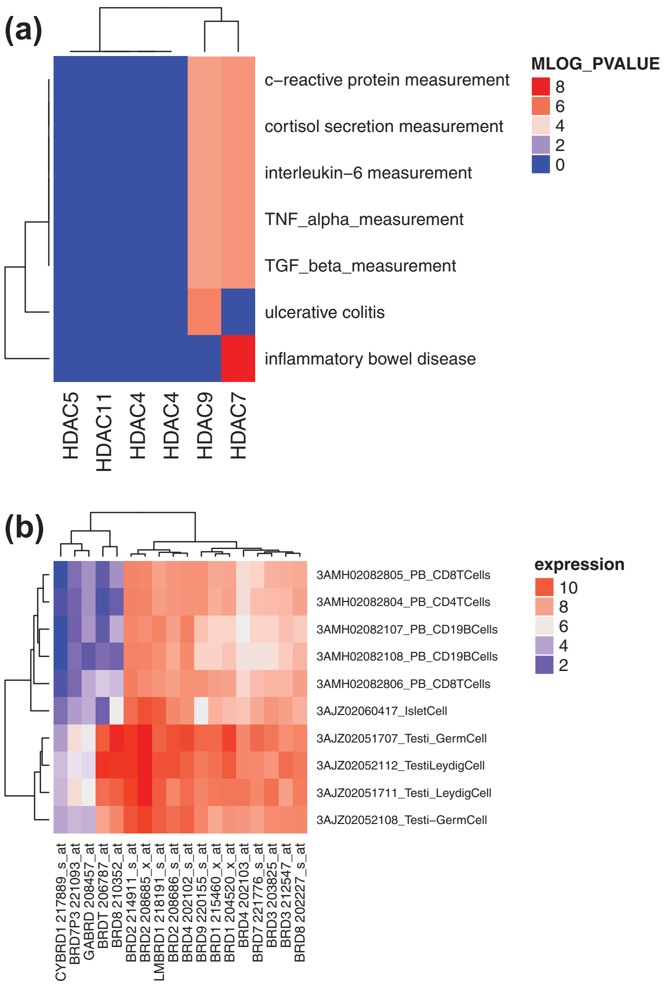


Figure 6. Analysis of the GWAS catalog and BioGPS data to illustrate the advantages of isoform selective modulation of HDACs and BETs respectively. (a) HDAC7 and HDAC9 show high association (in terms of MLOG_PVALUE) with various immune response phenotypes. The HDAC sub-family gene-disease associations were extracted from the GWAS catalog database and analyzed using the Bioconductor package 'gwascat' under the R programming environment. (b) High expression of BRDT in testis cells in comparison to inflammatory-like cells suggests that BRDT inhibition may play no role in anti-inflammatory effects and further reduces the development of testis causing temporary infertility-like effects. The BRD sub-family gene-tissue expression values were extracted from the BioGPS gene expression data set. The BioGPS expression data, GSE1133, was downloaded using the Bioconductor package 'GEOquery' and heatmaps were generated using the 'complexHeatmap' package in R.

amino acid provides structural features for inhibitor interaction and thus makes these enzymes attractive targets for intervention by small-molecule inhibitors [58].

Among many PKMTs, DOT1L and EZH2 are examples of attractive targets for epigenetic therapy. DOT1L is involved in inappropriate methylation of H3K79 and expression of HOX genes that drive leukemia [58]. EPZ004777 was one of the first SAM-competitive inhibitors intended to alter DOT1L, but has poor pharmacokinetic properties. Thus, second-generation DOT1L inhibitors with improved properties were made, such as EPZ-5676, which has successfully completed a phase I clinical trial [61]. In addition, hypermethylation of H3K27 by EZH2 promotes transcriptional silencing whereas high expression of EZH2 is associated with many types of cancer [58]. The first-generation EZH2 inhibitor 3-deazaneplanocin-A (DZNep) targets the S-adenosyl homocysteine activity and leads to alterations of methionine metabolism [62]. Furthermore, several potent SAM-competitive inhibitors including CPI-1205, GSK2816126, and EPZ6438 have been discovered and are currently undergoing clinical trials for treatment of hematological malignancies (http://clinicaltrials.gov). In addition, a growing number of compounds have been tested against other families of lysine methyltransferase including histone-lysine N-methyltransferase, H3 lysine-9 specific 3 (CHEMBL6032), and histone-lysine N-methyltransferase MLL (CHEMBL1293299) (Table 1).

PRMTs are structurally distinct with a conserved methyltransferase domain, a β-barrel specific to PRMTs, and a dimerization domain [59]. Although a number of PRMTs families have been associated with cancer, neurodegenerative diseases, and inflammatory diseases [63], the development of small-molecule inhibitors targeting PRMTs are still limited. An attempt was directed to identify a potent inhibitor targeting CARM1 (PRMT4) that catalyzes the methylation of H3R17. Compound RM65 is a drug-like inhibitor that induced hypomethylation in HepG2 cells [64]. Later, an inhibitor derived from plants, namely TBBD (ellagic acid) has been identified as a specific inhibitor of CARM1 [65]. Recently, a potent and selective inhibitor of PRMT5 with anti-proliferative activity has been characterized. This compound, EPZ015666 (GSK3235025), has entered phase I clinical trial for the treatment of solid tumor and non-Hodgkin's lymphoma [66]. DZNep induces apoptosis in breast cancer MCF7 and colorectal HCT116 cells, where it promotes the depletion of the polycomb-repressive complex-2 proteins (e.g. EZH2) and inhibits methylation of H3K27 [41]. Additionally, the arginine-specific HMT inhibitor AMI-1 (arginine N-methyltransferase inhibitor-1) is believed to inhibit PRMT1, PRMT3, PRMT4, and PRMT6 [67]. The fact that PRMT4 is overexpressed in hormone-dependent cancers may encourage research on these particular inhibitors [68]. Owing to structural similarities, analogs of the AMI-1 derivative AMI-5 can inhibit not only lysine and arginine-specific HMTs but also some HATs and sirtuins with the same potency, thus giving rise to the term 'epigenetic multiple ligands' [69].

3.7. Bromodomain inhibitors (BRDi)

BRD2, BRD3, and BRD4 are among the well-studied proteins of the bromo and extra terminal (BET) family with bioactivity data of 237, 190, and 873 compounds as reported in the ChEMBL database. This class of proteins bind to the acetylated lysine of histones and have been associated with a range of diseases spanning from cancer to inflammation and cardiovascular diseases. Currently, ten compounds capable of blocking the protein–protein interactions of BET bromodomains have entered clinical trials [70]. A phase III clinical trial candidate, RVX-208, developed by Resverlogix Corp. has been evaluated in a total of seven clinical trials for the treatment of atherosclerosis and associated cardiovascular disease. RVX-208 increased the levels of HDL-cholesterol and apolipoprotein A1 as well as decreased the incidences of major adverse cardiac events (MACE) in patients with diabetes mellitus. However, in a phase IIb, randomized, double-blind, multicenter, ASSURE trial, RVX-208 showed no significant increase in either apoA-I or HDL-C, nor an incremental regression of atherosclerosis than that observed with administration of a placebo [71]. OTX015, BMS-9861158, and GSK525762 have also reached phase II clinical trials [72]. OTX015, developed by OncoEthix and Merck, is involved in four different clinical trials for the treatment of acute leukemia and hematologic malignancies, advanced solid tumors (NCT02259114), recurrent *multiforme gliobastoma* and in combination with azacitidine for the treatment of patients with newly diagnosed acute myeloid leukemia that are not candidates for standard intensive induction therapy. BMS-986158 (i.e. structure undisclosed) has been tested for multiple cancer indications alone and in combination with paclitaxel. Finally, GSK525762, also known as I-BET762, is involved in two clinical trials: one to investigate the safety, pharmacokinetics, pharmacodynamics, and clinical activity in patients with NUT midline carcinoma and other cancers, and a second one directed toward patients having solid tumors and hematologic malignancies [70]. In addition to these four molecules, six other BET inhibitors have recently entered phase I clinical trials and are being studied for both solid tumors and hematological malignancies: two compounds with a very similar structure to (+)-JQ1, TEN-010 [73] and CPI-0610 [74]; GS-5829 (i.e. structure undisclosed); BAY1238097 (i.e. structure undisclosed); ABBV-075 (i.e. structure undisclosed); and INCB054329 (i.e. structure undisclosed) [70].

Despite the fact that several compounds are in clinical trials, the development of BET inhibitors having selectivity for individual BET proteins has remained a major challenge. Major motifs associated with bromodomain-containing proteins are made up of a conserved Asn, a conserved Tyr, the WPF motif, the ZA-loop and the BC loop. These motifs are conserved in most BET family isoforms including the testis-specific transcriptional regulator, BRDT [70]. Existing BET inhibitors are critically limited by the lack of isoform selectivity especially against BRDT, which is a source of unwanted adverse effects on male fertility. Benzodiazepines are a major class of high-affinity BRD inhibitors that suffer from numerous adverse effects associated with the promiscuity of this class of inhibitors.

4. Isoform selective modulation of epigenetic targets

A majority of the aforementioned HDAC and BET inhibitors nonselectively govern the activities of distinct classes of human HDAC and BET isoforms [75]. The reason for the lack of selectivity is certainly due to the high conservation of active site residues. Furthermore, the absence of isoform selective compounds (i.e. along with its corresponding phenotypic readouts and adverse effects data) that can be used as lead compounds, hampers further development. In the absence of sufficient chemical perturbagen data, we used the SNP and mRNA expression data present in the GWAS catalog [76] and BioGPS [77], respectively, as a proof of concept for identifying isoform selective modulation of epigenetic targets with improved efficacy and toxicity profiles against existing drugs. The epigenetic drugs discussed in the above sections generally produce global epigenetic changes (i.e. acetylate both disease relevant or adverse effect-associated genes), which could lead to unintended toxicities or reduced efficacy. In the following section, we discuss recent developments in chemical biology, especially the use of chemical probes to produce sequence-specific epigenetic changes in order to alleviate some of the toxicities associated with global epigenetic changes.

4.1. *Selective isoform inhibition can achieve better efficacy and toxicity profiles*

A recent trend in epigenetic drug discovery has been the discovery and development of isoform selective inhibitors [75]. Although the absence of a critical number of epigenetic drug-target-phenotype profile leads to an over-optimistic view of the current isoform selective inhibitors, they are of substantial interest as part of efforts to translate the findings obtained from human genetics into novel therapeutic strategies. A major promise provided by genomics in drug research is the elucidation of a set of protein targets so as to achieve efficacy and reduce toxicity. Hence, in the absence of sufficient small-molecule-induced readouts, we illustrate the utility of genetic data (i.e. GWAS and BioGPS) to highlight the relevance of selective isoform inhibition for improving efficacy and toxicity profiles. A genome-wide association study (GWAS) is an approach to rapidly scanning genetic variants (markers) across the genome (~0.5 M or 1 M) of many people (>2 K) to find genetic variations that are associated with a specific disease or trait. Such studies are particularly useful in finding genetic variations that contribute to common complex diseases such as asthma, cancer, diabetes, heart disease, and mental illnesses. GWAS studies are a source of target validation in humans. An analysis of the GWAS catalog data of the HDAC class of epigenetic targets reveal that HDAC7 and HDAC9 can synergistically produce relevant immune responses (Figure 6 (a)) [78]. Similarly, BioGPS is an mRNA expression data set encompassing a panel of 79 human tissues that can be used to analyze off-targets with unfavorable expression profiles. An analysis of the BioGPS data reveals that BRDT is an anti-target that should be avoided to reduce male contraceptive-like adverse effects of pan-BRD inhibitors that are currently being investigated as anti-inflammatory agents (Figure 6(b)). These analyses highlight the importance of isoform selective modulation of epigenetic targets for efficacy enhancement while reducing adverse effects.

4.2. *Sequence-specific targeting of epigenetic switches*

Recently, Pandian et al. developed a novel class of epigenetically active small molecules called SAHA-PIPs by conjugating selective DNA-binding pyrrole-imidazole polyamides (PIPs) with the histone deacetylase inhibitor SAHA, a pan-HDACi [79]. Through microarray studies and functional analysis, they demonstrated the remarkable ability of several SAHA-PIPs to trigger transcriptional activation of exclusive clusters of genes and noncoding RNAs, rather than inducing a whole genome-wide transcriptional regulation. These compounds called SAHA-PIPs can serve as chemical biology tools and help gain insight into unresolved mechanisms and may also be able to assign functions to uncharacterized genes. Since selected disease relevant gene clusters can be precisely targeted, the design and development of cell permeable sequence-specific epigenetic switches like SAHA-PIPs represents a major advance in epigenetic drug discovery [79].

5. Drug repurposing opportunities for epigenetic drug discovery

Drug repositioning or repurposing refers to the association of known authority-approved drugs to new indications (i.e. new diseases). Before the advent of the genomic era, epigenetic drug repurposing for specific targets was performed using computational ligand- and structural-based approaches [80]. By contrast, advances in bioinformatics techniques and the availability of numerous genome-wide measurement data sets has presented a more general, automated, and unbiased approach to drug repurposing [81]. The availability of drug–drug similarity, protein activity–drug, gene expression–drug, protein–protein interactions and gene/protein–disease data sets makes it possible to statistically prioritize new epigenetic drug–disease associations [82,83]. Genomics-based approaches seem really interesting as it has been shown to afford promising results for drug repurposing [81]. Still, cheminformatics and structural bioinformatics techniques are relevant and add value to genomic approaches. In fact, there exist many success stories for epigenetic drug repurposing [83]. Méndez-Lucio et al. [85] utilized cheminformatics analysis for identifying olsalazine (i.e. a drug that was previously approved by the FDA as an anti-inflammatory agent) as a DNA hypomethylating agent. Using a known hypomethylating agent (NSC14778) as a reference molecule, a similarity search approach was conducted by comparing the structure of the reference with the structures of 1582 FDA-approved compounds from the DrugBank database. The analysis led to the identification of Olsalazine (i.e. due to affording a high Tanimoto Combo score of 1.032 with NSC14778) as a good candidate for DNMT1 inhibition. The Tanimoto Combo score, as implemented in the Rapid Overlay of Chemical Structures (ROCS) package obtained from the OpenEye scientific software, is the sum of the Shape Tanimoto and the Chemical Functionalities Tanimoto (color Tanimoto). The range of the Tanimoto Combo score varies from 0 to 2 with a score exceeding 1.4 representing a high degree of similarity for a pair of compounds. An *in vivo* study conducted on HeLa cells adapted to report gene expression visually via the green fluorescent protein was used to experimentally establish the DNA hypomethylation ability of olsalazine by virtue of its interaction with DNMT1. The binding mode of olsalazine against DNMT1 and DNMT3b was further elucidated using a detailed docking study. In another interesting study by *de La*, *Cruz-Hernandez et al.* [86] also made use of cheminformatics for epigenetic drug repurposing of 3-deazaneplanocin A, a known inhibitor of SAM-dependent methyltransferase that targets the degradation of EZH2 and leads to apoptosis in various malignancies, was used as a reference compound for a chemical similarity search against FDA-approved and experimental drugs. The cheminformatic analysis identified ribavirin, a nucleoside drug approved for the treatment of hepatitis C virus infection, as having high structural similarity with 3-deazaneplanocin A. Experimental assays in various cell lines revealed that ribavirin could inhibit the expression of EZH2 and two other cancer-associated epigenetic targets.

Structural bioinformatics approaches also have great potential for drug repurposing via disease-associated epigenetic targets. Among several success stories of structure-based drug design approaches, the proteochemometric approach utilized by Dakshanamurthy et al. [87] was the most interesting in which they developed a new computational method called 'TMFS' that consisted of a docking score, ligand and receptor shape/topology descriptor scores and ligand–receptor contact point scores to predict 'molecules of best fit' and filter out most false-positive interactions. Using this method, they reprofiled 3671 FDA-approved/experimental drugs against 2335 human protein targets with a good prediction accuracy of 91% for the majority of drugs. Amongst the several novel associations, they experimentally validated that the anti-hookworm medication mebendazole could inhibit VEGFR2 and angiogenesis activity. Furthermore, they also found that the anti-inflammatory drug celcoxib and its analog DMC could bind CDH11 (i.e. a biomolecule that is very important in rheumatoid arthritis and poor prognosis malignancies, for which no targeted therapies currently exist). The advantages of the proteochemometric approach (i.e. analysis concomitantly involving both protein and chemical structural features) over a traditional cheminformatics approaches are multifold such as, the possibility to gain detailed insight into binding modes in addition to the discovery of novel drug–target associations.

The integration of genomics-assisted approaches to drug repurposing along the lines described by the above methods can address some of their drawbacks such as the limited applicability domain and the non-immediate disease relevance of ligand-based cheminformatics approaches and target-based structural bioinformatics approaches, respectively. The genomics-assisted approach to drug repurposing has some success stories in epigenetic drug discovery [88,89]. Drug repurposing through the genomics approach generally involves either correlating drug–drug gene expression profiles or drug–disease expressions using a range of statistical procedures to find useful patterns. An example is using the Kolmogorov–Smirnov statistical test as implemented in the connectivity map (CmAP), which is a searchable chemogenomic database containing thousands of gene-expression signatures of various cultured cancer cells as exposed to a large collection of small-molecule compounds, to find a pattern indicating a possible repurposing [90]. The database and statistical procedure represent a useful tool for the discovery of hitherto unexplored connections amongst small molecules with diseases in terms of Anatomical Therapeutic Chemical (ATC) codes. By comparing expression signatures, CmAP serves as a proxy to search for novel indications of all surveyed compounds. The correlation between a given gene expression profile and the various ranked gene expression profiles in the CMap is presented as the signed enrichment score. The signed enrichment score varies from +1 to −1.

Table 3. Known ATC associations of some epigenetic drugs mentioned in Figure 4.

DB ID	Name	Class	ATC code	ATC code description
DB00721	Procaine	DNMTi	C05AD05	Cardiovascular System, Vasoprotectives, Local Anesthetics
DB00721	Procaine	DNMTi	N01BA02, N01BA52	Nervous System, Anesthetics, Esters Of Aminobenzoic Acid
DB00721	Procaine	DNMTi	S01HA05	Sensory Organs, Ophthalmologicals, Local Anesthetics
DB00928	Azacitidine	DNMTi	L01BC07	Antineoplastic And Immunomodulating Agents, Antineoplastic Agents, Pyrimidine Analogues
DB01035	Procainamide	DNMTi	C01BA02	Cardiovascular System, Cardiac Therapy, Antiarrhythmics, Class Ia
DB01262	Decitabine	DNMTi	L01BC08	Antineoplastic And Immunomodulating Agents, Antineoplastic Agents, Pyrimidine Analogues
DB01275	Hydralazine	DNMTi	C02DB02	Cardiovascular System, Antihypertensives, Hydrazinophthalazine Derivatives
DB01275	Hydralazine	DNMTi	C02LG02	Cardiovascular System, Antihypertensives, Antihypertensives And Diuretics In Combination
DB02546	Vorinostat	HDACi	L01XX38	Antineoplastic And Immunomodulating Agents, Other Antineoplastic Agents
DB05015	Belinostat	HDACi	L01XX49	Antineoplastic And Immunomodulating Agents, Other Antineoplastic Agents
DB06176	Romidepsin	HDACi	L01XX39	Antineoplastic And Immunomodulating Agents, Other Antineoplastic Agents
DB06603	Panobinostat	HDACi	L01XX42	Antineoplastic And Immunomodulating Agents, Other Antineoplastic Agents
DB06819	Phenylbutyrate	HDACi	A16AX03	Alimentary Tract And Metabolism, Other Alimentary Tract And Metabolism Products, Various Alimentary Tract And Metabolism Products
DB00752	Tranylcypromine	HDMi	N06AF04	Nervous System, Psychoanaleptics, MAO Inhibitors, Non-Selective
DB00780	Phenelzine	HDMi	N06AF03	Nervous System, Psychoanaleptics, MAO Inhibitors, Non-Selective
DB01626	Pargyline	HDMi	C02KC01	Cardiovascular System, Antihypertensives, MAO Inhibitors
DB01626	Pargyline	HDMi	C02LL01	Cardiovascular System, Antihypertensives, MAO Inhibitors And Diuretics
DB00250	Allantodapsone	HMTi	D10AX05	Dermatologicals, Other Anti-Acne Preparations For Topical Use
DB00250	Allantodapsone	HMTi	J04BA02	Antiinfectives For Systemic Use, Antimycobacterials, Drugs For Treatment Of Lepra

ATC is the WHO recommended Anatomic Therapeutic Chemical classification system for drugs. As demonstrated in various sections of this article and the connectivity map analysis presented in Table 4, there are numerous indications that epigenetic drugs' expression signatures can be linked with various disease gene expression signatures. For instance, while known associations presented in this table links HDACi with LO1 (antineoplastic drugs), the connectivity map analysis presented in Table 4 associates a functionally similar second generation HDACi with numerous other ATCs. The DrugBank data set was downloaded as an XML file and parsed using 'xmlstarlet' package to extract the DB_ID and ATC codes to generate the table. Reproduced with permission from [93].

Table 4. Correlation of diseases and drugs on the basis of gene signature associations with 'ST7612AA1,' a novel second-generation oral HDAC inhibitor.

Name of drug, cell line or ATC code	Signed enrichment score
Vorinostat – MCF7	0.985
Trichostatin A – PC3	0.959
Trichostatin A – HL60	0.952
Trichostatin A – MCF7	0.931
Pioglitazone – PC3	−0.913
Sirolimus – PC3	0.911
Wortmannin – MCF7	0.821
Tanespimycin – HL60	0.805
Trifluoperazine – MCF7	0.796
LY-294002 – PC3	0.727
N05AB (Antipsychotic and Anxiolytic drugs)	0.647
N03AG (Fatty acid derivatives as Antiepileptic drugs)	0.47
L04AA (Selective immunosuppressants)	0.426
R06AX (Other antihistamines for systemic use)	0.411
A07DA (Intestinal anti-infectives)	0.854
N05AC (Hypnotics and sedatives drugs)	0.45
C01AA (Cardiac glycosides list)	0.636
L01CB (Plant alkaloids and other natural products as antineoplastic drugs)	0.831
B02AA (Antifibrinolytics)	−0.616
N05AG (Antipsychotic drugs)	0.591

ST7612AA1's gene expression dataset was downloaded from the NCBI GEO using id 'GSE62460'[94]. Connectivity map analysis on the drug-induced gene signatures from CMap database [90] was used to identify drugs and their ATCs whose expressions correlate with the top 250 up- and down-regulated genes of 'ST7612AA1.' The differentially expressed up- and down-regulated genes were extracted after Limma [95] analysis of 'GSE62460.' These types of *in silico* drug repurposing studies of epigenetic drugs can propose novel disease indications for experimental verification. Signed enrichment score were computed using the Kolmogorov–Smirnov (KS) test.

Using the CMap analysis, Zerbini et al. [91] presented a case study for the identification of compounds whose gene expression signatures were negatively enriched with the gene signatures of metastatic clear cell renal carcinoma (ccRC). The consensus top-scoring 8 drugs (those with a negative enrichment correlation between −0.7 and −1.0 in more than 50% of the patients) were selected to be tested *in vitro* and *in vivo*. Five of these drugs exhibited a strongly incremental rate of apoptosis in cancer cells; however, they did not affect the survival of normal cells. They also demonstrated that the status of VHS gene (whose mutation is known for causing ccRC) was strongly associated with the response. The best responses were observed in cells deficient in VHC. Furthermore, amitriptyline was seen to induce multiple myeloma apoptosis through the inhibition of cyclin D2 expression and also via repression of HDAC and consequently, its activity.

Using microarray technology, Claerhout et al. [88] generated a gene expression profile of human gastric cancer-specific genes from human gastric cancer tissue samples. They used this profile for CMap analysis as to identify candidate therapeutic compounds for gastric cancer. The histone deacetylase inhibitor vorinostat, emerged as the lead compound and thus a potential therapeutic drug for gastric cancer. Vorinostat has been experimentally shown to induce both apoptosis and autophagy in gastric cancer cell lines and it was further suggested that combination of vorinostat with autophagy inhibitors may be therapeutically synergistic. Moreover, gene expression analysis of gastric cancer identified a collection of genes (e.g. *ITGB5, TYMS, MYB, APOC1, CBX5, PLA2G2A*, and *KIF20A*) whose expressions were elevated in gastric tumor tissues and downregulated by more than twofold upon treatment with vorinostat in gastric cancer cell lines. In contrast, *SCGB2A1, TCN1, CFD, APLP1*, and *NQO1* manifested a reversed pattern.

Oprea and Overington [92] suggested a robust classification scheme, DREL, that can be used to evaluate drug repositioning projects according to the level of scientific evidence. Based on this scheme, the study by Zerbini et al. [91] and Méndez-Lucio et al. [85] can be classified as DREL-2 (i.e. animal studies with hypothetical relevance in man) whereas the studies by Claerhout et al. [88] and De la Cruz-Hernandez et al. [86] can be classified as DREL-1 (i.e. representing experimental validation in the form of *in vitro* studies with limited value for predicting *in vivo*/human situation).

To illustrate the utility of the CmAP approach for drug repurposing, we present a short case study involving the novel second-generation HDAC inhibitor ST7612AA1. The present study can be considered as an intermediate between DREL-0 and DREL-1 since it presents limited experimental validation in the form of *in vitro* drug-induced gene expression analysis and no phenotypic assays. As illustrated in Table 3, many epigenetic drugs map to multiple ATC classes, which classifies drugs according to the organ or system on which they act and their associated therapeutic, pharmacological and chemical properties. The ATC classification is hence a useful indicator of the drug's phenotype or its disease relevance. As illustrated in Table 3, the selected HDACi are primarily classified by ATC as Antineoplastic and Immunomodulating Agents (L01XX) and are presently widely used in anticancer therapies. However, our CmAP analysis correlating the ST7612AA1-induced gene expression signature in TMD8 and DOHH2 lymphoblastoma cell lines with various drug-induced gene expression values in the CmAP database reveals that HDACi could be repurposed for numerous other indications as anti-psychotics (N05AB with an enrichment score of 0.647), anti-infectives (A07DA with an enrichment score of 0.854) and cardiovascular agents (C01AA with an enrichment score of 0.636). The results presented in Table 4 corroborate the findings of numerous other studies including findings in clinical trials. In addition, the CMap analysis presented in Table 4 also shows the correlation of ST7612AA1-induced gene expression profiles with the expression signatures of other HDACi's such as vorinostat (enrichment score of 0.985) and trichostatin A (enrichment score of 0.959). The analysis reveals a positive correlation of ST7612AA1-induced gene expression profiles with anticancer compounds like sirolimus (also called rapamycin with an enrichment score of 0.911), wortmannin (enrichment score of 0.821), tanespimycin (enrichment score of 0.805), LY-294002 (enrichment score of 0.727), and antipsychotic compounds like trifluoperazine (enrichment score of 0.796). Interestingly, ST7612AA1-induced gene expression profile negatively correlates with pioglitazone (enrichment score of −0.913), which is a PPAR-gamma agonist that have been associated with a higher risk of cardiac events.

6. Conclusion

Epigenetics modulate the regulation of gene expression for the maintenance of homeostasis via the concerted actions of several epigenetic modifiers. The physiological functions of these modifiers are altered by external factors, which may lead to aberrant gene expression and diseases. Inherited (e.g. gender and racial), environmental (i.e. exposure to pollutants and chemicals, stress, etc.), and social (i.e. income, residence, occupation, education, culture, and malnutrition) factors are known to influence epigenetic regulations. Epigenomics has gained notable attention as a field that could provide answers on how external stimulus (e.g. environment, nutrition, and behavior) governs the development and progression of multifactorial diseases, as well as providing an explanation on the differential susceptibility to diseases amongst individuals. Moreover, epigenetic alterations have been implicated in a wide spectrum of diseases. Great progress has been made on identifying disease-relevant epigenetic targets, which has contributed to a better understanding of the pathogenesis and management of many complex diseases (e.g. metabolic and cardiovascular diseases, autoimmune diseases, psychological disorders, neurodegenerative and neurodevelopmental diseases, and cancers). This has involved the use of many advanced genomics, epigenomics, bioinformatics, and cheminformatics technologies, all of which has facilitated the discovery of several novel classes of epigenetic modifiers for therapeutic applications. As an example, a widely cited study by Jones and Baylin [96] reviewed advances in understanding how epigenetic alterations participate in the earliest stages of neoplasia, including stem/precursor cell contributions, and discuss the growing implications of these advances for strategies to control cancer. Naturally derived compounds are in the spotlight as an excellent source of active scaffolds for epigenetic drugs, while drug repositioning/repurposing demonstrates a powerful strategy for the discovery of novel indications for existing FDA-approved drugs. Discovery of novel epigenetic drugs may pave way for fulfilling several unsolved problems in multifactorial diseases. The field is a highly challenging one indeed. Of particular note is the distinct characteristics of the epigenome, which include long-lasting memory, transgenerational inheritance and environmental adaptations. Awareness of maternal and early life exposures to predisposing factors may decrease the risk of developing adult-onset diseases and developmental disorders. In addition, understanding the environmental adaptations of the epigenome renders adjustment of lifestyle and nutritional behavior as a potential path for disease prevention and health promotion.

7. Expert opinion

Extensive chemical biology and genomic studies have revealed druggable and clinically relevant epigenetic targets (e.g. DNMTs, HDACs, HATs, SIRTs, HDMs, BRDs, and PMTs). The clinical success of epigenetic modifiers has been demonstrated by the many drugs approved by the FDA. Therapeutic potential has been expressed most clearly in oncology where almost all types of epigenetic modifiers may have impact, whereas for cardiovascular and neurological disorders only a few modifiers have shown utility (i.e. BRDi, DNMTi, and SIRT modulators for the former; while HDMi for the latter).

The development of DNMTi has been primarily directed toward cytidine analogs. However, the CpG dinucleotide analogs (e.g. SGI-110) show promise as DNMTi owing to their superiority in resisting cytidine deaminase (i.e. a cytidine inactivating enzyme). Attention has also been directed toward compounds interfering with protein–protein interactions and compounds exhibiting DNMT inhibition via other mechanisms of action such as allosteric inhibitors, SAM mimicking compounds and DNA competitive substrates. For HDAC inhibitors, the discovery of new inhibitors has mostly been focused on hydroxamic acids and benzamides. However, the clinical outcome of these compounds are still uncertain and future direction could be emphasized toward the discovery of HDACi with novel mechanisms of action. Moreover, the development of novel hybrid molecules targeting HDAC inhibition and other oncogenic/inflammatory pathways has provided interesting results, especially those bearing adamantane moieties. Furthermore, the discovery of SIRT1 activators and HDMi are in the spotlight for neurodegenerative diseases. The development of SIRT1 activators has been focused toward naturally occurring phenol derivatives, especially resveratrols. Scaffolds possessing an inhibitory effect toward monoamine oxidases such as pargyline, phenelzine, and tranylcypromine have been suggested to be within the potential chemical spaces for the discovery of HDMi. The development of PMTi and HATi is currently limited to the area of oncology and is still in their infancy. Further research regarding these two types of epigenetic modifiers may extend the area of therapeutic epigenetics. Selectivity is of high concern regarding the development of BRDi as most of the bromodomain-containing proteins share similar structures but possess distinct structural differences and functions in biological pathways. Thus, the development of selective BRDi for reduced side effects is a challenging opportunity. In addition, advanced approaches employing availability of genomic data derived from GWAS and BioGPS expression data sets are underlined for the discovery of isoform selective inhibitors with improved efficacy and side effect profiles. The development of an epigenetically active hybrid molecule as a chemical biology tool to unravel insights, mechanisms, and functionally relevant genes of complex diseases is also marked as an area with great potential.

Natural compounds cannot be overlooked as attractive sources of novel scaffolds for the development of epigenetic modulators. For example, naturally derived polyphenols (e.g. EGCG, curcumin, and caffeic acids), flavonoids (e.g. genistein and quercetin), quinones (e.g. hypericin and laccaic acid), lycopene, and boswellic acid have been reported as DNMTi. Moreover, naturally occurring phenol derivatives are in the spotlight as SIRT activators especially resveratrol derivatives, which are suggested to act as allosteric activators. However, the development of these natural compounds may also pose similar problems as those observed in clinically useful drugs (e.g. poor absorption, metabolic stability, and pharmacokinetics).

Although, epigenetic drug discovery is increasingly directed toward selective epigenetic modifiers (i.e. inhibitors or modulators), the discovery strategies of promiscuous or pan-modulators are currently the most viable. Greater phenotypic responses of pan-modulators as observed by the broad range of factors are implicated via the etiology, pathogenesis, and progression of disease and the expensive screening techniques used to discover isoform selective modulators, renders isoform selective discovery programs both medically and financially inefficient [97]. In the present review, 'class-selective' modulators are not necessarily those which modulate a single target but instead modulate a subset of targets to produce the required phenotypic responses.

Drug repurposing is currently considered an attractive strategy for the discovery of new indications of the existing drug space. Specifically, it reduces the need for the costly and time-consuming preclinical pharmacology, formulation and toxicity testing, which are otherwise required for clinical trial approval. Computational approaches that play a crucial role in early stages of drug discovery have formed the core technology in drug repurposing. *In silico* analysis (e.g. cheminformatics, structural bioinformatics, and genomics) of relevant data sets (e.g. drug libraries, gene expression–disease, protein–drug, and protein–protein interactions) have proven capable of identifying novel epigenetic drug-disease associations.

The current interest in personalized medicine is largely due to recent insights into genomics and epigenomics. Epigenetic factors are responsible for phenotypic plasticity and are increasingly associated with individual specific disease etiologies and drug responses, and can be revealed by mining genomic and epigenomic data of individual patients. Hence, in an era of lifestyle-induced diseases where a complex myriad of individual and environmental factors exists that constantly modifies the individual's epigenetic landscape via external stimuli, there is an enormous potential for prevention and therapy. On top of this, the fields of nutritional and stress epigenomics that we have not covered here are on the rising trend for personalized diagnosis, prevention and control of cancer, cardiovascular, neurological and aging diseases. All of these points toward a great future for novel drugs directed to the many complex multifactorial diseases of humans, which are still often poorly understood and difficult to treat. In this regard, the redesign of routine lifestyle behaviors (i.e. involving alteration to nutrition, exercise and stress management) along with advanced studies in related areas (e.g. nutraceuticals and complementary medicine) should not be overlooked as key factors toward achieving good health and well-being.

Funding

C Phanus-Umporn is funded via a Royal Golden Jubilee Ph.D. Scholarship (No. PHD/0050/2558) from the Thailand Research Fund while R Pratiwi is supported by a Ph.D. Scholarship from the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia. This research is also supported by Research Career Development Grant (No. RSA5780031) from the Thailand Research Fund to C Nantasenamat. Partial support via the Swedish Research Links program (No. C0610701) from the Swedish Research Council to JES Wikberg and C Nantasenamat is also acknowledged.

Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties

ORCID

Veda Prachayasittikul http://orcid.org/0000-0001-6338-3721
Philip Prathipati http://orcid.org/0000-0002-7889-779X
Reny Pratiwi http://orcid.org/0000-0003-0210-1086
Chuleeporn Phanus-umporn http://orcid.org/0000-0001-5439-8462
Aijaz Ahmad Malik http://orcid.org/0000-0001-5132-1574
Nalini Schaduangrat http://orcid.org/0000-0002-0842-8277
Kanokwan Seenprachawong http://orcid.org/0000-0001-7138-7108

Prapimpun Wongchitrat http://orcid.org/0000-0001-7009-3028
Aungkura Supokawej http://orcid.org/0000-0002-3979-873X
Virapong Prachayasittikul http://orcid.org/0000-0001-7942-1083
Jarl E. S. Wikberg http://orcid.org/0000-0003-1916-3013
Chanin Nantasenamat http://orcid.org/0000-0003-1040-663X

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Wu C, Morris JR. Genes, genetics, and epigenetics: a correspondence. Science. 2001;293(5532):1103–1105.
2. Dupont C, Armant DR, Brenner CA. Epigenetics: definition, mechanisms and clinical perspective. Semin Reprod Med. 2009;27(5):351–357.
3. Allfrey V, Faulkner R, Mirsky A. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proc Natl Acad Sci USA. 1964;51:786–794.
4. Waddington CH. The epigenotype. 1942. Int J Epidemiol. 2012;41 (1):10–13.
   •• The coining of the breakthrough concept of epigenetics is introduced in this paper.
5. Noble D. Conrad Waddington and the origin of epigenetics. J Exp Biol. 2015;218(Pt 6):816–818.
6. Gold M, Hurwitz J, Anders M. The enzymatic methylation of RNA and DNA. I. Biochem Biophys Res Commun. 1963;11(2):107–114.
7. Hamm CA, Costa FF. The impact of epigenomics on future drug design and new therapies. Drug Discov Today. 2011;16(13–14):626–635.
   •• This paper provides a good synopsis on epigenetic drug design and its implications in treatments.
8. Gabory A, Attig L, Junien C. Developmental programming and epigenetics. Am J Clin Nutr. 2011;94(6 Suppl):1943S–1952S.
9. Bhutani N, Burns DM, Blau HM. DNA demethylation dynamics. Cell. 2011;146(6):866–872.
10. Davey CA, Sargent DF, Luger K, et al. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. J Mol Biol. 2002;319(5):1097–1113.
11. Franchini D-M, Schmitz K-M, Petersen-Mahrt SK. 5-Methylcytosine DNA demethylation: more than losing a methyl group. Annu Rev Genet. 2012;46:419–441.
12. Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes Dev. 2014;28(8):812–828.
13. Mariño-Ramírez L, Kann MG, Shoemaker BA, et al. Histone structure and nucleosome stability. Expert Rev Proteomics. 2005;2(5):719–729.
14. Meng F, Wang C, Wan W, et al. Discovery and development of small molecules targeting epigenetic enzymes with computational methods. In: Medina-Franco JL, Ed. Epi-Informatics: discovery and development of small molecule epigenetic drugs and probes. London: Elsevier Inc; 2016. p. 75–112.
15. Falkenberg KJ, Johnstone RW. Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. Nat Rev Drug Discov. 2014;13(9):673–691.
16. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21(3):381–395.
17. Fu X-D. Non-coding RNA: a new frontier in regulatory biology. Natl Sci Rev. 2014;1(2):190–204.
18. Mirbahai L, Chipman JK. Epigenetic memory of environmental organisms: a reflection of lifetime stressor exposures. Mutat Res Genet Toxicol Environ Mutagen. 2014;764-765:10–17.
19. Patkin EL, Sofronov GA. Population epigenetics, ecotoxicology, and human diseases. Russ J Genet Appl Res. 2013;3(5):338–351.
20. Prins GS, Birch L, Tang W-Y, et al. Developmental estrogen exposures predispose to prostate carcinogenesis with aging. Reprod Toxicol. 2007;23(3):374–382.
21. Cazaly E, Charlesworth J, Dickinson JL, et al. Genetic determinants of epigenetic patterns: providing insight into disease. Mol Med. 2015;21:400–409.
22. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945–D954.
23. Singh Nanda J, Kumar R, Raghava GPS. dbEM: a database of epigenetic modifiers curated from cancerous and normal genomes. Sci Rep. 2016;6:19340.
24. Huang Z, Jiang H, Liu X, et al. HEMD: an integrated tool of human epigenetic enzymes and chemical modulators for therapeutics. Plos ONE. 2012;7(6):e39917.
25. Nantasenamat C, Prachayasittikul V. Maximizing computational tools for successful drug discovery. Expert Opin Drug Discov. 2015;10(4):321–329.
   • This editorial provides a bird's eye view on selecting appropriate computational tools from the vast collection for tackling drug discovery projects.
26. Lapinsh M, Prusis P, Gutcaits A, et al. Development of proteochemometrics: a novel technology for the analysis of drug-receptor interactions. Biochim Biophys Acta. 2001;1525(1–2):180–190.
   • The concept of proteochemometrics is introduced in this paper to expand the one-target approach of QSAR to a multitargeted one whereby several target proteins and several compounds are considered in a unified model.
27. Cortés-Ciriano I, Ain QU, Subramanian V, et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *Med Chem Commun*. 2015;6(1):24–50.
   • This review article provides comprehensive coverage on the field of proteochemometrics along with examples of its usages.
28. Robertson KD. DNA methylation, methyltransferases, and cancer. Oncogene. 2001;20(24):3139–3155.
29. Cheng X, Roberts RJ. AdoMet-dependent methylation, DNA methyltransferases and base flipping. Nucleic Acids Res. 2001;29 (18):3784–3795.
30. Fellinger K, Rothbauer U, Felle M, et al. Dimerization of DNA methyltransferase 1 is mediated by its regulatory domain. J Cell Biochem. 2009;106(4):521–528.
31. Rai K, Chidester S, Zavala CV, et al. Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. Genes Dev. 2007;21(3):261–266.
32. Schaefer M, Pollex T, Hanna K, et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes Dev. 2010;24(15):1590–1595.
33. Jia D, Jurkowska RZ, Zhang X, et al. Structure of Dnmt3a bound to Dnmt3L suggests a model for *de novo* DNA methylation. Nature. 2007;449(7159):248–251.
34. Guianvarc'h D, Arimondo PB. Challenges in developing novel DNA methyltransferase inhibitors for cancer therapy. Future Medicinal Chemistry. 2014;6(11):1237–1240.
35. Copeland RA, Olhava EJ, Scott MP. Targeting epigenetic enzymes for drug discovery. Curr Opin Chem Biol. 2010;14(4):505–510.
   • This review article provides a succinct coverage on important epigenetic enzymes and the role of chemical biology approaches for the discovery of small-molecule modulators.
36. Erdmann A, Halby L, Fahy J, et al. Targeting DNA methylation with small molecules: what's next? J Med Chem. 2015;58(6):2569–2583.
37. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40 (Database issue):D1100–7.
38. Yoo CB, Jeong S, Egger G, et al. Delivery of 5-aza-2′-deoxycytidine to cells using oligodeoxynucleotides. Cancer Res. 2007;67 (13):6400–6408.
39. Griffiths EA, Choy G, Redkar S, et al. SGI-110: DNA methyltransferase inhibitor oncolytic. Drugs Future. 2013;38(8):535–543.
40. Cohen I, Poręba E, Kamieniarz K, et al. Histone modifiers in cancer: friends or foes? Genes Cancer. 2011;2(6):631–647.
41. Rodríguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. Nat Med. 2011;17(3):330–339.
42. Balasubramanyam K, Varier RA, Altaf M, et al. Curcumin, a novel p300/CREB-binding protein-specific inhibitor of acetyltransferase, represses the acetylation of histone/nonhistone proteins and histone acetyltransferase-dependent chromatin transcription. J Biol Chem. 2004;279(49):51163–51171.

43. Mukhopadhyay A, Banerjee S, Stafford LJ, et al. Curcumin-induced suppression of cell proliferation correlates with down-regulation of cyclin D1 expression and CDK4-mediated retinoblastoma protein phosphorylation. Oncogene. 2002;21(57):8852–8861.
44. Delcuve GP, Khan DH, Davie JR. Roles of histone deacetylases in epigenetic regulation: emerging paradigms from studies with inhibitors. Clin Epigenetics. 2012;4(1):5.
45. Thangapandian S, John S, Sakkiah S, et al. Ligand and structure based pharmacophore modeling to facilitate novel histone deacetylase 8 inhibitor design. Eur J Med Chem. 2010;45(10):4409–4417.
46. Ononye SN, Van Heyst M, Falcone EM, et al. Toward isozyme-selective inhibitors of histone deacetylase as therapeutic agents for the treatment of cancer. Pharm Pat Anal. 2012;1(2):207–221.
47. Nebbioso A, Carafa V, Benedetti R, et al. Trials with "epigenetic" drugs: an update. Mol Oncol. 2012;6(6):657–682.
48. West AC, Johnstone RW. New and emerging HDAC inhibitors for cancer treatment. J Clin Invest. 2014;124(1):30–39.
49. Mottamal M, Zheng S, Huang TL, et al. Histone deacetylase inhibitors in clinical studies as templates for new anticancer agents. Molecules. 2015;20(3):3898–3941.
50. Wang D. Computational studies on the histone deacetylases and the design of selective histone deacetylase inhibitors. Curr Top Med Chem. 2009;9(3):241–256.

•• **A chemical phylogenetic approach is presented for analyzing a structurally diversed panel of HDACi from which mechanistic insights on the functional classification of the HDAC family with these tool compounds were obtained. The study also identified the first pan-HDACi that may be useful as a chemical probe. Findings from this study may be used as guidelines for future design of HDACi.**

51. Bradner JE, West N, Grachan ML, et al. Chemical phylogenetics of histone deacetylases. Nat Chem Biol. 2010;6(3):238–243.
52. Chen K, Xu L, Wiest O. Computational exploration of zinc binding groups for HDAC inhibition. J Org Chem. 2013;78(10):5051–5055.
53. Mahajan SS, Leko V, Simon JA, et al. Sirtuin modulators. Handb Exp Pharmacol. 2011;206:241–255.
54. Walport LJ, Hopkinson RJ, Chowdhury R, et al. Arginine demethylation is catalysed by a subset of JmjC histone lysine demethylases. Nat Commun. 2016;7:11974.
55. https://www.ncbi.nlm.nih.gov/pubmed/15971202
56. Kim YZ. Protein methylation and demethylation in cancer. Int J Neurol Res. 2015;1(3):129–140.
57. Thinnes CC, England KS, Kawamura A, et al. Targeting histone lysine demethylases - progress, challenges, and the future. Biochim Biophys Acta. 2014;1839(12):1416–1432.
58. Copeland RA, Moyer MP, Richon VM. Targeting genetic alterations in protein methyltransferases for personalized cancer therapeutics. Oncogene. 2013;32(8):939–946.
59. Boriack-Sjodin PA, Swinger KK. Protein methyltransferases: a distinct, diverse, and dynamic family of enzymes. Biochemistry. 2016;55(11):1557–1569.
60. Schapira M, Arrowsmith CH. Methyltransferase inhibitors for modulation of the epigenome and beyond. Curr Opin Chem Biol. 2016;33:81–87.
61. Simó-Riudalbas L, Esteller M. Targeting the histone orthography of cancer: drugs for writers, erasers and readers. Br J Pharmacol. 2015;172(11):2716–2732.
62. Gelato KA, Shaikhibrahim Z, Ocker M, et al. Targeting epigenetic regulators for cancer therapy: modulation of bromodomain proteins, methyltransferases, demethylases, and microRNAs. Expert Opin Ther Targets. 2016;20(7):783–799.
63. Copeland RA. Protein methyltransferase inhibitors as personalized cancer therapeutics. Drug Discov Today: Ther Strateg. 2012;9(2–3):e83–e90.
64. Spannhoff A, Machmur R, Heinke R, et al. A novel arginine methyltransferase inhibitor with cellular activity. Bioorg Med Chem Lett. 2007;17(15):4150–4153.
65. Selvi BR, Batta K, Kishore AH, et al. Identification of a novel inhibitor of coactivator-associated arginine methyltransferase 1 (CARM1)-mediated methylation of histone H3 Arg-17. J Biol Chem. 2010;285(10):7143–7152.
66. Chan-Penebre E, Kuplast KG, Majer CR, et al. A selective inhibitor of PRMT5 with in vivo and in vitro potency in MCL models. Nat Chem Biol. 2015;11(6):432–437.
67. Cheng D, Yadav N, King RW, et al. Small molecule regulators of protein arginine methyltransferases. J Biol Chem. 2004;279(23):23892–23899.
68. El Messaoudi S, Fabbrizio E, Rodriguez C, et al. Coactivator-associated arginine methyltransferase 1 (CARM1) is a positive regulator of the Cyclin E1 gene. Proc Natl Acad Sci USA. 2006;103(36):13351–13356.
69. Mai A, Cheng D, Bedford MT, et al. Epigenetic multiple ligands: mixed histone/protein methyltransferase, acetyltransferase, and class III deacetylase (sirtuin) inhibitors. J Med Chem. 2008;51(7):2279–2290.
70. Galdeano C, Ciulli A. Selectivity on-target of bromodomain chemical probes by structure-guided medicinal chemistry and chemical biology. Future Med Chem. 2016;8(13):1655–1680.
71. Nicholls SJ, Puri R, Wolski K, et al. Effect of the BET protein inhibitor, RVX-208, on progression of coronary atherosclerosis: results of the phase 2b, randomized, double-blind, multicenter, ASSURE trial. Am J Cardiovasc Drugs. 2016;16(1):55–65.
72. Shu S, Polyak K. BET bromodomain proteins as cancer therapeutic targets. Cold Spring Harb Symp Quant Biol. 2017;81. doi:10.1101/sqb.2016.81.030908.
73. Romero FA, Taylor AM, Crawford TD, et al. Disrupting acetyl-lysine recognition: progress in the development of bromodomain inhibitors. J Med Chem. 2016;59(4):1271–1298.
74. Albrecht BK, Gehling VS, Hewitt MC, et al. Identification of a Benzoisoxazoloazepine Inhibitor (CPI-0610) of the Bromodomain and Extra-Terminal (BET) family as a candidate for human clinical trials. J Med Chem. 2016;59(4):1330–1339.
75. Bieliauskas AV, Pflum MKH. Isoform-selective histone deacetylase inhibitors. Chem Soc Rev. 2008;37(7):1402–1413.
76. Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin. 2015;8:57.
77. Wu C, Jin X, Tsueng G, et al. BioGPS: building your own mash-up of gene annotations and expression profiles. Nucleic Acids Res. 2016;44(D1):D313–6.
78. Reilly CM, Regna N, Mishra N. HDAC inhibition in lupus models. Mol Med. 2011;17(5–6):417–425.
79. Pandian GN, Taniguchi J, Junetha S, et al. Distinct DNA-based epigenetic switches trigger transcriptional activation of silent genes in human dermal fibroblasts. Sci Rep. 2014;4:3843.
80. Katsila T, Spyroulias GA, Patrinos GP, et al. Computational approaches in target identification and drug discovery. Comput Struct Biotechnol J. 2016;14:177–184.
81. Lussier YA, Chen JL. The emergence of genome-based drug repositioning. Sci Transl Med. 2011;3(96):96ps35.
82. Iwata H, Sawada R, Mizutani S, et al. Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. J Chem Inf Model. 2015;55(2):446–459.
83. Prathipati P, Mizuguchi K. Systems biology approaches to a rational drug discovery paradigm. Curr Top Med Chem. 2016;16(9):1009–1025.
84. Naveja JJ, Dueñas-González A, Medina-Franco JL. Drug repurposing for epigenetic targets guided by computational methods. In: Medina-Franco JL, Ed.. Epi-Informatics. London, UK: Elsevier Inc; 2016. p. 327–357.
85. Méndez-Lucio O, Tran J, Medina-Franco JL, et al. Toward drug repurposing in epigenetics: olsalazine as a hypomethylating compound active in a cellular context. ChemMedChem. 2014;9(3):560–565.
86. De La Cruz-Hernandez E, Medina-Franco JL, Trujillo J, et al. Ribavirin as a tri-targeted antitumor repositioned drug. Oncol Rep. 2015;33(5):2384–2392.

87. Dakshanamurthy S, Issa NT, Assefnia S, et al. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem. 2012;55(15):6832–6848.
88. Claerhout S, Lim JY, Choi W, et al. Gene expression signature analysis identifies vorinostat as a candidate therapy for gastric cancer. PLos ONE. 2011;6(9):e24662.

•• **An innovative utilization of proteochemometrics for repurposing of existing approved drugs to new indications.**

89. Wen Z, Wang Z, Wang S, et al. Discovery of molecular mechanisms of traditional Chinese medicinal formula Si-Wu-Tang using gene expression microarray and connectivity map. Plos ONE. 2011;6(3):e18278.
90. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–1935.
91. Zerbini LF, Bhasin MK, De Vasconcellos JF, et al. Computational repositioning and preclinical validation of pentamidine for renal cell cancer. Mol Cancer Ther. 2014;13(7):1929–1941.
92. Oprea TI, Overington JP. Computational and practical aspects of drug repositioning. Assay Drug Dev Technol. 2015;13(6):299–306.
93. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(Database issue):D668–672.
94. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
95. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
96. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128 (4):683–692.
97. Mencher SK, Wang LG. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). BMC Clin Pharmacol. 2005;5:3.

REVIEW ARTICLE

Towards Predicting the Cytochrome P450 Modulation: From QSAR to Proteochemometric Modeling

Watshara Shoombuatong[1,†], Philip Prathipati[2,†], Veda Prachayasittikul[1,†], Nalini Schaduangrat[1,†], Aijaz Ahmad Malik[1], Reny Pratiwi[1], Sompon Wanwimolruk[3], Jarl E. S. Wikberg[4], Matthew Paul Gleeson[5], Ola Spjuth[4] and Chanin Nantasenamat[1,*]

[1]Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; [2]National Institutes of Biomedical Innovation, Health and Nutrition, Osaka 567-0085, Japan; [3]Center for Research and Innovation, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; [4]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala 751 24, Sweden; [5]Department of Chemistry, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

ARTICLE HISTORY

Received: November 30, 2016
Revised: December 16, 2016
Accepted: January 17, 2017

DOI:
10.2174/1389200218666170320121932

Abstract: Drug metabolism determines the fate of a drug when it enters the human body and is a critical factor in defining their absorption, distribution, metabolism, excretion and toxicity (ADMET) characteristics. Among the various drug metabolizing enzymes, cytochrome P450s (CYP450) constitute an important protein family that aside from functioning in xenobiotic metabolism, is also responsible for a diverse array of other roles encompassing steroid and cholesterol biosynthesis, fatty acid metabolism, calcium homeostasis, neuroendocrine functions and growth regulation. Although CYP450 typically converts xenobiotics into safe metabolites, there are some situations whereby the metabolite is more toxic than its parent molecule. Computational modeling has been instrumental in CYP450 research by rationalizing the nature of the binding event (*i.e.* inhibit or induce CYP450s) or metabolic stability of query compounds of interest. A plethora of computational approaches encompassing ligand, structure and systems based approaches have been utilized to model CYP450-ligand interactions. This review provides a brief background on the CYP450 family (*i.e.* its roles, advantages and disadvantages as well as its modulators) and then discusses the various computational approaches that have been used to model CYP450-ligand interaction. Particular focus was given to the use of quantitative structure-activity relationship (QSAR) and more recent proteochemometric modeling studies. Finally, a perspective on the current state of the art and future trends of the field is also provided.

Keywords: Drug metabolism, cytochrome P450, CYP450, ADMET, pharmacokinetics, drug design, QSAR, proteochemometrics.

1. INTRODUCTION

Metabolism involves the biochemical transformation of molecules, representing an essential process for sustaining the many facets of life [1]. It is an essential component of the host defense system against foreign (and potentially toxic) xenobiotic substances and acts by converting them into metabolites that are more easily cleared *via* excretion. Drug metabolism can be classified as belonging to either phase I or II [2]. In the former phase, polar functional groups are added to the drug so as to increase hydrophilicity whereas in the latter case, the drug is conjugated to functional moieties (*e.g.* acetylation, methylation, glucuronidation, sulphation, *etc.*) to increase its molecular size and hydrophilicity [3]. In addition, an individual’s drug metabolism is influenced by several internal factors (*i.e.* age, gender, genetic polymorphism and disease states including within the kidneys and liver) and external factors (*i.e.* smoking, nutrition and alcohol consumption) [4]. Such biotransformation processes generate metabolites that have altered physicochemical, pharmacokinetic and toxicological properties [3]. Pharmacokinetics relates to the fate of a substance that enters a living organism. These includes parameters such as absorption, distribution, metabolism and excretion or collectively known as ADME. These parameters determine the drug efficacy, drug-drug interaction and play a key role in toxicity [5].

2. CYTOCHROME P450

A wide range of protein families exist for xenobiotic metabolism and this includes dehydrogenases, flavin- containing monooxygenases, glutathione S -transferases, hydrolases, peroxidases, sulfotransferases, UDP- glucuronosyltransferases and cytochrome P450 (CYP450) enzymes. Of these, CYP450 represents a large group of enzymes playing pivotal roles in critical life functions including steroid and cholesterol biosynthesis, fatty acid metabolism, calcium homeostasis, neuroendocrine functions and growth regulation [6]. The number 450 represents the characteristic wavelength (nm) at which, the members of this superfamily maximally absorb light owing to the heme-coordinating axial cysteine ligand [7]. It is encoded by the P450 gene superfamily and historical evidences suggest that CYP450 genes can be found in almost all organisms (*e.g.* bacteria, yeast, fungi, plants, animals and human) [8].

[*]Address correspondence to this author at the Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand; Tel: +66 2 441 4371 ext. 2715; Fax: +66 2 441 4380; E-mail: chanin.nan@mahidol.edu

[†] These authors contributed equally to this work.

1389-2002/17 $58.00+.00

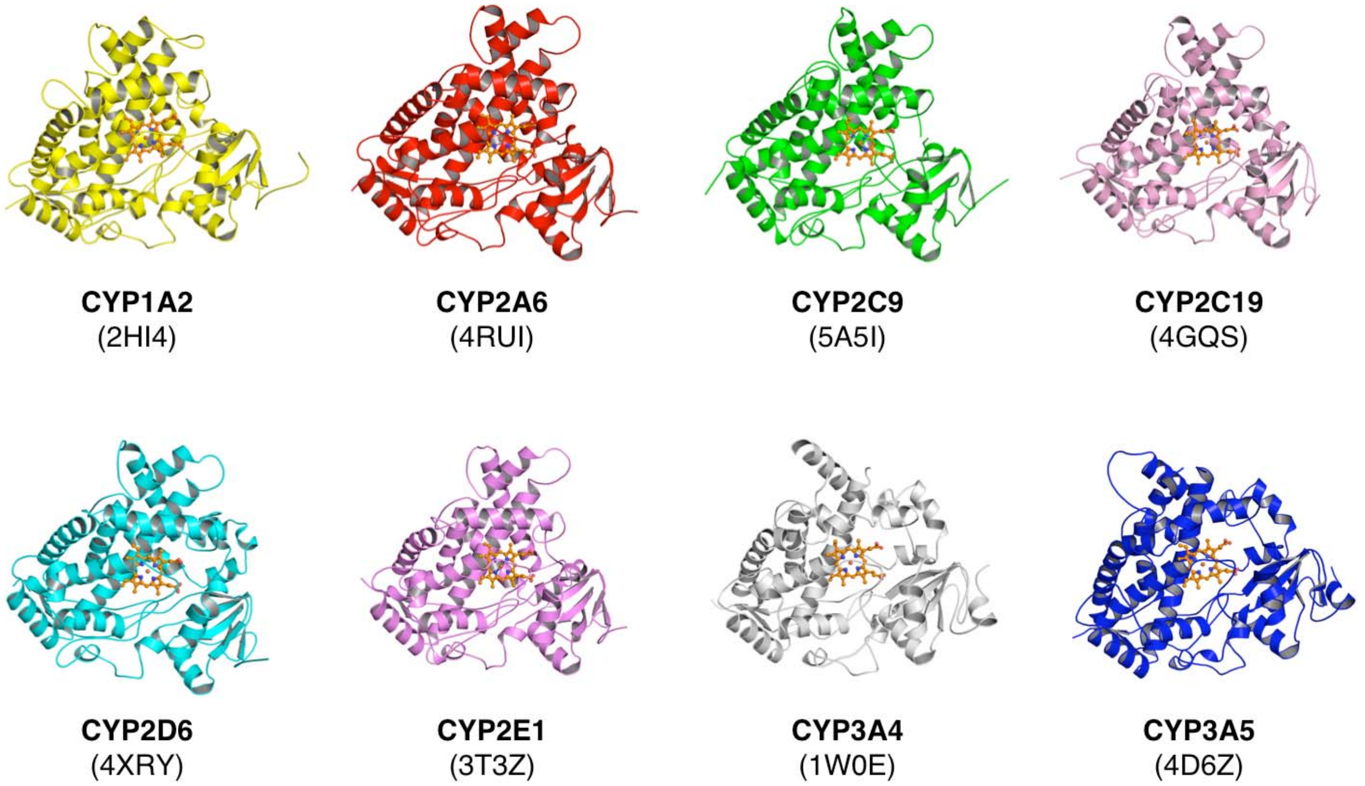


Fig. (1). Protein structure of selected members from the CYP450 protein family. Isoform name and their respective Protein Data Bank (PDB) identification number are shown in bold text and parentheses, respectively.

The naming convention of CYP450 genes (*e.g. CYP3A4*) is comprised of four main components [9]: (i) *CYP* represents the superfamily, (ii) the first Arabic number *3* indicates the family (*i.e.* its members contain sequence identity greater than 40%), (iii) the letter *A* indicates the subfamily (*i.e.* its members contain sequence identity greater than 55%) and (iv) the last Arabic number *4* defines the individual gene. In humans, there are 18 known families, 44 subfamilies and 57 functional genes [3, 10]. The structures of common CYP450 isoforms are shown in Fig. (**1**).

CYP450 are heme *b*-containing monooxygenases in which the heme molecule is linked to an apoprotein *via* a conserved cysteine [7]. This prosthetic heme contains an iron (Fe) atom coordinated to nitrogen atoms of the porphyrin ring. Moreover, CYP450 catalyzes a wide range of oxidation reactions [6] involving multiple steps for transferring oxygen (O_2) and proton (H^+) to the substrate. This catalytic activity requires nicotinamide adenine dinucleotide phosphate (NADP+) in the reduced state (NADPH) as an electron donor:

$$\mathrm{NADPH} + \mathrm{H}^+ + \mathrm{O}_2 + \mathbf{RH} \rightarrow \mathrm{NADP}^+ + \mathrm{H}_2\mathrm{O} + \mathbf{ROH},$$

where **RH** represents the substrate drug to be oxidized while **ROH** represents the hydroxylated metabolite product. Its substrates are structurally diverse encompassing fatty acids, steroids, terpenes, prostaglandins, heteroaromatic and polyaromatic compounds [11]. The broad recognition of CYP450s is also extends to many drugs, pesticides, carcinogens and toxicants [5]. As CYP450 enzymes are expressed in many organs, especially in the liver, they are considered to be one of the most important drug metabolizing enzymes [8]. Furthermore, a single drug or compound may be metabolized by several CYP450 enzymes (*i.e.* belonging to either the same or different isoforms) and the reaction may occur at either the same or different sites of the enzyme [3]. The relative rates of metabolism by each CYP450 are clinically important in terms of drug clearance, drug-drug interaction and treatment outcomes [8]. Therefore, dose adjustments and periodic monitoring of drug serum levels are recommended for maintaining the therapeutic serum levels as well as to minimize adverse effects [5]. Among all human CYP450s, only 6 isoforms are highlighted as major drug metabolizing enzymes consisting of CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1 and CYP3A4 [3]. CYP3A4 attracts the most attention as it metabolizes approximately 50% of the clinically prescribed drugs [4]. It has also been reported that the therapeutic failure of orally administered drugs is due to the extensive first-pass gastrointestinal metabolism that is mediated by CYP3A4 [5]. Also, the enzymes belonging to the CYP2 family should not be overlooked because they have been reported to have a high rate of genetic polymorphism, especially CYP2D6 and CYP2C19 [12].

CYP450s are classified according to their electron donating redox proteins into two classes [11]. Despite having different intermediate electron transferring molecules, both classes of CYP450s have the same electron donor (*e.g.* NADPH) and final electron acceptor (*i.e.* the Fe atom). Class I CYP450s are found in bacteria and eukaryotic mitochondria whereby electrons are donated from a NADPH to a membrane-bound flavoprotein ferridoxin reductase (FDXR), a soluble ferridoxin (FDX) and class I CYP450, respectively [6]. On the other hand, Class II CYP450s are found in eukaryotic endoplasmic reticulum (ER) [13]. Unlike class I CYP450s, the electrons from NADPH are donated to flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN) containing P450 reductase or POR [13]. The acceptance of electrons leads to a conformational change of the POR which allows FAD and FMN units to move closer for optimal transferring of electrons [6]. After the electrons are transferred to FMN, the POR returns to its original conformation to facilitate electron transfer from FMN to class II P450 [6]. For both classes, the electrons are ultimately transferred from CYP450s to the Fe atom, constituting the final electron acceptor which passes electrons to the molecular oxygen thereby giving rise to the catalytic activity of the enzyme [11] (Fig. **2**).

Understanding inter-individual differences to drug response is another area of great interest in the age of personalized medicine [14].

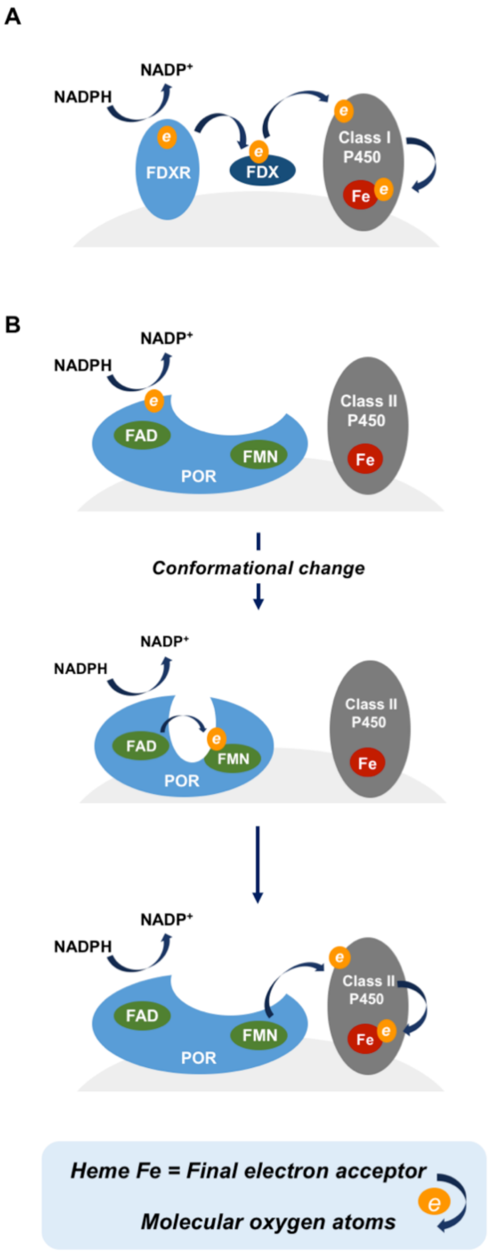


Fig. (2). Mechanism of catalytic activity of CYP450. Class I CYP450 (**A**) and Class II CYP450 (**B**) shares the same electron donor, which is a reduced form of nicotinamide adenine dinucleotide phosphate (NADPH), and final electron acceptor, (Fe) atoms. In panel A, ferredoxin reductase (FDXR) and ferredoxin (FDX) are intermediate electron transferring molecules of class I CYP450. In panel B, cytochrome P450 reductase (POR) serves as an electron transferring unit where conformational change facilitates the electron transfer from FAD and FMN as well as from POR to CYP450.

Inter-individual differences in drug response is affected by factors influencing CYP450 expression and function (*i.e.*, genetic polymorphism, epigenetics, host factors) [15]. Genetic polymorphism of CYP450 is considered as a key factor that can be used to explain the differences to drug responses (*i.e.*, efficacy, toxicities, drug-drug interactions) and the carcinogenic potential of xenobiotics among individuals [4]. In addition, many human CYP450 enzymes are polymorphic, in other words, possess more than one variant of the gene [12]. Understanding genetic polymorphism can also be useful towards tailoring the design of a drug and its optimum dosage; in other words, giving the right drug in the right dose in order to treat the right disease catered towards the right person [14]. Genetic polymorphism of CYP450 can be generally classified into three classes: (a) extensive metabolizer or those with normal CYP450 enzyme activity, (b) poor metabolizer or those with completely absent CYP450 activity, and (c) ultra-extensive metabolizer or those with enhanced CYP450 activity [4]. Furthermore, the lack of enzyme activity observed in poor metabolizers is due to two alleles of the gene that are present but exhibit a lack of function (null alleles) [15]. This lack of drug metabolism leads to reduced drug clearance and high plasma drug concentrations which have the potential to lead to toxicity if it reaches the maximum tolerated dose of the drug. In contrast, therapeutic outcome is impaired in poor metabolizers if the administered drug is a prodrug that requires metabolism in order to obtain its pharmacologically active form [16]. Among all CYP450 isozymes, CYP2D6 is highlighted as the most polymorphic enzyme with recessive poor metabolizer phenotype [17]. For example, the lack of CYP2D6 in individuals causes a predisposition to antidepressant or neuroleptic induced drug toxicity [18]. Another potentially disastrous polymorphism can be seen in people exhibiting deficient activity of CYP2C9. Patients possessing this enzyme variant are ineffective in clearing of an anticoagulant warfarin [19]. In contrast to poor metabolizers, therapeutic failure due to enhanced enzyme activity, increased drug clearance, decreased plasma drug concentration and reduced bioavailability is observed in the ultra-extensive metabolizers carrying one common allele and one amplified allele [15].

3. MODULATORS OF CYP450

Knowing whether an administered drug is a substrate, inducer or inhibitor of CYP450 is useful for preventing potential clinical drug-drug interactions, improving therapeutic outcomes, and minimizing adverse effects [5]. The metabolizing function of CYP450 enzymes can be modulated by a wide range of structurally diverse molecules. The molecules that enhance the enzyme's activity are classified as inducers whereas those that inhibit the enzyme's activity are categorized as inhibitors [4]. If a co-administered drug acts as an inhibitor of the relevant CYP enzymes, enzyme activity is inhibited which results in a decreased metabolism of one (substrate drug) or both drugs (substrate and inhibitor) ultimately increasing its/their levels in the serum. This increase may subsequently result in a variety of minor as well as major adverse events and toxicity [20]. On the other hand, if a co-administered drug acts as an inducer, the activity of the enzyme is enhanced leading to rapid drug clearance and therefore, a decreased efficacy [3].

The high clinical importance of CYP450 enzymes may be attributed to their promiscuous nature [17]. Unlike other enzymes that specifically target a particular type of structurally related compounds, a single CYP450 isozyme may be able to metabolize a wide range of structurally unrelated drugs [17]. In addition, one drug can be metabolized by multiple CYP450 isozymes [21]. The metabolism occurs either at the same or different site with the same or different rate of metabolism [3]. In this regard, the relative rates of metabolism incited by all CYP450 isozymes allow the determination of the dominating isozyme inducing drug efficacy, clearance and toxicity [3]. In particular, the isozyme whose metabolism rate is relatively faster than others is considered as the most clinically important [3]. Moreover, a drug can simultaneously act as a substrate for a particular CYP450 isozyme while being an inducer or inhibitor of the others [21]. Therefore, drug-drug interactions due to CYP450 are another issue that needs consideration when two or more drugs are co-administered [17]. In addition, the drug serum level is governed by metabolism of the substrate drug itself as well as those that are co-administered [22]. Co-administration of drugs sharing the same metabolic pathways may potentiate drug-drug interactions (DDI) [5]. The co-administered drugs act as additional substrates which can either be the inhibitor or inducer of the responsible CYP450 enzymes [16].

3.1. CYP450 Substrates

Broad substrate recognition and promiscuity of metabolizing CYP450 enzymes mean they play a key role in drug metabolism and as a consequence, other factors including efficacy, toxicity as well as drug-drug interaction. As these factors may be intertwined with one another, the modulation or alteration of one may impact the other. Drug metabolism influences the therapeutic efficacy, adverse effects and toxicities of an administered drug by controlling its therapeutic serum level [22]. An appropriate serum concentration is a key factor for determining the desirable therapeutic effects and toxicities and is controlled by both drug metabolism and drug clearance [23]. The appropriate rates of metabolism and clearance are required to maintain an apt drug serum concentration that can produce pharmacological effects while minimize adverse effects or toxicities [4]. After exerting its pharmacological effect, the administered drug can be directly eliminated by means of urine and other excretory fluids as well as can be inactivated through biotransformation as facilitated by metabolizing enzymes (*e.g.* CYP450s) as to allow clearance from the body [3]. Apart from inactivation of drugs to facilitate clearance, the metabolizing function of CYP450 enzymes are essential for converting prodrugs (*i.e.* the inactive form) to the active drug metabolites required for its therapeutic actions [3]. For example, codeine is a prodrug that needs to be activated by CYP2D6 to afford the active metabolite namely, morphine.

3.2. CYP450 Inhibitors

Broad substrate specificity of the CYP450 families makes them likely to be inhibited by a wide spectrum of structurally diverse molecules [17]. Inhibition of CYP450 is one of the most common mechanisms resulting in clinically relevant drug-drug interactions [24]. Most DDI related adverse effects are attributable to alterations in the CYP450 metabolic pathways. The catalytic cycle whereby the CYP enzymes regulate metabolism, involves 6 primary steps. Out of those, 3 steps of the CYP catalytic cycle are particularly prone to inhibition and they are: (i) substrate binding, (ii) molecular oxygen binding to ferrous enzyme (Fe^{+2}) and (iii) transfer of activated oxygen from the heme iron to the substrate, constituting the main catalytic step [25]. CYP450 inhibition can be generally classified into two types according to its mechanism of inhibition: (i) reversible inhibition and (ii) irreversible inhibition [17]. For reversible inhibition, inhibitors act by competing for the metabolic activity of the same CYP thereby blocking access of the substrate drug without causing structural change to the enzyme [17]. Reversible inhibition constitutes competitive binding at the CYP active site, usually involved in the substrate binding step of the catalytic cycle and is also the major form of inhibition observed. In contrast, irreversible inhibition involves an alteration of the other two aforementioned catalytic cycle steps [20, 25] and is achieved when the inhibitor binds to CYP enzyme and forms a reactive intermediate that subsequently destroys the enzyme *via* covalent modification [17]. Irreversible CYP450 inhibitors are generally further classified into three classes based on the site or type of interaction of the produced reactive intermediate and the enzyme; (i) compounds producing reactive intermediate that covalently binds the apoprotein, (ii) compounds producing reactive intermediate that covalently binds the prosthetic heme group, and (iii) compounds producing reactive intermediate that forms a complex with the enzyme (metabolic intermediate complex) leading to the destruction of the heme prosthetic group (*i.e.* quasi-irreversible inhibition) [20, 25]. However, the type of inhibition may not be able to clearly distinguish whether the inhibitor is tightly bound and slowly released from the enzyme [17]. Furthermore, irreversible inhibition is also called time-dependent inhibition (TDI). TDI of cytochrome P450 is of particular concern because *de novo* synthesis of the enzyme is required in order to restore activity.

It should be noted that the active sites of some CYP450 isozymes (*e.g.* CYP3A4 and CYP2C9) are relatively large and can therefore accommodate multiple substrates simultaneously [7]. This characteristic could lead to partial inhibition of the enzyme upon simultaneous binding of multiple compounds [17]. For example, if compounds a and b bind to an enzyme simultaneously, one compound a could either act as a competing substrate or as an inhibitor of compound b. Both phenomena decrease enzyme activity against the affected compound (compound b) thereby impairing its metabolism [26].

3.3. CYP450 Inducers

Inducers result in increased metabolism of the drugs involved thereby leading to diminished drug effects that ultimately lead to treatment failure. Thus, many drugs have been taken off the market due to adverse DDI. For example, Saldane, a non-sedating antihistamine was withdrawn from the market by the U.S FDA as its metabolic inhibition led to life-threatening arrhythmias [27]. Moreover, inducers enhance the activity of a particular CYP450 isozyme by modulating its gene expression level [4]. CYP induction is commonly mediated by three transcription factors *i.e.* aryl hydrocarbon receptor (AHR), pregnane X receptor (PXR) and constitutive androstane receptor (CAR) [28]. The primary mechanism of CYP450 induction *via* increased gene transcription typically occurs through nuclear receptor activation. While the first evidence of enzyme induction may occur after multiple doses *in vivo* (AUC and $T_{1/2}$ reduction), there are a number of *in vitro* systems that can be used. Among them, nuclear receptor transactivation assays using a stably transfected human hepatoma cell line and a luciferase reporter gene assay are often used. However, regulatory network of CYP induction is complex because the induction is not only dependent on ligand binding, but also on other factors affecting translocation of key transcription factors to the nucleus as well as the activity of cofactors (*i.e.*, nuclear translators). Of note, the complexity of gene expressing regulatory network may be amplified with cross-regulation and overlapping receptor modulation by other signaling pathways [16].

Apart from simultaneous substrate binding, active sites of some CYP450 enzymes (*i.e.*, CYP3A4 and CYP2C9) are highlighted for their atypical kinetics. Atypical kinetics of active sites allows the enzyme to be activated by effector molecules besides its substrates (heterotropic activation) [17]. Heterotropic activation can therefore be defined as an increased enzyme activity in the presence of another compound (inducers) which cause structural or electronic changes to the enzyme [17]. The mechanisms behind heterotropic activation has been proposed to be due to the presence of multiple enzyme binding sites and conformations, the presence of an enzyme allosteric site as well as enzyme conformational changes [29]. It should not be overlooked that the enzyme inducer can also be an enzyme substrate [21]. In this situation, metabolism of the compound may or may not be inhibited by another substrate [21]. Moreover, region-selectivity is indicated for some metabolizing reactions, in other words, the enzyme reaction produces more than one type of product from a single substrate and these products can possibly activate one oxidative pathway while inhibiting another [17].

With regard to the aforementioned context, CYP450 induced drug-drug interactions in human is complex and unpredictable as the outcome of multiple drugs bound to CYP450 enzymes can range from no effect, activation and inhibition [21]. If the modulator binds at other sites apart from the substrate binding site without altering the binding site of the substrate, the enzyme activity may not change or may even increase due to heterotropic activation. In contrast, if the modulator competitively binds at the substrate binding site or binds at the site that blocks the access of the substrate to its binding site, the enzyme activity is inhibited and the metabolism rate of the substrate is impaired [17].

Additionally, CYP-dependent metabolism may occasionally cause the formation of toxic or carcinogenic intermediates which are ultimately cleared *via* the phase-II enzyme dependent conjugation reactions [30]. However, many a times, although uncommon, some of those byproducts are not effectively cleared. Nevertheless, as previously mentioned, several prodrugs are activated by the actions of CYP, making them effective in cancer chemotherapy [31]. Therefore, the involvement of CYP in both activation and inactivation of anticancer drugs, infers the possibility of its association in the etiology of cancer [32]. In addition, substantial clinical consequences are also observed with CPY variability, not only in different populations and ethnicities, but also in different age groups. Such as the CYP1A2 isoform is not expressed in neonates, making them particularly susceptible to toxicities from drugs like caffeine [33].

3.4. Natural Products

Extensive experimental data conducted over the past decade indicates the active role of natural polyphenols in the modulation of CYP catalytic activity. These polyphenols are of special interest as they are micronutrients of plant origin that constitute a substantial proportion of human diet and medicinal herbs [34]. Dietary polyphenols work by potentially altering the activity of enzymes through modulation of their protein expression levels or through binding with their active sites, directly [34]. For example, the ability of CYP-catalyzed arachidonic acid oxidation, plays a significant role in the development of cardiovascular diseases [35]. Concerns regarding CYP450 mediated drug interactions are also extended to many active ingredients in food and beverages [36]. The first recognized herbal induced drug interaction was observed in grapefruit induced CYP3A4 inhibition [37, 38]. Herbal/food induced drug interaction has gained continual attention as several food, herbs, and their metabolites have been reported to interfere with the catalytic activities of CYP450s. For example, cruciferous vegetables such as cauliflower and cabbage are reported inducers of CYP1 family (*i.e.*, CYP1A1 and CYP1A2) [39] while an active flavonoid from grape namely resveratrol is an inhibitor of CYP1A1 and CYP3A4 [40]. Additionally, a medicinal plant called St. John's Wort and its bioactive compound namely hyperforin have also been reported as inducers of the CYP2 family (*i.e.*, CYP2B6, CYP2C9, and CYP2C19) as well as CYP3A4 [15].

3.5. Physicochemical Properties of CYP450 Modulators

The active sites of the major forms of CYP450 superfamily rely heavily on hydrophobic interactions with their substrates, ion-pair and

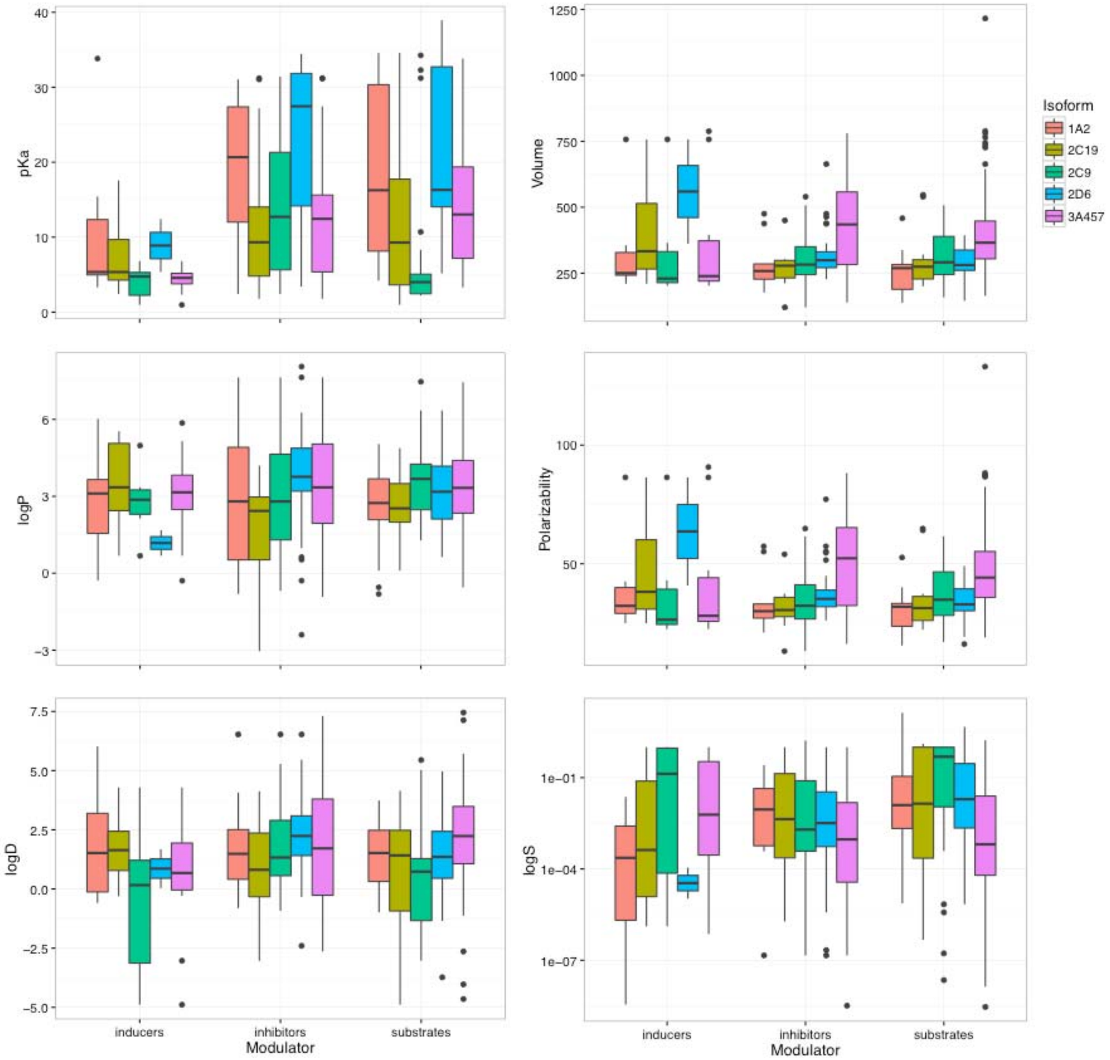


Fig. (3). Box plots depicting distributions of six major physicochemical properties (pKa, logP, logD, volume, polarizability and logS) for inducers, inhibitors and substrates of important CYP isoforms 1A2, 2C19, 2C9, 2D6 and 3A4,5,7. The compound-CYP interaction data was taken from the website maintained by David A. Flockhart [41, 42]. The Flockhart Table's on drug-CYP interaction together with the extracted structure and computed physicochemical properties are presented as a Supplementary Table **1**. Compound names were converted to chemical structures using the *CTSgetR* package and physicochemical properties were computed using the ChemAxon's calculator plugins. Box plots were generated using the *ggplot2* package in R after parsing the data. The interquartile range between the first (Q1) and third (Q3) quartiles covers the central 50% of the data and is presented as a box. A line inside the box that is not necessarily central depicts the median. Black dots outside the boxes are outliers with respect to the interquartile regions.

Table 1. Summary of the bioactivity data of selected CYP450 available from ChEMBL (version 22).

Gene Names	Uniprot Entry Name	UniProt Entry	ChEMBL_ID	Compounds	Bioactivities
CYP1A2	CP1A2_HUMAN	P05177	CHEMBL3356	20766	22889
CYP2C9	CP2C9_HUMAN	P11712	CHEMBL3397	22689	27752
CYP2C19	CP2CJ_HUMAN	P33261	CHEMBL3622	20685	26114
CYP2D6	CP2D6_HUMAN	P10635	CHEMBL289	23549	29376
CYP3A4	CP3A4_HUMAN	P08684	CHEMBL340	28464	45340
CYP3A5	CP3A5_HUMAN	P20815	CHEMBL3019	207	376
CYP3A7	CP3A7_HUMAN	P24462	CHEMBL3341582	15	28

The SQL queries [select count(a.activity_id) from activities a,assays ass,target_dictionary td where a.assay_id=ass.assay_id and ass.tid=td.tid and td.chembl_id='CHEMBL_target_ID';) and [select count(distinct a.molregno) from activities a,assays ass,target_dictionary td where a.assay_id=ass.assay_id and ass.tid=td.tid and td.chembl_id='CHEMBL_target_ID';] were used to query locally an installed SQL database of ChEMBL, for the number of unique compounds and bioactivities associated with selected CYP isoforms, respectively.

hydrogen bonding interactions. Although CYP inhibition and metabolism have a strong dependence on molecular recognition beyond relatively simple physicochemical properties, the chemical space mapping in terms of these properties provides a general overview of physicochemical preferences of various CYP isoform modulators. Hence, in the following section we attempt to analyze the physicochemical properties of CYP inhibitors, modulators and inducers using box plots. The compound-CYP interaction data was extracted from the website maintained by David A. Flockhart [41, 42].

Solubility impacts various stages of the drug discovery process. While low solubility correlates with low gastrointestinal absorption, it also negatively impacts the therapeutic dose. As expected, Fig. (**3**) reveals that most CYP modulators generally have poor solubility driven by their dependence of hydrophobicity (*i.e.* logs of most CYP modulators is lower than 1 mol/L) with inhibitors and substrates of CYPs 3A4,5,6 demonstrating the worst aqueous solubility measures. Many computational models of the CYP3A4 substrates and inhibitors as well as our analysis also implicate LogD in inhibitory and substrate interactions given the lipophilic nature of the CYP3A4 active site. Consequently, the introduction of a polar substitution is often beneficial to the inhibitors and substrates of CYP3A4,5,6 as it improves water solubility and reduces LogD.

Neutral, nonpolar species have a spherically symmetric arrangement of electrons in their electron cloud. When in the presence of an electric field, their electron cloud can be distorted. The ease of this distortion is defined as the polarizability of the atom or molecule. The distorted electron cloud causes the originally nonpolar molecule or atom to acquire a dipole moment. Generally, polarizability increases as the volume occupied by electrons increases. Consequently, the inhibitors and substrates of CYP3A4,5,7 have higher polarizability. Furthermore, as the binding pocket of CYP3A4 is quite large, it can accommodate molecules with a high molecular volume. Interestingly, CYP1A2 inhibitors and substrates have the lowest polarizability since the distortion of the electron cloud of a compounds aromatic system is detrimental for maintaining aromatic pi stacking interactions with Phe226 and Phe260 of CYP1A2 (Fig. **4**). As a result, some strategies that reduce CYP3A4 inhibition or substrate interactions are in turn reducing the volume and polarizability. However, increasing the polarizability of CYP1A2 inhibitors by introducing F or CF3 to the aromatic groups greatly reduces their CYP liabilities.

Acidity, basicity and hydrophobic features arising from the functional groups of the substrates are also responsible for demonstrating CYP isoform specificity. Our analysis presented in Fig. (**3**) emphasizes the importance of the basicity for inhibitory and substrate interactions with Asp301 and/or Glu216 of CYP2D6 (Fig. **4**) which corroborates with other quantitative structure-activity relationship (QSAR) and pharmacophore modelling studies. The importance of the acidic fragments in modulating substrate interactions with Arg108 of CYP2C9 (Fig. **4**) is also highlighted in our analysis and substantiates other theoretical studies on CYP2C9 inhibitors and substrates. Hence, reducing the basicity of CYP2D6 inhibitors and substrates as well as the acidity of CYP2C9 inhibitors and substrates can lead to a beneficial CYP modulatory effect.

3.6. Importance of Predicting Drug Metabolism

The high attrition rate in drug development is mainly due to severe toxicity and efficacy. Therefore, the screening of ADME properties as well as the understanding of metabolizing enzymes and metabolic stability for compounds of interests are strongly recommended measures that should be taken in the early developmental phase as to increase the success rate as well as reduce time and cost [5]. The clinical importance of CYP450 enzymes has been reiterated as they are the major metabolizing enzymes responsible for more than 75% of drug metabolism [16]. It is estimated that 23-27% of candidate compounds are discarded owing to biotransformation-related toxicity as well as unfavorable phase I metabolism caused by CYP450 enzymes [16]. Therefore, the *in vitro* screening of CYP450 inhibition of drugs is useful for predicting potential metabolism-mediated drug-drug interaction that is likely to occur *in vivo* [17]. Currently, many *in vitro* assays are available for providing information relating to CYP induction and inhibition, drug-drug interaction, CYP isoform identification and metabolic stability of compounds [4]. These *in vitro* assays are recommended by the FDA for initial assessment on the effect of drugs on metabolic pathways and drug-drug interaction potential [4]. The *in vitro* assays provide information relating to the key enzymes responsible for metabolism of the candidate compound, such as inhibitory potency, metabolic stability and toxicity, which is useful for further drug development process (*i.e.*, selection of *in vivo* assays and prediction of *in vivo* and clinical trials results) [4]. However, harmonized and definitive guidelines for evaluation of CYP inhibition are still unavailable [16] and drug-drug interaction information obtained from *in vitro* studies may not always correlate with or may not be appropriate to predict those of *in vivo* [17]. In addition, misinterpretation and miscalculation of the inhibitory potency of compounds may exclude safe and potent drugs from further development process or may prevent potent CYP inhibitors from reaching patients [17]. Moreover, kinetic parameters or inhibitory potency for compounds of interests are influenced by many factors, for instance, biochemical environment, non-specific protein binding and substrate-dependent factors, which could be the cause of the underestimated *in vitro* results thereby leading to serious errors in the prediction of *in vivo* drug-drug interaction [17]. Therefore, the accurate prediction of CYP inhibition and CYP-induced drug-drug interaction is indeed a challenging task [16, 17]. So far, great attention has been given to the utilization of *in silico* approaches, including molecular modeling and QSAR, for understanding drug response due to CYP450 enzymes. While molecular modeling is highlighted for understanding active sites and possible interactions, QSAR is utilized for prediction of drug-drug interactions, especially in the areas of cancer therapy and polypharmacology. However, most of the current QSAR studies relating to CYP450s are limited to ligand-based approaches which still rely on existing experimental data [43]. Therefore, the problems of over-predictive models as well as the aforementioned approach [44] are gaps that need to be filled [16].

4. COMPUTATIONAL APPROACHES FOR PREDICTION OF DRUG METABOLISM

The prediction of the rate of metabolism or the nature of the metabolites is of great interest in drug discovery. Drug metabolism is a complex biochemical network, which consists of many different parts and reactions. It is one of the most complicated pharmacokinetic properties to be understood and predicted. A majority of the drugs react with CYPs to produce metabolites which can be benign, pharmacologically active or produce adverse effects. Additionally, small molecules can also induce or inhibit CYPs and are implicated in drug-drug interactions. The drug-CYP interactions can be studied experimentally by (i) incubating drugs with individual drug-metabolising enzymes *e.g.* CYPs, UGTs or hepatocytes, (ii) specific reactive metabolite trapping in microsomal incubations, trapping of soft nucleophiles using *e.g.* glutathione or cysteine, and hard nucleophiles with, for instance, cyanide, (iii) conventional animal models (rat) and (iv) newer genetically modified humanized or “chimeric” mouse models.

The metabolic behavior of drugs depends not only on the physicochemical properties of compounds, but also on the structural characteristics of the involved metabolizing enzymes, whose expression depends in turn on a number of genetic and environmental factors. Hence, a range of *in silico* methods have been developed for the prediction of CYP reaction, inhibition and induction [45]. These can be classified into physics-based and empirical models and also into local and global models. Physics-based methods such as quantum chemical (QC) calculations and free energy perturbations are generally used for the prediction of the site of metabolism (SoM) and the rates of the metabolic reactions. Empirical methods, based on existing experimental data without knowledge of the physics of the system may be divided to ligand-based and target-based approaches and are generally used for the prediction of CYP inhibition or induction. They include methods such as quantitative structure-activity relationship (QSAR), pharmacophore and certain class of docking and scoring. In ligand-based empirical methods, structures of known active and inactive compounds are modeled to derive QSAR and pharmacophore models. Also various rule-based expert systems belong to this category. In target-based methods, the structure of the enzyme is the starting point for model generation. Models integrating both ligands and enzymes are known as combined or mechanism-based methods.

QSARs are linear and non-linear relationships that relate chemical structure (encoded by calculated descriptors) with, for this application, metabolic properties. 3D-QSARs are among the most widely used whereby they relate the metabolic activity of a set of aligned compounds with their 3D electronic, steric and hydrophobic properties. Pharmacophore generation is another interpretable model that captures the common structural features and their 3D spatial arrangements, from a series of molecules characterized by similar metabolic properties. Protein-based methods form the other end of the spectrum of *in silico* approaches and rely upon the structural information extracted from the X-ray crystallographic and/or homology protein structures. These also include docking techniques for the exploration of possible binding modes of a ligand to a given enzyme or receptor. This approach predicts energetically favorable conformations of ligands and also reveals key groups or atoms for binding. With crystal structures available for the major human CYPs (Fig. **4**), protein-ligand docking methods are increasingly used for the analysis and prediction of CYP-ligand interactions. However, docking poorly accounts for substrate reactivity [46, 47]. High promiscuity with regards to substrates, high flexibility and clinically significant genetic polymorphism of the CYP enzymes makes the application of target based approaches to modeling of CYP inhibition, a challenging task.

Today, *in silico* methods used to evaluate CYP-ligand interactions typically combine techniques from physics-based and empirical models. With appropriate combinations, the strengths of individual *in silico* methods complement each other. The most promising of these approaches are the ligand-protein interaction-based (mixed) approaches, whereby a synthesis of the information on both ligands and proteins is attempted, in relation to the relevant metabolic property [48]. Moreover, expert systems mimic human reasoning and formalize existing knowledge. These are programs in which a computer solves problems by applying rules from a knowledge base. Such rules may be a combination of factual and heuristic types, and are usually non-numerical. In most cases, 3D structures of compounds are not required. Metabolic pathways are sometimes very different even in closely related mammalian species, thus some expert systems allow filtering of specific subsets of the data to a specific species [47, 48]. In addition, expert systems exploit the extensive databases of experimentally derived metabolic pathways. Examples of such databases include the BIOVIA Metabolite database [49] and Fujitsu ADME database [50].

Furthermore, various *in silico* approaches for modelling the xenobiotic-CYP interactions are presented in Table **2**. Given the diverse modes of actions described in Figs. (**2** and **4**), along with the computational costs associated with quantum chemical studies and protein structure based docking studies, makes QSAR highly relevant in modelling the various types of xenobiotic-CYP interactions. In addition to the low computational costs, recent advances in machine learning allow QSARs to model non-linear and xenobiotic synergistic [51] and antagonistic interactions. QSARs models can be applied to screen large compound databases which are interpretable and can assist medicinal chemists with design strategies to overcome CYP liabilities [52]. Also, inverse QSAR [53] and multiobjective QSARs [54] may hold great promise in the design and optimization of compounds with improved therapeutic dose and reduced metabolic liabilities. Moreover, simple SMILES-based descriptors have extensively been shown to afford robust and yet interpretable QSAR models for small molecules [55, 56] and peptides/proteins [57, 58].

Amongst the various *in silico* approaches described in Table **1**, QSARs approaches may be better suited for modelling and predicting the CYP-inhibitor interactions since (i) the binding cavities of CYPs can be large and flexible, inhibitor molecules can coordinate directly to heme, bind close to heme or at a distant site in the protein (Fig. **4**), (ii) several ligands may bind simultaneously, (iii) the inhibitor may be oxidized to an electrophilic reactive intermediate, which forms covalent bonds with the CYP protein thereby causing mechanism-based inhibition.

Regioselectivity and liability of CYP450 metabolism is of central importance to drug design. Drug metabolism in most instances results in loss of therapeutic drug efficacy, and may even cause toxicities and other adverse effects [59]. Various expert systems primarily focused on physics based and ligand based- approaches have been used for the prediction of labile sites of metabolism. Some of the widely used methods are discussed in the following paragraphs.

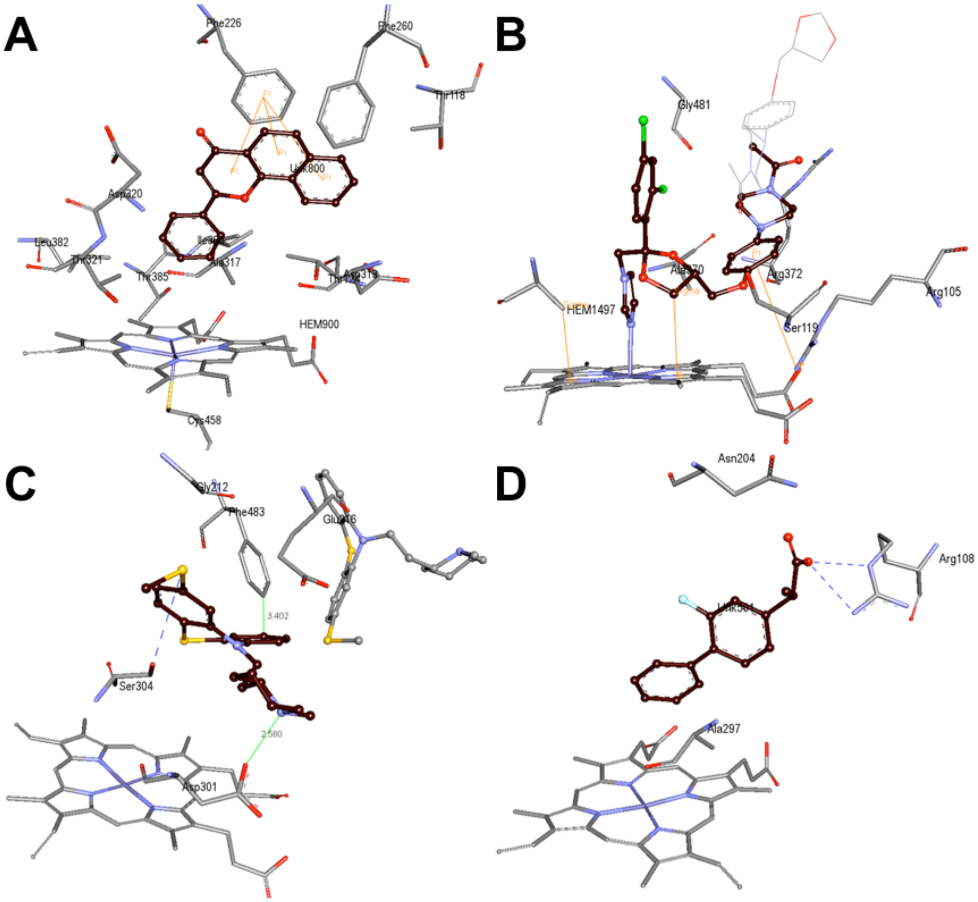


Fig. (4). Diverse substrate/inhibitory binding modes of various CYP450 isoforms with representative xenobiotics. A) The reversible inhibitory interactions of alpha-naphthoflavone with CYP1A2 as revealed by x-ray crystallographic coordinates of the PDB ID 2HI4, **B**) The irreversible inhibitory interactions of ketoconazole with CYP3A4 elucidated in PDB ID 2V0M, **C**) The reversible inhibitory interactions of two thioridazines with CYP2D6 elucidated in PDB ID 3TBG and **D**) The reversible inhibitory interactions of flurbiprofen with CYP2C9 elucidated in PDB ID 1R9O.

SMARTCyp [60] is primarily a CYP3A4 SoM lability prediction model and a general reactivity model which is applicable to all P450 isoforms. SMARTCyp predicts the site of metabolism directly from the 2D structure of a molecule, without requiring calculation of electronic properties or generation of 3D structures. In view of its advantages in terms of computational costs, the SMARTCyp methodology has been used by several groups to enhance other tools such as RS-Predictor.

XenoSite [61, 62] is another web-based tool for predicting the atomic sites at which xenobiotics will undergo metabolic modifications by CYP450 enzymes. XenoSite outputs are interpretable as probabilities, which reflects both the confidence of the model that a particular atom is metabolised and the statistical likelihood that its prediction for that atom is correct. XenoSite supports models addressing several different enzymes and mechanisms.

FAst MEtabolizer (FAME) [63] is a fast and accurate predictor of sites of metabolism (SoMs) and is a collection of random forest models trained on a comprehensive and highly diverse data set of 20,000 small molecules annotated with their experimentally determined sites of metabolism taken from multiple species (rat, dog and human). Using a comprehensive set of available data, FAME aims to assess metabolic processes from a holistic point of view. It is not limited to a specific enzyme family or species. Besides a global model, dedicated models are available for human, rat, and dog metabolism where specific prediction of phase I and II metabolism is also supported. FAME is able to identify at least one known SoM among the top-1, top-2, and top-3 highest ranked atom positions in up to 71%, 81%, and 87% of all cases tested, respectively.

StarDrop's [64, 65] CYP450 metabolism predictions incorporate both pathway-specific reactivity and isoform-specific accessibility considerations. Semiempirical quantum mechanical (QM) simulations, parameterized using experimental data and *ab initio* calculations, estimate the reactivity of each potential SoM in the context of the whole molecule. Ligand-based models, trained using high-quality regioselectivity data are correct for orientation and steric effects of the different CYP isoform binding pockets. In addition, in order to predict the relative proportion of metabolite formation at each site, these methods estimate the activation energy at each site, from which additional information can be derived regarding their lability in absolute terms.

Global models are used to predict the metabolism of any molecule exposed to a complex biological system. These models are often rule-based and use an extensive database of known biotransformations. MetaDrug [66] and Meteor [67, 68] are the prototypical global models/systems that use a series of rules together with a series of QSAR models to predict metabolic transformations, and includes both phase I and phase II metabolism. Many transformations including C, N, S and P-oxidation, including dealkylation, hydroxylation, double bond peroxidation, quinone formation, reduction (*e.g.* nitro, carbonyl, azo and sulphur), hydrolysis (*e.g.* esters, amides, phosphates and epoxides), glucoronidation, sulphation, glutathione conjugation, methyl transferases and amino acid conjugation are described.

5. QSAR MODELING

Structure-activity relationships (SAR) are useful for optimization of lead compounds in order to improve their metabolic stability

Table 2. Summary of computational approaches used for studying CYP450 inhibition, induction and reaction.

Modulation	Focus	Local Models			Global Models
		Physics-based	**Ligand-based**	**Target-based**	**Expert Systems**
CYP450 Reaction	SoM	● SMARTCyp [69] ● XenoSite [61] ● FAME [63] ● StarDrop [64, 65] ● CypScore [70] ● RS-Predictor [71]	● QSAR [72] ● Pharmacophore [73]	● QC-MM Docking [74] ● Tethered docking [75]	● MetaDrug [68]
	Metabolites	● QC [76]	● MetaPrint2D [77] ● RASCAL [78] ● MMRS [79]	NA	● FAME [63] ● TIMES [80]
	Metabolic rates	● Semi-empirical QC [81]	● QSAR [81]	NA	NA
CYP450 Inhibition	Target competitive inhibitors	NA	● QSAR [82] ● Pharmacophore [83]	● WhichCyp [84]	NA
	Target competitive irreversible inhibitors	NA	NA	● Docking [85]	NA
	Target allosteric inhibitor	NA	NA	● MD followed by analysis of allosteric site [86]	NA
	Target functional inhibitors	NA	● QSAR on DRUGMATRIX gene expression data set [87]	NA	NA
CYP450 Induction	Target functional activators	NA	● QSARs relating protein/gene activity or expression levels [88]	● Docking [88]	● Systems biology network models [89]

MD, molecular dynamics; NA, not available; QC, quantum chemical; QSAR, quantitative structure-activity relationship; SoM, site of metabolism.
The metabolic behavior of drugs depends not only on the physicochemical properties of compounds, but also on their structural characteristics and the expression levels of the involved metabolizing enzymes. Hence, a range of *in silico* methods are currently used for predicting the 3 major CYP-drug modulatory actions such as the induction, inhibition and reaction. Physics based *in silico* methods form one end of the spectrum with target based approaches occupying the other end. In more recent years, various expert systems and combined approaches are increasingly being developed and deployed in early stages of drug discovery research.

and ADME profiles [90, 91]. With improved knowledge of the CYP enzyme pathways, pharmaceutical companies are extensively screening new drug candidates for CYP450 drug interactions since the early drug discovery phase. Although our understanding of both CYP inhibitors and inducers has greatly improved over the past decade, the accurate prediction of DDI in terms of its occurrences and related consequences, prove to be a continued challenge. Of note, computational approaches serve as powerful tools for facilitating drug development as they could provide information useful for guiding the design of improved drugs [5].

A wealth of information is currently available in the literature as well as public databases on the bioactivity of CYP450s. In regards to the latter, ChEMBL (version 22) [92] lists more than 20,000 compounds for each of the five isoforms (*e.g.* CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4) with bioactivity data points in the range of 20,000 to 45,000. However, it should be noted that the underlying data may be heterogeneous in that some are inhibitors or inducers of CYP450. In addition to releasing the ChEMBL database, the EBI had also published the ADME SARfari database [93] that provides pharmacokinetic data for query compounds of interest. Furthermore, ever since DrugBank 4.0 [94] has been expanded to provide coverage on drug metabolism data, version 4.0 contains data on more than 1200 drug metabolites as well as more than 1300 drug metabolism reactions (*e.g.* metabolizing enzymes and reaction types). Moreover, aside from the bioactivity data that it provides, PubChem [95] is also a good resource for data on drug metabolism and metabolites. Particularly, the information page of an FDA-approved drug contains information on the metabolism and metabolites of the query drug of interest as text mined from the literature. Furthermore, the NCATS Chemical Genomics Center (NCGC) also deposited high-throughput screening data on PubChem for several CYP450 isoforms consisting of 1A2, 2C9, 2C19, 2D6 and 3A4. The aforementioned databases are public and free to use while there are several more databases that are commercially available and a list of such resources had previously been discussed in a review article by Peach *et al.* [96].

5.1. Predicting CYP450 Inhibition

There have been extensive studies using machine learning approaches for predicting the inhibition of various classes of compounds against several important drug-metabolizing enzymes. Particularly, several QSAR models have been proposed as fast filters that can be used in the initial steps of the discovery process for modeling the inhibition data that are expressed categorically or quantitatively (*i.e.* in terms of IC_{50} or pIC_{50}). One such example is the work of Korhonen *et al.* [97] in which a 3D-QSAR model of IC_{50} values was reported for inhibitors of CYP2B6 inhibitors. Comparative molecular field analysis (CoMFA) was performed as to discern the relationship between electrostatic and steric properties with IC_{50} values. The partial least squares (PLS) method was evaluated by a leave-group-out cross-validation (LGO-CV) scheme. To avoid random bias, calculations were repeated for 20 iterations as to verify the stability of the obtained Q^2 values. The cross-validated PLS model gave a good Q^2 value of 0.607. Three compounds proved to be very potent and selective competitive inhibitors of CYP2B6 *in vitro* namely 4-(4-chlorobenzyl)pyridine (CBP), 4-(4-nitrobenzyl)pyridine (NBP), and 4-benzylpyridine (BP).

Gleeson *et al.* [98] determined the physical characteristics required for the inhibition of CYP 1A2, 2C9, 2C19, 2D6, and 3A4 as a function of a set of interpretable descriptors and modeled by PLS and regression trees (RT). The newfound knowledge could then be applied to guide chemical synthesis. An experimental campaign was undertaken *via* the use of 384 oral drugs and 1,152 compounds as selected from the AstraZeneca compound collection. This set of 1,536 compounds represented a diverse set of compounds spanning many chemotypes. Bulk property and molecular fragments were selected for QSAR model construction. The prediction results revealed that the PLS model afforded Q^2 and RMSE of 0.36 and 0.58, respectively, for CYP3A4 as well as 0.29 and 0.56, respectively, for CYP2C9.

Roy and Roy [99] investigated the inhibitory potencies of 26 naphthalenes and 16 non-naphthalenes against human CYP2A6 and mouse CYP2A5 enzymes using the experimental data reported by Rahnasto *et al.* [100]. These compounds were first represented with quantum-chemical descriptors before further analysis. In order to verify the true predictive power of a QSAR model, the data set was divided into training and test sets (with a ratio of 3:1 for training and testing) using K-means clustering as applied on the standardized descriptor matrix. Genetic function approximation (GFA) and genetic partial least-squares (G/PLS) were used for constructing QSAR models. In regards to the CYP2A6 model, GFA and G/PLS provided Q^2 values of 0.667 and 0.714, respectively. As for the CYP2A5 model, Q^2 values for GFA and G/PLS were 0.914 and 0.869, respectively. In 2012, Gharaghani *et al.* [101] also used the same data set as that used by Roy and Roy [99] to provide understanding on the structural basis of the interactions between CYP2A6 with naphthalene and non-naphthalene. Furthermore, they also investigated the effect of inhibitor on the conformation of the enzyme *via* the use of docking and MD simulation. Molecular descriptors as computed by the Dragon software was used to quantitatively represent the chemical structures of the studied compounds. In this study, multiple linear regression (MLR) and least squares support vector regression (LS-SVR) models were applied for the construction of QSAR models and evaluated using a leave-one-out cross-validation (LOO-CV) scheme, which afforded Q^2 values of 0.695 and 0.728, respectively, for MLR and LS-SVR.

In 2009, furanocoumarin derivatives (FCs) were first used to analyze their pharmacokinetic interactions by Uesawa *et al.* [102] such as increased absorption of various drugs because the constituents exhibited inhibitory effects on drug metabolizing activities of CYP3A. Multiple linear regression analyses in conjunction with a variety of structural, physicochemical, and quantum chemical descriptors were used to create a QSAR model for predicting IC_{50} values. The IC_{50} values were collected from the literature and used as parameters that indicate the CYP3A inhibitory effect of the 37 kinds of FCs [103]. The constructed QSAR models were validated with leave-one-out cross validation and applicable regression diagnostic methods. The predictive results of MLR were Q^2 values of 0.812 (LOO) and 0.775 (bootstrap validation with 100 times). In addition, the simple linear regression analysis for each explanatory descriptor, whereby the descriptors involved in molecular size such as molecular volume (MV), molecular surface area (MA), molecular weight (MW), and heat of formation (E) indicated a good correlation with the log IC_{50} values.

CODESSA was used to obtain the equation relationship between the structural feature of ligand and their inhibitory propency on CYP2D6. Saraceno *et al.* [104] made an effort to develop a QSAR model with a good predictive ability for ligand binding to the very important drug-metabolizing CYP2D6. The initial data set consisted of 51 compounds with known CYP2D6 inhibitors grouped into seven subclasses according to the basis of their structural similarity. Different combinations of training ⁄ test sets were used to develop QSAR models. For example, if a QSAR model is constructed by a training set of 45 molecules, a test will contain 6 molecules. To obtain a more powerful QSAR model, 2D descriptors, 3D descriptors and their combination were considered. The predictive results revealed that QSAR models based solely on 2D molecular descriptors gave unsatisfactory predictions while QSAR models built with 3D descriptors provided good predictive power.

Sridhar *et al.* [105] developed a QSAR model based on their inhouse database of CYP1A2 inhibitor. The 36 molecules with known inhibition activity against CYP1A2 were used. To create a QSAR model, three techniques were used: quantitative comparative molecular field analysis (CoMFA), comparative molecular similarity analysis (CoMSIA), and hologram QSAR (HQSAR). CoMFA utilizes the shapes of the non-covalent fields surrounding the molecules as descriptors, while CoMSIA uses hydrophobic, hydrogen bond acceptor and hydrogen bond donor similarity fields. And, HQSAR uses substructural fragment fingerprints (molecular holograms). The prediction results showed that HQSAR provided the highest Q^2 values of 0.652 (LOO-CV) while the second and third highest performances were obtained from CoMFA (0.667) and CoMSIA (0.616).

5.2. Predicting Drug-Drug Interactions

Several computational approaches exists for the evaluation of drug-drug interactions (DDI). These approaches are divided into network-based methods and structure-activity relationship modeling. However, the limitations (*i.e.* the suppositions of DDIs based on different approaches and usually concern with adverse reactions caused by a single drug) associated with network-based approaches significantly decrease the accuracy of their predictions [106]. On the other hand, QSAR modeling is also used to study the adverse effects associated with DDI. Recently, Zakharov *et al.* [106] employed QSAR models such as HiT QSAR, RF, GUSAR and RBF-SCR for predicting the combinations of drugs likely to cause DDI using binary combinations of marketed drugs. Briefly, 642,411 binary combinations were generated from 1,134 drugs classified by the anatomical therapeutic chemical (ATC) system that were extracted from the DrugBank for construction of a comprehensive DDI data set. The results indicated a sensitivity of 0.9 with the ability to predict clinically unsafe DDIs with a total of 4,500 confirmations [106].

CYP2C9 enzyme plays a major role in the oxidation of both endogenous and xenobiotic compounds, metabolizing over 100 drugs. Owing to high genetic polymorphism of this enzyme, unanticipated changes in the enzyme activity of CYP2C9 can result in toxicity even at therapeutic doses. By the start of the 21st century, the understanding of the CYP2C9 active site and its substrate-inhibitor specificity has steadily progressed through a number of approaches, such as those conducted by Jones *et al.* [107], Mancy *et al.* [108] and Mo *et al.* [109]. The first CoMFA model enabled the construction of a refined homology model for the study of the CYP2C9 active site [110]. Furthermore, Rao *et al.* [111] used a 3D-QSAR model, namely CoMFA, in order to predict potential drug interactions of CYP2C9. The results were interpreted according to the similarities in K_i values and therapeutic concentrations. Hence, for two drugs to potentially interact, their K_i values for CYP2C9 and their therapeutic concentrations must be comparable. The study concluded that majority of the drug interactions were observed to be competitive. Similarly, *Hudelson et al.* utilized four different models namely, LWRP, NERP, Gravity and SUBDUE to predict the DDI of CYP2C9 using 50 test molecules and 11 external validation inhibitors based on binding affinity. The authors chose to use a consensus of 3 out of 4 models in order to accurately predict the binding accuracy, which was 90% and comparable to *in vitro* K_i values [112].

In addition, Lill *et al.* [113] used a docking approach combined with multi-dimensional QSAR for 48 structurally diverse molecules (38 as training set and 10 as test set) for quantification of small-molecule binding (direct or indirect) and possible DDI for CYP34A. Raptor, a 4D-QSAR software, allows representation of ligand molecules as an ensemble of conformations, orientations, stereoisomers and protonation states which decrease the identification bias of bioactive conformers. This approach was applied for the analysis in terms of hydrophobic interactions (ΔG_{HPhob}) and hydrogen-bonding (ΔG_{Hbond}) which indicated that significant contributions were involved in the ligand binding affinity for both of the above parameters. Therefore, the potential for screening new and hypothetical compounds for CYP binding and possible DDI was shown.

As previously stated in this review, the early prediction of ADMET drug properties during a drug discovery and development process has the potential to shorten the time while increasing the chances for identification of a new drug candidate. A key determinant of ADMET, drug metabolism, is known to be involved in various *in vivo* drug processes such as DDI, toxicity and metabolic stability [114]. Manga *et al.* built a simple yet interpretable model using FIRM method using physicochemical descriptors of 96 currently marketed drugs to construct a hierarchical decision tree for the determination of CYP450 enzymes predominantly responsible for a drug's metabolism. The result showed a remarkable 94% accurate classification of compounds [115]. Furthermore, Korhonen *et al.* [97] constructed a 3D-QSAR model (CoMFA) using IC_{50} values that were previously determined. A total of 41 compounds were screened for CYP2B6 inhibition. In addition, the CoMFA model generated was also used to determine the IC_{50} values of compounds not present in the training set. CoMFA correlated electrostatic and steric properties to those obtained by biological experiments. The results indicated a high predictive power of the constructed ComFA model when estimated against the pIC_{50} values, for example, the residuals between predicted and tested pIC_{50} values were 0.08 - 0.33 log units. Thus, the model was able to predict inhibitor potencies of structurally unrelated compounds.

5.3. Predicting Regioselectivity and Metabolites

CYP-mediated metabolism may give rise to toxic intermediates and therefore they are of great importance. Over the past decade, QSAR modeling has seen tremendous growth as indicated by its wide utilization in various scientific disciplines (*e.g.* chemistry, biology, medicine and toxicology). However, predicting metabolites arising from the metabolism process is dependent on several factors that may differ dramatically amongst species as well as within the same population. Therefore, the complexity of metabolism prediction requires a probabilistic approach that shapes the metabolic distribution under specific conditions. One such study conducted by Mekenyan *et al.* [80] uses a tissue metabolism simulator (TIMES), which generates a metabolic map *via* a library of abiotic reactions and biotransformations in order to estimate probabilities. Particularly, best-fit transformation probabilities were then used to determine the metabolic activation of a toxicity pathway.

Furthermore, in order to accurately predict metabolites, defining the sites of metabolism (SoMs) is a good starting point. [78, 116-118]. Although, there are *in vitro* assays for explicitly determining a molecule's SoMs, computational methods are quicker, less expensive, and frequently used in drug development. Thus, there is a need for reliable computational approaches to predict SoMs of a molecule. In the current state of the art, the approaches for the metabolic regioselectivity prediction may be divided into six groups: (i) reactivity-based approaches, (ii) fingerprint-based data mining approaches, (iii) machine learning approaches, (iv) molecular interaction fields, (v) shape-focused approaches and (vi) protein-ligand docking [47, 71, 116].

Many researchers have taken advantage of various fingerprint-based data mining and machine learning approaches for SoM prediction of different CYP450 isoforms with some of the major contributions and drawbacks summarized hereafter. An example of a machine learning approach is MetaPrint2D [77], which is a Java-based SoM predictor. MetaPrint2D was trained on the Accelrys Metabolite database [119, 120], which contains more than 100,000 metabolic transformation. This approach makes SoM prediction (*i.e.* SoM versus non-SoM) based on occurrence counts of atomic fingerprints within the database. If the prediction result is SoM, this approach provides the user with the overall occurrences for all similar transformation data and fingerprints that are stored for each query atom. However, the predictive capability of this approach may be limited and may not extrapolate well beyond its domain of applicability. In 2010, Carlsson *et al.* implemented MetaPrint2D with the graphical user interface of Bioclipse [121]. Similarly, SMARTCyp [60] is a Java-based SoM predictor containing a database of pre-calculated density functional theory (DFT) activation energies for diverse pre-defined ligand fragments. SMARTCyp, a 2D ligand structure-based method, is able to predict SoM reactivity of CYP3A4 and CYP2D6 substrate. SMARTCyp performed well on benchmarking tests of 394 3A4 substrates that provided at least one metabolic site in the top two ranked positions (Top-2 metric) 76% of the time. A few years later, Liu *et al.* [122] developed SMARTCyp as 2D SMARTCyp to predict the metabolic hot spots for CYP3A4, 2D6, 2C9, 2C19, and 1A2 substrates. This extended approach was trained by the impact of a key substrate-receptor recognition feature of each enzyme as a correction term to the SMARTCyp reactivity. For the prediction results, the observed sites of CYP1A2 and CYP2C9 metabolism were among the top-ranked 1, 2, and 3 positions in the range of 67-68%, 80-86% and 83-87%, respectively.

Many type of descriptors of xenobiotic structure were selected as features for constructing predictors due to its importance for SoM prediction [116]. One such example is RegioSelectivity (RS)-predictor [71, 123] which is based on the SVM method to predict SoM. RS-predictor was constructed with a combination of descriptors (*e.g.* 148 topological, 382 quantum chemical and atom-specific descriptors). This approach was reported to provide an acceptable SoM prediction on the comprehensive public data set of CYP substrates and metabolites. Xenosite [61] is one of the machine learning approaches used to predict SoMs by constructing a neural network approach. This approach was learned by using topological, quantum chemical and fingerprint descriptors. The SoM output from Xenosite is represented with the form of the probability of oxidation scores, as opposed to RS-predictor which provides a rank-ordering of the SoMs contained in the same substrate.

Almost all of the above-mentioned SoM predictors are based on information from the 3D structure of the enzyme and/or the quantum-chemical characteristics of the substrate for model construction or prediction. Therefore, there is no SoM predictor that is based solely on 2D structural formulas of substrates. In 2015, Rudik *et al.* [116] proposed a novel approach to predict SOMs of CYP 1A2, 2C9, 2C19, 2D6, and 3A4 by using the PASS (prediction of activity spectra for substances) algorithm incorporated with the 2D structural information and xenobiotic biotransformations catalyzed by P450 isoforms. To create the SoM predictor, this approach did not require a 3D structural information of enzymes and substrates and/or quantum chemical characteristics of the substrates. This allows the approach to be faster than those methods depending on information regarding 3D structures. The experimental results of an average invariant accuracy of prediction (IAP) calculated by the leave-one-out cross-validation (LOO) and external test procedure was 0.9 and 0.95, respectively. All experimental results demonstrated that this proposed method yielded higher accuracies of SoM predictions by RS-Predictor for CYP1A2, CYP2D6, CYP2C9, CYP2C19, and CYP3A4 and is comparable to or better than SMARTCyp for CYP 2C9 and 2D6. In addition, this work is also freely available as a web-server called SOMP at http://www.way2drug.com/SOMP [124]. Tyzack *et al.* took advantage of using simple 2D structural information and xenobiotic biotransformations catalyzed by P450 isoforms [78]. Additionally, a 2D topological fingerprint of atomic sites for the prediction of metabolites was conducted by Tyzack *et al.* [78]. The authors used three probabilistic classifiers namely, NB, PRW and a novel method known as RASCAL (Random Attribute Sampling Classification Algorithm) as applied to a publicly available data of CYP 3A4, 2D6 and 2C9 to predict SoMs. The methods were able to identify SoMs in the top two predictions for 85%, 91% and 88% of the CYP 3A4, 2D6 and 2C9 data sets respectively, using PRW. Similarly, with RASCAL the performance prediction of 83%, 91% and 88%, respectively for CYP 3A4, 2D6 and 2C9 data sets was observed. These results put PRW and RASCAL performance ahead of NB which gave a much lower classification performance of 51%, 73% and 74%, respectively.

Recently, a Microsomal Metabolic Reaction System (MMRS) had been developed by He *et al.* [79] in which they integrated information of SoMs and enzymes. The data was used to predict metabolism as mediated by CYP3A4, 2D6 and 2C9 using various feature selections (*e.g.* CHI, IG, GR) and classification procedures (*e.g.* BN, IBK, RF, J48 and SVM). The authors defined the system formed by the chemical bond data during biotransformation with its corresponding metabolic enzyme as MMRS. This allowed the authors to take into account the relationships between substrate and metabolic enzymes, without the need of crystal models. As a result, 87.7% of potential SoMs were correctly identified by MMRS. Furthermore, the University of Minnesota Biocatalyst/Biodegradation Database (UM-BBD) [125] is a free online database that uses a rule based pathway prediction system called UM-PPS to describe microbial metabolic pathways in more detail. Moreover, Fast Metabolizer (FAME) [63] is a predictor of SoMs based on a collection of RF models trained on a diverse set of small molecule data set of experimental SoMs. As a result, FAME delivers a competitive 85% prediction value for the top-2 and/or top-3 rates.

6. PROTEOCHEMOMETRIC MODELING OF CYP450 INHIBITION

Proteochemometric modeling (PCM) is a bioactivity modeling technique founded on the description of both small molecules (*i.e.* the ligands) and proteins (*i.e.* the targets). By combining the two elements associated with a ligand-target interaction, proteochemometric techniques model the interaction complex or the full ligand-target interaction space. In addition, they are able to quantify the similarities between both ligands and targets simultaneously. Given the high conservation of active site residues among various CYP isoforms, PCM is a highly relevant and potentially useful technique for modeling the complex network of interactions between ligands and various CYP isoforms.

Kontijevskis *et al.* [126] described a novel approach based on the principles of proteochemometrics for the generalized concomitant modeling of multiple CYP isoforms and their inhibitors. They created a predictive and statistically valid proteochemometric SVM model for CYP enzymes by combining data from a large number of publicly available reports that describe the interactions of 14 CYP enzyme subtypes and 375 structurally diverse inhibitors. It was demonstrated that the models had a greater applicability domain than traditional QSARs and are capable of predicting new drug-CYP interactions.

Lapins *et al.* [127] further extended this model by considering a data set with 16,359 compounds associated with five CYP isoforms. The five CYP enzymes were encoded by the alignment-independent descriptors of composition and transition of amino acid properties in the protein primary sequences and 16,359 compounds were encoded by molecular signatures of heights one, two and three computed in the Bioclipse software. Support vector machines and random forest which consider non-linear relationships yielded models with very good predictive performances, having accuracies ranging between 84-88%, and AUC being above 0.9 for both cross-validation and external predictions. These type of generalized PCM models are highly adaptive and can analyze drug interaction data with multiple genetically diverse CYP populations, thus enabling *a priori* predictions of individuals and populations based on their genetic makeup that might respond adversely or even with idiosyncratic drug reactions to drug combinations as well as a drug in development.

7. CURRENT TRENDS AND FUTURE DIRECTIONS

The accuracy of predictive models intrinsically depends on the size and quality of the data that models are based on, and the current open data is unfortunately not as large as commercial databases or those with the Pharma industry. An important future direction is the buildup of large, open, public databases of hiqh quality, and crowdsourcing could be one way of achieving this. But it is also important that the data in such databases contains in-depth information of *e.g.* atom-atom mapping for chemical reactions that can be used for predictive models. The Xenobiotics Metabolism Database (XMetDB) [128] is a recent initiative that provides this, and aims to become a hub for crowd-sourcing efforts. However, so far the uptake has been relatively low, possibly because the experiments underlying such data are time-consuming, and the reported data in literature is not always complete with the necessary details.

The seminal work of Cros [129] on the relationship between the water solubility of a set of primary aliphatic alcohols with their water solubility and their toxicity set the stage for further developments in the field that ultimately led to the founding and coining of modern QSAR by Corwin Hansch [90]. Classical QSAR approaches results in simple and highly interpretable models that are constructed using a few molecular descriptors. Furthermore, such models are based on a congeneric series of compounds against a single target protein. However, advancements in understanding as well as science and technology has brought about the development of more detailed biological screening methods and more sophisticated machine learning algorithms (*e.g.* deep learning), bioactivity datasets (*e.g.* big data) as well as software packages and tools for generating descriptors (*i.e.* that may be or may not be interpretable and readily usable by medicinal chemists).

Polypharmacology is now recognised as a key issue in drug discovery which is driven by multiple protein-ligand interactions. These far surpass the capability of traditional QSAR and this naturally led to the concept of multi-target QSAR and proteochemometric approaches, which consider the interaction space of a series of compounds against a series of target proteins. The wealth of available polypharmacology data in public databases (*e.g.* ChEMBL, Binding DB, *etc.*) are good starting points for continued developments in the field, especially for the development of unified models of several CYP isoforms. It of course should be noted that as more and more public bioactivity data becomes available, the quality and robustness of these modelling approaches increases.

Another important future direction for CYP450 modeling is to provide confidence measures for predictions. In the field of QSAR this has long been overlooked, with the resulting predictions not being trusted by non-expert users. Recently, the discussion of a model's applicability domain [130] has been intensified. The problem stems from the fact that the predictive performance of a QSAR model is based on a finite data set, and the predictive performance is assessed using either a hold out set, an external test set, or by cross-validation. Furthermore, there is an uncertainty at the time of prediction regarding how different a query molecule is from the dataset the model was trained on. If the chemical structure of the query is very different, can the result of the model still be trusted? A new methodology that addresses this is conformal prediction [131], which returns individual values per class (classification) or prediction intervals (regression) which depend on the confidence the user requests in the prediction and a nonconformity measure to devise how similar the query compound is to the data the model is based on. If users requests a higher confidence, the prediction intervals will be larger. It has been proven that conformal predictions are always valid, but the efficiency (*e.g.* prediction interval) naturally depends on the underlying algorithm. The result is an object-based, valid prediction interval and therefore makes the discussion on applicability domain unnecessary. We believe that conformal prediction will be an important step towards establishing trust in predictive models, and that users will start to appreciate the ability to select confidence levels prior to prediction.

Another important direction is to ensure reproducibility of models. The inability to take a published model and run it, or reproduce the study, makes many published models useless to others. We believe it is important to strive towards FAIR (Findable, Accessible, Interoperable, Reproducible) [132], and that models and studies should be FAIR for machines as well as people so that they can be consumed and integrated on-demand. There are initiatives working towards this, such as the QSAR-ML standard [133] and QsarDB [134]. The latter has developed both a data format and a public repository. Developing and publishing models on open data derived from study data and analysis described with metadata according to the FAIR principles, is concomitant with the use of methodologies that report confidence in predictions and will pave the way towards iteratively better science and ultimately better models and more confidence in predictions.

Finally, a key issue facing QSAR practitioners is how to balance the need for highly predictive models over those that are easily interpretable. Recent modeling studies have focused on sophisticated machine learning methods as well as making use of large quantities of molecular descriptors, which may be statistically robust but practically may provide little guidance for medicinal chemists. Thus, it is recommended that, where possible, practitioners should make use of simple and interpretable descriptors and if possible employ the so-called "white box" learning methods (*e.g.* decision trees, MLR, PLS, *etc.*) so as to obtain models that are biologist and chemist friendly. If descriptors are used that can be interpreted as chemical substructures, it is possible to highlight them based on their contributions to the prediction [135]. Studies have shown that this can be done even when using non-linear methods [136, 137] and using conformal prediction [138].

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This research is supported by a Research Career Development Grant (No. RSA5780031) from the Thailand Research Fund to CN and the Swedish Research Links program (No. C0610701) from the Swedish Research Council to JESW and CN. Partial support from the New Scholar Research Grant (No. MRG5980220) to WS is also acknowledged.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Kirchmair, J.; Göller, A.H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I.D.; Glen, R.C.; Schneider, G. Predicting drug metabolism: Experiment and/or computation? *Nat. Rev. Drug. Discov.*, **2015**, *14*, 387-404.

[2] Iyanagi, T. Molecular mechanism of phase I and phase II drug-metabolizing enzymes: implications for detoxification. *Int. Rev. Cytol.*, **2007**, *260*, 35-112.

[3] McKinnon, R.A.; Sorich, M.J.; Ward, M.B. Cytochrome P450 part 1: Multiplicity and function. *J. Pharm. Prac. Res.*, **2008**, *38*, 55-57.

[4] Gunaratna, C. Drug metabolism & pharmacokinetics in drug discovery: A primer for bioanalytical chemists, Part I. *Curr. Sep.*, **2000**, *19*, 17-23.

[5] Ogu, C.C.; Maxa, J.L. Drug interactions due to cytochrome P450. *Proc. (Bayl. Univ. Med. Cent.)*, **2000**, *13*, 421-423.

[6] Nebert, D.W.; Wikvall, K.; Miller, W.L. Human cytochromes P450 in health and disease. *Philos. Trans. R. Soc. Lond., B., Biol. Sci.*, **2013**, *368*, 20120431.

[7] Guengerich, F.P.; Waterman, M.R.; Egli, M. Recent structural insights into cytochrome P450 function. *Trends Pharmacol. Sci.*, **2016**, *37*, 625-640.

[8] Werck-Reichhart, D.; Feyereisen, R. Cytochromes P450: A success story. *Genome Biol.*, **2000**, *1*, Reviews3003.

[9] Nelson, D.R. Cytochrome P450 nomenclature, 2004. *Methods Mol. Biol.*, **2006**, *320*, 1-10.

[10] Tanaka, E. Clinically important pharmacokinetic drug-drug interactions: Role of cytochrome P450 enzymes. *J. Clin. Pharm. Ther.*, **1998**, *23*, 403-416.

[11] Urlacher, V.B.; Girhard, M. Cytochrome P450 monooxygenases: An update on perspectives for synthetic application. *Trends Biotechnol.*, **2012**, *30*, 26-36.

[12] Belpaire, F.M.; Bogaert, M.G. Cytochrome P450: Genetic polymorphism and drug interactions. *Acta Clin. Belg.*, **1996**, *51*, 254-260.

[13] Pandey, A.V.; Flück, C.E. NADPH P450 oxidoreductase: Structure, function, and pathology of diseases. *Pharmacol. Ther.*, **2013**, *138*, 229-254.

[14] Vaiopoulou, A.; Gazouli, M.; Karikas, G.A. Pharmacogenomics: Current applications and future prospects towards personalized therapeutics. *JBUON*, **2013**, *18*, 570-578.

[15] Zanger, U.M.; Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.*, **2013**, *138*, 103-141.

[16] Tralau, T.; Luch, A. "Drugs on oxygen": An update and perspective on the role of cytochrome P450 testing in pharmacology. *Expert Opin. Drug Metab. Toxicol.*, **2012**, *8*, 1357-1362.

[17] Wienkers, L.C.; Heath, T.G. Predicting *in vivo* drug interactions from *in vitro* drug discovery data. *Nat. Rev. Drug Discov.*, **2005**, *4*, 825-833.

[18] Johansson, I.; Ingelman-Sundberg, M. Genetic polymorphism and toxicology-with emphasis on cytochrome P450. *Toxicol. Sci.*, **2011**, *120*, 1-13.

[19] Biss, T.T.; Avery, P.J.; Williams, M.D.; Brandao, L.R.; Grainger, J.D.; Kamali, F. VKORC1 and CYP2C9 genotype is associated with over-anticoagulation during initiation of warfarin therapy in children. *J. Thromb. Haemost.*, **2012**, 11(2), 373-375.

[20] Lin, J.H.; Lu, A.Y. Inhibition and induction of cytochrome P450 and the clinical implications. *Clin. Pharmacokinet.*, **1998**, *35*, 361-390.

[21] Lynch, T.; Price, A. The efffect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am. Fam. Physician*, **2007**, *76*, 391-396.

[22] Prachayasittikul, V.; Worachartcheewan, A.; Shoombuatong, W.; Songtawee, N.; Simeon, S.; Prachayasittikul, V.; Nantasenamat, C. Computer-aided drug design of bioactive natural products. *Curr. Top. Med. Chem.*, **2015**, *15*, 1780-1800.

[23] Kang, J.S.; Lee, M.H. Overview of therapeutic drug monitoring. *Korean J. Intern. Med.*, **2009**, *24*, 1-10.

[24] Niel, N.; Rechencq, E.; Muller, A.; Vidal, J.P.; Escale, R.; Durand, T.; Girard, J.P.; Rossi, J.C.; Bonne, C. Synthesis and contractile activity of new pseudopeptido and thioaromatic analogues of leukotriene D4. *Prostaglandins*, **1992**, *43*, 45-54.

[25] Hollenberg, P.F. Characteristics and common properties of inhibitors, inducers, and activators of CYP enzymes. *Drug Metab. Rev.*, **2002**, *34*, 17-35.

[26] Lynch, T.; Price, A. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am. Fam. Physician*, **2007**, *76*, 391-396.

[27] Dresser, G.K.; Spence, J.D.; Bailey, D.G. Pharmacokinetic-pharmacodynamic consequences and clinical relevance of cytochrome P450 3A4 inhibition. *Clin. Pharmacokinet.*, **2000**, *38*, 41-57.

[28] Sinz, M.; Wallace, G.; Sahi, J. Current industrial practices in assessing CYP450 enzyme induction: Preclinical and clinical. *AAPS J.*, **2008**, *10*, 391-400.

[29] Davydov, D.R.; Halpert, J.R.; Renaud, J.-P.; Hui Bon Hoa, G. Conformational heterogeneity of cytochrome P450 3A4 revealed by high pressure spectroscopy. *Biochem. Biophys. Res. Commun.*, **2003**, *312*, 121-130.

[30] Rodriguez-Antona, C.; Ingelman-Sundberg, M. Cytochrome P450 pharmacogenetics and cancer. *Oncogene*, **2006**, *25*, 1679-1691.

[31] Rooseboom, M.; Commandeur, J.N.M.; Vermeulen, N.P.E. Enzyme-catalyzed activation of anticancer prodrugs. *Pharmacol. Rev.*, **2004**, *56*, 53-102.

[32] Oyama, T.; Kagawa, N.; Kunugita, N.; Kitagawa, K.; Ogawa, M.; Yamaguchi, T.; Suzuki, R.; Kinaga, T.; Yashima, Y.; Ozaki, S.; Isse, T.; Kim, Y.-D.; Kim, H.; Kawamoto, T. Expression of cytochrome P450 in tumor tissues and its association with cancer development. *Front. Biosci.*, **2004**, *9*, 1967-1976.

[33] Van den Anker, J.N.; Schwab, M.; Kearns, G.L. Developmental pharmacokinetics, In: *Pediatric Clinical Pharmacology, Handbook of Experimental Pharmacology; Seyberth, H.W.; Rane, A.; Schwab, M.; Eds.*, **2011**, Vol. *205*, pp. 51-75.

[34] Korobkova, E.A. Effect of natural polyphenols on CYP metabolism: Implications for diseases. *Chem. Res. Toxicol.*, **2015**, *28*, 1359-1390.

[35] Aspromonte, N.; Monitillo, F.; Puzzovivo, A.; Valle, R.; Caldarola, P.; Iacoviello, M. Modulation of cardiac cytochrome P450 in patients with heart failure. *Expert Opin. Drug Metab. Toxicol.*, **2014**.

[36] Wanwimolruk, S.; Prachayasittikul, V. Cytochrome P450 enzyme mediated herbal drug interactions (Part 1). *EXCLI J.*, **2014**, *13*, 347-391.

[37] Bailey, D.G.; Dresser, G.K.; Leake, B.F.; Kim, R.B. Naringin is a major and selective clinical inhibitor of organic anion-transporting polypeptide 1A2 (OATP1A2) in grapefruit juice. *Clin. Pharmacol. Ther.*, **2007**, *81*, 495-502.

[38] Bailey, D.G.; Malcolm, J.; Arnold, O.; Spence, J.D. Grapefruit juice-drug interactions. *Br. J. Clin. Pharmacol.*, **1998**, *46*(2), 101-110.

[39] Murray, M. Altered CYP expression and function in response to dietary factors: Potential roles in disease pathogenesis. *Curr. Drug. Metab.*, **2006**, *7*, 67-81.

[40] Chan, W.K.; Delucchi, A.B. Resveratrol, a red wine constituent, is a mechanism-based inactivator of cytochrome P450 3A4. *Life Sci.*, **2000**, *67*, 3103-3112.

[41] Flockhart, D.A.; Oesterheld, J.R. Cytochrome P450-mediated drug interactions. *Child Adolesc. Psychiatr. Clin. N. Am.*, **2000**, *9*, 43-76.

[42] Flockhart, D.A. Drug Interactions: Cytochrome P450 Drug Interaction Table http://medicine.iupui.edu/clinpharm/ddis/ (Accessed Dec 13, **2016**).

[43] Roy, K.; Roy, P.P. QSAR of cytochrome inhibitors. *Expert Opin. Drug. Metab. Toxicol.*, **2009**, *5*, 1245-1266.

[44] Zakrzewski-Jakubiak, H.; Doan, J.; Lamoureux, P.; Singh, D.; Turgeon, J.; Tannenbaum, C. Detection and prevention of drug-drug interactions in the hospitalized elderly: Utility of new cytochrome P450-based software. *Am. J. Geriatr. Pharmacother.*, **2011**, *9*, 461-470.

[45] Raunio, H.; Kuusisto, M.; Juvonen, R.O.; Pentikäinen, O.T. Modeling of interactions between xenobiotics and cytochrome P450 (CYP) enzymes. *Front. Pharmacol.*, **2015**, *6*, 123.

[46] Roncaglioni, A.; Toropov, A.A.; Toropova, A.P.; Benfenati, E. *In silico* methods to predict drug toxicity. *Curr. Opin. Pharmacol.*, **2013**, *13*, 802-806.

[47] Kirchmair, J.; Williamson, M.J.; Tyzack, J.D.; Tan, L.; Bond, P.J.; Bender, A.; Glen, R.C. Computational prediction of metabolism: Sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.*, **2012**, *52*, 617-648.

[48] Sridhar, J.; Liu, J.; Foroozesh, M.; Stevens, C.L.K. Insights on cytochrome P450 enzymes and inhibitors obtained through QSAR studies. *Molecules*, **2012**, *17*, 9283-9305.

[49] BIOVIA. BIOVIA Metabolite http://accelrys.com/products/collaborative-science/databases/bioactivity-databases/biovia-metabolite.html (Accessed Nov 23, **2016**).

[50] Fujitsu Kyushu Systems. Fujitsu ADME Database http://www.fujitsu.com/jp/group/kyushu/en/solutions/industry/lifescience/admedatabase/ (Accessed Nov 23, **2016**).

[51] Sánchez-Gómez, S.; Japelj, B.; Jerala, R.; Moriyón, I.; Fernández Alonso, M.; Leiva, J.; Blondelle, S.E.; Andrä, J.; Brandenburg, K.; Lohner, K.; Martínez de Tejada, G. Structural features governing the activity of lactoferricin-derived peptides that act in synergy with antibiotics against *Pseudomonas aeruginosa in vitro* and *in vivo*. *Antimicrob. Agents Chemother.*, **2011**, *55*, 218-228.

[52] Wendt, B.; Mulbaier, M.; Wawro, S.; Schultes, C.; Alonso, J.; Janssen, B.; Lewis, J. Toluidinesulfonamide hypoxia-induced factor 1 inhibitors: Alleviating drug-drug interactions through use of PubChem data and comparative molecular field analysis guided synthesis. *J. Med. Chem.*, **2011**, *54*, 3982-3986.

[53] Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR analysis for chemical structure generation (from y to x). *J. Chem. Inf. Model.*, **2016**, *56*, 286-299.

[54] Nicolotti, O.; Gillet, V.J.; Fleming, P.J.; Green, D.V.S. Multiobjective optimization in quantitative structure-activity relationships: deriving accurate and interpretable QSARs. *J. Med. Chem.*, **2002**, *45*, 5069-5080.

[55] Worachartcheewan, A.; Mandi, P.; Prachayasittikul, V.; Toropova, A.P.; Toropov, A.A.; Nantasenamat, C. Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. *Chemometr. Intell. Lab. Sys.*, **2014**, *138*, 120-126.

[56] Worachartcheewan, A.; Prachayasittikul, V.; Toropova, A.P.; Toropov, A.A.; Nantasenamat, C. Large-scale structure-activity relationship study of hepatitis C virus NS5B polymerase inhibition using SMILES-based descriptors. *Mol. Divers.*, **2015**, *19*, 955-964.

[57] Toropov, A.A.; Toropova, A.P.; Raska, I.; Benfenati, E.; Gini, G. QSAR modeling of endpoints for peptides which is based on representation of the molecular structure by a sequence of amino acids. *Struct. Chem.*, **2012**, *23*, 1891-1904.

[58] Toropova, M.A.; Veselinović, A.M.; Veselinović, J.B.; Stojanović, D.B.; Toropov, A.A. QSAR modeling of the antimicrobial activity of peptides as a mathematical function of a sequence of amino acids. *Comput. Biol. Chem.*, **2015**, *59 Pt A*, 126-130.

[59] Toropov, A.A.; Toropova, A.P.; Raska, I.; Leszczynska, D.; Leszczynski, J. Comprehension of drug toxicity: Software and databases. *Comput. Biol. Med.*, **2014**, *45*, 20-25.

[60] Rydberg, P.; Gloriam, D.E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett.*, **2010**, *1*, 96-100.

[61] Zaretzki, J.; Matlock, M.; Swamidass, S.J. XenoSite: Accurately predicting CYP-mediated sites of metabolism with neural networks. *J. Chem. Inf. Model.*, **2013**, *53*, 3373-3383.

[62] Matlock, M.K.; Hughes, T.B.; Swamidass, S.J. XenoSite server: A web-available site of metabolism prediction tool. *Bioinformatics*, **2015**, *31*, 1136-1137.

[63] Kirchmair, J.; Williamson, M.J.; Afzal, A.M.; Tyzack, J.D.; Choy, A.P.K.; Howlett, A.; Rydberg, P.; Glen, R.C. FAst MEtabolizer (FAME): A rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes. *J. Chem. Inf. Model.*, **2013**, *53*, 2896-2907.

[64] Optibrium Ltd. *StarDrop*; Optibrium: Cambridge, UK, **2016**.

[65] Tyzack, J.D.; Hunt, P.A.; Segall, M.D. Predicting regioselectivity and lability of cytochrome P450 metabolism using quantum me-

chanical simulations. *J. Chem. Inf. Model.*, **2016**, 56(11), 2180-2193.

[66] Ekins, S.; Andreyev, S.; Ryabov, A.; Kirillov, E.; Rakhmatulin, E.A.; Bugrim, A.; Nikolskaya, T. Computational prediction of human drug metabolism. *Expert Opin. Drug Metab. Toxicol.*, **2005**, *1*, 303-324.

[67] Testa, B.; Balmat, A.-L.; Long, A. Predicting drug metabolism: Concepts and challenges. *Pure Appl. Chem.*, **2004**, *76, 907-914*.

[68] Testa, B.; Balmat, A.-L.; Long, A.; Judson, P. Predicting drug metabolism - An evaluation of the expert system METEOR. *Chem. Biodivers.*, **2005**, *2*, 872-885.

[69] Rydberg, P.; Gloriam, D.E.; Olsen, L. The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics*, **2010**, *26*, 2988-2989.

[70] Hennemann, M.; Friedl, A.; Lobell, M.; Keldenich, J.; Hillisch, A.; Clark, T.; Göller, A.H. CypScore: Quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory. *ChemMedChem*, **2009**, *4*, 657-669.

[71] Zaretzki, J.; Rydberg, P.; Bergeron, C.; Bennett, K.P.; Olsen, L.; Breneman, C.M. RS-predictor models augmented with SMARTCyp reactivities: Robust metabolic regioselectivity predictions for nine CYP isozymes. *J. Chem. Inf. Model.*, **2012**, *52*, 1637-1659.

[72] Korzekwa, K.R.; Jones, J.P.; Gillette, J.R. Theoretical Studies on Cytochrome P-450 mediated hydroxylation: A predictive model for hydrogen atom abstractions. *J. Am. Chem. Soc.*, **1990**, *112*, 7042-7046.

[73] De Groot, M.J.; Ackland, M.J.; Horne, V.A.; Alex, A.A.; Jones, B.C. Novel approach to predicting P450-mediated drug metabolism: Development of a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.*, **1999**, *42*, 1515-1524.

[74] De Groot, M.J.; Alex, A.A.; Jones, B.C. Development of a combined protein and pharmacophore model for cytochrome P450 2C9. *J. Med. Chem.*, **2002**, *45*, 1983-1993.

[75] Tyzack, J.D.; Williamson, M.J.; Torella, R.; Glen, R.C. Prediction of cytochrome P450 xenobiotic metabolism: Tethered docking and reactivity derived from ligand molecular orbital analysis. *J. Chem. Inf. Model.*, **2013**, *53*, 1294-1305.

[76] Crivori, P.; Zamora, I.; Speed, B.; Orrenius, C.; Poggesi, I. Model based on GRID-derived descriptors for estimating CYP3A4 enzyme stability of potential drug candidates. *J. Comput. Aided. Mol. Des.*, **2004**, *18*, 155-166.

[77] Boyer, S.; Arnby, C.H.; Carlsson, L.; Smith, J.; Stein, V.; Glen, R.C. Reaction site mapping of xenobiotic biotransformations. *J. Chem. Inf. Model.*, **2007**, *47*, 583-590.

[78] Tyzack, J.D.; Mussa, H.Y.; Williamson, M.J.; Kirchmair, J.; Glen, R.C. Cytochrome P450 site of metabolism prediction from 2D topological fingerprints using GPU accelerated probabilistic classifiers. *J. Cheminform.*, **2014**, *6*, 29.

[79] He, S.-B.; Li, M.-M.; Zhang, B.-X.; Ye, X.-T.; Du, R.-F.; Wang, Y.; Qiao, Y.-J. Construction of metabolism prediction models for CYP450 3A4, 2D6, and 2C9 based on microsomal metabolic reaction system. *Int. J. Mol. Sci.*, **2016**, *17*.

[80] Mekenyan, O.G.; Dimitrov, S.D.; Pavlov, T.S.; Veith, G.D. A systematic approach to simulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework. *Curr. Pharm. Des.*, **2004**, *10*, 1273-1293.

[81] Olsen, L.; Rydberg, P.; Rod, T.H.; Ryde, U. Prediction of activation energies for hydrogen abstraction by cytochrome P450. *J. Med. Chem.*, **2006**, *49*, 6489-6499.

[82] Lewis, D.F.V.; Modi, S.; Dickins, M. Structure-activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metab. Rev.*, **2002**, *34*, 69-82.

[83] Ekins, S.; Stresser, D.M.; Andrew Williams, J. *In vitro* and pharmacophore insights into CYP3A enzymes. *Trends Pharmacol. Sci.*, **2003**, *24*, 161-166.

[84] Rostkowski, M.; Spjuth, O.; Rydberg, P. WhichCyp: Prediction of cytochromes P450 inhibition. *Bioinformatics*, **2013**, *29*, 2051-2052.

[85] Shahrokh, K.; Cheatham, T.E.; Yost, G.S. Conformational dynamics of CYP3A4 demonstrate the important role of Arg212 coupled with the opening of ingress, egress and solvent channels to dehydrogenation of 4-hydroxy-tamoxifen. *Biochim. Biophys. Acta*, **2012**, *1820*, 1605-1617.

[86] Sgrignani, J.; Bon, M.; Colombo, G.; Magistrato, A. Computational approaches elucidate the allosteric mechanism of human aromatase inhibition: A novel possible route to small-molecule regulation of CYP450s activities? *J. Chem. Inf. Model.*, **2014**, *54*, 2856-2868.

[87] Fernald, G.H.; Altman, R.B. Using molecular features of xenobiotics to predict hepatic gene expression response. *J. Chem. Inf. Model.*, **2013**, *53*, 2765-2773.

[88] Khandelwal, A.; Krasowski, M.D.; Reschly, E.J.; Sinz, M.W.; Swaan, P.W.; Ekins, S. Machine learning methods and docking for predicting human pregnane X receptor activation. *Chem. Res. Toxicol.*, **2008**, *21*, 1457-1467.

[89] Yamashita, F.; Sasa, Y.; Yoshida, S.; Hisaka, A.; Asai, Y.; Kitano, H.; Hashida, M.; Suzuki, H. Modeling of Rifampicin-induced CYP3A4 activation dynamics for the prediction of clinical drug-drug interactions from *in vitro* data. *PLoS ONE*, **2013**, *8*, *8*, e70330.

[90] Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. A practical overview of quantitative structure-activity relationship. *EXCLI J.*, **2009**, *8*, 74-88.

[91] Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Prachayasittikul, V. Advances in computational methods to predict the biological activity of compounds. *Expert Opin. Drug. Discov.*, **2010**, *5*, 633-654.

[92] Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J.P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **2012**, *40*, D1100-7.

[93] Davies, M.; Dedman, N.; Hersey, A.; Papadatos, G.; Hall, M.D.; Cucurull-Sanchez, L.; Jeffrey, P.; Hasan, S.; Eddershaw, P.J.; Overington, J.P. ADME SARfari: Comparative genomics of drug metabolizing systems. *Bioinformatics*, **2015**, *31*, 1695-1697.

[94] Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A.C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z.T.; Han, B.; Zhou, Y.; Wishart, D.S. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.*, **2014**, *42*, D1091-7.

[95] Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S.H. PubChem substance and compound databases. *Nucleic Acids Res.*, **2016**, *44*, D1202-13.

[96] Peach, M.L.; Zakharov, A.V.; Liu, R.; Pugliese, A.; Tawa, G.; Wallqvist, A.; Nicklaus, M.C. Computational tools and resources for metabolism-related property predictions. 1. Overview of publicly available (free and commercial) databases and software. *Future Med. Chem.*, **2012**, *4*, 1907-1932.

[97] Korhonen, L.E.; Turpeinen, M.; Rahnasto, M.; Wittekindt, C.; Poso, A.; Pelkonen, O.; Raunio, H.; Juvonen, R.O. New potent and selective cytochrome P450 2B6 (CYP2B6) inhibitors based on three-dimensional quantitative structure-activity relationship (3D-QSAR) analysis. *Br. J. Pharmacol.*, **2007**, *150*, 932-942.

[98] Gleeson, M.P.; Davis, A.M.; Chohan, K.K.; Paine, S.W.; Boyer, S.; Gavaghan, C.L.; Arnby, C.H.; Kankkonen, C.; Albertson, N. Generation of *in-silico* cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models. *J. Comput. Aided Mol. Des.*, **2007**, *21*, 559-573.

[99] Roy, K.; Roy, P.P. Exploring QSAR and QAAR for inhibitors of cytochrome P450 2A6 and 2A5 enzymes using GFA and G/PLS techniques. *Eur. J. Med. Chem.*, **2009**, *44*, 1941-1951.

[100] Rahnasto, M.; Raunio, H.; Poso, A.; Wittekindt, C.; Juvonen, R.O. Quantitative structure-activity relationship analysis of inhibitors of the nicotine metabolizing CYP2A6 enzyme. *J. Med. Chem.*, **2005**, *48*, 440-449.

[101] Gharaghani, S.; Khayamian, T.; Keshavarz, F. Docking, molecular dynamics simulation studies, and structure-based QSAR model on cytochrome P450 2A6 inhibitors. *Struct. Chem.*, **2012**, *23*, 341-350.

[102] Uesawa, Y.; Mohri, K. Quantitative structure-activity relationship (QSAR) analysis of the inhibitory effects of furanocoumarin derivatives on cytochrome P450 3A activities. *Pharmazie*, **2010**, *65*, 41-46.

[103] Guo, L.Q.; Taniguchi, M.; Xiao, Y.Q.; Baba, K.; Ohta, T.; Yamazoe, Y. Inhibitory effect of natural furanocoumarins on human microsomal cytochrome P450 3A activity. *Jpn. J. Pharmacol.*, **2000**, *82*, 122-129.

[104] Saraceno, M.; Massarelli, I.; Imbriani, M.; James, T.L.; Bianucci, A.M. Optimizing QSAR models for predicting ligand binding to the drug-metabolizing cytochrome P450 isoenzyme CYP2D6. *Chem. Biol. Drug. Des.*, **2011**, *78*, 236-251.

[105] Sridhar, J.; Foroozesh, M.; Stevens, C.L.K. QSAR models of cytochrome P450 enzyme 1A2 inhibitors using CoMFA, CoMSIA and HQSAR. *SAR QSAR Environ. Res.*, **2011**, *22*, 681-697.

[106] Zakharov, A.V.; Varlamova, E.V.; Lagunin, A.A.; Dmitriev, A.V.; Muratov, E.N.; Fourches, D.; Kuz'min, V.E.; Poroikov, V.V.; Tropsha, A.; Nicklaus, M.C. QSAR modeling and prediction of drug-drug interactions. *Mol. Pharm.*, **2016**, *13*, 545-556.

[107] Jones, B.C.; Hawksworth, G.; Horne, V.A.; Newlands, A.; Morsman, J.; Tute, M.S.; Smith, D.A. Putative active site template model for cytochrome P4502C9 (tolbutamide hydroxylase). *Drug Metab. Dispos.*, **1996**, *24*, 260-266.

[108] Mancy, A.; Broto, P.; Dijols, S.; Dansette, P.M.; Mansuy, D. The substrate binding site of human liver cytochrome P450 2C9: An approach using designed tienilic acid derivatives and molecular modeling. *Biochemistry*, **1995**, *34*, 10365-10375.

[109] Mo, S.-L.; Zhou, Z.-W.; Yang, L.-P.; Wei, M.Q.; Zhou, S.-F. New insights into the structural features and functional relevance of human cytochrome P450 2C9. Part I. *Curr. Drug Metab.*, **2009**, *10*, 1075-1126.

[110] Jones, J.P.; He, M.; Trager, W.F.; Rettie, A.E. Three-dimensional quantitative structure-activity relationship for inhibitors of cytochrome P4502C9. *Drug Metab. Dispos.*, **1996**, *24*, 1-6.

[111] Rao, S.; Aoyama, R.; Schrag, M.; Trager, W.F.; Rettie, A.; Jones, J.P. A refined 3-dimensional QSAR of cytochrome P450 2C9: Computational predictions of drug interactions. *J. Med. Chem.*, **2000**, *43*, 2789-2796.

[112] Hudelson, M.G.; Ketkar, N.S.; Holder, L.B.; Carlson, T.J.; Peng, C.-C.; Waldher, B.J.; Jones, J.P. High confidence predictions of drug-drug interactions: Predicting affinities for cytochrome P450 2C9 with multiple computational methods. *J. Med. Chem.*, **2008**, *51*, 648-654.

[113] Lill, M.A.; Dobler, M.; Vedani, A. Prediction of small-molecule binding to cytochrome P450 3A4: Flexible docking combined with multidimensional QSAR. *ChemMedChem*, **2006**, *1*, 73-81.

[114] Li, H.; Sun, J.; Fan, X.; Sui, X.; Zhang, L.; Wang, Y.; He, Z. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J. Comput. Aided Mol. Des.*, **2008**, *22*, 843-855.

[115] Manga, N.; Duffy, J.C.; Rowe, P.H.; Cronin, M.T.D. Structure-based methods for the prediction of the dominant P450 enzyme in human drug biotransformation: Consideration of CYP3A4, CYP2C9, CYP2D6. *SAR QSAR Environ. Res.*, **2005**, *16*, 43-61.

[116] Rudik, A.V.; Dmitriev, A.V.; Lagunin, A.A.; Filimonov, D.A.; Poroikov, V.V. Metabolism site prediction based on xenobiotic structural formulas and PASS prediction algorithm. *J. Chem. Inf. Model.*, **2014**, *54*, 498-507.

[117] Crivori, P.; Poggesi, I. Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. *Eur. J. Med. Chem.*, **2006**, *41*, 795-808.

[118] Susnow, R.G.; Dixon, S.L. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1308-1315.

[119] Borodina, Y.; Rudik, A.; Filimonov, D.; Kharchevnikova, N.; Dmitriev, A.; Blinova, V.; Poroikov, V. A new statistical approach to predicting aromatic hydroxylation sites. Comparison with model-based approaches. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1998-2009.

[120] Borodina, Y.; Sadym, A.; Filimonov, D.; Blinova, V.; Dmitriev, A.; Poroikov, V. Predicting biotransformation potential from molecular structure. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1636-1646.

[121] Carlsson, L.; Spjuth, O.; Adams, S.; Glen, R.C.; Boyer, S. Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinformatics*, **2010**, *11*, 362.

[122] Liu, R.; Liu, J.; Tawa, G.; Wallqvist, A. 2D SMARTCyp reactivity-based site of metabolism prediction for major drug-metabolizing cytochrome P450 enzymes. *J. Chem. Inf. Model.*, **2012**, *52*, 1698-1712.

[123] Zaretzki, J.; Bergeron, C.; Huang, T.; Rydberg, P.; Swamidass, S.J.; Breneman, C.M. RS-WebPredictor: A server for predicting CYP-mediated sites of metabolism on drug-like molecules. *Bioinformatics*, **2013**, *29*, 497-498.

[124] Rudik, A.; Dmitriev, A.; Lagunin, A.; Filimonov, D.; Poroikov, V. SOMP: Web server for *in silico* prediction of sites of metabolism for drug-like compounds. *Bioinformatics*, **2015**.

[125] Gao, J.; Ellis, L.B.M.; Wackett, L.P. The University of Minnesota Biocatalysis/Biodegradation database: Improving public access. *Nucleic Acids Res.*, **2010**, *38*, D488-91.

[126] Kontijevskis, A.; Komorowski, J.; Wikberg, J.E.S. Generalized proteochemometric model of multiple cytochrome P450 enzymes and their inhibitors. *J. Chem. Inf. Model.*, **2008**, *48*, 1840-1850.

[127] Lapins, M.; Worachartcheewan, A.; Spjuth, O.; Georgiev, V.; Prachayasittikul, V.; Nantasenamat, C.; Wikberg, J.E.S. A unified proteochemometric model for prediction of inhibition of cytochrome P450 isoforms. *PLoS ONE*, **2013**, *8*, e66566.

[128] Spjuth, O.; Rydberg, P.; Willighagen, E.L.; Evelo, C.T.; Jeliazkova, N. XMetDB: An open access database for xenobiotic metabolism. *J. Cheminform.*, **2016**, *8*, 47.

[129] Cros, A.F.A. Action de L'alcohol Amyliquesur L'organisme. Doctoral dissertation, **1863**.

[130] Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.*, **2005**, *45*, 839-849.

[131] Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *J. Chem. Inf. Model.*, **2014**, *54*, 1596-1603.

[132] Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; Bouwman, J.; Brookes, A.J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C.T.; Finkers, R.; Gonzalez-Beltran, A.; Mons, B. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, **2016**, *3*, 160018.

[133] Spjuth, O.; Willighagen, E.L.; Guha, R.; Eklund, M.; Wikberg, J.E. Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J. Cheminform.*, **2010**, *2*, 5.

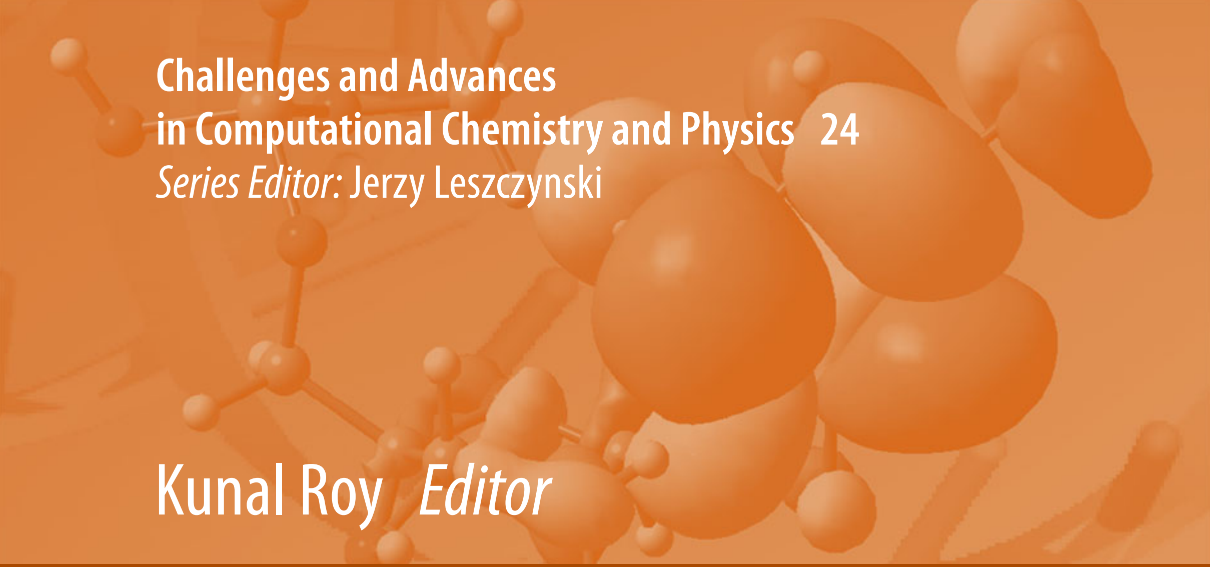
[134] Ruusmann, V.; Sild, S.; Maran, U. QSAR DataBank Repository: Open and linked qualitative and quantitative structure-activity relationship models. *J. Cheminform.*, **2015**, *7*, 32.

[135] Spjuth, O.; Eklund, M.; Ahlberg Helgee, E.; Boyer, S.; Carlsson, L. Integrated decision support for assessing chemical liabilities. *J. Chem. Inf. Model.*, **2011**, *51*, 1840-1847.

[136] Carlsson, L.; Helgee, E.A.; Boyer, S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J. Chem. Inf. Model.*, **2009**, *49*, 2551-2558.

[137] Rosenbaum, L.; Hinselmann, G.; Jahn, A.; Zell, A. Interpreting linear support vector machine models with heat map molecule coloring. *J. Cheminform*, **2011**, *3*, 11.

[138] Ahlberg, E.; Spjuth, O.; Hasselgren, C.; Carlsson, L. Interpretation of conformal prediction classification models. In: *Statistical Learning and Data Sciences*; Gammerman, A.; Vovk, V.; Papadopoulos, H., Eds.; Springer International Publishing: Cham, Switzerland, **2015**; Vol. 9047, pp. 323-334.



Challenges and Advances
in Computational Chemistry and Physics 24
Series Editor: Jerzy Leszczynski

Kunal Roy *Editor*

Advances in QSAR Modeling

Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences

Springer

Contents

Part I Theory

Towards the Revival of Interpretable QSAR Models 3
Watshara Shoombuatong, Philip Prathipati, Wiwat Owasirikul, Apilak Worachartcheewan, Saw Simeon, Nuttapat Anuwongcharoen, Jarl E.S. Wikberg and Chanin Nantasenamat

The Use of Topological Indices in QSAR and QSPR Modeling. 57
John C. Dearden

Which Performance Parameters Are Best Suited to Assess the Predictive Ability of Models? 89
Károly Héberger, Anita Rácz and Dávid Bajusz

Part II Methods

Structural, Physicochemical and Stereochemical Interpretation of QSAR Models Based on Simplex Representation of Molecular Structure 107
P. Polishchuk, E. Mokshyna, A. Kosinskaya, A. Muats, M. Kulinsky, O. Tinkov, L. Ognichenko, T. Khristova, A. Artemenko and V. Kuz'min

The Maximum Common Substructure (MCS) Search as a New Tool for SAR and QSAR 149
Azadi Golbamaki, Alessio Mauro Franchi and Giuseppina Gini

Generative Topographic Mapping Approach to Chemical Space Analysis. 167
Dragos Horvath, Gilles Marcou and Alexandre Varnek

Part III Applications

On Applications of QSARs in Food and Agricultural Sciences: History and Critical Review of Recent Developments 203
Supratik Kar, Kunal Roy and Jerzy Leszczynski

Quantitative Structure-Epigenetic Activity Relationships. 303
Mario Omar García-Sánchez, Maykel Cruz-Monteagudo
and José L. Medina-Franco

QSAR/QSPR Modeling in the Design of Drug Candidates with Balanced Pharmacodynamic and Pharmacokinetic Properties 339
George Lambrinidis, Fotios Tsopelas, Costas Giaginis
and Anna Tsantili-Kakoulidou

Strategy for Identification of Nanomaterials' Critical Properties Linked to Biological Impacts: Interlinking of Experimental and Computational Approaches. 385
Iseult Lynch, Antreas Afantitis, Georgios Leonis,
Georgia Melagraki and Eugenia Valsami-Jones

In Silico Approaches for the Prediction of In Vivo Biotransformation Rates. 425
Ester Papa, Jon A. Arnot, Alessandro Sangion and Paola Gramatica

Development of Monte Carlo Approaches in Support of Environmental Research 453
Alla P. Toropova, Andrey A. Toropov, Emilio Benfenati,
Robert Rallo, Danuta Leszczynska and Jerzy Leszczynski

Environmental Toxicity of Pesticides, and Its Modeling by QSAR Approaches. 471
Mabrouk Hamadache, Abdeltif Amrane, Othmane Benkortbi,
Salah Hanini, Latifa Khaouane and Cherif Si Moussa

Counter-Propagation Artificial Neural Network Models for Prediction of Carcinogenicity of Non-congeneric Chemicals for Regulatory Uses 503
N. Fjodorova, M. Novic, S. Zuperl and K. Venko

Big Data in Structure-Property Studies—From Definitions to Models. 529
Jaroslaw Polanski

Index 553

Towards the Revival of Interpretable QSAR Models

Watshara Shoombuatong, Philip Prathipati, Wiwat Owasirikul, Apilak Worachartcheewan, Saw Simeon, Nuttapat Anuwongcharoen, Jarl E. S. Wikberg and Chanin Nantasenamat

Abstract Quantitative structure-activity relationship (QSAR) has been instrumental in aiding medicinal chemists and physical scientists in understanding how modification of substituents at different positions on a molecular structure exert its influence on the observed biological activity and physicochemical property, respectively. QSAR has received great attention owing to its predictive capability and as such efforts had been directed toward obtaining models with high prediction performance. However, to be useful QSAR models need to be informative and interpretable in which the underlying molecular features that contribute to the increase or decrease of the biological activity are revealed by the model. Thus, the aim of this chapter is to briefly review the general concepts of QSAR modeling, its development and discussions on key issues influencing and contributing to the interpretability of QSAR models.

W. Shoombuatong and P. Prathipati
These authors contributed equally to this work.

W. Shoombuatong · S. Simeon · N. Anuwongcharoen · C. Nantasenamat (✉)
Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology,
Mahidol University, Bangkok 10700, Thailand
e-mail: chanin.nan@mahidol.edu

P. Prathipati
National Institutes of Biomedical Innovation, Health and Nutrition,
Osaka 567-0085, Japan

W. Owasirikul
Department of Radiological Technology, Faculty of Medical Technology,
Mahidol University, Bangkok 10700, Thailand

A. Worachartcheewan
Department of Community Medical Technology, Faculty of Medical Technology,
Mahidol University, Bangkok 10700, Thailand

J.E.S. Wikberg
Department of Pharmaceutical Biosciences, BMC, Uppsala University,
SE-751 24 Uppsala, Sweden

K. Roy (ed.), *Advances in QSAR Modeling*, Challenges and Advances
in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_1

Keywords Quantitative structure-activity relationship · Quantitative structure-property relationship · Proteochemometrics · Data mining · Machine learning · Cheminformatics · Chemogenomics · QSAR · QSPR · Interpretable · Drug discovery · Drug design

1 Introduction

Quantitative structure-activity relationship (QSAR) can be considered to be one of the pillars for driving drug discovery efforts forward by enabling practitioners to make sense of the big data from bioactivity assays of chemical library (Nantasenamat et al. 2009, 2010; Cherkasov et al. 2014). Computer-aided drug design or simply computational drug design is essentially comprised of four major levels: (i) fragment, (ii) ligand, (iii) structure and (iv) systems based approaches (Nantasenamat and Prachayasittikul 2015). QSAR is a ligand-based approach meaning that it primarily makes use of information derived from ligands that does not require the need for details of the target protein. Thus, ligand-based approaches are particularly suited in situations where there is negligible information on the biological target. The reasons for using QSAR and quantitative structure-property relationship (QSPR) models are many: (i) to reduce time and cost; (ii) to rationally predict biological, pharmaceutical, physical and chemical activities/properties; (iii) to aid experimental scientists by providing the collective wisdom learned from previous big data; (vi) to shed light on the mechanism of action for biological activities of interest. QSAR/QSPR has found wide applications in the life sciences (Prachayasittikul et al. 2015) (e.g. biology, agriculture and medicine) as well as the physical sciences (Katritzky et al. 2010) (e.g. organic chemistry, physical chemistry, materials sciences). In drug discovery, QSAR has been successfully applied in the prediction of logP and pK_a values as well as absorption, distribution, metabolism, excretion and toxicity (ADMET) properties (Khan and Sylte 2007). It is indeed a difficult task to design a drug that exert activity toward the target protein(s) of interest while at the same time show proper uptake, metabolism, excretion and be devoid of toxicity. To aid medicinal chemists in understanding the origin of ADMET properties Gleeson proposed a set of simple and interpretable rules through the use of principal component analysis of simple descriptors (e.g. molecular weight, logP, ionization state, etc.) (Gleeson 2008).

The robustness of QSAR relies on its capability to predict the biological activities or chemical properties of interests by learning from retrospective experimental data sets. Particularly, each compound in a chemical library is quantitatively or qualitatively described by a set of molecular descriptors and such vector of descriptors (also known as independent variables in statistics) are mathematically correlated with the biological or chemical endpoint of interest (i.e. pIC_{50}, log$P$, etc.) via traditional multivariate analysis or machine learning algorithm. However, it is worthy to note that QSAR models is only as good as the data that was used to train it and in spite of its predictive capability it should not be viewed as a replacement of domain knowledge of scientists but rather should be considered as a complementary tool for aiding the decision-making process.

In spite of its widespread usage, it seems that the full potential for QSAR models has not yet been achieved as current efforts are localized on generating models with good predictive performance at the cost of vague or uninterpretable models. Most robust machine learning algorithms are so-called *black box* since the underlying features contributing to the variation in the endpoint values are not accessible to practitioners. To be of benefit for the experimental biologist or chemist, models need to be transparent such that the underlying important features are revealed. Moreover, features describing the general or unique characteristics of compounds needs to be unambiguous, interpretable and easily comprehensible. Upstream to the issue of interpretability is the accessibility or the know-how on the development of robust QSAR models. Nowadays, the construction of QSAR models may seem to be a trivial and mainstream task in computational drug design. However, a robust, reliable and reproducible model can only be achieved through careful data curation and analysis, which certainly requires the expertise of trained practitioners. This is particularly true as not all starting data set is *modelable* or may not always yield promising results right out of the box owing to several inherent issues that will be discussed in this chapter.

2 Brief History of QSAR

More than a century ago, QSAR was developed by several research groups. The precursor to the birth of QSAR began in 1863 when Cros (Cros 1863) observed that there exists an inverse correlation between toxicity and water solubility. Particularly, the toxicity of alcohols toward mammals increased as the water solubility of alcohols decreased. Shortly after, Crum-Brown and Fraser (1868) reported that there was a correlation between chemical substituents and their physiological properties. Later in the 1890s, Hans Horst Meyer reported that the toxicity of organic compounds depended on their lipophilicity (Borman 1990; Lipnick 1991). Subsequently, the linear correlation between lipophilicity (e.g. oil-water partition coefficients) and biological properties was investigated. Louis Hammett (Hansch et al. 1991) investigated the relationship between electronic properties of organic acids and bases with their equilibrium constants and reactivity. These early studies form the basis for the development of modern QSAR by establishing the idea that molecular structures directly influenced the endpoint (i.e. biological activity and chemical property) of interest. In 1962, Hansch et al. (1962) formally coined the term QSAR and laid its initial foundations by investigating the structure-activity relationship (SAR) of plant growth regulators and pesticides and their dependency on Hammett constants (Hammett 1937) and hydrophobicity (Gallup et al. 1952).

The Free-Wilson model (Free and Wilson 1964) is a simple and efficient method for the quantitative description of SAR. It explains the variation in a series of congeneric compounds using the presence or absence of substituents or functional groups as molecular descriptors. It is the only numerical method that directly relates structural features with biological properties, which is in contrast to Hansch analysis where physicochemical properties are correlated with biological activity values (Kubinyi 1988). Nevertheless, both approaches are closely interrelated, not only from a theoretical point of view but also in their practical applicability (Kubinyi 1988). In many cases both models were combined to a afford a mixed approach that includes Free-Wilson type parameters for describing the activity contributions of certain structural modifications and physicochemical parameters for describing the effect of substituents on the biological activity (Kubinyi 1988; Wei et al. 2001). Many successful applications, especially from the work of Hansch and his group (Verma and Hansch 2009; Hansch et al. 2002; Kurup et al. 2000; Gao et al. 1999; Selassie et al. 2002; Kurup et al. 2001; Hansch and Gao 1997; Kurup et al. 2001; Hansch et al. 1996; Hadjipavlou-Litina et al. 2004; Garg et al. 1999, 2003) on the SAR of enzyme inhibitors, demonstrated that this combined model affords stellar performance for classical QSAR (Hansch 2011). Several variations to Free-Wilson approach have been developed and recently found useful applications in fragment-based drug design (Eriksson et al. 2014; Chen et al. 2013; Radoux et al. 2016).

The field of QSAR modeling had evolved progressively and this encompasses two radical transformations as follows:

1. Paradigm shift from the *classical* to the *non-classical* QSAR approach (Fujita and Winkler 2016). The former is based on a small set of congeneric series of compounds that usually have a single mode of action while the latter is based on large, heterogeneous and non-congeneric data set that may contain several mode of actions.
2. Paradigm shift of QSAR models (Nantasenamat et al. 2009, 2010; Cherkasov et al. 2014) that considers the SAR of *several compounds against a single target protein* to the so-called proteochemometric model (Cortes-Ciriano et al. 2015; Qiu et al. 2016) (sometimes referred to as computational chemogenomics) that investigates the SAR of *several compounds against several target proteins*.

3 How Far Can QSAR Take Us: Can It Really Bring a Drug to Market?

QSAR modeling have evolved from concept to initial hype followed by skepticism thereby leading to the identification of their pitfalls and caveats to a moderation of their expectations (Doweyko 2008). QSAR models are routinely used in the prediction of physicochemical properties (e.g. log$P$, pK_a and solubility) as well as pharmacokinetic and toxicity endpoints (e.g. permeability, plasma protein binding, liver toxicity, carcinogenicity, seizure and off-target activities). However, their usage for actual lead identification and optimization phase has remained quite limited. The skepticism from medicinal chemists towards QSAR models stems from the inability of descriptor based QSAR models (constructed using fingerprints and various topological descriptors) to rationalize activities in terms of simple, meaningful and constructive ways that can clearly provide details on what modifications should be made to the chemical structure that can afford activity enhancement. Furthermore, with better ability to assimilate data from human readable patents and publications of SAR data in concomitant with better understanding of the isosteric concept, medicinal chemists are better able to capture the underlying principles of SAR and make synthetically feasible and conservative predictions. However, many encouraging signs are beginning to appear as more robust machine learning algorithm and interpretable molecular descriptors are being developed. It is still early to predict the potential of QSAR modeling for bringing a drug to market since they are used in the early stages of a drug discovery project. With the ever increases in the availability of clinical and adverse effect data, the use of QSAR modeling together with complementary computational approaches (e.g. cheminformatics, computational chemistry, molecular docking, molecular dynamics, etc.) helps improve the odds of bringing a drug to market. QSAR modeling in combination with other computer-aided drug design techniques have already shown numerous success stories as summarized in an excellent report by Kubinyi (2006).

3.1 Why Does QSAR Fail?

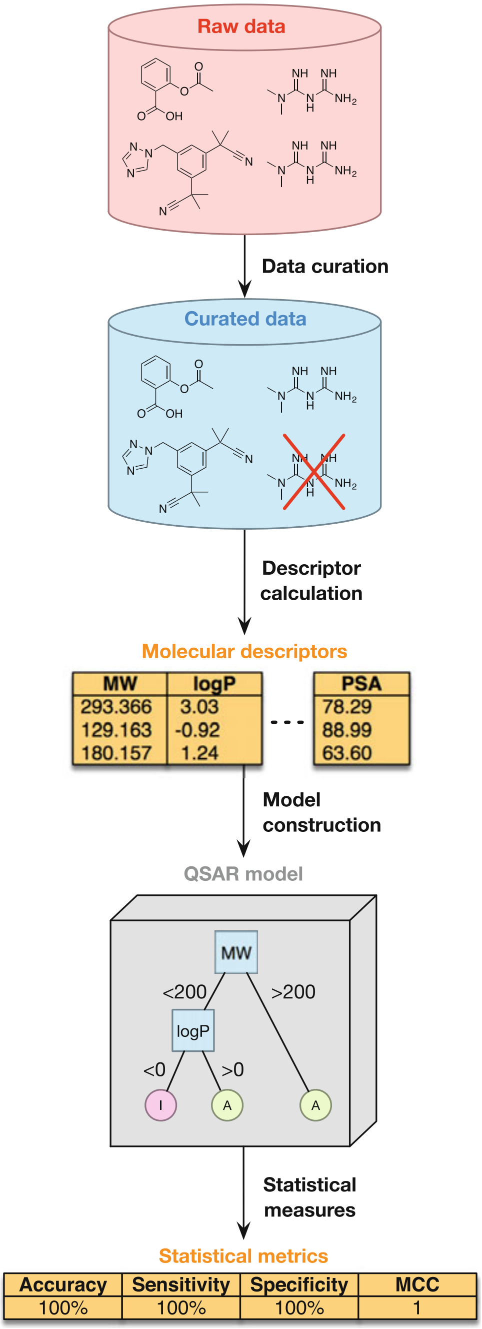
QSAR modeling, like many other research disciplines, has had its fair share of ups and downs. Many predicted the eventual demise of QSAR due to the advances in synthetic chemistry techniques (e.g. combinatorial chemistry) and assay attributes (e.g. automation and miniaturization). Drug discovery researchers dissolution with QSARs is rooted in the fact that it has yet to demonstrate a robust ability to predict the desired biological activities. The disappointing results from QSAR models in certain situation can be attributed to features obtained by chance correlation, rough response surfaces, incorrect functional forms and overtraining (Johnson 2008; Doweyko 2008). Particularly, rough response surfaces are an inherent characteristic of SAR data sets that nevertheless significantly affect the QSAR model predictions. For instance, most aminergic GPCR ligands' agonistic activities correlate with their pK_a and in many instances an order of magnitude change in the pK_a results in a comparable or even an multi-fold change in the biological activity. Such conservative change in the chemical structure leading to a large change in the activity are often not captured by QSAR models which rely heavily on statistical approaches to capture the features that cause the biological responses. On the other hand, a chemist quickly grasps the trend using rational thought, controlled experiments and personal observation assisted by prior knowledge of the protein's structure-function relationships. This over-reliance on statistical procedures by QSAR researchers for feature selection and data modeling has led to the identification of features that may have no mechanistic role in modulating the activities but might have correlated by chance. The excessive emphasis on machine learning has also resulted in model overfitting, models that uses the incorrect functional forms and/or highly predictive models with vague or little interpretability. Hence, the resulting QSAR models do not reflect the reality of the binding or modulation event, which causes the predictions to eventually fail. Thus, to derive meaningful hypothesis, practitioners should not blindly rely on results from computational models but should view the results as hints or guides for supporting their own decision-making process (Nantasenamat and Prachayasittikul 2015). Thus, it is recommended to implement some form of expert knowledge guided component in the QSAR workflows such that new solutions are built upon prior knowledge of targets and their modulation (Saxena and Prathipati 2003). In fact, such data-driven approach as implemented in the HADDOCK docking software (Vries et al. 2010) relies on prior biochemical and biophysical data to drive the docking simulations. Moreover, several recent blinded genomic challenges for phenotype prediction such as sbvImprover (Tarca et al. 2013) and DREAM (Costello et al. 2014) also suggests that the inclusion of prior knowledge can significantly enhance the predictive power while consuming minimal computational resources. In this context, the use of interpretable molecular descriptors aided by transparent machine learning models can greatly alleviate the existing problems of QSAR models.

4 Recommendations for Building Robust QSAR Models

In practice, the development of QSAR models can be carried out to reveal the relationship between the chemical structures and their respective endpoint through the use of various types of mathematical and statistical methods for constructing predictive models that can reveal the origin of bioactivity of interest. A typical $m \times n$ data matrix is comprised of m descriptors and n compounds. A closer look at the M descriptors revealed that it is typically comprised of a set of $\mathbf{X}_{ij}$ descriptors and an $\mathbf{y}_i$ endpoint. In a nutshell, a typical QSAR model is essentially described by an equation the form of $\mathbf{Y} = \mathrm{f}(\mathbf{X}) + \mathit{error}$ that can be used to predict the endpoint for new compounds in lieu of cost and time-consuming approaches. The classical QSAR modeling workflow can be broken down into five prime steps as demonstrated in Fig. 1.

Thus far, several thousands of QSAR models have been developed for various endpoints and these models are created using different model construction schemes (e.g. stringency of data pre-processing, descriptor types, learning methods and evaluation metrics) and published in the public domain (i.e. this is not including the thousands of QSAR models developed in pharmaceutical companies that are not ever published). The variability in the methods used for the QSAR models and their quality may obviously give rise to different outcome for the conclusions possible to draw from them. To further complicate the picture, the reproduction of QSAR models by following the often rather vague instructions in the Methodology sections of research articles do not always yield the same outcome as in the original article owing to the aforementioned factors.

Fig. 1 General workflow of QSAR modeling. Raw data compiled from the literature or public databases are often noisy and dirty and therefore requires curation to clean the data. In this example, redundant chemical structure is removed followed by descriptor calculation, model building and model performance evaluation



MW	logP		PSA
293.366	3.03	- - -	78.29
129.163	-0.92		88.99
180.157	1.24		63.60

Accuracy	Sensitivity	Specificity	MCC
100%	100%	100%	1

Table 1 Summary of the OECD principles for QSAR modeling

No.	OECD principles	Description
1	Defined endpoint	To ensure that all endpoint values within a given data set are consistent
2	Unambiguous algorithm	To ensure transparency and reproducibility of the proposed QSAR model
3	Defined applicability domain	To determine the boundaries in which the model is robust for predicting query compounds
4	Measures of model's predictive potential	To evaluate the internal and external predictive power of the model
5	Mechanistic interpretation	To ensure that the underlying mechanism of action of compounds can be elucidated

Thus, owing to such lack of standards in QSAR/QSPR modeling, the OECD principles was established to address such issues. This first draft initially took place in Setubal, Portugal in 2002 and a revised version in Paris, France in 2004 at the *Workshop on Regulatory Acceptance of QSAR Modelling for Human Health and Environmental Endpoints* and *37th Joint Meeting of Chemicals Committee and Working Party on Chemicals, Pesticides & Biotechnology*, respectively (Worth and Cronin 2004). It has been mandated that to facilitate the consideration of a QSAR model for regulatory purposes, the model should conform to the five principles summarized in Table 1.

Moreover, the integrity of a QSAR model could be pursued by following suggested sets of standards and best practices (Dearden et al. 2009; Tropsha 2010; Tropsha et al. 2003; Dimova and Bajorath 2016; Spjuth et al. 2010) in the development of robust QSAR models. Particularly, Tropsha et al. stressed the importance of leave-many-out validation, bootstrapping, Y-scrambling test and external validation. Moreover, conflicting viewpoints exist on whether to evaluate the robustness of QSAR models on the basis of external validation in which Hawkins et al. (2003) is against this while Esbensen and Geladi (2010) is in support of this. Moreover, recent investigations clearly favor cross-validation over a single external one (Gütlein et al. 2013; Rácz et al. 2015).

In a nutshell, the development of robust QSAR models should address the following key issues:

1. *Data curation*—The curation or pre-processing of data sets prior to performing any form of data analysis is of utmost importance for QSAR modeling. Raw data sets are often *noisy* or *dirty* in the sense that they may inherently contain redundant compounds, redundant descriptors, incorrect representation of the chemical structure or molecular charge. Curation helps to *clean* and increases the reliability of the data set for subsequent analysis.

2. *Modelability*—Modelability is an a priori estimate of the feasibility to obtain externally predictive QSAR models. Modelability is based on the fact that QSAR models are influenced either by data set characteristics (i.e. size, chemical diversity, activity distribution, presence of activity cliffs, etc.), or by modeling workflow steps (e.g. data set curation, feature selection, external validation, consensus modeling, applicability domain, etc.). Particularly, influences arising from the composition of the modeling workflow can be quantified and can be varied given the wide range of molecular descriptors and machine learning methods that are available. However, effects of data set characteristics can be rather difficult to quantify. While size and chemical diversity are subjective attributes of a data set and are difficult to quantify, recent advances have provided methods for objective quantification of activity cliffs (Guha and Drie 2008; Seebeck et al. 2011; Bajorath 2014; Stumpfe et al. 2014; Hu et al. 2012). Building on the earlier proposed concept of the activity cliffs, Golbraikh et al. (2014) proposed a novel modelability index (MODI) that can be easily computed for any dataset at the onset of any QSAR investigation.
3. *Reproducibility*—This important issue is often overlooked by the QSAR community. This is particularly true as often times, QSAR models are built using proprietary software or code that are often restricted to a selected few and not accessible to the general public thereby precluding further attempts to make use of these models. Moreover, the reproduction of QSAR models is a very difficult task indeed as the construction of QSAR models employs different data sets (e.g. different version of the same bioactivity databases such as ChEMBL 19, 20 or 21; it is also highly likely that data sets focused on the same target protein and performed by different laboratory tend to contain different compounds as they may be compiled from different papers), descriptor types, learning methods and evaluation metrics. Spjuth et al. (2010) examines this issue by proposing an open XML format known as QSAR-ML to formalize QSAR data sets with meta-data, which will facilitate the exchange and reproducibility of the model.
4. *Model validation*—The robustness of QSAR models is reliant on stringent validation of QSAR models. Several validation strategies including (1) randomization of the modelled property also known as Y-scrambling, (2) *k*-fold cross-validations and (3) external validation using rational division of a data set into training and test sets are currently the de facto standard for ensuring the utility of a model for virtual screening (Tropsha et al. 2003).
5. *Outliers*—Outlying compounds are those molecules which have unexpected biological activity and do not fit in a QSAR model owing to the fact that such compounds may be acting in a different mechanism or interact with its respective target molecules in different modes (Nantasenamat et al. 2009; Verma and Hansch 2005). Similarly, conformational flexibility of target protein binding site (Kim 2007a) and unusual binding mode are attributed as the possible source of outliers (Kim 2007b). Mathematically speaking, an outlier is essentially a data point that has high standardized residual in absolute value when compared to the other samples of the data set. Furthermore, the building of robust and reliable QSAR models generally emphasizes two major aspects: (1) feature selection and

(2) outlier detection. The two problems are interrelated as outlier definitions are dependent on the selected features. In the realm of QSAR, outliers can be classified as belonging to the following two types: (1) those that fall outside the applicability domain or (2) activity cliffs as discussed in the next section. As the applicability domain considers both chemical and biological space, therefore outliers with respect to biological space can be safely eliminated from QSAR models. However outliers defined based on the chemical space needs further attention. Recent methods such as those from Cao et al. (2011) have argued in support for simultaneously performing variable subset selection and outlier detection using the idea of statistical distribution that can be simulated by the establishment of many cross-predictive linear models. Their approaches build on the concept that the distribution of linear model coefficients provides a mechanism for ranking and interpreting the effects of variables while the distribution of prediction errors provides a mechanism for differentiating the outliers from normal samples (Cao et al. 2011).

6. *Applicability domain*—The applicability domain (AD) (Sahigara et al. 2013) of a QSAR model defines the model limitations with respect to its structural subspace and response space. AD is an indication of the degree of generalization of a given predictive model. AD associated with an endpoint prediction is often well defined if the endpoint prediction for a chemical structure is within the scope of the model. The AD is thus critically reliant on the sampling of chemical subspace and the range of biological readouts that are used for the model development (Sheridan 2015). A commonly overlooked aspect in AD is also the influence of molecular descriptors, generally degenerate and transparent molecular descriptors such as logP, pK_a, etc. afford better degree of generalization to the model while lacking the superior predictive abilities of the more recent topological graph-based descriptors. The various approaches for AD determination are classified as range-based (e.g. bounding box, principal component analysis bounding box and convex hull) and geometric methods (e.g. k-nearest neighbours, DTs, probability density based methods) (Sahigara et al. 2012).
7. *Structure-activity cliffs*—Compounds within a congeneric series whose subtle differences in the chemical structure lead to striking differences in the observed bioactivity are called activity cliffs (Bajorath 2014). Although, the activity cliffs are appealing to medicinal chemists their presence may be detrimental to QSAR models. The inclusion should be carefully reviewed after analyzing for filters such as PAINS (Baell and Holloway 2010) as unusual activity could be due to a wide range of mechanisms such as outliers of different kinds or even the presence of reactive functional groups (Saxena and Prathipati 2006). However, these compounds belonging to the *activity cliffs* are currently categorized as outliers and frequently removed from QSAR models (Guha and Drie 2008). The MODI quantifies the extent of activity cliffs and serves as a guide to the modelability of a data set (Golbraikh et al. 2014).
8. *Feature selection*—The number of molecular descriptors that can capture various aspects of a chemical structure have proliferated in recent years (Todeschini and Consonni 2008). Hence, feature or variable selection is an important and hot

area of research (Guyon 2003; Eklund et al. 2014; Goodarzi et al. 2013). In the context of QSAR studies, feature selection improves interpretability by neglecting non-significant effects thereby reducing noise, enhancing generalization by reducing overfitting (also known as reduction of variance), increasing the models' predictive ability and speeds up the QSAR model building process (Saxena and Prathipati 2003). Some widely used and relevant approaches for QSAR studies includes: (1) all subset models (ASM), (2) sequential search (SS), (3) stepwise methods (SW), (4) genetic algorithm (GA), (5) particle swarm optimization (PSO), (6) ant colony optimization (ACO), (7) least absolute shrinkage and selection operator (LASSO), (8) elastic net and (9) variables importance on PLS projections (VIP) (Eklund et al. 2014), (10) correlation-based feature selection (CFS) (Hall 1999), (11) simulated annealing (Siedlecki and Sklansky 1988), (12) sequential feature backward selection (Pudil et al. 1994), (13) sequential feature forward selection (Pudil et al. 1994), (14) minimum-redundancy-maximum-relevance (mRMR) (Peng et al. 2005), (15) ReliefF (Liu and Motoda 2007), (16) Tikhonov regularization (Destrero et al. 2009), (17) recursive feature elimination (RFE) (Guyon et al. 2002), (18) random forest (RF) (Breiman 2001), (19) decision tree (DT) (Quinlan 1993), etc.

9. *Class imbalance*—Class imbalance in supervised machine learning is a major confounding problem for the construction of QSAR models (Li et al. 2009). In a classification setting, the size of the active and inactive sets of compounds may be significantly disproportional and may therefore lead to biased predictive models. Several solutions that include artificially undersampling the overrepresented class or oversampling the underrepresented class or using one class learning or cost-sensitive training have all been suggested as possible remedies to address this issue (Zakharov et al. 2014; Capuzzi et al. 2016).
10. *Chance correlation*—Objectivity is a critical component of any hypothesis generating workflow including QSAR. It has been stressed that causation and correlation are indeed two different things and that a model's performance may possibly arise by chance. A possible remedy is to apply Y-scrambling (Rucker et al. 2007) to evaluate model robustness.
11. *Confidence/reliability of the model*—QSAR models are not universally applicable as predictions may fail under certain conditions. QSAR models are based on mathematical formulations for modeling the bioactivity as well as to draw conclusions from. Their utilization in medicinal chemistry encompasses idea generation, virtual screening and knowledge discovery. Hence, the confidence in the predictions derived from QSAR model should be accessible. Substantial efforts have been devoted to research on this topic within the QSAR community over the last decade and a number of methods have been suggested for estimating the confidence of QSAR predictions. These confidence estimates are typically based on the very loosely defined concept of a QSAR models applicability domain (AD), which is described as the response and chemical structure space in which the model makes predictions with a given reliability. The assumption is that the further away a molecule is from a QSAR models AD,the less reliable the prediction becomes. This confidence measure can be afforded by an approach

known as conformal prediction (Shafer et al. 2008), which has been successfully applied in QSAR modeling (Eklund et al. 2012). The conformal prediction framework provides a unified view of the different approaches for estimating a QSAR models AD. Moreover, conformal prediction provides a natural and intuitive way of interpreting the AD estimates as prediction intervals with a given confidence.

12. *Interpretability of the model*—Perhaps, the most important contribution of QSAR modeling lies in their ability to propose a hypotheses to rationalize the binding/function modulation phenomenon via interpretation of the model's features. In view of its critical role in fulfilling the objectives of QSAR modeling, we focus our chapter on their interpretability. The hypothesis gleaned from QSAR models can benefit biologists and chemists by providing insights into the cause-effect relationships between molecular features and bioactivity measures. These insights can aid medicinal chemists to design future SAR studies objectively and comprehensively. They can also assist molecular and structural biologists in proposing candidates for site-directed mutagenesis and related structure-function experiments. This chapter proposes the use of interpretable molecular descriptors together with interpretable machine learning methods. Recent interest in the field had also shifted towards making the black box learning methods more transparent and amenable to interpretations, which will be covered in the forthcoming sections.

5 Trade-Offs Between Performance and Interpretability

Over the past decades, many QSAR studies had predominantly focused on enhancing and improving the predictive performance instead of the interpretability of the model (Fujita and Winkler 2016). The shift can be seen in QSAR model descriptors moving away from the physicochemical and indicator variables of Hansch-Fujita and Free-Wilson approaches towards highly non-degenerate and continuous molecular descriptors which offer high predictive power. However, improved understanding of the concepts of bioisosterism and the molecular recognition events, identification of problems associated with capturing molecular structures and errors in assay data of widely used SAR databases give credence to the use of moderately degenerate and interpretable 1D or fingerprint based molecular descriptors as expanded elsewhere in this chapter. Learning methods in QSAR modeling have evolved from simple interpretable methods such as linear regression as used by Hansch and Fujita to the complex black box approaches such as neural networks and deep learning. While many experts agree with the obvious improvements (i.e. approximately 10%) to the predictive power from these complex machine learning methods, they argue that the loss of interpretability of the feature contributions are not worth the gain in predictive power. Hence, in Sect. 8.1.4 we expand upon the recent advances in rule extraction techniques that help to provide enhanced interpretation of the complex black box approaches. This section also presents several recent enhancements that significantly improve the predictive power of the white box learning approaches. Hence, this chapter presents and advance the case for interpretable QSAR models in drug discovery research. We argue that a simple and interpretable QSAR model with modest predictive performance would be more valuable to experimental scientists than a highly predictive but black box model since no or minimal insights can be gained from it.

6 Reverse Engineering of QSAR models

Designing new molecules corresponding to the given biological activity is invaluable to the chemical, material and pharmaceutical industries. The traditional approaches of computer-aided molecular design based on QSAR modeling can be used to solve two main problems: (i) *forward QSAR problem*, which identifies the compounds' structural and physicochemical features related to the experimental readout using machine learning (ii) *inverse QSAR problem* that seeks to reconstruct compounds' structures which correspond to the specific features related with the readout (Faulon et al. 2005; Brown et al. 2006).

The inverse problem is generally addressed as a subgraph construction. Previously, there were five types of approaches to solve the inverse problem: random search, heuristic enumeration, mathematical programming, knowledge-based system, and graphical reconstruction methods. The inverse QSAR analysis is quite challenging for various reasons: combinatorial complexity of the search space, design knowledge acquisition difficulties, nonlinear structure property correlations, and problems in incorporating higher level chemical and biological knowledge (Venkatasubramanian et al. 1995). Thus, it is not surprising that constructing new structural compound given a desired activity is a long-standing problem. In practice, the inverse QSAR method can be divided into the common four steps (Skvortsova et al. 1993; Wong and Burkowski 2009; Churchwell et al. 2004; Visco et al. 2002; Weis et al. 2005). Firstly, a QSAR equation is constructed to derive a forward QSAR model that essentially discerns the relationship between a set of descriptors and their activities. The second step is to generate the set of constraint equations with integer coefficients. The constraints are used for ensuring that the constructed compounds afford the desired activities. There are two types of constraint equations: graphical and consistent equations, which are then solved in the third step. Finally, the compound structures are enumerated and constructed to afford the desired activity while their activities are predicted using the forward QSAR model described in the first step.

Until now, there are relatively few studies providing computational-based models for solving this problem (Visco et al. 2002). Almost all of the proposed computational-based methods that are used are essentially a stochastic model in nature and use either genetic algorithm (GA) or Monte Carlo simulated annealing approach to construct new chemical compounds. In 1995, Venkatasubramanian et al. (1995) and Sheridan and Kearsley (1995) proposed a stochastic model based on Monte Carlo. GA is a general purpose approach based on the Darwinian principle for natural selection and evolution, which are used for stochastic, evolutionary search, and optimization strategies. The main advantage of GA lies in its ability to allow a dynamically evolving population of molecules to gradually improve by competing for the best performance. However, the problem from these studies represent a combinatorial explosion (Kvasnicka and Pospichal 1996). In order to analyze a huge number of compounds, Kvasnicka and Pospichal (1996) developed a new approach based on a random search that not only afford all solutions but also provide users with a high probability of deriving the correct solution. In 2002, Visco et al. introduced the use of signature descriptors to represent compounds as molecular graphs. In this study, a set of 121 HIV-1 protease inhibitors were analyzed by comparing the proposed QSAR model with other descriptor types consisting of connectivity indices, KierHall shape indices, fragments, electrotopological states and information indices. This work also revealed that signature descriptors are particularly well suited for tackling the inverse problem (also see the work from Faulon 1994, 1996; Faulon et al. 2003; Churchwell et al. 2004; Faulon et al. 2004; Weis et al. 2005). Also from the same group, Churchwell et al. (2004) applied the inverse QSAR approach to a small set of peptide inhibitors that targets the leukocyte functional antigen-1 (LFA-1)/intercellular adhesion molecule-1 (ICAM-1) complex. Their prediction results showed that the predicted IC_{50} values were very close to that of the experimental IC_{50} values. Practically, the inverse QSAR problem is relatively difficult when compared to the forward QSAR problem because the molecular descriptors used for constructing the inverse QSAR model must adequately address the forward QSAR model for the activity or property of a given data, if the subsequent recovery phase is to be meaningful. Additionally, a major problem is to reconstruct and enumerate the chemical structures from its extracted descriptors. To solve such problem, Wong and Burkowski proposed (Wong and Burkowski 2009) a new workflow using a vector space model molecular descriptor (VSMMD) to represent the chemical structures. Their proposed inverse QSAR model consists of five key steps: (i) calculating the VSMMD for each compound from the training set; (ii) apply the kernel function (i.e. more detail is discussed in a subsequent section) to map each VSMMD from the input space (i.e. low dimension) to the feature space (i.e. high dimension); (iii) designing a new point in the feature space using a kernel function algorithm; (iv) map the new point from the feature space and trace back to the input space using a pre-image approximation algorithm and (v) building the chemical structures using the VSMMD recovery algorithm.

As can be seen, inverse QSAR models has great potential for obtaining desirable compounds directly from the trained QSAR model. Further work in this area is highly encouraged as to help steer towards the practical utility of QSAR models for building promising chemical structures aside from making predictions of their bioactivity values or class label.

7 Interpretable Molecular Descriptors

7.1 Role of Molecular Descriptors in Post-genomic Drug Discovery

Molecular descriptors encode the physical and chemical properties of molecules of interest and are central to QSAR/QSPR studies (Danishuddin 2016). The availability and the use of high quality, interpretable descriptors can greatly contribute to the formulation of an intuitive model for retrospective and prospective analysis of life or material sciences data (Cherkasov et al. 2014). As depicted in Fig. 2, molecular descriptors play a critical role in enabling mathematical and statistical analysis for relating chemical structure with biological data. While human intuitive molecular graphics depictions use the atom, bond, angle coordinates together with charge

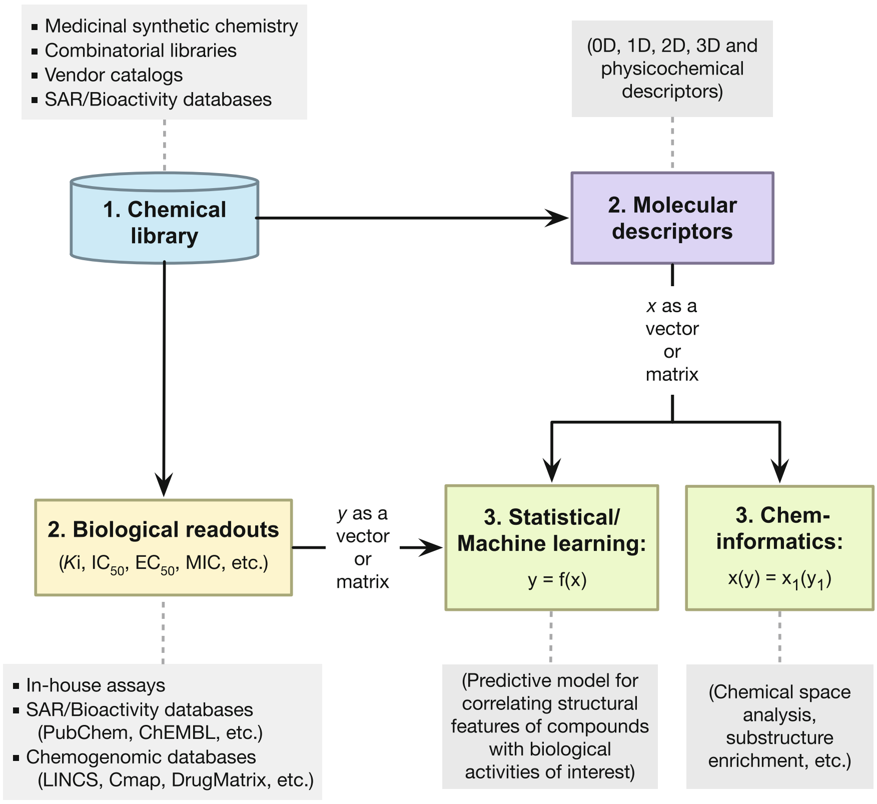

▪ Medicinal synthetic chemistry
▪ Combinatorial libraries
▪ Vendor catalogs
▪ SAR/Bioactivity databases
(0D, 1D, 2D, 3D and physicochemical descriptors)
1. Chemical library
2. Molecular descriptors
x as a vector or matrix
2. Biological readouts
(*K*i, IC$_{50}$, EC$_{50}$, MIC, etc.)
y as a vector or matrix
3. Statistical/ Machine learning:
y = f(x)
3. Chem-informatics:
x(y) = x$_1$(y$_1$)
▪ In-house assays
▪ SAR/Bioactivity databases (PubChem, ChEMBL, etc.)
▪ Chemogenomic databases (LINCS, Cmap, DrugMatrix, etc.)
(Predictive model for correlating structural features of compounds with biological activities of interest)
(Chemical space analysis, substructure enrichment, etc.)

Fig. 2 General schematic diagram depicting the importance of molecular descriptors for capturing the details of chemical structures (from a chemical library) as vectors and matrices; hence enabling mathematical and statistical procedures for QSAR and other chemoinformatics analysis

information to reconstruct the chemical structures as 2D and 3D projections, encoding chemical structure as machine readable matrices and vectors is required for performing mathematical and statistical analysis. In this regard molecular descriptors play a key role in establishing QSARs and in performing chemo-informatics tasks such as chemical space mapping, substructure analysis, etc. In the pre-genomic era, biological readouts were available as single vectors, however advances in miniaturization, robotics and automation in the post-genomic era presented QSAR researchers with a complex array of biological data as matrices. The complex biological matrices include both the traditional target and phenotypic measurements and the recent clinical chemistry and histopathology findings and microarray and proteomics data (Prathipati and Mizuguchi 2016a). These data were generated in standardized high-throughput format and are available in databases such as LINCS, Open TG-Gates, CEBS, DrugMatrix and CMap (Prathipati and Mizuguchi 2016a). Several advanced multi-label statistical techniques (such as network-based inference) and complex molecular descriptors (such as proteo-chemometric) are presently under development which can capture both the biological data's relationship with the chemical structure together with complex relationships among the biological readouts and the chemical structures (Prathipati and Mizuguchi 2016a). Thus a range of machine learning methods are under consideration for multi-label QSAR models depending on the data types such as support vector machines (SVMs), neural networks (NN), *k*-nearest neighbors (*k*NN), boosting methods for unrelated multi-label datasets and similarity based approaches such as DT-hybrid, kernel regression methods such as lasso or elastic nets or pairwise kernel method (PKM) for related multi-label datasets (Prathipati and Mizuguchi 2016a). While some of these machine learning methods are discussed in Sect. 8, in the following subsections we expand upon the range of molecular descriptors and their attributes and their utility for modelling the wide array of biological readouts.

7.2 Interpretability of Molecular Descriptors Advances Ligand-Based Approaches

The continuing appeal of QSAR models as part of ligand-based approaches in the face of the ever increasing structural data of target proteins and advancements in structure-based approaches is an interesting conundrum (Prathipati and Mizuguchi 2016a). Although structure-based approaches are highly interpretable and intuitive to drug researchers, their efficiency and effectiveness is limited by several factors including ambiguity in pose prediction, limitations of scoring functions at capturing the molecular recognition event, limitations of existing methods in considering bridging water molecules and induced fit phenomenon (Prathipati et al. 2007; Prathipati and Mizuguchi 2016b). Furthermore, drug targets such as nuclear receptors, G protein-coupled receptors (GPCRs) and kinases are known to have multiple conformational states that exists in equilibrium in the absence of their cognate ligands (Spyrakis and Cavasotto 2015; Zhao et al. 2014; Rueda et al. 2009, 2010). Most often the X-ray structures of one or the other of these conformational states are difficult to obtain. For instance, several kinases are known to exist in at least 4 different conformational states (e.g. DFG-in, DFG-out, A-loop-out and A-loop-in) in recognizing type -I, -II and -III inhibitors (Chiu et al. 2013). The DGF-out inactive conformational state of a kinase is quite flexible and is quite difficult to crystallize where the catalytically important p-loop is most often difficult to resolve (Kufareva and Abagyan 2008). Similarly, GPCRs too exist in the active, inactive and apo conformational states. While the inactive GPCR conformational states are easy to crystallize owing to its rigidity as conferred by the strong salt-bridge interactions between the helices (e.g. helices 3 and 6 and helices 2 and 5), the active conformational state stabilized in the presence of an agonist disrupts these interactions through charge neutralization, hence becomes flexible and is difficult to crystallize and resolve (Standfuss et al. 2011). Conversely, ligand-based QSAR models are quick and can be dynamically adapted to model both target and phenotypic endpoints as well as different types of chemotypes with relatively little effort (Prathipati and Saxena 2005). QSAR models derived using molecular descriptors were shown to provide high predictive power and were successfully used for hit identification (Krasavin 2015; Geronikaki et al. 2008; Poroikov et al. 2003). The disadvantages of this approach is their comparatively low intuitiveness and their difficulty for interpretation (Saxena and Prathipati 2006). Hence, we shall attempt to discuss the pros and cons of various descriptors in terms of their quality and interpretability.

7.3 *Assessing the Quality and Interpretability of a Molecular Descriptor*

Historically, the Hammett equation (Hammett 1937) describes one of the earliest known mathematical formulations relating structures with the property of interest (i.e. reactivity in this instance) and remains the most widely used and understood mathematical equation to date. It describes a linear free-energy relationship relating rate or equilibrium of a reaction with a substituent's position and electronic property (i.e. withdrawing or donating) captured as 'Sigma' (Hammett 1937). The molecular descriptor 'Sigma' as proposed by Hammett (1937) to explain the acidity of substituted benzoic acids also serves as useful guidepost in evaluating the quality and interpretability of a molecular descriptor. 'Sigma', also called the substituent constant, has several features that makes it an excellent molecular descriptor, particularly it has (1) high structural interpretation, (2) good correlation with biological or physical property (i.e. pK_a in this case), (3) can be applied to local structure (substructures), (4) uses the familiar structural and electronic concepts (e.g. electronegativity and polarizability), (5) high sensitivity (i.e. varies with structures; even isomers) and (6) size dependence (i.e. changes with molecular weight). However, the original implementation of Hammett involves using experimental properties and makes the

Table 2 Summary of the strengths and weaknesses of the various dimensions of molecular descriptors. The number of stars denotes the strengths and weaknesses for each characteristics while the exclamation mark designate that caution should be taken

Characteristics	0D	1D	2D	3D	PC
Simplicity	★ ★ ★	★ ★ ★	★★	★	★ ★ ★
Calculation efficiency	★ ★ ★	★ ★ ★	★★	!	★
Structural interpretation	★	★★	★	★ ★ ★	★ ★ ★
Correlation with biological property	★	★★	★★	★ ★ ★	★ ★ ★★
Applicable to local structure (substructures)	★	★ ★ ★	★★	★	★★
Use familiar structural and electronic concepts	★	★	★★	★ ★ ★	★ ★ ★★
Sensitivity (discriminate different structures including isomers)	!	★	★★	★ ★ ★	★★
Size dependency (varies with MW)	★	★★	★★	★★	★★

0D: zero-dimensional descriptors, 1D: one-dimensional descriptors,
2D: two-dimensional descriptors, 3D: three-dimensional descriptors,
PC: physicochemical descriptors

computation of sigma highly inefficient and hence not practical for high-throughput virtual screening workflows. We shall discuss the importance of physicochemical properties descriptors and the 4 major class of structural descriptors in light of the features discussed above (Table 2). Furthermore, several novel applications of QSAR such as the modelling of peptides, nucleotides and nanostructures for biologics-based drug discovery research requires the availability of novel descriptors. Hence, Table 4 presents the list of free software along with availability of various descriptor types (Table 3).

7.4 *Trade-Offs Between Descriptor Quality and Interpretability*

Thus, it should be noted that a descriptor's quality and its interpretability, together with the use of an appropriate machine learning method can greatly produce a practical and interpretable QSAR model that scientists can use. The *sensitivity* or the *degeneracy* of a molecular descriptor is the measure of its ability to avoid equal values for different molecules. This is the most critical attribute of a descriptor's quality. Furthermore, a descriptor's interpretability can be defined as its ability to elucidate and rationalize the underlying structural and physicochemical properties responsible for the biological response.

3D descriptors which most accurately encode the structural and physicochemical properties that are responsible for the investigated endpoint are presently regarded to afford robust quantitative descriptions of molecular structures. They have high

Table 3 Summary of model techniques used in QSAR modeling and their advantages and disadvantages

Method[a]	Interpretable	Linear	Supervised learning?	Advantage(s)	Disadvantage(s)
MLR	Yes	Yes	Yes	Good interpretability	Problem of learning dichotomous variables
LR	Yes	Yes	Yes	Deal with dichotomous variables	Perform poor on complex data
ELM	Yes	Yes	Yes	Interpretability	Perform poor on multiclass data
PCA	Yes	Yes	No	Dimension reduction	Unsupervised learning
PLSR	Yes	Yes	Yes	Interpretable and reducted dimension	Linear model
DT	Yes	No	Yes	High interpretability	Overfitting
RF	Yes	No	Yes	High interpretability/tolerant to overfitting	Long training time
ANN	No	No	Yes	Perform well on complex data	Poor interpretability
DL	No	No	Yes	Hierarchical features learning	Poor interpretability
SVM	No	No	Yes	Good generalization performance	Poor interpretability

[a] MLR: multiple linear regression, LR: logistic regression, ELM: efficient learning method,
PCA: principal component analysis, PLSR: partial least squares regression,
DT: decision tree, RF: random forest, ANN: artificial neural network, DL: deep learning,
SVM: support vector machine

sensitivity and present different values of different isomers and other subtle structural variations. Some 3D descriptors such as those based on the GRID concept or obtained from quantum chemical computations provide causal insights while those based on the graph concept akin to the 2D graph-based descriptors present very little causal interpretation. Furthermore, 2D graph-based descriptors are equally as degenerate as a 3D descriptor and can also be regarded as a descriptor of high quality. However, most medicinal chemistry SAR data are not highly sensitive to small changes in the structure (i.e. the addition of substructures to non-pharmacophoric areas) and are shown to have moderate complexity (Schuffenhauer et al. 2006). Furthermore,the assay data too are prone to experimental artifacts (e.g. aggregation, reactive functional groups induced assay readouts) and errors (i.e. standard deviation of technical replicates) (Feng et al. 2005; Feng and Shoichet 2006; Feng et al. 2007; McGovern et al. 2002; Thorne et al. 2010). The moderate complexity of the chemical space can be attributed to the difficulties in their synthesis and purification as well as the characterization of stereo- and regioisomers.

In light of the moderately complex chemical space, 1D or fingerprint descriptors having moderate sensitivity (e.g. non-degenerativity) and interpretability, have become the de facto standard in chemoinformatics both for a prospective and retrospective QSAR analysis (Schuffenhauer et al. 2006). The compact nature of the bit-vector representation makes them amenable to not only QSAR modeling but also for a wide range of computations such as similarity searching (Prathipati et al. 2008), clustering (Prathipati et al. 2008), substructure searching and the inverse QSAR problems (Rosenbaum et al. 2011). As to address issues such as assay errors, artifacts and heterogeneity of assay methods, the use of classification models has been proposed as a promising solution and as such its usage has steadily increased in recent years.

7.5 Dimensions of Molecular Descriptors

7.5.1 0D Descriptors

The 0D descriptors (Todeschini and Consonni 2008) capture the counts of atoms (e.g. number of carbon atoms, number of nitrogen atoms, etc.) and bonds as well as their constitution (e.g. hybridization states and bond orders). In addition, 0D descriptors also encode the sum or average of the atomic properties such as weight, volume, polarizability, electronegativity, etc. These descriptors are easily calculated and naturally interpreted but they may not be very sensitive to subtle changes in molecular structures (e.g. isoforms). However, this class of descriptors have successfully been used in explaining the variation effect of structures on activity/property of several data sets as has extensively been shown by the research group of Andrey Toropov and Alla Toropova (Toropov and Benfenati 2007a, b; Toropov et al. 2010).

Particularly, the research group of Toropov and Toropova proposed the SMILES-based descriptors for the easy computation and interpretation of the importance of features followed by QSAR modeling using the Monte Carlo approach. This computational methodology has been produced as a free software called the CORrelation And Logic (CORAL) (http://www.insilico.eu/coral) (Toropov and Benfenati 2007a, b; Toropov et al. 2010). The SMILES notation is used to directly extract 1D molecular features (e.g. atom, bond and other elements) from the chemical structures without the need for external software for descriptor calculation. It can be used for the development of regression and classification based predictive models using the Monte Carlo technique for biological activities (Worachartcheewan et al. 2015; Masand et al. 2014), chemical properties (Toropova and Toropov 2014; Gobbi et al. 2016) and nanomaterial properties (Toropov et al. 2013). CORAL requires an input file consisting of the compound name, SMILES notation and the bioactivity values or class labels. Compounds from the data set are separated into training, invisible training, calibration sets (i.e. used as visible data set) and validation set (i.e. used as invisible data set that is not used during the model construction). Moreover, such data subsets are generated for three or more independent data splits as to evaluate variability from the prediction models. The performance of such models can be derived from statistical parameters such as R^2, Q^2, $R^2 - Q^2$ (Worachartcheewan et al. 2014).

A set of local and global molecular features can be derived from the SMILES notations as follows:

$$\begin{aligned} abcdef &\rightarrow a + b + c + d + e + f(S_k) \\ abcdef &\rightarrow ab + bc + cd + de + ef(SS_k) \\ abcdef &\rightarrow abc + bcd + cde + def(SSS_k) \end{aligned} \tag{1}$$

These are the examples of local descriptors that represents the elements in the SMILES notation. In addition, global descriptors are also encoded designated as *BOND*, *PAIR*, *NOSP* and *HALO* as follows:

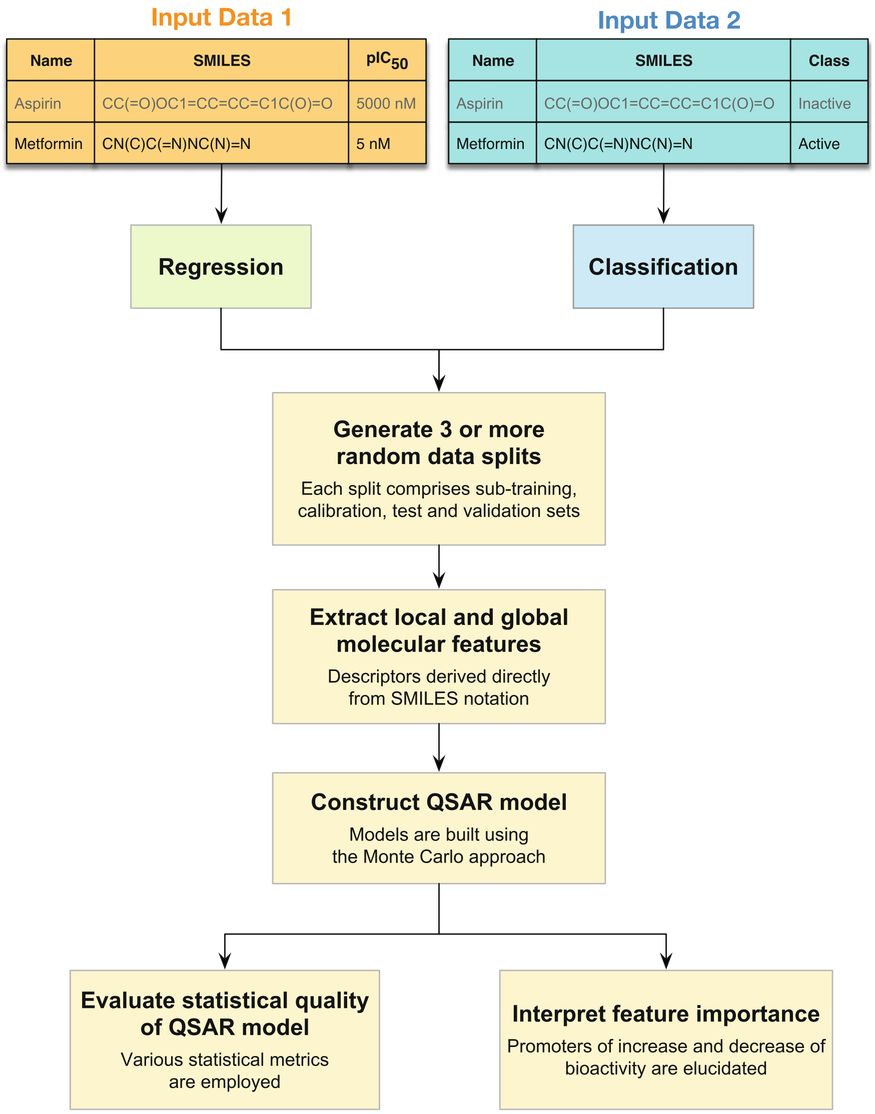
- *BOND* is presence/absence of bond in the SMILES input such as double bond (=), triple bond (#) and stereo chemical bond (@)
- *PAIR* is the co-incidence of two elements of the following: F, Cl, Br, I, N, O, S, P, #, = and @
- *NOSP* is presence/absence of N, O, S and P
- *HALO* is presence/absence of halogens

In the software, optimized parameters include threshold and correlation weights (CW). An example of equation of SMILES-based optimal attributes, was calculated by the following equation:

$$\begin{aligned} DCW(Threshold, N_{epoch}) = \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) + \\ \sum CW(BOND) + \sum CW(NOSP) + \sum CW(HALO) + \sum CW(PAIR) \end{aligned} \tag{2}$$

The biological/chemical endpoint can be calculated as follows:

$$Endpoint = C_0 + C_1 \times DCW(Threshold, N_{epoch}) \tag{3}$$



Input Data 1

Name	SMILES	pIC_{50}
Aspirin	CC(=O)OC1=CC=CC=C1C(O)=O	5000 nM
Metformin	CN(C)C(=N)NC(N)=N	5 nM

Input Data 2

Name	SMILES	Class
Aspirin	CC(=O)OC1=CC=CC=C1C(O)=O	Inactive
Metformin	CN(C)C(=N)NC(N)=N	Active

Regression
Classification
Generate 3 or more random data splits
Each split comprises sub-training, calibration, test and validation sets
Extract local and global molecular features
Descriptors derived directly from SMILES notation
Construct QSAR model
Models are built using the Monte Carlo approach
Evaluate statistical quality of QSAR model
Various statistical metrics are employed
Interpret feature importance
Promoters of increase and decrease of bioactivity are elucidated

Fig. 3 Workflow of the CORAL software for constructing QSAR modelings using SMILES-based descriptors

where C_0 is the intercept and C_1 is the slope or correlation coefficient.

Furthermore, the molecular fragments obtained from the software can give knowledge of important chemical feature influencing their activities as promoters for increasing or decreasing biological activity. The summary of development of predictive models using SMILES-based descriptors by CORAL software are outlined in Fig. 3.

Recently, Filimonov et al. (2009) proposed a novel QNA-based Star Track QSAR approach in which any molecule is represented as a set of points in 2D space of QNA descriptors. The Star Track approach is in contrast with the classical QSAR method and does not require the use of feature selection. This approach is implemented in the GUSAR software package and is based on a self-consistent regression, QNA descriptors and the topological length and volume of a molecule. This approach predicts quantitative values of biological activity of compounds on the basis of their structural formula and does not require the use of information about the 3D structures of ligands and/or target proteins. The Star Track QSAR approach compares favorably with different 3D and 2D QSAR methods on various gold standard data sets and does not select models based on Q^2 values. Thus, the Star Track QSAR approach as implemented in the GUSAR software package is a potentially useful approach for the derivation of statistically robust, interpretable and fast QSAR models.

7.5.2 1D Descriptors

1D descriptors, also referred to as fingerprints, essentially capture the counts and properties of functional groups and substructural fragments (Todeschini and Consonni 2008). A fundamental difference between 1D descriptors and fingerprints is that the former uses a predefined set of keys (i.e. functional groups and substructures) to generate the descriptors while the latter uses either a predefined set or a set of keys generated on the fly. The older generation of fingerprints consisting of MACCS (Durant et al. 2002), PubChem, and SMARTS still uses a predefined set of keys (Hinselmann et al. 2011) for generating fingerprints and are critically limited at capturing the domain (target- and ligand-) specific structural features responsible for variation in activities. For instance, predefined fingerprints may capture too few or too many correlating features which may have moderate value in QSAR studies. However, recent advances in computer science led to the concept of hashed fingerprints where a set of patterns are generated by gathering atom environment information or subgraph information or both. The generated context dependent patterns are then transformed into hash codes (i.e. a fixed size vector) using hashing algorithm. These hash codes can then be transformed into bit strings using a random number generation of a defined length (i.e. size of the fingerprint). The presence and absence of a pattern is marked as being either *1* and *0*, respectively. Extended connectivity fingerprint (ECFP) (Rogers and Hahn 2010) is a prototypical example of a hashed fingerprint. A major advantage of 1D descriptors or hashed fingerprints is their ability to capture complex structural patterns in uniform fixed bit vectors, which can be quickly computed (Rogers and Hahn 2010). These bit vectors are amenable for molecular similarity/substructure analysis problems, show little degeneracy, are naturally interpreted and are widely used in chemoinformatics (Prathipati et al. 2008). In view of the intuitive concepts of substructures' and functional groups' contributions to drug design and their efficient computation, the 1D descriptors or fingerprints were primarily used for the inverse QSAR problems (Rosenbaum et al. 2011) as discussed in the Introduction.

7.5.3 2D Descriptors

2D or topological descriptors (Gozalbes et al. 2002) are computed by encoding the atoms and their connectivity as a graph. Several variations to the graph-theoretic representation of atoms and their connectivities led to the wide plethora of methods for the generation of 'graph-theoretic' descriptors such as Kier and Hall (1976), Broto et al. (1984), Balaban (1982), Randic (1975), MEDV etc. Although they lack in interpretability, 2D descriptors can be considered good descriptors in many aspects (as listed in Table 2). However, the poor interpretability of this class of descriptor critically limits its usage in retrospective QSAR analysis (Gozalbes et al. 2002). Furthermore, since correlation does not always imply causality, models derived using these class of descriptors are difficult to prioritize from a pool of models that offer very similar statistical significance (Saxena and Prathipati 2006). There are two excellent techniques to mitigate this problem and discriminate seemingly equivalent models via the generalized pairwise correlation method (GPCM) (Héberger and Rajkó 2002) and the sum of ranking differences (Heberger and Skrbic 2012). However, QSAR models derived from these descriptors are ideally suited for a prospective virtual screening analysis as they can be efficiently computed and generally have very low levels of degeneracy (Saxena and Prathipati 2006).

Among the various topological indices, the molecular electronegativity distance vector based on 13 atomic types called the MEDV-13, is a fast, easy to use, reproducible and predictable descriptor for QSAR studies. The studies by Liu et al. (2001) show the performance of MEDV-13 models were comparable to 3D QSAR studies and are also applicable to QSARs of peptides. MEDV-13 descriptor in addition employs information about an element atom type, valence electronic state, and chemical bond type from 2D molecular topology and requires no information related to 3D structures or physicochemical properties or molecular alignments.

7.5.4 3D Descriptors

3D descriptors characterize the 3D structure of a molecule in terms of their shape, steric and electronic features (Kubinyi 1993). While shape-based 3D descriptors (e.g. volume, *RDF* (Gonzlez et al. 2005), *autocorrelation3D* (Sliwoski et al. 2016), etc.) are highly relevant in explaining SAR data, they remain difficult to interpret. Furthermore, the 3D descriptors comprising of *RDF* (Gonzlez et al. 2005), *3D-MoRSE* (Devinyak et al. 2014), *WHIM* (Bravi et al. 1997) and *GETAWAY* (Consonni et al. 2002) descriptors share many similarities with 2D descriptors as described above. While the latter encodes atoms and their connectivity as simple graphs, the 3D shape-based descriptors capture these features together with their distances and angles as part of a complex graphs. On the other end, the 3D descriptor spectrum includes descriptors such as steric and electrostatic fields that are computed using semi-empirical quantum chemical methods as part of the GRID concept (Sippl 2006). The 3D QSAR paradigm asserts the importance of conformational preferences of compounds for molecular recognition to its target protein in addition to structural and physicochemical features as described above. The CoMFA/CoMSIA methods (Cramer et al. 1988) to date remains the prototypical examples of this paradigm and several leading publications reported seemingly interpretable retrospective analysis of both target-based (Prathipati et al. 2005) and phenotype-based SAR data. However, in a seminal paper, Doweyko (2004) debunked the commonly asserted illusion and showed that the so-called significant regions are subject to the vagaries of alignment and that the nature of possible interactions heavily depends on the eye of the beholder. Furthermore, the arbitrary nature of both the alignment paradigm and atom description lends itself to capricious models, which in turn can lead to distorted conclusions (Doweyko 2004). In spite of limitations of the 3D QSAR approach, this class of descriptors demonstrates very low levels of degeneracy (i.e. extremely sensitive to changes in the structure) and is considered as the gold standard amongst the QSAR modelling techniques. Although, the 3D steric and electrostatic fields have been very intuitive both for explaining the SAR data and for guiding several novel designs, a potential limitation is their rationalization is limited to a congeneric series of compounds. Hence, 3D-QSAR models are not typically used for large-scale prospective virtual screening analysis (Doweyko 2004). Although, several variations of the Tripos CoMFA/CoMSIA (Cramer et al. 1988) have emerged in recent years, the only known freeware is Open3DQSAR (Tosco et al. 2011), which is potentially an interesting addition to the growing number of 3D QSAR software.

7.5.5 Physicochemical Properties

Physicochemical properties are considered to be one of the most relevant descriptors for drug design (Brustle et al. 2002; Taskinen and Yliruusi 2003). While they are mostly measured quantities, they are calculated based on parameterization with measured data. Thus, these descriptors differ from others in that they are not derived from first principles but are obtained from models trained using either 0D, 1D, 2D, 3D (e.g. 3D quantum chemical descriptors calculated using the GRID approach) to fit with experimentally obtained physical and chemical properties such as log$P$, pK_a and solubility measures (Taskinen and Yliruusi 2003). Hence, in contrast to some molecular descriptor software and reviews, which had categorized this class of molecular descriptors as 0D, 1D, 2D or 3D. Thus, in this chapter we have placed this class of descriptors separately. These descriptors (e.g. log$P$, log$D$, pK_a) play a major role in both pharmacodynamic and pharmacokinetic properties of compounds (Taskinen and Yliruusi 2003). Furthermore, they have now become a part of the standard checklist for assessing the drug-likeness (e.g. Lipinski's rule-of-five) and other pharmacokinetic liabilities. Moreover, they are also widely used in explaining the variation of target-based SAR data. Most proteins' structure-function modulation is mediated via salt-bridges and small molecules typically modulate the function of a protein via charge neutralization thereby leading to the disruption of salt-bridges followed by a consequent change in the structure and function of the protein (Prathipati and Saxena 2005). In this context, physicochemical properties like pK_a and other quantum chemically derived electronic properties are widely used (Manallack 2008). In spite of their widespread usage, intuitive appeal and interpretability, these descriptors remain difficult to compute. Given the importance of modelling various electronic effects (e.g. inductive, mesomeric, polar) (Thornber 1979; Patani and LaVoie 1996; Jelfs et al. 2007; O'Boyle et al. 2017b; Harding et al. 2009; Morgenthaler et al. 2007; Xing et al. 2003; Manallack 2008), it should be noted that computationally-expensive quantum chemical descriptors are often used to train models that can predict the pK_a, polarizability, etc. Thus, the development of software for computing these descriptors is an area of active research. Improvements in GPU technology have greatly accelerated the utilization of quantum chemical simulations (Patani and LaVoie 1996) for the prediction of physicochemical properties and biological activities.

8 Interpretable Learning Algorithms

8.1 Black Box Learning Methods

Kurgan et al. (2009) used the term black box models to describe the fact that machine learning models do not identify the underlying associations of individual features with the specific outcome as well as not revealing which features provide essential contribution to the observed prediction accuracy. Black box have demonstrated success in modeling a wide range of bioactivities and properties (Charoenkwan et al. 2013; Shoombuatong et al. 2015; Simeon et al. 2016a, b; Shoombuatong et al. 2015; Nantasenamat et al. 2005, 2007a).

8.1.1 Support Vector Machine

Support vector machine (SVM) (Cortes and Vapnik 1995; Burges 1998; Barakat and Bradley 2010) is a statistical learning approach and a well-known maximum margin classifier that is based on the principles of structural risk minimization (SRM). The SRM principle is utilized to seek a hypothesis function with low capacity from a nested sequence of functions that can simultaneously minimize both the true error rate (i.e. prediction error on the external set) and the empirical error rate (i.e. prediction error on the training set) as illustrated in Fig. 4.

Given a training set $D_{Tr}^{m} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_m, \mathbf{y}_m)\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in -1, +1$, the SVM classifier finds the optimal separating hyperplane that has the largest margin and satisfies the following conditions:

$$\begin{aligned} \mathbf{w}^T \varphi(\mathbf{x}_i) + \mathbf{b} \geq +1, \quad & \mathit{for} \quad \mathbf{y}_i = +1 \\ \mathbf{w}^T \varphi(\mathbf{x}_i) + \mathbf{b} \leq +1, \quad & \mathit{for} \quad \mathbf{y}_i = -1 \end{aligned} \tag{4}$$

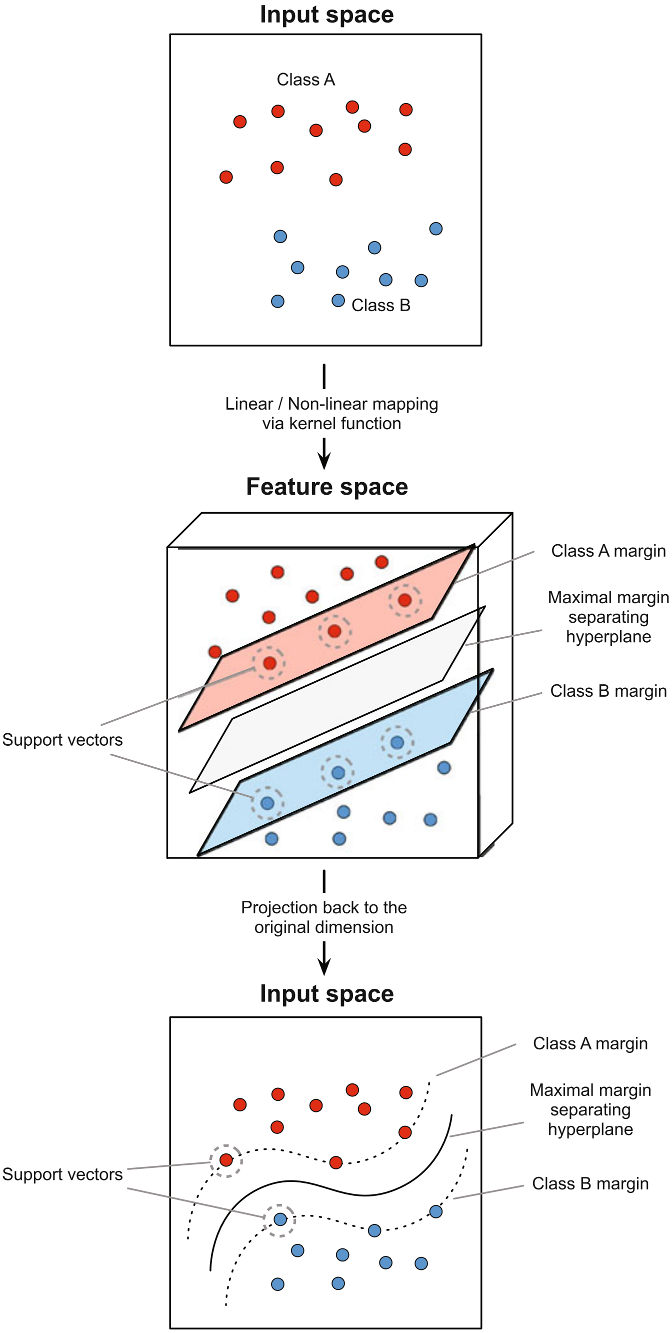


Fig. 4 Illustration of the SVM learning process. Initially, the input space is transformed to a higher dimensional feature space via the use of kernel functions whereby the maximal margin separating hyperplane is obtained after defining the margins of the two classes. It should be noted that compounds (denoted by *circles*) lying on the margin represents the support vectors

which is equivalent to:

$$\mathbf{y}_i[\mathbf{w}^T\varphi(\mathbf{x}_i) + \mathbf{b}] \geq +1, \quad i = 1, 2, ..., m \tag{5}$$

The non-linear function maps the input space to a higher dimensional space called the feature space. The mapping function $\varphi(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^p$, where $n << p$, is performed by defining the inner product between two samples through kernel function $K(\mathbf{x},\mathbf{y})$. Practically, the kernel function $K(\mathbf{x},\mathbf{y})$ is expressed with a similarity measurement between two samples in the data set, which is defined as Burges (1998):

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \varphi(\mathbf{x}) \cdot \varphi(\mathbf{y}) \\ &= \sum_i \varphi(\mathbf{x})_i \varphi(\mathbf{y})_i \end{aligned} \tag{6}$$

For the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, the most popular kernel function includes: the linear kernel $\varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j)$; the polynomial kernel $(1+\varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j))^d$, where d = 2, 3, and 4 (i.e. it should be noted that $d = 1$ for linear kernel); and the radial basis function (RBF) kernel $\exp(-\gamma(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|))$, where C (the penalty factor), γ (trading off error predictions against margin width) and ε (the percentage of support vectors in the SVM model) are parameters to be optimized. Kernel functions are often used in SVM because of the scalar product in the dual form. In fact, these approaches can also be used for other machine learning algorithms, but they are not tied to the SVM formalism. It should be noted that the RBF kernel has been widely used in SVM modelling. The decision function of the SVM classifier is given by:

$$y(x) = sign[\sum_{i=1}^{m} \alpha_i \mathbf{y}_i K(\mathbf{x}_i, \mathbf{x}) + \mathbf{b}] \tag{7}$$

where α is the parameter solved by the Lagrangian algorithm and $\mathbf{x} = (x_1, x_2, ..., x_M)$.

This method was not originally developed as a tool for statistical prediction by Cortes and Vapnik (1995). However, Vapnik enabled the original SVM to solve regression problems also known as support vector regression (SVR), by choosing a suitable cost function (ε-insensitive loss function) that enables a sparse set of support vectors to be obtained. The standard regression procedures is to identify a function f(x) that provides the least square error between predicted and actual observed responses for all training data set. In contrast, SVR attempts to minimize the generalization error bound for achieving higher generalization performance. This generalization error bound is derived from the combination of the training error and a regularization term controlling the complexity of hypothesis space. The first term is calculated by the ε-insensitive losses. The ε-insensitive loss function for SVR method (Drucker et al. 1996; Song et al. 2002) is defined as follows:

$$L_\varepsilon(\mathbf{y}, f(\mathbf{x}, \beta)) = \begin{cases} |\mathbf{y} - f(\mathbf{x}, \beta)| - \varepsilon, & |\mathbf{y} - f(\mathbf{x}, \beta)| \geq \varepsilon \\ 0 \qquad\qquad\qquad , & |\mathbf{y} - f(\mathbf{x}, \beta)| < \varepsilon \end{cases} \tag{8}$$

where $\mathbf{y}$ is the actual value, $f(\mathbf{x}, \beta)$ is the predicted value (i.e. in which the simple form is f($\mathbf{x}$), where $\mathbf{x} = (x_1, x_2, ..., x_n)$) and ε is the insensitivity parameter.

8.1.2 Artificial Neural Network

Artificial neural network (ANN) is a well-established machine learning algorithm for establishing QSAR models (Nantasenamat et al. 2005, 2007a, b, 2008; Worachartcheewan et al. 2009). ANN represents biologically inspired prediction and classification methods whose original development was based on the structure and function of the network of neurons (Zurada 1992). A typical ANN is established with three major components, namely the transfer function, the learning rule and the connection formula (Simpson 1990) as illustrated in Fig. 5. Until now, the feed-forward ANN (FF-ANN) is the most popular ANN that has been used in real-life situation (Ebrahimi et al. 2016). Among many learning algorithm for estimating the parameter of FF-ANN, the back-propagation (BP) algorithm is the most extensively used for finding the optimal parameters, which is carried out by minimizing the error of the network through the derivatives of the error function. For a given training set D_{Tr}^{m} in a BP-ANN task, the input layer starts to propagate the signal through the connection

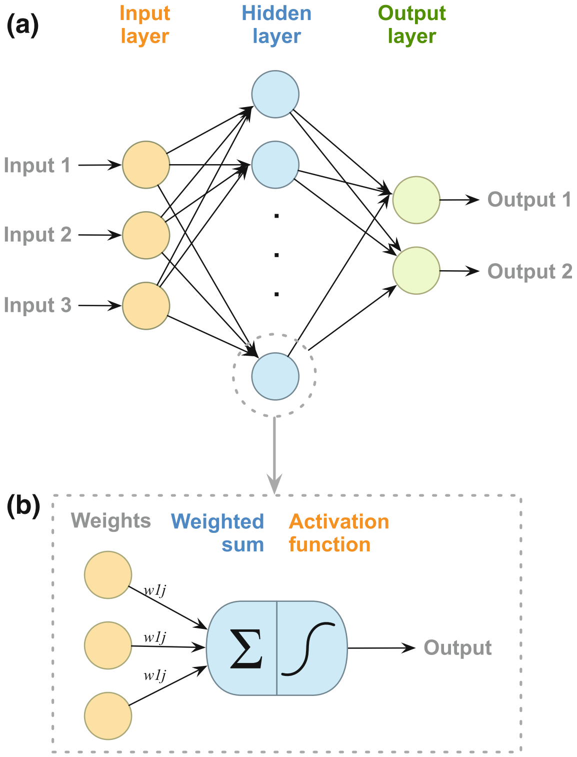


Fig. 5 Illustration of the architecture of artificial neural network (**a**) and inner working of neurons in a hidden layer (**b**)

weights and the transfer function to produce the output for each neuron. The output or predicted value is then compared to the actual value and the differences in the value between the predicted and actual values is minimized by the BP algorithm. Practically, the delta rule is used to optimize the weights via the BP algorithm:

$$W_{ij}^{new} = W_{ij}^{old} + \triangle W_{ij} \tag{9}$$

$$\triangle W_{ij} = -\mu \frac{\partial E_p}{\partial W_{ij}} out_j \tag{10}$$

where out_j is the output of the jth neuron, μ is the training rate and E_p is the error. The output layer of ANN can be represented mathematically as:

$$O = f(\sum_{i=1}^{M} \mathbf{w}_i \mathbf{x}_i + \mathbf{b}) \tag{11}$$

8.1.3 Deep Learning

Owing to the limitations of FF-ANN, a deep learning (DL) method was proposed by three separate groups (Hinton et al. 2006; Raiko 2012; Bengio 2009) for solving the process of training models in many layers. In 2006, DL also known as deep neural network has become increasingly popular for parameter approximation by allowing computational models to learn from representations of data using multiple levels of abstraction (Hinton et al. 2006, 2012). Many research groups reported that there are many different points between ANN and DL (Xing et al. 2003; Leung et al. 2014; Ma et al. 2015). Firstly, each layer of the neural network is constructed from a row of neurons while DL is built from several layers of neurons. Layers in a DL consist of three main layers: (i) the input layer (i.e. the bottom layer), where the descriptors of a molecule are entered; (ii) the output layer (i.e. the top layer), where prediction results are created; (iii) the hidden (middle) layers, where the word "deep" in DL implies that there is more than one hidden layer, as illustrated in Fig. 6. There are two popular choices of activation functions (f) that are used in the hidden (f_H) and output (f_O) layers, namely the sigmoid function and the rectified linear unit (ReLU) function. Secondly, the output layer of ANN basically has one or more neurons and each output neuron generates prediction for a separate endpoint while DL can naturally model multiple endpoints at the same time. Finally, DL employs ReLU instead of sigmoids (i.e. usually used in ANN) as activation functions in order to overcome the vanishing gradient problem. These activation functions have non-vanishing derivative.

Previously, many reports suggested that the predictive performance of DL has dramatically improved as compared to that of standard ANN. The strength of DL lies in its ability to manipulate the intricate structure in large training set by using the backpropagation algorithm. Presently, DL is being applied to many domains of science, business and government. For instance, in the domain of bioinformatics,

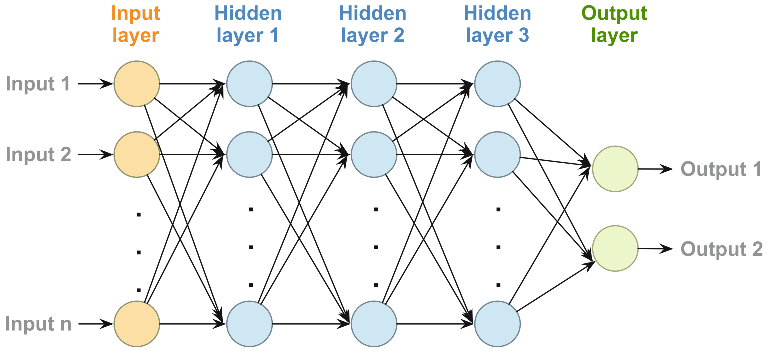


Fig. 6 Illustration of the architecture of deep learning algorithm

DL has been compared with other conventional machine learning algorithm for predicting the activity of potential drug molecules (Ma et al. 2015), analysing particle accelerator data (Ciodaro et al. 2012), reconstructing brain circuits (Helmstaedter et al. 2013) and predicting the effects of mutations in non-coding DNA on gene expression and disease (Xing et al. 2003; Leung et al. 2014). DL has also yielded promising results in natural language processing (NLP) (Collobert et al. 2011), especially for topic classification, sentiment analysis, question answering and language translation (Bordes 2014; Sutskever et al. 2014).

8.1.4 Towards Opening the Black Box

The classical QSAR approach developed by Hansch in the 1960s (Hansch et al. 1962) has a long history in predicting biological activities and physical properties. The original model used a simple, transparent and interpretable MLR model and provided excellent mechanistic interpretation of the biological activity. However, QSAR models are expected to provide both quick predictions (i.e. in a prospective manner) and mechanistic interpretation (i.e. through its features in a retrospective manner). The superior performance of SVM and ANN models vis-a-vis other computational-based models in a variety of application areas is widely known. The high accuracy and robustness of these methods can be attributed to their ability to build non-linear, black-box models that can account for the complexity of the input data. This inability to provide an explanation or comprehensible justification for the predicted solutions critically limits their application to several areas. In application areas such as medical diagnosis, it is highly desirable to give a clear mechanistic interpretation associated with the classification decisions in order to aid the compliance by both the physician and the patient. To mitigate this problem, methods that can aid the interpretation of significant features used by the model can be obtained via the use of rule extraction methods as had recently been shown for ANNs (Fung et al. 2005; Andrews et al. 1995; Setiono et al. 2002) and SVMs (Andrews et al. 1995; Barakat and Bradley

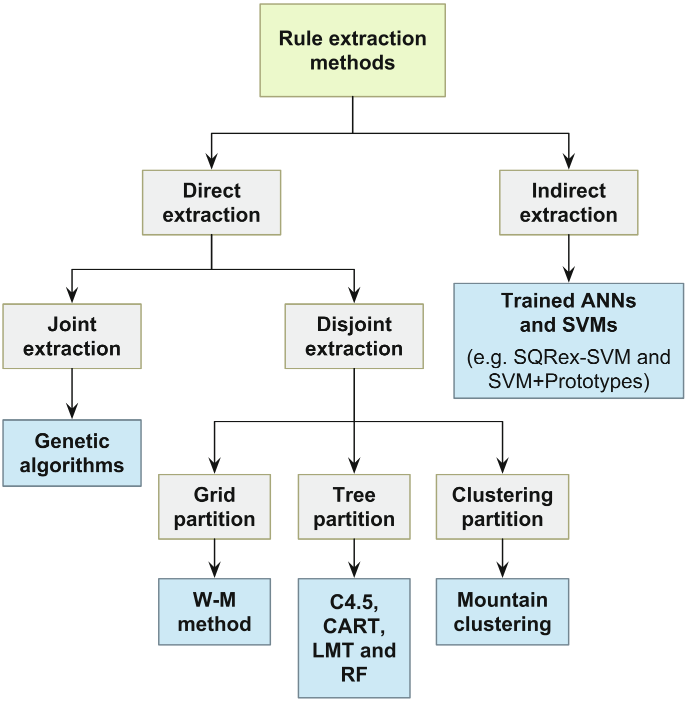


Fig. 7 Taxonomy of rule extraction techniques

2010; Núñez et al. 2002; Zhang et al. 2005; Fu et al. 2004; Barakat and Diederich 2004, 2005).

In recent years, many rule extraction techniques were developed to extract easy-to-understand regularities from data. Figure 7 illustrates the taxonomy of those methods that are derived from the data mining research community. Firstly, they are divided into direct and indirect methods according to the approach that rules are reasoned out. As mentioned, indirect rule extraction methods (e.g. SQRex-SVM and SVM+Prototypes) have been developed for providing explanations as well as affording prediction. Direct rule extraction methods are more widely studied in theory and applied in practice. The direct extraction of rules contains two critical tasks namely antecedent (i.e. representing the condition part of rules) and consequent (i.e. defining the behavior within each region) identifications. Based on the approach that these two tasks are carried out, methods to extract rules are further divided into two groups consisting of joint methods and disjoint methods. Joint methods, such as GA (Lawrence 1991), simultaneously identifies the antecedent and consequent by exceeding the capabilities of most optimization algorithms as they can afford the capability of finding global optimal solutions by mimicking biological evolution. As for disjoint methods, the divide and conquer approach is used as the strategy for optimizing the following two tasks: separating and identifying advantages over joint

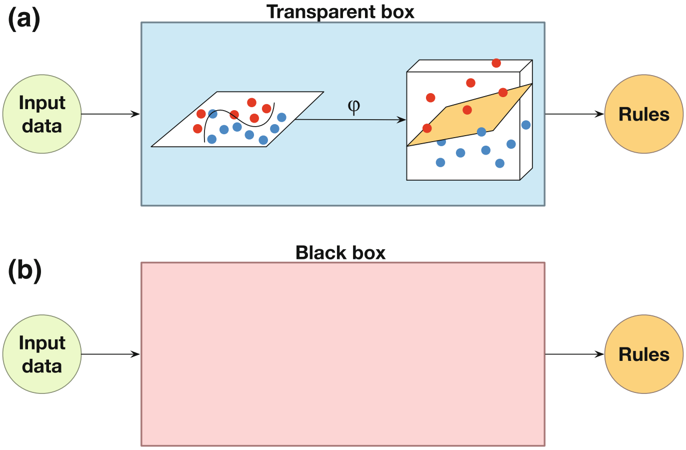


Fig. 8 System flowchart of decompositional and pedagogical rule extraction techniques

ones in computational efficiency. There are three methods that are widely used for partition namely grid (e.g. Wang-Mendel (WM) method (Wang and Mendel 1992)), tree partition (e.g. C4.5 (Quinlan 1993), classification and regression trees (CART), logistic model tree (LMT) and random forest (RF) (Breiman 2001)) as well as clustering (e.g. Mountain clustering and its extension, subtractive clustering (Yager and Filev 1994)).

The interpretability of ANNs and SVMs can be obtained by extracting symbolic rules from the trained model. The rule extraction techniques are used to open up the black box approach by generating symbolic, comprehensible descriptions while maintaining the same predictive power (Martens et al. 2007). Andrews et al. (Andrews 1974; Andrews et al. 1995) proposed an approach for the rule extraction from ANN that can be easily extended to SVMs. Two approaches exist to extract rules from the black-box ANN and SVM models (Martens et al. 2007) which are the decompositional and pedagogical approaches. The decompositional approach determines rules by utilizing information from the internal components of the constructed SVM model while the pedagogical approach considers SVM model as a black box and derives its rules by relating the inputs with the outputs of the SVM model. The difference between the decompositional and pedagogical rule extraction techniques is schematically illustrated in Fig. 8.

For the *decompositional approach*, Setiono and Liu (1995) firstly proposed an approach to understand the ANN's results. Understanding the ANN's results through rule extraction was obtained via the use of a three-phase algorithm as follows: (i), a weight-decay back-propagation network is built such that important connections are reflected by the larger weight values; (ii) the network is pruned by deleting non-informative connections while still maintaining its predictive accuracy; (iii) rules are extracted and produced. In 1997, the decompositional technique NeuroLinear (Setiono and Liu 1997) was developed to extract oblique classification rules from neural networks comprising of one hidden layer. Kim and Lee (2000) have proposed an algorithm for feature extraction and feature combination by utilizing multilayer perceptron networks with sigmoid functions. A few years later, Gupta et al. (1999) had proposed an analytical framework for classifying existing rule extraction methods for FF-ANN. This method extracts rules by directly interpreting the strengths of the connection weights in a trained network. In the case of the decompositional method, a few research have been published for extracting rules from SVMs. For instance, Núñez et al. (2002) proposed the SVM+Prototypes method for extracting rules from SVMs. The basic idea of this approach consists of: (i) determining the decision function by means of SVM while a clustering algorithm is used to determine prototype vectors for each class; (ii) defining regions in the input space that can be transferred to if-then rules. In 2007, Barakat and Bradley (2007) proposed a novel algorithm for the rule extraction from SVMs known as SQRex-SVM. After training the SVM model, SQRex-SVM directly extracts rules from the support vectors (SVs) by using a modified sequential covering algorithm. Rules are then produced by using the rank of the most discriminative features as measured by the interclass separation.

For the *pedagogical approach*, there are a large number of studies focused on opening the black box nature of ANN as to improve their interpretability. In 1988, Saito and Nakano (1988) have proposed a workflow for medical diagnosis using rule extraction from a modified ANN. A few years later, the BRAINNE system was proposed (Sestito and Dillon 1992) for extracting rules from ANN using backpropagation algorithm. The major contribution of the BRAINNE system is that it can directly deal with continuous data as inputs without requiring discretization. Shortly afterwards, Thrun (1993) proposed the VIA method for extracting rules by mapping inputs directly to the output through the use of a generate-and-test procedure for extracting symbolic rules from ANN trained by the backpropagation algorithm. Furthermore, details on how to improve the interpretability of the black box ANN have been discussed previously (Zhou and Chen 2002; Andrews et al. 1995; Augasta and Kathirvalavakumar 2012). Similar to the case of the decompositional method, only a few studies have been reported for improving the interpretability of ANN via the pedagogical approach. For example, Trepan (Craven and Shavlik 1996) was the first to introduce the pedagogical tree extraction algorithm by extracting decision trees from trained neural networks having an arbitrary architecture. In constructing a tree, this method makes use of the best first expansion strategy to build a tree via recursive partitioning. Trepan allowed splits with at least M-of-N type of tests. At each step, a queue of leaves is further expanded into sub-trees until a stopping criterion is met. In 2007, Martens et al. (2007) proposed the use of an SVM model as an oracle to generate rules. For the convenience of the vast majority of scientists, a MATLAB toolbox for generating rules using any black box model as oracle has been implemented and made publicly available. Previously, many researchers reported that ANN and SVM rule extraction approaches had equal or higher performance when compared with the original ANN and SVM methods (Barakat and Bradley 2007; Augasta and Kathirvalavakumar 2012; Gong et al. 2008).

8.2 White Box Learning Methods

8.2.1 Multiple Linear Regression

MLR is one of the most basic method for performing regression in QSAR modeling. Given a matrix *X* of a compound of interest, the MLR model assumes that the expected value of *Y* could be expressed in the form of a linear equation as summarized below:

$$\mathbf{y}_i = \sum_{i=1}^{m} \mathbf{b}_i \mathbf{x}_i + \mathbf{b}_0 \tag{12}$$

Generally, this approach is favored for its simplicity and ease of interpretation as the model assumes that there exists a linear relationship between a set of molecular descriptors and the bioactivity. When using MLR, regression coefficients can be obtained via the use of the least squares method. The size of the coefficient may reveal the degree of influence that molecular descriptors has on the bioactivity. Moreover, a positive coefficient indicate that the respective molecular descriptors contributes positively to the bioactivity and vice versa for the negative coefficient. However, in the presence of collinear descriptors, these interpretations may be error prone. A general rule of thumb states that the sample size (i.e. number of compounds in the data set) should be at least five times the number of descriptors that are used.

8.2.2 Logistic Regression

The transformation of MLR to a logistic regression (LR), can be easily performed by representing the *Y* variable via the conditional probability of *Y* given *X* variables ($\pi(X)$) when the logistic distribution is used (Hosmer et al. 2013). The specific formula of LR is defined as follows:

$$\pi(X) = \frac{e^{b_0 + b_1 x_1 + + b_2 x_2 + \cdots + + b_M x_M}}{1 + e^{b_0 + b_1 x_1 + + b_2 x_2 + \cdots + + b_M x_M}} \tag{13}$$

where $\mathbf{b}_i$ represents the transformation of $\pi(X)$. Furthermore, the logit transformation is defined in terms of $\pi(X)$:

$$\begin{aligned} g(X) &= ln[\frac{\pi(X)}{1 - \pi(X)}] \\ &= b_0 + b_1 x_1 + + b_2 x_2 + \cdots + + b_M x_M \end{aligned} \tag{14}$$

For the MLR method, the least square approach is used to estimate unknown parameters $\mathbf{b}_i$. The basic idea of this method is to minimize the sum of square error between predicted *Y* and actual *Y* values. Unfortunately, the least square approach cannot be used to optimize $\mathbf{b}_i$ on a data having a dichotomous variable (i.e. variables that have a value of 0 or 1). As for the LR method, the maximum likelihood estimator is used to alleviate the problem of dichotomous variables. A convenient way to represent the likelihood probability function for $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} = (x_1, x_2, ..., x_M)$ and $\mathbf{y} = (y_1, y_2, ..., y_M)$ can be defined as follows:

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{15}$$

Since the data set (X, Y) is assumed to be independent variables, the likelihood probability function is used to estimate β_i in expressions summarized as follows:

$$l(b_i) = \prod_{i=1}^{M} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{16}$$

In the binomial case, where the outputs of LR is close to 0 and 1, respectively, indicates low and high probability of occurrences.

8.2.3 Efficient Linear Method

Efficient linear method (ELM) is a general-purpose learning method proposed by Shoombuatong et al. (2015) that can be used for performing both classification and regression tasks. This approach was first applied in the QSAR study of the bioactivity of aromatase inhibitors (AIs) where it has been shown to afford an interpretable model in which significant features are transparent and can be used to provide insights pertaining to the origin of its bioactivity. The main procedures of the ELM method entails the following steps:

Step 1: Prepare a training data set D_{Tr}^{M} consisting of positive and negative samples.

Step 2: Formulate a predictive model with a weighted summation $f(C)$ in the form of a linear model as follows:

$$f(C) = \sum_{i=1}^{m} \mathbf{b}_i\mathbf{x}_i + \mathbf{b}_0 \tag{17}$$

Step 3: Select informative features using the fitness function of the Akaike information criterion (AIC). Finally, features affording high feature usage is selected for the construction of a predictive model.

Step 4: Estimate the optimal parameter **b** by using the genetic algorithms (GA) with the Andrews' sine function *fitness(x)* (Andrews 1974). To obtain a reliable parameter, the fitness function utilizes a 10-fold cross-validation (10-fold CV) scheme.

Step 5: Predict the unknown P with the scoring function ($Pred(C)$) using the weighted summation and subsequently discriminate it using only the threshold as obtained from:

$$Pred(C) = \begin{cases} positive, & f(C) > threshold \\ negative, & otherwise \end{cases} \tag{18}$$

8.2.4 Principal Component Analysis

The aforementioned learning approach are supervised (e.g. MLR, ANN, or SVM) in which the SAR is discerned from a list of compounds in the training set using the function in the form of $Y = f(X)$ (i.e. Y can be computed as a function of X descriptors). As a counterpart, unsupervised learning methods aim to characterize the underlying patterns of X variables without the need for Y variable. Principal component analysis (PCA) is one of the most commonly used unsupervised learning method for multivariate data analysis that can help reveal details from the high-dimensional information hidden inside the array of numerical descriptors (Jolliffe 2002). PCA analyzes the high-dimensional and intercorrelated X variables and compresses its information into a few dimensions without much loss of the core information while filtering out the noise. Briefly, the first principal component (PC) lies along the direction of maximal data variance capturing the most variability of all possible linear combinations. Because PCA seeks the linear combination of X variables that are uncorrelated with maximal variability, the assumption can be made that the first PC contains the most core information while much of the last PCs contain the noise. PCA focuses on identifying the data structures based on measurement scales and the resulting PC weights will be larger for X variables with higher variation. Two of the most useful features of PCA are the loadings and scores values.

8.2.5 Partial Least Squares Regression

Partial least squares regression (PLSR) is a commonly used learning method for the analysis of large data sets owing to its inherent ability to handle large redundant features and readily produce interpretable regression coefficients from the predictive model. In PCA, only the X variables are considered in the multivariate analysis as it does not take into account the biological properties of compounds (i.e. the Y variable). However, PLSR makes use of the information of Y variables to maximize inter-class variance (Helland 1988). PLSR is a widely used method for constructing predictive models in which features are compressed into orthogonal latent variable or PCs. The origins of PLSR can be traced back to the non-linear iterative partial least squares (NIPALS) algorithm as proposed by Herman Wold (Helland 2001). For the principle assumption of PLSR methods, a data set with intercorrelated variables is generated and then the latent structure are projected by means of PLSR. This learning method can be used for both regression and classification tasks where dimension reduction of the original feature space is an integral part of its modeling process.

8.2.6 Decision Tree

Decision trees (DT) are tree-like graphs that model a decision, which are commonly learned by recursively splitting the set of training instances into subsets based on the instances' values for the explanatory variables (Quinlan 1993). It uses the conditional statement consisting of if-then statement, which allows us to make a prediction. In short, DT constitutes a series of split points that are known as nodes. To make a prediction, we start at the top-most root node, which represents the most important feature. From this root node, a decision threshold value leads to divergence of two subsequent nodes in which the value of the feature of interest is greater than or less than the threshold value. This process is repeated at each subsequent inner nodes until we reach one of the terminal leaf nodes, which are the prediction class (i.e. whether the compound's bioactivity is classified as either being active or inactive).

8.2.7 Random Forest

Random Forest (RF) is an ensemble of unpruned classification and regression tree (Breiman et al. 1984; Breiman 2001). RF takes advantage of two efficient machine learning methods (e.g. bagging and random feature selection). RF is a further development of bagging. Instead of using all features, RF randomly selects two-third of a training data set to build the predictor and the other one-third of the training data set, known as the out-of-bag (OOB) data set, is utilized to evaluate the performance of the predictor. Predictions are derived from the majority vote or averaging the output of all trees for classification and regression problems, respectively. To evaluate the importance for each feature f_i, the values of features f_i in the OOB data set are randomly permutated and the feature importance for f_i can then be evaluated by measuring the decrease of prediction performance of the permutated OOB data set. The prediction performance can be measured by using accuracy or Gini index. The Gini index is calculated by using the impurity of each feature that is capable of separating samples of two (or more) classes. The size of the feature subsets used is a fixed number in which the number of different features tried at each split ($\mathrm{m}_{\mathrm{try}}$) are set at $p^{1/2}$ and $p/3$ for classification and regression problems.

9 Resources and Software for Performing QSAR Modeling

In this section, we present some of the software that can be used for the construction of QSAR models. This spans molecular descriptor software, multivariate analysis software and integrated software that typically lowers the steep learning curve that are usually required to get up and running in developing QSAR models.

Prior to the construction of QSAR models, the molecular features of compounds can be discerned via the use of software for computing molecular descriptors. Table 4

Table 4 List of open source descriptor calculation software

	0D	1D	2D	3D	PCP	Availability	Ref.
CDK	✓	✓	✓	✓	✓	Java, R and Python	Guha (2017), rcdk (2017), O'Boyle and Hutchison (2008)
RCPI	✓	✓	✓	✓	✓	R	Xiao et al. (2017)
ChemmineR	✓	✓	✓		✓	R	Girke (2017)
PaDEL	✓	✓	✓	✓	✓	Java, Standalone	Yap (2017), Yap (2011)
ChemDes	✓	✓	✓	✓	✓	Web server	Cao (2017a), Dong et al. (2015)
jCompoundMapper		✓	✓	✓		Java	Hinselmann et al. (2017), Hinselmann et al. (2011)
QuBiLs-MAS			✓			Standalone	Ponce (2017a), Medina Marrero et al. (2015)
QuBiLs-MIDAS				✓		Standalone	Ponce (2017b), Garcia-Jacas et al. (2014)
Chemical Descriptors Library (CDL)	✓	✓	✓	✓	✓	C++ library	Molplex Ltd. and Sykora (2017)
ChemoPy	✓	✓	✓	✓	✓	Web server	Cao (2017b), Cao et al. (2013)
Pybel	✓	✓	✓	✓	✓	Python	O'Boyle et al. (2017a), Oldham et al. (2008)
Babel	✓	✓	✓	✓	✓	Standalone	O'Boyle et al. (2017b, 2011)

Table 5 Summary of software for performing QSAR modeling

Software	Description	Standalone	Online	Ref.
AutoWeka	Automated data mining software based on Weka machine learning package	✓		Nantasenamat et al. (2015)
AZOrange	Open source high performance machine learning in a graphical environment	✓		Stalring et al. (2011)
CDK-Taverna	Platform independent workflow environment for cheminformatics	✓		Kuhn et al. (2010)
CHARMMing	Aside from ligand docking this suite of tools supports QSAR model building		✓	Miller et al. (2008)
ChemBench	Web platform for building QSAR models		✓	Walker et al. (2010)
ChemMine	Cheminformatics and data mining tools for small molecule data analysis		✓	Backman et al. (2011)
CORAL	Software for building QSAR models using SMILES-based descriptors via Monte Carlo	✓		Benfenati et al. (2011)
DMax Chemistry Assistant	Data mining tool for QSAR, compound data analysis and virtual screening	✓		DTAI Research Group (2017)
MOE Cheminformatics and QSAR	Module for performing cheminformatics and QSAR modeling	✓		Chemical Computing Group Inc. (2017)
OCHEM	Online platform for building QSAR models	✓	✓	Sushko et al. (2011)
OCED QSAR Toolbox	QSAR application toolbox for assessing hazards of chemicals	✓		Dimitrov et al. (2016)
PASS Online	Predicts the biological activity spectra of query compounds	✓	✓	Filimonov et al. (2014)
QSARINS	QSAR modeling tool in agreement with OECD principles	✓		Gramatica et al. (2013)
QSAR Workbench	QSAR workflow tool with numerical and graphical results	✓		Cox et al. (2013)
Toxtree	Toxicity estimation using decision tree	✓		Patlewicz et al. (2008)

Table 6 Summary of software for multivariate analysis

Software	Description	License	Ref.
Benchware	Data mining software for analyzing biological and chemical data	Commercial	Certara (2017)
ChemmineR	Cheminformatics package for analyzing drug-like small molecule data in R	Free	Cao et al. (2008)
IBM SPSS	Statistical and data mining software for multivariate data analysis	Commercial	IBM (2017)
KEEL	Java-based software for performing various	Free	Alcal-Fdez et al. (2011)
KNIME	Modular data exploration and mining platform that allow users to create data flows and extend functionality via modular API	Free	Mazanetz et al. (2012)
LIBSVM	Data mining software based on SVM algorithm	Free	Chang and Lin (2011)
Neuralware	Platform for developing and deploying empirical modeling based on neural networks	Commercial	NeuralWare (2017)
Neural Network Toolbox	MATLAB package providing algorithms, functions, and tools to create, train, visualize and simulate neural networks	Commercial	The MathWorks, Inc. (2017a)
MAPLE	Mathematical and computational engine with an intuitive user interface	Commercial	Maplesoft (2017)
MATLAB	Interactive environment and programming language for performing computationally intensive tasks visual programming on Python scripting	Commercial	The MathWorks, Inc. (2017b)
PyChem	Python package for chemometric for univariate and multivariate data analysis	Free	Jarvis et al. (2006)
R	Comprehensive statistical environment for data analysis and graphics visualization	Free	Ripley (2017)
RapidMiner	Open source system for data mining with an intuitive graphical user interface	Free	RapidMiner, Inc. (2017)

(continued)

Table 6 (continued)

Software	Description	License	Ref.
SAS Enterprise Miner	Reveal insights from data mining analysis	Commercial	SAS Institute Inc. (2017)
Scikit-learn	Python package for data mining analysis	Free	Pedregosa et al (2017)
SNNS	Software simulator for neural networks on Unix workstations	Free	Zell et al. (2017)
SOM Toolbox	MATLAB package for implementation the self-organizing map algorithm and more	Free	Kohonen (2017)
Spotfire S+	Statistical programming environment for analysis large scale data as well as an interactive graphics system for creation of statistical charts	Commercial	TIBCO Software Inc. (2017)
The Unscrambler	Chemometric software for data analysis and design of experiments	Commercial	CAMO Software AS (2017)
WEKA	Java-based software for data analysis via a wide range of machine learning algorithm	Free	Frank et al. (2017)

Table 7 Comparison of machine learning packages and modules from R, Python's scikit-learn and WEKA

Methods	R package	Python's scikit-learn	Weka
SVM	e1071	SVC, NuSVC and LinearSVC	LibSVM
ANN	neuralnet	MLPClassifier and MLPRegressor	MultilayerPerceptron
DL	deeplearning	–	–
MLR	car	LinearRegression	LinearRegression
LR	logistf	LogisticRegression	Logistic
ELM	*R script*[a]	–	–
PCA	princomp	PCA	PrincipalComponents
PLSR	pls	PLSRegression	PartialLeastSquares
DT	C50	DecisionTreeClassifier and DecisionTreeRegressor	J48graft
RF	randomForest	RandomForestClassifier and RandomForestRegressorr	RandomForest

[a] http://dx.doi.org/10.6084/m9.figshare.1274030

summarizes the available software along with the dimensional type of descriptor that can be computed.

A wide range of software and tools for performing QSAR modeling are available as either standalone desktop-based application or as web-based application as summarized in Table 5.

Table 6 lists some of the software for performing multivariate analysis for computer savvy scientists as the software may require a steeper learning curve than those listed in Table 5.

Table 7 summarizes the comparison between three popular machine learning packages in three popular languages namely R, Python and Java.

10 Conclusion

In spite of certain inherent flaws, the QSAR paradigms inevitably is one of the driving forces contributing to the advancements in drug discovery and design. As with all technologies, QSAR is not perfect, however, its weaknesses and flaws are continuously being identified, solved and reformed to help shape a more robust QSAR model. Particularly, the present chapter argues for the increased use of interpretable QSAR models in drug discovery research. QSAR models were originally intended to assist medicinal chemists with design ideas that are often overlooked as a useful approach; one reason is that chemists and biologists do not understand the underlying assumptions of the predictions. Hence, we have presented several concepts pertaining to inverse QSAR techniques that can reconstruct a chemical structure with good synthetic feasibility based on features identified by QSAR models. We have also presented concepts on rule extraction methods that can unravel the black box and make interpretations of machine learning approaches. Furthermore, we reviewed the utility of various molecular descriptors in the post-genomic era of the biological data deluge. Moreover, the concept of conformal prediction have also been discussed as a novel and potentially powerful approach that can define the relative confidence or reliability of predictions made. The inherent heterogeneity and vagueness of details describing the construction of QSAR models in the literature may hinder further progress. Therefore, markup language such as QSAR-ML have been suggested as a means to solve the reproducibility of QSAR models by standardizing and demystifying the underlying details of QSAR models (i.e. addition of metadata on the source of the data set, the type of descriptors used, the machine learning employed, software names and version that are used, etc.) as well as making them exchangeable (i.e. in the context that they can be shared and readily be used by the scientific community). The availability of interpretable molecular descriptors and transparent machine learning methods presents a positive outlook for the utility of QSARs in drug discovery research. The application of several key sets of standards in QSAR modeling will further help to enhance their generalization and acceptance by the wider drug research community.

Acknowledgements This work is supported by a Research Career Development Grant (No. RSA5780031) to CN from the Thailand Research Fund; the New Scholar Research Grant (No. MRG5980220) to WS from the Thailand Research Fund; and the Swedish Research Links program (No. C0610701) to CN and JESW from the Swedish Research Council.

References

Alcal-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garca, S., Snchez, L., et al. (2011). *Journal of Multiple-Valued Logic and Soft Computing*, *17*, 255.

Andrews, D. F. (1974). *Technometrics*, *16*(4), 523.

Andrews, R., Diederich, J., & Tickle, A. B. (1995). *Knowledge-Based Systems*, *8*(6), 373.

Augasta, M. G., & Kathirvalavakumar, T. (2012). *Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, Salem, Tamilnadu* (pp. 21–23).

Backman, T. W., Cao, Y., & Girke, T. (2011). *Nucleic Acids Research*, *39*, W486.

Baell, J. B., & Holloway, G. A. (2010). *Journal of Medicinal Chemistry*, *53*(7), 2719.

Bajorath, J. (2014). *Molecular Informatics*, *33*(6–7), 438.

Balaban, A. T. (1982). *Chemical Physics Letters*, *89*(5), 399.

Barakat, N. H., & Bradley, A. P. (2007). *IEEE Transactions on Knowledge and Data Engineering*, *19*(6), 729.

Barakat, N., & Bradley, A. P. (2010). *Neurocomputing*, *74*(1), 178.

Barakat, N., & Diederich, J. (2004). *14th International Conference on Computer Theory and Applications (ICCTA'2004)*. Alexandria, Egypt.

Barakat, N., & Diederich, J. (2005). *International Journal of Computational Intelligence*, *2*(1), 59.

Benfenati, E., Toropov, A. A., Toropova, A. P., Manganaro, A., & Gonella, D. R. (2011). *Chemical Biology and Drug Design*, *77*(6), 471.

Bengio, Y. (2009). *Foundations and Trends in Machine Learning*, *2*(1), 1.

Borman, S. (1990). *Chemical and Engineering News*, *68*(8), 20.

Bordes, A., Chopra, S., & Weston, J. (2014). arXiv preprint: arXiv:1406.3676.

Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R., & Zaliani, A. (1997). *Journal of Computer-Aided Molecular Design*, *11*(1), 79.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. New York, USA: CRC Press.

Breiman, L. (2001). *Machine Learning*, *45*(1), 5.

Broto, P., Moreau, G., & Vandycke, C. (1984). *European Journal of Medicinal Chemistry*, *19*(1), 66.

Brown, N., McKay, B., & Gasteiger, J. (2006). *Journal of Computer-Aided Molecular Design*, *20*(5), 333.

Brustle, M., Beck, B., Schindler, T., King, W., Mitchell, T., & Clark, T. (2002). *Journal of Medicinal Chemistry*, *45*(16), 3345.

Burges, C. J. (1998). *Data Mining and Knowledge Discovery*, *2*(2), 121.

Cao, D. S. (2017a). ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. http://www.scbdd.com/chemdes.

Cao, D. S. (2017b). ChemoPy Descriptor Calculator. http://www.scbdd.com/chemopy_desc/index/.

Cao, Y., Charisi, A., Cheng, L. C., Jiang, T., & Girke, T. (2008). *Bioinformatics*, *24*(15), 1733.

Cao, D., Liang, Y., Xu, Q., Yun, Y., & Li, H. (2011). *Journal of Computer-Aided Molecular Design*, *25*(1), 67.

Cao, D. S., Xu, Q. S., Hu, Q. N., & Liang, Y. Z. (2013). *Bioinformatics*, *29*(8), 1092.

CAMO Software AS. (2017). The Unscrambler. http://www.camo.com/rt/Products/Unscrambler/unscrambler.html.

Capuzzi, S. J., Politi, R., Isayev, O., Farag, S., & Tropsha, A. (2016). *Frontiers of Environmental Science*, *4*, 3.

Certara. (2017). Benchware 3D Explorer. https://www.certara.com/software/molecular-modeling-and-simulation/benchware-3d-explorer/.

Chang, C. C., & Lin, C. J. (2011). *ACM Transactions on Intelligent Systems and Technology*, *2*(27), 1.

Charoenkwan, P., Shoombuatong, W., Lee, H. C., Chaijaruwanich, J., Huang, H. L., & Ho, S. Y. (2013). *PLoS One*, *8*(9), e72368.

Chemical Computing Group Inc. (2017). Molecular Operating Environment (MOE). https://www.chemcomp.com/MOE-Cheminformatics_and_QSAR.htm.

Chen, H., Carlsson, L., Eriksson, M., Varkonyi, P., Norinder, U., & Nilsson, I. (2013). *Journal of Chemical Information and Modeling*, *53*(6), 1324.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). *Journal of Medicinal Chemistry*, *57*(12), 4977.

Chiu, Y. Y., Lin, C. T., Huang, J. W., Hsu, K. C., Tseng, J. H., You, S. R., et al. (2013). *Nucleic Acids Research*, *41*(Database issue), D430.

Churchwell, C. J., Rintoul, M. D., Martin, S., Visco, D. P., Kotu, A., Larson, R. S., et al. (2004). *Journal of Molecular Graphics and Modelling*, *22*(4), 263.

Ciodaro, T., Deva, D., De Seixas, J., & Damazio, D. (2012). *Journal of Physics: Conference Series*, *368*, 012030. IOP Publishing.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). *Journal of Machine Learning Research*, *12*, 2493.

Consonni, V., Todeschini, R., & Pavan, M. (2002). *Journal of Chemical Information and Computer Sciences*, *42*(3), 682.

Cortes-Ciriano, I., Ain, Q. U., Subramanian, V., Lenselink, E. B., Mendez-Lucio, O., IJzerman, A. P., et al. (2015). *Medicinal Chemical Communications*, *6*, 24.

Cortes, C., & Vapnik, V. (1995). *Machine Learning*, *20*(3), 273.

Costello, J. C., Heiser, L. M., Georgii, E., Gonen, M., Menden, M. P., Wang, N. J., et al. (2014). *Nature Biotechnology*, *32*(12), 1202.

Cox, R., Green, D. V., Luscombe, C. N., Malcolm, N., & Pickett, S. D. (2013). *Journal of Computer-Aided Molecular Design*, *27*(4), 321.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). *Journal of the American Chemical Society*, *110*(18), 5959.

Craven, M. W., & Shavlik, J. W. (1996). *Advances in neural information processing systems* (pp. 24–30). Cambridge, USA: MIT Press.

Cros, A. F. A. (1863). Action de lalcohol amylique sur lorganisme. Ph.D. thesis, University of Strasbourg.

Crum-Brown, A., & Fraser, T. (1868). *Transactions of the Royal Society of Edinburgh*, *25*, 151.

Danishuddin, A. U. K. (2016). *Drug Discovery Today*, *21*(8), 1291.

Dearden, J., Cronin, M., & Kaiser, K. (2009). *SAR and QSAR in Environmental Research*, *20*(3–4), 241.

de Vries, S. J., van Dijk, M., & Bonvin, A. M. (2010). *Nature Protocols*, *5*(5), 883.

Destrero, A., Mosci, S., De Mol, C., Verri, A., & Odone, F. (2009). *Computational Management Science*, *6*(1), 25.

Devinyak, O., Havrylyuk, D., & Lesyk, R. (2014). *Journal of Computer-Aided Molecular Design*, *54*, 194.

Dimova, D., & Bajorath, J. (2016). *Molecular Informatics*, *35*(5), 181.

Dimitrov, S. D., Didericj, R., Sobanski, T., Pavlov, T. S., Chapkov, G. V., Chapkonov, A. S., et al. (2016). *SAR and QSAR in Environmental Research*, 1–17.

Dong, J., Cao, D. S., Miao, H. Y., Liu, S., Deng, B. C., Yun, Y. H., et al. (2015). *Journal of Cheminformatics*, *7*, 60.

Doweyko, A. M. (2004). *Journal of Computer-Aided Molecular Design*, *18*(7), 587.

Doweyko, A. M. (2008). *Journal of Computer-Aided Molecular Design*, *22*(2), 81.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., Vapnik, V. (1996). *Proceedings of the 9th International Conference on Neural Information Processing Systems, NIPS'96* (pp. 155–161). Cambridge, MA, USA: MIT Press.

DTAI Research Group (2017). DMax Chemistry Assistant. https://dtai.cs.kuleuven.be/software/dmax/.

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). *Journal of Chemical Information and Computer Sciences*, *42*(6), 1273.

Ebrahimi, E., Monjezi, M., Khalesi, M. R., & Armaghani, D. J. (2016). *Bulletin of Engineering Geology and the Environment*, *75*(1), 27.

Eklund, M., Norinder, U., Boyer, S., & Carlsson, L. (2012). In L. Iliadis, I. Maglogiannis, H. Papadopoulos, K. Karatzas, & S. Sioutas (Eds.), *Artificial Intelligence Applications and Innovations: AIAI 2012 International Workshops: AIAB, AIeIA, CISE, COPA, IIVC, ISQL, MHDW, and WADTMB, Halkidiki, Greece, September 27–30, 2012, Proceedings, Part II* (pp. 166–175). Berlin, Germany: Springer.

Eklund, M., Norinder, U., Boyer, S., & Carlsson, L. (2014). *Journal of Chemical Information and Modeling*, *54*(3), 837.

Eriksson, M., Chen, H., Carlsson, L., Nissink, J. W., Cumming, J. G., & Nilsson, I. (2014). *Journal of Chemical Information and Modeling*, *54*(4), 1117.

Esbensen, K. H., & Geladi, P. (2010). *Journal of Chemometrics*, *24*(3–4), 168.

Faulon, J. L. (1994). *Journal of Chemical Information and Computer Sciences*, *34*(5), 1204.

Faulon, J. L. (1996). *Journal of Chemical Information and Computer Sciences*, *36*(4), 731.

Faulon, J. L., Churchwell, C. J., & Visco, D. P. (2003). *Journal of Chemical Information and Computer Sciences*, *43*(3), 721.

Faulon, J. L., Collins, M. J., & Carr, R. D. (2004). *Journal of Chemical Information and Computer Sciences*, *44*(2), 427.

Faulon, J. L., Brown, W. M., & Martin, S. (2005). *Journal of Computer-Aided Molecular Design*, *19*(9–10), 637.

Feng, B. Y., Shelat, A., Doman, T. N., Guy, R. K., & Shoichet, B. K. (2005). *Nature Chemical Biology*, *1*(3), 146.

Feng, B. Y., Simeonov, A., Jadhav, A., Babaoglu, K., Inglese, J., Shoichet, B. K., et al. (2007). *Journal of Medicinal Chemistry*, *50*(10), 2385.

Feng, B. Y., & Shoichet, B. K. (2006). *Nature Protocols*, *1*(2), 550.

Filimonov, D. A., Zakharov, A. V., Lagunin, A. A., & Poroikov, V. V. (2009). *SAR and QSAR in Environmental Research*, *20*(7), 679.

Filimonov, D. A., Lagunin, A. A., Gloriozova, T. A., Rudik, A. V., Druzhilovskii, D. S., Pogodin, P. V., et al. (2014). *Chemistry of Heterocyclic Compounds*, *50*(3), 444.

Frank, E., Hall, M. & Trigg, L. Weka. http://www.cs.waikato.ac.nz/ml/weka/.

Free, S. M., & Wilson, J. W. (1964). *Journal of Medicinal Chemistry*, *7*(4), 395.

Fu, X., Ong, C., Keerthi, S., Hung, G. G., & Goh, L. (2004). In *Proceedings of IEEE International Joint Conference on Neural Networks* (pp. 291–296). Budapest, Hungary: IEEE.

Fujita, T., & Winkler, D. A. (2016). *Journal of Chemical Information and Modeling*, *56*(2), 269.

Fung, G., Sandilya, S., & Rao, R. B. (2005). *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 32–40). New York, USA: ACM.

Gallup, G. A., Gilkerson, W., & Jones, M. (1952). *Transactions of the Kansas Academy of Science*, *55*(2), 232.

Gao, H., Katzenellenbogen, J. A., Garg, R., & Hansch, C. (1999). *Chemical Reviews*, *99*(3), 723.

Garcia-Jacas, C. R., Marrero-Ponce, Y., Acevedo-Martinez, L., Barigye, S. J., Valdes-Martini, J. R., & Contreras-Torres, E. (2014). *Journal of Computational Chemistry*, *35*(18), 1395.

Garg, R., Gupta, S. P., Gao, H., Babu, M. S., Debnath, A. K., & Hansch, C. (1999). *Chemical Reviews*, *99*(12), 3525.

Garg, R., Kurup, A., Mekapati, S. B., & Hansch, C. (2003). *Chemical Reviews*, *103*(3), 703.

Geronikaki, A. A., Lagunin, A. A., Hadjipavlou-Litina, D. I., Eleftheriou, P. T., Filimonov, D. A., Poroikov, V. V., et al. (2008). *Journal of Medicinal Chemistry*, *51*(6), 1601.

Girke, T. (2017). ChemmineR: Cheminformatics toolkit for R. https://www.bioconductor.org/packages/release/bioc/html/ChemmineR.html.

Gleeson, M. P. (2008). *Journal of Medicinal Chemistry*, *51*(4), 817.

Gobbi, M., Beeg, M., Toropova, M. A., Toropov, A. A., & Salmona, M. (2016). *Toxicology Letters*, *250*, 42.

Golbraikh, A., Fourches, D., Sedykh, A., Muratov, E., Liepina, I., & Tropsha, A. (2014). *Practical aspects of computational chemistry III* (pp. 187–230). Boston, USA: Springer.

Gong, R., Huang, S. H., & Chen, T. (2008). *IEEE Transactions on Industrial Informatics*, *4*(3), 198.

Gonzlez, M. P., Tern, C., Fall, Y., Teijeira, M., & Besada, P. (2005). *Bioorganic and Medicinal Chemistry*, *13*(3), 601.

Goodarzi, M., Heyden, Y. V., & Funar-Timofei, S. (2013). *Trends in Analytical Chemistry*, *42*, 49.

Gozalbes, R., Doucet, J. P., & Derouin, F. (2002). *Current Drug Targets Infectious Disorders*, *2*(1), 93.

Gramatica, P., Chirico, N., Papa, E., Cassani, S., & Kovarich, S. (2013). *Journal of Computational Chemistry*, *34*(24), 2121.

Guha, R. (2017). CDK Descriptor Calculator GUI (version 1.4. 6). http://www.rguha.net/code/java/cdkdesc.html.

Guha, R., & Van Drie, J. H. (2008). *Journal of Chemical Information and Modeling*, *48*(8), 1716.

Gupta, A., Park, S., & Lam, S. M. (1999). *IEEE Transactions on Knowledge and Data Engineering*, *11*(6), 985.

Gütlein, M., Helma, C., Karwath, A., & Kramer, S. (2013). *Molecular Informatics*, *32*(5–6), 516.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). *Machine Learning*, *46*(1–3), 389.

Guyon, I. (2003). *Journal of Machine Learning Research*, *3*, 1157.

Hadjipavlou-Litina, D., Garg, R., & Hansch, C. (2004). *Chemical Reviews*, *104*(9), 3751.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato.

Hammett, L. P. (1937). *Journal of the American Chemical Society*, *59*(1), 96.

Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). *Nature*, *194*, 178.

Hansch, C., Leo, A., & Taft, R. (1991). *Chemical Reviews*, *91*(2), 165.

Hansch, C., Hoekman, D., & Gao, H. (1996). *Chemical Reviews*, *96*(3), 1045.

Hansch, C., Hoekman, D., Leo, A., Weininger, D., & Selassie, C. D. (2002). *Chemical Reviews*, *102*(3), 783.

Hansch, C. (2011). *Journal of Computer-Aided Molecular Design*, *25*(6), 495.

Hansch, C., & Gao, H. (1997). *Chemical Reviews*, *97*(8), 2995.

Harding, A. P., Wedge, D. C., & Popelier, P. L. (2009). *Journal of Chemical Information and Modeling*, *49*(8), 1914.

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). *Journal of Chemical Information and Computer Sciences*, *43*(2), 579.

Héberger, K., & Rajkó, R. (2002). *Journal of Chemometrics*, *16*(8), 436.

Heberger, K., & Skrbic, B. (2012). *Analytica Chimica Acta*, *716*, 92.

Helland, I. S. (1988). *Communication in Statistics: Simulation and Computation*, *17*(2), 581.

Helland, I. S. (2001). *Chemometrics and Intelligent Laboratory*, *58*(2), 97.

Helmstaedter, M., Briggman, K. L., Turaga, S. C., Jain, V., Seung, H. S., & Denk, W. (2013). *Nature*, *500*(7461), 168.

Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., & Zell, A. (2011). *Journal of Cheminformatics*, *3*(1), 3.

Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., & Zell, A. (2017). jCompoundMapper: An open source java library and command-line tool for chemical fingerprints. http://jcompoundmapper.sourceforge.net/.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). *Neural Computing*, *18*(7), 1527.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. (2012). arXiv preprint arXiv:1207.0580.

Hosmer, D. W, Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (pp. 1–33). New Jersey, USA: Wiley.

Hu, X., Hu, Y., Vogt, M., Stumpfe, D., & Bajorath, J. (2012). *Journal of Chemical Information and Modeling*, *52*(5), 1138.

IBM. (2017). IBM SPSS Software. http://www.ibm.com/analytics/us/en/technology/spss/.

Jarvis, R. M., Broadhurst, D., Johnson, H., O'Boyle, N. M., & Goodacre, R. (2006). *Bioinformatics*, *22*(20), 2565.

Jelfs, S., Ertl, P., & Selzer, P. (2007). *Journal of Chemical Information and Modeling*, *47*(2), 450.

Johnson, S. R. (2008). *Journal of Chemical Information and Modeling*, *48*(1), 25.

Jolliffe, I. (2002). *Principal component analysis*. New York, USA: Springer.

Katritzky, A. R., Kuanar, M., Slavov, S., Hall, C. D., Karelson, M., Kahn, I., et al. (2010). *Chemical Reviews*, *110*(10), 5714.

Khan, M. T., & Sylte, I. (2007). *Current Drug Discovery Technologies*, *4*(3), 141.

Kier, L. B., & Hall, L. H. (1976). *Molecular connectivity in chemistry and drug research*. New York, USA: Academic Press.

Kim, K. H. (2007a). *Journal of Computer-Aided Molecular Design*, *21*(8), 421.

Kim, K. H. (2007b). *Journal of Computer-Aided Molecular Design*, *21*(1–3), 63.

Kim, D., & Lee, J. (2000). In López de Mántaras and Plaza (Eds.), *Proceedings of the 11th European conference on machine learning* (pp. 211–219). London, UK: Springer.

Kohonen, T. (2017). SOM: Self-Organization Map. http://www.cis.hut.fi/somtoolbox/.

Krasavin, M. (2015). *European Journal of Medicinal Chemistry*, *97*, 525.

Kubinyi, H. (1988). *Quantitative Structure-Activity Relationship*, *7*(3), 121.

Kubinyi, H. (1993). *3D QSAR in drug design: Volume 1: Theory methods and applications* (Vol. 1). Dordrecht, Netherlands: Springer Science & Business Media.

Kubinyi, H. (2006). In S. Ekins (Ed.) *Computer applications in pharmaceutical research and development* (pp. 377–424). New Jersey, USA: Wiley.

Kufareva, I., & Abagyan, R. (2008). *Journal of Medicinal Chemistry*, *51*(24), 7921.

Kuhn, T., Willighagen, E. L., Zielesny, A., & Steinbeck, C. (2010). *BMC Bioinformatics*, *11*, 159.

Kurgan, L., Razib, A. A., Aghakhani, S., Dick, S., Mizianty, M., & Jahandideh, S. (2009). *BMC Structural Biology*, *9*, 50.

Kurup, A., Garg, R., & Hansch, C. (2000). *Chemical Reviews*, *100*(3), 909.

Kurup, A., Garg, R., Carini, D. J., & Hansch, C. (2001). *Chemical Reviews*, *101*(9), 2727.

Kurup, A., Garg, R., & Hansch, C. (2001). *Chemical Reviews*, *101*(8), 2573.

Kvasnicka, V., & Pospichal, J. (1996). *Journal of Chemical Information and Computer Sciences*, *36*(3), 516.

Lawrence, D., et al. (1991). *Handbook of genetic algorithms*. New York, USA: Van No Strand Reinhold.

Leung, M. K., Xiong, H. Y., Lee, L. J., & Frey, B. J. (2014). *Bioinformatics*, *30*(12), i121.

Li, Q., Wang, Y., & Bryant, S. H. (2009). *Bioinformatics*, *25*(24), 3310.

Lipnick, R. L. (1991). *Studies of narcosis*. Dordrecht, Netherlands: Springer.

Liu, S. S., Yin, C. S., Li, Z. L., & Cai, S. X. (2001). *Journal of Chemical Information and Computer Sciences*, *41*(2), 321.

Liu, H., & Motoda, H. (2007). *Computational methods of feature selection*. Boca Raton, Florida: CRC Press.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). *Journal of Chemical Information and Modeling*, *55*(2), 263.

Manallack, D. T. (2008). *Perspectives in Medicinal Chemistry*, *1*, 25.

Maplesoft. (2017). Maple. https://www.maplesoft.com/products/Maple/.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). *European Journal of Operational Research*, *183*(3), 1466.

Masand, V. H., Toropov, A. A., Toropova, A. P., & Mahajan, D. T. (2014). *Current Computer-Aided Drug Design*, *10*, 75.

Mazanetz, M. P., Marmon, R. J., Reisser, C. B., & Morao, I. (2012). *Current Topics in Medicinal Chemistry*, *12*(8), 1965.

McGovern, S. L., Caselli, E., Grigorieff, N., & Shoichet, B. K. (2002). *Journal of Medicinal Chemistry*, *45*(8), 1712.

Medina Marrero, R., Marrero-Ponce, Y., Barigye, S. J., Echeverria Diaz, Y., Acevedo-Barrios, R., Casanola-Martin, G. M., et al. (2015). *SAR and QSAR in Environmental Research*, *26*(11), 943.

Miller, B. T., Singh, R. P., Klauda, J. B., Hodoscek, M., Brooks, B. R., & Woodcock, H. L. (2008). *Journal of Chemical Information and Modeling*, *48*(9), 1920.

Molplex Ltd., & Sykora, V. (2017). Chemical Descriptors Library (CDL). https://sourceforge.net/projects/cdelib/.

Morgenthaler, M., Schweizer, E., Hoffmann-Roder, A., Benini, F., Martin, R. E., Jaeschke, G., et al. (2007). *ChemMedChem*, *2*(8), 1100.

Nantasenamat, C., Naenna, T., Isarankura-Na-Ayudhya, C., & Prachayasittikul, V. (2005). *Journal of Computer-Aided Molecular Design*, *19*(7), 509.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., & Prachayasittikul, V. (2007a). *Biosensors and Bioelectronics*, *22*(12), 3309.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Tansila, N., Naenna, T., & Prachayasittikul, V. (2007b). *Journal of Computational Chemistry*, *28*(7), 1275.

Nantasenamat, C., Piacham, T., Tantimongcolwat, T., Naenna, T., Isarankura-Na-Ayudhya, C., & Prachayasittikul, V. (2008). *Journal of Biological Systems*, *16*(02), 279.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., & Prachayasittikul, V. (2009). *EXCLI Journal*, *8*(7), 74.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., & Prachayasittikul, V. (2010). *Expert Opinion on Drug Discovery*, *5*(7), 633.

Nantasenamat, C., Worachartcheewan, A., Jamsak, S., Preeyanon, L., Shoombuatong, W., Simeon, S., et al. (2015). In H. Cartwright (Ed.), *Artificial neural networks* (pp. 119–147). New York, NY, USA: Springer.

Nantasenamat, C., & Prachayasittikul, V. (2015). *Expert Opinion on Drug Discovery*, *10*(4), 321.

NeuralWare. (2017). NeuralWare. http://www.neuralware.com/.

Núñez, H., Angulo, C., & Català, A. (2002). *10th European Symposium on Artificial Neural Networks (ESANN)*, pp. 107–112.

O'Boyle, N. M., & Hutchison, G. R. (2008). *Chemistry Central Journal*, *2*, 24.

O'Boyle, N. M., Morley, C., & Hutchison, G. R. (2008). *Chemistry Central Journal*, *2*, 5.

O'Boyle, N. M., Morley, C. & Hutchison, G. R. (2017a). Pybel. https://openbabel.org/docs/dev/UseTheLibrary/Python_Pybel.html.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). *Journal of Cheminformatics*, *3*, 33.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2017b). Open Babel: The open source chemistry toolbox. http://openbabel.org/.

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., et al. (2008). *Nature Neuroscience*, *11*(11), 1271.

Patani, G. A., & LaVoie, E. J. (1996). *Chemical Reviews*, *96*(8), 3147.

Patlewicz, G., Jeliazkova, N., Safford, R. J., Worth, A. P., & Aleksiev, B. (2008). *SAR and QSAR in Environmental Research*, *19*(5–6), 495.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. et al. (2017). Scikit-learn. http://scikit-learn.org/.

Peng, H., Long, F., & Ding, C. (2005). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226.

Poroikov, V. V., Filimonov, D. A., Ihlenfeldt, W. D., Gloriozova, T. A., Lagunin, A. A., Borodina, Y. V., et al. (2003). *Journal of Chemical Information and Computer Sciences*, *43*(1), 228.

Prachayasittikul, V., Worachartcheewan, A., Shoombuatong, W., Songtawee, N., Simeon, S., Prachayasittikul, V., et al. (2015). *Current Topics in Medicinal Chemistry*, *15*(18), 1780.

Prathipati, P., Pandey, G., & Saxena, A. K. (2005). *Journal of Chemical Information and Modeling*, *45*(1), 136.

Prathipati, P., Dixit, A., & Saxena, A. K. (2007). *Journal of Computer-Aided Molecular Design*, *92*, 29.

Prathipati, P., Ma, N. L., & Keller, T. H. (2008). *Journal of Chemical Information and Modeling*, *48*(12), 2362.

Prathipati, P., & Mizuguchi, K. (2016a). *Current Topics in Medicinal Chemistry*, *16*(9), 1009.

Prathipati, P., & Mizuguchi, K. (2016b). *Journal of Chemical Information and Modeling*, *56*(6), 974.

Prathipati, P., & Saxena, A. K. (2005). *Journal of Computer-Aided Molecular Design*, *19*(2), 93.

Pudil, P., Novovičová, J., & Kittler, J. (1994). *Pattern Recognition Letters*, *15*(11), 1119.

Ponce, Y. M. (2017a). QuBiLs-MAS. http://tomocomd.com/qubils-mas.

Ponce, Y. M. (2017b). QuBiLs-MIDAS. http://tomocomd.com/qubils-midas.

Qiu, T., Qiu, J., Feng, J., Wu, D., Yang, Y., Tang, K., et al. (2016). *Briefings in Bioinformatics*, *18*(1), 125.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, USA: Morgan Kaufmann Publishers Inc.

RapidMiner, Inc. (2017). RapidMiner. https://rapidminer.com/.

rcdk: Interface to the CDK Libraries. https://cran.r-project.org/web/packages/rcdk/index.html.

Rácz, A., Bajusz, D., & Héberger, K. (2015). *SAR and QSAR in Environmental Research*, *26*(7–9), 683.

Radoux, C. J., Olsson, T. S., Pitt, W. R., Groom, C. R., & Blundell, T. L. (2016). *Journal of Medicinal Chemistry*, *59*(9), 4314.

Raiko, T., Valpola, H., & LeCun, Y. (2012). In *Proceedings of the Fifteenth Internation Conference on Artificial Intelligence and Statistics (AISTATS). JMLR Workshop and Conference Proceedings* (Vol. 22, pp. 924–932).

Randic, M. (1975). *Journal of the American Chemical Society*, *97*(23), 6609.

Ripley, B. D. (2017). The R project in statistical computing. https://www.stats.ox.ac.uk/pub/bdr/LTSN-R.pdf.

Rogers, D., & Hahn, M. (2010). *Journal of Chemical Information and Modeling*, *50*(5), 742.

Rosenbaum, L., Hinselmann, G., Jahn, A., & Zell, A. (2011). *Journal of Cheminformatics*, *3*(1), 11.

Rucker, C., Rucker, G., & Meringer, M. (2007). *Journal of Chemical Information and Modeling*, *47*(6), 2345.

Rueda, M., Bottegoni, G., & Abagyan, R. (2009). *Journal of Chemical Information and Modeling*, *49*(3), 716.

Rueda, M., Bottegoni, G., & Abagyan, R. (2010). *Journal of Chemical Information and Modeling*, *50*(1), 186.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). *Molecules*, *17*(5), 4791.

Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2013). *Journal of Cheminformatics*, *5*(1), 27.

Saito, K., & Nakano, R. (1988). In *IEEE International Conference on Neural Networks, 1988* (pp. 255–262). IEEE.

SAS Institute Inc. (2017). SAS Enterprise Miner. http://www.sas.com/en_th/software/analytics/enterprise-miner.html.

Saxena, A. K., & Prathipati, P. (2003). *SAR and QSAR in Environmental Research*, *14*(5–6), 433.

Saxena, A. K., & Prathipati, P. (2006). *SAR and QSAR in Environmental Research*, *17*(4), 371.

Schuffenhauer, A., Brown, N., Selzer, P., Ertl, P., & Jacoby, E. (2006). *Journal of Chemical Information and Modeling*, *46*(2), 525.

Seebeck, B., Wagener, M., & Rarey, M. (2011). *ChemMedChem*, *6*(9), 1630.

Selassie, C. D., Garg, R., Kapur, S., Kurup, A., Verma, R. P., Mekapati, S. B., et al. (2002). *Chemical Reviews*, *102*(7), 2585.

Sestito, S., & Dillon, T. (1992). *Proceedings of the 12th International Conference on Expert Systems and their Applications (AVIGNON'92)* (pp. 645–656).

Setiono, R., & Liu, H. (1995). *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 1, IJCAI'95* (pp. 480–485). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Setiono, R., & Liu, H. (1997). *Neurocomputing*, *17*(1), 1.

Setiono, R., Leow, W. K., & Zurada, J. M. (2002). *IEEE Transactions on Neural Networks*, *13*(3), 564.

Shafer, G., & Vovk, V. (2008). *Journal of Machine Learning Research*, *9*, 371.

Sheridan, R. P. (2015). *Journal of Chemical Information and Modeling*, *55*(6), 1098.

Sheridan, R. P., & Kearsley, S. K. (1995). *Journal of Chemical Information and Computer Sciences*, *35*(2), 310.

Shoombuatong, W., Prachayasittikul, V., Prachayasittikul, V., & Nantasenamat, C. (2015). *EXCLI Journal*, *14*, 452.

Shoombuatong, W., Prachayasittikul, V., Anuwongcharoen, N., Songtawee, N., Monnor, T., Prachayasittikul, S., et al. (2015). *Drug Design. Development and Therapy*, *9*, 4515.

Siedlecki, W., & Sklansky, J. (1988). *International Journal of Pattern Recognition and Artificial Intelligence*, *2*(02), 197.

Simeon, S., Möller, R., Almgren, D., Li, H., Phanus-umporn, C., Prachayasittikul, V., et al. (2016a). *Chemometrics and Intelligent Laboratory Systems*, *151*, 51.

Simeon, S., Spjuth, O., Lapins, M., Nabu, S., Anuwongcharoen, N., Prachayasittikul, V., et al. (2016b). *PeerJ*, *4*, e1979.

Simpson, P. K. (1990). *Artificial neural system: Foundation, paradigm, application and implementations*. Pennsylvania, USA: Windcrest/McGraw-Hill.

Sippl, W. (2006). *Molecular interaction fields* (pp. 145–170). KGaA: Wiley-VCH Verlag GmbH & Co.

Skvortsova, M. I., Baskin, I. I., Slovokhotova, O. L., Palyulin, V. A., & Zefirov, N. S. (1993). *Journal of Chemical Information and Computer Sciences*, *33*(4), 630.

Sliwoski, G., Mendenhall, J., & Meiler, J. (2016). *Journal of Computer-Aided Molecular Design*, *30*(3), 209.

Song, M., Breneman, C. M., Bi, J., Sukumar, N., Bennett, K. P., Cramer, S., et al. (2002). *Journal of Chemical Information and Computer Sciences*, *42*(6), 1347.

Spjuth, O., Willighagen, E. L., Guha, R., Eklund, M., & Wikberg, J. E. (2010). *Journal of Cheminformatics*, *2*, 5.

Spyrakis, F., & Cavasotto, C. N. (2015). *Archives of Biochemistry and Biophysics*, *583*, 105.

Stalring, J. C., Carlsson, L. A., Almeida, P., & Boyer, S. (2011). *Journal of Cheminformatics*, *3*, 28.

Standfuss, J., Edwards, P. C., D'Antona, A., Fransen, M., Xie, G., Oprian, D. D., et al. (2011). *Nature*, *471*(7340), 656.

Stumpfe, D., Hu, Y., Dimova, D., & Bajorath, J. (2014). *Journal of Medicinal Chemistry*, *57*(1), 18.

Sushko, I., Novotarskyi, S., Krner, R., Pandey, A. K., Rupp, M., et al. (2011). *Journal of Computer-Aided Molecular Design*, *25*(6), 533.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K., & Q. Weinberger (Ed.) *Advances in neural information processing systems 27* (pp. 3104–3112). Curran Associates, Inc.

The MathWorks, Inc. (2017a). Neural Network Toolbox. http://www.mathworks.com/products/neural-network/.

The MathWorks, Inc. (2017b). MATLAB. https://www.mathworks.com/products/matlab/.

TIBCO Software Inc. (2017). TIBCO Spotfire S+. http://spotfire.tibco.com/discover-spotfire/who-uses-spotfire/by-role/statisticians.

Tarca, A. L., Than, N. G., & Romero, R. (2013). *Systems Biomedicine*, *1*(4), 217.

Taskinen, J., & Yliruusi, J. (2003). *Advanced Drug Delivery Reviews*, *55*(9), 1163.

Thornber, C. W. (1979). *Chemical Society Reviews*, *8*(4), 563.

Thorne, N., Auld, D. S., & Inglese, J. (2010). *Current Opinion in Chemical Biology*, *14*(3), 315.

Thrun, S. (1993). *Extracting provably correct rules from artificial neural networks*. Bonn, Germany: University of Bonn.

Todeschini, R., & Consonni, V. (2008). *Handbook of molecular descriptors*. Weinheim, Germany: Wiley-VCH Verlag GmbH.

Toropov, A. A., Toropova, A. P., Benfenati, E., Leszczynska, D., & Leszczynski, J. (2010). *Journal of Computational Chemistry*, *31*(2), 381.

Toropov, A. A., Toropova, A. P., Puzyn, T., Benfenati, E., Gini, G., Leszczynska, D., et al. (2013). *Chemosphere*, *92*(1), 31.

Toropova, A. P., & Toropov, A. A. (2014). *European Journal of Pharmaceutical Sciences*, *52*, 21.

Toropov, A. A., & Benfenati, E. (2007a). *European Journal of Medicinal Chemistry*, *42*(5), 606.

Toropov, A. A., & Benfenati, E. (2007b). *Current Drug Discovery Technologies*, *4*(2), 77.

Tosco, P., Balle, T., & Shiri, F. (2011). *Journal of Computer-Aided Molecular Design*, *25*(8), 777.

Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). *QSAR and Combinatorial Science*, *22*(1), 69.

Tropsha, A. (2010). *Molecular Informatics*, *29*(6–7), 476.

Venkatasubramanian, V., Chan, K., & Caruthers, J. M. (1995). *Journal of Chemical Information and Computer Sciences*, *35*(2), 188.

Verma, R. P., & Hansch, C. (2005). *Bioorganic and Medicinal Chemistry*, *13*(15), 4597.

Verma, R. P., & Hansch, C. (2009). *Chemical Reviews*, *109*(1), 213.

Visco, D. P., Pophale, R. S., Rintoul, M. D., & Faulon, J. L. (2002). *Journal of Molecular Graphics and Modelling*, *20*(6), 429.

Walker, T., Grulke, C. M., Pozefsky, D., & Tropsha, A. (2010). *Bioinformatics*, *26*(23), 3000.

Wang, L. X., & Mendel, J. M. (1992). IEEE Transactions on Systems. *Man and Cybernetics: Systems*, *22*(6), 1414.

Wei, D. B., Zhang, A. Q., Han, S. K., & Wang, L. S. (2001). *SAR and QSAR in Environmental Research*, *12*(5), 471.

Weis, D. C., Faulon, J. L., LeBorne, R. C., & Visco, D. P. (2005). *Industrial and Engineering Chemistry*, *44*(23), 8883.

Wong, W. W., & Burkowski, F. J. (2009). *Journal of Cheminformatics*, *1*, 4.

Worachartcheewan, A., Nantasenamat, C., Naenna, T., Isarankura-Na-Ayudhya, C., & Prachayasittikul, V. (2009). *European Journal of Medicinal Chemistry*, *44*(4), 1664.

Worachartcheewan, A., Mandi, P., Prachayasittikul, V., Toropova, A. P., Toropov, A. A., & Nantasenamat, C. (2014). *Chemometrics and Intelligent Laboratory Systems*, *138*, 120.

Worachartcheewan, A., Prachayasittikul, V., Toropova, A. P., Toropov, A. A., & Nantasenamat, C. (2015). *Molecular Diversity*, *19*(4), 955.

Worth, A. P., & Cronin, M. T. (2004). *Alternatives to Laboratory Animals*, *32*, 703.

Xiao, N., Cao D. S., & Xu, Q. (2017). Rcpi: Toolkit for compound-protein interaction in drug discovery. http://bioconductor.org/packages/release/bioc/html/Rcpi.html.

Xing, L., Glen, R. C., & Clark, R. D. (2003). *Journal of Chemical Information and Computer Sciences*, *43*(3), 870.

Yager, R. R., & Filev, D. P. (1994). *Journal of Intelligent & Fuzzy Systems*, *2*(3), 209.

Yap, C. W. (2011). *Journal of Computational Chemistry*, *32*(7), 1466.

Yap, C. W. (2017). PaDEL-Descriptor. http://www.yapcwsoft.com/dd/padeldescriptor.

Zakharov, A. V., Peach, M. L., Sitzmann, M., & Nicklaus, M. C. (2014). *Journal of Chemical Information and Modeling*, *54*(3), 705.

Zell, A., Mache, N., Hubner, R., Mamier, G., Vogt, M., Döring, S., et al. (2017). SNNS: Stuttgast neural network simulator. http://www.ra.cs.uni-tuebingen.de/SNNS/.

Zhao, Z., Wu, H., Wang, L., Liu, Y., Knapp, S., Liu, Q., et al. (2014). *ACS Chemical Biology*, *9*(6), 1230.

Zhang, Y., Su, H., Jia, T., & Chu, J. (2005). In Ho T. B., Cheung D., Liu H. (Eds.), *Advances in knowledge discovery and data mining: 9th Pacific-Asia conference on knowledge discovery and data mining* (pp. 61–70). Berlin/Heidelberg, Germany: Springer.

Zhou, Z. H., & Chen, S. F. (2002). *Journal of Research and Development*, *39*(4), 398.

Zurada, J. M. (1992). *Introduction to artificial neural systems* (Vol. 8). Minnesota, USA: West Publishing Co.

Symposium: ISBS2015 - Towards Novel Blood Transfusion Therapies

Date: **June 21st-24th, 2015**

Venue: **Palaestra et Odeum, Lund University**
Paradisgatan 2
223 50 Lund, Sweden

Palaestra et Odeum

University Building

Sunday June 21st, 2015

16.00 – 18.00 Registration in the University Building

17.30 Welcome address (Leif Bülow)

Monday June 22nd, 2015

Production of red blood cells for transfusion medicine (Palaestra)

8:00 – Registration and poster set-up

9:00 – 9:10 Opening message: Professor Gunilla Westergren-Thorsson (Dean of the Medical Faculty at Lund University), Johan Flygare and Kenichi Miharada

9:10 – 9:40 **Opening lecture:** Martin Olsson MD, PhD, Lund University, Sweden
Transfusion Medicine in the 21st Century - Current Challenges and Changing Concepts

Session 1: ***In vitro* red blood cell production I (Palaestra)**
Chair: Johan Flygare

9:40 – 10:30 James Palis, MD, University of Rochester Medical Center, NY, USA
"TBA"

10:30 – 11:00 Coffee Break

11:00 – 11:50 Luc Douay, MD PhD, Université Pierre et Marie Curie-Paris, France
Progress in in-vitro red cell generation - are we there yet

11:50 – 12:00 Photo session

12:00 – 13:30 Lunch (served in Palaestra)

12:15 – 13:00 Round Table Discussion *The Dilemmas of the Academic Entrepreneur - License the Technology to Big Pharma or Start a New Company?*
Chairman: Bo Hedlund, Biomedical Frontiers Inc, Minneapolis, Minnesota

Session 2: *In vitro* red blood cell production II (Palaestra)
Chair: Johan Flygare

13:30 – 14.00 Emile van den Akker, PhD, Sanquin Institute, The Netherlands
"*TBA*"
14.00– 14.30 Kenichi Miharada, PhD, Lund University, Sweden
Production of mature red blood cells from immortalized erythroid cell lines
14:30 – 15:00 Joanne Mountford, PhD, University of Glasgow, UK
"TBA"

15:00 – 15:30 Coffee Break

Session 3: *In vitro* red blood cell production (Palaestra)
Chair: Kenichi Miharada

15:30 – 16:00 Marieke von Lindern, PhD, Sanquin Institute, The Netherlands
"TBA"
16:00 – 16:30 Johan Flygare, MD PhD, Lund University, Sweden
Generation of red blood cells from reprogrammed fibroblasts
16:30 – 17:00 Ashley Toye, PhD, University of Bristol, UK
"TBA"

17:00 – 17:15 Adam Sidaway, University of Bristol, UK
Analysis of the transcriptome of erythroid cells to identify factors that promote proliferation

17:15 – 17:30 Jan Frayne, PhD, University of Bristol, UK
"The first human immortalised cell line generated from adult erythroid cells"

17:30 – 17:45 Marieangela Wilson, PhD, University of Bristol, UK
"The development of methodology for proteome-wide profiling, and

comparative analysis of transcription factors in erythroid cells”

Free Evening

Session 4: ***HBOCs from industrial laboratories, in parallel with Sessions 2 and 3 (Eden).***

Chairs: Thomas Chang and Leif Bülow

13.30 – 13.55 Gord Adamson, Therapure Biopharma Inc., Mississauga, Canada
Hemoglobin-Based Products for Drug Delivery and Organ Perfusion

13.55 – 14.20 Franck Zal, Hemarina SA, Morlaix, France
Optimisation of Mesenchymal Stem Cell functions and proliferation: Investigation of the benefits of a new oxygen carrier, HEMOXCell®, in platelet lysate-supplemented media

14.20 – 14.45 Peter Keipert, Keipert Corp, San Diego, USA
Clinical evaluation of MP4OX as an oxygen therapeutic adjunct to acute resuscitation of trauma patients in severe hemorrhagic shock

14.45 – 15.10 Jan Simoni, Texas HemoBioTherapeutics & BioInnovation Center, Lubbock, USA
Current Status of Clinical Development of Hb-Based Oxygen Therapeutic with Pharmacologic Properties of ATP, Adenosine and Reduced Glutathione (GSH)

15.10 – 15.40 Coffee break

15.40 - 16.05 Carleton Hsia, NanoBlood LLC, Sioux Falls, USA
The Second Quantum Revolution - nanoQuantum Medicine - The Introduction of nano Red Blood Cell (nanoRBC) for Use in Critical and Chronic Care Medicine

16.05 - 16.30 Kim D. Vandegriff, NovoSang, Inc., San Diego, CA, USA
Hemoglobin extravasation in the brain of rats exchange-transfused with hemoglobin-based oxygen carriers

16. 30 – 16.55 Abraham Abuchowski, Prolong Pharmaceuticals. South Plainfield, USA
Sanguinate™: A CO/O2 delivery therapeutic for the treatment of anemic and ischemic disorders

16.55-17.20 Jonathan S Jahr, Anesthesiology and Perioperative Medicine, David Geffen School of Medicine at UCLA, USA

Post Marketing Analysis of Safety and Efficacy of Hemoglobin Glutamer-250 (Bovine), [HEMOPURE®, HBOC-201)] in South Africa, On Behalf of the HEMOPURE South African Task Force

17.20 - Abdu I Alayash, Center for Biologics Evaluation and Research, Food and Drug Administration, USA
Tribute - Dr. Joseph Fratantoni

Free evening

Tuesday June 23rd, 2015

Hemoglobin Production, Engineering and HBOC Design

Session 5:
Chairs: Michael Wilson and Chengmin Yang

8.30-9.00 Abdu I Alayash, Center for Biologics Evaluation and Research, Food and Drug Administration, USA
Oxidative Pathways in Hemoglobin: Do They Really Matter?

9.00-9.30 John S Olson, Rice University, USA
Protein Engineering Strategies for Recombinant Hb-Based Oxygen Carriers (rHBOCs): Compromises between Hemoglobin Stability, Production, Function, and Toxicity In Vivo

9.30-9.45 José Luis Martinez, Chalmers, Sweden
Production of recombinant human hemoglobin by microbial fermentation through yeast metabolic engineering

9.45-10.00 Khuanpiroon Ratanasopa, Lund University, Sweden
Genetically Linked Human fetal Hemoglobin: Strategy to enhance recombinant hemoglobin production and produce HBOCs.

10.00 – 10.30 Coffee break

Chairs: Ken W Olsen and Peter Keipert

10.30-11.00 Chengmin Yang, Institute of Blood Transfusion, CAMS & PUM, PR China
Some Thoughts on Research and Development of HBOCs in China

11.00-11.30 Hiromi Sakai, Nara Medical University, Japan
Translational Research of Hemoglobin-vesicles as a Transfusion Alternative

11.30-12.00 Lian Zhao, Transfusion Medicine, Academy of Military Medical Sciences, Beijing, PR China
Construction of blood-compatible hemoglobin-loaded nanoparticles as oxygen carriers for in vivo oxygenation

12.00-12.15 Ka Zhang, Lund University, Sweden
Characterization of human hemoglobin-imprinted polymer beads–recognition and protection against lipid peroxidation

12.15 – 13.10 **Lunch (served in Palaestra)**

12.20 – 13.05 **Round Table Discussion** *"The Dilemmas of the Academic Entrepreneur - License the Technology to Big Pharma or Start a New Company?"* **Chairman:** Bo Hedlund, Biomedical Frontiers Inc, Minneapolis, Minnesota

Session 6:
Chairs: Chris Cooper and Lian Zhao

13.10-13.40 Thomas MS Chang, McGill University, Montreal, Canada
Red Blood Cell replacement or nanobiotherapeutics with enhanced red blood cell functions?

13.40-14.05 Jiaxin Liu, Blood Transfusion, CAMC, Chengdu, P. R. China
Resuscitation with polymerized human placenta hemoglobin improves tissue oxygenation and survival in a rat hemorrhagic shock model

14.05-14.30 Hans Bäumler, Transfusion Medicine, Charité-Universitätsmedizin Berlin, Germany
Novel Hemoglobin Particles—Promising New-Generation Hemoglobin-Based Oxygen Carriers

14.30-14.55 Hae Won Kim, Brown University, Providence, USA
A new collaboration model for facilitated HBOC development

14.55 – 15.20 Coffee break

Chairs: Hiromi Sakai and Hans Bäumler

15.20-15.40 Ken W Olsen, Loyola University Chicago, USA
Development of Inside-Out PEGylated Crosslinked Hemoglobin Polymers

15.40-15.55 Sandeep Chakane, Lund University, Sweden
A recombinant unstructured polypeptide to increase stability of fusion HbF: an alternative to PEGylation

15.55-16.15 Chanin Nantasenamat, Mahidol University, Bangkok, Thailand
Predicting the oxygen affinity of human hemoglobin

16.15-16.35 YanJon Guo, National Center for Nanoscience and Technology, Beijing PR China
Application of Raman spectroscopy in the structure and function

study of hemoglobin

16.35-17.00 Leif Bülow, Lund University, Sweden
From the white buses to the white trucks - Or how to convert HBOCs into viable products

17.45 Cathedral concert

18.30 Buses for Conference Dinner leaving outside the Cathedral

19.15 -22.00 Kulturens Östarp

Wednesday June 24th, 2015

Extracellular hemoglobin and Protection Proteins

Session 7:
Chairs: Abdu Alayash, Brandon Reeder

8.00- Registration

8.30 - 8.40 Magnus Gram, Skåne University Hospital, Lund, Sweden
Extracellular hemoglobin, why is it toxic?

8.40 – 9.10 Stefan Hansson, Lund University, Lund, Sweden
Clinical Situation: Hemoglobin toxicity in Preeclampsia

9.10 – 9.40 Faikah Güler, Hannover Medical School, Hannover, Germany
Clinical Situation: Kidneys – a major target for hemoglobin and heme toxicity

9.40 - 10.10 Jozsef Balla, University of Debrecen, Debrecen, Hungary
Heme, heme oxygenase, and ferritin:how the vascular endothelium and smooth muscle cells survive (and die) in an iron-rich environment.

10.10 – 10.40 Coffee Break

10.40 - 11.10 Ann Smith, University of Missouri, Kansas City, USA
Mechanisms of heme toxicity and hemopexin as a therapeutic

11.10 - 11.40 Sören Moestrup, Aarhus University, Aarhus, Denmark
Receptor systems for hemoglobin and heme – role in inflammation

11.40 – 12.10 Isaac K Quaye, University of Namibia School of Medicine, Windhoek, Namibia
Plasticity in macrophage function: A pathway for intervention in hemoglobin release?

12.10 - 13.10 Lunch (served in Palaestra)

13.10 - 13.40 Bo Åkerström, Lund University, Lund, Sweden
A1M – an extravascular heme-binding and tissue cleaning protein with therapeutic potential

13.40 - 14.10 Chris Cooper, Essex University, Essex, UK
Modification of Tyrosine Electron Transfer Pathways in Hemoglobin decreases oxidative reactivity

14.10 - 14.40 Willem A Buurman, Maastricht University, Netherlands
Hemolysis during cardiac surgery

14.40-14.50 Closing remarks

14.50 -15.00 Thomas MS Chang
ISBS2017

Predicting the oxygen affinity of human hemoglobin

Watshara Shoombuatong [(1)], Virapong Prachayasittikul [(2)],
*Leif Bülow [(3)], Chanin Nantasenamat [(1)]**

[(1)]Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology,
Mahidol University, 2 Prannok Rd, Bangkok 10700, Thailand
[(2)]Department of Clinical Microbiology, Faculty of Medical Technology,
Mahidol University, 2 Prannok Rd, Bangkok 10700, Thailand
[(3)]Pure and Applied Biochemistry, Chemical Center,
Lund University, Getingevägen 60, Lund 221 00, Sweden

*Corresponding author: chanin.nan@mahidol.ac.th

Abstract

Human hemoglobin (Hb) is instrumental in the transportation of oxygen from lungs to tissues. In spite of several decades of investigations, the structural basis of oxygen binding has yet to be fully elucidated. Therefore, a comprehensive study on the physicochemical properties contributing to oxygen binding was performed herein on a large set of 331 human Hb variants harboring a single point mutation. Statistical and multivariate analysis was performed to gain insights into the origins of low and high oxygen binding affinities in human Hb. This study proposes the use of a Random Forest classifier [1] for predicting the oxygen affinity of Hb variants as a function of their sequence-based descriptors comprising of amino acid composition, dipeptide composition and physicochemical properties from the AAIndex [2]. Prediction results from 5-fold cross-validation showed that our proposed method performed well with average accuracy, precision and recall in excess of 80. Results revealed that informative physicochemical properties were pertaining to beta-strands, aperiodic indices for alpha/beta-proteins and hydrophobicity. The investigation presented herein provide useful insights on physicochemical properties giving rise to oxygen binding affinities which may guide further structure-based design of novel human Hb variants with desired oxygen binding characteristics.

[1] Breiman L., Random forests. Machine Learning, 45: 5–32, 2001

[2] Kawashima, S., Ogata, H., and Kanehisa, M., AAindex: amino acid index data base. Nucleic Acids Res, 27, 368-369,1999.

Keywords: hemoglobin; oxygen; oxygen affinity; data mining.

สถาบันชีววิทยาศาสตร์โมเลกุล มหาวิทยาลัยมหิดล
25/25 ถนนพุทธมณฑล สาย 4
ตำบลศาลายา อำเภอพุทธมณฑล
จังหวัดนครปฐม 73170

ที่ ศธ 0517.22/425

วันที่ 8 เมษายน 2559

เรื่อง ขอเรียนเชิญเป็นวิทยากร

เรียน รองศาสตราจารย์ ดร. ชนินทร์ นันทเสนามาตร์

สิ่งที่ส่งมาด้วย (ร่าง) กำหนดการประชุมวิชาการ

ด้วยสถาบันชีววิทยาศาสตร์โมเลกุล มหาวิทยาลัยมหิดล กำหนดให้มีการประชุมวิชาการ ประจำปี 2559 ภายใต้หัวข้อ *"Systems Biosciences : frontiers in integrative research"* ในวันที่ 19 พฤษภาคม 2559 เวลา 9.00-16.00 น. ณ ห้องประชุม ณ ห้องประชุมใหญ่ (ศาสตราจารย์เกียรติคุณสิรินทร์ พิบูลนิยม) ชั้น 1 สถาบันชีววิทยาศาสตร์โมเลกุล มหาวิทยาลัยมหิดล เพื่อเป็นเวทีในการแลกเปลี่ยนเรียนรู้ประสบการณ์ ความคิดเห็นทางด้านชีววิทยาระบบ (Systems biology)

ในการนี้ สถาบันชีววิทยาศาสตร์โมเลกุล มหาวิทยาลัยมหิดล พิจารณาแล้วเห็นว่าท่านเป็นผู้ทรงคุณวุฒิที่มีความรู้ ความสามารถ ด้าน Systems biology ดังนั้น ทางสถาบันฯ จึงใคร่ขอเรียนเชิญท่านเป็นวิทยากรบรรยาย(ภาษาอังกฤษ) ช่วง **Symposium I:** Systems Bioscience and Medicine **เวลา 10.30-11.00 น.** ในหัวข้อเรื่อง "Towards Interactive and Reproducible QSAR Models for Studying the Origins of Bioactivity" ดังรายละเอียดตาม (ร่าง) กำหนดการที่แนบมาพร้อมนี้

จึงเรียนมาเพื่อโปรดให้เกียรติเป็นวิทยากร ตามวัน เวลา และสถานที่ ดังกล่าวข้างต้น ด้วย จะเป็นพระคุณยิ่ง

ขอแสดงความนับถือ

(ศาสตราจารย์ นพ.ประเสริฐ เอื้อวรากุล)
ผู้อำนวยการสถาบันชีววิทยาศาสตร์โมเลกุล
มหาวิทยาลัยมหิดล

ผู้ประสานงาน : น.ส.ใกล้รุ่ง ศรีก๊กเจริญ
โทรศัพท์ 0 2441 9003-7 ต่อ 1208
อีเมลล์: klairung.sri@mahidol.ac.th

Mahidol University
Institute of Molecular Biosciences

Systems Biosciences
: Frontiers in Integrative Research

May 19th, 2016 | 8.30 a.m. - 4.00 p.m.

at the Professor Emeritus Sirin Piboonniyom Auditorium, Institute of Molecular Biosciences, Mahidol University

09.00 a.m. The 14th Prof. Emeritus Sirin Piboonniyom Keynote Speech on
"The Road from Traditional Biochemistry to Modern Trends in Bioscience Research"
by Prof. Emeritus Dr. M.R. Jisnuson Svasti, Chulabhorn Research Institute.

10.30 a.m. **Symposium I: Systems Biosciences and Medicine**

- "Towards Interactive and Reproducible QSAR Models for Studying the Origins of Bioactivity"
  by Assoc. Prof. Dr. Chanin Nantasenamat, Faculty of Medical Technology, Mahidol University
- "In Depth, Multidisciplinary & Big Picture of Ligand Design: MANORAA.org"
  by Dr. Duangrudee Tanramluk, Institute of Molecular Biosciences, Mahidol University
- "Personalized Cancer Medicine in Thailand: A Focus on Colorectal Cancer Gene Sequencing"
  by Dr. Natini Jinawath, Faculty of Medicine Ramathibodi Hospital, Mahidol University

01.00 p.m. **Symposium II: Frontiers in Integrative Research**

- "Proteomics: Mass Spectrometer for Peptide and Peptidome Research"
  by Dr. Sittiruk Roytrakul, National Center for Genetic Engineering and Biotechnology (BIOTEC) and National Science and Technology Development Agency (NSTDA)
- "Epigenomics: Long Non-Coding RNA and Cellular Diversity: An Epigenomic Approach"
  by Dr. Patompon Wongtrakoongate, Faculty of Science, Mahidol University
- "Microbiomes: Embrace Our Second Genome"
  by Dr. Poochit Nonejuie, Institute of Molecular Biosciences, Mahidol University

03.00 p.m. **Symposium III: Applications in Systems Biology**

- "Omics in Biomedical Sciences"
  by Prof. Emeritus Dr. Suthat Fuchareon, Institute of Molecular Biosciences, Mahidol University
- "Proteomics in Agricultural Sciences"
  by Assoc. Prof. Chartchai Krittanai, Institute of Molecular Biosciences, Mahidol University

For more information and online registration please visit :
www.mb.mahidol.ac.th/conference2016

Conference Schedule

"Systems Biosciences: Frontiers in Integrative Research"

on May 19th, 2016

At the Professor Emeritus Serene Piboonniyom Auditorium,
Institute of Molecular Biosciences, Mahidol University

08.30 - 08.45am	**Registration**
08.45 - 09.00am	**Opening Ceremony** By *Prof. Prasert Auewarakul, M.D.* Director, Institute of Molecular Biosciences, Mahidol University.
09.00 - 10.00am	**The 14th Prof. Emeritus Serene Piboonniyom Keynote Speech on "The Road from Traditional Biochemistry to Modern Trends in Bioscience Research"** Keynote Speaker: *Prof. Emeritus Dr. M.R. Jisnuson Svasti,* Chulabhorn Research Institute.
10.00 - 10.15am	**Presentation of the 2015 Distinguished Alumni Awards**
10.15 - 10.30am	**Break**
10.30 - 12.00pm	**Symposium I: Systems Biosciences and Medicine**
	• **"Towards Interactive and Reproducible QSAR Models for Studying the Origins of Bioactivity"** Speaker: *Assoc. Prof. Dr. Chanin Nantasenamat* Faculty of Medical Technology, Mahidol University • **"In Depth, Multidisciplinary & Big Picture of Ligand Design: MANORAA.org"** Speaker: *Dr. Duangrudee Tanramluk* Institute of Molecular Biosciences, Mahidol University • **"Translational Cancer Research in Thailand: a Focus on Colorectal Cancer"** Speaker: *Dr. Natini Jinawath, M.D.* Faculty of Medicine Ramathibodi Hospital, Mahidol University
	Moderator: *Dr. Natini Jinawath, M.D.* Faculty of Medicine Ramathibodi Hospital, Mahidol University
12.00 - 13.00pm	**Lunch**

13.00 - 14.30pm	**Symposium II: Frontiers in Integrative Research**
	• **"Proteomics: Mass Spectrometer for Peptide and Peptidome Research"** Speaker: *Dr. Sittiruk Roytrakul* National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA) • **"Epigenomics: Long Non-Coding RNA and Cellular Diversity: An Epigenomic Approach"** Speaker: *Dr. Patompon Wongtrakoongate* Faculty of Science, Mahidol University • **"Microbiomes: Embrace Our Second Genome"** Speaker: *Dr. Poochit Nonejuie* Institute of Molecular Biosciences, Mahidol University
	Moderator: *Dr. Nawapol Kunkeaw* Institute of Molecular Biosciences, Mahidol University
14.30 - 15.00pm	**Break**
15.00 - 16.00pm	**Symposium III: Applications in Systems Biology**
	• **" 'Omics in Biomedical Sciences"** Speaker: *Prof. Emeritus Suthat Fucharoen, M.D.* Institute of Molecular Biosciences, Mahidol University • **"Proteomics in Agricultural Sciences"** Speaker: *Assoc. Prof. Dr. Chartchai Krittanai* Institute of Molecular Biosciences, Mahidol University
	Moderator: *Assoc. Prof. Dr. M.L. Saovaros Svasti* Institute of Molecular Biosciences, Mahidol University
16.00 – 16.10pm	**Closing Ceremony**

Symposium I: Systems Biosciences and Medicine

- **"Towards Interactive and Reproducible QSAR Models for Studying the Origins of Bioactivity"**
  Speaker: *Assoc. Prof. Dr. Chanin Nantasenamat*
  Faculty of Medical Technology, Mahidol University

- **"In Depth, Multidisciplinary & Big Picture of Ligand Design: MANORAA.org"**
  Speaker: *Dr. Duangrudee Tanramluk*
  Institute of Molecular Biosciences, Mahidol University

- **"Translational Cancer Research in Thailand: a Focus on Colorectal Cancer"**
  Speaker: *Dr. Natini Jinawath, M.D.*
  Faculty of Medicine Ramathibodi Hospital, Mahidol University

Moderator: *Dr. Natini Jinawath, M.D.*
Faculty of Medicine Ramathibodi Hospital, Mahidol University

Towards Interactive and Reproducible QSAR Models for Studying the Origins of Bioactivity

Assoc. Prof. Dr. Chanin Nantasenamat
Center of Data Mining and Biomedical Informatics
Faculty of Medical Technology, Mahidol University
E-mail: chanin.nan@mahidol.ac.th

Quantitative structure-activity/property relationship (QSAR/QSPR) has been instrumental in unraveling the origins of the mechanism of action for biological activity of interest. QSAR is essentially a mathematical framework that relates the physicochemical description of chemical structures with their observed biological activity. Some of the questions that QSAR can answer includes: (i) what substructure/functional group are favorable/unfavorable for the biological activity, (ii) will the query compound of interest be able to inhibit the investigated protein/enzyme, (iii) aside from inhibiting protein X, what other protein will the query compound inhibit? (iv) how can the query compound be modified to improve its biological activity?, etc. The inherent heterogeneity in the format, quality and preparation of QSAR data sets poses a major barrier for reproducibility. Prior effort from the Blue Obelisk initiative had laid the important foundations for interoperable QSAR data sets via the use of a QSAR markup language (QSAR-ML) in which the meta data is incorporated with the data set. Although useful, but QSAR-ML considers only the pre-modeling phases primarily encompassing only the data set compilation and descriptor calculation while the modeling phases were not yet addressed. Here we describe the use of the Jupyter notebook (i.e. the successor of the iPython notebook) for storing codes (e.g. Python, R, etc.) that performs all of the procedures in a typical QSAR workflow encompassing the pre-processing, construction, validation and evaluation of the robustness of the QSAR model. Consequently, this naturally facilitates reproducible construction of QSAR models as the precise protocol, learning function, learning parameters and performance metrics are clearly documented in the Jupyter notebook. Such notebook can be used by other researchers to readily reproduce the output of the QSAR model while being intuitive to serve as a learning tool for students and newcomers of QSAR modeling. Herein, we propose a public database called iQSAR (an acronym for interactive QSAR) where users can contribute their QSAR models in the form of Jupyter notebooks. This repository can be accessed at http://www.iqsar.com.