



รายงานวิจัยฉบับสมบูรณ์

โครงการ : การศึกษาเชิงทฤษฎีของอัลบูมินในเซรัมของมนุษย์เพื่อเป็นโปรตีนขนส่งใน
พลาสมาซึ่งเกี่ยวข้องกับสมบัติทางเภสัชจลนศาสตร์

Theoretical Studies on Human Serum Albumin. A key Plasma Transport Protein with widespread
ADMET implications

โดย ดร. แมทธิว พอล กลีสัน

มิถุนายน 2557

รายงานวิจัยฉบับสมบูรณ์

โครงการ : การศึกษาเชิงทฤษฎีของอัลบูมินในเซรัมของมนุษย์เพื่อเป็นโปรตีนขนส่งใน
พลาสมาซึ่งเกี่ยวข้องกับสมบัติทางเภสัชจลนศาสตร์

Theoretical Studies on Human Serum Albumin. A key Plasma Transport Protein with widespread
ADMET implications

ผู้วิจัย

ดร. แมทธิว พอล กลีสัน

สังกัด

ภาควิชาเคมี คณะวิทยาศาสตร์

มหาวิทยาลัยเกษตรศาสตร์

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย
และ มหาวิทยาลัยเกษตรศาสตร์

(ความเห็นในรายงานผลการวิจัยเป็นของผู้วิจัย สกว. ไม่จำเป็นต้องเห็นด้วยเสมอไป)

กิตติกรรมประกาศ
(Acknowledgements)

I would like to grateful acknowledge the support from the Thailand Research Fund who enabled this research to be undertaken. I would also like to acknowledge the support of Kasetsart University and my collaborators both within and outside Thailand.

Abstract

Project code : RSA5480016
Project Title : Theoretical Studies on Human Serum Albumin. A key Plasma Transport Protein with widespread ADMET implications
Investigator : Dr. Matthew Paul Gleeson Kasetsart University
E-mail Address : paul.gleeson@ku.ac.th
Project period : June 2011 –June 2014

The goal of this research project is to develop predictive models for studying absorption, distribution, metabolism excretion and toxicity (ADMET) phenomenon. In the first phase of the project we have reviewed in great detail the available methodologies, publishing a number of reviews in the area. Subsequently, we have developed quantum mechanical (QM) based methods for understanding ADMET phenomenon.

Initial studies were focused on validating the methodology using the available kinase X-ray crystal structures. The studies were extended to models of protein reactivity, as given by the local lymph node assay (LLNA). Our calculations shows that the sensitivity of aromatic halides were directly proportional to their reactivity as predicted by high level QM calculations.

Simulations on HSA, a protein important for drug distribution, show that differences in fatty acid binding can have a dramatic effect on the flexibility of the protein and also the pocket characteristics. We discuss how the remarkable selectivity of the HSA pockets, towards both endogenous fatty acids and exogenous drug molecules, is highly complex and is not simply driven by the number of H-bonds, or VDW contacts or even solvent accessibility. Further QM/MM calculations on HSA suggest that Lys199, His242, Arg257 give rise to the experimentally observed esterase activity and that the most catalytically efficient active site configuration requires that both Lys199 and Aspirin are in their neutral forms. The abundance of HSA in the body suggest the protein might be a suitable target for the computational guided design of acetyl based pro-drugs of acidic molecule that often displayed limited oral exposure due to their unmasked ionizable substituent.

Keywords: Human serum albumin (HSA), skin sensitization, Simulations, MD, QM/MM

บทคัดย่อ

รหัสโครงการ : RSA5480016
ชื่อโครงการ : การศึกษาเชิงทฤษฎีของอัลบูมินในเซรัมของมนุษย์เพื่อเป็นโปรตีนขนส่งในพลาสมา
ซึ่งเกี่ยวข้องกับสมบัติทางเภสัชจลนศาสตร์
ชื่อนักวิจัย : ดร. แมทธิว พอล กลีสัน มหาวิทยาลัยเกษตรศาสตร์
E-mail Address : paul.gleeson@ku.ac.th
ระยะเวลาโครงการ : มิถุนายน 2554 – มิถุนายน 2557

วัตถุประสงค์ของโครงการวิจัยนี้ เพื่อพัฒนาแบบจำลองการทำนายสำหรับการศึกษาปรากฏการณ์การดูดซึม การกระจาย การเผาผลาญอาหาร การขับถ่าย และความเป็นพิษ หรือที่เรียกว่า ADMET โดยในระยะแรกของโครงการ ได้ศึกษารายละเอียดเกี่ยวกับระเบียบวิธีและตีพิมพ์ผลงานในสาขาที่เกี่ยวข้อง ซึ่งต่อมาได้พัฒนาพื้นฐานวิธีกลศาสตร์ควอนตัม (QM) เพื่อให้เข้าใจปรากฏการณ์ ADMET

การศึกษาในเบื้องต้น จะเน้นถึงการตรวจสอบวิธีการ โดยใช้โครงสร้างผลึกรังสีเอ็กซ์ (X-ray) ของโคเนสที่มีอยู่ นอกจากนี้ยังได้จำลองแบบความว่องไวในการทำปฏิกิริยาของโปรตีน ตามวิธี LLNA การคำนวณแสดงให้เห็นว่าความว่องไวของหมู่ aromatic halides แปรผันโดยตรงกับความว่องไวที่ทำนายโดยการคำนวณด้วยกลศาสตร์ควอนตัมในระดับสูง

การจำลองบน HSA ซึ่งเป็นโปรตีนที่สำคัญสำหรับการกระจายของยา แสดงให้เห็นว่าความแตกต่างของการยึดของกรดไขมัน ส่งผลกระทบอย่างมากต่อความยืดหยุ่นของโปรตีนและลักษณะเฉพาะของตำแหน่งที่เกิดปฏิกิริยา (HAS pockets) ได้อธิบายถึงการเลือกที่สำคัญของ HSA pockets โดยผ่านกรดไขมันที่เกิดขึ้นภายในสิ่งมีชีวิตและที่เกิดขึ้นภายนอกโมเลกุลของยา ซึ่งมีความซับซ้อนมากและไม่ถูกจับได้ง่าย ๆ เพียงแค่จำนวนพันธะไฮโดรเจน หรือติดต่อกับวันเดอร์วาลส์ หรือแม้แต่การเข้าถึงของตัวทำละลาย การคำนวณ QM/MM ขั้นต่อไปบน HSA ได้ชี้ให้เห็นว่า Lys199, His242 Arg257 ก่อให้เกิดความว่องไวของ esterase ที่สังเกตได้จากการทดลอง และมีประสิทธิภาพการเป็นตัวเร่งของการจัดเรียงตัวของบริเวณที่ว่องไวมากที่สุด จำเป็นต้องใช้ Lys199 และ Aspirin ในรูปที่เป็นกลาง จากปริมาณของ HSA ในร่างกายจำนวนมาก ชี้ให้เห็นว่าโปรตีนอาจจะเป็นเป้าหมายที่เหมาะสมสำหรับเป็นแนวทางการออกแบบทางการคำนวณของยาที่มี acetyl เป็นพื้นฐานของโมเลกุลที่เป็นกรดซึ่งมักจะปรากฏข้อจำกัดที่จะเกิดขึ้นในช่องปาก เนื่องจากแสดงหมู่แทนที่ที่แลกเปลี่ยนไฮออนได้

Keywords: Human serum albumin (HSA), skin sensitization, Simulations, MD, QM/MM

เนื้อหาทางวิจัย

1. บทนำ (Introduction)

The research plans proposed over the course of the 3 year project have been undertaken in full. Some aspects of the project have proved more difficult than others; however all of the topics proposed have been undertaken and have resulted in publications. In particular, due to the very large size and flexibility of HSA, QM and MD calculations proved very challenging, requiring additional unexpected validation work to ensure the calculations could be completed in a timeframe consistent with the project plan. These final issues have been resolved through careful validation using smaller model systems and the final binding and reactivity studies of HSA have been submitted for publication.

Year 1

- Validate proposed methods via extensive literature review.
 - Completed, 2 publications obtained.
- Validate QM models to assess subtle binding interactions
 - Completed, 1 publication obtained.

Year 2

- Validate the MD methodologies for use with HSA.
 - Completed, 1 publication obtained.
- Validate the QM methods to study excited state reaction intermediates.
 - Completed, 2 publications obtained.

Year 3

- Use MD to assess HSA ligand binding and origin of affinity.
 - Completed, 1 publication submitted for publication.
- Use QM/MM to study the catalytic reactions of HSA.
 - Completed, 1 publication submitted for publication.

The goal of this research project is to develop predictive models for studying absorption, distribution, metabolism excretion and toxicity (ADMET) phenomenon. In the first phase of the project we have reviewed in great detail the available methodologies, publishing a number of reviews in the area. Subsequently, we have developed quantum mechanical (QM) based methods for understanding ADMET phenomenon. Initial studies were focused on validating the methodology using the available kinase X-ray crystal structures. The studies were extended to models of protein reactivity (as given by the local lymph node assay – LLNA) to validate the methods utility in explaining experimental biological data.

The final phase of the project has focused on studying binding and reactivity phenomena associated with human serum albumin. As described in the proposal, the initial focus of this study has been directed towards understanding aspects of ligand binding using the available X-ray crystal structures from the RCSB including: (a) the differing affinity of the two key drug binding pockets (b) the differing affinity of the seven key fatty acid binding pockets (c) the impact of concomitant fatty acid or ligand binding to the HSA-drug complex and (d) the newly suggested esterase mechanism. The reason for this is to better understand the methods needed to study the HSA protein-ligand complex.

The original proposal specified between 3-5 papers as output, being depending on the scientific findings. To date, the project has resulted in 6 international publications. Two additional publications are expected, 1 of which has been submitted for publication and another which is due to be submitted in the coming weeks. Thus it is expected that the project will result in 8 scientific papers in total.

2. วิธีการทดลอง (Method)

Both the AMBER and Gromacs 4.0 programs were assessed. Due to the dramatic speed advantage of the latter found from test calculations, it was used for all the MD simulations. However, the AMBER 99SB force field was used instead of the inbuilt GROMOS forcefield as the former was found to be more reliable for simulating drug-like molecules.

To try and better understand what MD simulations conditions are best for proteins such as HSA, simulations were undertaken on a smaller, simpler protein kinase system to begin with (the same as used in the test QM system calculations). These results have been written up and submitted for publication. Following these simulations, studies on HSA were undertaken. Simulations of HSA with different sites occupied with fatty acid were undertaken to better understand the dynamics characteristics of the protein and how fatty acid binding affects this, as well as ligand binding.

Quantum chemical calculations have been performed as described in the original proposal. Cluster models of protein ligand complexes based on kinase crystal structures have been used as the models systems to study binding interactions. All calculations have used the Gaussian 09 program.

The effect of using a number of different theoretical methods to optimize QM active site models were investigated. MP2/6-31+G**, M062X/6-31G*, HF/6-31G*, HF/3-21G, AM1 and CHARMM methods have been employed, and considered the effect of BSSE and the inclusion of an implicit solvation model. The M062X method showed the best balance between speed and reliability so was taken forward for further validation. The method was shown to also be very suitable for studying reaction

mechanisms, predicting excited state species with good accuracy compared to higher computational cost methods. As such the M062X/6-31G* method was used for QM/MM simulations of HSA, with singles point being performed with the 6-31+G** basis set.

Hybrid QM/MM calculations to explore HSA reactivity have been undertaken using a similar methodology as used in the earlier gasphase calculations (M062X). The reaction profile of the protein with active site residues in 3 distinct protonation states were investigated. Minima and transition states along the reaction coordinates were obtained and confirmed from vibrational frequency analysis.

3. ผลการทดลองและวิจารณ์ผลการทดลอง (Results and Discussions)

PART 1: Probing the Binding Site Characteristics of HSA: A Combined Molecular Dynamics and Cheminformatics Investigation.

ABSTRACT

Human serum albumin is a remarkable protein found in high concentrations in the body. It contains at least 7 distinct fatty acid binding sites and 2 principle sites for drugs. Its primary function is to act as a fatty acid transport system, but it also shows the capacity to bind a diverse range of acidic, neutral and zwitterionic drug molecules. In this paper we investigate the ligand binding selectivity of HSA using cheminformatics analyses and molecular dynamics simulations. We compare and contrast the known ligand binding specificities as obtained from X-ray structural data using PCA, with additional direct analyses of the 7 key binding pockets using analyses derived from molecular simulations. We assess both the fatted and defatted states of HSA using 100 ns simulations of the APO and HOLO forms as well as structures containing 1, 3 and 7 myristic acid molecules. We find that differences in fatty acid binding can have a dramatic effect on the flexibility of the protein and also the pocket characteristics. We discuss how the remarkable selectivity of the HSA pockets, towards both endogenous fatty acids and exogenous drug molecules, is highly complex and is not simply driven by the number of H-bonds, or VDW contacts or even solvent accessibility.

1. Introduction

The binding of molecules to plasma proteins is important from a pharmaceutical perspective since this property can be used to estimate the total body clearance in man.¹ There are two main proteins implicated in the binding of drug molecules in plasma, namely human serum albumin (HSA) and α 1-acid glycoprotein (AGP). The former is the most abundant protein found in plasma, making up

approximately 60% of the total protein present. HSA is an important transport protein that binds numerous molecules of endogenous and exogenous molecules. Fatty acid (FA) binding to HSA has been long studied due to the critical role these molecules play in energy metabolism and membrane synthesis.² HSA is well known for its capacity to bind drug molecules and understanding the structure and function of the protein is thus essential for future drug development. As a result, the binding characteristics and specificity of HSA have received much attention both experimentally and computationally. To date, approximately 100 HSA structures have been determined by X-ray.^{3, 4} In addition, NMR⁵⁻⁷, EPR⁸ and biochemical assays have been widely used to explore aspects of ligand binding to the protein.^{9, 10}

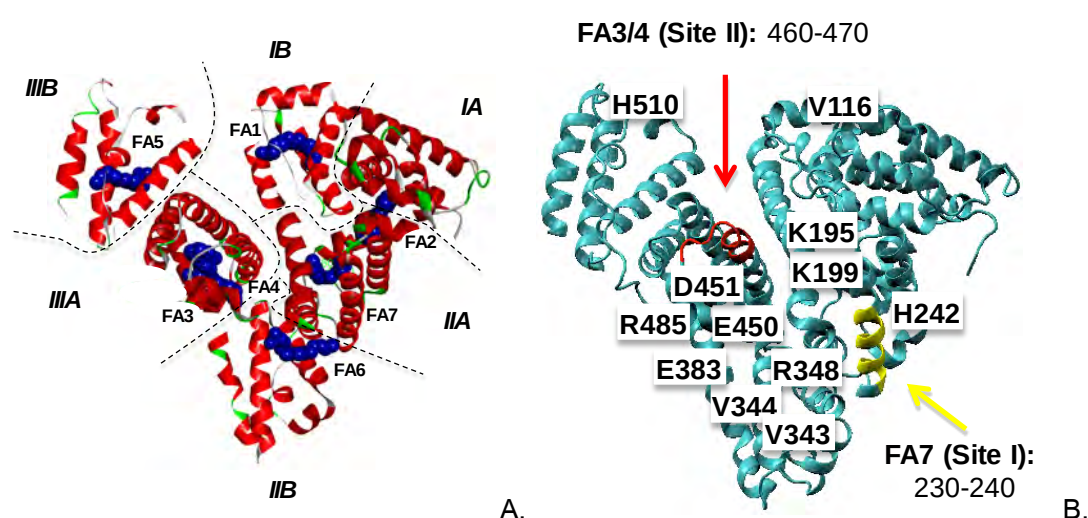


Figure 1 Cartoon views of the key regions and binding sites (E) and important residues in HSA (F)

HSA has a molecular mass of 66 kDa (585 amino acids). It exists as a monomer with three similar α -helical domains (I–III) (Figure 1). Each domain can be divided further into subdomains A (4 α -helices) and B (6 α -helices).^{3, 4, 11} HSA contains numerous sites for long chain fatty acids, distributed across the protein, and 2 primary sites for drugs known as drug site I and II.^{6, 12} Simard *et al.* observed that such 7 FA sites exist for FAs with chain lengths >14 carbons.^{5, 6} Kragh-Hansen *et al.* noted that Lauric acid (12 carbons) binds to 8 unique sites and decanoic acid (10 carbons) can bind to 10.¹³ Subsequent studies revealed that three of the sites displayed higher affinities for FA (sites FA2, FA4 and FA5, where FA5 is the highest).^{5, 6} Even so, even the lower affinity site FA1 can bind very large acidic molecules such as HEME with high affinity.¹⁴ The presence of FA bound to HSA is also known to induce conformational change in the protein and this affects drug binding.¹⁵ Thus,

understanding the origin of the FA-binding differences among these 7 sites remains a source of major interest.^{6, 14-17}

QSAR approaches have been extensively used to rationalise both HSA and whole plasma protein binding.^{10, 18, 19} However, the specific interactions and pocket shapes give rise to ligand selectivity¹⁵ which bulk physico-chemical property QSAR models cannot easily account for. However, incorporating specific details of the binding interactions at a molecular level may help improve this situation in the future.²⁰⁻²² Furthermore, there are an increasing number of computational studies on HSA in which molecular docking or molecular dynamics (MD) simulations have been used to gain insight into HSA-ligand binding. For example, Fujikara *et al.* employed MD simulations and principle component analysis (PCA) to evaluate the possible conformational changes of HSA in the presence of bound myristic acids.²³ MD simulations have also been used to postulate drug-binding sites or understand the drug-binding mechanisms at specific HSA pockets (i.e. warfarin,²⁴ hydroxyquinoline derivatives,²⁵ betulinic acid²⁶). More recently Castellanos *et al.* have investigated the role of disulphide bonds on the structural flexibility of the protein.²⁷ Nevertheless, previous studies have focused their attention on the macroscopic conformational changes of HSA in the presence or absence of natural or synthetic ligands are bound. The microscopic characteristics of the individual binding pocket remain largely unstudied computationally. For an extensive review of the molecular modelling studies conducted on HSA readers are referred to reference.²⁸

In this study we aim to shed further light on the structural characteristics and binding pocket selectivities of HSA. We have extracted and analysed known drug molecules bound to each of the 7 principle binding pockets of HSA. PCA was then used to understand what physical properties are most important in defining the selectivity for each site. We then contrast this result with those derived from 100 ns MD simulations of HSA, performed using a number of different states, to assess how the 7 different pockets evolve over time and under different conditions. Simulations have been performed on the closed form of HSA (APO model), the open form with 7 myristic acid (MYR) molecules bound (FA(1-7) model). We have also assessed intermediate states by removing a portion of FA molecules from the open protein structure, with only those occupying (a) the 3 highest affinity sites (FA(2,4,5) model), (b) the single highest affinity site (FA(5) model) (c) zero FA molecules bound (HOLO model).

2.0 Computational Procedures

2.1 MD Simulations

HSA models for the closed and open form or HSA were created from the coordinates 1e78 and 1e7g, respectively,²⁹ which were sourced from the RCSB databank.³⁰ The closed structure, corresponding to the non-liganded form, was used to generate the APO HSA model. The open

structure, containing bound MYR molecules, was used to create a HSA model containing ligands at the 7 principle FA sites (FA(1-7) model). The latter X-ray structure was also used to create an additional model of HSA in the open form but containing no fatty acids (HOLO model). Additionally, two other HSA models were created which contained either 1 or 3 fatty acids at the most high affinity sites. These are termed the FA(2) and FA(2,4,5) models. The N- and C-terminal ends of each model were capped with acetyl (ACE) and methyl amino (NME) groups, respectively, and the overall quality confirmed using PROCHECK v3.5.4.¹ The protonation state of ionisable residues were determined using PROPKa³¹ and a visual analysis of the environment surrounding each ionisable residue. Residues were treated as HID unless otherwise stated: HIE (39, 288, 40) and HIP (3, 67, 105, 128, 146, 247, 338, 367, 510). Each model was then placed in a cubic simulation box containing water and neutralized with counter ions by randomly replacing water molecules.

Simulations were carried out using GROMACS v4³² with the AMBER99SB forcefield³³ Ligand topology files were generated using ACPYPE script⁶ and the GAFF force field³⁴. Electrostatic charges were generated using a HF/6-31G(d) optimized structures (Gaussian 09 Revision C01)³⁵ and the RESP protocol of Ambertool1.5³⁶ Long range electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method with a 0.12 nm cut-off³⁷ while van der Waals interactions utilized a 1 nm cut-off. Simulations were performed at constant temperature, pressure and number of particles (NPT). The temperature of the protein and solvent were each coupled separately. The Berendsen thermostat³⁸ was applied at 300 K with a coupling constant of 0.1 ps.³⁹ Coordinates and velocities were saved every 2 ps. The LINCS algorithm was used to restrain bond lengths and a 2 fs timestep was used for integration.⁴⁰ Five ns of protein backbone-restrained dynamics were employed initially, followed by 100 ns production runs. Simulations for the APO, HOLO and FA(1-7) models were performed in duplicate to assess the sensitivity of the results to different starting conditions. Analyses on the resulting MD coordinates were performed using GROMACS routines, Discovery Studio,⁴¹ and locally written scripts. Molecular graphics images were produced using VMD.⁴² The statistical significance of any reported differences have been confirmed using paired Student's T-test or F-test, respectively at >95% confidence level.

2.2 Cheminformatics Analyses

The physico-chemical determinants of selectivity for the 7 HSA binding pockets were assessed by curating a list of ligands known to bind in each pocket. This was done based on an analysis of structures deposited in the Protein databank structures. A range of commonly used physico-chemical descriptors were calculated for the list of 57 inhibitors identified as binding to at least one of the known binding sites. All descriptors were calculated using the Chemaxon JChem suite⁴³.

A PCA model was also generated using data reported in Table 1 to Table 3 to (a) investigate the structural differences between the difference proteins simulations and (b) investigate the differences in pockets of HSA and the effect of different simulation conditions.

All models were generated in SIMCA-P10⁴⁴ using the default settings (i.e., mean centred and scaled descriptors, with the auto-fitting of components) and report in the form of either the traditional loading and scores plot or a combined loadings bi-plot (i.e. both the scores and loadings projected onto the same plot).

Table 1 Comparison of key parameters obtained for each of the 5 simulations. All values reported are average values obtained over the course of the 100nS simulations. Standard deviation given in parenthesis. * results derived from a single simulations.

Model	RMSD	RMSF	H-bonds	Ryd. Gyr.	Area
APO	0.38 (0.07)	0.18 (0.12)	456 (11.4)	2.72 (0.03)	0.59 (0.39)
HOLO	0.31 (0.07)	0.17 (0.07)	459 (11.4)	2.76 (0.03)	0.60 (0.38)
FA (1-7)	0.32 (0.08)	0.15 (0.06)	461 (11.3)	2.77 (0.02)	0.61 (0.37)
FA (2,4,5)*	0.24 (0.04)	0.16 (0.06)	461 (11.9)	2.77 (0.02)	0.60 (0.38)
FA (5)*	0.27 (0.04)	0.19 (0.08)	462 (10.4)	2.78 (0.02)	0.60 (0.38)

Table 2 Center of Mass (COM) distances between key residues within HSA. Distances are reported in nm. * results derived from a single simulations.

Descriptor	APO	HOLO	FA(1-7)	FA (2,4,5)*	FA (5)*
H510-V116	2.12 (0.67)	2.09 (0.39)	2.63 (0.27)	2.40 (0.29)	2.28 (0.24)
E383-R485	0.74 (0.06)	0.85 (0.15)	0.78 (0.08)	0.89 (0.11)	0.83 (0.09)
K195-D451	0.81 (0.08)	0.69 (0.05)	0.69 (0.05)	0.82 (0.14)	0.84 (0.12)
R348-E450	0.91 (0.17)	1.24 (0.12)	1.18 (0.04)	0.89 (0.05)	0.87 (0.03)
V343-E450	0.89 (0.07)	0.76 (0.12)	0.66 (0.04)	0.79 (0.04)	0.83 (0.06)
V344-E450	0.62 (0.06)	0.69 (0.10)	0.67 (0.05)	0.58 (0.02)	0.59 (0.04)
Y138-Y161	0.37 (0.02)	0.51 (0.03)	0.45 (0.10)	0.47 (0.07)	0.90 (0.06)

Table 3 The mean number of H2O contacts, H-bonds, Total contacts observed for FA bound in the FA(1-7) simulations. Also reported are the mean pocket sizes for each simulation.

Simulation	Descriptor	FA Site.1	FA Site.2	FA Site.3	FA Site.4	FA Site.5	FA Site.6	FA Site.7
FA(1-7)	H2O-contacts	5.58 (0.75)	4.10 (0.93)	2.58 (0.75)	6.14 (0.99)	5.84 (0.66)	7.75 (0.97)	8.43 (0.83)
FA(1-7)	H-bonds	3.45 (0.38)	3.35 (0.29)	3.24 (0.11)	2.29 (0.53)	1.75 (0.20)	1.83 (0.31)	1.83 (0.32)
FA(1-7)	Total-contacts	182.1 (14.2)	203.3 (16.5)	212.9 (14.2)	185.3 (17.1)	214.4 (14.6)	184.4 (14.7)	169.7 (15.2)
FA(1-7)	FA length	0.88 (0.18)	1.4 (0.09)	1.24 (0.05)	1.3 (0.07)	1.34 (0.03)	1.16 (0.10)	0.78 (0.12)
FA(1-7)	Volume	5.74 (0.75)	5.89 (0.74)	5.27 (0.53)	5.28 (0.56)	4.76 (0.61)	4.44 (0.45)	5.01 (0.58)
HOLO	Volume	5.35 (0.71)	5.94 (0.77)	5.42 (0.53)	5.29 (0.52)	4.73 (0.58)	4.53 (0.46)	4.91 (0.60)
APO	Volume	5.58 (0.70)	5.99 (0.76)	5.09 (0.52)	5.04 (0.48)	4.78 (0.58)	4.36 (0.45)	4.93 (0.60)
FA(5)	Volume	5.02 (0.19)	5.97 (0.23)	5.14 (0.21)	5.10 (0.17)	4.88 (0.16)	4.60 (0.13)	4.96 (0.19)
FA(2,4,5)	Volume	4.76 (0.62)	5.52 (0.69)	5.00 (0.47)	5.32 (0.51)	4.88 (0.58)	4.59 (0.44)	5.01 (0.50)

3. Results and Discussion

We begin this section with a discussion of the structural features of the HSA. In the subsequent section we discuss the use of FA molecules to assess the dynamic characteristics of the 7 pockets of HSA over the course of the MD simulations. This of course assumes that FA molecules are suitable probes for the range of substrates that FAs bind. As such, in the final section we employ a complementary, ligand-based approach to further probe the binding site specificity of the protein. We compare and contrast the calculated physico-chemical descriptors of the known substrates of each pocket, to the binding site parameters obtained from the MD simulations. In this way we can gain insight into the unique characteristics of the pockets and assess their dynamic character to a level of detail yet reported.

3.1 HSA Structural Characteristics

The overall size, shape, and dynamics of the HSA protein are thought to influence the behaviour of its binding pockets, which in turn affect the binding affinity of fatty acids and other drug molecules⁴⁵. Based on a series of simulations performed in this study, the global dynamic properties of HSA under each condition were investigated and compared in order to elucidate the key dynamic features that are essential for structure and function. Moreover, we pay particular attention to the microscopic details of the binding pockets that contribute to the binding of fatty acid.

The constant total energy of each simulation indicates the convergence of each run after 5-ns pre-equilibration. Both C_{α} RMSDs and RMSFs of the proteins show the significant flexibility of HSA, with the greatest fluctuation being found at the interface between the different sub-domains (Figure 2). The RMSDs of the duplicate simulations performed on the APO and HOLO structures show slight differences in the backbone flexibility, this being related to when the inter-domain movement begins. Unlike the APO and HOLO models, the presence of 7 bound FA molecules is found to reduce the magnitudes of both the RMSDs and RMSFs significantly, which indicates a more rigid protein. We observe that the key structural features are preserved well although certain individual residues or loop movements can differ between duplicates.⁴ We see a subtle difference between the two FA(1-7) simulations due to a change in the interaction pattern of R117. R117 sidechain is so flexible that it can form a salt-bridge with either E119 located in domain I or E520 on domain III.

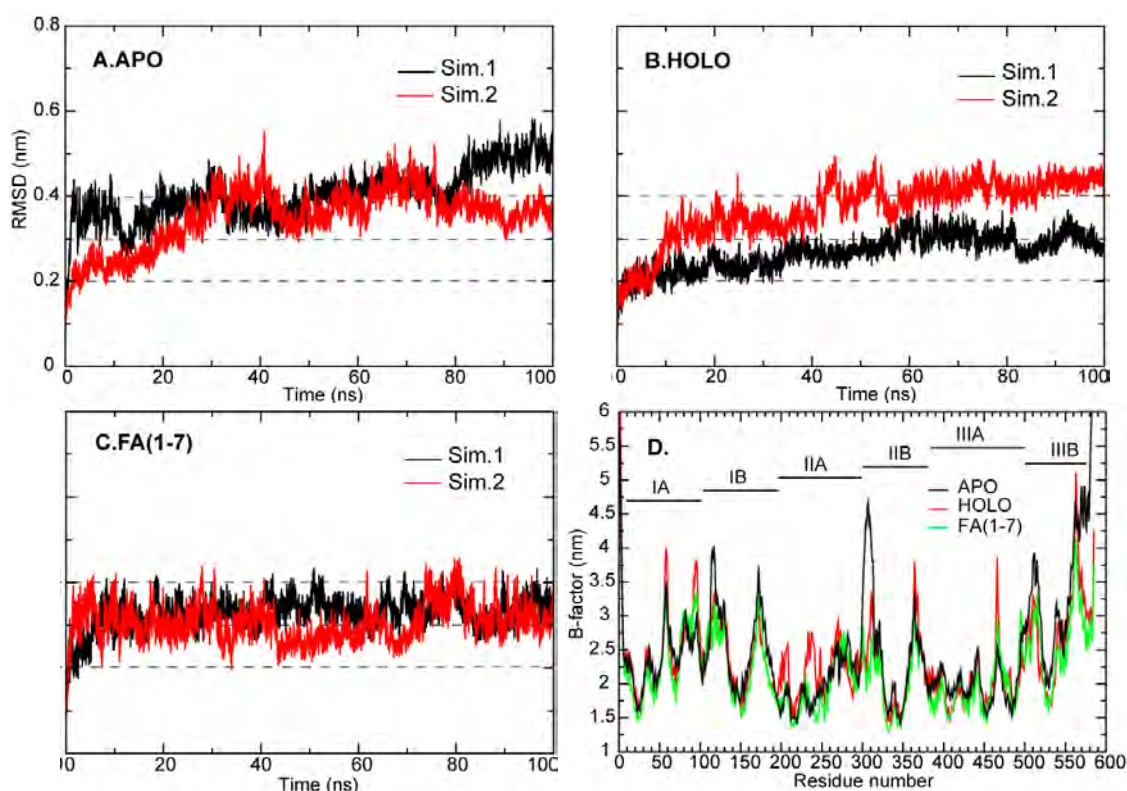


Figure 2 (A-C) C α RMSDs observed in the APO (open conformation with no FA), HOLO (closed conformation with FAs removed) and FA(1-7) (closed conformation containing 7 FAs) models. (D) B-factor for APO, HOLO, and FA(1-7). (E-F) show cartoon views of the key regions and binding sites (E) and important residues in HSA (F).

Simulations with either 1 or 3 FA molecules at the highest binding sites (FA(5) and FA(2,4,5) models) are more akin to the physiologically FA binding conditions experienced in the body.^{41,42} Experimental X-ray crystallographic studies have shown that HSA in this state is found to adopt a structure intermediate between the open and closed forms of the protein.²⁹ It appears that these configurations can also enhance the protein rigidity (Table 1). This higher rigidity is observed over the closed protein due to the additional ~3-6 hydrogen bonds found in the presence of the FA in the open conformation. Taken overall, the similarities between the number of hydrogen bonds among amino acids, the radius of gyration (indicating the compactness of protein structure), and the residue surface area, suggest the partially and fully FA-bound HSA show similar global properties (Table 1). Furthermore, it also appears that the HOLO structure does not transition to the closed conformation, as might be expected. The 100ns timescale appears to be insufficiently long to observe the complete conformational change from FA-bound to native HSA structure.

Nevertheless, a number of interactions have been identified as key indicators of the conformational state of the protein in the fatted and de-fatted states.^{5, 6, 15} The H510-V116

distance is such a parameter, being indicative of the distance between interface of domain IIIB and IB (H510 on domain IIIB and V116 on domain IB, Figure 1B). A change in the distance between H510-V116 can indicate the movement of both domains, which is an indicator for the open and closed states. When FA binds, the H510-V116 distance increases from 2.12nm (APO) to 2.63nm for FA(1-7), 2.4nm for FA(2,4,5), and 2.28nm for FA(5) (Table 2), respectively. Binding of FA molecules clearly induces the movement of domain IIIB and IB, which agree well with previous studies^{46, 47} and shows that the partially occupied structures are in a more intermediate state as expected experimentally. This direction of movement can be captured by PCA analysis. The dominant motion obtained from the PCA model (i.e component 1) demonstrates a scissor-like motion of domain IIB and IB in both the absence and presence of bound FAs. This is in a good agreement with previous studies.^{47, 48} Compared to domain IB and IIIB, the other domains seem to be rather rigid. Furthermore, Ghuman *et al.*¹⁵ noted that FA binding at site FA3 in domain IIIA led to the disruption of a salt-bridge between R348-E450, resulting in a stronger H-bond interaction between E450 with the backbone of V343/V344 (Figure 1). The relocation of E450 sidechain also force the displacement of D451, resulting in an interaction with K195. From our simulations, we clearly observed an increase in R348-E450 distance in the HOLO and FA(1-7) (~1.2nm) models, while the two FA3-free simulations (FA(5) and FA(2,4,5)) display shorter distances of ~0.9nm (Table 2). The disruption of R348-E450 when the FA3 site is occupied allows E450 to interact with V343, leading to closer contacts between the two (~0.6nm), and an improved interaction between K195-D451 (reducing from ~0.8nm in FA3-free models to ~0.69nm in FA(1-7) (Table 2).

3.2 HSA Binding Site Characteristics from MD

The binding of ligands to the Sudlow sites I (FA 3&4) and II (FA7) have been extensively reported in the literature from a structural, biochemical and computational perspective.^{12, 15, 16} Of course, other binding sites play a role in ligand binding, although often secondary, and are therefore less well studied. For some sites the intricacies of the interactions have not been elaborated on, or for others the primary amino acids that bind substrates have not been confirmed.⁵ In this section we aim to describe the different binding pockets of HSA in terms of the size, shape and interactions, and how they evolve over the course of the simulations. We also discuss the effect of FA binding on the characteristics of these sites and the implications this will have on more rigorous calculations to estimate binding affinities.

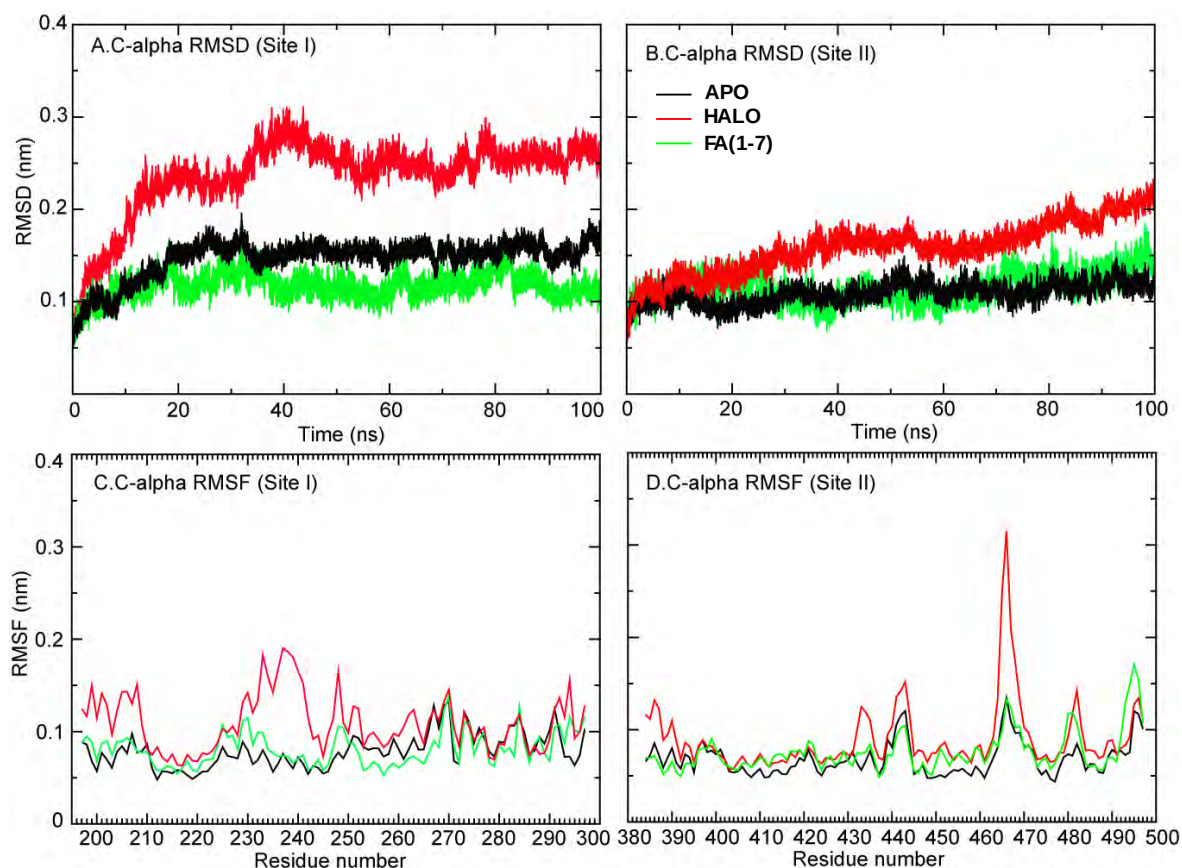


Figure 3 The average RMSD and RMSF values associated with Sudlow sites 1 and 2 sites for the APO, HOLO and FA(1-7) simulations of HSA.

The C_{α} RMSDs of residues surrounding both binding pockets for the APO, HALO and FA(1-7) have been analysed in Figure 3. RMSDs indicate that the Sudlow site II is much less sensitive to the conformation state of the protein. The RMSD of surrounding site I are broadly comparable for the APO and FA(1-7) structures. However, removing the FA from the latter (i.e. the HALO structure) leads to a dramatically larger RMSD and RMSF. In contrast, site II shows a smaller effect, with a smaller deviation between the different HSA configurations. This suggests site II is less sensitive to dynamic effects (i.e. the open, closed or intermediate conformation).

All of the 7 known fatty acids sites can bind drug-like molecules, albeit most with much lower frequency than the Sudlow sites. It is of interest to us to explore the differences between these sites in greater detail and how they are affected by random dynamic fluctuation. To better characterise each pocket, the distances between each FA and its neighbouring key residues, hydrogen bonds, water and total contacts were extracted and reported in Table 3 as well as graphically in Figure 4, Figure 5.

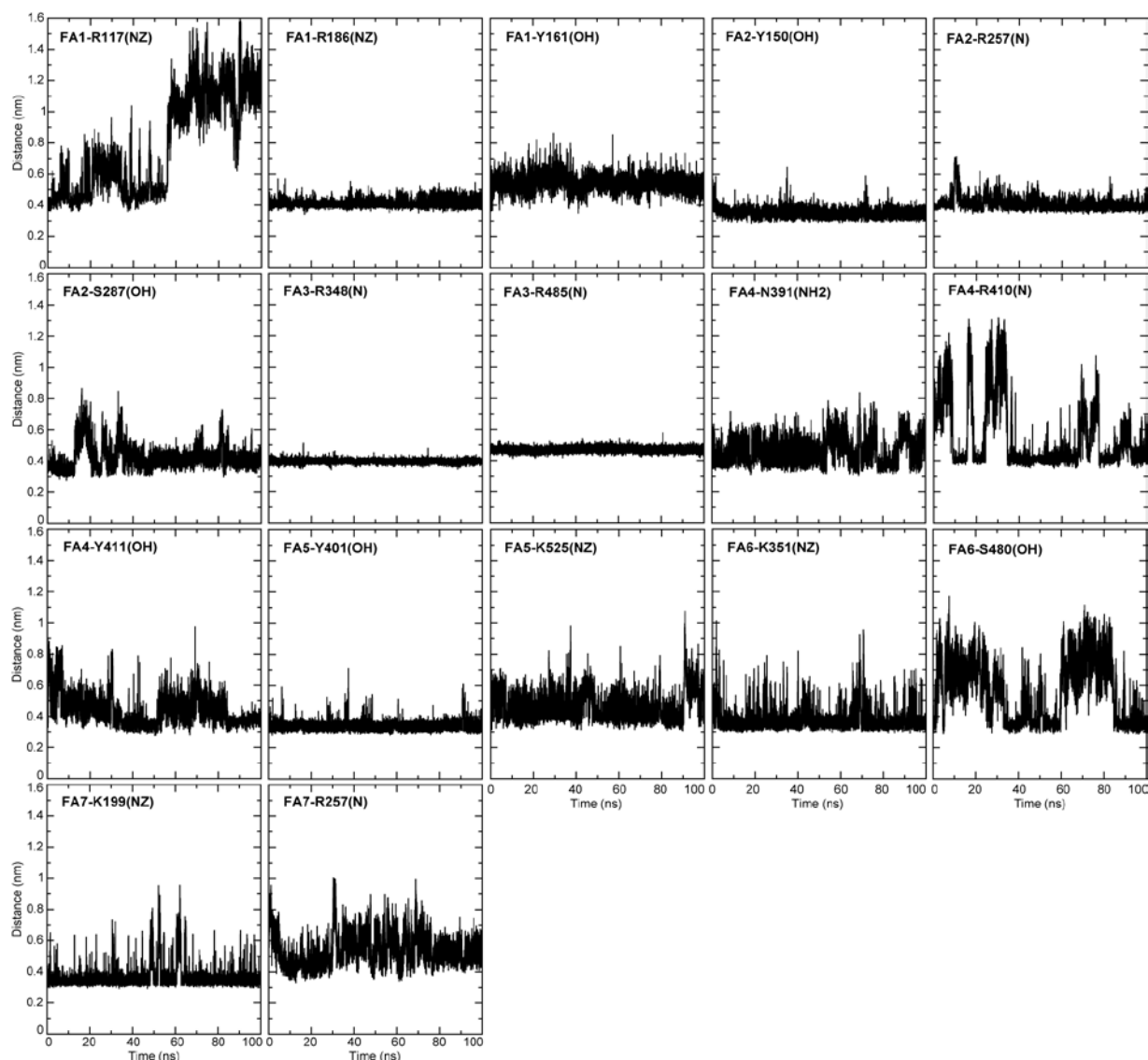


Figure 4 Distances between the carboxylic carbon of each FA in FA(1-7) and the center atom in functional group on a sidechain of key neighbours in an individual pocket (NZ is an amine nitrogen, N is a quaternary nitrogen, and OH is a hydroxyl oxygen).

The FA1 pocket, located in subdomain IB, is a curved cavity of large size with moderately solvent accessibility. Within the cavity, the FA molecules are on average found to maintain ~3 non-covalent bonds. Distances of ~0.4 nm between FA1 and R117 (before 70ns), R186, and Y161 indicate the presence of 2 salt bridges and 1 hydrogen bond, where R186 and Y161 interact with FA1 consistently (Figure 5 and Table 3). The fluctuating distance after 70ns of R117 sidechain indicates the high flexibility which enables interactions with residues in subdomain IB as mentioned above. FA2 is also a large linear pocket, located between subdomains IA and IIA. It is

the 2nd most enclosed of the HSA sites and maintains 1 salt bridge (R257) and 2 H-bonds with Y150 and S287 over the course of the FA(1-7) 100 ns simulation (Figure 5). FA3 is a moderately sized pocket located in subdomain IIIA. The lowest water contacts of ~2.58 demonstrates the most buried site for FA3. However, FA3 can make 2 salt bridges with R348 and R485.

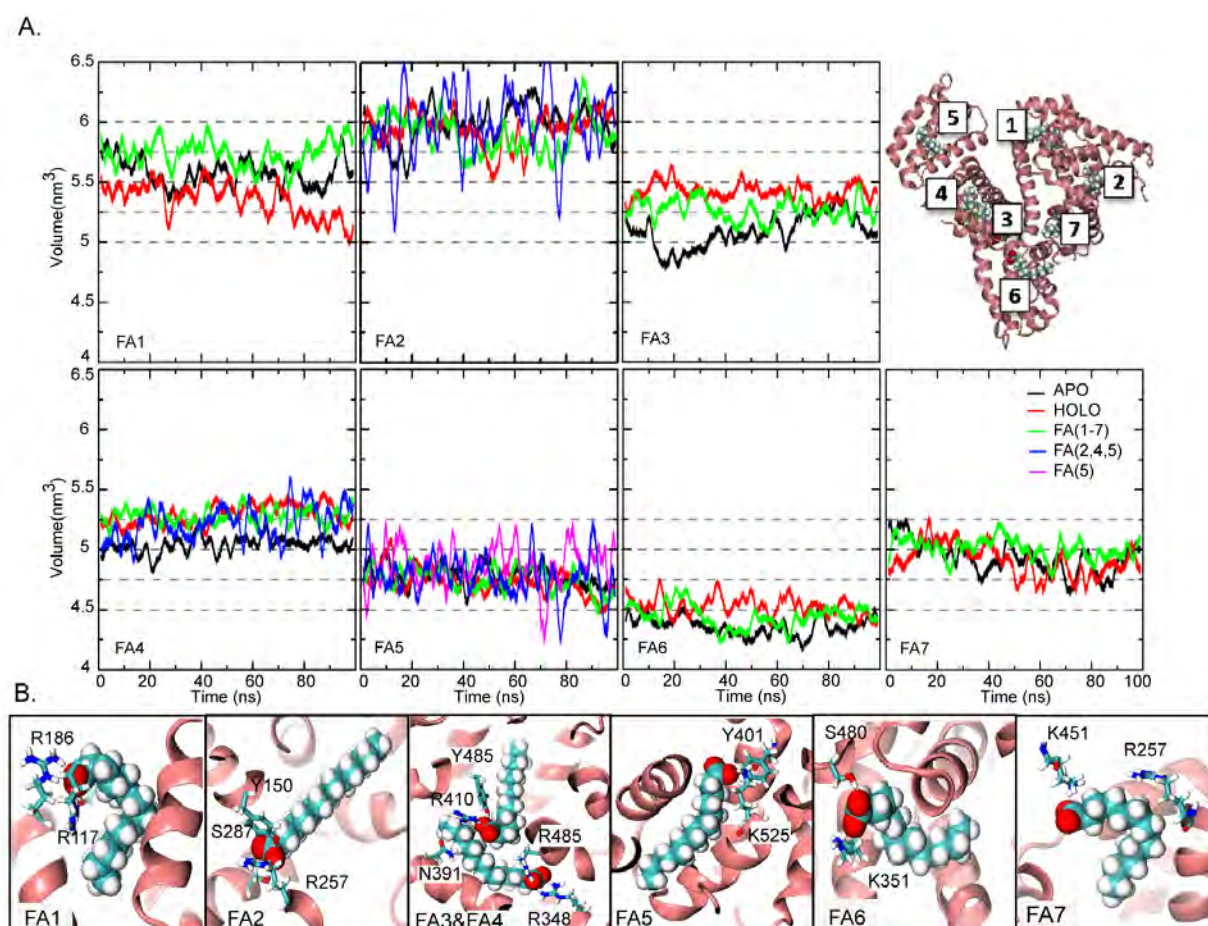


Figure 5 (A) Computed volume of fatty acid binding sites 1 to 7 for APO, HOLO, FA(1-7), FA(2,4,5), and FA(5) models. (B) Final snapshot of FA alignment in each pocket with key residues for FA(1-7).

FA4 sits in subdomain IIIA, whose site is a moderately sized linear pocket. This site makes up Sudlow site II in combination with FA3. FA4 site is rather solvent accessible with the water contacts of 6.14 (Table 3). Based on closed distances and computed H-bonds, FA4 can form ~2 H-bonds with N391 and Y411 and a salt bridge to R410 even though such interactions are transient. FA5 is a small linear pocket located in subdomain IIIB. It is however moderately accessible with the ~5.84 water contacts. H-bond with Y401 and a salt bridge with K525 are observed.

FA6 is a small pocket in HSA, located in domain IIB. However, it is also the 2nd most solvent accessible of the sites. From earlier X-ray and NMR results it was not clear which residues

helped to anchor the bound carboxyl group of the bound FA.⁵ Here, FA6 is found to make a salt bridge with K351 consistently and a H-bond to S480 to a lesser extent. FA7, located in subdomain IIA, is a moderately large pocket and is also known as the Sudlow site II. It is the most solvent accessible FA pocket with 2 salt bridges with K199 and R257. Despite the small pocket size, and low number of protein-FA interactions, the high degree of water exposure permit FA6 and FA7 to easily diffuse out of a pocket resulting in low binding affinity.

These results are broadly in line with previous NMR reports which suggested that 2,4,5 are the most enclosed sites and this helps to promote their high affinity for FA.⁵ The findings also support more recent NMR displacement studies which showed that FA bound to the low affinity FA6 and FA7 sites were easily displaceable.⁷ Our simulation results also appear to confirm that electrostatic interactions (H bonds and salt bridges) are certainly important for anchoring the acidic groups within the binding pockets. All sites are found to have at least 2 electrostatic bonds (at least one has to be salt bridge and another can be either a H-bond or salt bridge). However, the number of electrostatic interactions or H-bonds does not correlate well with the FA binding affinities. Indeed, more detailed calculations of the strength of the electrostatic interactions with the program APBS⁴⁹ appears to confirm this. We find the strength of the electrostatic interactions have the following order FA7 > FA6 = FA3 > FA5 > FA2 = FA4 > FA1.

FA sites 2,4, and 5 are known to be the most high affinity FA sites, yet site 1, 2, and 3 have the greatest number of H-bonds on average, sites 4 6 and 7 are the most solvent accessible, and sites 1, 2 are the largest in size. Furthermore, sites FA1, FA2 and FA3 show the greatest degree of volume fluctuation between the different protein configurations (APO, HALO, FA(1-7)). However, we can see that FAs bound to FA1, FA6, and FA7 bind in a more distorted, non-linear fashion (Table 3 & Figure 4B), indicating greater internal strain. In addition, these ligands show greater fluctuation, as given by the standard deviation, suggesting there are less effectively bound. FA3 lies in an intermediate state, presumably adversely affected by the bonding of FA to the adjacent site 4. It would appear that the selectivity of the HSA pockets towards much more diverse exogenous drug molecules will be much more complex, and not simply driven by the number of H-bonds, or VDW contacts or even the accessibility (i.e. difficult to predict).

The effect of concomitant ligand binding is important over the course of our simulations (Table 3). It is found that the FA2 pocket of the FA (1-7) model shows much greater fluctuation than in the partially occupied FA(2,4,5) model. In contrast, for FA5, ligand binding to this pocket increases the distances between two the ring centers of the π -stacked residues Y138 and Y161 significantly (0.37nm in Apo to ~0.9nm in FA(5) (Table 2), increasing the volume of the pocket,

and presumably also having an impact of the structural flexibility of the protein due to its location at the hinge point for the inter-domain movement of III and I.

3.3 Cheminformatics Analysis of HSA Ligand Binding Sites

Thus far, the focus of this study has been on assessing the unique characteristics of the seven different HSA binding sites using atomic information derived from the MD simulation data. A limitation of the univariate discussion above is that it neglects the correlation between the multiple different parameters that were extracted. To rectify this limitation we have assessed (a) the structural features and (b) the pocket characteristics derived from each simulation using PCA. We also used an orthogonal, ligand based approach to assess the features of the different sites.

3.3.1 PCA Analysis of the MD Structural Information

A PCA model was generated using the 8 distinct simulations performed and the 18 common simulation derived descriptors (Table 1-Table 3). The model fitted 3 components which describe 87% of the total variation in the dataset. The information content of each component corresponds to 42%, 36% and 13% respectively. In addition to the correlation between different models, this analysis also offers us a way to assess the impact the different starting configurations have on an individual simulation. The results for the two key components are presented in as a loadings bi-plot Figure 6A.

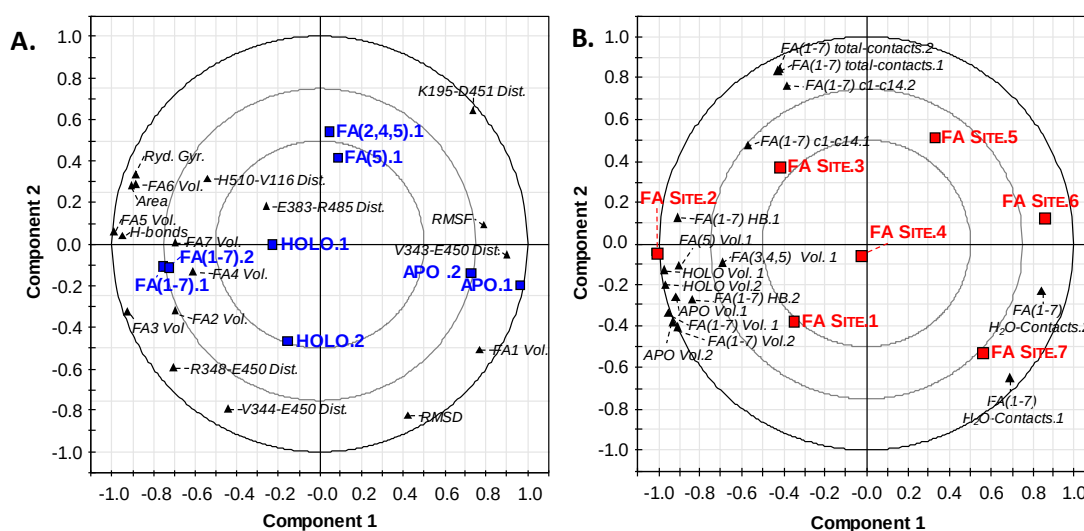


Figure 6 Loadings bi-plot for the PCA models describing the differences between the different protein configurations (right) and the differences between the binding pockets (left)

The clustering of the simulations in different areas of the plots is indicative of their relative similarity or dissimilarity. On the first component we see that the APO and FA(1-7) simulations are separated. Both replicates are closely clustered together indicating the two simulations differ only to a limited degree compared to the difference between the different simulations. The key descriptors that give rise to the separation are found at the extreme of the X-axis. For example, the radius of gyration, area, total hydrogen bonds etc are all greater in the FA(1-7) model, which are indicative of the open conformation. In contrast, the HOLO, FA(5) and FA(2,4,5) all occupy a position intermediate between the fatted and de-fatted states as is expected. Component 2, as described by the X-axis, describes the physical differences between the FA(1-7) and APO from the FA(5), FA(2,4,5) and HOLO structures. It should be noted that the HOLO structure shows the greatest deviation in terms of the distance between replicates. This difference reflects the observations in Figure 2, where there is a dramatic change in RMSD due to domain movements between IIB and IB caused by the change of a salt bridge between R117 with either E119 or E520 located nearby. Duplicate simulations were not undertaken for FA(5) and FA(2,4,5) were not undertaken, however, but we appear to observe the opposite behaviour with significantly lower RMSDs than either of the FA(1-7) simulations. It is also apparent from Figure 6A that the change in fatty acid binding has a dramatic change on the flexibility of the structure, key distances and the pocket volumes. This has dramatic implications for simulations of HSA since it is known that the effect of fatty acid can change the affinity of exogenous ligands for HSA.^{50, 51}

3.3.2 PCA Analysis of the MD based Pocket data

We next assessed the similarities and differences between the different pockets using an analysis of the FA site data extracted from the MD simulations. The FA site PCA model was generated using information from the 7 distinct simulations performed for which 12 descriptors were extracted. The model fitted 3 components which describe 94% of the total variation in the dataset. The information content of each component corresponds to 67%, 19% and 8% respectively.

The multivariate analysis of the FA binding pockets presented in Figure 6B shows the similarities between the pockets in terms of the volume, H-bond interactions, total interactions and water contacts over the course of the 8 different simulations performed. While this analysis does not take into account the specific 3D volume, or the precise nature of the H-bonds (i.e. salt bridge), they provide at least a crude estimate of their overall properties. Component 1 is dominated by the volume and H-bond, and appears to be a descriptor for the overall size of the pockets (i.e. the total no. of H-bonds correlates with the volume). In contrast, the total VDW

contacts and the H₂O contacts dominate the second component. As discussed above FA2 displays the largest volume and FA5 and FA6 the smallest. Sites FA1 and FA7 make the most VDW contacts with sites FA3 and FA5 being the least. The solvent exposure descriptor is encompassed in both Components 1 and 2, lying in the bottom right quadrant. Thus Sites FA6 and FA7 have the greatest exposure and FA2 and FA3 the least. Considering Figure 6B holistically, sites FA5 and FA6 cluster together roughly, as do sites FA1, FA4 and FA7, while sites FA3 and FA2 seem more unique.

3.3.3 PCA Analysis of the X-ray based ligand-pocket data

An orthogonal approach to exploring the HSA binding pockets is to consider the ligand properties of those molecules that are confirmed binders. Thus, a second PCA model was generated using 57 different crystallographic determined ligands of HSA (at different sites) for which 14 physico-chemical descriptors were calculated, for contrasting with the orthogonal MD based approach described above. The model fitted 4 components which describe 81% of the total variation in the dataset. The information content of each component corresponds to 37%, 20%, 13% and 11% respectively. Again, we limit our discussion to the two most significant components.

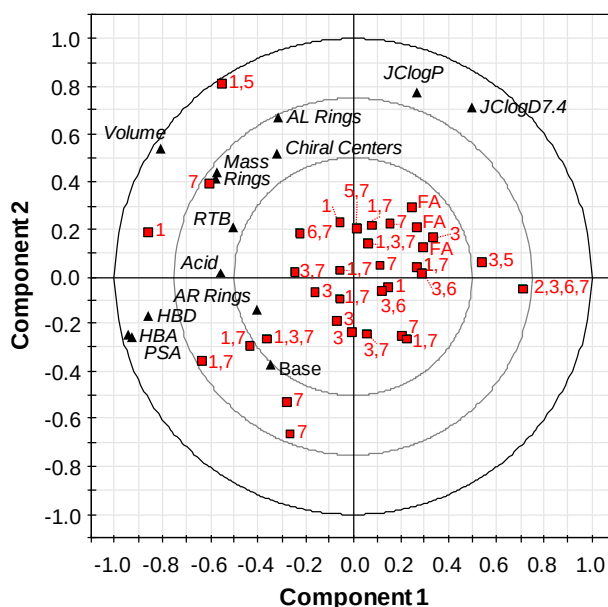


Figure 7 PCA scores (top) and loadings plots (bottom) generated from the key physicochemical properties of substrates known to bind at each binding site. RTB=rotatable bonds, PSA=polar surface area, HBA=H-bond acceptor, HBD=H-bond donor, AR Rings=aromatic rings, AL rings=aliphatic rings, while Acid and Base or ionisation state indicator variables.

Of the 57 ligands identified as binding to HSA,²⁸ 43 are acidic, 10 are neutral and 3 are zwitterionic. HEME, a forth known zwitterionic molecule was not included in the study due to the inability to calculate a number of descriptors due to the presence of the Fe^{3+} metal cofactor. The zwitterions Lysophosphatidylethanolamine and dansyl-L-arginine, are found to bind to site 1 while the latter is also found to bind to site 7. The 6 unique neutral molecules are found to occupy sites FA1, FA2, FA3, FA6 and FA7. No basic molecules were identified from the crystallographic studies. The multivariate analysis of the HSA ligands broken down by binding site is presented in Figure 7. Firstly, Component 1 primarily encode for H-bonding characteristics, with larger molecules capable of more H-bonds or salt bridge interactions (i.e. Acid class indicator) being found to the left of the x-axis. Component 2 of the PCA plot is dominated by descriptors that encode hydrophobicity, with polar molecule found at the bottom of the Y axis and hydrophobicity (i.e. JCllogP). Molecular size descriptors contribute to both component 1 and 2, with large, hydrophobic molecule capable of multiple H-bonds being found in the top left quadrant. Three non-polar FA molecules (C14, C15 and C16), which contain 1 H-bonding substituent were added as references, and these are located in the upper right hand quadrant.

In contrast to the model presented in Figure 6, we do not see a dramatic separation of the ligands that bind to HSA based on the pockets that they are found to occupy (Figure 7). Indeed, analysis of the ligands in greater detail showed that it was not possible to separate them based on bulk properties, suggesting that binding to HSA must involve a large degree of pocket specificity. Interestingly, from Figure 7 it can be seen that drug-like molecules that bind to FA1 often do so in combination with FA7 or FA3. This is consistent with the PCA model results based on the known ligands as these two sites have very similar bulk properties, which leads them to cluster together. While FA3 does not cluster with FA1 on both components, the pockets are very similar on Component 1 which indicates they have similar H-bonding and volume characteristics.

4.0 Conclusions

Ligand binding to HSA is a remarkably exciting, but very difficult task due to rationalize⁵, let alone predict due to (a) the diverse conformational states the protein can occupy (open, intermediate and closed), (b) the seven distinct binding pockets known for endogenous and exogenous ligands and (c) the interaction effect other ligands binding at other pockets can have on the binding affinity (d) good configurational sampling of protein-ligand space via replicate runs and longer sampling times⁵²⁻⁵⁴. While bulk ligand molecule properties can describe binding to a moderate degree of accuracy from QSAR analyses,^{10, 18, 19} it is clear that greater understanding of

the specificity of binding is needed before we can have confidence in predicting either the likely binding pocket or overall affinity from molecular dynamics.

In this work we have shown that different starting configurations of HSA will have implications for simulation results obtained. Indeed, this is consistent with experimental data since it is known that the effect of fatty acid can change the overall structure or affect the affinity of ligands for HSA. Additionally, the overall size, shape and dynamics of the protein is expected to have a dramatic effect on the binding pockets within HSA, and this will affect the affinity fatty acids and other drug molecules have for them. Our results confirm that differences in fatty acid binding can have a dramatic effect on the flexibility of the structure and the pocket volumes. We see that over the course of a 100ns simulation, varying the amount of FA bound to the protein allows the open form of the protein to adopt the experimentally expected intermediate conformational state, between the open and closed. Interestingly, FA sites 2, 4 and 5 are known to be the most high affinity FA sites, yet site 1,2 and 3 have the greatest number of H-bonds on average, sites 4 6 and 7 are the most solvent accessible and sites 1,2 are the largest in size. However, sites FA1, FA2 and FA3 show the greatest degree of volume fluctuation between the different simulations. Analysis of the linearity of the pockets reveals that FA2, FA4 and FA5 can bind long chained fatty acids with the lowest strain and this would help to explain their higher affinity status.

Our comparison of the properties of the HSA pockets using MD parameters or physic-chemical properties of known pocket ligands provides interesting results. In terms of their gross properties, sites FA5 and FA6 cluster together, as do sites FA1, FA4 and FA7, while sites FA3 and FA2 appear more unique. In contrast, a ligand based analysis of bound drug-like molecules shows that we do not see a dramatic separation of the ligands according to physiochemical properties and the pockets that they occupy. This suggests that binding to HSA must involve a large degree of specificity and is not heavily influenced by the bulk properties of its ligands. It therefore appears that the remarkable selectivity of the HSA pockets towards exogenous drug molecules is more complex, and is not driven as a simple function of the number of H-bonds, VDW contacts, pocket linearity, or even solvent accessibility.

Acknowledgments

This work was supported by the Thailand Research Fund grant RSA5480016 (MPG).

PART 2: Elucidating the Origin of the Esterase Activity of HSA using QM/MM Calculations.

ABSTRACT

Human serum albumin (HSA) is a critical plasma protein found, accounting for ~60% of the total protein content in blood. Remarkably, this transport protein is also found to displays esterase catalytic activity. In this study we apply theoretical studies to elucidate the origin of the esterase-like activity arising from Sudlow Site I. Using MD and QM/MM calculations we investigate which active site residues are involved in the reaction, and the precise mechanistic sequence of events. Our results suggest Lys199, His242, Arg257 give rise to the esterase activity and that the most catalytically efficient active site configuration requires that both Lys199 and Aspirin are in their neutral forms. The abundance of HSA in the body suggest the protein might be a suitable target for the computational guided design of acetyl based pro-drugs of acidic molecule that often displayed limited oral exposure due to their unmasked ionizable substituent.

Introduction

Human serum albumin (HSA) is an extremely important transport protein in blood plasma, accounting for approximately 60% of the total protein content overall. HSA is a monomeric protein which consists of 585 amino acids with molecular mass of 66.5 KDa.³ There are three structural domains, each domain can be divided into 2 subdomains, termed IA, IB, IIA, IIB, IIIA and IIIB (Figure 1). HSA binds a range of endogenous and exogenous substrates, including fatty acids¹¹ and molecules such as Warfarin¹⁶ and acetyl-salicylic acid (Aspirin).¹⁵ Long chained fatty acid molecules can bind to 7 sites distributed across the protein^{6, 29} while drug-like molecules are typically found at 1 of 2 sites, termed the Sudlow sites I and II.¹² Site I is located within subdomain IIA while Site II is located within subdomain IIIA.^{12, 15}

The primary function of HSA is to transport endogenous compounds such as fatty acids and steroids throughout the body. However, it also display remarkable ability to deacylate Aspirin to its active metabolite salicylic acid.^{55, 56} Crystallographic studies by Yang *et al.*⁵⁷ have shed light on the origin of the esterase activity arising from Site I. The structural coordinates reveal that Lys199 is in an acylated state with Aspirin bound nearby, suggesting adjacent residues must facilitate the generation of an amine nucleophile and help to stabilize any excited state species along the reaction coordinate. Moreover, it opened up the possibility that HSA could be used to activate specifically designed prodrugs⁵⁷. This strategy would be attractive from a safety point of view, given the abundance of HSA in the body means that drug-drug interactions are not likely to arise.^{12,13} Furthermore, the rational design of prodrugs would also offer drug discovery scientist a useful way of overcoming exposure limitations associated with acidic compounds.⁵⁸⁻⁶⁰

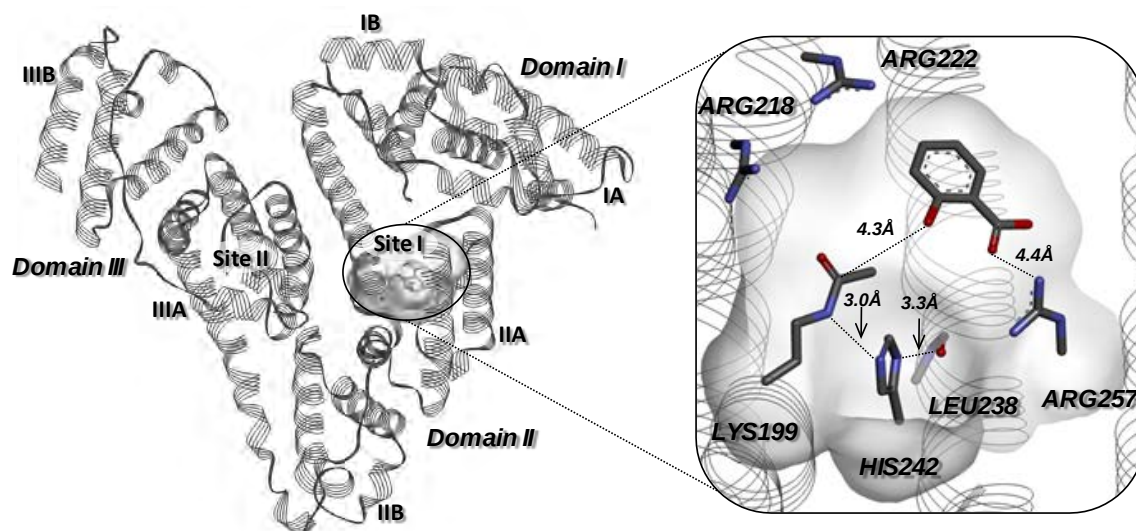


Figure 8 Illustration of the HSA protein structure (PDB 2I2Z). Structural data confirms Sudlow Site I plays a role in the deacylation of Aspirin based on X-ray crystallography data (PDB ID 2I2Z).

The esterase activity possessed by HSA is very similar to that catalyzed by the acetylcholine esterase (Ache) family of proteins.⁶¹ Ache contains the so called catalytic triad consisting of a serine nucleophile, a histidine base and an acidic group such as glutamate that can stabilize the histidine carbocation on generating the negatively charged serine nucleophile required for C-O bond breaking (Figure 2A).⁶² The serine nucleophile can attack the carbonyl of the substrate forming a negatively charged tetrahedral intermediate which is stabilized by 2 hydrogen bonds from adjacent glycine residues (the rate determining step). Decomposition of this intermediate leads to the breaking of the substrate C-O bond, through the loss of the hydroxide leaving group. This results in the formation of the acylated serine residue and the deprotonation of histidine to give the corresponding hydroxyl moiety (Figure 3). Analysis of Site I HSA reveals that if the imidazole of His242 is rotated 180° then a catalytic triad is observed,⁶² consisting of a lysine nucleophile⁶³, a histidine base and a carbonyl group from the protein backbone capable of stabilizing a histidine carbocation (Figure 2B). Site I also contains a number of basic residues and bound water molecules, however these are not in the same optimal position as the oxyanion hole found in Ache. Thus, while this non-standard catalytic triad⁶² could be the origin of catalytic activity, is not totally clear what is the sequence of events during the Aspirin deacylation reaction, and the range of residues involved.

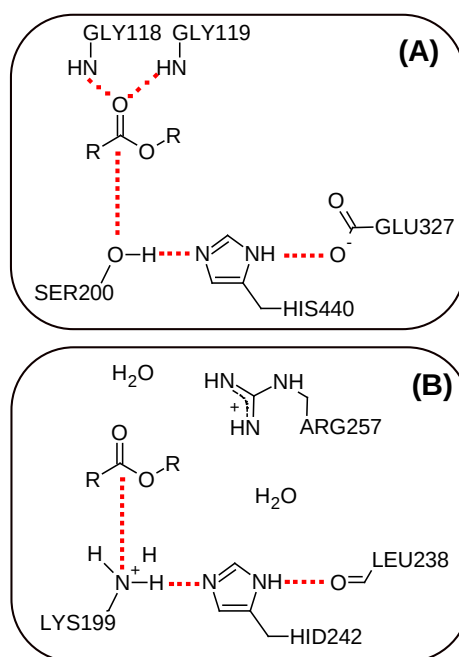


Figure 9 Illustration of similarities and differences between Acetylcholine Esterase (2C4H)⁶⁴ vs HSA (2I2Z)⁵⁷. The latter corresponds to the classic catalytic triad configuration.⁶²

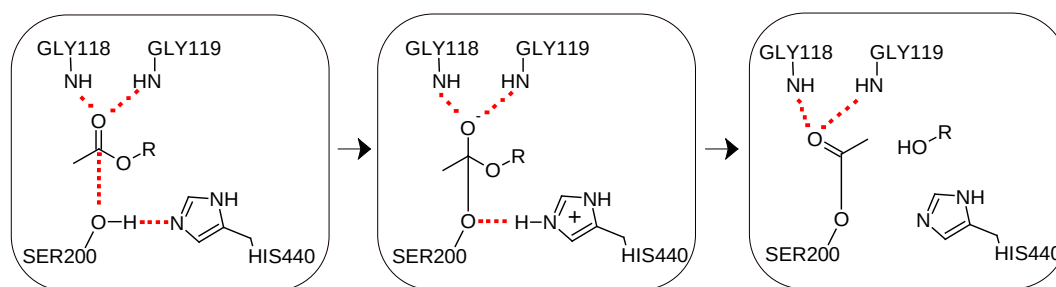


Figure 10 Catalytic reaction involved in the ester bond breakage by AChE.^{62, 64}

Molecular simulations have been employed to study a wide variety of phenomena associated HSA.²⁸ The methods employed have included quantitative structure activity relationships (QSAR)^{10, 18, 19}, molecular dynamics (MD)^{22-24, 27, 65, 66} and ligand docking.^{22, 25, 26} For example, Diaz *et al.*⁶⁵ used MD simulations to study the protonation state of Lys195 and Lys199 ins Site I and its implications for ligand binding and selectivity. Castellanos *et al.*²⁷ investigated how specific disulphide bond are critical for the structural flexibility of the protein. Fujiwara *et al.*²⁴ who used MD simulations to study the effect fatty acid binding has on warfarin affinity for HSA. Also worth noting is the work of Alvarez *et al.*⁶⁷ used MD to assess the binding free energy of Aspirin to HSA. Very recently, Etienne *et al.*⁶⁸ used quantum mechanical/molecular mechanical (QM/MM) methods to study the absorption spectra of the protein with good accuracy. QM/MM is a hybrid technique that allows the reactivity of large biomolecules to be simulated.^{69, 70} In this approach the substrate

and catalytic residues located in the protein active site are treated using QM while the remainder is described using as MM forcefield. In this way bond breaking and formation can be simulated in the presence of the extended protein framework and solvent. Aspirin works by selectively acetylates the hydroxyl group of Ser530⁷¹ on cyclooxygenase (COX) and this has been studied from a QM/MM perspective.⁷²

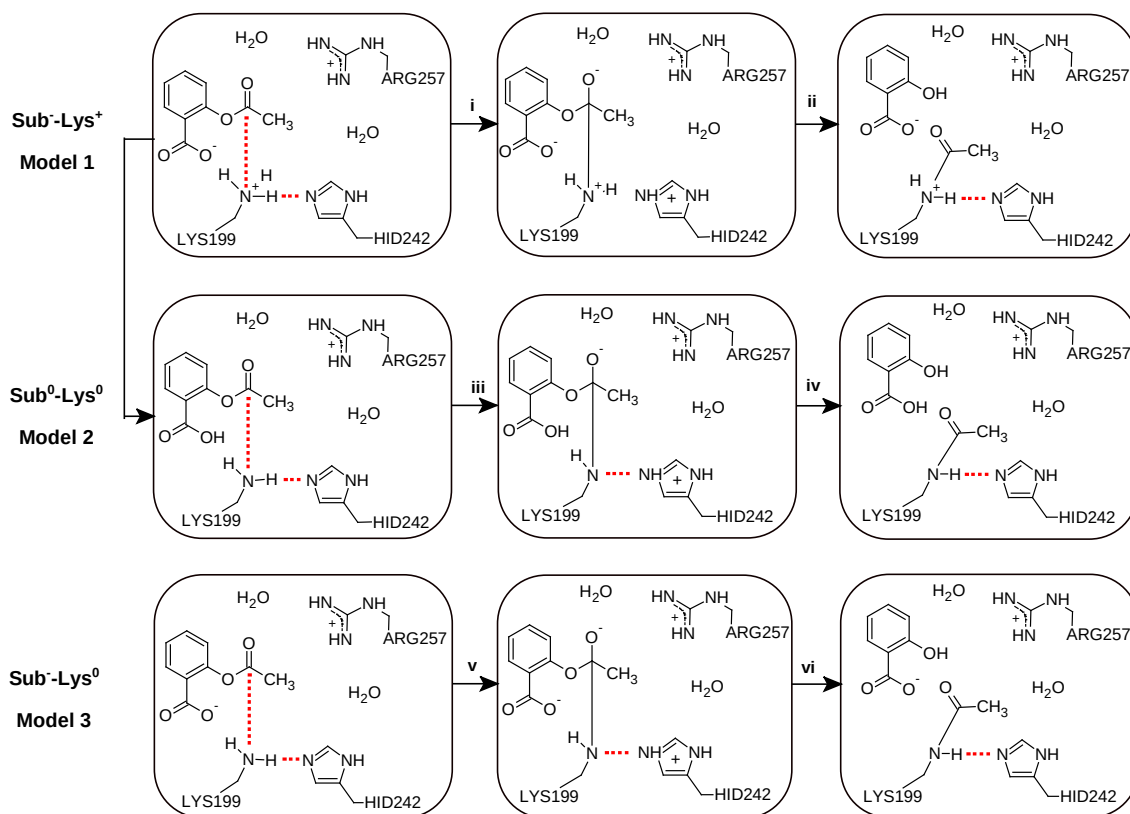


Figure 11. Three possible HSA reaction mechanisms leading to the deacylation of Acetyl salicylic acid. Each of the three models simulations have active sites with different protonation configurations.

In this work we report the use of MD and QM/MM methods to study the deacylation mechanism of Aspirin at the Site I of HSA. As noted by Diaz *et al.*⁶⁵, multiple protonation configuration are possible within the active site. We therefore simulate 3 different configurations where Lys199 is either protonated (Lys⁺) or deprotonated (Lys⁰) and Aspirin is either protonated (Sub⁰) or deprotonated (Sub⁻). These are termed the Sub⁻-Lys⁺, Sub⁻-Lys⁰ and Sub⁰-Lys⁰ models. We assess which HSA configuration is more likely to contribute to the esterase activity by performing MD simulations with Aspirin bound to see which leads to the greatest number of reactive confirmations over a 10 ns timeframe. We subsequently assess the reactivity of the three different configurations using hybrid QM/MM methods to compute the complete reaction coordinate (Figure 4). The key goal of this work is to further our understanding of HSA reactivity

and to try assess whether this abundant protein could conceivably be used in a safe,⁷³ effective prodrug activation strategy.⁷⁴

Computational Methods

HSA models were constructed from PDB structure 2I2Z⁵⁷ which was downloaded from the RCSB databank.³⁰ The acetyl salicylic acid substrate was generated by removing the acetyl group from Lys199. Fatty acid molecules bound to HSA were retained. The imidazole ring of His242 was rotated 180° such that it made strong interactions with both Lys199 and the backbone carbonyl of Leu238. The N- and C-terminal ends of each model were capped with acetyl (ACE) and methyl amino (NME) groups, respectively, and the overall quality confirmed using PROCHECK v3.5.4.¹ The protonation state of ionisable residues were determined using PROPKa³¹ and a visual analysis of the environment surrounding each ionisable residue. Residues were treated as HID unless otherwise stated: HIE (39, 288, 40) and HIP (3, 67, 105, 128, 146, 247, 338, 367, 510). Finally, the protonation state of the substrate and Lys199 were manually altered to create the 3 separate models: Sub⁻-Lys⁺, Sub⁰-Lys⁰ and Sub⁻-Lys⁰ (Figure 4). Sub⁻-Lys⁺ differs from Sub⁰-Lys⁰ only in the position of one proton while the Sub⁻-Lys⁰ system has 1 less proton. Each model was then placed in a 12 Å rectangular box containing water. The system was neutralized with counter ions (0.15 mM of NaCl) by randomly replacing water molecules.

2.1 MD Simulations

Simulations were carried out using GROMACS v4³² with the AMBER99SB forcefield.³³ Ligand topology files were generated using ACPYPE script⁶ and the GAFF force field³⁴. The TIP3P water model was employed. Ligand electrostatic charges were generated using HF/6-31G(d) optimized structures (Gaussian 09 Revision d01)⁷⁵ and the RESP protocol of Ambertool1.5.³⁶ Long range electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method with a 0.12 nm cut-off³⁷ while van der Waals interactions utilized a 1 nm cut-off. Simulations were performed at constant temperature, pressure and number of particles (NPT). The temperature of the protein and solvent were each coupled separately. The Berendsen thermostat³⁸ was applied at 300 K with a coupling constant of $\tau = 0.1$ ps.³⁹ Coordinates and velocities were saved every 2 ps. The LINCS algorithm was used to restrain bond lengths and a 2 fs timestep was used for integration.⁴⁰ Two hundred ps of protein backbone-restrained dynamics were employed initially, followed by 10 ns production runs. The total energy, RMSD, RMSF and the key distances associated with hydrogen transfer (His242-Lys199, ND---HZ) and nucleophilic attack (Lys199-substrate, NZ---C) were extracted for further analysis.

2.2 QM/MM Calculations

The energy minimized HSA structures were used as input for the ONIOM QM/MM calculations. All but two water molecule bound within the Site I cavity were excluded from the charge neutral system for reasons of computational efficiency. The QM region consisted of acetyl-salicylic acid, the side chains of Lys199, His242, Arg257 and the two water molecules bound within the cavity. Arg257 was included as it adopts a position close to where the oxyanion hole of AChE is found.⁶² All MM atoms were treated flexibly except for residues >12 Å from the active site and link atoms.

All ONIOM calculations were performed using the electrical embedding scheme as implemented in Gaussian 09 (d01).⁷⁵ All stationary points were optimized at the M062X/6-31G+(d,p) level of theory and confirmed by vibrational frequency analysis. Minima showed no negative frequencies and transition states displayed a single high intensity negative frequency. ΔG values were estimated by adding the vibrational and thermal corrections to the ΔH .

3.0 Results and Discussion

The preferred protein-ligand active site configuration required for reactivity⁵⁷ was initially evaluated using the results from the MD simulations. This was done by monitoring the two key interactions that define the reaction coordinate (Figure 5). We evaluate the H-bond distance between Lys199 and HID242 (NH---:N) and the distance between Lys199 NZ atom and the substrate carbonyl carbon atom (HN:---C=O). The active site model that maintains these distances in an optimal configuration is at a distinct advantage in terms of the probability of the reaction occurring (Figure 6).

Subsequently, we assess the reactivity of the 3 different protein-ligand models towards the deacylation of the substrate using the hybrid QM/MM scheme previously described (Figure 8 & Figure 7). We then use the computed rate determining step in conjunction with the data derived from the MD simulations to determine whether one particular protein-ligand configuration plays a dominant role in the catalytic process.

3.1 MD Results

Minimal variation in the total energy found over the course of the three simulation indicates they had converged after the 200 ps of pre-equilibration. In addition, the change in C $_{\alpha}$ RMSDs after the 10 ns simulation is consistent with other reports on HSA (Figure S2) suggesting the protein is dynamically well behaved.²³

For a reaction to occur within HSA, the two key coordinates illustrated in Figure 2 (proton transfer and nucleophilic attack) must be in a desirable configuration to enable reaction. Thus,

along with mean values, we assessed the frequency with which each model systems maintained a H-transfer coordinate (NH---:NH) distance < 0.25 nm, and the nucleophilic attack coordinate (HN---:C=O) maintains a distance < 0.40 nm (Figure 5). These values were chosen as within these cut-offs subtle translation of the flexible lysine side chain or the unbound ligand is sufficient to result in optimal starting geometry (i.e. NH---:N ~ 0.20 and HN---:C=O ~ 0.30 nm). We found that the NH---:N distances were 0.44±0.08, 0.56±0.18, and 0.59±0.12 nm for the Sub⁻-Lys⁺, Sub⁰-Lys⁰ and Sub⁻-Lys⁰ simulations, well beyond the ideal values. We also found that the HN---:C=O was not in a range that would facilitate reaction easily, with average distances of 0.85±0.14, 0.55±0.10, and 1.04±0.12 nm respectively for the Sub⁻-Lys⁺, Sub⁰-Lys⁰, and Sub⁻-Lys⁰ models. The percentage of cases where the NH---:NH distances were within an ideal range, were 2.2, 8.5, and 2.5%. The corresponding values for the HN---:C=O distance were 0.0% 6.6% and 0.0%, respectively. Furthermore, the angle for proton transfer N-H-N ~180° and nucleophilic attack N:-C=O ~120° are also found to be optimal for Sub⁰-Lys⁰.

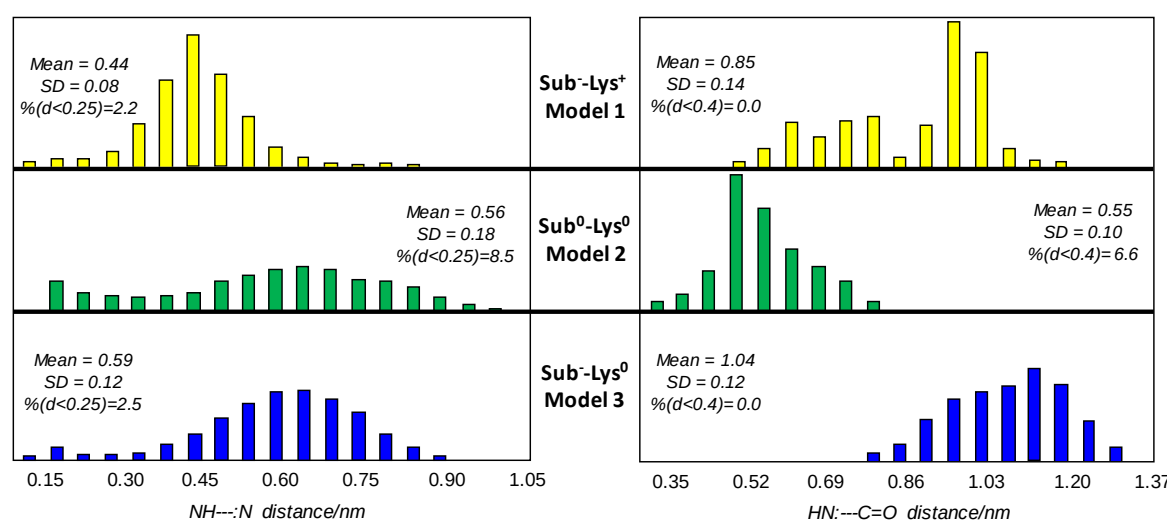


Figure 12. Distributions associated with the nucleophilic attack (left) and proton transfer distances (right) for the Sub⁻-Lys⁺ (top), Sub⁰-Lys⁰ (middle) and Sub⁻-Lys⁰ models (bottom) over the course of the MD simulations.

A visual analysis of the simulations results revealed that the substrate showed marked structural freedom to rotate and translate within the rather large Site I pocket.⁵⁷ The substrate was found to rotate and translate within the pocket, maintaining optimal interaction between its acidic substituent and at least 1 adjacent basic residues (Site I contains; Lys195, Lys199, Arg218, Arg222 & Arg257). The dramatic instability of the catalytic triad is primarily due to the rather weak interaction between the Lys-His triad units being easily broken during the simulation, in preference to salt-bridge interactions, particularly between the acidic group of the substrate which

can easily rotate and translate to interact with Lys199. Both the $\text{HN}:\cdots\text{C}=\text{O}$ and $\text{HN}:\cdots\text{N}$ distance must be in the necessary configuration consistent with other forms of catalytic triads for HSA to have activity. The frequency with which both of these interactions are formed will be indicative of the probability of reaction and from (Figure 6) it is clear that the $\text{Sub}^0\text{-Lys}^0$ model has a distinct advantage over both the $\text{Sub}^-\text{-Lys}^+$ and $\text{Sub}^-\text{-Lys}^0$ models. The fact that neither model is in an ideal configuration all of the time is consistent with the known catalytic activity of HSA, which is a much weaker metabolizer of Aspirin than its principle activator human butyrylcholinesterase (BChE).⁷⁶

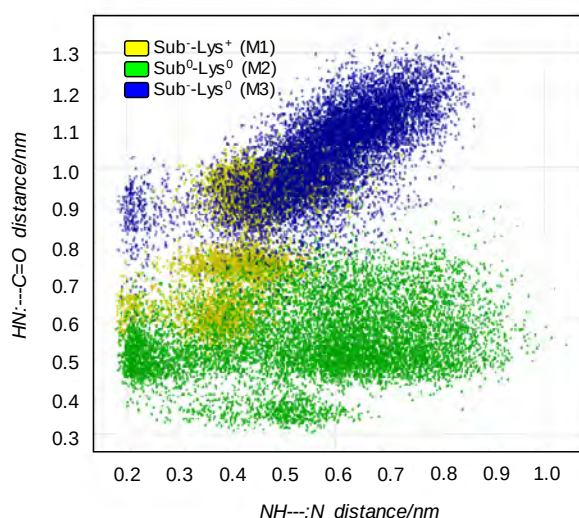


Figure 13. Plot of the distance corresponding to the nucleophilic attack (Y-axis) versus proton transfer distance (X-axis) for the $\text{Sub}^-\text{-Lys}^+$, $\text{Sub}^0\text{-Lys}^0$ and $\text{Sub}^-\text{-Lys}^0$ model for each time point of the MD simulation.

These results would suggest that Lys199 and the substrate may be in neutral form prior to the reaction commencing. This is consistent with the fact that the acetoxyl group of Aspirin must position itself in such a way that it faces toward Lys199.⁵⁷ Furthermore, the finding that Lys199 in the neutral state is preferred is consistent with earlier studies by Diaz *et al.*⁶⁵ Their MD investigations suggested the presence of water bridges within Site I can potentially promote a Lys195 - Lys199 proton-transfer process. This would add weight to the proposal that the catalytically active configuration involves Lys199 in the neutral form. As for the protonation state of the substrate, it is clear that the neutral form results in a catalytically more preferable active site configuration. This must be balanced with the fact that its acidic pKa favors the ionized form, although the effect of the surrounding environment will certainly modify this.^{77, 78} Furthermore, it must be noted that nucleophilic attack of an electron rich acidic molecule will be less favorable than the attack of the neutral form.

3.2 QM/MM Results

The esterase activity of HSA was subsequently assessed according to the proposal presented in Figure 4. All structures were fully optimized at the M062X/6-31+G** and transition states were then determined. It was quickly identified that the pKa miss-match between Lysine (basic pKa ~ 10) and Histidine (basic pKa ~7) were too large to allow for full proton transfer between. Analysis of the resulting Mulliken charge distribution shows that His242 plays a role in stabilizing the excited state species on the reaction coordinate. Thus, the Sub⁻-Lys⁺ configuration was found to be unreactive. His242 is unable to accept a proton from Lys199 and Lys199 is therefore not sufficiently nucleophilic to attack the Aspirin carbonyl group to form the tetrahedral intermediate. The Sub⁻-Lys⁺ reactant is however capable of interconverting into the Sub⁰-Lys⁰ reactant (3.8 kcal/mol more exothermic) with a barrier of just 6.0 kcal/mol (Figure 7 & Table 1). The proton on Lys199 can be transferred to carboxylate group of Aspirin by a proton shuttle mechanism involving an adjacent water molecule. The distance between the proton being transferred from Lys199 to water drops from 0.18 to 0.12 nm in the transition state, while the distance between the water proton and the Aspirin carboxyl group goes from 0.17 to 0.12 nm (Figure 8).

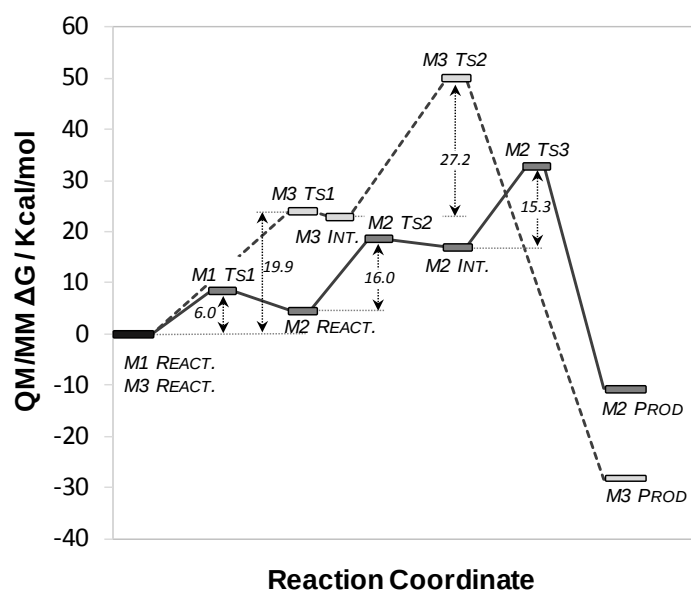


Figure 14. Predicted QM/MM Reaction Coordinate for HSA deacylation of Aspirin acid to salicylic acid.

Nucleophilic attack of the Acetyloxy group of Aspirin can only occur within the Sub⁰-Lys⁰ and Sub⁻-Lys⁺ models. In the former case, nucleophilic attack results in a barrier of 16 kcal/mol while in the former, a barrier of 19.9 kcal/mol is observed. This is the rate determining step for the Sub⁰-Lys⁰ model. In both models, the transition state does not see proton transfer from Lys199 to

His242 given since the NH₂ moiety is already reasonably nucleophilic itself. However, we do observe charge transfer between the Lys199 and His242 with the net charge on the histidine residue increasing by +0.19 and +0.13, respectively on the Sub⁰-Lys⁰ and Sub⁻-Lys⁺ models, respectively. The HN:---C=O distances observed in the transition state are 0.19 nm to 0.21 nm while the corresponding NH---:N distance decreases slightly from 0.19 nm to 0.20 nm, for both models, respectively (Figure 8 and Figure S1). This indicates the former system provides greater charge stabilization.

The tetrahedral intermediate is found to be only weakly stable, with a barrier to reverse reaction of < 2 kcal/mol for both models. QM/MM studies on the deacylation mechanism of Aspirin by Ser350 in COX were unable to locate a stable tetrahedral intermediate.⁷² We observed this artifact using a smaller 6-31G* basis set, as used in that study. However, calculations performed with additional diffuse functions, as is the case for M062X/6-31+G** used here, allow for more effective delocalization of the negative charge in the tetrahedral intermediate.⁷² Lys199 is found in the protonated form, which is again expected given its higher basic pKa compared to His242. The HN:---C=O distances in the intermediates is 0.16 and 0.17 nm for the Sub⁰-Lys⁰ and Sub⁻-Lys⁺ models, and NH---:N distance is 0.17 and 0.19 nm, respectively. Analysis of the charge distributions shows the charge on Arg257 changes by -0.10 for the former model, as opposed to +0.04 for the former. This indicates the Sub⁰-Lys⁰ is more capable of stabilizing the zwitterionic intermediate.

The barrier to formation of the acylated Lys199 residue is found to be the rate determining step in the Sub⁻-Lys⁺ model (27.2 kcal/mol). The energy of the Sub⁰-Lys⁰ is just 15.3 kcal/mol, slightly lower than that of the first step. The transition state associated with this step involves the breaking of the C-O acetyloxy bond of Aspirin and the protonation of the resulting hydroxide group. The huge discrepancy between the two different models is that in the Sub⁻-Lys⁺, as the C-O bond breaks, the resulting hydroxyl anion is destabilized by the already ionized carboxyl group present in the molecule. In contrast, the protonated carboxyl group in the Sub⁰-Lys⁰ is capable of stabilizing the resulting hydroxide anion as it forms through an intermolecular H-bond (Figure 8). The net charge on the substrate is -0.76 for the Sub⁰-Lys⁰, while for Sub⁻-Lys⁺ the value is -1.83, respectively.

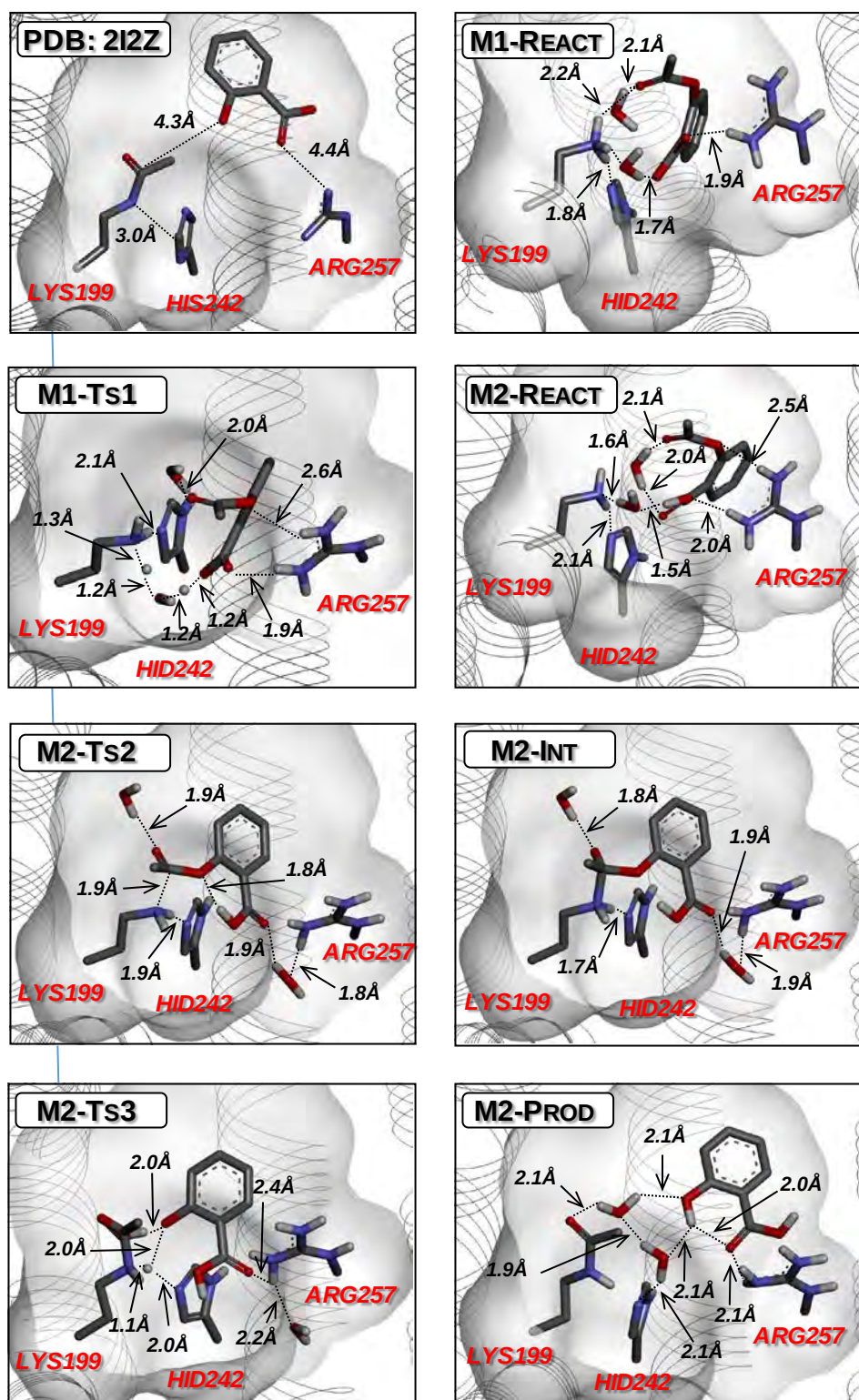


Figure 15. Graphical illustration of the QM/MM optimized minima and transition states for mechanism 2.

The corresponding C-O distance in the transition states are found to be 0.21 and 0.22, while the O-H transfer distances are 0.20 and 0.18, for both models respectively. The proton being transferred comes from Lys199 since the resulting amide is non-basic in character, and involves

a distorted 4-membered arrangement. The result of traversing the final transition state is an energetically favorable product for both models. As the bond breaks, acet yl-salicylic translates 0.1 nm, which completes the proton transfer and leads to the rotation of the carbonyl to allow for an internal H-bond. The Sub⁰-Lys⁰ is found to be exothermic by -11.3 compared with -30.6 kcal/mol for Sub⁻-Lys⁺. This is a reflection of the ability of the negatively charged salicylic acid molecule to form a strong salt-bridge interaction with Arg257 (Figure S1). This is completely consistent with the ligand selectivity displayed by HSA.¹⁵

The main finding from the QM/MM calculations is that Site I in the Sub⁰-Lys⁰ configuration results in a dramatically lower barrier to reaction. This is due to electronic destabilization of the resulting transition states associated with both nucleophilic attack and the departure of the hydroxide leaving group. The rate determining barrier of 16.0 kcal/mol is correctly predicted to be higher in energy than that found in AChE (12.4 kcal/mol)⁷⁹, a protein specifically evolved for esterase activity. Additionally, the MD results show that not all conformations will be catalytically active which, taken together, is consistent with its lower esterase activity ($\sim 10^{-3} \text{ s}^{-1}$) compared to AChE ($\sim 10^4 \text{ s}^{-1}$).⁶¹

Conclusions

MD and QM/MM simulations were employed to shed greater light on the esterase activity of HSA derived from the Site 1 pocket. We assessed the 3 different protonation states that could contribute to the catalytic activity. We found that a model consisting of Lys199 and the substrate in their neutral forms (a) resulted in active site configurations more capable of undergoing reaction and (b) led to dramatically lower rate determining barriers for the deacylation of Aspirin. The computed rate determining energy barrier of 16.0 kcal/mol is higher than that of AChE (12.4 kcal/mol), a protein evolved specifically for its esterase activity.⁷⁹ We found that His242 did act to stabilize the reaction however it could not function as a base during the reaction due to the mismatch in pKa with Lys199.

Structural studies Yang *et al.* led to the suggestion that molecules could be specifically designed for activation at HSA (i.e. diflunisal derivatives).⁵⁷ This proposal could be particularly useful to overcome exposure limitations associated with acidic compounds.⁵⁸⁻⁶⁰ This idea could be extended further with computationally driven methods to identify or design molecules to be activated by the protein^{57, 74}. Employing HSA in this strategy would be attractive from a safety point of view given the abundance of the protein in the body would mean that drug-drug interactions were less likely to arise as they do for other metabolic pathways.^{73, 80} Furthermore,

Site II display different ligand selectivity than Site I,⁸¹ and it too displays esterase activity (p-nitro-phenyl-acetate via Arg410 and Tyr411)^{82, 83}, offering further opportunity for design.

Acknowledgments

M.P. Gleeson would like to acknowledge financial support from The Thailand Research Fund (TRF) (RSA5480016). O. Phuangsawai would like to acknowledge a Development and Promotion of Science and Technology Grant (DPST) for supporting her studies.

4. เอกสารอ้างอิง (REFERENCES)

1. Rowland, M.; Benet, L.; Graham, G. Clearance concepts in pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics* **1973**, 1, 123-136.
2. Hamilton, J. A. NMR reveals molecular interactions and dynamics of fatty acid binding to albumin. *Biochimica et Biophysica Acta* **2013**, 1830, 5418-26.
3. Carter, D.; He, X. Structure of human serum albumin. *Science* **1990**, 249, 302-303.
4. Curry, S.; Mandelkow, H.; Brick, P.; Franks, N. Crystal structure of human serum albumin complexed with fatty acid reveals an asymmetric distribution of binding sites. *Nat Struct Biol* **1998**, 5, 827-35.
5. Simard, J. R.; Zunszain, P. A.; Hamilton, J. A.; Curry, S. Location of high and low affinity fatty acid binding sites on human serum albumin revealed by NMR drug-competition analysis. *J Mol Biol* **2006**, 361, 336-51.
6. Simard, J. R.; Zunszain, P. A.; Ha, C. E.; Yang, J. S.; Bhagavan, N. V.; Petitpas, I.; Curry, S.; Hamilton, J. A. Locating high-affinity fatty acid-binding sites on albumin by x-ray crystallography and NMR spectroscopy. *Proc Natl Acad Sci U S A* **2005**, 102, 17958-63.
7. Krenzel, E. S.; Chen, Z.; Hamilton, J. A. Correspondence of fatty acid and drug binding sites on human serum albumin: a two-dimensional nuclear magnetic resonance study. *Biochemistry* **2013**, 52, 1559-67.
8. Junk, M. J.; Spiess, H. W.; Hinderberger, D. The distribution of fatty acids reveals the functional structure of human serum albumin. *Angew Chem Int Ed Engl* **2010**, 49, 8755-9.
9. Valko, K.; Nunhuck, S.; Bevan, C.; Abraham, M. H.; Reynolds, D. P. Fast gradient HPLC method to determine compounds binding to human serum albumin. Relationships with octanol/water and immobilized artificial membrane lipophilicity. *J. Pharm. Sci.* **2003**, 92, 2236-48.
10. Colmenarejo, G.; Alvarez-Pedraglio, A.; Lavandera, J. L. Cheminformatic models to predict binding affinities to human serum albumin. *J. Med. Chem.* **2001**, 44, 4370-8.
11. Curry, S.; Brick, P.; Franks, N. P. Fatty acid binding to human serum albumin: new insights from crystallographic studies. *Biochim Biophys Acta* **1999**, 1441, 131-40.
12. Sudlow, G.; Birkett, D. J.; Wade, D. N. The characterization of two specific drug binding sites on human serum albumin. *Mol Pharmacol* **1975**, 11, 824-32.
13. Kragh-Hansen, U.; Watanabe, H.; Nakajou, K.; Iwao, Y.; Otagiri, M. Chain length-dependent binding of fatty acid anions to human serum albumin studied by site-directed mutagenesis. *J Mol Biol* **2006**, 363, 702-12.

14. Zunszain, P. A.; Ghuman, J.; Komatsu, T.; Tsuchida, E.; Curry, S. Crystal structural analysis of human serum albumin complexed with hemin and fatty acid. *BMC Struct Biol* **2003**, 3, 6.
15. Ghuman, J.; Zunszain, P. A.; Petitpas, I.; Bhattacharya, A. A.; Otagiri, M.; Curry, S. Structural basis of the drug-binding specificity of human serum albumin. *J Mol Biol* **2005**, 353, 38-52.
16. Petitpas, I.; Bhattacharya, A. A.; Twine, S.; East, M.; Curry, S. Crystal structure analysis of warfarin binding to human serum albumin: anatomy of drug site I. *J Biol Chem* **2001**, 276, 22804-9.
17. Ryan, A. J.; Ghuman, J.; Zunszain, P. A.; Chung, C. W.; Curry, S. Structural basis of binding of fluorescent, site-specific dansylated amino acids to human serum albumin. *J Struct Biol* **2011**, 174, 84-91.
18. Gunturi, S. B.; Narayanan, R.; Khandelwal, A. In silico ADME modelling 2: computational models to predict human serum albumin binding affinity using ant colony systems. *Bioorg Med Chem* **2006**, 14, 4118-29.
19. Estrada, E.; Uriarte, E.; Molina, E.; Simon-Manso, Y.; Milne, G. W. An integrated in silico analysis of drug-binding to human serum albumin. *J Chem Inf Model* **2006**, 46, 2709-24.
20. van Westen, G. J. P.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm* **2011**, 2, 16-30.
21. Gleeson, M. P.; Modi, S.; Bender, A.; L. Marchese Robinson, R.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; C. Glen, R. The challenges involved in modeling toxicity data in silico: a review. *Curr. Pharm. Des.* **2012**, 18, 1266-1291.
22. Deeb, O.; Rosales-Hernández, M. C.; Gómez-Castro, C.; Garduño-Juárez, R.; Correa-Basurto, J. Exploration of human serum albumin binding sites by docking and molecular dynamics flexible ligand–protein interactions. *Biopolymers* **2010**, 93, 161-170.
23. Fujiwara, S.; Amisaki, T. Molecular dynamics study of conformational changes in human serum albumin by binding of fatty acids. *Proteins* **2006**, 64, 730-9.
24. Fujiwara, S.; Amisaki, T. Steric and allosteric effects of fatty acids on the binding of warfarin to human serum albumin revealed by molecular dynamics and free energy calculations. *Chem Pharm Bull (Tokyo)* **2011**, 59, 860-7.
25. Abou-Zied, O. K.; Al-Lawatia, N.; Elstner, M.; Steinbrecher, T. B. Binding of hydroxyquinoline probes to human serum albumin: combining molecular modeling and Forster's resonance energy transfer spectroscopy to understand flexible ligand binding. *Journal of Physical Chemistry B* **2011**, 115, 1062-74.
26. Subramanyam, R.; Gollapudi, A.; Bonigala, P.; Chinnaboina, M.; Amooru, D. G. Betulinic acid binding to human serum albumin: a study of protein conformation and binding affinity. *Journal of Photochemistry and Photobiology. B, Biology* **2009**, 94, 8-12.
27. Castellanos, M. M.; Colina, C. M. Molecular Dynamics Simulations of Human Serum Albumin and Role of Disulfide Bonds. *The Journal of Physical Chemistry B* **2013**, 117, 11895-11905.
28. Fujiwara, S.; Amisaki, T. Fatty acid binding to serum albumin: molecular simulation approaches. *Biochim Biophys Acta* **2013**, 1830, 5427-34.
29. Bhattacharya, A. A.; Grune, T.; Curry, S. Crystallographic analysis reveals common modes of binding of medium and long-chain fatty acids to human serum albumin. *J Mol Biol* **2000**, 303, 721-32.
30. RCSB Protein Data Bank. <http://www.rcsb.org/>

31. Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Structure, Function, and Bioinformatics* **2005**, 61, 704-721.
32. Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular modeling* **2001**, 7, 306-317.
33. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **2006**, 65, 712-725.
34. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **2006**, 25, 247-260.
35. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*, Gaussian, inc: Wallingford CT, 2004.
36. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; et al.; . AMBER 12, Francisco, CA:
37. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J Chem Phys* **1993**, 98, 10089-10092.
38. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **1984**, 81, 3684-3690.
39. Nosé, S. A Molecular-Dynamics Method for Simulations in the Canonical Ensemble. *Molecular Physics* **1984**, 52, 255-268.
40. Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput Chem* **1997**, 18, 1463-1472.
41. Discovery Studio 2.5, Accelrys
42. Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, 14, 33-38.
43. ChemAxon JChem.
44. Umetrics. *SIMCA-P 10*, Umetrics: Uvitevägen 48, Box 7960, SE-907 19 Umeå, Sweden.
45. Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: a perspective. *J Med Chem* **2010**, 53, 3862-86.

46. Curry, S. Lessons from the crystallographic analysis of small molecule binding to human serum albumin. *Drug Metabolism and Pharmacokinetics* **2009**, 24, 342-57.
47. Fujiwara, S. I.; Amisaki, T. Molecular dynamics study of conformational changes in human serum albumin by binding of fatty acids. *Proteins-Structure Function and Bioinformatics* **2006**, 64, 730-739.
48. Fujiwara, S. I.; Amisaki, T. Fatty acid binding to serum albumin: Molecular simulation approaches. *Biochimica et Biophysica Acta* **2013**.
49. Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences* **2001**, 98, 10037-10041.
50. Vorum, H.; Honoré, B. Influence of Fatty Acids on the Binding of Warfarin and Phenprocoumon to Human Serum Albumin with Relation to Anticoagulant Therapy. *Journal of Pharmacy and Pharmacology* **1996**, 48, 870-875.
51. Birkett, D. J.; Myers, S. P.; Sudlow, G. Effects of Fatty Acids on Two Specific Drug Binding Sites on Human Serum Albumin. *Molecular Pharmacology* **1977**, 13, 987-992.
52. Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *The Journal of Physical Chemistry B* **2010**, 114, 10235-10253.
53. de Ruiter, A.; Oostenbrink, C. Free energy calculations of protein–ligand interactions. *Current Opinion in Chemical Biology* **2011**, 15, 547-552.
54. Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. Basic ingredients of free energy calculations: A review. *Journal of Computational Chemistry* **2010**, 31, 1569-1582.
55. Kurono, Y.; Yamada, H.; Ikeda, K. Effects of drug binding on the esterase-like activity of human serum albumin. V. Reactive site towards substituted aspirins. *Chem Pharm Bull (Tokyo)* **1982**, 30, 296-301.
56. Honma, K.; Nakamura, M.; Ishikawa, Y. Acetylsalicylate-human serum albumin interaction as studied by NMR spectroscopy--antigenicity-producing mechanism of acetylsalicylic acid. *Mol. Immunol.* **1991**, 28, 107-113.
57. Yang, F.; Bian, C.; Zhu, L.; Zhao, G.; Huang, Z.; Huang, M. Effect of human serum albumin on drug metabolism: structural evidence of esterase activity of human serum albumin. *J Struct Biol* **2007**, 157, 348-55.
58. Amidon, G.; Lennernäs, H.; Shah, V.; Crison, J. A Theoretical Basis for a Biopharmaceutic Drug Classification: The Correlation of in Vitro Drug Product Dissolution and in Vivo Bioavailability. *Pharmaceutical Research* **1995**, 12, 413-420.
59. Leeson, P. D.; St-Gallay, S. A.; Wenlock, M. C. Impact of ion class and time on oral drug molecular properties. *Med. Chem. Comm.* **2011**, 2, 91-105.
60. Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, 51, 817-34.
61. Quinn, D. M. Acetylcholinesterase: enzyme structure, reaction dynamics, and virtual transition states. *Chemical Reviews* **1987**, 87, 955-979.
62. Dodson, G.; Wlodawer, A. Catalytic triads and their relatives. *Trends in biochemical sciences* **1998**, 23, 347-352.

63. Somoza, J. R.; Skene, R. J.; Katz, B. A.; Mol, C.; Ho, J. D.; Jennings, A. J.; Luong, C.; Arvai, A.; Buggy, J. J.; Chi, E.; Tang, J.; Sang, B.-C.; Verner, E.; Wynands, R.; Leahy, E. M.; Dougan, D. R.; Snell, G.; Navre, M.; Knuth, M. W.; Swanson, R. V.; McRee, D. E.; Tari, L. W. Structural Snapshots of Human HDAC8 Provide Insights into the Class I Histone Deacetylases. *Structure* **2004**, 12, 1325-1334.
64. Colletier, J. P.; Fournier, D.; Greenblatt, H. M.; Stojan, J.; Sussman, J. L.; Zaccai, G.; Silman, I.; Weik, M. *Structural insights into substrate traffic and inhibition in acetylcholinesterase*. 2006; Vol. 25, p 2746-2756.
65. Díaz, N.; Suárez, D.; Sordo, T. L.; Merz, K. M. Molecular Dynamics Study of the IIA Binding Site in Human Serum Albumin: Influence of the Protonation State of Lys195 and Lys199. *Journal of Medicinal Chemistry* **2000**, 44, 250-260.
66. Fujiwara, S.; Amisaki, T. Identification of high affinity fatty acid binding sites on human serum albumin by MM-PBSA method. *Biophys J* **2008**, 94, 95-103.
67. Alvarez, H. A.; McCarthy, A.; N., s.; Grigera, J. R.; #xfa. A Molecular Dynamics Approach to Ligand-Receptor Interaction in the Aspirin-Human Serum Albumin Complex. *Journal of Biophysics* **2012**, 2012, 7.
68. Etienne, T.; Assfeld, X.; Monari, A. QM/MM calculation of absorption spectra of complex systems: The case of human serum albumin. *Computational and Theoretical Chemistry* **2014**, online.
69. van der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, 52, 2708-2728.
70. Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angewandte Chemie-International Edition* **2009**, 48, 1198-1229.
71. Vane, J. R.; Botting, R. M. The mechanism of action of aspirin. *Thrombosis Research* **2003**, 110, 255-258.
72. Tóth, L.; Muszbek, L.; Komáromi, I. Mechanism of the irreversible inhibition of human cyclooxygenase-1 by aspirin as predicted by QM/MM calculations. *Journal of Molecular Graphics and Modelling* **2013**, 40, 99-109.
73. Park, B. K.; Boobis, A.; Clarke, S.; Goldring, C. E. P.; Jones, D.; Kenna, J. G.; Lambert, C.; Lavery, H. G.; Naisbitt, D. J.; Nelson, S.; Nicoll-Griffith, D. A.; Obach, R. S.; Routledge, P.; Smith, D. A.; Tweedie, D. J.; Vermeulen, N.; Williams, D. P.; Wilson, I. D.; Baillie, T. A. Managing the challenge of chemically reactive metabolites in drug development. *Nature Reviews Drug Discovery* **2011**, 10, 292-306.
74. Huttunen, K. M.; Raunio, H.; Rautio, J. Prodrugs—from Serendipity to Rational Design. *Pharmacological Reviews* **2011**, 63, 750-771.
75. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Jr., J. A. M.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V.

G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09, Revision D.01*, Gaussian Inc.: Wallingford CT, 2009.

76. Gilmer, J. F.; Moriarty, L. M.; Lally, M. N.; Clancy, J. M. Isosorbide-based aspirin prodrugs: II. Hydrolysis kinetics of isosorbide diaspirinate. *European Journal of Pharmaceutical Sciences* **2002**, 16, 297-304.

77. Onufriev, A.; Case, D. A.; Ullmann, G. M. A Novel View of pH Titration in Biomolecules†. *Biochemistry* **2001**, 40, 3413-3419.

78. Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. Protein Ionizable Groups: pK Values and Their Contribution to Protein Stability and Solubility. *Journal of Biological Chemistry* **2009**, 284, 13285-13289.

79. Zhou, Y.; Wang, S.; Zhang, Y. Catalytic Reaction Mechanism of Acetylcholinesterase Determined by Born–Oppenheimer Ab Initio QM/MM Molecular Dynamics Simulations. *The Journal of Physical Chemistry B* **2010**, 114, 8817-8825.

80. Thompson, R. A.; Isin, E. M.; Li, Y.; Weaver, R.; Weidolf, L.; Wilson, I.; Claesson, A.; Page, K.; Dolgos, H.; Kenna, J. G. Risk assessment and mitigation strategies for reactive metabolites in drug discovery and development. *Chem Biol Interact* **2011**, 192, 65-71.

81. Bhattacharya, A. A.; Grüne, T.; Curry, S. Crystallographic analysis reveals common modes of binding of medium and long-chain fatty acids to human serum albumin. *Journal of Molecular Biology* **2000**, 303, 721-732.

82. Gerig, J. T.; Katz, K. E.; Reinheimer, J. D. Reactions of 2,6-dinitro-4-trifluoromethylbenzenesulfonate with human serum albumin. *Biochim. Biophys. Acta.* **1978**, 21, 196-209.

83. Watanabe, H.; Tanase, S.; Nakajou, K.; Maruyama, T.; Kragh-Hansen, U.; Otagiri, M. Role of arg-410 and tyr-411 in human serum albumin for ligand binding and esterase-like activity. *Biochem. J.* **2000**, 349, 813-819.

OUTPUT

OUTPUT จากโครงการวิจัยที่ได้รับทุนจาก สกว.

1. ผลงานตีพิมพ์ในวารสารวิชาการนานาชาติ (List of publications):

1. Gleeson, M.P.*; Modi, S.; Bender, A.; Marchese Robinson, R. L.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. The challenges involved in Modeling Toxicity Data in silico: A Review. Curr. Pharm. Des. **2012**, 8, 1226-1291.
 - *Extensive literature review to identify best methods for studying ADMET based endpoints from a computational perspective.*
2. Gleeson, D.; Tehan, B.*; Gleeson, M. P.*; Limtrakul, J. Evaluating the Enthalpic Contribution to Ligand Binding using QM Calculations: Effect of Methodology on Geometries and Interaction Energies. Org. Biomol. Chem. **2012**, 10, 7053-7061.
 - *Validation of the quantum mechanical method on much smaller protein kinase target to determine optimum methods for the subsequent HSA calculations*
3. Gleeson, M. P.*; Montanari, D. Strategies for the generation, validation and application of in silico ADMET models in lead generation and optimization. Expert Opinion Drug Metabol. Toxicol. **2012**, 8, 1435-1446.
 - *Extensive literature review to identify best methods for studying ADMET based endpoints from a computational perspective.*
4. Songatee, N.; Gleeson M. P.*; Choowongkamon, K.* Computational Study of EGFR Inhibition: Molecular Dynamics Studies on the Active and Inactive Protein Conformations. J. Mol. Model., **2013**, 19 (2), 497-509.
 - *Validation of the Molecular dynamics parameters on much smaller protein kinase target to determine optimum methods for the subsequent HSA calculations (i.e. AMBER forcefield, GROMACS program).*

5. Boonyarattanakalin, S.; Ruchirawat, S.; Gleeson, M. P.* Ring opening polymerization of mannosyl tricyclic orthoesters: rationalising the stereo and regioselectivity of glycosidic bond formation using quantum chemical calculations. *Med. Chem. Comm.* **2013**, 4, 265-268.
- *Validation of the quantum mechanical method for a charged-species organic chemistry reaction to determine optimum methods for the subsequent HSA reaction mechanism studies.*
6. Promkatkaew, M.; Gleeson, D.; Hannongbua, S.; Gleeson, M. P.* Skin Sensitization Prediction Using Quantum Chemical Calculations: A Theoretical Model for the SNAr Domain. *Chem. Res. Toxicol.*, **2014**, 27 (1), 51–60.
- *Validation of the quantum mechanical method for an ADMET based endpoint to determine optimum methods for the subsequent HSA reaction mechanism studies.*
7. Pongprayoon, P.; Gleeson, M. P.* Probing the Binding Site Characteristics of HSA: A Combined Molecular Dynamics and Cheminformatics Investigation. *J. Chem. Inf. Model.* **2014**, submitted.
- *Molecular dynamics simulations on HSA, exploring the active site selectivities and the origin of the differing fatty acid affinities using the methodologies identified from previous validation studies.*
8. Phuangsawai, O.; Hannongbua, S.; Gleeson, M. P.* Elucidating the Origin of the Esterase Activity of HSA using QM/MM Calculations. *Biochemistry*, **2014**, submitted.
- *Hybrid QM/MM studies to evaluate the reactivity of HSA using the methodologies identified from previous validation studies.*

2. การนำผลงานวิจัยไปใช้ประโยชน์ (เชิงวิชาการ):

3. อื่น ๆ

(1) การได้รับเชิญไปเป็นวิทยากรให้เสนอผลงานในที่ประชุมวิชาการ ในต่างประเทศ

1. งาน 9th Annual Conference Intelligent Compound Libraries 2013.
ระหว่างวันที่ 28 – 29 October 2013 ประเทศ Germany (เอกสารแนบในภาคผนวก)
2. งาน DSI-RSC Conference: Overcoming the Bottlenecks in Drug Discovery.
ระหว่างวันที่ 19 – 22 March 2014 ประเทศ India (เอกสารแนบในภาคผนวก)

(2) การเชื่อมโยงทางวิชาการกับนักวิชาการอื่น ๆ ทั้งในและต่างประเทศ Collaborations with other researchers (both domestic and international).

In the course of this research grant I have collaborated with the following experts. These collaborations have been undertaken to leverage key skills and resources available to these individuals, which has facilitated more rapid progression of my research project.

1. Sandeep Modi: Unilever, United Kingdom.
2. Andreas Bender: University of Cambridge, United Kingdom.
3. Ben Tehan: Heptares Therapeutics, United Kingdom.
4. Dino Montanari: Lundbeck Pharmaceuticals, Denmark.
5. Supa Hannongbua: Kasetsart University, Thailand.
6. Jumras Limtrakul: Kasetsart University, Thailand.
7. Kiattawee Choowongkamon: Kasetsart University, Thailand.
8. Duangkamol Gleeson: KMITL, Thailand.
9. Somsak Ruchirawat, Chulabhorn Research Institute, Thailand.
10. Poonsakdi Ploypradith: Chulabhorn Research Institute, Thailand.
11. Siwarutt Boonyarattanakalin, SIIT, Thailand

(3) การเชื่อมโยงกับต่างประเทศหรือรางวัลที่ได้รับ (Collaboration-international only, Awards, Prize, etc.)

1. Sandeep Modi: **Unilever**, United Kingdom.

- Please see the publication: Gleeson et al. *Curr. Pharm. Des.* 2012 (appendix) for the topic and scope of the collaboration.

2. Andreas Bender: **University of Cambridge**, United Kingdom.

- Please see the publication: Gleeson et al. *Curr. Pharm. Des.* 2012 (appendix) for the topic and scope of the collaboration.

3. Ben Tehan: **Heptares Therapeutics Ltd** (Pharmaceuticals), United Kingdom.

- Please see the publication: Gleeson et al. *Org. Biomol. Chem.* 2012 (appendix) for the topic and scope of the collaboration.

4. Dino Montanari: **Lundbeck Pharmaceuticals**, Denmark.

- Please see the publication: Gleeson et al. *Expert Opinion Drug Metabol. Toxicol.* 2012, for the topic and scope of the collaboration.

ภาคผนวก

Appendix (Reprint, manuscript, awards, collaboration)

Reprints :

1. Gleeson, M.P.*; Modi, S.; Bender, A.; Marchese Robinson, R. L.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. The challenges involved in Modeling Toxicity Data in silico: A Review. *Curr. Pharm. Des.* **2012**, 8, 1226-1291.
2. Gleeson, D.; Tehan, B.*; Gleeson, M. P.*; Limtrakul, J. Evaluating the Enthalpic Contribution to Ligand Binding using QM Calculations: Effect of Methodology on Geometries and Interaction Energies. *Org. Biomol. Chem.* **2012**, 10, 7053-7061.
3. Gleeson, M. P.*; Montanari, D. Strategies for the generation, validation and application of in silico ADMET models in lead generation and optimization. *Expert Opinion Drug Metabol. Toxicol.* **2012**, 8, 1435-1446.
4. Songatee, N.; Gleeson M. P.*; Choowongkamon, K.* Computational Study of EGFR Inhibition: Molecular Dynamics Studies on the Active and Inactive Protein Conformations. *J. Mol. Model.*, **2013**, 19 (2), 497-509.
5. Boonyarattanakalin, S.; Ruchirawat, S.; Gleeson, M. P.* Ring opening polymerization of mannosyl tricyclic orthoesters: rationalising the stereo and regioselectivity of glycosidic bond formation using quantum chemical calculations. *Med. Chem. Comm.* **2013**, 4, 265-268.
6. Promkatkaew, M.; Gleeson, D.; Hannongbua, S.; Gleeson, M. P.* Skin Sensitization Prediction Using Quantum Chemical Calculations: A Theoretical Model for the SNAr Domain. *Chem. Res. Toxicol.*, **2014**, 27 (1), 51–60.

Manuscripts:

1. Pongprayoon, P.; Gleeson, M. P.* Probing the Binding Site Characteristics of HSA: A Combined Molecular Dynamics and Cheminformatics Investigation. *J. Chem. Inf. Model.* **2014**, submitted (*see results section*)
2. Phuangswai, O.; Hannongbua, S.; Gleeson, M. P.* Elucidating the Origin of the Esterase Activity of HSA using QM/MM Calculations. *Biochemistry*, **2014**, submitted. (*see results section*)

Conferences Letters and Schedule:

1. Invitation letters and Conference Schedule :
 - a. IQPC Conference Berlin, Germany
 - b. RSC-DSI Conference, Gurgaon, India

The Challenges Involved in Modeling Toxicity Data *In Silico*: A Review

M. Paul Gleeson^{1*}, Sandeep Modi², Andreas Bender³, Richard L. Marchese Robinson³, Johannes Kirchmair³, Malinee Promkatkaew¹, Supa Hannongbua¹ and Robert C. Glen³

¹Department of Chemistry, Faculty of Science, Kasetsart University, 50 Phaholyothin Rd, Chatuchak, Bangkok 10900, Thailand;

²Unilever, Colworth, Sharnbrook, Bedford, MK44 1LQ, United Kingdom; ³Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Abstract: The percentage of failures in late pharmaceutical development due to toxicity has increased dramatically over the last decade or so, resulting in increased demand for new methods to rapidly and reliably predict the toxicity of compounds. In this review we discuss the challenges involved in both the building of *in silico* models on toxicology endpoints and their practical use in decision making. In particular, we will reflect upon the predictive strength of a number of different *in silico* models for a range of different endpoints, different approaches used to generate the models or rules, and limitations of the methods and the data used in model generation. Given that there exists no unique definition of a 'good' model, we will furthermore highlight the need to balance model complexity/interpretability with predictability, particularly in light of OECD/REACH guidelines. Special emphasis is put on the data and methods used to generate the *in silico* toxicology models, and their strengths and weaknesses are discussed. Switching to the applied side, we next review a number of toxicity endpoints, discussing the methods available to predict them and their general level of predictability (which very much depends on the endpoint considered). We conclude that, while *in silico* toxicology is a valuable tool to drug discovery scientists, much still needs to be done to, firstly, understand more completely the biological mechanisms for toxicity and, secondly, to generate more rapid *in vitro* models to screen compounds. With this biological understanding, and additional data available, our ability to generate more predictive *in silico* models should significantly improve in the future.

Keywords: Toxicity, QSAR, *in silico* modeling, descriptors.

1.0. INTRODUCTION

The use of chemicals which display significant levels of toxicity is difficult to countenance in either the pharmaceutical or consumer products industry for ethical reasons, and also due to legal responsibility and commitments of the manufacturer to ensure products are safe. Significant effort is therefore spent assessing the toxic liabilities of large numbers of chemicals for a variety of different endpoints. In the consumer products industry and in late drug development, *in vivo* and *in vitro* methods are commonly used to determine the potential for genotoxicity, carcinogenicity or skin sensitization, for example, while in early drug discovery toxicity endpoints associated with inhibition of the human ether-a-go-go-related gene ion-channel (hERG) or cytochrome P450 (abbreviated P450 henceforth), as well as metabolite formation assessment, are more prevalent.

The percentage of failures in late pharmaceutical development due to toxicity has increased dramatically over the last decade. From reports by Kola *et al.* [1] it was revealed that the number of failures due to absorption, distribution, metabolism and excretion (ADME) problems were 40% in 1991 while those due to combined safety/toxicity issues were approximately 20%. However, following significant investments in *in vivo*, *in vitro* and *in silico* ADME infrastructure [1-3], the proportion failures due to ADME in 2000 dropped to around 10%, mainly due to better ADME screening at early stages during last decade. Failures due to safety/toxicity reasons however, increased to around 30% [1]. This noticeable increase might be a reflection of the strengthened regulatory oversight in late development, or a reflection of the lack of investment in toxicity based testing in early drug discovery, at least relative to ADME. Whatever the precise cause, it is undeniable that there is an increased emphasis in early drug discovery to minimize toxicity failures, in particular in later, more expensive clinical trials [4].

In an effort to minimize the number of candidate drug failures due to off-target toxicity (as opposed to toxicity associated with the primary pharmacology of the target) [5], a variety of *in vivo* and *in vitro* models have been developed to screen development compounds [6, 7]. For example, the mutagenic potential of a compound can be assessed using the Ames test [8], skin sensitization risk assessment can be performed by use of the Organisation for Economic Co-operation and Development (OECD) Mouse Local Lymph Node Assay (LLNA) test [9-11] and the potential to cause cardiac arrhythmia/Torsades de pointes can be evaluated using the perfused rabbit heart assay, patch clamp or hERG-dofetilide binding assays [12].

As a result of the continuous experimental assessment of compounds in such assays, large databases of measurements exist within many pharmaceutical companies, and also in the public domain to a lesser degree. The *in vitro* hERG inhibition assay used at Pfizer had screened almost 60,000 compounds by 2005 [13] while the *in vitro* P450 2D6 inhibition assay used at GSK was reported to have assessed at least 50,000 compounds in 2008 [14]. In contrast, publicly available databases for toxicity are much smaller, with no more than 8000 measurements recently reported in the *in vitro* Ames assay in 2011 [15], just 500 for the *in vivo* LLNA measure of sensitization [16], while many other datasets contain much fewer measurements [17].

The existence of sizeable, diverse databases of measurements has opened up the possibility of generating so called '*in silico*' models of toxicity. *In silico* models involve the generation of a prediction of a given toxicity response using computers alone. Since the toxicity response is due to the molecular structure of the compounds under investigation, we can try to generate predictive models to assess the likely response of compounds not yet tested. The prediction of the toxic potential from molecular structure alone is highly desirable due to: (a) the high speed at which a prediction can be generated, (b) the reduced costs associated with the method and (c) ethical considerations due to the lack of animals needed for the assessment to be made. For an *in silico* prediction to be made on an untested molecule, we would typically build a statistical, rule-

*Address correspondence to this author at the Department of Chemistry, Faculty of Science, Kasetsart University, 50 Phaholyothin Rd, Chatuchak, Bangkok 10900, Thailand; Tel: +662-562-5555; Ext: 2210; Fax: +662-579-3955; E-mail: paul.gleeson@ku.ac.th

based or atomistic model using the available experimental data. These models must be built using some representation of a molecule's physical properties and its chemical characteristics, which is then related to their toxicological response. *In silico* models can in principle be built using 1D representations of molecules including descriptors such as molecular weight or atom counts, 2D descriptors, which incorporate information about a molecule's topological structure, such as predicted lipophilicity, connectivity indices or fingerprints, or 3D descriptions such as pharmacophoric points, 3D surfaces or grids (discussed in detail below). The relationship between structure and response must then be fitted (or trained) using the available dataset which can be achieved using relatively simple methods such as "read-across" or multiple linear regression to more complex non-linear machine learning methods such as non-linear Support Vector Machines (SVM) or Artificial Neural Networks (ANNs). The precise choice of descriptors for a given model is generally dictated by: (1) the physical process that underlies the toxicity event, (2) regulatory considerations such as OECD QSAR guidelines [18] which prefer mechanistically interpretable models and (3) the personal preference of the model developer. For example, drug-induced phospholipidosis, the potentially toxic excessive accumulation of phospholipids in cells/tissues, may be described with simple descriptors, such as the presence of a positive charge/basic substituent and high lipophilicity [19] (although more sophisticated approaches have also been investigated [20]). In contrast, P450 inhibition is clearly a complex receptor-mediated process, which arguably requires a complex chemical description such as that afforded by receptor docking, pharmacophores [21, 22] or multiple descriptors in conjunction with complex non-linear modeling methods [23] (and which is further complicated by the flexibility and promiscuity of the proteins involved). In an effort to conform with the OECD QSAR regulatory guidelines [18], *in silico* methods to predict skin sensitization, which is believed to be a toxic response due to covalent modification of unknown proteins in the epidermis, rely on mechanistically interpretable models, at least in terms of their encoding of chemical reactivity. This means that models based on relatively simple chemical descriptors are generally preferred (such as the presence or absence of certain fragment indicator variables) [24-26].

2.0. IN SILICO TOXICOLOGY MODELING

In 2003, the European Commission adopted a legislative proposal for creation of a new chemical management system called REACH (registration, evaluation and authorization of chemicals), and one of the key goals was to reduce the number of compounds that require toxicity testing in animals through the use of alternative methods [27, 28]. A key consideration of REACH was that methods such as *in silico* based quantitative structure activity relationships (QSARs) should be used where possible to avoid the use of animal testing. For a QSAR model to be appropriate in a regulatory environment, OECD guidelines were developed, suggesting the following be used [18]:

- A defined endpoint.
- An unambiguous algorithm.
- A defined domain of applicability.
- Appropriate measures of goodness-of-fit, robustness and predictivity.
- A mechanistic interpretation, if possible.

The carefully considered, pragmatic guidelines proposed by the OECD should be followed in the generation of all *in silico* models to ensure that any subsequent model(s) generated will be reliable in use, and acceptable to regulatory agencies.

In this review, we discuss published QSAR and modeling studies in the toxicity field, making references to the above guidelines. We first review the general characteristics of toxicity datasets that are used to generate the QSAR models. Subsequently, we summa-

rize chemical descriptors that can be used to underpin an *in silico* model. This is followed by a review of the basic statistical techniques used to determine the relationship between the experimental toxicological response and computed chemical descriptors. Finally, we discuss some specific *in silico* toxicology models that have been reported in the literature and how they can be used in a screening setting. As we cannot possibly cover every area of the *in silico* toxicology field in any great detail, nor cover every possible *in silico* toxicology model for those areas that we do cover, we refer readers to Table 1 where a list of either very notable papers or comprehensive reviews are listed.

Table 1. A List of Comprehensive Reviews or Notable Papers in the *In Silico* Toxicology Field

Toxicity Endpoint	References
General Reviews	[29-40]
Genotoxicity/Carcinogenicity	[25, 26, 41-46]
Skin sensitization	[24, 47-50]
Cardiotoxicity	[51-54]
P450 inhibition/Drug-drug interactions	[21, 55-61]
Metabolite prediction	[61-68]
Reproductive/Developmental toxicity	[69-71]

It should be noted that the *in silico* toxicology models discussed here are hazard identification methods, and in the majority of cases they do not take consider the dose and exposure into account. Therefore, these methods will not be capable of predicting the absolute toxicity in isolation, but provide useful supplementary information for the overall risk assessment process.

2.1. Toxicity Data

The major challenge in toxicity modeling is that all molecules are toxic at some level. It is therefore necessary that an *in silico* method can predict both the type of toxicity, as well as the level of toxicity of a compound. Since there are numerous different ways in which toxicity (related to the primary pharmacology or many secondary pathways) can arise, the prediction of the absolute toxic potential of a compound, either from animal models, or *via in vitro* or *in silico* methods, is extremely challenging [5]. In addition, the level of toxicity for a given toxicity mechanism may also depend on the level and distribution of drug in the body (ADME processes), making accurate toxicity predictions even more difficult [51, 72-75].

Experimental toxicity screening in drug discovery is a multi-tiered approach, with *in silico* models towards the front, followed by chemical and *in vitro* biological assays, and finally, *in vivo* animal assays. Assays deployed towards the rear are generally considered to be the most predictive of man, but also the most labor and cost intensive [72]. As a result of the REACH guidelines on animal testing, there is now increased pressure for methods towards the front, where appropriate, to be used more extensively as alternatives to *in vivo* assays.

In silico techniques would appear to be the best choice to screen compounds with unknown toxic potential due to the rapid predictions given, and the limited physical resources needed. However, *in silico* modeling of toxicity data is extremely challenging due to the complexity, and often considerable variability of the data itself. In addition, there are still question marks over the ability of *in vitro* data to predict the likely toxic response of a chemical in man. Animal tests have been known to show poor concordance with results

in man on occasions [76], which would explain why many approved drugs have subsequently been withdrawn from the market due to previously undetected toxicity issues [77].

Improved understanding of the different mechanisms by which chemicals affect biological structures, processes and pathways, and thus the physiological response, is critical aspect of toxicology. Increased knowledge will help in predicting the toxicity of chemicals, and to implement strategies to reduce the overall exposure to toxic compounds, and develop ways to minimize their affect. However, both the complexity in the biological response and the lack of publically available mechanistic data are key reasons why *in silico* approaches to date have had limited success in delivering *in vivo* relevant predictions.

2.1.1. Implications of Experimental Errors for Modeling

All data contain errors and it is important to understand the frequency or magnitude of the errors involved in a particular toxicity dataset before beginning any analyses. This is particularly important in the case of *in silico* modeling since we need to understand how accurate or noisy a dataset is as this will dictate the type of model to be built. It can also give us a rough idea of how good (or bad) a model we could theoretically produce with such a dataset. To highlight this point, we consider three toxicity endpoints.

- The *in vitro* inhibition and metabolite assays used to assess for drug-drug interactions (DDI) are not totally predictive of a molecule's effect in man [72, 73, 76]. Complicating factors include known inter-species and inter-patient differences due to P450 polymorphism [78, 79]. Researchers have also noted that routine P450 inhibition assays commonly employed cannot easily discriminate competitive/non competitive inhibitor behavior, compounds that are substrates or inducers, the precise binding site involved in inhibition, as well as time-dependent effects [41]. Experimental errors for such assays are expected to be ~2 fold, although ketoconazole was reported to have up to 17-fold variation in a 2D6 inhibition assay [80].
- The Ames test [8] for mutagenicity has experimental errors that are not negligible [81] and an experimental error rate of ~15% between laboratories has been quoted [82]. The method's ability to predict rodent carcinogenicity is good at between 77-90% [8]. However, while predicted Ames test mutagens have a high probability of being carcinogens, compounds that are predicted to be non-mutagens still have an equal probability of being either entirely non-carcinogenic, or potentially carcinogenic via a different mechanism (*i.e.*, non-genotoxic carcinogens) [43, 83].
- The gold standard assay in skin sensitization assessment, the *in vivo* LLNA assay [84], is reported to display a relatively low error rate of ~10% [85]. This quantitative assay is considered more reliable than the earlier Guinea pig assay where the toxic potential was scored based on visual analysis [85]. This, along with species differences, would explain the considerable discrepancies between the two assays as noted by Lalko *et al.* [48].

A consideration of the experimental variability of an assay, and its relevance in predicting the human toxicological endpoint, is therefore desirable before beginning any *in silico* modeling. It also helps us to understand: (a) what level of predictive ability is theoretically possible for a given dataset and (b) determine what level of predictive ability from a model would be deemed acceptable. *In silico* models cannot predict the experimental result with absolute accuracy, since the experimental results themselves have associated errors, so an understanding of this variability is important in setting our expectations. It is also unlikely that an *in silico* model will predict the *in vivo/in vitro* assay with an error rate equivalent to its experimental error (a perfect model) since there are other errors that are introduced during the descriptor calculations and model fitting

procedures. An apparently perfectly fitted model would most likely have been fitted to some of the experimental variation also, *i.e.* it would actually be overfitted in practice.

2.1.2. Compound Screening and Curation Issues

An important factor affecting data generated *via in vitro* assays is whether the compound being tested is soluble enough to be assayed, or whether storage has resulted in compound degradation [7, 86, 87]. As an example, following the merger of SmithKline-Beecham and GlaxoWellcome, a large scale analytical exercise to assess the company's combined screening collection showed that 80% of the entries were considered both pure and confirmed as having the correct structure based on mass spectrometry - which in turn means that 1/5 of the compounds did not fulfill those requirements [88].

Another issue can arise as a result of data archiving. Transcription errors are known to arise when data reported in the public domain are curated and added into publically accessible databases. Young *et al.*, for example, highlighted data translation errors ranging from 0.1 to 3.4% depending on the database in question [89]. In addition, Fourches *et al.* have shown that these transcription errors can complicate the generation of quantitative structure activity relationship (QSAR) models [90].

2.1.3. Importance of Data Treatment

Data treatment is an important aspect of modeling toxicity data. Based on a consideration of the data error, a modeler may decide to build a model on either the continuous output from a given toxicity assay (LD50, IC50, EC3 etc), or a categorical measure (*e.g.* a two class system distinguishing toxic and non-toxic predictions). Defining a class based output for data coming from an assay reporting a continuous output, in effect, discards useful data. However, it can be advantageous in modeling studies as categorization, in effect, reduces the complexity of fitting a QSAR relationship and in practice this often leads to models with a better performance (*i.e.* a model only has to determine how the features separate the classes, not a continuum).

Categorization of data can also be advantageous in modeling studies since data from related assay systems can be more easily combined. This, of course, requires that the errors introduced due to assay differences are more than compensated for by building a model on a larger, more diverse dataset. Examples of this approach include modeling studies by Schultz *et al.* who used both LLNA and glutathione binding data to build models for skin sensitization [91]. A similar approach was taken in the TIMES-SS program where three different sources of *in vivo* data, derived from multiple animals, were used in modeling skin sensitization [92].

2.2. Determining the Relationship between Chemical Structure and Toxicity

Accurate and rapid *in silico* models of toxicological endpoints are the most ideal solution to the proposals for animal testing. *In silico* models/rules are quantitative or qualitative relationships between descriptors encoding molecular structure and a response variable encoding bioactivity and can take a variety of forms (Table 2).

Before generating a model on a particular toxicological endpoint, some key questions must be considered to assess the viability of model building and setting our expectations as to the likely predictive ability, namely:

- Is the *in vitro* or *in vivo* assay sufficiently predictive of toxic events in man? If it is weakly predictive then it is questionable how useful an *in silico* model generated for this endpoint will be.
- Is the data sufficiently reproducible to allow for model building? If the assay variability is large over successive experiments, or between laboratories, then any *in silico* model will not perform better than the reproducibility of the data.

Table 2. A list of Common Techniques Employed in In Silico Toxicology Studies

Method	Advantages	Disadvantages
Similarity (Read across)	- A variety of methods can be used to determine the molecular similarity between pairs of molecules (i.e. untested molecules with those that have been experimentally tested). A query compound is expected to have similar activity to those that display high similarity to it in the measured set. This can be done using 1D, 2D or 3D descriptors. - Advantages are simplicity, speed, and the ability to directly compare the compound at hand with the reference compound for interpretation.	The main premise in computational chemistry is that similar molecules act similarly. It is known, however, that this does not hold true in many circumstances, with molecules that differ only very subtly showing dramatically different biological activities.
Substructure Alerts/ Reaction schemes	- Computational methods can be used to assess for the presence of a particular functional group/molecular scaffolds which are known to form specific metabolites. - Computational enumeration can be used to generate all possible combinations of metabolites. The toxic potential of each metabolite can then be assessed. - The main advantage of this method is its interpretability, which is immediately apparent to a chemist.	- Software packages such as Meteor [99] can predict large numbers of metabolites for a particular molecule based on expert curation of the available literature data. - As with SAR data, the lack of contextual molecular information in such rule bases can mean that the majority of predictions are not actually relevant [100]. - While these method might be acceptable when looking qualitatively at individual compounds on a case by case basis, they are less useful when trying to assess very large sets of molecules in an automated fashion.
SAR	A qualitative observation that certain molecular characteristics (including substructures) can give rise to a certain toxicological outcomes. This SAR can be molecular property based or substituent/fragment based.	- Qualitative rules can be problematic as they generally lack contextual molecular information. For example the presence of a liability such as functional group may be acceptable if found in combination with some other feature [100]. - Absolute estimation of risk is difficult or even impossible
QSAR	- A quantitative statistical relationship between a compound's structure and a toxicological response fitted, prior to the assessment of an unknown molecule, using a training set of compounds with known toxicity. - QSARs can be used to predict an absolute value for a given toxicity endpoint, a class, or probability of class membership.	- QSAR models have received much criticism of late, although much of this is due to inappropriate model use and/or modeling building errors. - While undeniably useful, such models are unable to reliably extrapolate to chemical structures which are significantly different to those used to generate the model and, in addition, care must be taken to avoid capturing non-generalisable patterns in the underlying data. - A general rule is that QSAR models need validation for each molecular series that is to be predicted and this should be continuously assessed over time.
3D Receptor models	- Many toxicity events require a degree of recognition between a molecule and a protein. This suggests that a consideration of the 3D shape of a molecule, and the 3D structure of the protein could be beneficial. - MetaSite [101] is an example of 3D-based tool for predicting the likely site of metabolism of a molecule. - Hybrid quantum mechanical/molecular mechanical (QM/MM) models have also been used to assess the intrinsic reactivity of molecules with particular 3D protein conformations [98]. - The main advantage of these methods is their ability to cover different chemical series, since an arrangement in 3D protein space potentially more relevant to the toxic mechanism is used to derive the model/prediction (instead of non-context specific chemical descriptors/features).	- The application of 3D methods to understand toxicity is generally done in a retrospective way at present. - These calculations are very time-consuming compared to other methods, and require extensive expertise. - In some cases, no X-ray structures are available for the protein in question (or a protein-ligand complex for the series in question). Therefore, one needs to use molecular docking and/or construct less reliable homology models - The introduction of shape, while increasing the information content within the model, also introduces additional noise (e.g. due to conformation sampling), and whether the final model will be superior to simpler models depends on the signal gained (the ratio of information to noise introduced into the model when considering 3D information).

- Has a diverse set of compounds been screened in the assay? If the dataset screened in the assay is small, or unrepresentative, then our ability to make predictions for a diverse set of untested compounds will be questionable.
- Is the toxic event/mechanism understood? If the toxic event is due to multiple processes (e.g. tissue permeation, followed by protein binding, then reaction), each with different dependencies on a compound's molecular structure, then obtaining an accurate relationship between structure and measured toxicity is much more challenging.
- Is it apparent what type of descriptors/*in silico* method should be used to model the toxicity parameter in question? If the process requires binding to a particular protein to achieve the toxic event then this level of specificity should in principle require a 3D model of the interaction/event [61] or a quantum chemical consideration of the reaction [93]. Nevertheless, 3D methods are still not sufficiently mature to be used in a black box manner [94, 95], meaning more simplistic 2D and 1D methods are often preferred.
- Is it known what the relationship between structure and function looks like - i.e., whether there are multiple binding modes, which might benefit from non-linear modeling methods? The modeling algorithm linking molecular descriptors to the output variable needs to be chosen so that it takes the complexity of the particular relationship into account, otherwise overfitting (in case of too complex a modeling procedure used) or an insufficiently predictive model (in case of too simple a modeling procedure used) may result.

A wide range of different computational methods have been used to build *in silico* models of toxicological endpoints and these are summarised briefly in Table 2. The methods range from rather simple similarity measures to known toxic/non-toxic molecules, the presence of undesirable substituents in a molecule, or simple SAR, to more complex multivariate QSAR models or, at the other extreme, atomistic simulations of the toxic events. In this review, we try to cover the theory used in a variety of these methods and discuss a number of different applications to specific toxicological endpoints.

The computational resources needed to generate a prediction of toxicity can range significantly for the methods listed in Table 1. Methods relying on functional group identification, or similarity, are highly automated and can take seconds to perform. QSAR predictions using conventional 2D descriptors are also rapid, however, comparative molecular field analysis (COMFA) based 3D QSAR [96] are much more time consuming due to the alignment step needed. Quantum mechanical based descriptors can be very time consuming (minutes to hours); however this is very dependent on the level of theory used. Semi-empirical calculations are generally rapid, while more accurate density functional theory calculations are considerably slower [97]. Protein-ligand simulations are generally the most time consuming due to the number of atoms involved. Molecular dynamics simulations (MD) can take days to months depending on the amount of sampling needed to reach convergence [97]. Hybrid quantum mechanical/molecular mechanical (QM/MM) methods, which are commonly used to study reactions within proteins [98], can take similar amounts of time as MD simulations.

2.2.1. Encoding Chemical Information in *in silico* Models

A fundamental part of any data driven, predictive method is how we choose to encode the chemical information of the toxic and non-toxic molecules in question. If we cannot accurately describe a molecule using some type of relevant, computationally-based descriptor(s), then it will not be possible to generate a useful model. Molecular descriptors can be rapidly calculated from a known molecular structure using a wide variety of methods reported in the literature [102-112]. These are commonly grouped into three distinct categories: 1D, 2D and 3D descriptors. 1D molecular descrip-

tors (sometimes also referred to as '0D' descriptors) represent certain bulk properties of compounds, such as the number of atoms of a particular element, the molecular weight etc, and these can usually be calculated solely based on the molecular formula of a molecule (i.e. do not necessarily require a knowledge of the precise connectivity of the molecule). The so-called 2D molecular descriptors are computed from a representation of a molecule which takes into account the precise connectivity. Examples include the presence of particular functional groups [26], or path-based descriptors/indices [113] and fingerprints [113-115]. Finally, 3D molecular descriptors capture structural information based on the three-dimensional structure of a molecule (or set of conformations, sometimes referred to as 4D descriptors [116]). These descriptors encompass those such as 3D polar surface area [117], pharmacophore-based [97] approaches or descriptors derived from docking into a receptor [118, 119]. It could be argued that 3D molecular descriptors might perform better than 2D variants in modeling molecule-target binding; however, it has been shown that combinations of 1D and 2D descriptors are often more successful [120]. This appears to be due to the number of approximations involved in the former methods and these are discussed in later sections of the paper.

The toxicity of a compound can arise for different reasons, and the choice of descriptors for a modeling study very much depends on the particular type of toxicity under investigation. For example, toxicity due to the physicochemical properties of a compound may be caused by its ability to interact with cell membranes, such as binding to [19], or increasing the intrinsic permeability through, the cell itself [121]. In many cases, the amphiphilic character of a compound is the cause of the issue; hence, descriptors capturing the lipophilicity and charge of a compound will be needed. In addition, physicochemical properties such as lipophilicity increase the likelihood of a compound being metabolized. This has the potential to lead to toxicity due to drug-drug interactions, caused by the blocking of pathways of metabolizing enzymes, such as those of the P450 family.

The binding of a molecule to a protein target is another potential cause of toxicity. The toxic event may be a side effect of binding to the therapeutically relevant target, a result of binding to a related target with high affinity, or due to inhibition of other unrelated targets. In these cases, descriptors which can be used for virtual screening can also be applied [122-126]. Recently, a large-scale analysis of the relationships between adverse effects of drug molecules and their interactions with protein targets has been published [127].

Toxicity due to the chemical reactivity of particular functional groups found in a molecule is another cause of toxicity. These types of toxic events are defined by particular structural groups of 'reactive functional groups' for which multiple filters have been published [128]. A key limitation of these methods however is that they lack molecular context. For example, certain undesirable substituents will not cause problems due to the presence of additional features in a molecule. A recent matched molecular pairs analysis shows that the effect of certain undesirable substructures can be modulated by other features found in the molecule [129].

2.2.2. Knowledge-Based Rules and Alerts

A wide variety of knowledge-based rules or systems have been generated for toxicity endpoints based on the observation that certain types of molecules frequently display similar toxicological outcomes. These rules are generally derived based on an expert analysis of toxicity data and chemical structures, rather than by the use statistical models and computed descriptors (i.e. data mining/machine learning). In this review, we have separated substructure-based and similarity-based methods from those that employ QSAR. It is worth noting that software packages such as MultiCASE [17] and Toxtree [130, 131] employ hybrid systems that

make QSAR predictions on subsets of compounds that display a particular substructure.

Structural Alerts

A number of commercial and freely available tools are available to virtually screen for structural features that can give rise to toxicity. An example of the former is the DEREK program (Deductive Estimation of Risk from Existing Knowledge) [132, 133]. DEREK uses a set of structural rules derived from experts in industry, academia, and government to give a qualitative assessment of the probability a compound will display toxicity for a variety of endpoints. In many studies that have evaluated the performance of DEREK, the absence of an alert has been associated with a negative (toxicity free) prediction. However, this is not how DEREK was designed to be used [64]. The absence of any alert does not mean the molecule will not display toxicity; rather, an alert for a toxic substructure may not have been compiled yet. Conversely, the presence of an alert does not necessarily mean that a molecule will display toxicity either, since other properties determined by its overall structure are not considered in this type of analysis (such as distribution of the compound). An alternative to DEREK is the freely available Toxtree program [134]. This program can predict the carcinogenicity and mutagenicity potential of molecules as well as other endpoints listed in Table 3.

Table 3. Alerts Implemented in Toxtree [134]

Cramer decision tree approach to estimate toxic hazard [135]
Verhaar scheme for modes of toxic action [136]
German Federal Institute for Risk Assessment (BfR) and skin irritation corrosion rules estimation tool (SICRET) rules to predict skin irritation and corrosion [137]
BfR rules to predict eye irritation and corrosion [138]
Benigni-Bossa rules to predict mutagenicity and carcinogenicity [134]
Cytochrome P450 mediated drug metabolism alerts [139]
Skin sensitisation alerts [24]
Alerts for Michael Acceptors [140]

Similarity Searching / Read-across

Similarity searching is a popular tool in virtual screening [114, 115, 125]. The method involves finding molecules that are similar, in some chemical manner, to a query molecule that is of biological interest. The similarity can be computed using fingerprints, pharmacophores or shape based descriptors and the most similar molecules are generally considered to be more likely to display similar biological characteristics to the query. However, even molecules that display very high similarity do not necessarily share similar levels of activity. This is because it is well known that very small differences in structure can result in dramatic change in response for a particular process, or activity at a particular receptor (*i.e.* activity cliffs) [100, 125].

Similarity methods are commonly framed in terms of molecular fingerprints. These fingerprints consist of a bit string that defines a molecule in terms of the fragments, atom paths or pharmacophoric points present in the molecule. Common fingerprint implementations include: Daylight fingerprints [141], MDL keys/fragments [142] and 2D/3D pharmacophoric fingerprints [113, 125]. The relative performance of these methods have been compared recently by Bender *et al* [120]. Similarity between molecules can be assessed using the Tanimoto or Tversky metrics or distance measures (such as the Euclidean distance) [97].

Similarity concepts can also be used to assess the toxic potential of molecules and the term “read-across” is commonly used to describe the process in the *in silico* toxicology field. In this case, similarity to the toxic query molecule is considered undesirable, unlike in traditional virtual screening applications. Similarity-based methods have been implemented in the freely available Toxmatch program described in detail by Gallegos-Saliner *et al.* [131].

Of course, in some cases the overall similarity of two molecules may not be a relevant consideration, but rather the presence of particular functional groups might be defining factor. Therefore, whether the overall similarity concept can be applied to toxicity prediction needs to be considered in every individual case (both in terms of the toxicity endpoint and the molecular series in question).

2.2.3. QSAR Modeling Methods

A wide variety of methods may be employed to derive QSAR models of toxicological endpoints. We present a brief summary of those techniques which have found widespread applicability in the QSAR literature. We shall discuss some of their key strengths and weaknesses and, in particular, we consider the relative merits of linear and non-linear modeling approaches. The following discussion of QSAR is by no means exhaustive, so readers are referred elsewhere for greater detail on this topic [23, 90, 122, 143-152].

An important initial consideration in any *in silico* modeling investigation is which of many different statistical methodologies will be used to determine a QSAR between the structures in a dataset, which are encoded using chemical descriptors, and their toxicological response. QSAR methods can be broadly split into two basic types, namely (i) linear models, which assume the relationship between the descriptors and response value is linear (which is not always true) and (ii) non-linear methods, which make no such assumption about this relationship.

Non-linear methods would appear to be the most sensible option in most circumstances; however, the fact that fewer assumptions are made about the relationship between structure and toxicity means that many possible relationships between the descriptors and response can be explored. While this situation is more likely to find the correct relationship, it is also potentially more capable of fitting a model to the idiosyncracies of a particular dataset, especially if the dataset is rather small and lacking in chemical diversity. In addition, the complexity of non-linear statistical methods means the resulting models are difficult, if not impossible, to interpret (as discussed in later sections).

2.2.3.1. QSAR Model Building and Validation

QSAR models or rules are derived using pre-existing experimental data. To allow for the generation and validation of the model in question, this data needs to be split into a training set, containing the majority of the data, and a smaller external test/validation set (~10-25%). The former set (which may itself be split into a training and test set for internal validation purposes, see Fig. (1)) is used to build the *in silico* model, while the latter set is used to determine its predictivity, since this data was not used in the building process. The splitting of a dataset can be done either in a random fashion or using a variety of selection techniques [153] (Fig. (1)).

The size and nature of the dataset will dictate whether the model can be classed as a local or global model. Global models are models built on very large, diverse datasets and these models cover a larger area of chemical space and therefore have a larger coverage or domain of applicability [154-158]. This essentially means that the model is more likely to maintain its performance when used to predict more diverse compounds. Local models, on the other hand, are generated on relatively small, often congeneric series of compounds and are generally only used to predict related molecules such as other members of that series. One cannot reliably describe the chemical space beyond that of the rather limited dataset used to

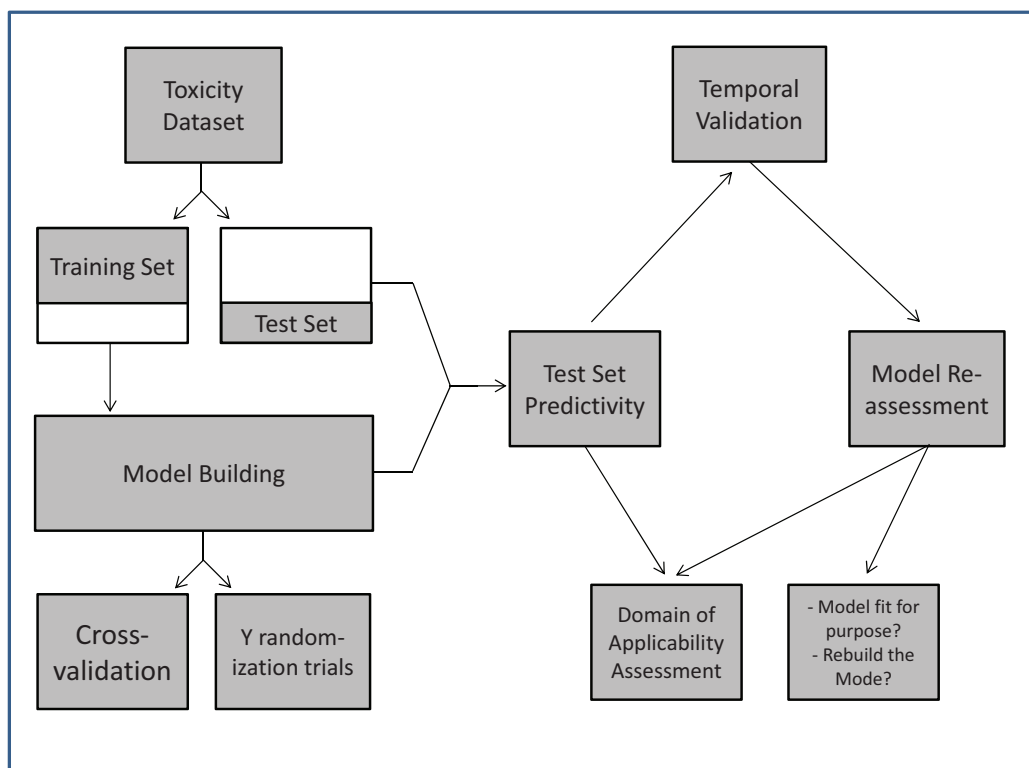


Fig. (1). A graphical representation of a QSAR model building strategy.

build the model [159]. Local models are extensively used in toxicity models implemented in MultiCASE [17] and Toxmatch [131], but only for compounds identified as having particular toxic functional groups (i.e. compounds within their domain of applicability).

The potential always exists for statistical methods to overfit (i.e. to fit both the real variation and experimental error/idiosyncracies in a dataset) so it is critical that predictivity of the model is assessed in a rigorous way. Ideally, external validation, on data neither used to train the model, nor select the descriptors or model parameters, should be used [158, 160, 161]. Benigni *et al.* argue that internal cross-validation (i.e. validation that assesses the effect of building models with different subsets of the training set, such as leave one out (LOO) or leave many out (LMO)), where the validation data is used for model selection purposes, is often of little value, particularly with very small datasets [42, 149], and that external validation is a better measure of true model predictivity [146, 162, 163]. Ideally, external validation should be carried out using a large, independent test set. However, cross-validation may yield unbiased estimations of model performance when it has not been used for model selection [160]. Another useful test to assess the reliability of the model is Y-randomization [20, 164]. In this process, the Y values of the compounds are randomized many times within the training set and random models are built. If the performance of the random models approaches that of the true model, then the model must be considered unreliable.

Given that an independent test set is often a random subset of the data set, or a hypothetically optimal representation of the data set if selection methods are used, then the test set might be expected to perform as well in prediction as the training set (at least for large datasets). This means that temporal validation may be needed to get a better estimate of the true predictivity of the model in real world use [153, 156, 165, 166]. Data generated after the initial batch of data used in the model building and validation process typically consists of more diverse chemical types due to the nature of scientific research. This dataset would be expected to be more challeng-

ing to predict and application of the model to this dataset would be closer to how the model would be used in practice.

A consideration of QSAR model metrics is important for determining if a model is sufficiently predictive for the purpose it was generated for. We therefore discuss a number of commonly used metrics in later sections. It is also important that we consider whether there exists a relationship between the distance of a query compound to the training set and its prediction error [158]. If we can predict which compounds are more reliably predicted by the model [155], then we can employ the model with greater confidence, trusting only predictions for those that lie sufficiently within the space the model is qualified to make predictions for.

2.2.3.2. Linear QSAR Methods

In linear models, the output of the model (either the predicted bioactivity for regression models, or a value used for class assignment), takes the following general form [167-169].

$$y_{pred} = \sum_{i=1}^M w_i x_i + w_0 \quad \text{Equation 1}$$

where y_{pred} is the model output, x_i and w_i the i th descriptor (out of a total of M) and corresponding coefficient (weight) respectively, and w_0 an offset term.

Multiple Linear Regression

The original QSARs pioneered by Hansch and Fujita were developed for congeneric (structurally related) chemical series [170-172]. These studies derived QSARs using Multiple Linear Regression (MLR), for which the coefficients in Equation 1 are found by minimizing the sum of the squared deviations between the experimental bioactivity values and the corresponding values of the model output, for all N compounds in the training set, with respect to these coefficients. Considerable overfitting is expected when descriptors are correlated (even incompletely) so this situation is not advised. It

is further recommended that the number of observations (N) is much greater than the number of descriptors (M) [168].

Principal Components Regression

To overcome the problem of correlated descriptors in MLR, Principal Components Regression (PCR) may be employed. This applies the MLR procedure after estimating a new set of variables (the principal components) for use in place of the original descriptors. The principal components are linear combinations of the original descriptors which are designed to be orthogonal (*i.e.* entirely uncorrelated). More precisely, the coefficients of the original descriptors in each of these linear combinations are given by the eigenvectors of the co-variance matrix describing the correlation between the descriptors in the training set. Since using all M principal components would be equivalent to applying MLR to the original descriptors, a subset of principal components is selected for fitting the regression equation; commonly, this would entail ranking the principal components by their eigenvalues (or variances) and observing how the model performance (estimated on a subset of compounds not used to fit the model) varied with respect to the number of retained principal components [168].

Partial Least Squares (PLS) Regression

PLS entails the application of the same procedure as per PCR, save for initially weighting the descriptor values for each compound by its experimental bioactivity, prior to computing the covariance matrix. This is designed to yield explanatory variables for which the highest variance is associated with the highest correlation with bioactivity [168]. Recently, Gavaghan and co-workers at AstraZeneca derived various 'hierarchical' PLS regression models for hERG pIC₅₀ values (*i.e.* the outputs of one set of models were used as the inputs for the final model). By binning the experimental and predicted pIC₅₀ values, they also interpreted their models as classifiers [153]. PLS may also be used to directly generate classification models. For example, class labels A and B may be assigned values 1 and 0 and a PLS regression model derived, followed by the selection of a threshold output value to enable class assignment [173]. This procedure was recently applied by Clark and Wiseman to derive models for discriminating between drugs with and without the potential to induce Torsades de Pointes (a potentially fatal cardiac arrhythmia related to hERG inhibition) [174].

Linear Discriminant Analysis

Linear classifiers may also be derived from Linear Discriminant Analysis (LDA). A variety of approaches to determine a linear discriminant separation of two classes exist [167, 175] (*i.e.* a model with the form of Equation 1, along with a threshold delimiting predictions for one class from predictions for the other). However, LDA is usually used to denote the determination of such a linear discriminant *via* adjusting the coefficients of the descriptors in order to maximise the Fisher criterion (which effectively entails maximizing the ratio of between-class-variance to within-class-variance within the training set, under certain data assumptions) [167, 175, 176]. Recently, Doddareddy *et al.* [177] used LDA to derive binary classifiers for discriminating hERG blockers from non-blockers, while also comparing this approach to non-linear Support Vector Machines (see below). Like MLR, LDA cannot handle $M > N$ (see above) [173].

Naive Bayes

This method estimates the relative probabilities of a compound belonging to a given bioactivity class by assuming that all descriptors contribute independently to these probabilities. When logarithms of these probabilities are taken (yielding scores which can be used for class assignment), and the classifier is used with binary descriptors (*e.g.* based on occurrences of substructures) [178], this method yields a linear classifier. A commonly used Laplacian modified version of the Naive Bayes classifier was successfully

employed by O'Brien and de Groot to identify hERG inhibitors [13].

2.2.3.3. Non-Linear QSAR Methods

Non-linear methods are inherently more versatile than linear methods in that they do not assume a linear relationship between structure and a biological response. For large, diverse datasets this flexibility should be beneficial; however for small datasets it may lead to overfitting [179]. Concerns have also been raised regarding the lack of interpretability of such models [180].

Recursive Partitioning

In this approach, a decision tree model is generated using the training set. Starting from the entire training set (*i.e.* the "root node"), each descriptor is searched for "cutpoints" which partition the training set compounds within the current "parent node" into K "daughter nodes" [181], such that the separation in the experimental bioactivities of the subgroups of the data passed to different daughter nodes is maximized according to some measure of separation. Commonly only one split criterion is sought per descriptor (*i.e.* $K=2$) [181]. For classification, this separation might be determined *via* the decrease in Gini impurity [182], or a t -test might be used for continuous bioactivities [183]. A variant on the standard approach is to select cutpoints from linear or non-linear combinations of descriptors [184]. Partitioning continues until some stopping criterion (*e.g.* all compounds in the current node belong to the same class) is met. Predictions are generated by passing compounds through the tree, and assigning the majority class or the average bioactivity value for training set compounds in the final ("leaf") node.

Recursive partitioning is notably prone to overfitting [184]. Even small changes to the training set can yield changes to one cutpoint, appreciably changing the structure of the decision tree [181]. Overfitting can be limited to some extent *via* pruning, that is removing branches from the fully grown tree, with the optimal depth of the tree determined using internal cross-validation [185]. Recursive partitioning may also be particularly sensitive to the use of unbalanced training sets (*i.e.* when there are unequal numbers of compounds in different classes) [185].

Higher predictivity, but lower interpretability [181], may be obtained using ensembles of decision trees, such as Breiman's Random Forest (RF) [186], which uses the CART algorithm [187] to grow decision trees on multiple random samples of the training set [173].

Artificial Neural Networks

This class of methods may be used to generate either regression or classification models. As comprehensively discussed by Peterson [188], Artificial Neural Networks (ANNs) are comprised of a series of interlinked layers of "neurons" which transform weighted input variables (signals) into a new signal which may be passed to subsequent neurons. The original input variables are descriptor values and the final output signal(s) are used to make predictions. During training, the training set compounds are sequentially, and randomly, passed through the network, commonly multiple times, and the weights adjusted according to a set of learning rules. Learning may be "unsupervised" (no information about the experimental bioactivities is used to update the weights), "supervised" (the difference between the output and the desired output, based on the experimental bioactivity, for the current training set compound, is used to adjust the weights) or, as an alternative/in addition to supervised learning, "reinforcement learning" uses information about how well the network is currently performing (*e.g.* *via* an error rate) in order to adjust the weights [188].

A plethora of network types exist. In the commonly used "backpropagation" class of networks, the difference between the output and the desired output is not only used to adjust the weights associated with the inputs to the neurons in the output layer, but also to calculate, by computing partial derivatives of the output

error with respect to the weights, how the input weights in preceding layers should be updated to minimize this difference [169, 188]. In "counterpropagation" networks [188, 189], the weights in the Kohonen layer are adjusted on the basis of the difference between the descriptor values and the weights for a "winning" neuron, and the weights for the output layer are adjusted on the basis of the experimental bioactivities. Counterpropagation networks tend to be faster to train compared to backpropagation networks [188]; however, care must be taken to optimize the number of neurons in the Kohonen layer [189]. If this is greater than or equal to the size of the training set, considerable overfitting may occur [188]. Given the tendency of ANNs to overfit [23], a variety of strategies have been developed to prevent overfitting including "early stopping" (training ceases when the performance on a validation set starts to decrease) and *regularization* (see below) [169]. It has also been suggested that, to avoid overfitting, the number of training instances must exceed the total number of weights in the network, with a ratio of 3:2 (training samples:weights) proposed as a "realistic minimum" [190].

Support Vector Machines

Support Vector Machines (SVMs) aim to determine a linear decision boundary (or "hyperplane") which will optimally separate classes in some "feature space" [191]. If the training set is perfectly separable, SVMs find the "maximum margin" hyperplane in order to avoid overfitting (*i.e.* the hyperplane, out of the set of possible hyperplanes separating compounds in both classes, corresponding to the maximum distance, in the perpendicular direction to the hyperplane, between correctly classified points on either side of the hyperplane). If the data are not perfectly separable, SVMs balance the need to maximise this distance, in order to limit overfitting, with the need to minimize the extent of training set misclassification (*i.e.* the trade-off between both objectives is determined *via* an adjustable parameter) [192-193]. Control of overfitting by adding a penalty term to the error function, which limits the extent to which the model can adapt to the training data, is known as *regularization*, and is a generally applicable approach for methods which seek to minimize an error term during training [169]. Importantly, this feature space may be a non-linear projection of the original descriptor space, such that the linear decision boundary in the feature space corresponds to a non-linear boundary in the original descriptor space. The determination of the decision function in the feature space can be achieved *via* the use of a kernel function. If a linear kernel is used, the hyperplane is determined in the original descriptor space (*i.e.* a linear classifier is generated). For non-linear kernels, such as the popular Gaussian Radial Basis Function (RBF) kernel, the shape of the decision boundary in the descriptor space is controlled using an additional parameter [192].

A regression model may be developed using an extension of this approach, Support Vector Regression (SVR) [191-194]. As before, overfitting is controlled using regularization and a kernel function allows for a (potentially) non-linear model to be generated from the descriptors. Importantly, deviations between the experimental bioactivities and the model output for the training set are usually only considered errors if their magnitude is greater than an adjustable parameter. Hence, for SVR, three (or two, if a linear kernel is used) parameters need to be determined, controlling the extent of regularization, the form of the model, and the size of deviations not treated as errors.

It is important to recognize that these approaches are capable of overfitting and that the correct choice of the parameters is *crucial* for their successful application. These parameters should be chosen by assessing model predictivity on an internal portion of the training set not used to fit the model [193] (possibly followed by subsequent training of the model, using the "optimised" parameters, on the entire training set).

***k*-Nearest Neighbours (*k*-NN)**

This approach, which may be used for classification or regression, entails the generation of predictions for experimentally untested compounds using the experimentally determined bioactivities of the *k* most similar compounds in the training set (as determined using some distance metric computed from the descriptors). For regression, a (distance weighted) average over the values for these neighbours may be used as a prediction [162], whilst, for classification, majority voting is employed (*k* is typically an odd number) [195]. *k*-NN approaches can be seen as an extension of the molecular similarity principle, where the distance to the set of neighbors, and hence their contributions to the predicted output, can be weighted in different ways.

2.3. Quantifying Model Performance and Applicability Domain

2.3.1. Assessing the Predictive Ability of Categorical Models

To judge the predictive power of *in silico* models, some, or all, of the following statistical parameters should be reported. These statistics can be easily calculated based on a confusion matrix (*C*) with elements *C_{ij}* denoting the number of compounds belonging to class *i* and predicted to belong to class *j* [196]. For simplicity, we only define these statistical parameters for the case of a 2-class model, which is commonly encountered in *in silico* toxicology. In addition to the simple metrics described in Table 4 and Table 5, the Matthews' correlation coefficient (Equation 2) [197] and Cohen's kappa coefficient (Equation 3) [198] can also be used to gain insight into model predictivity. These measures of agreement for categorical items are generally thought to be a more appropriate measure than a simple percentage agreement calculation, since both these measures take into account the agreement expected to occur by chance. The popular Matthews' correlation coefficient takes into account true and false positives and negatives and is regarded as a balanced measure, which can be used even if the classes are of different sizes. However, every performance measure has its advantages and disadvantages, many of which are discussed in the following reference [196].

$$MCC = \frac{(TP+TN) - (FP+FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{Equation 2}$$

$$kappa = \frac{(TP+TN) - \left(\frac{(TP+FP) \cdot (TP+FN) + (TN+FN) \cdot (TN+TP)}{N} \right)}{N - \left(\frac{(TP+FP) \cdot (TP+FN) + (TN+FN) \cdot (TN+TP)}{N} \right)} \quad \text{Equation 3}$$

2.3.2 Assessing the Predictive Ability of Continuous Response Models

For regression models, the Pearson's correlation coefficient (*r*), the Coefficient of Determination (*R*²) and the Root Mean Square Error (*RMSE*) may be used to quantify predictive performance [199-201]. Common definitions for these metrics are provided below, where *y_{i,obs}* and *y_{i,model}* represent the experimental and predicted bioactivity values for compound *i*, *Y_{obs}* and *Y_{model}*, their respective arithmetic means, and *N* the total number of compounds on which the model is tested.

$$R^2 = 1 - \frac{\sum_i^N (y_{i,obs} - y_{i,model})^2}{\sum_i^N (y_{i,obs} - Y_{obs})^2} \quad \text{Equation 4}$$

$$r = \frac{\sum_i^N (y_{i,obs} - Y_{obs})(y_{i,model} - Y_{model})}{\sqrt{\sum_i^N (y_{i,obs} - Y_{obs})^2 \sum_i^N (y_{i,model} - Y_{model})^2}} \quad \text{Equation 5}$$

$$RMSE = \sqrt{\frac{\sum_i^N (y_{i,obs} - y_{i,model})^2}{N}} \quad \text{Equation 6}$$

Table 4. Classification Confusion Matrix for a 2 Class Model. Which Class of Correctly Predicted Instances (True Positives or True Negatives) and Falsely Predicted Instances (False Positives and False Negatives) One is Particularly Interested in Depends on the Property the User Attempts to Predict

OBS/PRED	Observed +ve	Observed -ve
Predicted. +ve	True Positive (TP)	False Positive (FP)
Predicted -ve	False Negative (FN)	True Negative (TN)

Table 5. A List of Common Classification Statistics used to Assess the Predictive Ability of Models

Statistic	Description	Formula
Total accuracy	Proportion of compounds correctly predicted to be positive and negative relative to total number of predictions.	$(TP+TN)/(TP+FP+TN+FN)$
Sensitivity	Proportion of compounds correctly predicted to be positive relative to all compounds experimentally determined to be positive.	$TP/(TP+FN)$
Specificity	Proportion of compounds correctly predicted to be negative relative to all compounds experimentally determined to be negative.	$TN/(TN+FP)$
Positive predictive power	Proportion of compounds correctly predicted to be positive relative to all predictions categorized as positive.	$TP/(TP+FP)$
Negative predictive power	Proportion of compounds correctly predicted to be negative relative to all predictions categorized as negative.	$TN/(TN+FN)$
Dataset Prevalence	Proportion of compounds that are observed in the positive (or negative) class. For a dataset that contains 95% actives, a model predicting all compounds "active" will therefore have a 95% overall success rate even though it offers no discrimination whatsoever.	$(TP+FN)/(TP+FP+TN+FN)$

2.3.3. The Applicability Domains of in silico Models

The training set used to derive a QSAR model cannot completely cover all of chemical space. Thus, an *in silico* QSAR model will only be predictive of a finite variety of chemical structures. Consequently, there is a need for additional metrics to determine whether a model can be expected to make a reliable prediction for a given query compound. This desire has led to the development of the so called domain of applicability, which may be defined using the distance to model concept. In this method, the similarity of a compound is assessed with respect to the training set, in the context of a specific model, and, if it is found to be sufficiently similar, the prediction can be trusted [155, 156, 158 202-206].

The "applicability domain" (AD) of the model is widely understood to define the range of chemical structures to which the model is 'applicable'. More precisely, a report for the European Centre for the Validation of Alternative Methods (ECVAM) has defined the applicability domain as: "the response and chemical structure space in which the model makes predictions with a given reliability" [207]. When making predictions for an experimentally untested compound, its inclusion within the applicability domain may, of course, only be assessed with respect to its location within chemical structure space. Whilst this may be interpreted as a range of chemical structures for which the expected model performance is well characterized [180], the "applicability domain" is commonly interpreted as a region of chemical structure space in which the model is known to exhibit desirable predictivity [155, 156, 158 205 208].

A distinction may be made between those approaches which simply try to categorize compounds as inside the applicability domain (AD) or outside the AD, and those which seek to directly assess the expected performance of the model for a particular compound

[205]. In the context of predictive toxicology, where the "mechanism of toxic action" is understood, the former approach may be informed by mechanistic reasoning [207]. For example, skin sensitizers may be categorized as belonging to different "natural mechanistic domains", on the basis of the mechanism *via* which they covalently bind to skin proteins and hence trigger sensitization [47]. Models may then be developed that are specific to one such domain. Unfortunately, automated assignment to such domains on the basis of chemical structures is not trivial: For example, aldehydes may either belong to the Schiff base or Michael Acceptors domain. Moreover, for many endpoints the "mechanism of toxic action" may be poorly understood [207]. Alternatively, an approach based upon molecular fragments may be employed [155 207]. The simplest approach is to simply categorize compounds with unknown fragments (*i.e.* fragments unseen in any training set compound) as being outside the AD. This raises various questions, such as which fragments should be used and whether or not to consider additional information such as the frequency with which the fragment occurs within and across compounds in the training set. Kühne *et al.* recently explored some of these questions with respect to Atom-Centered Fragments [208].

An alternative approach has been to relate a distance between the query compound and compounds in the training set to the prediction error. A number of authors have assessed the ability of numerous metrics (*e.g.* distance to the K nearest training set neighbours, leverage etc.) to discriminate between well and poorly predicted compounds [148, 156, 158 204 205]. Sushko *et al.* [205] emphasize the value of considering metrics based on the *output* of the model - in particular, measures based upon the variation in the predictions yielded by an ensemble of models. Similarly, Dragos *et*

al. found that a metric based on the variation in predictions across an ensemble of models was amongst the top three "mistrust scores" for a range of datasets [148].

3.0. TOXICITY MODEL EXAMPLES

The use of *in silico* toxicology models early in the life of a discovery project can prove useful to weed out compounds that are highly likely to cause problems. It is unarguable however, that such models are not sufficiently predictive to make definitive decisions in late studies where costs are significant both in terms of product termination or financial liabilities that might arise later in the life span of a product due to the discovery of significant toxicity.

Nevertheless, *in silico* methods can be effectively used for the following purposes. (a) Models may be used to prioritise compounds for testing in particular toxicity assays. For example, lipophilic bases often cause phospholipidosis [19], or significant hERG inhibition [209], the presence of unsubstituted aniline substituents is an alert for genotoxicity [43], while the presence of an aldehyde attached to an aromatic ring is a Schiff base alert for skin sensitization [24]. (b) In cases where a compound tests positive in a toxicity assay, models can be used to shed light on the mechanism/molecular origin of the toxicity, and then used to suggest ways to remove the liability. (c) In cases where a compound is predicted to be toxic in lead generation, or early lead optimization, but an equally acceptable alternative is available that has no predicted risks, the latter could be prioritized for further development work, with potentially expensive experimental testing performed later if the compound passes other development hurdles.

3.1. hERG Inhibition

A hERG inhibition assay is an early hurdle a compound/series in drug discovery must overcome. Inhibition of this ion-channel has been linked to a potentially fatal cardiac arrhythmia, Torsades de Pointes (TdP) [153]. A variety of different experimental assays are routinely employed to measure hERG inhibitory potential, including competitive binding assays, as well as both manual and, higher throughput, automated electrophysiological assays [210, 211]. From recent reports, it is apparent that significant hERG inhibition does not necessarily mean that a drug will induce Torsades de Pointes [51, 75]. This may be in part due to the fact that the plasma concentration of the compound is an important additional determinant of its physiological effect. Nevertheless, not least due to FDA guidelines, there is considerable pressure to minimize the likelihood of this form of toxicity occurring, meaning hERG inhibition is still routinely screened for in drug discovery [153].

A wide range of experimental protocols are employed for assessing hERG inhibition [210]. It has also been suggested that the level of variability for replicate electrophysiological measurements is typically of the order of 0.10 to 0.50 log units [153]. The resulting variation in the IC_{50} values reported in the literature for the same compounds, commonly measured using different protocols, means that the generation and validation of *in silico* models using literature mined datasets is a recognized challenge [161, 177, 212, 213]. Furthermore, an additional challenge is the small size of many literature datasets [161, 214, 215]. Whilst it has been suggested that training on larger literature mined datasets may decrease predictive performance due to increased data inconsistency [213], Doddareddy *et al.* have argued that combining data from patch-clamp and binding assays to yield larger training sets may actually be beneficial [177].

The prediction of hERG inhibition potential has been extensively tackled using QSAR approaches. Both regression and classification based approaches have been employed for this purpose. For the most part, the latter approaches have entailed the construction of binary classifiers designed to discriminate between inhibitors and non-inhibitors. The experience of Yao and co-workers within GlaxoSmithKline is that compounds with an $IC_{50} < 1 \mu M$ usually

induce QT prolongation (a surrogate marker for TdP) in *in vivo* assays, and their development is usually discontinued, while those with $IC_{50} > 10 \mu M$ usually do not and are typically progressed [216]. These cut-offs are generally accepted values elsewhere in the pharmaceutical industry [153, 217]. In keeping with this, various modeling studies have sought to develop binary classifiers based on either a $1 \mu M$ or $10 \mu M$ threshold [161, 185, 212, 217]. However, a variety of other thresholds (between $130 nM$ and $40 \mu M$) has also been proposed [217]. Some studies have sought to develop models designed to discriminate between strong ($IC_{50} < 1 \mu M$) and weak inhibitors ($IC_{50} > 10 \mu M$) [185, 217, 218]. Whilst this type of strategy may have some justification, in light of possible class ambiguities arising due to experimental uncertainty [177], these models might be of limited value in drug discovery, where many compound series exhibit moderate inhibition values (i.e. IC_{50} in the range $1-10 \mu M$) [153]. To our knowledge, only Dubus *et al.* [185] and Thai and Ecker [189, 219] have specifically sought to develop ternary classifiers discriminating between strong, moderate and weak inhibitors, with only the latter authors having sought to externally validate their models; Gavaghan *et al.* also considered how their regression model might be used to categorize compounds as strong, weak and moderate inhibitors [153].

Modeling Studies

A diverse set of computational approaches has been employed to predict hERG inhibition *in silico*. These approaches include the use of pharmacophores to define inhibitors of hERG [220, 221], PLS regression based on docking scores [222] (structure-based approaches using homology models, in the absence of a crystal structure [223]), as well as a variety of ligand-based QSAR studies which are our primary focus here. QSARs based on 2D [189, 217, 219], including variants using 2D molecular fingerprints [13, 161, 177], 3D [219] and 4D [213] descriptors have been described in the literature.

All manner of statistical algorithms have been used to relate the types of descriptors described above to hERG activity, including PLS regression [153], PLS classification [213, 214, 224], LDA [177], linear [212] and non-linear SVMs [161, 177, 212], linear [194] and non-linear [225] SVRs, Laplacian modified Naive Bayes classifiers [13], and ANNs [189, 219, 224]. Many of the earlier QSAR studies were reviewed by Demel *et al.* [226], while a summary of the performance of different classification models was also presented more recently by Thai and Ecker [219] and Marchese Robinson *et al.* [161].

Obrezanova and Segal reported binary classifiers based on the non-linear Gaussian Processes method [214]. This approach has also recently been used to build hERG regression models [227]. Recently, Marchese Robinson *et al.* reported, to the best of our knowledge, the first hERG classification models based on the Winnow method [161]. Both these relatively underused methods were compared, in conjunction with the same descriptors, to other widely used statistical methods. Obrezanova and Segal [214] found that both their binary classifiers, based on a $10 \mu M$ threshold, generated using Gaussian Processes, outperformed RF, SVM (using a RBF kernel), a Decision Tree and PLS, using the same descriptor sets. They presented the following kappa values for these methods using an external test set (save for their Decision Tree model, which was optimised using this test set): 0.66 and 0.60 (Gaussian Processes), 0.54 (RF and SVM), 0.58 (Decision Tree) and 0.43 (PLS). Marchese Robinson *et al.* found that binary classifiers ($1 \mu M$ threshold) built using Winnow (a linear method) performed comparably well to both RF and SVM (using a RBF kernel) models generated using the same high dimensional descriptor sets and trained on the same data. They reported MCC values, on randomly selected external test sets, across the following ranges: 0.43-0.59 (Winnow), 0.40-0.55 (SVM) and 0.44-0.52 (RF). [161] However, in keeping with many studies based on literature mined data [226], these stud-

ies validated their models on small test sets, comprised of (up to) 50 [214] and 148 [161] literature mined compounds respectively.

It is of interest to note that studies on large propriety pharmaceutical datasets by Johnson *et al.* [228] and Gavaghan *et al.* [153], both obtained some RMSE values, from external validation, less than 0.65, with linear models which are readily interpretable. However, studies which also assessed their models on large amounts of data, and which compared non-linear and linear approaches, appear to suggest that linear approaches may be somewhat outperformed by non-linear approaches. In the studies by Doddareddy *et al.* [177] and O'Brien and de Groot [13], the comparison between the linear and non-linear approaches is confounded by the lack of *regularization* to control overfitting for the linear LDA model [177] and the use of different descriptors [13] respectively. Neither of these confounding factors exists when comparing the linear Ridge Regression model to the non-linear approaches employed by Hansen *et al.* [225].

A common theme echoed by both Gavaghan *et al.* [153], as well as Thai and Ecker [189], working with large in-house datasets and linear methods, and considerably smaller, literature mined datasets and neural networks respectively, is the inherent difficulty in discriminating moderate inhibitors from strong and weak inhibitors (as previously defined). Whilst Thai and Ecker's approaches to discriminating between these three classes showed some promise, it must be noted that their predictive power was only assessed on small quantities of data (various 20-50% subsets of 285 compounds [189], or a set of 58 compounds [219] were used to test their models) and that the performance of some of their models was clearly appreciably different when different training and validation approaches were employed. For example, the (average) fraction of moderate inhibitors they were able to identify with one of their models fell from 86% to 46% depending upon the manner in which their data was partitioned into training and test sets [189].

Model Interpretation

The study by Gavaghan *et al.* [153] also serves to emphasise how *in silico* models for hERG inhibition can be used. (1) To serve as a screening tool for eliminating potent hERG inhibitors from chemical libraries in early drug discovery. This could serve as an alternative to more expensive, time consuming experimental assessment, or the model could be used to prioritise sending predicted inhibitors for experimental assay. In addition, virtual libraries could be filtered prior to synthesis. (2) If the model is interpretable, the contributions to hERG inhibition for a particular compound may be discerned, suggesting synthetic strategies to the medicinal chemist for the reduction of hERG activity. It is conceivable that a suitably interpretable model may allow for an "inverse QSAR" approach, whereby one seeks to predict the specific change in molecular structure required for a specific change in biological activity [129]. The examination of electrostatic complementarities between a docked inhibitor and a hERG homology model may also be used for this task [229].

Various studies, including those which have sought to evaluate fragment contributions to hERG models [161 194, 230], have suggested structural fragments contributing to or reducing hERG blockade. For example, tertiary amines (less so those on the molecular periphery) [161], as well as fluorinated phenyl rings, are suggested to contribute to hERG blockade, while amides are suggested to reduce it. As discussed in these studies, the first of these relationships is well known in the literature, and various mechanistic interpretations - cation- π interactions [12], non-classical hydrogen bond facilitation [231], or non-specific electrostatics [223] - have been proposed, based on the fact that amines are expected to mainly exist in a protonated form at physiological pH. The typical increase in hERG inhibition with molecular size, as well as lipophilicity, as well as the significance of ionisation state, was also emphasised in a recent study utilising thousands of compounds

[14]. These clear relationships may suggest that descriptors encoding these physiochemical properties should be included in models for hERG blockade. However, the most appropriate descriptors to encode these is not always obvious and the contributions of descriptors designed to do so may not always be as expected; for example, a recent study found that two descriptors, both encoding hydrophilic character, made both positive and negative contributions to predicted hERG inhibition [213].

3.2. Genotoxicity / Carcinogenicity

Carcinogenic chemicals are divided into two broad categories based on their presumed mode of action: genotoxic and non-genotoxic. Genotoxic carcinogens cause damage by interacting/binding directly with DNA (mutagens), whereas non-genotoxic carcinogens cause changes in cellular processes, differing considerably from the process of binding directly to DNA [232]. Chemicals are defined as carcinogenic if they induce tumors, increase tumor incidence or shorten the time to tumor occurrence. The carcinogenic potential of a chemical also depends on the conditions of exposure (*e.g.*, route, level, pattern and duration of exposure).

The bacterial reverse mutation assay (Ames test) is commonly used to detect for mutagenicity and has been widely used as an early alerting system for potential genotoxicity. This assay was designed to detect and identify genetic damage caused by chemicals in bacterial cells [8, 233-235]. In the Ames test, frame-shift mutations or base-pair deletions may be detected by exposure of histidine dependent strains of *Salmonella typhimurium* and/or *Escherichia coli* to a test compound. When these strains are exposed to a mutagen, reverse mutations that restore the functional capability of the bacteria to synthesize histidine enable bacterial colony growth on a medium deficient in histidine. These altered bacteria are referred to as "revertants". Since many chemicals interact with genetic material only after metabolic activation by enzyme systems not available in the bacterial cell, the test compounds are, in many cases, additionally examined in the presence of a mammalian metabolizing system, which contains liver microsomes (S9 mix). A compound is classified as Ames positive (otherwise negative) if it significantly induces revertant colony growth in any of the (usually five) strains, tested either in the presence or absence of an S9 mix. As a consequence of this definition, Ames-negative compounds in the benchmark dataset which have not been tested in all recommended strains may turn out to cause reverse mutations when examined in additional strains.

The main sources of publically available data for this assay are; the US Food and Drug Administration (FDA), Centre for Food Additives and Applied Nutrition (CFSAN), Food Additive Resource Management system (FARM), Chemical Carcinogenicity Research Information System (CCRIS) [236], the National Toxicology Programs (NTP) Genetic Toxicology database [237] and the Tokyo-Eiken database [238]. As there is quite a large dataset available for this test, it is not surprising that this is the most commonly modeled endpoint for genotoxicity.

Kirkland *et al.* [239] evaluated the ability of a battery of *in vitro* genotoxicity tests (Ames, mouse lymphoma assay, *in vitro* micronucleus (MN) and chromosomal aberrations) to discriminate rodent *in vivo* carcinogens and non-carcinogens. It was found that a combination of two or three test systems had greater sensitivity than individual tests, resulting in sensitivities of around 90% or more, depending on the test combination. The sensitivity of individual methods was between 59% (for Ames, over more than 500 chemicals) and 79% (for MN, over more than 80 chemicals). The specificity of the Ames test was reasonable (73.9%), but all mammalian cell tests had a low specificity (below 45%), and this was reduced in combinations of two and three test systems. This highlights the difficulties in the current ability to extrapolate from *in vitro* genotoxicity results to *in vivo* carcinogenicity.

As discussed above, *in silico* toxicology methods are hazard identification methods and in most of cases they don't take dose and exposure into account unless an exposure-response relationship has been studied. Therefore, these methods cannot predict toxicity in isolation, but can provide useful supplementary information for the overall risk assessment process. For example, the aromatic nitro group is a well known fragment that triggers a structural alert for carcinogenicity, but if a chemical containing this fragment has very low exposure, or bioavailability, it is questionable whether the toxic prediction will be correct. Therefore, whenever possible, internal exposure (i.e. the amount taken up and distributed as given by the free plasma concentration) should be taken into account by use of either *in silico* or *in vitro* ADME data. Ideally, the results of the predictions should be combined with other evidence and data for consideration for the risk assessments.

Modeling Studies

In silico predictive models for genotoxicity fall into two principal categories: rule based (expert systems) and QSARs. The former approach is associated with the local reactivity of chemicals, i.e., reactivity of functional groups or structural alerts. The key step in the development of this approach is defining chemical categories for genotoxicity and defining the organic chemistry associated with the formation of a covalent bond between DNA and an exogenous chemical. In this approach, a well-defined reactive group which has the potential to interact with DNA is identified. Several rule-based systems have been developed [240-242] which help to summarize the relationships between specific chemical substructures and observed mutagenicity outcomes. This technique can be an invaluable tool in the *in silico* prediction of genotoxicity, as it is very simple and highlights the presence of certain substructures (toxicophores) within the molecule, and can be related to the Ames test outcome. This alert approach may also provide mechanistic understanding of the Ames test outcome.

It should be noted that this approach is generally used to predict if a chemical to be Ames positive only. If no alert is triggered, it does not mean that the chemical under investigation will be Ames negative [243]. A recent comprehensive review [232] of different *in silico* models and approaches for predictions of genotoxic outcome, shows that most of the earlier approaches described for the prediction of Ames mutagenicity produced good specificity and sensitivity values (prediction accuracy of up to 85%). Depending on the descriptors and the statistical methods used, some of the models offer simple SAR information [82, 244], whilst others are harder to interpret due to the choice of chemical descriptors derived from structural information [245, 246].

QSAR models are widely used to predict genotoxicity. One of the advantages of these approaches is that they can be used to predict both positive and negative outcomes (i.e. unlike structural alerts which focus on toxic/positive molecules). Different QSAR and machine learning methods have been used to derive *in silico* predictions about the Ames outcome of the chemicals. These include Ames test QSAR models using PLS, NN, RF, and SVM [46, 244-250].

There have been several attempts to generate *in silico* models from rodent *in vivo* carcinogenicity data [46, 251]. However, as with many toxicity measures, there are difficulties in modeling this endpoint due to the diversity of carcinogenicity pathways and the relatively small number of compounds available. Several complex pathways contribute to carcinogenicity for which reliable data sets are in general unavailable. Most carcinogenicity models have primarily focused on rodent bioassays. Using a dataset extracted from the Carcinogenicity Potency Database CPDB, consisting of 805 chemicals with rat TD₅₀ values, Fjodorova *et al.* [251] generated a classification model using a Counter-Propagation ANN method. Although a good accuracy (93%) for classification was reported for the training set, the corresponding value for the test set dropped to

approximately 70%. While this model may still be of value, the result suggests that overfitting is a possibility, and that applicability domain considerations are likely to be important and must be taken into account.

At present, QSAR methods are more reliable for predicting genotoxic potential than carcinogenic potential. As discussed above, carcinogenicity prediction represents a considerable challenge due to the multitude of possible mechanisms of toxic action. One of the main aspects these models do not include is a consideration of ADME properties, which could be critical steps in the carcinogenic process. It is crucial we understand how a molecule is distributed across different organs to fully understand their toxic potential.

3.3. P450 Inhibition

Drug-drug interactions (DDIs) arise from either inhibition/blockage or induction of certain metabolic pathways, causing substantial variations in drug concentrations present in the systemic circulation [252]. Inhibition of one or more cytochrome P450 isoforms may potentially block the metabolism of a drug molecule, which may in turn lead to its accumulation in the body. In the case of metabolism-activated prodrugs, inhibition of an enzyme needed for its activation may lead to a loss of pharmacological efficacy.

The prediction of DDIs is a non-trivial, complex problem that has traditionally been addressed using elaborate clinical studies [253]. Indeed there are some major uncertainties regarding the extrapolation of *in vitro* assay data to *in vivo* effects [254]. Several *in vitro* P450 inhibitors, such as clotrimazole and other imidazoles have in fact been observed to actually induce these proteins *in vivo* [255]. One major challenge in predicting systemic effects derives from the crosstalk between receptors regulating metabolizing enzymes [256]. Though assays are becoming more readily available, and greater insight into the mechanisms of inhibition and induction of metabolic enzymes has been gathered, a complete framework that will allow the accurate prediction of enzyme inhibition and induction is still missing [252].

P450 enzymes play a pivotal role in drug metabolism and DDIs. This is particularly observed for drugs where metabolism is dependent on a single P450 isoform. Polymorphism of certain P450 isoforms such as P450 2D6, P450 2C9 or P450 2C19 adds an additional layer of complexity to the problem [257, 258]. Inhibition is, in general, considered to be more problematic than P450 induction due to its potential to cause toxic effects due to compound accumulation. Inhibition is commonly evaluated by determining the IC₅₀ or K_i using human liver microsomes, cDNA-expressed microsomes or recombinant protein systems [58]. In addition, as a result of the availability of 3D structures for a variety of P450s, a wide range of 3D modeling techniques have been used to model P450 inhibition.

Modeling Studies

A plethora of P450 QSAR modeling studies have been reported in the literature, reflecting the considerable interest in predicting interactions of small organic molecules with P450s [60]. The quality of training and test data is critical in defining the performance of a computational model. Assay data may be inaccurate or noisy and, in the case of highly promiscuous interaction sites like those observed for P450 3A4, may mean assay data based on a single probe may be not sufficient to map the binding properties of the enzyme or its inhibitors [56]. Potent inhibitors of P450 are often found to coordinate directly with the heme iron, as illustrated in Fig. (2).

The extent to which coverage of chemical space can be achieved by a model, which is dictated by the diversity of the molecules used for training and testing, is particularly critical for modeling P450 activity given the promiscuity of these proteins for a diverse range of molecules. Despite this, only a minority of reported P450 QSAR studies have considered the applicability domain



Fig. (2). Cytochrome P450 3A4 with metyrapone (PDB accession code 1W0G). The protein is denoted using a cartoon representation, the heme using sticks, and metyrapone using spheres.

concept. The commonly observed superior prediction accuracy of local models when compared to more general, global models comes at the cost of a much smaller domain of applicability [159]. In addition, performance differences between linear and non-linear approaches are in general smaller for local models, likely due to the fact that local phenomena are more easily described by linear relationships. For more information on challenges for QSAR modeling of P450 data the reader is referred to [22].

Local models have been reported for the major CYPs of interest in drug development. Examples include models for the prediction of flavonoids interacting with P450 1A2 using MLR and back-propagation ANN models [259], as well as MLR, PLS, genetic function approximation, and genetic PLS [260]. Saraceno *et al.* [261] recently reported QSAR models for P450 2D6 based on 51 known inhibitors. The authors found that only models including 3D descriptors in addition to 2D descriptors obtained high prediction accuracy. Susnow and Dixon used 2D structural descriptors to identify inhibitors of P450 2D6 [262], using a recursive partitioning approach and training data of 100 compounds with known inhibition constants. Correct classification was obtained in 75 to 80% of the cases. For P450 2B and 3A, a local, bilinear model for azoles was reported, which points out the importance of logP for binding affinity [263].

Global models based on PLS, MLR, classification and regression tree (CART) and Bayesian ANNs models were developed for P450 1A2 [264]. The training set consisted of 109 compounds and a validation set of 249 orally applied administered drugs, respectively. Similar results were obtained for the four methods, with R^2 values of the training set ranging from 0.72 to 0.84 and correct classification in 83% of all cases for a consensus model. A complicating factor was the bias of inactive molecules in the test set making true validation more difficult. For the same isoform, recently a number of classification models using a set of approximately 400 inhibitors and 7000 non-inhibitors were reported [265]. Binary QSAR, SVM, RF, k-NN and decision tree models were developed using Volsurf [266] and MOE descriptors [104], with between 73-76% of compounds included in the test set predicted correctly. Burton *et al.* employed recursive partitioning to classify inhibitors and non-inhibitors of P450 1A2 and P450 2D6, with data sets of 306 and 498 molecules, respectively. Again, MOE descriptors were employed and accuracy of >80% was reported [267].

Recursive partitioning was also used in a study by Ekins *et al.* to model the inhibition potential of more than 1750 molecules for P450 2D6 and P450 3A4 (Spearman's rho of 0.61 and 0.48 for CYP2D6 and CYP3A4 on the test set, respectively) [268]. Both

isoforms were also modeled by Jensen *et al.*, who used a Gaussian kernel-weighted k-NN approach employing extended connectivity fingerprints (ECFPs) using 1153 P450 data points for 2D6, and 1382 values for P450 3A4, respectively [269]. Correct classification was obtained for 82% to 88% of the test set molecules, with 10-14% of the compounds not being classified.

Examples of QSAR studies for the prediction of inhibitory potential for P450 3A4 include the PLS binary classification model by Zuegge *et al.* [270] based on a training set of 311 compounds and 333 descriptors. Correct predictions were obtained for 95% of the training data and for 90% of a semi-independent validation set of 50 compounds. In a further study, 807 drug-like compounds were used to train SVM classification models for P450 3A4 inhibition [271]. A three-class model based on 2D descriptors yielded correct predictions for 70% of a comprehensive test set. Using structural fingerprints and topological indices, SVMs were also found to outperform various algorithms such as recursive partitioning, Bayesian classifier, logistic regression, k-NN in a study based on 4470 inhibitors and non-inhibitors of P450 3A4 [272]. This investigation demonstrated the correlation between the prediction accuracy and the structural similarity of the query molecule to the training data. A drop in predictive power was identified for molecules sharing a Tanimoto similarity index lower than 0.8 with compounds of the training set. Similar observations were also reported by Zhou *et al.*, who used a set of 826 P450 3A4 inhibitors and 873 non-inhibitors to train a SVM binary classification model [273].

Several QSAR efforts aimed at the prediction of inhibition for multiple isoforms. Regression QSAR models were developed for six P450 isoforms relevant to xenobiotic metabolism, P450 1A2, P450 2B6, P450 2C9, P450 2C19, P450 2D6 and P450 3A4 [274]. The authors of this study report correlation coefficients for this approach between 0.94 and 0.99, but it should be noted that the datasets were of very small size. Interestingly, it was found that the consideration of hydrogen bonding and pi-pi interaction capability is crucial for model performance. Gleeson *et al.* used PLS and regression trees in combination with a number of relatively interpretable descriptors to develop global classification models for P450 1A2, P450 2C9, P450 2C19, P450 2D6 and P450 3A4 [275] based on a high quality data set of approximately 1500 compounds. Interestingly, the P450 3A4 model was found to outperform models of other isoforms such as P450 2D6 and P450 2C19, which the authors infer is due to its greater promiscuity/softer molecular recognition features. Another recent study employed PLS regression along with 18 non-linear methods on eight P450 isoforms including P450 2D6, P450 1A2, P450 3A4 and P450 2C9. The predictions of pIC_{50} values displayed R^2 values between 0.69 and 0.94 for a subset of six descriptors, as well as R^2 values between 0.78 and 0.99 for a subset of 6 and 15 descriptors, respectively [276].

Hammann *et al.* employed k-NN, decision trees, RFs, ANNs and SVMs using different kernel functions to classify 335 substrates, inhibitors of P450 1A2, P450 2D6 and P450 3A4 isoforms [277]. Using 188 descriptors and 10-fold cross validation, classification performances were reported between 82 to 94% for the different isoforms. In reality a significant proportion of xenobiotics are metabolized by more than one isoform. This was addressed in a recent study employing multi-label classifiers such as SVMs, multi-label k-NNs and ANNs [278] based on a set of 580 substrates of seven P450 isoforms (approx. 15% of these substrates are known to bind to up to five enzyme variants). While single-label and multi-label classifiers in general seem to obtain comparable prediction rates, the multi-label approach more adequately represents the real situation and avoids information loss on isoform specificity.

3D QSAR Methods

3D QSAR methods such as Comparative Molecular Field Analysis (CoMFA) have been used to predict P450 inhibition of small organic molecules. While these 3D methods introduce an-

other level of complexity by their dependency on an accurate ligand alignment, the 3D visualization of the observed stereochemical and electronic properties responsible for the exhibited bioactivity can provide valuable insights for the interpretation of the underlying protein-inhibitor interactions. Examples of the application of 3D QSAR methods to analyze P450 inhibition include the derivation of CoMFA models for P450 1A2 [279], P450 2A5 [280], P450 2A6 [280], P450 2B6 [281] and P450 19A1 (aromatase) [282].

Molecular interaction fields (MIFs) encode the variation in interaction energies between a target structure and a chemical probe in three-dimensional space. GRID [283], which has been developed as a tool to assess protein surfaces for areas of favorable interaction energies with ligands, has been extensively used for P450 enzymes in combination with several computational approaches. The GRID/GOLPE approach was used to rationalize the inhibitory activity of compounds on P450 2C9 [284], P450 2A5 [280] and P450 2A6 [280] and P450 1A2 [279]. A potential drawback of GRID and related descriptors is their dependence on the pre-alignment of the molecules. This of course can be bypassed using GRID alignment independent descriptors (GRIND) [285, 286]. These descriptors were successfully employed to generate a PLS-derived model for quantitative predictions ($r^2 = 0.77$, $q^2 = 0.60$) of P450 2C9 inhibition [287]. Another GRIND-based 3D-pharmacophoric model was reported for P450 3A4, where a data set of 331 compounds with inhibition data was used to derive a 3D-QSAR model from MIFs [288, 289].

3D Modeling Studies

A pharmacophore defines a pattern of chemical and steric features essential for a ligand to exhibit biological activity. A major reason for the popularity of this approach is its simple underlying concept that relates closely to the way medicinal chemists intuitively consider SAR. Pharmacophore models for the qualitative and quantitative assessment of substrate and inhibitor binding have been reported for a large number of P450 isoforms relevant to drug discovery, including; P450 1A2, P450 2A6, P450 2B6, P450 2C9, P450 2D6, P450 3A4, P450 3A5 and P450 3A7. These models have been discussed and summarized in a number of comprehensive reviews elsewhere [63, 290-293].

A collection of eleven structure-based and ligand-based pharmacophore models for P450 substrates and inhibitors has been published by Schuster *et al.* [294]. Pharmacophore models of P450 2C9 substrates typically include a hydrophobic/aromatic and a negatively charged interaction feature [290, 291]. There are, however, also non-anionic substrates of 2C9 known that have been considered by Locuson *et al.* [295]. Pharmacophores of P450 2D6 include an aromatic interaction as well as a characteristic positive charge about 5 to 7 Å distant from the oxidation site [291]. Contrary to more specific P450 isoforms, the promiscuity of P450 3A4 is reflected in the apparent absence of pharmacophoric requirements [292]. Mao *et al.* showed that for modeling QSARs of P450 3A4 inhibitors a number of local pharmacophore models is required to adequately represent distinct binding modes [296]. A more recent study used three individually derived pharmacophore models to identify substrates and inhibitors of P450 2A6 in an attempt to account for protein flexibility. This approach was combined with an SVM to increase prediction accuracy [297].

Protein-ligand docking has been used to predict potential drug-drug interactions via P450 2D6, which may affect the therapeutic success of anti-cancer treatments [298]. A homology model of P450 2D6 was derived to dock 20 drugs commonly prescribed to cancer patients using GOLD [119]. 13 compounds were identified *in silico* to potentially bind to P450 2D6. For eleven of these compounds, binding to P450 2D6 was confirmed experimentally.

A wide variety of techniques can be applied, from simple SAR, complex multivariate QSARs, to more complex 3D modeling methods traditionally used in structure based design. P450 inhibi-

tion is probably the most accessible area of toxicology to mainstream computational chemistry due to the wealth of experimental inhibitor screening and X-ray structural data, and is likely to remain so, at least in the near future.

3.4. Metabolite Prediction

The prediction of metabolites formed by a particular molecule *in vivo* is of critical importance, as the presence of these in the body may give rise to undesirable toxic side effects due to a wide variety of reasons (protein reactivity, P450 inhibition etc). The experimental determination of these metabolites is very resource intensive, meaning *in silico* methods for prediction are highly sought after [299].

Within the past few years a wide variety of computational approaches and tools have been developed that attempt to pinpoint the most likely site of metabolism (SOM) of a molecule and the resulting metabolites. These can be classified into QSARs, fingerprint-based methods, shape-focused techniques, molecular interaction fields, protein-ligand docking and reactivity-based methods. Most of these approaches consider one particular aspect of a metabolic reaction such as the reaction energy barrier, geometrical properties or pharmacokinetic aspects. In reality, however, an ensemble of properties is required for a metabolic reaction to take place and, hence, combined approaches have been developed to enable more realistic assessments. Readers are referred to the excellent review of Tarcsay and Keserü on SOM prediction for more detail on this area [300].

Expert systems were the earliest computational approaches to predict metabolic liability in rational drug development. They are based on dictionaries of biotransformations, encoding rules for metabolic reactions derived from *in vivo* and *in vitro* experimental data. These dictionaries can in general be adapted and extended by the user and several of the software tools come with a dedicated programming language for setting up such rules. Examples of expert systems include a number of well known products including META [301], MetabolExpert [302], METEOR [64], SyGMa [303] and TIMES [92, 304]. Due to their rule-based approach, expert systems tend to have a high false positive hit rate. In order to filter these false positives, Tarcsay *et al.* [305] recently proposed a combination of MetabolExpert with the docking protocol Glide [118].

Fingerprint-based descriptors allow for the encoding of the chemical environment of SOMs as well as metabolically stable atom environments. Mining metabolic reaction databases such as the Accelrys Metabolite Database [306] allows for the derivation of probability scores that an atom in a defined environment will be involved in a metabolic reaction. MetaPrint2D is a software tool to mine these (and other) biotransformation resources and enables the prediction of the most likely SOMs for xenobiotics, providing a normalized probability score [307, 308]. However, this score does not account for the absolute probability of a metabolic reaction. The approach is not limited to P450-based reactions and an additional module, MetaPrint2D-React allows one to predict the metabolites based on SMIRKS pattern-encoded chemical reactions [141]. MetaPrint2D is open source software [309]. Also, the PASS method (Prediction of Activity Spectra for Substances) uses fingerprints (Multi-Level Neighborhood of Atoms) to predict the metabolic liability of atom positions of small organic compounds [310]. This software tool reports the name of the biotransformation, but does not explicitly highlight the SOM or generate structures of potential metabolites.

The P450 Regioselectivity Module, provided by ACDlabs, is a recently introduced commercial software tool for the prediction of SOMs, based on a comprehensive training set of more than 900 compounds. The method assigns a probability score for a biotransformation to occur at a defined atom position considering N- and O-dealkylation, aliphatic and aromatic hydroxylation, as well as S-oxidation. Also, isoform-specific predictions can be performed on

five of the cytochromes most relevant to xenobiotic metabolism using additional modules from ACDlabs [311].

Well-known approaches and products for SOM prediction are based on MIFs. MetaSite [101], which evolved from earlier developments using GRID and ALMOND [312, 313], combines several different computational approaches to address metabolic liability: A MIF-based module for the assessment of protein and ligand properties and a quantum chemical component to simulate metabolic reactivity. More recently, a knowledge-based module was introduced, which accounts for preferred reaction types for the individual P450 isoforms. The authors of MetaSite report a success rate of 85% for correctly identifying the known SOM amongst the two top-ranked positions, and an 80% success rate for the identification of the enzyme isoform involved in the biotransformation [101]. Comparing several methods for the prediction of SOMs of P450 3A4-based biotransformation, Zhou *et al.* found that MetaSite correctly predicted at least one SOM within the top three atom positions in 78% of all cases [314]. In a more recent prospective assessment, SOMs for atom positions ranked first by MetaSite were confirmed as correct in 55% of all cases, increasing to 84% when considering the top three atom positions [315].

Protein-ligand docking is another frequently employed approach to predict SOMs based on the orientation of the ligand within the catalytic binding site of P450s. Atom positions closest to the heme iron are considered the most likely SOMs. Unwalla *et al.* used Glide to identify SOMs of P450 2D6. Predictions were deemed correct in cases where a known SOM is located within a 4.5 Å distance from the catalytic center among the five highest-ranked poses. This was found to be correct for 85% of the 16 investigated compounds [316]. This study points out the common observation that, due to considerable conformational rearrangements of the protein upon ligand binding, P450 apo structures generally yield inferior performance compared to holo structures. Related studies were reported using GOLD [317, 318], which also investigated the impact of water molecules placed in the binding site. The effect of the latter was systematically investigated for 19 P450 structures [319] using AutoDock [320], FlexX [321], and GOLD and, in a study by Santos *et al.*, again using GOLD [322].

P450 enzymes show a remarkable degree of flexibility within the binding site and, hence, make it a prime target for ensemble-based docking approaches. Teixeira *et al.*, for example, used molecular dynamics simulations to derive an ensemble of P450 3A4 structures [323]. While the original X-ray structures did not allow for the identification of the known SOM for any of the 16 ligands under investigation, the ensemble-based technique was able to successfully predict the correct SOM for all 16 compounds in at least one of the investigated MD-derived protein conformations. Also, using docked poses as a starting point for MD simulations may considerably improve prediction accuracy [324].

Shape-based methods are becoming increasingly popular as a technique for similarity-based virtual screening [325-330] and bioactivity profiling [331]. They are based on the observation that compounds of comparable shape are more likely to exhibit similar bioactivity at the same receptors. A recently published study employed the shape-based screening engine ROCS (Rapid Overlay of Chemical Structures) [332] to identify the most likely SOMs for P450 2C9 substrates [333]. In this study, flurbiprofen was extracted in its protein-bound conformation from an X-ray structure and used as a query for alignment of 70 known P450 2C9 substrates. The authors reported the correct alignment of the known SOMs of these substrates to the known SOM of flurbiprofen in 60% of all cases, with 89% out of the 44 top-ranked compounds predicted correctly.

Quantum chemical (QC) methods allow for prediction of liability to metabolism based on theoretical reactivity measures. For aliphatic hydroxylation, aldehyde hydroxylation and dealkylation reactions, the hydrogen abstraction energy from a carbon atom

correlates with metabolic reactivity [98]. Modern approaches include SMARTCyp, which ranks SOMs based on pre-computed transition state energies, in conjunction with an accessibility descriptor (SPAN) [334, 335]. Recently, the RS-Predictor program was developed, a method that uses topological and QC descriptors to represent the reactivity potential of SOMs [336]. The tool was found to correctly predict known SOMs for 78% of all cases between the two top-ranked atom positions, outperforming SMARTCyp and StarDrop (see below). For further information the reader is referred to [61].

One of the earliest approaches combining several different techniques to identify SOMs employed ligand-based pharmacophores, homology modeling and molecular orbital (MO) calculations [337]. The experimental SOM was identified for six out of eight investigated compounds. Related to MetaSite, which has been already discussed above, StarDrop identifies SOMs via a combination of approaches, including a QC-based, global P450 model. In addition, StarDrop features P450 isoform-specific, ligand-based models derived from known substrates, and considers logP to improve prediction accuracy. Recently, another combined approach was presented that uses five QC descriptors related to reaction energies, plus the energies calculated using the SmartCyp approach, SASA and pharmacophoric constraints, to reflect the properties of the P450 3A4 binding site [338]. The best performing model was obtained using all three different descriptor types and RF, which was found to correctly identify known SOMs in 82% of all cases among the two highest-ranked atom positions. Ligand docking methods do not account for reactivity and, hence, are frequently combined with QC-based methods in particular.

Kuhn *et al.* analyzed the P450 3A4-catalyzed hydroxylation and O-dealkylation of sirolimus (rapamycin) and everolimus (RAD-001) employing a QC/docking/MD simulations workflow [339]. This led to valuable insights on how to potentially reduce the metabolic liability of these compounds. MLite is a model for the prediction of P450 3A4-mediated metabolism based on a docking algorithm and a QC method developed by Korzekwa *et al.* [340], and Jones *et al.* [341] for the estimation of reactivity [342]. With the optimized model, correct predictions of the SOM were obtained for 76% of the 25 investigated compounds, when taking into account the two highest-ranked atom positions for each molecule.

Finally, one of the latest methods employed to predict the regioselectivity of P450 3A4 substrates considers four metabolic reaction types, which are aliphatic and aromatic hydroxylation, N- and O-dealkylation [338]. 61 data points were considered, 51 thereof were used to develop a pharmacophore model and 10 were used as a test set. Geometric accessibility of any atoms of a molecule was encoded by its solvent accessible surface area (SASA) and metabolic reactivity was represented by calculated reaction energetics which was calculated using activation energy, electronegativity equalization sigma charge and sigma Fukui function, as well as Huckel pi charge and pi Fukui function. A RF model was generated, which obtained 82% correct predictions on the test set.

The general conclusion is that the tools available to predict SOM provide users with useful insight into the potential sites of metabolism. This ability can be useful to rationalize results from experimental metabolic assays or in the design of new molecules with reduced metabolic liability. At present the methods discussed here are not sufficiently accurate to be used to predict whether a molecule is likely to suffer significant metabolism or not.

3.5. Skin Sensitization

For the many different types of toxicity known, the underlying mechanism for skin sensitization is one of the best understood. It is generally believed that it is a result of the formation of chemical adducts formed by the reaction of unknown proteins and the chemical in question [47]. The skin sensitization potential of chemicals can be assessed using *in vivo* methods and the mouse local lymph

node assay (LLNA) is the method of first choice at present [84]. The LLNA assay has been validated as an alternative to the guinea pig models for the identification of potential skin sensitizers. A substance is classified as a sensitizer if it induces a threefold stimulation index (EC3) or greater at one or more test concentrations. As with all biological endpoints, variability within skin-sensitization assays, and between different assay types is not insignificant [85]. A more recent, novel approach to assess sensitization potential *in vitro* involves measuring the chemical reactivity of molecules with small chain peptides as a protein mimic [44, 343].

The ability of a chemical to react covalently with carrier protein nucleophiles relates to electrophilicity, and likely shape and hydrophobicity, given protein binding is required. A recent published database of more than 200 chemicals tested in the LLNA assay has been examined with regards to various chemical reaction domains known to be associated with sensitization [344]. LLNA based datasets are relatively small compared to other endpoints (e.g. P450, hERG inhibition or genotoxicity) which use higher-throughput assays.

However, the development of reliable alternative approaches to assess skin sensitization, such as QSARs, will require additional testing of a broad range of chemicals, covering the major chemical mechanisms for skin sensitization, as well as an appropriate balance between confirmed skin sensitizers and non-sensitizers. As discussed above the use of categorical models (i.e. predictions of sensitizers vs non sensitizers) will be of limited use in the the risk assessment of compounds unless an exposure-response relationship has also been studied. This is necessary if reliable non-animal approaches, such as those based on structure-activity relationship (SAR), quantitative structure-activity relationships (QSAR), read-across, quantitative mechanistic modeling (QMM) and experimental chemistry based models, are to be developed. For a much more in-depth review of this area see reference [50].

Modeling Studies

A hybrid QSAR study by Dimitrov on skin metabolism and formation of protein conjugation was able to correctly classify about 80% of the chemicals with significant sensitizing effect and 72% of non-sensitizing chemicals [345]. QSAR analyses of 16 Schiff base compounds with LLNA measurements were undertaken by Aptula *et al.* [346]. A good model could be generated to predict the EC3 value using just the Taft parameter, a model electrophilicity, and the logarithm of the partition coefficient (log P). The presence of an aliphatic carbonyl group was common to all compounds in the set, suggesting the model has a limited applicability domain.

Yuan *et al.* [347] recently modelled the skin sensitization potential of a larger set of 162 compounds with LLNA measurements, and 92 compounds with guinea pig maximization test (GPMT) measurements, using a particle swarm optimized Support Vector Machine. The particle swarm optimization algorithm was implemented for feature selection from a large number of molecular descriptors. The classification accuracies (sensitizer vs non-sensitizer) reported were 95.37% and 88.89% for the training and the test sets, respectively. For the GPMT data set, the classification accuracies were 91.80% and 90.32% for the training and the test sets, respectively. Chaudhry *et al.* reported the development of two global QSAR models using a 209 compounds dataset for skin sensitization using two different computational approaches: Adaptive Fuzzy Partition (AFP) and Neural Network (NN) models. The best model shows classification success rates of 84 and 71% for the training and test set respectively [49].

Enoch *et al.* [24] adopted a different approach in their study, focusing on the development of an expert system/applicability domain. They reported a comprehensive series of SMARTS (Smiles Arbitrary Target Specification) [141] patterns capable of classifying 210 chemicals from the LLNA assay database into potential mechanisms of action classes, which were originally assigned based on

on expert knowledge [344]. The results showed that the SMARTS patterns provided an excellent method of identifying potential electrophilic mechanisms, in a rapid manner. Given the small dataset size it is probable that the rules will need to reviewed and updated as new compounds are identified and potential new mechanisms/substructures are identified. Also important is a consideration of the domain of applicability of such models, which has been discussed recently by Ellison *et al.* [348].

An alternative strategy to predict skin sensitization EC3 values of the LLNA assay uses a mechanism-based read across approach. For example, 40 compounds were tested in the LLNA assay [349] and previous determined results, in conjunction with a read-across approach, led to a good classification of the newly tested compounds. The prediction concordance was found to be 83%. In addition, Enoch *et al.* assessed the utility of read-across for 19 Michael acceptor chemicals in conjunction with an electrophilicity index. These were classified in good agreement with their experimentally determined values [350].

A substantial research program was initiated by Unilever, to critically evaluate the feasibility of a new conceptual approach for consumer safety risk assessment [351], including a computational model of skin sensitization using system biology [352]. Insights from these modelling exercises led to the focus of subsequent non-animal test development upon the identified toxicity pathways, namely skin bioavailability [353], protein binding [343], skin inflammation/Dendritic cell (DC) maturation and T cell proliferation.

4. CONCLUSIONS

There still exists a need for new methods to rapidly and accurately determine the toxic potential of both drug molecules and molecules contained in consumer product goods. *In silico* toxicology models, such as those discussed in this review, fit many of these criteria, and have seen widespread use in drug discovery applications. Nevertheless, there is a general consensus in the literature that *in silico* toxicology models are not presently capable of accurately and reliably predicting the toxicological fate of molecules [34]. This is perhaps not surprising given that researchers also question the level predictive ability of *in vitro* or *in chemico* methods to predict toxicity in man [48, 72, 73, 76, 81]. While *in silico* predictions are not sufficiently accurate to replace experimentally based assays, they can be used to filter large virtual libraries prior to synthesis or prioritize compounds in drug discovery programs for risk assessment at the very beginning of a program. A compound that is predicted toxic *in silico* can be prioritized for subsequent *in vitro* (or *in vivo*) screening. If it is confirmed to be toxic, steps can be taken to remove the liability, be it a particular functional group, or some generic property of the compound in question. Alternatively, models can be used purely in the post-rationalization of experimental results, and to suggest ways to remove the liability in question.

Another aspect requiring consideration is that typical *in silico* toxicology models (and many *in vitro* methods for that matter) do not take dose or exposure into account. Therefore, these models cannot predict the absolute level of toxicity, but may still provide useful supplementary information for the overall risk assessment process. Therefore, whenever possible, internal exposure should be taken into account by use of either *in silico* or *in vitro* ADME data (i.e. the amount taken up and distributed as free plasma concentration within an organism). There is also a need for these computational chemistry tools to align with other information sources (for example from systems biology, metabolite information and exposure) to facilitate the development of virtual models of tissues, organs and physiological processes that could be used for the toxicological assessments. Understanding how the different mechanisms of action of chemicals can affect biological structures, processes and pathways, and thus how they impact on physiological responses is an important aspect of toxicology. This knowledge could help

improve our ability to predict the level of toxicity of chemicals, and to develop strategies to prevent exposure to toxic compounds, or ways to minimize their effects. Presently, both the complexity of the biological response for a given toxicity, and the lack of publicly available mechanistic data are key reasons why *in silico* approaches to date have had limited success in delivering accurate *in vivo* relevant predictions.

In view of these issues, a number of projects have undertaken by several groups aimed at delivering improvements in the modeling of *in vivo* human toxicity based on data available during early stages of the innovation pipeline. These include (a) the European Commission's seventh framework joint research programmes (FP7) [354]; (b) the Innovative Medicines initiative IMI, funded by with European Federation of pharmaceutical industries and association (EFPIA) [355]; and (C) and the Safety Evaluation Ultimately Replacing Animal Testing (SEURAT1 project) funded with European Cosmetic, Toiletry and Perfumery Association (Colipa) [356]. Furthermore mechanistic understanding of human toxicity forms [357] a central component of the National Research Council (NRC) vision whose roadmap is described in "Toxicity Testing in the 21st Century (TT21C): A Vision and a Strategy" [358]. This vision is summarized as follows: "Advances in toxicogenomics, bioinformatics, systems biology, epigenetics, and computational toxicology could transform toxicity testing from a system based on whole-animal testing to one founded primarily on *in vitro* methods that evaluate changes in biologic processes using cells, cell lines, or cellular components, preferably of human origin." The National Toxicology program (NTP) High Throughput Screening Initiative and the EPA Toxcast programme [359] are two efforts that aim to utilise the technological advances in molecular biology and computational science.

We conclude that, while *in silico* toxicology models are valuable tools for drug discovery scientists, much still needs to be done to firstly, understand more completely the biological mechanisms for toxicity, and secondly, to generate more rapid *in vitro* models to screen compounds. With this biological understanding and additional data available, our ability to generate more predictive *in silico* models should significantly improve in the future.

5. ACKNOWLEDGEMENTS

M.P. Gleeson would like to acknowledge financial support from The Thailand Research Fund (TRF) (RSA5480016) and from the Faculty of Science at Kasetsart University. R.L. Marchese Robinson acknowledges financial support from the Engineering and Physical Sciences Research Council (EPSRC) and Unilever. J. Kirchmair and A. Bender thank Unilever for support. S. Hannongbua would like to acknowledge financial support provided by the TRF (RSA5380010), and the National Center of Excellence in Petroleum, Petrochemicals and Advanced Materials for providing research facilities.

REFERENCES

- [1] Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 2004; 3: 711-16.
- [2] Riley RJ, Martin IJ, Cooper AE. The Influence of DMPK as an Integrated Partner in Modern Drug Discovery. *Curr Drug Metab* 2002; 3: 527-50.
- [3] Hop CE, Cole MJ, Davidson RE, *et al.* High throughput ADME screening: practical considerations, impact on the portfolio and enabler of *in silico* ADME models. *Curr Drug Metab* 2008; 9: 847-53.
- [4] Kramer JA, Sagartz JE, Morris DL. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat Rev Drug Discov* 2007; 6: 636-49.
- [5] Smith DA, Schmid EF. Drug withdrawals and the lessons within. *Curr Opin Drug Discov Devel* 2006; 9: 38-46.
- [6] Krewski D, Andersen ME, Mantus E, Zeise L. Toxicity testing in the 21st century: implications for human health risk assessment. *Risk Anal* 2009; 29: 474-9.
- [7] Shukla SJ, Huang R, Austin CP, Xia M. The future of toxicity testing: a focus on *in vitro* methods using a quantitative high-throughput screening platform. *Drug Discov Today* 2010; 15: 997-1007.
- [8] Mortelmans K, Zeiger E. The Ames Salmonella/microsome mutagenicity assay. *Mutat Res* 2000; 455: 29-60.
- [9] Kimber I, Hilton J, Botham PA. Identification of contact allergens using the murine local lymph node assay: comparisons with the Buehler occluded patch test in guinea pigs. *J Appl Toxicol* 1990; 10: 173-80.
- [10] Kimber I, Hilton J, Weisenberger C. The murine local lymph node assay for identification of contact allergens: a preliminary evaluation of *in situ* measurement of lymphocyte proliferation. *Contact Dermatitis* 1989; 21: 215-20.
- [11] Kimber I, Weisenberger C. A murine local lymph node assay for the identification of contact allergens. Assay development and results of an initial validation study. *Arch Toxicol* 1989; 63: 274-82.
- [12] Jamieson C, Moir EM, Rankovic Z, Wishart G. Medicinal chemistry of hERG optimizations: Highlights and hang-ups. *J Med Chem* 2006; 49: 5029-46.
- [13] O'Brien SE, de Groot MJ. Greater Than the Sum of Its Parts: Combining Models for Useful ADMET Prediction. *J Med Chem* 2005; 48: 1287-91.
- [14] Gleeson MP. Generation of a set of simple, interpretable ADMET rules of thumb. *J Med Chem* 2008; 51: 817-34.
- [15] TOXNET [homepage on the internet]. United States National Library of Medicine; [updated 2011 Oct 17; cited 2011 Oct 17]. Available from <http://toxnet.nlm.nih.gov/>
- [16] ICCVAM. LLNA Database [homepage on the internet]. National Toxicology Program; [updated 2011 Aug 08; cited 2011 Oct 17]. <http://iccvam.niehs.nih.gov/methods/immunotox/rLLNA.htm>
- [17] Saiakhov RD, Klopman G. MultiCASE expert systems and the REACH initiative. *Toxicol Mechanisms Methods* 2008; 18: 159-75.
- [18] OECD Quantitative Structure-Activity Relationships Project [(Q)SARs] [homepage on the internet] Organisation for Economic Co-operation and Development; c2011 [updated 2011 Oct 17; cited 2011 Oct 17]. Available from http://www.oecd.org/document/23/0,3746,en_2649_34379_33957015_1_1_1_1,00.html
- [19] Nonoyama T, Fukuda R. Drug-induced Phospholipidosis - Pathological Aspects and Its Prediction. *J Toxicol Pathol* 2008; 21: 9-24.
- [20] Lowe R, Glen RC, Mitchell JBO. Predicting Phospholipidosis Using Machine Learning. *Molecular Pharmaceutics* 2010; 7: 1708-14.
- [21] 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.
- [22] Li H, Sun J, Fan X, *et al.* Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J Comput Aided Mol Des* 2008; 22: 843-55.
- [23] Michielan L, Moro S. Pharmaceutical perspectives of nonlinear QSAR strategies. *J Chem Inf Model* 2010; 50: 961-78.
- [24] Enoch SJ, Madden JC, Cronin MT. Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach. *SAR QSAR Environ Res* 2008; 19: 555-78.
- [25] Benigni R, Bossa C. Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutat Res* 2008; 659: 248-61.
- [26] Devillers J, Mombelli E, Samsera R. Structural alerts for estimating the carcinogenicity of pesticides and biocides. *SAR QSAR Environ Res* 2011; 22: 89-106.
- [27] ECB REACH Guidelines [homepage on the internet]. European Commission; [updated 2011 Sept 28; cited 2011 Oct 17]. Available from http://ec.europa.eu/environment/chemicals/reach/reviews_en.htm
- [28] Schaafsma G, Kroese ED, Tieleman ELJP, Van de Sandt JJM, Van Leeuwen CJ. REACH, non-testing approaches and the urgent need for a change in mind set. *Reg Toxicol and Pharmacol* 2009; 53: 70-80.
- [29] Valerio LG, Jr. *In silico* toxicology for the pharmaceutical sciences. *Toxicol Appl Pharmacol* 2009; 241: 356-70.

- [30] Muster W, Breidenbach A, Fischer H, Kirchner S, Muller L, Pahler A. Computational toxicology in drug development. *Drug Discov Today* 2008; 13: 303-10.
- [31] Tsakovska I, Lessigiarska I, Netzeva T, Worth AP. A Mini Review of Mammalian Toxicity (Q)SAR Models. *QSAR Comb Sci* 2008; 27: 41-48.
- [32] Boyer S. The use of computer models in pharmaceutical safety evaluation. *Altern. Lab. Anim.* 2009; 37: 467-75.
- [33] Vedani A, Smiesko M. In silico toxicology in drug discovery - concepts based on three-dimensional models. *Altern Lab Anim* 2009; 37: 477-96.
- [34] Egan WJ, Zlokarnik G, Grootenhuis PDJ. In silico prediction of drug safety: despite progress there is abundant room for improvement. *Drug Discov Today: Technologies* 2004; 1: 381-7.
- [35] Vedani A, Dobler M, Lill MA. The challenge of predicting drug toxicity in silico. *Basic Clin Pharmacol Toxicol* 2006; 99: 195-208.
- [36] Merlot C. Computational toxicology--a tool for early safety evaluation. *Drug Discov Today* 2010; 15: 16-22.
- [37] Cronin MTD, Ellison CM, Madden JC, Judson P. Using In Silico Tools in a Weight of Evidence Approach to Aid Toxicological Assessment. *Mol Informatics* 2010; 29: 97-110.
- [38] Helma C. In silico predictive toxicology: The state-of-the-art and strategies to predict human health effects. *Curr Opin Drug Discov Dev* 2005; 8: 27-31.
- [39] Scheiber J, Chen B, Milik M, *et al.* Gaining Insight into Off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis. *J Chem Inf and Model* 2009; 49: 308-17.
- [40] Scheiber J, Jenkins JL, Sukuru SC, *et al.* Mapping adverse drug reactions in chemical space. *J Med Chem* 2009; 52: 3103-7.
- [41] Hillebrecht A, Muster W, Brigo A, Kansy M, Weiser T, Singer T. Comparative Evaluation of in Silico Systems for Ames Test Mutagenicity Prediction: Scope and Limitations. *Chem Res Toxicol* 2011.
- [42] Benigni R, Bossa C. Predictivity and reliability of QSAR models: the case of mutagens and carcinogens. *Toxicol Mech Methods* 2008; 18: 137-47.
- [43] Benigni R, Bossa C. Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. *Chem Rev* 2011; 111: 2507-36.
- [44] Enoch SJ, Cronin MT. A review of the electrophilic reaction chemistry involved in covalent DNA binding. *Crit Rev Toxicol* 2010; 40: 728-48.
- [45] Benigni R, Bossa C. Alternative strategies for carcinogenicity assessment: an efficient and simplified approach based on *in vitro* mutagenicity and cell transformation assays. *Mutagenesis* 2011; 3: 455-60.
- [46] Benfenati E, Benigni R, Demarini DM, Helma C, Kirkland D, Martin TM, Mazzatorta P, Ouedraogo-Arras G, Richard AM, Schilter B, Schoonen WG, Snyder RD, Yang C. Predictive models for carcinogenicity and mutagenicity: frameworks, state-of-the-art, and perspectives. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 2009; 27: 57-90.
- [47] Aptula AO, Roberts DW. Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity. *Chem Res Toxicol* 2006; 19: 1097-105.
- [48] Lalko JF, Kimber I, Dearman RJ, Gerberick GF, Sarlo K, Api AM. Chemical reactivity measurements: potential for characterization of respiratory chemical allergens. *Toxicol In vitro* 2011; 25: 433-45.
- [49] Chaudhry Q, Piclin N, Cotterill J, Pintore M, Price NR, Chretien JR, Roncaglioni A. Global QSAR models of skin sensitizers for regulatory purposes. *Chem Cent J* 2010; 4 Suppl 1: S5.
- [50] Schwobel JAH, Koleva YK, *et al.* Measurement and Estimation of Electrophilic Reactivity for Predictive Toxicology. *Chem Rev* 2011; 111: 2562-96.
- [51] Raschi E, Ceccarini L, De Ponti F, Recanatini M. hERG-related drug toxicity and models for predicting hERG liability and QT prolongation. *Expert Opin Drug Metab Toxicol* 2009; 5: 1005-21.
- [52] Aronov AM. Ligand structural aspects of hERG channel blockade. *Curr Top Med Chem* 2008; 8: 1113-27.
- [53] Aronov AM. Predictive in silico modeling for hERG channel blockers. *Drug Discov Today* 2005; 10: 149-55.
- [54] Matthews EJ, Frid AA. Prediction of drug-related cardiac adverse effects in humans-B: Use of QSAR programs for early detection of drug-induced cardiac toxicities. *Reg Toxicol and Pharmacol* 2010; 56: 276-89.
- [55] Foti RS, Wienkers LC, Wahlstrom JL. Application of cytochrome P450 drug interaction screening in drug discovery. *Comb Chem High Throughput Screen* 2010; 13: 145-58.
- [56] Refsgaard HHF, Jensen BF, Christensen IT, Hagen N, Brockhoff PB. In Silico Prediction of Cytochrome P450 Inhibitors. *Drug Devel Res* 2006; 67: 417-29.
- [57] Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010; 26: 246-54.
- [58] Zhang L, Zhang YD, Zhao P, Huang SM. Predicting drug-drug interactions: an FDA perspective. *AAPS J* 2009; 11: 300-6.
- [59] Stjernschantz E, Vermeulen NP, Oostenbrink C. Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. *Expert Opin Drug Metab Toxicol* 2008; 4: 513-27.
- [60] Roy K, Roy PP. QSAR of cytochrome inhibitors. *Expert Opin Drug Metab Toxicol* 2009; 5: 1245-66.
- [61] Tarceay A, Keseru GM. In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin Drug Metab Toxicol* 2011; 7: 299-312.
- [62] Madden JC, Cronin MT. Structure-based methods for the prediction of drug metabolism. *Expert Opin Drug Metab Toxicol* 2006; 2: 545-57.
- [63] Ekins S, Andreyev S, Ryabov A, *et al.* Computational prediction of human drug metabolism. *Expert Opin Drug Metab Toxicol* 2005; 1: 303-24.
- [64] Marchant CA, Briggs KA, Long A. In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic. *Toxicol Mech Methods* 2008; 18: 177-87.
- [65] Sun H, Scott DO. Structure-based drug metabolism predictions for drug design. *Chem Biol Drug Des* 2010; 75: 3-17.
- [66] Furge LL, Guengerich FP. Cytochrome P450 enzymes in drug metabolism and chemical toxicology: An introduction. *Biochem Mol Biol Educ* 2006; 34: 66-74.
- [67] Guengerich FP. Cytochrome p450 and chemical toxicology. *Chem Res Toxicol* 2008; 21: 70-83.
- [68] Boyer S, Zamora I. New methods in predictive metabolism. *J Comput Aided Mol Des* 2002; 16: 403-13.
- [69] Piparo L, Worth AP. Review of QSAR Models and Software Tools for predicting Developmental and Reproductive Toxicity. *JRC Report* 2011. Models and Software Tools for predicting Developmental and Reproductive Toxicity. Accessed October 7 2010.
- [70] Hewitt M, Ellison CM, Enoch SJ, Madden JC, Cronin MT. Integrating (Q)SAR models, expert systems and read-across approaches for the prediction of developmental toxicity. *Reprod Toxicol* 2010; 30: 147-60.
- [71] Scialli AR. The challenge of reproductive and developmental toxicology under REACH. *Regul Toxicol Pharmacol* 2008; 51: 244-50.
- [72] McKim JM. Building a Tiered Approach to *In vitro* Predictive Toxicity Screening: A Focus on Assays with *In vivo* Relevance. *Comb Chem & High Throughput Screen* 2010; 13: 188-206.
- [73] Kremers P. Can Drug-Drug Interactions Be Predicted from *In vitro* Studies? . *The Scientific World Journal* 2002; 2: 751-66.
- [74] Adler S, Basketter D, Creton S, *et al.* Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. *Arch Toxicol* 2011; 85: 367-485.
- [75] Lagrutta AA, Trepakova ES, Salata JJ. The hERG channel and risk of drug-acquired cardiac arrhythmia: an overview. *Curr Top Med Chem* 2008; 8: 1102-12.
- [76] Kalgutkar AS, Didiuk MT. Structural Alerts, Reactive Metabolites, and Protein Covalent Binding: How Reliable Are These Attributes as Predictors of Drug Toxicity? *Chem Biodiversity* 2009; 6: 2115-37.
- [77] Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. Timing of New Black Box Warnings and Withdrawals for Prescription Medications. *AMA, J Am Med Assoc* 2002; 287: 2215-20.
- [78] Thompson RA, Isin EM, Li Y, *et al.* Risk assessment and mitigation strategies for reactive metabolites in drug discovery and development. *Chem Biol Interact* 2011; 192: 65-71.

- [79] Waring JF, Anderson MG. Idiosyncratic toxicity: mechanistic insights gained from analysis of prior compounds. *Curr Opin Drug Discov Devel* 2005; 8: 59-65.
- [80] Gola J, Obrezanova O, Champness E, Segall M. ADMET Property Prediction: The State of the Art and Current Challenges. *QSAR Comb Sci* 2006; 25: 1172-80.
- [81] Snyder RD, Smith MD. Computational prediction of genotoxicity: room for improvement. *Drug Discov Today* 2005; 10: 1119-24.
- [82] Kazius J, McGuire R, Bursi R. Derivation and validation of toxicophores for mutagenicity prediction. *J Med Chem* 2005; 48: 312-20.
- [83] Benigni R, Bossa C, Tcheremenskaia O, Giuliani A. Alternatives to the carcinogenicity bioassay: in silico methods, and the *in vitro* and *in vivo* mutagenicity assays. *Expert Opin Drug Metabol Toxicol* 2010; 6: 809-19.
- [84] Ikarashi Y, Tsuchiya T, Nakamura A. Detection of contact sensitivity of metal salts using the murine local lymph node assay. *Toxicol Lett* 1992; 62: 53-61.
- [85] Basketter DA, Kimber I. Skin sensitization, false positives and false negatives: experience with guinea pig assays. *J Applied Toxicol* 2010; 30: 381-86.
- [86] Thorne N, Auld DS, Ingles J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr Opin Chem Biol* 2010; 14: 315-24.
- [87] Di L, Kerns EH. Biological assay challenges from compound solubility: strategies for bioassay optimization. *Drug Discov Today* 2006; 11: 446-51.
- [88] Lane SJ, Eggleston DS, Brinded KA, Hollerton JC, Taylor NL, Readshaw SA. Defining and maintaining a high quality screening collection: the GSK experience. *Drug Discov Today* 2006; 11: 267-72.
- [89] Young D, Martin T, Venkatapathy R, Harten P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb Sci* 2008; 27: 1337-45.
- [90] Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J Chem Inf and Model* 2010; 50: 1189-204.
- [91] Schultz TW, Rogers K, Aptula AO. Read-across to rank skin sensitization potential: subcategories for the Michael acceptor domain. *Contact Dermatitis* 2009; 60: 21-31.
- [92] Patlewicz G, Dimitrov SD, Low LK, *et al.* TIMES-SS-A promising tool for the assessment of skin sensitization hazard. A characterization with respect to the OECD validation principles for (Q)SARs and an external evaluation for predictivity. *Reg Toxicol Pharmacol* 2007; 48: 225-39.
- [93] Bathelt CM, Mulholland AJ, Harvey JN. QM/MM modeling of benzene hydroxylation in human cytochrome P450 2C9. *J Phys Chem A* 2008; 112: 13149-56.
- [94] Warren GL, Andrews CW, Capelli AM, *et al.* A critical assessment of docking programs and scoring functions. *J Med Chem* 2006; 49: 5912-31.
- [95] Leach AR, Shoichet BK, Peishoff CE. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J Med Chem* 2006; 49: 5851-5.
- [96] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 1988; 110: 5959-67.
- [97] Leach AR. *Molecular Modelling. Principles and Applications.* Harlow: Longman 1996.
- [98] Shaik S, Cohen S, Wang Y, Chen H, Kumar D, Thiel W. P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations. *Chemrev* 2010; 110: 949-1017.
- [99] Greene N, Judson PN, Langowski JJ, Marchant CA. Knowledge-Based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR and METEOR. *SAR QSAR Env Res* 1999; 10: 299-314.
- [100] Wassermann AM, Bajorath Jr. Chemical Substitutions That Introduce Activity Cliffs Across Different Compound Classes and Biological Targets. *J Chem Inf Mod* 2010; 50: 1248-56.
- [101] Cruciani G, Carosati E, De Boeck B, *et al.* MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J Med ChemMed Chem* 2005; 48: 6970-9.
- [102] Todeschini R, Lasagni M, Marengo E. New molecular descriptors for 2D and 3D structures. *Theory J Chemometr J Chemometr* 1994; 8: 263-72.
- [103] Xue L, Bajorath J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Comb Chem & High Throughput Screen* 2000; 3: 363-72.
- [104] Labute P. A widely applicable set of descriptors. *J Mol Graph Model* 2000; 18: 464-77.
- [105] Hong H, Xie Q, Ge W, *et al.* Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* 2008; 48: 1337-44.
- [106] Randic M. Novel shape descriptors for molecular graphs. *J Chem Inf Comput Sci* 2001; 41: 607-13.
- [107] Mansfield ML, Covell DG, Jernigan RL. A new class of molecular shape descriptors. 1. Theory and properties. *J Chem Inf Comput Sci* 2002; 42: 259-73.
- [108] Arteca GA. *Molecular Shape Descriptors. Reviews in Computational Chemistry.* John Wiley & Sons, Hoboken NJ 2007.
- [109] Todeschini R, Consonni V. *Molecular descriptors for chemoinformatics,* Wiley VCH Verlag GmbH, Weinheim 2009
- [110] Todeschini R, Lasagni M, Marengo E. New molecular descriptors for 2D and 3D structures. *J Chemom.* 1994; 8: 263-72.
- [111] Xue L, Bajorath J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High-Throughput Screen.* 2000; 3: 363-72.
- [112] Hong H, Xie Q, Ge W, *et al.* Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem Inf. Model.* 2008; 48: 1337-44.
- [113] Bender A, Glen RC. Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem* 2004; 2: 3204-18.
- [114] Gillet VJ, Willett P, Bradshaw J. Similarity Searching Using Reduced Graphs. *J Chem Inf Comput Sci* 2003; 43: 338-45.
- [115] Willett P, Barnard JM, Downs GM. Chemical Similarity Searching. *J Chem Inf Comput Sci* 1998; 38: 983-96.
- [116] Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ. 4D-fingerprints, universal QSAR and QSPR descriptors. *J Chem Inf Comput Sci* 2004; 44: 1526-39.
- [117] Clark DE. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J Pharm Sci* 1999; 88: 815-21.
- [118] Friesner RA, Banks JL, Murphy RB, *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* 2004; 47: 1739-49.
- [119] Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 1997; 267: 727-48.
- [120] Bender A, Jenkins JL, Scheiber J, Sukuru SC, Glick M, Davies JW. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Inf Model* 2009; 49: 108-19.
- [121] Maget-Dana R, Peypoux F. Iturins, a special class of pore-forming lipopeptides: biological and physicochemical properties. *Toxicology* 1994; 87: 151-74.
- [122] Clark RD, Norinder U. Two personal perspectives on a key issue in contemporary 3D QSAR. *WIREs Comput Mol Sci* 2011; doi: 10.1002/wcms.69.
- [123] Nicholls A, McGaughey GB, Sheridan RP, *et al.* Molecular shape and medicinal chemistry: a perspective. *J Med Chem* 2010; 53: 3862-86.
- [124] Cavasotto CN, Orry AJ. Ligand docking and structure-based virtual screening in drug discovery. *Curr Top Med Chem* 2007; 7: 1006-14.
- [125] Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* 2007; 12: 225-33.
- [126] Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr Opin Chem Biol* 2006; 10: 194-202.
- [127] Bender A, Scheiber J, Glick M, *et al.* Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure *Chem Med Chem* 2007; 2: 861-73.
- [128] Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 2010; 53: 2719-40.
- [129] Papadatos G, Alkharouri M, Gillet VJ, *et al.* Lead optimization using matched molecular pairs: inclusion of contextual information for

- enhanced prediction of HERG inhibition, solubility, and lipophilicity. *J Chem Inf Model* 2010; 50: 1872-86.
- [130] Patlewicz G, Jeliaskova N, Gallegos Saliner A, Worth AP. Toxmatch-a new software tool to aid in the development and evaluation of chemically similar groups. *SAR QSAR Environ Res* 2008; 19: 397-412.
- [131] Gallegos-Saliner A, Poater A, Jeliaskova N, Patlewicz G, Worth AP. Toxmatch--A chemical classification and activity prediction tool based on similarity measures. *Reg Toxicol and Pharmacol* 2008; 52: 77-84.
- [132] Sanderson DM, Earnshaw CG. Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum Exp Toxicol* 1991; 10: 261-73.
- [133] Ridings JE, Barratt MD, Cary R, et al. Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. *Toxicology* 1996; 106: 267-79.
- [134] The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree 2008, [pdf on the internet]. European Commission Joint Research Centre, Ispra; c2008 [updated 2008; cited 2011 Oct 17]. Available from http://ihcp.jrc.ec.europa.eu/our_labs/computational_toxicology/doc/EUR_23241_EN.pdf.
- [135] Patlewicz G, Jeliaskova N, Safford RJ, Worth AP, Aleksiev B. An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *Sar and Qsar in Environmental Research* 2008; 19: 495-524.
- [136] Verhaar HJM, C. VL, JLMH. Classifying environmental pollutants. 1: Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere* 1992; 25: 471-91.
- [137] Gerner I, Schlegel K, Walker JD, Hulzebos E. Use of Physicochemical Property Limits to Develop Rules for Identifying Chemical Substances with no Skin Irritation or Corrosion Potential. *QSAR Comb Sci* 2004; 23: 726-33.
- [138] Gerner I, Liebsch M, Spielmann H. Assessment of the eye irritating properties of chemicals by applying alternatives to the Draize rabbit eye test: The use of QSARs and *in vitro* tests for the classification of eye irritation. *Atla-Alternatives to Laboratory Animals* 2005; 33: 215-37.
- [139] Olsen L, Rydberg P, Gloriam DE, Zaretski J, Breneman C. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *Acs Medicinal Chemistry Letters* 2010; 1: 96-100.
- [140] Schultz TW, Yarbrough JW, Hunter RS, Aptula AO. Verification of the structural alerts for Michael acceptors. *Chem Res Toxicol* 2007; 20: 1359-63.
- [141] Daylight Fingerprints [homepage on the internet]: Daylight Chemical Information Laguna Niquel CA, [updated 2008; cited 2011 Oct 17]. Available from www.daylight.com
- [142] Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Comput Sci* 2002; 42: 1273-80.
- [143] Mussa HY, Hawizy L, Nigsch F, Glen RC. Classifying Large Chemical Data Sets: Using A Regularized Potential Function Method. *J Chem Inf and Model* 2011; 51: 4-14.
- [144] Gleeson MP, Hersey A, Hannongbua S. In-silico ADME models: a general assessment of their utility in drug discovery applications. *Curr Top Med Chem* 2011; 11: 358-81.
- [145] Madden JC. In Silico Approaches for Predicting Adme Properties. *Recent Advances in QSAR Studies*. Springer Netherlands 2010.
- [146] Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* 2010; 29: 476-88.
- [147] Scior T, Medina-Franco JL, Do QT, Martinez-Mayorga K, Yunes Rojas JA, Bernard P. How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr Med Chem* 2009; 16: 4297-313.
- [148] Dragos H, Gilles M, Alexandre V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J Chem Inf and Model* 2009; 49: 1762-76.
- [149] Benigni R, Bossa C. Predictivity of QSAR. *J Chem Inf Model* 2008; 48: 971-80.
- [150] Hansch C, Leo A, Mekapati SB, Kurup A. QSAR and ADME. *Bioorg Med Chem* 2004; 12: 3391-400.
- [151] Lombardo F, Gifford E, Shalaeva MY. In silico ADME prediction: data, models, facts and myths. *Mini Rev Med Chem* 2003; 3: 861-75.
- [152] Dearden JC. In silico prediction of drug toxicity. *J Comput Aided Mol Des* 2003; 17: 119-27.
- [153] Gavaghan CL, Arnby CH, Blomberg N, Strandlund G, Boyer S. Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J Comput-aided Mol Des* 2007; 21: 189-206.
- [154] Xu Y-j, Gao H. Dimension related distance and its application in QSAR/QSPR model error estimation. *QSAR Comb Sci* 2003; 22: 422-9.
- [155] Tetko IV, Bruneau P, Mewes H-W, Rohrer DC, Poda GI. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov Today* 2006; 11: 700-7.
- [156] Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model* 2008; 26: 1315-26.
- [157] Bruneau P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J Chem Inf Comput Sci* 2001; 41: 1605-16.
- [158] Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J Chem Inf Comput Sci* 2004; 44: 1912-28.
- [159] Helgee EA, Carlsson L, Boyer S, Norinder U. Evaluation of Quantitative Structure-Activity Relationship Modeling Strategies: Local and Global Models. *J Chem Inf and Model* 2010; 50: 677-89.
- [160] Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci* 2004; 44: 1-12.
- [161] Marchese Robinson RL, Glen RC, Mitchell JBO. Development and Comparison of hERG Blocker Classifiers: Assessment on Different Datasets Yields Markedly Different Results. *Mol Informatics* 2011; 30: 443-58.
- [162] Golbraikh A, Tropsha A. Beware of q²! *J Mol Graph Model* 2002; 20: 269-76.
- [163] Johnson SR. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J Chem Inf Model* 2008; 48: 25-6.
- [164] Rucker C, Rucker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 2007; 47: 2345-57.
- [165] Rodgers SL, Davis AM, van de Waterbeemd H. Time-Series QSAR Analysis of Human Plasma Protein Binding Data. *QSAR Comb Sci* 2007; 26: 511-21.
- [166] Rodgers SL, Davis AM, Tomkinson NP. QSAR modeling using automatically updating correction libraries: Application to a human plasma protein binding model. *J Chem Inf and Model* 2007; 47: 2401-07.
- [167] Fukunaga K. Introduction to Statistical Pattern Recognition. Academic Press: United Kingdom 1972.
- [168] Jensen F. Introduction to Computational Chemistry. 2nd ed. WILEY: Chichester, England 2007.
- [169] Bishop C. Pattern Recognition and Machine Learning. Springer Science and Business Media: New York, USA 2006.
- [170] Hansch C, Quinlan JE, Lawrence GL. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J Org Chem* 1968; 33: 347-50.
- [171] Hansch C, Fujita T. ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J Am Chem Soc* 1964; 86: 1616-26.
- [172] Hansch C, Fujita T. Additions and Corrections - ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J Am Chem Soc* 1964; 86: 5710-10.
- [173] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 2003; 43: 1947-58.
- [174] Clark M, Wiseman JS. Fragment-Based Prediction of the Clinical Occurrence of Long QT Syndrome and Torsade de Pointes. *J Chem Inf and Model* 2009; 49: 2617-26.
- [175] Aberg KM, Jacobsson SP. On the separation of classes: can the Fisher criterion be improved upon when classes have unequal variance-covariance structure? *J Chemometr J Chemometr* 2010; 24: 650-54.
- [176] Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 1936; 7: 179-88.
- [177] Doddareddy MR, Klaasse EC, Shagufta, Ijzerman AP, Bender A. Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases. *ChemMedChem* 2010; 5: 716-29.

- [178] Nigsch F, Bender A, Jenkins JL, Mitchell JBO. Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J Chem Inf and Model* 2008; 48: 2313-25.
- [179] Hewitt M, Cronin MT, Enoch SJ, Madden JC, Roberts DW, Dearden JC. In silico prediction of aqueous solubility: the solubility challenge. *J Chem Inf Model* 2009; 49: 2572-87.
- [180] Worth AP, Lapenna S, Piparo EL, Mostrag-Szlichtyng A, Serafimova R. A Framework for assessing in silico Toxicity Predictions: Case Studies with selected Pesticides [pdf on the internet]: European Commission Joint Research Centre, Ispra; c2011 [updated 2011; cited 2011 Oct 17]. Available from http://ihcp.jrc.ec.europa.eu/our_labs/computational_toxicology/doc/EUR_24705_EN.pdf
- [181] Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychol Methods* 2009; 14: 332-48.
- [182] Menze BH, Kelm BM, Masuch R, *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 2009; 10: 213.
- [183] Rusinko A, Farnen MW, Lambert CG, Brown PL, Young SS. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J Chem Inf Comput Sci* 1999; 39: 1017-26.
- [184] Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput Chem* 2001; 26: 5-14.
- [185] Dubus E, Ijjaali I, Petitot F, Michel A. In silico classification of hERG channel blockers: a knowledge-based strategy. *ChemMedChem* 2006; 1: 622-30.
- [186] Breiman L. Random forests. *Mach. Learn.* 2001; 45: 5-32.
- [187] Breiman LBRCHC. Classification and Regression Trees. Chapman & Hall/CRC: Boca Raton 1984.
- [188] Peterson KL. Artificial Neural Networks and Their Use in Chemistry In: Reviews in Computational Chemistry; Lipkowitz, K.B., Boyd, D.B., Eds.; Wiley-VCH: New York, 2000, Vol. 16, 53-128.
- [189] Thai KM, Ecker GF. Classification models for hERG inhibitors by counter-propagation neural networks. *Chem Biol Drug Des* 2008; 72: 279-89.
- [190] Cartwright H. Development and Uses of Artificial Intelligence in Chemistry In: Reviews in Computational Chemistry; Lipkowitz, K.B., Cudari T.R. Eds.; Wiley-VCH: New York, 2007, Vol. 25, 349-389.
- [191] Ivanciuc O. Applications of Support Vector Machines in Chemistry In: Reviews in Computational Chemistry; Lipkowitz, K.B., Cudari T.R. Eds.; Wiley-VCH: New York, 2006, Vol. 23, 291-400.
- [192] Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on* 2001; 12: 181-201.
- [193] Keerthi SS, Lin CJ. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput* 2003; 15: 1667-89.
- [194] Song MH, Clark M. Development and evaluation of an in silico model for hERG binding. *J Chem Inf and Model* 2006; 46: 392-400.
- [195] Zheng W, Tropsha A. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *J Chem Inf Comput Sci* 2000; 40: 185-94.
- [196] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000; 16: 412-24.
- [197] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 1975; 405: 442-51.
- [198] Cohen J. Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968; 70: 213-20.
- [199] Kononov Da, Llewellyn LE, Vander Heyden Y, Coomans D. Robust cross-validation of linear regression QSAR models. *J Chem Inf Model* 2008; 48: 2081-94.
- [200] Gleeson MP. Plasma protein binding affinity and its relationship to molecular structure: an in-silico analysis. *J Med Chem* 2007; 50: 101-12.
- [201] Gleeson MP, Waters NJ, Paine SW, Davis AM. In silico human and rat Vss quantitative structure-activity relationship models. *J Med Chem* 2006; 49: 1953-63.
- [202] Todeschini R, Consonni V, Pavan M. A distance measure between models: a tool for similarity/diversity analysis of model populations. *Chemometr Intell Lab* 2004; 70: 55-61.
- [203] Schultz TW, Hewitt M, Netzeva TI, Cronin MTD. Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action. *QSAR Comb Sci* 2007; 26: 238-54.
- [204] Bruneau P, McElroy NR. logD7.4 modeling using Bayesian Regularized Neural Networks. Assessment and correction of the errors of prediction. *J Chem Inf Model* 2006; 46: 1379-1387.
- [205] Sushko I, Novotarskyi S, Kormer R, *et al.* Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J Chem Inf and Model* 2010; 50: 2094-111.
- [206] Tetko IV, Sushko I, Pandey AK, *et al.* Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J Chem Inf and Model* 2008; 48: 1733-46.
- [207] Netzeva TI, Worth A, Aldenberg T, *et al.* Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Alternatives to laboratory animals : ATLA 2005; 33: 155-73.
- [208] Kuhne R, Ebert RU, Schuurmann G. Chemical Domain of QSAR Models from Atom-Centered Fragments. *J Chem Inf and Model* 2009; 49: 2660-69.
- [209] Vaz RJ, Li Y, Rampe D. Human ether-a-go-go related gene (HERG): a chemist's perspective. *Prog Med Chem* 2005; 43: 1-18.
- [210] Polak S, Wisniewska B, Brandys J. Collation, assessment and analysis of literature *in vitro* data on hERG receptor blocking potency for subsequent modeling of drugs' cardiotoxic properties. *J Applied Toxicol* 2009; 29: 183-206.
- [211] Bridgland-Taylor MH, Hargreaves AC, Easter A, *et al.* Optimisation and validation of a medium-throughput electrophysiology-based hERG assay using IonWorks(TM) HT. *J Pharmacol Toxicol* 2006; 54: 189-99.
- [212] Li Q, Jorgensen FS, Oprea T, Brunak S, Taboureau O. hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol Pharm* 2008; 5: 117-27.
- [213] Su BH, Shen MY, Esposito EX, Hopfinger AJ, Tseng YJ. In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. *J Chem Inf Model* 2010; 50: 1304-18.
- [214] Obrezanova O, Segall MD. Gaussian processes for classification: QSAR modeling of ADMET and target activity. *J Chem Inf Model* 2010; 50: 1053-61.
- [215] Sen S, Sinha N. Predicting hERG activities of compounds from their 3D structures: Development and evaluation of a global descriptors based QSAR model. *European J Med Chem* 2011; 46: 618-30.
- [216] Yao X, Anderson DL, Ross SA, *et al.* Predicting QT prolongation in humans during early drug development using hERG inhibition and an anaesthetized guinea-pig model. *Br J Pharmacol* 2008; 154: 1446-56.
- [217] Thai KM, Ecker GF. A binary QSAR model for classification of hERG potassium channel blockers. *Bioorg Med Chem* 2008; 16: 4107-19.
- [218] Chekmarev DS, Kholodovych V, Balakin KV, Ivanenkov Y, Ekins S, Welsh WJ. Shape signatures: new descriptors for predicting cardiotoxicity in silico. *Chem Res Toxicol* 2008; 21: 1304-14.
- [219] Thai KM, Ecker GF. Similarity-based SIBAR descriptors for classification of chemically diverse hERG blockers. *Mol Divers* 2009; 13: 321-36.
- [220] Ekins S, Crumb WJ, Sarazan RD, Wikel JH, Wrighton SA. Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *J Pharmacol Exp Ther* 2002; 301: 427-34.
- [221] Aronov AM. Common pharmacophores for uncharged human ether-a-go-go-related gene (hERG) blockers. *J Med Chem* 2006; 49: 6917-21.
- [222] Du LP, Li MY, Xia L, You QD. A novel structure-based virtual screening model for the hERG channel blockers. *Biochemical and Biophysical Research Communications* 2007; 355: 889-94.
- [223] Recanatini M, Cavalli A, Masetti M. Modeling hERG and its Interactions with Drugs: Recent Advances in Light of Current

- Potassium Channel Simulations. *Chem Med Chem* 2008; 3: 523-35.
- [224] Roche O, Trube G, Zuegge J, Pflimlin P, Alanine A, Schneider G. A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *Chem Bio Chem* 2002; 3: 455-9.
- [225] Hansen K, Rathke F, Schroeter T, *et al.* Bias-correction of regression models: a case study on hERG inhibition. *J Chem Inf Model* 2009; 49: 1486-96.
- [226] Demel MA, Ecker GF, Janacek AGK, Thai KM, Gansterer WN. Predictive QSAR models for polyspecific drug targets: The importance of feature selection. *Curr Comput Aided Drug Des* 2008; 4: 91-110.
- [227] Obrezanova O, Csanyi G, Gola JM, Segall MD. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J Chem Inf Model* 2007; 47: 1847-57.
- [228] Johnson SR, Yue HW, Conder ML, Shi H, Doweiko AM, Lloyd J, Levesque P. Estimation of hERG inhibition of drug candidates using multivariate property and pharmacophore SAR. *Bioorg Med Chem* 2007; 15: 6182-92.
- [229] Davenport AJ, Moller C, Heifetz A, Mazanetz MP, Law RJ, Ebneith A., Gemkow MJ. Using Electrophysiology and In Silico Three-Dimensional Modeling to Reduce Human Ether-a-go-go Related Gene K(+) Channel Inhibition in a Histamine H3 Receptor Antagonist Program. *Assay and Drug Development Technologies* 2010; 8: 781-9.
- [230] Gleeson P, Bravi G, Modi S, Lowe D. ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg Med Chem* 2009; 17: 5906-19.
- [231] Imai YN, Ryu S, Oiki S. Docking Model of Drug Binding to the Human Ether-a-go-go Potassium Channel Guided by Tandem Dimer Mutant Patch-Clamp Data: A Synergic Approach. *J Med Chem* 2009; 52: 1630-38.
- [232] Serafimova R, Gatnik MF, Worth AP. A Review of QSAR Models and Software Tools for Predicting Genotoxicity and Carcinogenicity [pdf on the internet]. European Commission Joint Research Centre, Ispra; c2010 [updated 2010; cited 2011 Oct 17]. Available from http://toxnet.nlm.nih.gov/http://ihcp.jrc.ec.europa.eu/our_databases/jrc-qsar-database/review-qsar-models/at_multi_download/file?name=EUR_24427_EN.pdf
- [233] Ames BN, Lee FD, Durston WE. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proc Natl Acad Sci U S A* 1973; 70: 782-6.
- [234] Ames BN, McCann J, Yamasaki E. Methods for detecting carcinogens and mutagens with the *Salmonella/mammalian-microsome* mutagenicity test. *Mutat Res* 1975; 31: 347-64.
- [235] Ames BN. The detection of environmental mutagens and potential carcinogens. *Cancer* 1984; 53: 2034-40.
- [236] Chemical Carcinogenesis Research Information System (CCRIS) [homepage on the internet]. United States National Library of Medicine; [updated 2011 Oct 17; cited 2011 Oct 17]. Available from <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS>
- [237] Tennant RW. The genetic toxicity database of the National Toxicology Program: evaluation of the relationships between genetic toxicity and carcinogenicity. *Environ Health Perspect* 1991; 96: 47-51.
- [238] Mutagenicity of food additives dataset [homepage on the internet]. Tokyo Metropolitan Institute of Public Health; [updated 2008 May 16; cited 2011 Oct 17]. Available from www.tokyo-eiken.go.jp/henigen/index.htm
- [239] Kirkland D, Aardema M, Henderson L, Muller L. Evaluation of the ability of a battery of three *in vitro* genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity. *Mutat Res* 2005; 584: 1-256.
- [240] Ashby J, Tennant RW. Chemical structure, *Salmonella* mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat Res* 1988; 204: 17-115.
- [241] Ashby J, Tennant RW. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat Res* 1991; 257: 229-306.
- [242] Ashby J. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ Mutagen* 1985; 7: 919-21.
- [243] Cariello NF, Wilson JD, Britt BH, Wedd DJ, Burlinson B, Gombar V. Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. Deductive Estimate of Risk from Existing Knowledge. Toxicity Prediction by Komputer Assisted Technology. *Mutagenesis* 2002; 17: 321-9.
- [244] Helma C, Cramer T, Kramer S, De Raedt L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J Chem Inf Comput Sci* 2004; 44: 1402-11.
- [245] Brinn M, Walsh PT, Payne MP, Bott B. Neural network classification of mutagens using structural fragment data. SAR and QSAR in *Environ Res* 1993; 2-3: 189-210.
- [246] Contrera JF, Matthews EJ, Kruhlak NL, Benz RD. In silico screening of chemicals for bacterial mutagenicity using electrotopological E-state indices and MDL QSAR software. *Reg Toxicol and Pharmacol* 2005; 43: 313-23.
- [247] Klopman G, Chakravarti SK, Harris N, Ivanov J, Saiakhov RD. In-silico screening of high production volume chemicals for mutagenicity using the MCASE QSAR expert system. *SAR QSAR Env Res* 2003; 14: 165-80.
- [248] Maran U, Sild S. QSAR modeling of genotoxicity on non-congeneric sets of organic compounds. *Artif Intell Rev* 2003; 20: 13-38.
- [249] Mekenyan O, Dimitrov S, Serafimova R, Thompson E, Kotov S, Dimitrova N, Walker JD. Identification of the structural requirements for mutagenicity by incorporating molecular flexibility and metabolic activation of chemicals I: TA100 model. *Chem Res in Toxicol* 2004; 17: 753-66.
- [250] Mekenyan O, Serafimova R, Todorov M, Pavlov T, Kotov S, Jacob E, Aptula A. Identification of the structural requirements for mutagenicity, by incorporating molecular flexibility and metabolic activation of chemicals. II. General Ames mutagenicity model. *Chem Res in Toxicol* 2007; 20: 662-76.
- [251] Fjodorova N, Vracko M, Novic M, Roncaglioni A, Benfenati E. New public QSAR model for carcinogenicity. *Chem Central J* 2010; 4: S3.
- [252] Almond LM, Yang J, Jamei M, Tucker GT, Rostami-Hodjegan A. Towards a quantitative framework for the prediction of DDIs arising from cytochrome P450 induction. *Curr Drug Metabol* 2009; 10: 420-32.
- [253] Hewitt NJ, de Kanter R, LeCluyse E. Induction of drug metabolizing enzymes: a survey of *in vitro* methodologies and interpretations used in the pharmaceutical industry-do they comply with FDA recommendations? *Chem-biol interact* 2007; 168: 51-65.
- [254] Wienkers LC, Heath TG. Predicting *in vivo* drug interactions from *in vitro* drug discovery data. *Nat Rev Drug Discov* 2005; 4: 825-33.
- [255] Zhou S-F. Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4. *Curr Drug Metabol* 2008; 9: 310-22.
- [256] Pelkonen O, Turpeinen M, Hakola J, Honkakoski P, Hukkanen J, Raunio H. Inhibition and induction of human cytochrome P450 enzymes: current status. *Arch Toxicol* 2008; 82: 667-715.
- [257] Zhang L, Zhang YD, Zhao P, Huang S-M. Predicting drug-drug interactions: an FDA perspective. *AAPS J* 2009; 11: 300-6.
- [258] Chu V, Einolf HJ, Evers R, Kumar G, Moore D, Ripp S, Silva J, Sinha V, Sinz M, Skerjanec A. *In vitro* and *in vivo* Induction of Cytochrome P450: A Survey of the Current Practices and Recommendations: A Pharmaceutical Research and Manufacturers of America Perspective. *Drug Metabol Dispos* 2009; 37: 1339-54.
- [259] Moon T, Chi MH, Kim D-H, Yoon CN, Choi Y-S. Quantitative Structure-Activity Relationships (QSAR) Study of Flavonoid Derivatives for Inhibition of Cytochrome P450 1A2. *Quant Structure-Act Rel* 2000; 19: 257-63.
- [260] Roy K, Roy PP. Comparative QSAR Studies of CYP1A2 Inhibitor Flavonoids Using 2D and 3D Descriptors. *Chem Biol Drug Des* 2008; 72: 370-82.
- [261] Saraceno M, Massarelli I, Imbriani M, James TL, Bianucci AM. Optimizing QSAR Models for Predicting Ligand Binding to the Drug-Metabolizing Cytochrome P450 Isoenzyme CYP2D6. *Chem Biol Drug Des* 2011; 78: 236-51.
- [262] Susnow RG, Dixon SL. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J Chem Inf Comput Sci* 2003; 43: 1308-15.
- [263] Itokawa D, Nishioka T, Fukushima J, Yasuda T, Yamauchi A, Chuman H. Quantitative Structure-Activity Relationship Study of Binding Affinity of Azole Compounds with CYP2B and CYP3A. *QSAR Comb Sci* 2007; 26: 828-36.

- [264] Chohan KK, Paine SW, Mistry J, Barton P, Davis AM. A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *J Med Chem* 2005; 48: 5154-61.
- [265] Vasanathanan P, Taboureau O, Oostenbrink C, Vermeulen NPE, Olsen L, Jørgensen FS. Classification of Cytochrome P450 1A2 Inhibitors and Noninhibitors by Machine Learning Techniques. *Drug Metabol Dispos* 2009; 37: 658-64.
- [266] Cruciani G. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur J Pharm Sci* 2000; 11: S29-S39.
- [267] Burton J, Ijjaali I, Barberan O, Petitot F, Vercauteren DP, Michel A. Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset. *J Med Chem* 2006; 49: 6231-40.
- [268] Ekins S, Berbaum J, Harrison RK. Generation and Validation of Rapid Computational Filters for CYP2D6 and CYP3A4. *Drug Metabol Dispos* 2003; 31: 1077-80.
- [269] Jensen BF, Vind C, Brockhoff PB, Refsgaard HHF. In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted k-Nearest Neighbor and Extended Connectivity Fingerprints, Including Structural Fragment Analysis of Inhibitors versus Noninhibitors. *J Med Chem* 2007; 50: 501-11.
- [270] Zuegge J, Fechner U, Roche O, Parrott NJ, Engkvist O, Schneider G. A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quantitative Structure-Activity Relationships* 2002; 21: 249-56.
- [271] Kriegl JM, Eriksson L, Arnhold T, Beck B, Johansson E, Fox T. Multivariate modeling of cytochrome P450 3A4 inhibition. *Eur J Pharm Sci* 2005; 24: 451-463.
- [272] Arimoto R. Computational Models for Predicting Interactions with Cytochrome p450 Enzyme. *Curr Topics Med Chem* 2006; 6: 1609-18.
- [273] Zhou DS, Liu RF, Otmani SA, Grimm SW, Zauhar RJ, Zamora I. Rapid classification of CYP3A4 inhibition potential using support vector machine approach. *Lett Drug Design & Discov* 2007; 4: 192-200.
- [274] Lewis DF, Modi S, Dickins M. Quantitative structure-activity relationships (QSARs) within substrates of human cytochromes P450 involved in drug metabolism. *Drug Metabol Drug Interact* 2001; 18: 221-42.
- [275] Gleeson MP, Davis AM, Chohan KK, Paine SW, Boyer S, Gavaghan CL, Amby CH, 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: 559-73.
- [276] Dagliyan O, Kavakli IH, Turkay M. Classification of cytochrome P450 inhibitors with respect to binding free energy and pIC50 using common molecular descriptors. *J Chem Inf and Model* 2009; 49: 2403-11.
- [277] Hammann F, Gutmann H, Baumann U, Helma C, Drewe J. Classification of cytochrome p(450) activities using machine learning methods. *Mol pharmaceutics* 2009; 6: 1920-16.
- [278] Michielan L, Terfloth L, Gasteiger J, Moro S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *J Chem Inf and Model* 2009; 49: 2588-605.
- [279] Korhonen LE, Rahnasto M, Mahonen NJ, Wittekindt C, Poso A, Juvonen RO, Raunio H. Predictive three-dimensional quantitative structure-activity relationship of cytochrome P450 1A2 inhibitors. *J Med Chem* 2005; 48: 3808-15.
- [280] Poso A, Gynther J, Juvonen R. A comparative molecular field analysis of cytochrome P450 2A5 and 2A6 inhibitors. *J Comput-aided Mol Design* 2001; 15: 195-202.
- [281] Korhonen LE, Turpeinen M, Rahnasto M, Wittekindt C, Poso A, Pelkonen O, Raunio H, Juvonen RO. 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: 932-42.
- [282] Oprea TI, Garcia AE. Three-dimensional quantitative structure-activity relationships of steroid aromatase inhibitors. *J Comput-aided* 1996; 10: 186-200.
- [283] Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 1985; 28: 849-57.
- [284] Afzelius L, Zamora I, Ridderström M, Andersson TB, Karlén A, Masimirembwa CM. Competitive CYP2C9 inhibitors: enzyme inhibition studies, protein homology modeling, and three-dimensional quantitative structure-activity relationship analysis. *Mol Pharmacol* 2001; 59: 909-19.
- [285] Pastor M, Cruciani G, McLay I, Pickett S, Clementi S. GRIND-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem* 2000; 43: 3233-43.
- [286] Fontaine F, Pastor M, Sanz F. Incorporating molecular shape into the alignment-free Grid-Independent Descriptors. *J Med Chem* 2004; 47: 2805-15.
- [287] Afzelius L, Masimirembwa CM, Karlén A, Andersson TB, Zamora I. Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. *J Comput-aided Mol Design* 2002; 16: 443-58.
- [288] Gabriele C, Mannhold R, Kubinyi H, Folkers G. Molecular Interaction Fields - Applications in Drug Discovery and ADME Prediction. Wiley-VCH, Weinheim 2005.
- [289] Cianchetta G, Li Y, Singleton R, Zhang M, Wildgoose M, Rampe D, Kang J, Vaz RJ. Molecular Interaction Fields in ADME and Safety. In: Gabriele C, Mannhold R, Kubinyi H, Folkers G, ed., *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*. Wiley: Weinheim 2005; pp. 197-216.
- [290] de Groot MJ, Ekins S. Pharmacophore modeling of cytochromes P450. *Adv Drug Deliv Rev* 2002; 54: 367-83.
- [291] de Graaf C, Vermeulen NPE, Feenstra KA. Cytochrome p450 in silico: an integrative modeling approach. *J Med Chem* 2005; 48: 2725-55.
- [292] de Groot MJ. Designing better drugs: predicting cytochrome P450 metabolism. *Drug Discov Today* 2006; 11: 601-6.
- [293] Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: applications to targets and beyond. *Br J Pharmacol* 2007; 152: 21-37.
- [294] Schuster D, Laggner C, Steindl TM, Langer T. Development and validation of an in silico P450 profiler based on pharmacophore models. *Curr Drug Discov Technol* 2006; 3: 1-48.
- [295] Locuson CW, Rock Da, Jones JP. Quantitative binding models for CYP2C9 based on benzobromarone analogues. *Biochem* 2004; 43: 6948-58.
- [296] Mao B, Gozalbes R, Barbosa F, Migeon J, Merrick S, Kamm K, Wong E, Costales C, Shi W, Wu C, Froloff N. QSAR modeling of *in vitro* inhibition of cytochrome P450 3A4. *J Chem Inf and Model* 2006; 46: 2125-34.
- [297] Leong MK, Chen Y-M, Chen H-B, Chen P-H. Development of a new predictive model for interactions with human cytochrome P450 2A6 using pharmacophore ensemble/support vector machine (PhE/SVM) approach. *Pharm Res* 2009; 26: 987-1000.
- [298] Yu J, Paine MJ, Maréchal J-D, Kemp CA, Ward CJ, Brown S, Sutcliffe MJ, Roberts GCK, Rankin EM, Wolf CR. In silico prediction of drug binding to CYP2D6: identification of a new metabolite of metoclopramide. *Drug Metabol Dispos: the biological fate of chemicals* 2006; 34: 1386-92.
- [299] Watt A, Mortshire-Smith R, Gerhard U, Thomas S. Metabolite identification in drug discovery., *Curr Opin Drug Discov Devel* 2003; 6: 57-65.
- [300] Tarcsay Á, Keseru GM. In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin Drug Metab Toxicol* 2011; 7: 299-312.
- [301] Klopman G, Dimayuga M, Talafofus J. META 1. A Program for the Evaluation of Metabolic Transformation of Chemicals. *J Chem Inf and Model* 1994; 34: 1320-25.
- [302] Darvas F. Predicting metabolic pathways by logic programming. *J Mol Graph Mod* 1988; 6: 80-6.
- [303] Ridder L, Wagoner M. SyGMA: combining expert knowledge and empirical scoring in the prediction of metabolites. *Chem Med Chem* 2008; 3: 821-32.
- [304] Mekenyan OG, Dimitrov SD, Pavlov TS, Veith GD. A Systematic Approach to Simulating Metabolism in Computational Toxicology. I. The TIMES Heuristic Modelling Framework. *Current Pharm Des* 2004; 10: 1273-93.
- [305] Tarcsay A, Kiss R, Keserü GM. Site of metabolism prediction on cytochrome P450 2C9: a knowledge-based docking approach. *J Comput-aided Mol Des* 2010: 399-408.
- [306] Accelrys Metabolite Database. [homepage on the internet]. Accelrys Inc. San Diego CA; c2011 [updated 2011 Oct 17; cited

- 2011 Oct 17]. Available from <http://toxnet.nlm.nih.gov/httpaccelrys.com/products/databases/bioactivity/metabolite.html>
- [307] Adams SE. Molecular Similarity and Xenobiotic Metabolism. [document on the internet]. University of Cambridge, UK; c2010 [updated 2010; cited 2011 Oct 17] <http://www.dspace.cam.ac.uk/bitstream/1810/225225/1/Adams,%20SE%20-%20Molecular%20Similarity%20and%20Xenobiotic%20Metabolism.pdf>
- [308] Carlsson L, Spjuth O, Adams S, Glen RC, Boyer S. Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC bioinformatics* 2010; 11: 362.
- [309] MetaPrint2D. [homepage on the internet]. University of Cambridge, UK; c2008-2011 [updated 2011 Oct 17; cited 2011 Oct 17]. Available from <http://toxnet.nlm.nih.gov/http://www-metaprint2d.ch.cam.ac.uk/metaprint2d>
- [310] Lagunin A, Stepanchikova A, Filimonov D, Poroikov V. PASS: prediction of activity spectra for biologically active substances. *Bioinformatics* 2000; 16: 747-48.
- [311] ACD Labs P450 Regioselectivity Tool [homepage on the internet]. Advanced Chemistry Development, Inc; c1996-2011 [updated 2011 Oct 17; cited 2011 Oct 17]. Available from <http://toxnet.nlm.nih.gov/www.acdlabs.com>
- [312] Zamora I, Afzelius L, Cruciani G. Predicting drug metabolism: a site of metabolism prediction tool applied to the cytochrome P450 2C9. *J Med Chem* 2003; 46: 2313-24.
- [313] Berellini G, Cruciani G, Mannhold R. Pharmacophore, drug metabolism, and pharmacokinetics models on non-peptide AT1, AT2, and AT1/AT2 angiotensin II receptor antagonists. *J Med Chem* 2005; 48: 4389-99.
- [314] Zhou DS, Afzelius L, Grimm SW, Andersson TB, Zauhar RJ, Zamora I. Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metabol Dispos* 2006; 34: 976-83.
- [315] Trunzer M, Fallner B, Zimmerlin A. Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC-MS/MS validation. *J Med Chem* 2009; 52: 329-35.
- [316] Unwalla RJ, Cross JB, Salaniwal S, Shilling AD, Leung L, Kao J, Humblet C. Using a homology model of cytochrome P450 2D6 to predict substrate site of metabolism. *J Comput-aided Mol Des* 2010; 24: 237-56.
- [317] Vasanathanathan P, Hritz J, Taboureaux O, Olsen L, Jørgensen FS, Vermeulen NPE, Oostenbrink C. Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. *J Chem Inf and Model* 2009; 49: 43-52.
- [318] de Graaf C, Oostenbrink C, Keizers PHJ, van Der Wijst T, Jongejan A, Vermeulen NPE. Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J Med Chem* 2006; 49: 2417-30.
- [319] de Graaf C, Pospisil P, Pos W, Folkers G, Vermeulen NPE. Binding mode prediction of cytochrome p450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J Med Chem* 2005; 48: 2308-18.
- [320] Goodsell DS, Olson AJ. Automated docking of substrates to proteins by simulated annealing. *Proteins* 1990; 8: 195-202.
- [321] Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 1996; 261: 470-89.
- [322] Santos R, Hritz J, Oostenbrink C. Role of water in molecular docking simulations of cytochrome P450 2D6. *J Chem Inf and Model* 2010; 50: 146-54.
- [323] Teixeira VH, Ribeiro V, Martel PJ. Analysis of binding modes of ligands to multiple conformations of CYP3A4. *Biochimica et biophysica acta* 2010; 1804: 2036-45.
- [324] Keizers PHJ, de Graaf C, de Kanter FJJ, Oostenbrink C, Feenstra KA, Commandeur JNM, Vermeulen NPE. Metabolic regio- and stereoselectivity of cytochrome P450 2D6 towards 3,4-methylenedioxy-N-alkylamphetamines: in silico predictions and experimental validation. *J Med Chem* 2005; 48: 6117-27.
- [325] Carrieri A, Pérez-Nueno VI, Fano A, Pistone C, Ritchie DW, Teixidó J. Biological profiling of anti-HIV agents and insight into CCR5 antagonist binding using in silico techniques. *ChemMedChem* 2009; 4: 1153-63.
- [326] Pérez-Nueno VI, Ritchie DW, Borrell JJ, Teixidó J. Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape-based virtual screening approach: further evidence for multiple binding sites within the CCR5 extracellular pocket. *J Chem Inf and Model* 2008; 48: 2146-65.
- [327] Pérez-Nueno VI, Pettersson S, Ritchie DW, Borrell JJ, Teixidó J. Discovery of novel HIV entry inhibitors for the CXCR4 receptor by prospective virtual screening. *J Chem Inf and Model* 2009; 49: 810-823.
- [328] Venkatraman V, Pérez-Nueno VI, Mavridis L, Ritchie DW. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J Chem Inf and Model* 2010; 50: 2079-93.
- [329] Kirchmair J, Ristic S, Eder K, Markt P, Wolber G, Laggner C, Langer T. Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J Chem Inf and Model* 2007; 47: 2182-96.
- [330] Kirchmair J, Distinto S, Markt P, Schuster D, Spitzer GM, Liedl KR, Wolber G. How to optimize shape-based virtual screening: choosing the right query and including chemical information. *J Chem Inf and Model* 2009; 49: 678-92.
- [331] Perez-Nueno VI, Ritchie DW. Using Consensus-Shape Clustering To Identify Promiscuous Ligands and Protein Targets and To Choose the Right Query for Shape-Based Virtual Screening. *J Chem Inf and Model* 2011; 6: 1233-48.
- [332] Nicholls A, Grant JA. Molecular shape and electrostatics in the encoding of relevant chemical information. *J Comput-aided Mol Des* 2005; 19: 661-86.
- [333] Sykes MJ, McKinnon RA, Miners JO. Prediction of metabolism by cytochrome P450 2C9: alignment and docking studies of a validated database of substrates. *J Med Chem* 2008; 51: 780-91.
- [334] Rydberg P, Gloriam D, Olsen L. The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* 2010; 26: 2988-89.
- [335] Rydberg P, Gloriam DE, Zaretski J, Breneman C, Olsen L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *Med Chem Lett* 2010; 1: 96-100.
- [336] Zaretski J, Bergeron C, Rydberg P, Huang T-W, Bennett KP, Breneman CM. RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4. *J Chem Inf and Model* 2011; 7: 1667-89.
- [337] de Groot MJ, Ackland MJ, Home VA, Alex AA, Jones BC. Novel approach to predicting P450-mediated drug metabolism: development of a combined protein and pharmacophore model for CYP2D6. *J Med Chem* 1999; 42: 1515-24.
- [338] Hasegawa K, Koyama M, Funatsu K. Quantitative Prediction of Regioselectivity Toward Cytochrome P450 / 3A4 Using Machine Learning Approaches. *Mol Informatics* 2010; 29: 243-249.
- [339] Kuhn B, Jacobsen W, Christians U, Benet LZ, Kollman Pa. Metabolism of sirolimus and its derivative everolimus by cytochrome P450 3A4: insights from docking, molecular dynamics, and quantum chemical calculations. *J Med Chem* 2001; 44: 2027-34.
- [340] Korzekwa KR, Jones JP, Gillette JR. Theoretical studies on cytochrome P-450 mediated hydroxylation: a predictive model for hydrogen atom abstractions. *J Am Chem Soc* 1990; 112: 7042-46.
- [341] Jones JP, Mysinger M, Korzekwa KR. Computational Models for Cytochrome P450: A Predictive Electronic Model for Aromatic Oxidation and Hydrogen Atom Abstraction. *Drug Metabol Dispos* 2002; 30: 7-12.
- [342] Oh WS, Kim DN, Jung J, Cho K-H, No KT. New combined model for the prediction of regioselectivity in cytochrome P450/3A4 mediated metabolism. *J Chem Inf and Model* 2008; 48: 591-601.
- [343] Aleksic M, Thain E, Roger D, Saib O, Davies M, Li J, Aptula A, Zazzaroni R. Reactivity Profiling: Covalent Modification of Single Nucleophile Peptides for Skin Sensitization Risk Assessment. *Toxicol Sci* 2009; 108: 401-11.
- [344] Roberts DW, Patlewicz G, Kern PS, Gerberick F, Kimber I, Dearman RJ, Ryan CA, Basketter DA, Aptula AO. Mechanistic applicability domain classification of a local lymph node assay dataset for skin sensitization. *Chem Res Toxicol* 2007; 20: 1019-30.
- [345] Dimitrov SD, Low LK, Patlewicz GY, Kern PS, Dimitrova GD, Comber MH, Phillips RD, Niemela J, Bailey PT, Mekenyan OG. Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates. *Int J Toxicol* 2005; 24: 189-204.

- [346] Roberts DW, Aptula AO, Patlewicz G. Mechanistic applicability domains for non-animal based prediction of toxicological endpoints. QSAR analysis of the schiff base applicability domain for skin sensitization. *Chem Res Toxicol* 2006; 19: 1228-33.
- [347] Yuan H, Huang JP, Cao CZ. Prediction of Skin Sensitization with a Particle Swarm Optimized Support Vector Machine. *Int J Mol Sci* 2009; 10: 3237-54.
- [348] Ellison CM, Sherhod R, Cronin MTD, Enoch SJ, Madden JC, Judson PN. Assessment of Methods To Define the Applicability Domain of Structural Alert Models. *J Chem Inf and Model* 2011; 51: 975-85.
- [349] Roberts DW, Patlewicz G, Dimitrov SD, Low LK, Aptula AO, Kern PS, Dimitrova GD, Comber MIH, Phillips RD, Niemela J, Madsen C, Wedebye EB, Bailey PT, Mekenyan OG. TIMES-SS - A mechanistic evaluation of an external validation study using reaction chemistry principles. *Chem Res in Toxicol* 2007; 20: 1321-30.
- [350] Enoch SJ, Madden JC, Cronin MTD, Schultz TW. Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via Michael addition. *Chem Res Toxicol* 2008; 21: 513-20.
- [351] Fentem J, Carmichael P, Maxwell G, Pease C, Reynolds F, Warner G, Westmoreland C. Exploring New Approaches to assess Safety without Animal Testing. *AATEX* 2008; 14: 15-20.
- [352] Maxwell G, Mackay C. Application of a systems biology approach to skin allergy risk assessment. *Altern Lab Anim* 2008; 36: 521-56.
- [353] Davies M, Pendlington RU, Page L, Roper CS, Sanders DJ, Bourner C, Pease CK, MacKay C. Determining epidermal disposition kinetics for use in an integrated non-animal approach to skin sensitisation risk assessment. *Toxicol Sci* 2010; 2: 308-18.
- [354] 7th framework joint research programmes (FP7) [homepage on the internet]. European Commission, Brussels; c2011 [updated 2011 Oct 17; cited 2011 Oct 17]. Available from http://toxnet.nlm.nih.gov/http://ec.europa.eu/research/fp7/index_en.cfm
- [355] Innovative Medicines Initiative [homepage on the internet]. Innovative Medicines initiative Brussels; c2011 [updated 2011 Oct 17; cited 2011 Oct 17]. Available from <http://toxnet.nlm.nih.gov/http://www.imi.europa.eu/>
- [356] SEURAT1 project [homepage on the internet]. The European Cosmetics Association, Brussels; c2010 [updated 2011 Oct 17; cited 2011 Oct 17]. Available from <http://toxnet.nlm.nih.gov/http://www.colipa.eu/news-a-events.html>
- [357] Schmidt CW. TOX 21: new dimensions of toxicity testing. *Environ Health Perspect* 2009; 117: 348-353.
- [358] Andersen ME, Krewski D. Toxicity testing in the 21st century: bringing the vision to life. *Toxicol Sci* 2009; 107: 324-30.
- [359] Knudsen TB, Houck KA, Sipes NS, Singh AV, Judson RS, Martin MT, Weissman A, Kleinstreuer NC, Mortensen HM, Reif DM, Rabinowitz JR, Setzer RW, Richard AM, Dix DJ, Kavlock RJ. Activity profiles of 309 ToxCast chemicals evaluated across 292 biochemical targets. *Toxicology* 2011; 282: 1-15.

EXPERT OPINION

1. Introduction
2. Generating in silico models
3. Expert opinion

informa
healthcare

Strategies for the generation, validation and application of *in silico* ADMET models in lead generation and optimization

Matthew Paul Gleeson[†] & Dino Montanari

[†]Kasetsart University, Faculty of Science, Department of Chemistry, Chatuchak, Bangkok, Thailand

Introduction: The most desirable chemical starting point in drug discovery is a hit or lead with a good overall profile, and where there may be issues; a clear SAR strategy should be identifiable to minimize the issue. Filtering based on drug-likeness concepts are a first step, but more accurate theoretical methods are needed to i) estimate the biological profile of molecule in question and ii) based on the underlying structure–activity relationships used by the model, estimate whether it is likely that the molecule in question can be altered to remove these liabilities.

Areas covered: In this paper, the authors discuss the generation of ADMET models and their practical use in decision making. They discuss the issues surrounding data collation, experimental errors, the model assessment and validation steps, as well as the different types of descriptors and statistical models that can be used. This is followed by a discussion on how the model accuracy will dictate when and where it can be used in the drug discovery process. The authors also discuss how models can be developed to more effectively enable multiple parameter optimization.

Expert opinion: Models can be applied in lead generation and lead optimization steps to i) rank order a collection of hits, ii) prioritize the experimental assays needed for different hit series, iii) assess the likelihood of resolving a problem that might be present in a particular series in lead optimization and iv) screen a virtual library based on a hit or lead series to assess the impact of diverse structural changes on the predicted properties.

Keywords: ADMET, *in silico* models, lead generation, lead optimization, virtual screening

Expert Opin. Drug Metab. Toxicol. [Early Online]

1. Introduction

Candidate attrition remains a major concern in the pharmaceutical industry, running at approximately 90% [1,2]. This dramatic failure rate has led to a concerted effort to screen development compounds, not only for their target potency and selectivity, but also their absorption–distribution–metabolism–excretion–toxicity (ADMET) profile. Large datasets of compounds, ordering on 100,000s, have been generated for some of these measures within the industry, with 1000s now contained in databases such as ChEMBL [3], Pubchem [4] or commercial variants such as Aureus [5] and WOMBAT-PK [6]. Given the massive costs associated with the determination of these measurements, it is highly desirable that these measurements are utilized to help with the selection of leads, and guide their optimization into clinical candidates.

The selection of the initial lead molecule is of critical importance in the early stages of drug discovery. Lead optimization generally tends to increase molecular properties such as molecular weight and lipophilicity [7]. In addition, it is often

Article highlights.

- Large amounts of ADMET data are available for data mining.
- Highly desirable to utilize huge investment costs in screening for future discovery efforts.
- Models of many different types are available. It is important to consider how they are used.
- Models can be used in lead generation for tasks such as focused set selection, hit prioritization and highlighting possible ADMET issues.
- Models can be used in lead optimization to help guide SAR studies, enable the screening and prioritization of virtual libraries and to facilitate multiple parameter optimization.

This box summarizes key points contained in the article.

not possible to alter the key pharmacophoric features associated with a lead series, without losing activity at the target, making multivariate optimization of activity and other developability parameters challenging, if not impossible. Thus, the final candidate is generally quite similar in terms of structure or properties to the starting lead compound. This is an area where computational ADMET models, or so-called *in silico* models, could prove useful to guide the selection and optimization of such leads.

Drug-likeness concepts based on molecular physicochemical properties are popular as first filters in early drug discovery [8,9]. However, many compounds in what might be considered ideal physicochemical property space will still have liabilities, although it is less probable. These liabilities must therefore be identified and subsequently designed out if possible. Experimental screening dominates this assessment process at present, yet the possibility of using computational (i.e., *in silico*) tools has the potential to streamline costs and speed up the assessment process, if they can be more effectively employed as early filters [10-12]. That is, the models do not replace the experimental methods, rather they supplement them. Models could have much greater impact in lead generation and early lead optimization if more effectively integrated into decision-making processes, and used, where suitable, as surrogates for experimental measures.

'All models are wrong, but some are more useful than others: George Box 1987'. The key issue with the application of *in silico*-based methods in drug discovery is that one must understand in which situations, and for what decisions, a particular *in silico* method can be sensibly used [13]. In this paper, the authors consider this issue, taking into account the sources of error, the modeling methods, model assessment and its validation. They discuss the criteria required by models that are to be applied in the lead generation and lead optimization steps. The uses discussed include: i) the rank ordering of hits from high-throughput screening (HTS), ii) the prioritization of the experimental assays needed for different hit series, iii) assessing the likelihood of resolving a problem that might

be present in a particular series and iv) the screening of virtual libraries of a given hit or lead series to assess the impact of diverse structural changes on the predicted properties.

2. Generating *in silico* models

In silico ADMET models are theoretically derived functions based on chemical descriptors, offering a means to virtually predict the outcome of chemicals for a particular ADMET measure (Figure 1). The methods rely on leveraging previously derived results by training a model to relate the known assay responses to a theoretical description of a molecule's structure. *In silico* predictions can be made for many ADMET assays types, however, their accuracy can vary significantly due to the nature of the end point itself, the quality and amount of data available, and the methods used to build and validate the model(s).

This paper discusses the different steps involved in building an *in silico* ADMET model. The authors discuss aspects of model building and use, including the data collation steps, the different statistical methodologies and descriptors available, the model validation step and finally how and where one might choose to use the validated model.

2.1 Data considerations

An *in silico* model cannot predict with 100% accuracy measurements obtained from the experimental assay that it is based on. This is because *in silico* models not only must deal with the experimental errors in the assay output, but also approximations used in the computation of the theoretical chemical descriptors, as well as model fitting errors arising from non-ideal, non-diverse datasets (in terms of both molecular structure and activities). Before embarking on an *in silico* model building exercise, it is therefore advisable to consider i) how reliable/reproducible the experimental assay output is and ii) how large and structurally diverse a dataset has been screened in the experimental assay?

If the available database does not contain a large, diverse range of activities, then the utility of any derived model to predict in the underrepresented region is likely to be questionable. In addition, if the diversity of the compounds screened in the assay is low, there also exists questions as to whether the model will be able to predict with any accuracy the outcome of a molecule which deviates significantly in terms of molecular structure [14,15]. A number of computational methods exist to estimate the reliability of a quantitative structure-activity relationship (QSAR) prediction. These are based on the similarity of the compound being predicted to members of the training set. The so-called domain of applicability, has been reviewed elsewhere [16].

2.1.1 Data variability

A significant challenge in the development of *in silico* ADMET models relates to data variability [17,18]. Even within individual laboratories, methods used to measure parameters

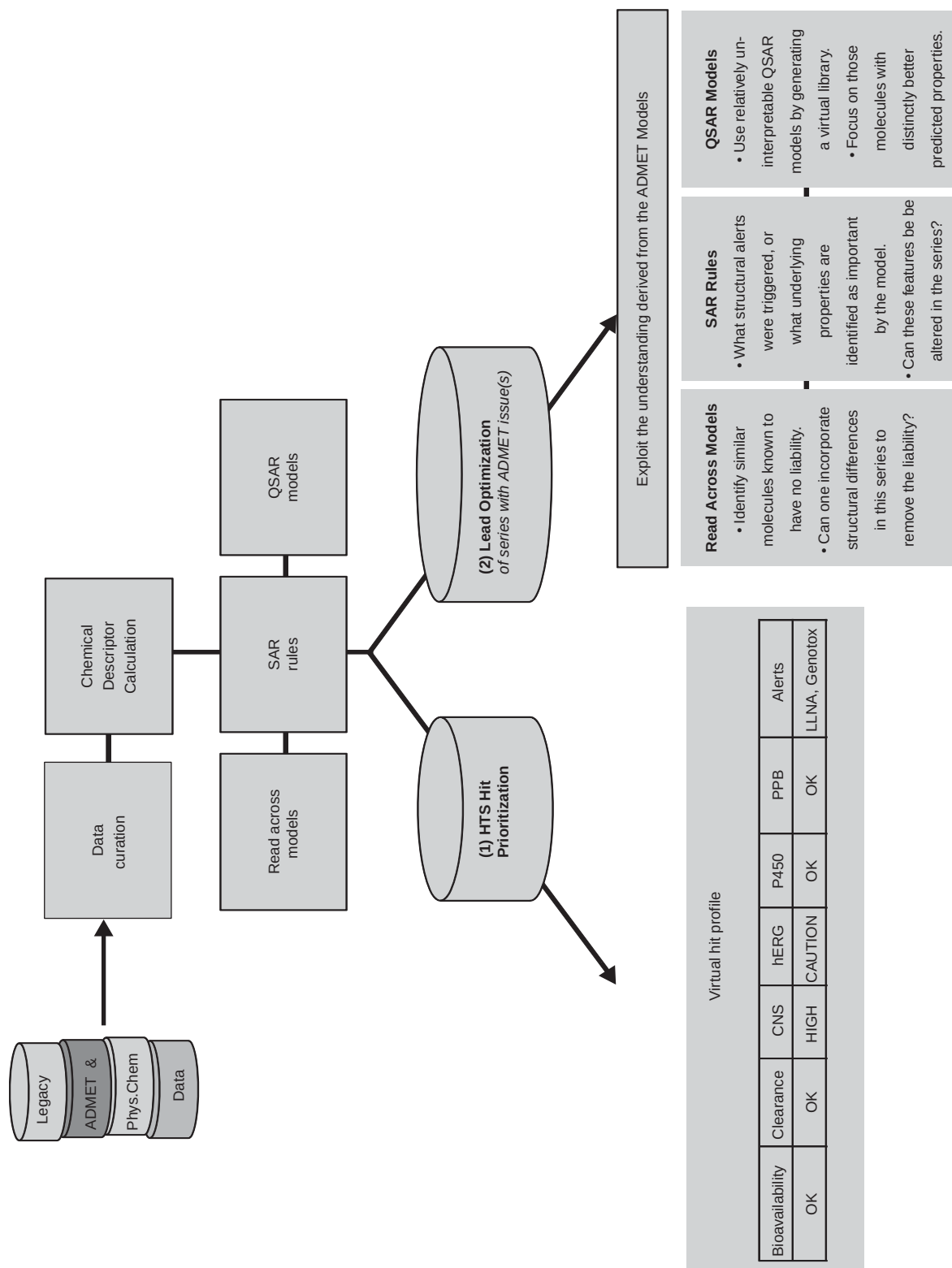


Figure 1. The use of ADMET and physicochemical data to generate models that can be used (1) for use in hit prioritization and to determine the priority of follow-up assays and (2) the use in the optimization of a lead series by exploiting the underlying structure-activity relationships to design new, more tolerable molecules.

can evolve over time, which has implications for ADMET model building and validation. This is because experimental data are implicitly assumed to be error free.

To highlight why *in silico* models cannot be 100% predictive, the authors show some examples of how relatively small changes in ADMET assay conditions can introduce non-negligible errors. Reports based on P450 inhibition screening at GlaxoSmithKline exemplify the point [19]. The effect of a change in the P450 protein source used in the inhibition assays, from Cypex Bactosome to Gentest cDNA microsomes, has been systematically analyzed. One hundred and twenty-four compounds were measured in one of five P450 isoforms, using the same probes, which allowed the researchers to assess the effect of the protein source on variability. The researchers observed a coefficient of determination (r^2) of 0.84 with a root mean squared error (RMSE) of 0.40 (2.5-fold error) between the two pIC_{50} measurements of the combined group of 124 compounds. For 23 compounds with P450 3A4 measurements in a different protein source, and using a different probe molecule (DEF vs. PPR), the r^2 decreased to 0.74, and the RMSE increased to 0.77 (5.9-fold error).

Scientists from Schering-Plough reported the impact of using fresh and frozen hepatocytes on intrinsic clearance experiments [20]. Analysis of the data in the reference shows there were a total of 27 compounds with measurements in both forms. The r^2 between the log Cl_{int} values, with three significant outliers removed, was 0.58, and the RMSE was 0.61 (threefold error).

In vivo assays display significant variability due to their even greater complexity. Indeed, errors are considered larger for the cassette dosing of compounds, which is now a much more popular method for ethical and cost reasons [21,22]. Researchers at GlaxoSmithKline systematically compiled a list of 115 compounds with a full set of discrete and cassette measurements for total clearance (CL), volume of distribution (VDSS), plasma half-life ($T_{1/2}$) and bioavailability (F) [22]. Analysis of the data provided in this reference shows that the square of the r^2 between the logarithms of the reported values are: 0.51 for F, 0.67 for $T_{1/2}$, 0.83 for VDSS and 0.89 for CL. The corresponding RMSEs are 0.31 (~ 2-fold error) for F, 0.24 (~ 1.8-fold) for $T_{1/2}$, 0.21 (~ 1.6-fold) for VDSS and 0.20 (~ 1.6-fold) for CL.

A common toxicity assay, the AMES mutagenicity test, displays a ~ 15% classification error rate between laboratories [23], and the overall classification success rate is between 77 and 90% [24]. Indeed, while the predicted AMES test mutagens have a high probability of being carcinogens, compounds that are predicted to be non-mutagens still have an equal probability of being either non-carcinogenic, or carcinogenic via a different mechanism (i.e., non-genotoxic carcinogens) [25,26].

Such examples help to highlight the challenge faced when generating *in silico* models on ADMET parameters. Consider the case of rat bioavailability, for example. Since the bioavailability between discrete and cassette experiments shows a roughly 50% concordance (the r^2 of 0.51 means ~ 51% of

the variance in one assay is explained by the other), then it is probable that *in silico* models, which are typically built on data using different methods, rat species, doses, etc., will in all probability, not do much better. Indeed, *in silico* errors of 0.5 log (~ threefold) units should be expected for bioavailability (in the authors' experience), and goes some way to explain why no quantitative models for bioavailability have been reported in the literature to date. Figure 2 illustrates the implications of making decisions with an *in silico* model for a parameter such as bioavailability (~ threefold), as compared with experimental results from discrete studies (~ twofold accuracy).

2.1.2 Data collation issues

Database entry errors, compound dispensing errors and compound degradation during storage are additional factors that can confound SARs. Researchers at GSK highlighted issues surrounding the maintenance of a good quality screening collection, from an extensive quality control study. Only 80% of the entries in the original GSK screening collection were considered both pure and confirmed as having the correct structure based on mass spectrometry [27]. Data generated from collections where compounds are poorly quality controlled will have significant detrimental effect on data quality, SAR rationalization and decision making.

Young *et al.* [28] reported on the problems associated with data entry based on an analysis of a number of public databases. These databases typically rely on secondary sources such as peer review journals for much of their information. Errors associated with the input structures were found to range from 0.1 to 3.4% depending on the database in question [28]. In addition, Fourches *et al.* reported on the quite extensive impact such errors can have during *in silico* model generation. They also discuss the steps required for reliable data collation prior to QSAR model generation [29].

2.2 Basic model types

A variety of different modeling approaches can be used to estimate the ADMET potential of new compounds [30,31]. All these methods rely on the derivation of a model using experimental data to train and validate the method. The success of the method is generally governed by i) the complexity of the underlying biological process, ii) the diversity of the compounds screened, iii) the diversity of the biological results obtained, iv) the ability of the theoretical descriptors to describe the physical events occurring and v) the ability of the statistical method to fit the correct relationship between the structure and the response.

One of the simplest approaches employs the similarity principle in the form of 'read across' [32]. The similarity of a compound to a set of known molecules can be calculated using fingerprint or fragment-based descriptors (see below). Molecules that are found to be very similar (i.e., 0.95 Tanimoto similarity) are more likely to display similar activity. The value of this approach is that it is very intuitive, however, it only provides qualitative information since the

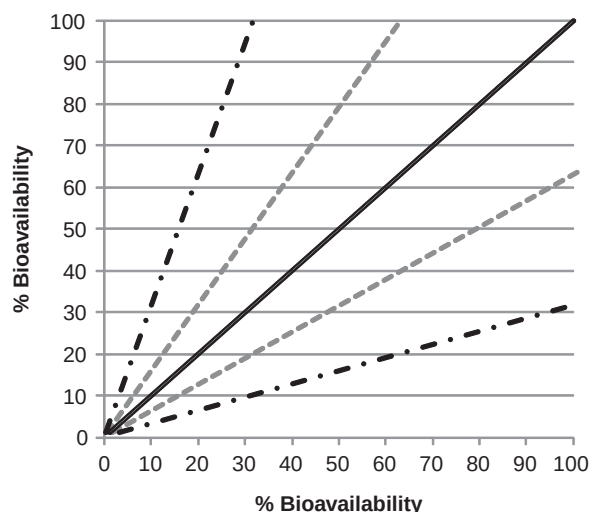


Figure 2. An illustration of the experimental error associated with *in vivo* bioavailability (dashed, twofold) and the corresponding *in silico* prediction errors (dash-dot, fivefold).

success of the similarity principle is dependent on the relevance of the descriptors chosen to describe chemical space of the physical event occurring. Indeed, it is well known that even very simple modifications to a particular molecular structure can result in dramatic differences in activity [33-35].

A second, qualitative approach involves the use of qualitative SAR. Compounds with known activities are analyzed and molecular features that can separate the active and inactive compounds are identified [36]. These can then be used to qualitatively apportion risk to molecules untested in the experimental assay [37,38].

An extension of SAR models, QSAR models attempt to predict the level of activity for a given ADMET end point [39,40]. Models are built in a quantitative fashion that relates the experimental activity to a set of molecular descriptors. These models can be fitted using linear or partial least squares regression, or a variety of machine learning algorithms. QSAR models can be built on small sets of congeneric series (local models) or large diverse sets (global models) [41-43].

A fourth type of model can also be defined. Rather than try to indirectly associate chemical structure with ADMET assay output, an atomistic representation of the ADMET event can be constructed and this used to generate a prediction [44]. An example of this approach can be found in the three-dimensional (3D) protein-ligand models, used for site of metabolism prediction, combining molecular docking and QM calculations [45].

2.3 Molecular descriptors

Theoretically derived molecular descriptors are needed to build *in silico* ADMET models [46,47]. A large number of different types exist, which can be broadly categorized as follows: one-dimensional (1D) molecular properties or substructure counts,

two-dimensional (2D) electro-topological and fingerprint descriptors, 3D pharmacophore and fingerprint descriptors, as well as many quantum chemical descriptors or those derived from atomic simulations [30]. Molecular descriptors can be computed in a wide variety of software packages, using a variety of different methodologies. A number of excellent reviews are available elsewhere [30,32,44,46,48].

The theoretical computation of molecular descriptors is a non-trivial process since molecules are dynamic in nature. As such, their physical and electronic properties can change depending on their ionization state or conformation, meaning computed values are somewhat limited surrogates of the true molecular structure [49]. Different ADMET end point will have varying dependencies on the different types of descriptors, so careful consideration should be given to the choice prior to model building studies. For example, the permeability of molecules through a biological membrane will show a greater dependence on simple molecule properties such as size and ionization state than on their molecular fingerprint. By contrast, molecular properties have less of an impact on the sites of primary metabolism. This is because the process is more affected by the precise molecular structure and recognition, with only certain P450 metabolizing molecules. In addition, quantum mechanical descriptors such as the HOMO-LUMO band gap are known to be an important descriptors for the barrier to oxidation [45].

Simple molecular descriptors such as molecular weight, lipophilicity and atom counts are very quick to calculate and have proved very useful for the derivation of ADMET rules. This includes the rule of 5 [8] and many others [50-53], as well as many QSARs based on both diverse dataset and individual molecular series [54-57]. More complex descriptors describing the connectivity, size and shape can also be employed. Where the ADMET process is controlled by a specific binding event, such as P450 inhibition or metabolism, ligand pharmacophores or shape-based methods could prove useful to identify likely issues [45].

2.4 Model validation

Irrespective of the method or descriptors used to generate a model, it is of fundamental importance that its performance be assessed in an independent manner before it is used in decision making. A variety of protocols for the generation and validating of *in silico* models have been proposed in the literature [31,40,43,58-60]. In most modeling studies, a hold-out set of between 10 and 30% of the data will be held back to test the independent performance of the model. If the model performs well on this set then it is typically deemed suitable for use in real-world decision making. However, this test set is not necessarily ideal since it will cover the same diversity as the original training set. Thus, it is necessary to monitor how new compounds that are predicted compare with the original training set using so-called domain of applicability method mentioned before [16]. As newer compounds are synthesized, the diversity of the compounds will

increase, so the performance of the model would be expected to drop. In this case, it is typically advisable to experimentally validate the model for a new series of compounds by assessing the concordance between experiment and prediction [15,61].

A number of common statistical metrics used to quantify the performance of *in silico* models have been reported in the literature [29,31,40]. A rigorous assessment of the model performance must be undertaken, using standard metrics, so as to characterize the level of predictivity, but also facilitate the communication of results. A number of papers have described the metrics suitable for characterizing models that output a class-based [31,62] or continuous prediction [31,39].

3. Expert opinion

As a general rule, extensive *in vivo*-based animal studies are restricted to very small numbers of heavily optimized molecules, late in the lead optimization phase. This provides the necessary high-quality information for the critical candidate selection decisions. *In vitro*-based studies are generally used in late-stage hit prioritization and early lead optimization, as a means to characterize a molecular series, and facilitate its optimization [63]. *In silico*-based methods are typically applied for the selection of HTS screening sets, post-HTS hit characterization and in lead optimization, to help rationalize SAR and guide the chemical modifications of a lead series.

Many different ADMET models are available; from web-based portals available in most large pharmaceutical companies, to freely accessible internet web portals [64], and stand alone commercial packages [65,66]. With so many tools available to the drug discovery scientist, it is important that one carefully considers how and when they should be incorporated into day-to-day decision making. It is generally accepted that *in silico*-based ADMET models are not sufficiently accurate to make the most critical decision in drug discovery. *In vitro*-based models are more reliable, however even these methods cannot be used alone to make candidate selection decisions. The key to using any predictive surrogate successfully is to understand its strengths and limitations, and use them for decisions where their accuracy and speed of determination is appropriate. This means that one must first consider the data used to build the model, the experimental assay errors, before employing a model for a particular task.

A great number of *in silico* ADMET models have been reported in the literature to date [26,31,45,55,57,66-74], and many papers describe how such models should be incorporated into drug discovery workflows [10-12]. However, rather few medicinal chemistry papers have been reported in the literature that describe the application of such models. Some examples include the design of a CB1 receptor antagonist series with reduced central nervous system (CNS) activity by focusing on high Polar Surface Area (PSA) [75], the reduction of P-glycoprotein (P-gp) efflux in β -secretase inhibitors by masking known recognition features [76], and the design of

histamine H3 receptor antagonist with reduced hERG using 3D-based models [77].

The reason for the relative scarcity of *in silico* ADMET applications in the literature is likely to be manifold. *In silico* ADMET models are certainly used in HTS triaging, but this is not often reported in literature SAR studies. This is possibly due to the fact that the primary focus of most publications is on the optimization step and the source of the lead is not of primary concern. This could arise where the lead was based on competitor structures, or was identified by another team within the same company, whose sole focus is lead generation. Also, it is generally only discovery projects that fail during development that are reported. It is also probable that many of these projects did not benefit from the use of *in silico* ADMET model usage given that development failures due to ADMET issues are known to be significant [1]. Indeed, those that did benefit, and overcame ADMET issues, are also unlikely to be published. Another important point is that it is often very difficult to optimize both potency and ADMET issues in unison [78].

The challenge of demonstrating the true value of computational methods has been noted by others [79,80]. A key issue is that when compounds are predicted as having poor ADMET properties, they are less likely to be synthesized, thereby making it difficult to fully validate said model. In addition, the focus on target potency [78,81] means that the structural changes necessary to remove a particular ADMET liability are unlikely to be pursued. For example, a basic, lipophilic template with the required dopaminergic activity might also have significant hERG affinity. However, any *in silico* model or rule will lead to the selection of compounds with lower lipophilicity and reduced basicity, thereby reducing the dopaminergic activity also.

In the following section, the authors describe situations where *in silico* models of varying degrees of predictivity can make valuable contribution to drug discovery. The discussion of the methods is split into two categories, those suitable for i) lead generation and ii) lead optimization. This categorization is somewhat subjective given that a considerable degree of optimization can occur during lead generation. The distinction is made simply because in earlier decision-making processes, lower accuracy in a given predictive method is tolerated better. Hence relatively crude, but useful rules, such as the rule of 5 [8], can be used to bias hit selection without the risk of losing too many good molecules. As the number of compounds under consideration drops, and as the chemical modifications of a lead series decrease with each medicinal chemistry iteration, more accurate methods are needed to guide the design of increasingly more subtle chemical modifications.

3.1 Using *in silico* ADMET models in lead generation

In silico models are highly suited for use in lead generation. This is because one cannot experimentally assess the approximately 2 million compounds found in a typical screening

collection with any rigour, even using the most high-throughput *in vitro* assays. Models can be used pre-HTS to reduce the number of compounds that need to be screened, or post-HTS, to reduce the number of hits obtained to a more reasonable size that is amenable for full IC₅₀ determination, or visual assessment. The definition of focused screening sets, or the filtering of datasets post-HTS, can be achieved using methods such as Lipinski's rule of 5 [8], or more lead-like [82] or fragment-like filters [83]. Alternatively, newer probabilistic-type functions that encode drug-likeness could be used to reduce the size of the sets to be screened [84].

Models can be used as an early, qualitative indicator of a compound or series ADMET tolerability, pre-*in vitro* assessment. QSAR models are not 100% predictive, so the absolute prediction could be used to classify compounds as high, low or indeterminate risk (i.e., if predicted in the middle of the scale) [85]. For example, if the *in silico* model error is ~ 0.5 log unit, predictions 0.5 log units beyond the upper or lower bound values are $\geq 68\%$ likely to lie either side of the cut-off. That is, there is $\geq 68\%$ probability that the difference between the predicted value and the cut-off is not 0 (based on a two-tailed t-test, which assumes data normality). Predictions at ≥ 1 log unit from the cut-off are $\geq 95\%$ likely to have a non-zero difference from the cut-off. Compounds predicted within 0.5 log of the cut-offs (or 1 log unit for greater accuracy) could then be classed as indeterminate, while those lying beyond are given a more reliable classification (Figure 3). The use of a model to conservatively predict what issues are likely to be present within a given compound, within a give series, allows scientists to select between different hits, or to bias the priority of assays during hit workup to be more specific for a particular compound or series.

One of the most simple but most used models developed is the rule of 5 [8]. Developed from an analysis of drugs and late development compounds, it was determined that compounds which fail two or more of the following rules should probably be excluded from further development due to the high likelihood of poor absorption (MWT > 500, calculated log P (clog P) > 5, hydrogen bond acceptors (HBA) > 10, hydrogen bond donors HBD > 5). More recently, Martin proposed the relatively intuitive bioavailability score, which also takes into account the effect of different ionization states on absorption [86].

More focused rules can be used to create focused screening sets, or during hit selection. As an example consider a situation where CNS penetration is required. Norinder and Haeberlein [87] have proposed the following simple rules: rule 1, if N + O (the number of nitrogen and oxygen atoms) in a molecule is ≤ 5 it has a high chance of entering the brain. Rule 2, if $\log P - (N + O)$ is > 0 then log BB is positive (i.e., greater in the brain than blood). More recently, Di *et al.* proposed the following physicochemical guidelines for compounds more likely to cross the blood-brain barrier (BBB); PSA < 60 – 70 Å², log D between 1 and 3 and MWT < 450 [88].

One can use knowledge-based rules to assess for possible metabolic or toxicity issues for a given compound or series [89].

While, these expert systems are only qualitative (i.e., known to display false positives and false negative predictions), they can still prove useful to highlight possible sites of metabolism [90,91]. If the liability is intrinsic to the template in question, rather than a part or position that is more easily functionalizable, then this could influence its prioritization at the hit selection stage. Examples of such models include structural alerts for both mutagenicity [26] and skin sensitivity [92] and reactive metabolites [91].

In summary, the principle uses for these models in lead generation are: i) definition of focused screening sets, pre-HTS, ii) prioritization of screening hits, post-HTS, iii) assessment of the possible ADMET issues for each unique template and iv) use to determine the priority of assays used for *in vitro* ADMET work-up.

3.2 Using *in silico* ADMET models in lead optimization

In silico models can also be used in a more quantitative fashion in lead optimization. The principle uses for these models in lead optimization are: i) to help discharge a confirmed ADMET liability by exploiting SAR knowledge from the *in silico* model, ii) to prioritize a diverse virtual library of all possible template modifications and substituents using *in silico* model(s) and iii) to perform multiple-optimization of confirmed ADMET issues, using underlying SAR from the relevant *in silico* models.

Models can be used for qualitative assessments of the likelihood of resolving an ADMET problem in a particular series. Alternatively, screening a virtual library based on the hit or lead series to assess the impact of diverse structural changes on the predicted properties. If the model has good discriminating power, then it suggests that the underlying SAR is relevant and can be used to guide further design. This can be achieved by exploiting the SAR for parameters that can positively affect the ADMET end point. In this instance, interpretable models that are physically meaningful are preferential to black box types. *In silico* models could also be used to screen a virtual library of all possible modifications to a series. This is ideal in cases where the *in silico* model complexity is high, hindering its use in guiding the design of new molecules directly.

QSAR models can also be used more quantitatively to facilitate the multi-objective optimization of leads [13,93]. Equally importantly, models can be used to assess whether it is realistically possible to discharge a particular ADMET liability or liabilities. Based on an analysis of the underlying descriptors within a model, or predictions made on a virtual library of possible modifications, it should be possible to assess whether the discharge of multiple liabilities will be facile or not. Another important use of such methods is to help facilitate termination decision on a series or project, more quickly. This is particularly important since it is often not possible to balance the structural requirements imposed by the primary target pharmacophore and the properties required for good experimental ADMET.

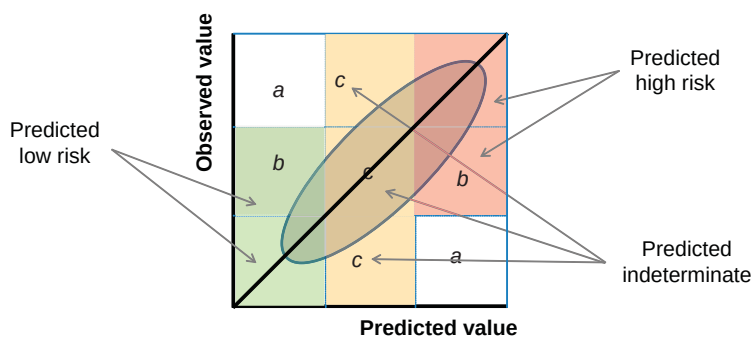


Figure 3. A three class model with appropriately selected cut-offs can be used to; (1) exclude the highest risk compounds (i.e. focus on predicted “low” and “indeterminate” compounds (red)) or (2) bias the selection of a set towards low risk compounds (i.e. focus on predicted “low” compounds only for example (green)). The latter selection will result in a smaller dataset, with a much lower percentage of high risk compounds. However, it will also exclude a percentage of observed low and moderate risk compounds also. Cut-offs that will result in no observation miss-classified by more than 2 classes (denoted a) and few by 1 class (b) are preferred. This however may result in many predictions being classified in the indeterminate category (c).

This is because the target affinity and the different ADMET parameters often show differing dependencies with simple physical properties, making the multi-objective optimization of many parameters at once challenging [37,38,78].

Some examples of quantitative, and quite predictive *in silico* ADMET models include volume of distribution at steady state (VDSS), plasma protein binding, solubility and CNS penetration. Using a dataset of 669 compounds with VDSS measurements, Lombardo and collaborators generated QSAR models with mean fold errors < 2 on their external test set [94]. A model of this accuracy is sufficiently accurate to evaluate the likely VDSS at an early stage. Plasma protein binding has also been modeled using large databases, of between 1000 and 10,000 observations in size. On external test sets these QSAR models show r^2 of ~ 0.5 , and RMSE of ~ 0.6 [61,95], certainly sufficient for the initial assessment of protein binding phenomena. Solubility has also been modeled extensively, using diverse literature-derived datasets of 1000s of compounds [96], to more focused sets consisting of ~ 100 drug-like compounds [97]. The prediction errors found based on independent test set based on the former datasets were ~ 0.6 log units, while for the latter they were observed to be ~ 1 log unit. CNS penetration QSAR models have been reported on relatively small datasets. Nevertheless, the models show reasonable performance using very small numbers of simple descriptors, giving greater confidence in their generalizability. A notable example is the QSAR model of Abraham *et al.* generated using 148 compounds in total. The model displayed an r^2 of 0.75 and a 2.2-fold mean error. CNS models of this accuracy are certainly useful for such tasks as focusing screening sets for improved CNS activity or other similar tasks in early drug discovery.

3.3 Application of *in silico* tools in drug discovery

A wide variety of ADMET models are available from commercial and open sources [65]. They can be also be generated from

the large amounts of data in publically accessible databases [3,4], or using the many curated datasets reported in the QSAR literature. The challenge however is to effectively incorporate these techniques into multivariate decision-making processes [10-12]. As the unhealthy focus on absolute potency begins to subside [78,81], it might be expected that greater unhealthy focus on ADMET models, or relevant properties, will probably be observed in the coming years.

In silico models cannot be used to definitively assign ADMET risks to a compound or lead series [36,98,99]. The methods should be used to discard only the highest risk compounds, or those that are predicted to possess multiple issues, so as to reduce the chance of throwing away good compounds. Should a compound display certain ADMET liabilities, but be an attractive lead for alternate reasons, the series should be followed up experimentally. Additional theoretical analyses could be done to assess how likely the liability can be discharged given the ADMET model SAR and the physical requirements of the target pharmacophore.

Lead optimization is, of course, a multi-objective process, and the critical parameters being optimized can be correlated, inversely correlated or orthogonal, making the optimization process itself extremely challenging [93,100]. Physical property-based models and substructure-based alerting tools can only get one so far, however. To use them most effectively in lead generation they should ideally be deployed alongside *in vitro* and *in vivo* ADMET methods, to reduce the amount of experimental testing where appropriate, to help rationalize SAR, and to guide chemical modification.

Declaration of interest

This work was supported by the Thailand Research Fund grant RSA5480016 and the Faculty of Science, Kasetsart University. The authors declare no other conflicts of interest.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates?. *Nat Rev Drug Discov* 2004;3(8):711–16
2. Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;9(3):203–14
3. ChEMBL. Available from: <https://www.ebi.ac.uk/chembl/>
4. Pubchem. Available from: <http://pubchem.ncbi.nlm.nih.gov/>
5. Aureus. Available from: <http://www.aureus-sciences.com>
6. WOMBAT-PK. Available from: <http://www.sunsetmolecular.com/>
7. Oprea TI, Davis AM, Teague SJ, Leeson PD. Is there a difference between leads and drugs? A historical perspective. *J Chem Inf Comput Sci* 2001;41(5):1308–15
8. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. 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
9. Lajiness MS, Vieth M, Erickson J. Molecular properties that influence oral drug-like behavior. *Curr Opin Drug Disc* 2004;7:470–7
10. Yu H, Adedoyin A. ADME-Tox in drug discovery: integration of experimental and computational technologies. *Drug Discov Today* 2003;8(18):852–61
11. Stahl M, Guba W, Kansy M. Integrating molecular design resources within modern drug discovery research: the Roche experience. *Drug Discov Today* 2006;11(7–8):326–33
12. Hop CE, Cole MJ, Davidson RE, et al. High throughput ADME screening: practical considerations, impact on the portfolio and enabler of *in silico* ADME models. *Curr Drug Metab* 2008;9(9):847–53
13. Segall M, Chadwick A. Making priors a priority. *J Comput Aided Mol Des* 2010;24(12):957–60
14. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK. Similarity to Molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J Chem Inf Comput Sci* 2004;44(6):1912–28
- **A paper that clearly demonstrates with a number of examples the relationship between model performance and similarity to the training set.**
15. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model* 2008;26(8):1315–26
16. Tetko IV, Bruneau P, Mewes H-W, et al. Can we estimate the accuracy of ADME-Tox predictions?. *Drug Discovery Today* 2006;11(15-16):700–7
17. Thorne N, Auld DS, Ingles J. Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr Opin Chem Biol* 2010;14(3):315–24
18. Gilbert M R. Reactive compounds and *in vitro* false positives in HTS. *Drug Discov Today* 1997;2(9):382–4
19. Kajbaf M, Longhi R, Montanari D, et al. A comparative study of the CYP450 inhibition potential of marketed drugs using two fluorescence based assay platforms routinely used in the pharmaceutical industry. *Drug Metab Lett* 2011;5(1):1872–3128
20. Lau YY, Sapidou E, Cui X, et al. Development of a novel *in vitro* model to predict hepatic clearance using fresh, cryopreserved, and sandwich-cultured hepatocytes. *Drug Metab Dispos* 2002;30(12):1446–54
21. He K, Qian M, Wong H, et al. N-in-1 dosing pharmacokinetics in drug discovery: experience, theoretical and practical considerations. *J Pharm Sci* 2008;97(7):2568–80
22. Nagilla R, Nord M, McAtee JJ, Jolivet LJ. Cassette dosing for pharmacokinetic screening in drug discovery: comparison of clearance, volume of distribution, half-life, mean residence time, and oral bioavailability obtained by cassette and discrete dosing in rats. *J Pharm Sci* 2011;100(9):3862–74
23. Kazius J, McGuire R, Bursi R. Derivation and validation of toxicophores for mutagenicity prediction. *J Med Chem* 2005;48(1):312–20
24. Mortelmans K, Zeiger E. The AMES salmonella/microsome mutagenicity assay. *Mutat Res* 2000;455(1-2):29–60
25. Benigni R, Bossa C, Tcheremenskaia O, Giuliani A. Alternatives to the carcinogenicity bioassay: *in silico* methods, and the *in vitro* and *in vivo* mutagenicity assays. *Expert Opin Drug Met* 2010;6(7):809–19
26. Benigni R, Bossa C. Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. *Chem Rev* 2011;111(4):2507–36
27. Lane SJ, Eggleston DS, Brinded KA, et al. Defining and maintaining a high quality screening collection: the GSK experience. *Drug Discov Today* 2006;11(5-6):267–72
- **This interesting paper highlights the importance of screening collection maintenance. Failure to do will have a major impact on the subsequent data quality.**
28. Young D, Martin T, Venkatapathy R, Harten P. Are the chemical structures in your QSAR correct?. *QSAR Comb Sci* 2008;27(11-12):1337–45
- **An interesting paper that highlights the problems when incorrect chemical structures are used in QSAR applications**
29. Fourches D, Muratov E, Tropsha A. Trust, But verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* 2010;50(7):1189–204
- **An excellent paper that describes common errors found in a range of popular databases. They highlight the need for effective dataset cleaning.**
30. van Westen GJP, 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
31. Gleeson MP, Modi S, Bender A, et al. The Challenges involved in modeling toxicity data *in silico*: a review. *Curr Pharm Des* 2012;18(9):1266–91
32. Nikolova N, Jaworska J. Approaches to measure chemical similarity – a review.

- QSAR Comb Sci 2003; 22(9-10):1006–26
33. Maggiora GM. On outliers and activity cliffs—why QSAR often disappoints. *J Chem Inf Model* 2006;46(4):1535
 - **A paper that describes the breakdown of similarity principle and its effect on QSAR predictions.**
 34. Wassermann AM, Bajorath JR. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J Chem Inf Model* 2010;50(7):1248–56
 35. Leung CS, Leung SSF, Tirado-Rives J, Jorgensen WL. Methyl effects on protein–ligand binding. *J Med Chem* 2012;55(9):4489–500
 36. Stepan AF, Walker DP, Bauman J, et al. Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the united states. *Chem Res Toxicol* 2011;24(9):1345–410
 37. Gleeson P, Bravi G, Modi S, Lowe D. ADMET rules of thumb II: a comparison of the effects of common substituents on a range of ADMET parameters. *Bioorgan Med Chem* 2009;17(16):5906–19
 38. Gleeson MP. Generation of a set of simple, interpretable ADMET rules of thumb. *J Med Chem* 2008;51(4):817–34
 39. Roy K, Mitra I. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb Chem High Throughput Screen* 2011;14(6):450–74
 - **A good review paper that describes in detail metrics suitable for QSAR model validation.**
 40. Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 2010;29(6-7):476–88
 - **A detailed paper that describes best practices in QSAR model development.**
 41. Feher M, Ewing T. Global or local QSAR: is there a way out?. *QSAR Comb Sci* 2009;28(8):850–5
 42. Wood DJ, Buttar D, Cumming JG, et al. Automated QSAR with a hierarchy of global and local models. *Mol Inform* 2011;30(11-12):960–72
 43. Stouch TR, Kenyon JR, Johnson SR, et al. In silico ADME/Tox: why models fail. *J Comput Aided Mol Des* 2003;17(2-4):83–92
 - **A paper that describes real-world cases where the generation and application of in silico models failed. The authors discuss the reason why this occurred.**
 44. Seifert MHJ, Wolf K, Vitt D. Virtual high-throughput in silico screening. *BIOSILICO* 2003;1(4):143–9
 45. Kirchmair J, Williamson MJ, Tyzack JD, et al. Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J Chem Inf Model* 2012;52(3):617–48
 46. Xue L, Bajorath J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb Chem High Throughput Screen* 2000;3:363–72
 47. Leach AR. Molecular modelling. Principles and applications. Longman limited; Harlow: 1996
 48. Livingstone DJ. The characterization of chemical structures using molecular properties. a survey. *J Chem Inf Comput Sci* 1999;40(2):195–209
 49. Senese CL, Duca J, Pan D, et al. 4D-fingerprints, universal QSAR and QSPR descriptors. *J Chem Inf Comput Sci* 2004;44(5):1526–39
 50. Clark DE. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J Pharm Sci* 1999;88(8):815–21
 51. Veber DF, Johnson SR, Cheng H-Y, et al. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* 2002;45(12):2615–23
 52. Seelig A. A general pattern for substrate recognition by P-glycoprotein. *Eur J Biochem* 1998;251(1-2):252–61
 53. Lewis D. Structural determinants of cytochrome P450 substrate specificity, binding affinity and catalytic rate. *Chem Biol Interact* 1998;115:175–99
 54. Hansch C, Fujita T. ρ - σ - π Analysis. A Method for the correlation of biological activity and chemical structure. *J Am Chem Soc* 1964;86(8):1616–26
 55. Lombardo F, Gifford E, Shalaeva MY. In silico ADME prediction: data, models, facts and myths. *Mini Rev Med Chem* 2003;3(8):861–75
 56. Hansch C, Leo A, Mekapati SB, Kurup A. QSAR and ADME. *Bioorg Med Chem* 2004;12(12):3391–400
 57. Chohan KK, Paine SW, Waters NJ. Advancements in predictive in silico models for ADME. *Curr Chem Biol* 2008;2(3):215–28
 58. Tropsha A, Gramatica P, Gombar VK. The Importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci* 2003;22:69–77
 59. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 2007;26:694–701
 60. Roy K, Mitra I, Kar S, et al. Comparative studies on some metrics for external validation of QSPR models. *J Chem Inf Model* 2012;52(2):396–408
 61. Rodgers SL, Davis AM, van de Waterbeemd H. Time-series QSAR analysis of human plasma protein binding data. *QSAR Comb Sci* 2007;26(4):511–21
 62. Baldi P, Brunak S, Chauvin Y, et al. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000;16(5):412–24
 - **An excellent paper that describes in detail metrics suitable for the validation of classification-based QSAR models.**
 63. Davis AM, Keeling DJ, Steele J, et al. Components of successful lead generation. *Curr Top Med Chem* 2005;5(4):421–39
 - **An excellent paper that describes the lead generation process in detail, including filters suitable for HTS selection.**
 64. Lagorce D, Sperandio O, Galons H, et al. FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics* 2008;9(1):396
 65. van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards

- prediction paradise?. *Nat Rev Drug Discov* 2003;2(3):192–204
66. Norinder U, Bergstrom CA. Prediction of ADMET Properties. *ChemMedChem* 2006;1(9):920–37
67. Gola J, Obrezanova O, Champness E, Segall M. ADMET Property prediction: the state of the art and current challenges. *QSAR Comb Sci* 2006;25(12):1172–80
68. Wang J, Hou T. Chapter 5 recent advances on in silico ADME modeling. In: Ralph AW. editor *Annual reports in computational chemistry*. Amsterdam, The Netherlands. Volume 5 Elsevier; 2009. p. 101–27
69. Madden JC. In silico approaches for predicting ADME properties. In: Puzyn T, Leszczynski J, Cronin MT. editors *Recent advances in QSAR studies*. 8. Springer; Netherlands: 2010. p. 283–304
70. Enoch SJ, Cronin MT. A review of the electrophilic reaction chemistry involved in covalent DNA binding. *Crit Rev Toxicol* 2010;40(8):728–48
71. Gleeson MP, Hersey A, Hannongbua S. In-silico ADME models: a general assessment of their utility in drug discovery applications. *Curr Top Med Chem* 2011;11(4):358–81
72. Thompson RA, Isin EM, Li Y, et al. Risk assessment and mitigation strategies for reactive metabolites in drug discovery and development. *Chem Biol Interact* 2011;192(1–2):65–71
73. Tarcsay Á, Keseru GM. In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin Drug Met* 2011;7:299–312
74. Piparo L, Worth AP. Review of QSAR Models and Software Tools for predicting Developmental and Reproductive Toxicity. JRC Report 2011
75. Dow RL, Carpino PA, Gautreau D, et al. Design of a potent CB1 Receptor antagonist series: potential scaffold for peripherally-targeted agents. *Acs Med Chem Lett* 2012;3(5):397–401
76. Weiss MM, Williamson T, Babu-Khan S, et al. Design and preparation of a potent series of hydroxyethylamine containing β -secretase inhibitors that demonstrate robust reduction of. central β -amyloid. *J Med Chem* 2012. DOI: 10.1021/jm300119p
77. Davenport AJ, Möller C, Heifetz A, et al. Using electrophysiology and in silico three-dimensional modeling to reduce human ether-à-go-go related gene k⁺ channel inhibition in a histamine h3 receptor antagonist program. *Assay Drug Dev Technol* 2010;8(6):781–9
78. Gleeson MP, Hersey A, Montanari D, Overington J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat Rev Drug Discov* 2011;10(3):197–208
- **This paper highlights why it can be challenging to optimize both potency and ADMET parameters in unison.**
79. Bajorath J. Progress in computational medicinal chemistry. *J Med Chem* 2012;55(8):3593–4
80. Schneider G. Virtual screening: an endless staircase?. *Nat Rev Drug Discov* 2010;9(4):273–6
81. Hann MM. Molecular obesity, potency and other addictions in drug discovery. *Med Chem Comm* 2011;2(5):349–55
82. Hann MM, Oprea TI. Pursuing the lead-likeness concept in pharmaceutical research. *Curr Opin Chem Biol* 2004;8(3):255–63
83. Congreve M, Carr R, Murray C, Jhoti H. A 'Rule of Three' for fragment-based lead discovery?. *Drug Discov Today* 2003;8(19):876–7
84. Bickerton GR, Paolini GV, Besnard J, et al. Quantifying the chemical beauty of drugs. *Nat Chem* 2012;4(2):90–8
85. Gleeson MP, Davis AM, Chohan KK, et al. Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models. *J Comput Aided Mol Des* 2007;21(10–11):559–73
86. Martin YC. A bioavailability score. *J Med Chem* 2005;48(9):3164–70
87. Norinder U, Haeberlein M. Computational approaches to the prediction of the blood-brain distribution. *Adv Drug Deliv Rev* 2002;54(3):291–313
88. Di L, Kerns EH, Carter GT. Strategies to assess blood-brain barrier penetration. *Expert Opin Drug Discov* 2008;3(6):677–87
89. Hawkins D. Use and value of metabolism databases. *Drug Discov Today* 1999;4:466–71
90. Testa B, Balmat AL, Long A, Judson P. Predicting drug metabolism—an evaluation of the expert system METEOR. *Chem Biodivers* 2005;2(7):872–85
91. Kalgutkar AS, Gardner I, Obach RS, et al. A comprehensive listing of bioactivation pathways of organic functional groups. *Curr Drug Metab* 2005;6(3):161–225
92. Enoch SJ, Madden JC, Cronin MT. Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach. *SAR QSAR Environ Res* 2008;19(5–6):555–78
93. Segall MD. Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr Pharm Des* 2012;18(9):1292–310
- **A very interesting paper that describes in detail how multi-parameter optimization can be performed in drug discovery using computational methods.**
94. Berellini G, Springer C, Waters NJ, Lombardo F. In Silico Prediction of volume of distribution in human using linear and nonlinear models on a 669 compound data set. *J Med Chem* 2009;52(14):4488–95
95. Gleeson MP. Plasma protein binding affinity and its relationship to molecular structure: an in-silico analysis. *J Med Chem* 2007;50(1):101–12
96. Huuskonen J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J Chem Inf Comput Sci* 2000;40(3):773–7
97. Hopfinger AJ, Esposito EX, Llinas A, et al. Findings of the challenge to predict aqueous solubility. *J Chem Inf Model* 2009;49(1):1–5
98. Che J, King FJ, Zhou B, Zhou Y. Chemical and biological properties of frequent screening hits. *J Chem Inf Model* 2012;52(4):913–26
99. 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

- **An excellent, extensive review of issues associated with virtual screening in drug discovery.**

100. Ekins S, Honeycutt JD, Metz JT.
Evolving molecules using multi-objective optimization: applying to ADME/Tox.
Drug Discov Today
2010;15(11-12):451–60

Affiliation

Matthew Paul Gleeson^{†1} & Dino Montanari^{2,3}

[†]Author for correspondence

¹Kasetsart University,
Faculty of Science,
Department of Chemistry,
50 Phaholyothin Rd,
Chatuchak, Bangkok 10900,
Thailand
Tel: +66 2 562 5555 extn 2210;
Fax: +66 2 5793955;
E-mail: paul.gleeson@ku.ac.th

²Aptuit,
Medicines Research Centre,
Via A. Fleming, 2, 37135,
Verona, Italy

³H. Lundbeck A/S,
Ottiliavej 9, 2500 Valby,
Denmark

Evaluating the enthalpic contribution to ligand binding using QM calculations: effect of methodology on geometries and interaction energies†

Duangkamol Gleeson,^a Ben Tehan,^{*b} M. Paul Gleeson^{*c} and Jumras Limtrakul^d

Received 2nd April 2012, Accepted 27th June 2012

DOI: 10.1039/c2ob25657f

As a result of research on ligand efficiency in the pharmaceutical industry, there is greater focus on optimizing the strength of polar interactions within receptors, so that the contribution of overall size and lipophilicity to binding can be decreased. A number of quantum mechanical (QM) methods involving simple probes are available to assess the H-bonding potential of different heterocycles or functional groups. However, in most receptors, multiple features are present, and these have distinct directionality, meaning very minimalist models may not be so ideal to describe the interactions. We describe how the use of gas phase QM models of kinase protein–ligand complex, which can more closely mimic the polar features of the active site region, can prove useful in assessing alterations to a core template, or different substituents. We investigate some practical issues surrounding the use of QM cluster models in structure based design (SBD). These include the choice of the method; semi-empirical, density functional theory or *ab-initio*, the choice of the basis set, whether to include implicit or explicit solvation, whether BSSE should be included, *etc.* We find a combination of the M06-2X method and the 6-31G* basis set is sufficiently rapid, and accurate, for the computation of structural and energetic parameters for this system.

1. Introduction

A variety of studies have helped to highlight the important contribution that individual interactions can have on the overall protein binding energy of a ligand. These include detailed studies on the characteristic interactions made by a variety of different functional groups^{1,2} with amino acid residues, the characteristics of π – π stacking,^{3–6} cation– π ⁷ and anion– π ⁸ interactions, halogen bond interactions,^{9,10} as well as the unique conformational preferences of different functional groups.^{11,12} Indeed, recent analyses of isothermal calorimetry data by Keseru *et al.*,^{13–15} who advocate the assessment of both the enthalpic and entropic contributions to the binding affinity, have noted how the focus on entropic gains in potency are not as productive

(*i.e.*, increasing lipophilicity and driving potency through the hydrophobic effect). Optimization efforts that focus on improving the enthalpic contribution to protein binding, by directly improving the polar interactions between the ligand and receptor, are preferable. In fact, the authors note that the undesirable focus on entropic potency gains is one of the key reasons for the increase in lipophilicity and molecular weight of drugs and drug candidates over time.¹⁵ In addition, it helps to explain the observation that historical drugs generally have lower potencies, lipophilicity and molecular weight compared to compounds in current, or recent development.^{16,17}

In light of the recent focus on ligand efficient molecules,^{18–23} there now appears to be a greater emphasis on improving the efficiency of the lead template/series, rather than achieving potency gains due to addition of lipophilicity. The latter is typically achieved by filling lipophilic pockets, displacing labile water, or incorporating extensive non-polar linkers to target more distant polar interactions, often resulting in questionable overall gain. This is because the resultant increase in overall lipophilicity and/or molecular weight to increase potency can have a significant detrimental effect on a wide variety of Adsorption Distribution Metabolism Excretion and Toxicity parameters (ADMET).^{24–26}

Understanding the interactions between functionality on a ligand with that in a protein active site is critical to improving potency in an efficient manner. A more ideal approach is to optimize the available enthalpic interactions present in a template, with the use of additional approaches to increase potency

^aDepartment of Chemistry, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

^bHeptares Therapeutics, BioPark, Broadwater Road, Welwyn Garden City, Herts, AL7 3AX, United Kingdom.

E-mail: ben.tehan@heptares.com; Tel: +44-1707-358646

^cDepartment of Chemistry, Faculty of Science, Kasetsart University, 50 Phaholyothin Rd, Chatuchak, Bangkok 10900, Thailand.

E-mail: paul.gleeson@ku.ac.th; Fax: +66-2-5793955;

Tel: +66-2-562-5555 extn 2210

^dDepartment of Chemistry, Faculty of Science, and Center for Advanced Studies in Nanotechnology and its Applications in Chemical, Food and Agricultural Industries, Kasetsart University, Bangkok 10900, Thailand

†Electronic supplementary information (ESI) available: All structural parameters and energies and energy correlation matrix plots and tables. See DOI: 10.1039/c2ob25657f

afterwards as needed.^{13–15} This is not a trivial task, but could be achieved by leveraging calorimetry binding data and structure based design (SBD) techniques. The latter technique is extensively used in drug discovery programs with structural data of the target, to rationally design increases in potency or selectivity into a lead series. The use of experimental structures derived from X-ray or NMR, can be used in isolation or in conjunction with computational chemistry.²⁷ The latter method presents program teams with a means to rationally design and test new molecules that can better leverage the interactions and steric features present in the protein.

Theoretical models of protein–ligand complexes can be generated in a number of different ways. Rapid, molecular mechanical (MM) methods can be used to sample whole protein models quickly (or over long timescales),²⁸ linear scaling semi-empirical methods can be used to simulate the whole protein system quantum mechanically (QM),^{29–32} or QM/MM methods^{33–40} can be used to simulate the active site using QM and the remainder using MM. Alternatively, smaller, approximate models can be used at higher levels of QM theory to evaluate particular regions of interest more rapidly.^{41–44}

Each of the methods discussed above offers distinct advantages in particular circumstances. For example, MM methods are very quick to evaluate, meaning extensive sampling is possible. However, non-standard templates, metals or certain interactions are not ideally described.^{9,32,45} Semi-empirical QM methods are relatively rapid, allowing large clusters to be considered or whole proteins in linear scaling form, but are not considered the most accurate as a result of the approximate method used.^{46,47} QM/MM allows the use of accurate QM methods to treat the important core regions, and take into account longer range effects using MM, however interactions across the boundary region can lead to issues.^{40,48} QM clusters allow the use of very accurate levels of theory to study the key interactions between a protein and ligand, however the effect of the surrounding protein is therefore completely neglected. QM cluster calculations are nevertheless employed for many tasks including the prediction of interaction strengths between model ligands and probes,^{42–44,49–51} to more complex tasks such as reaction mechanism elucidation^{52,53} and X-ray structure refinement.^{47,54}

In previous reports the authors have investigated the use of QM/MM methods to study protein kinase-inhibitor complexes, showing the distinct benefits of this method over traditional docking in ligand pose scoring.⁵⁵ A follow up to this study highlighted the potential use of this method in aiding refinement of the active site region where non-standard ligands are present.⁵⁶ Subsequent investigations were carried out on smaller, but more rapidly computable QM cluster models, consisting of the ligand and active site residues that make the key interactions.⁵⁷ While this approach neglects the effect of the protein and solvent, it allows a researcher to assess how optimal the interactions between the moieties present are, and whether they can be improved.

As illustrated in Fig. 1, if the ligand conformation or interactions in the optimized active site model differ significantly from the experimental protein–ligand structure: this suggests that either the conformation/interactions present are not optimal to make the best possible interactions, due to unfavourable sterics for example. Alternatively, the structure might change dramatically because the initial ligand parameters used in the refinement

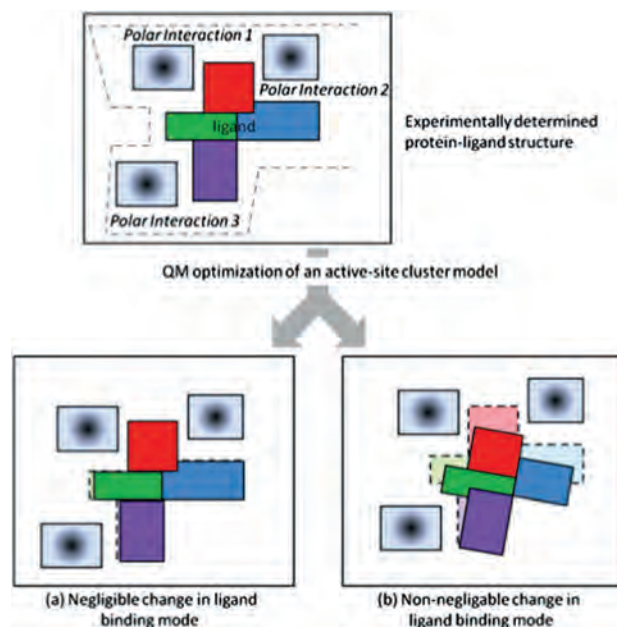


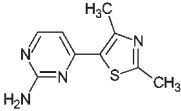
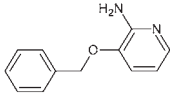
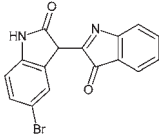
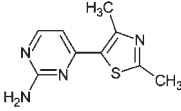
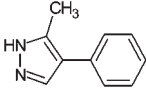
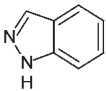
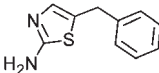
Fig. 1 An illustration of how QM active site models could be employed to aid in the optimization of the enthalpic contribution to overall binding energy. In case (a), a negligible change in the structure occurs on optimization suggesting the interactions present are optimal, since when the protein is removed they do not change dramatically. In case (b), the ligand conformation changes dramatically on optimization suggesting (1) the polar ligand interactions are not optimal and should be improved by alteration of the template substituents or (2) the refined ligand coordinates obtained from the experimental structure were sub-optimal.

step were sub-optimal. In contrast, a negligible change in structure without the surrounding protein present suggests that the enthalpic interactions between the ligand and the key residues are optimal. Thus, understanding the strength of interactions, and the preferred conformations adopted by a molecule in a receptor are important pre-requisites to allow the rational, efficient optimization of a lead series to be performed. A number of methods are available to predict the strength of interaction of individual functional groups which can prove extremely insightful in the design and modification of lead series.^{41,43,44,49–51}

In this work we consider the use of small QM models of receptors, consisting of the key polar active site interactions, rather than generic probes. The region selected is not as extensive as used in the approach of Gueto-Tettay, who used residues within 5 Å of the active site, which, due to the significant size, necessitates the use of the semi-empirical PM6 method.⁴¹ Here we investigate smaller, yet more interaction relevant active site models. We are not, *per se*, interested in predicting the much more challenging absolute binding free energy,²⁸ rather, the goal is to determine whether such methods could be used to assess the relative interaction strength of inhibitors with key polar elements of a receptor, with the view to using them to rapidly assess alternative modifications of lead series to improve the contribution to enthalpic binding.

For these initial studies, we have employed the cluster based approach using a variety of conditions to understand the impact

Table 1 Kinase-inhibitor structure used in this study. Reported are the PDB ID, inhibitor structure, resolution, kinase target, target pIC₅₀ and a description of the H-bonds mediated with the hinge. Outer (O), central (C) and inner (I) HBs correspond to those defined in Fig. 1. CH refers to a short interaction distance between a carbonyl group of the hinge and a CH hydrogen atom of the inhibitor

PDB ID	Inhibitor	Resolution	Target	Activity ^a	H bond pattern
1PXJ ⁶⁸		2.3	CDK2	IC ₅₀ = 6.5 uM ^{68,69}	O(CH), C, I
1W7H ⁷⁰		2.2	P38	IC ₅₀ = 1300 uM ⁷¹	O(CH), C, I
2BHE ⁷²		1.9	CDK2	IC ₅₀ = 2 uM ⁷²	O, C, I
2C5O ⁶⁹		2.1	CDK2	K _i = 6.5 uM ^{68,69}	O, C, I(CH)
2UVX ⁷³		2.0	PKA-B	IC ₅₀ > 100 uM	O(CH), C, I
2UW3 ⁷⁴		2.2	PKA-B	IC ₅₀ = 80 uM	C, I
2VTA ⁷⁵		2.0	CDK2	IC ₅₀ = 185 uM ⁷⁵	O, C, I(CH)
3DND ⁷⁶		2.3	CDK2	IC ₅₀ = 16 uM ⁷⁶	O(CH), C, I

^a SD in activity <1 log unit which means the binding energies of these molecules differ no more than 1.4 kcal mol⁻¹ on average.

of the choice of methods in such assessments. We have attempted to quantify the effect of using different methodologies on a set of cluster models generated from a set of 8 PDB structures we have previously reported on (Table 1). We consider a number of different factors in this study, including; (a) the choice of model system (*i.e.*, a QM active site model containing the key residues), (b) the choice of QM method (*i.e.*, semi-empirical, density functional theory or *ab-initio*), (c) the size of the basis set, (d) should solvation be included, (e) should basis set superposition error be considered when assessing binding energies.

2. Computational procedures

Crystal structures of the 8 protein-kinases listed were downloaded from the RCSB protein databank (www.rcsb.org) (Table 1). These structures were chosen such that the ligands only made polar interactions with the three amino acid residues that constitute the “hinge” region (*i.e.* no water mediated interactions were present). For a detailed description of the structural features of the protein kinase target class, see ref. 77.

The 8 truncated protein models consisted of the backbones of the 3 hinge amino acids involved in binding the adenine portion of ATP. The amino acid sidechains were replaced by hydrogen atoms. The QM representation used in this study is exemplified in Fig. 2 and has been employed by both us and others to elucidate aspects of non-bonded interactions in kinase-inhibitor complexes.⁵⁷ The C_α atoms of the truncated amino acids were frozen during geometry optimization.

Geometry optimization of QM models was performed using Gaussian 03⁵⁸ at the following levels of theory: MP2/6-31+G**, M06-2X/6-31G*, HF/6-31G*, HF/3-21G, AM1. These span the time-consuming, to very rapid methods. M06-2X is an increasingly popular, newer, DFT method that has performed better in recent benchmarking studies than the more common B3LYP method.^{59–61} For the purpose of comparison, a purely MM based approach was also investigated, consisting of the CHARMM force field as implemented in Discovery Studio 2.5 with empirically derived Momany-Rone atomic charges.⁶² The effect of including an implicit solvent model of water was also investigated for the M06-2X/6-31G* and HF/3-21G models using a polarizable continuum model (PCM).

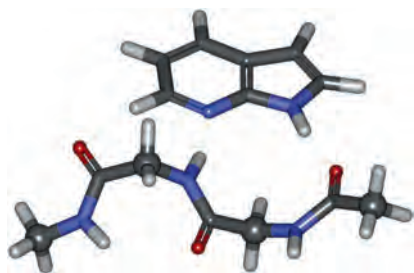


Fig. 2 An illustration of the QM model used in this study. The polar interactions of the proteins are denoted using the backbone atoms of 3 amino acids that constitute the hinge region. The amino acids sidechains were removed and replaced by hydrogen atoms. The AA chain was terminated one SP³ carbon atom after the nearest amide heteroatom. All atoms in the calculation were flexible except for the C_α atoms (denoted with a ball representation).

Computed interaction energies were obtained by subtracting the energy of the optimized, isolated ligand, and protein acid model, from the energy of the complex. In addition, the effect of correcting the energies for basis set superposition error (BSSE)⁶³ was considered (counterpoise correction) for the M06-2X/6-31G* and HF/3-21G models.

3. Results and discussion

We have evaluated a number of different methodologies that can be used to generate small QM models of the polar active site interactions found within typical protein ligand complexes. This was done in order to understand how ideal the polar interactions are in the absence of the extended protein environment, and discuss their suitability in terms of their RMSDs, in addition to analyzing H-bond distances and how they compare to their corresponding X-ray structures. It should be noted that minor changes in the distances and angles of a particular interaction can result in subtle differences in the positioning of a ligand within an active site pocket, and which in turn may significantly affect the choice of substituents, or where additional growth is considered.

A benefit of using such easy to construct, albeit approximate models, is that we can rapidly evaluate how ideal the interactions are between the polar active site features and the ligand. As discussed above, deviations in the interactions or binding conformation on removing the extended protein might suggest that the ligand binding mode is either sub-optimal due to high conformational strain or sub-optimal polar interactions due to steric constraints imposed by the extended protein, or potentially due to sub-optimal refinement.^{64,65}

As discussed in our previous study,⁵⁷ where we earlier reported the results at the MP2/6-31+G(d,p) level for this dataset alone, we found both of the scenarios above had occurred. Briefly, models for 3DND, and 1W7H, in particular, showed dramatic differences between the QM optimized models and X-ray results. Rotation of the non-polar benzyl and benzyloxy groups in the two complexes, respectively, led to lower energy, preferred conformations (*i.e.*, significant strain energy present). This also led to a dramatic change in the polar interactions in the case of the former, however this is also likely to be

affected by another factor. In 3DND, the ligand lies in a position rather distant from the hinge, but on QM optimization, the H-bonds (and the atypical C–H...O=C interaction that is frequently seen in kinases) decrease dramatically.

Other effects were independent of the protein and more likely due to issues regarding the ligand fitting to relatively poor density. In 2C5O for example, the pyrimidin-2-amine and thiazole group are planar with respect to each other. However, on optimization the groups adopt a more plausible angle of ~37°. Indeed, the same ligand was also found in the structure 1PXJ where it displayed an angle of 39°, apparently confirming that the refinement process led to the former result.

Furthermore, unusually short and long H-bond distances were observed in 2BHE, 3DND, 1PXJ and 2C5O. In particular 2BHE displays a very short H-bond to the central H-bond acceptor that on optimization increases to approximately 1.8 Å. In addition, 2C5O displays a very short C–H...O=C bond (~2.0 Å), which increases to the more realistic value (~2.4 Å) from an analysis of known kinase X-ray structures sourced from the PDB databank. These results also suggest that less attention was spent assessing the chemical accuracy of the interactions in question, compared to the empirical fitting to the density, which in these cases is not ideal.^{64–66}

It should be noted that the resolutions of the X-ray structures used here are typical of those used in SBD studies (~1.9–2.3 Å). However, those studied are not necessarily at an ideal standard to compare theoretical results to. This is because the structures are (a) not completely representative of a protein–ligand complex in solution, at 37° and (2) that the atomic coordinates that have been derived are not error free.^{36,64–66} Indeed, these structures typically lack any information regarding hydrogen atom positions, and sometimes contain poorly positioned ligands, especially in cases where inhibitors are non-standard,⁶⁶ or only weakly potent.⁶⁴ We therefore also make reference to higher resolution experimental structural data taken from comparable interactions⁵⁷ found in the Cambridge Structural Database (CSD) (www.ccdc.cam.ac.uk/products/csd). This contrasts to comparative studies by others who have used the original electron density as a reference.⁶⁵

3.1 Effect of methodology of QM active site structures

For molecular systems of the size employed here, the MP2/6-31+G(d,p) calculations are resource intensive, requiring days per complex to optimize on Intel Core i7 workstations. Thus, even these calculations might be prohibitive when used in SBD exercises, or in support refinement studies. In typical SBD applications, multiple template modifications or alternate substituents may require evaluation, and the use of less computationally expensive methods may therefore be warranted. For example, if 10 alternatives to the heterocyclic template were considered, and another 10 modifications in terms of the points substitution, or substituent types, 100 different calculations would be required. A solution would be to employ DFT, semi-empirical method, or MM based methods in such studies. We have therefore investigated the use of a number of different methodologies for use in probing active site models, ranging from the very slow (MP2/6-31+G(d,p)), to the very fast (AM1 and CHARMM calculations).

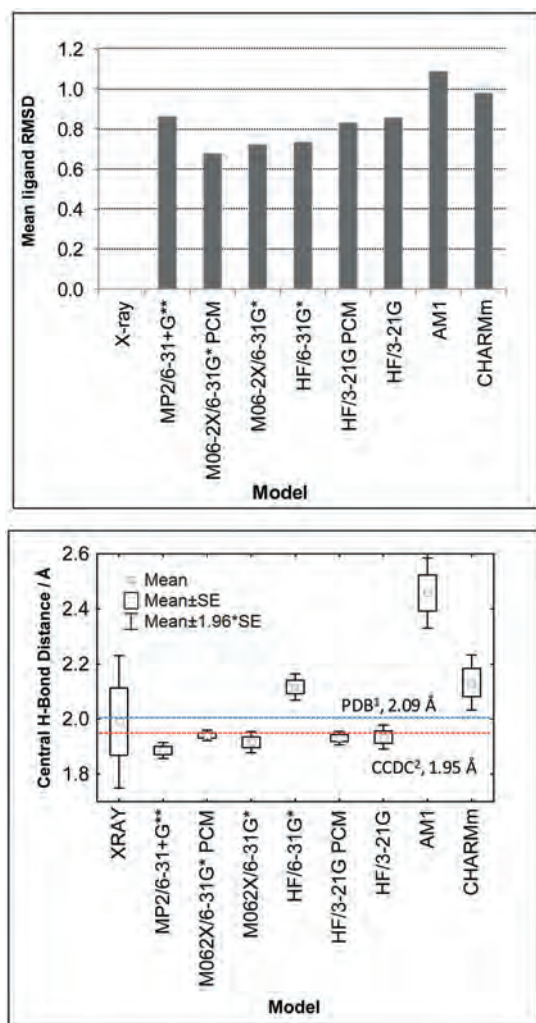


Fig. 3 (a) Plot of the mean RMSD of the optimized gas phase models to the original X-ray structures and (b) a box plot summary of the central hinge H-bond interactions. ¹Taken from analysis of PDB kinase complexes. ²Taken from analysis of CSD small molecule interactions.

A summary of the structural results obtained in this study is presented in Fig. 3 and 4, and the ESI Table S1.[†] We present the results in terms of the H-bond interaction distances from the different models and also the RMSD to the original X-ray coordinates (Fig. 3). We can also compare these distances to benchmark values obtained from a search of high resolution kinase X-ray structures containing heterocyclic inhibitors, and comparable interactions in high resolution, small molecule crystal structures. We observed that there is a general trend toward lower mean RMSDs with the increasing accuracy of the computation method used (Fig. 3a). However it is clear that the MP2 based results do not show the best agreement with the original X-ray results. Although the X-ray structures are not ideal standards, a general trend to lower RMSDs is still expected to be a reasonable measure of computational success, at least up to a point. A further measure that one can use to assess the overall quality of the method is the predicted interaction distances.

As a result of the ~2.0 Å resolution of the X-ray structures, and the lack of hydrogen information (atoms were added using

the AMBER forcefield⁵⁷), the X-ray structure derived distances are not ideal. We also make reference to distances derived from comparable high resolution small molecule crystal structures.⁷⁸ We can compare these values to (a) the mean value from the 8 X-ray structures used here, (b) the mean over comparable, high resolution structures reported in the PDB (c) and the mean distance between a heterocyclic nitrogen and an amide based on those reported in the small molecule Cambridge Structural Database (CSD).⁵⁷ In Fig. 3b it can be seen that the average distance between the ligand hetero-atom and the hinge H-bond donor, generally improves with increasing level of theory, albeit with the MP2 based method again being an outlier. AM1, CHARMm and HF/6-31G* show mean values higher than the mean of the original X-ray complexes, or benchmark values taken from the CSD and PDB sources. In contrast, the MP2/6-31+G**, M06-2X variants and HF/3-21G variants show lower means than the mean of the original X-ray complexes, or benchmark values from the PDB. Apart from the MP2/6-31+G** set, the mean values are very close to the values obtained from the CSD reference set suggesting the models here have lost a degree of their kinase character. These results also show that the neglect of the protein environment generally leads to a greater association between the protein model and the ligand. For the MP2 based result the effect is even more pronounced suggesting that the increased accuracy of the method is not beneficial since the structures deviate more significantly from those in the protein environment. Note, this does not mean that isolated QM active site models have no value in SBD. Indeed, if this was the case then data from the CSD would probably not prove useful in design efforts.⁷⁸ The value of a simplified QM model is that it represents the best case interaction between the moieties concerned, without external electrostatic or VDW constraints. Alteration of the real ligand in the protein environment, so that it can adopt the preferred low energy conformation observed in the gas phase, may help to maximise the intermolecular interaction.

Looking more broadly at the structural results, we can see that the trends identified using the computationally demanding MP2/6-31+G**,⁵⁷ are reproduced using the M06-2X methods and both HF 3-21G based models. AM1 and CHARMm models have generally larger, but also much more variable, distances between the ligand and protein hinge model compared to the other methods, and experimental benchmark values.⁵⁷ This is perhaps unsurprising in that rigorous charge derivation for MM methods is reported to be needed, or additional terms added. In addition, AM1 semi-empirical methods are being superseded by the newer PM6 variants, as well as PM6 with additional customization.^{46,54} HF/6-31G* models seem to systematically underestimate the association compared to the M06-2X and MP2 based models. The inclusion of water solvent was also investigated using an implicit PCM solvent model. The M06-2X/6-31G* and HF/3-21G models were reoptimized using the PCM model. The results in Fig. 3 show that the mean RMSD and central H-bond interaction at the hinge are slightly lower compared to the related gas phase optimization. Given the considerable computational overhead, such treatment may not therefore be warranted, at least in terms of an assessment of the structural features.

The results reported here indicate that the structures obtained can vary noticeably depending on the method used. Validation of the method for the system under investigation should be

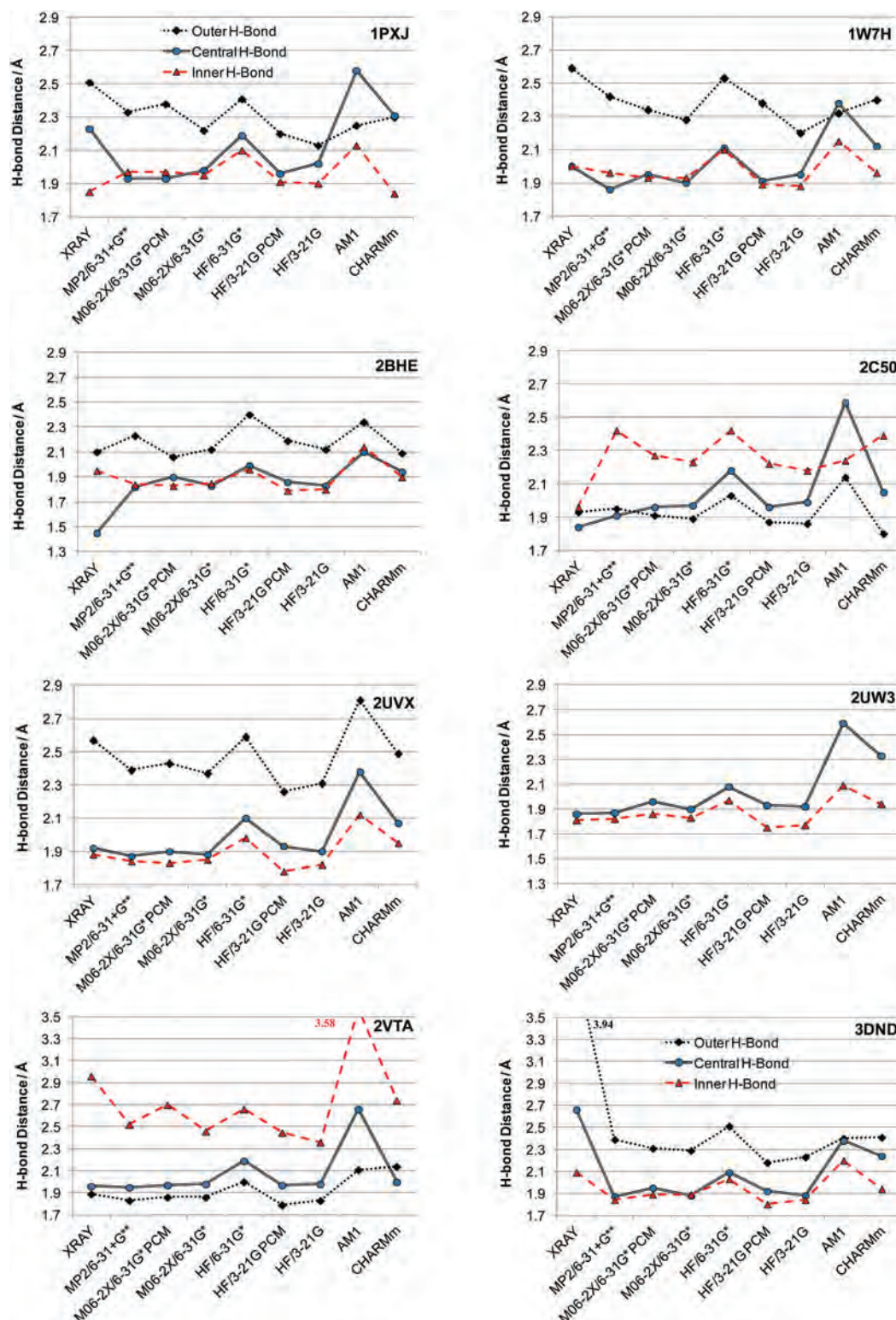


Fig. 4 A comparison of the H-bond distances obtained from 8 theoretical models and the original X-ray coordinates.

undertaken to ensure that the method can reliably account for the interactions present within the system in question. These results show that relatively rapid methods could be used to assess interactions of non-charged heterocycles. In particular, the well

validated M06-2X method with a modestly sized basis set, gave optimized structures with both the lowest RMSDs, and interaction distances closest to benchmark values, at relatively modest computational expense.

3.2 Effect of methodology of QM interaction energies

An additional application of QM cluster models is in the computation of enthalpies of binding. The goal of such a method is not to compute a realistic binding free energy, rather it is to try and assess the strength of polar interactions between a molecule and a probe (or active site representation in our case). For example, a number of methods are available to assess the H-bonding potential of different heterocycles or functional groups.^{41,43,44,49–51} The basis of such methods is that substituents or frameworks that have the optimal potential to interact with the polar features of an active site should lead to greater binding. However, in most receptors multiple features are present which have distinct directionality, meaning simple models may not be so ideal to compute interaction energies. The use of a model more closely mimicking the polar features of the active site might prove advantageous in assessing alterations to a core template, or different templates completely.

It is not expected that simplistic enthalpies will correlate strongly with the experimental free energy related parameters such as the K_i or IC_{50} , especially for such a diverse set of templates, across a range of protein kinases, as sampled in this study. Indeed, in this study it should be noted that the dataset chosen here consists of molecules with moderate to low potency for their particular kinase (Table 1). The observed standard deviation of 0.6 log units corresponds to just a 0.91 kcal mol⁻¹ difference in energy according to the Arrhenius equation, which is below the accuracy of many theoretical methods. Thus, even in the best case scenario, a correlation between the predicted interaction energy and the activity would not be expected (especially since the contribution of hydrophobic effects also need to be considered in any evaluation). Nevertheless, in this study we are interested in examining the magnitude of the differences in interaction energies for the different methods assessed, as each method treats H-bond interactions, bond lengths and angles *etc.* to different degrees of accuracy. These differences will have a dramatic effect on the rank ordering, which is especially pertinent if used in a design setting. For example, diffuse functions are suggested in cases where negative charges are present as delocalized can occur within the higher orbitals. The presence of halogen bonds necessitates additional parameters for PM6, and will be poorly described using MM methods for example.⁹

It is important to note that the MP2/6-31+G** energies are the most rigorous that have been obtained here. However, the optimized geometries deviate slightly more from the X-ray coordinates than those from M06-2X for example. Nevertheless, they are expected to be the most suitable here in terms of describing the interactions and conformational energies in the systems under investigation. Thus, we compare the interaction energies of all methods to these benchmark values (Table S2†).

The correlation between the energies obtained at the MP2/6-31+G**, M06-2X/6-31G*, HF/6-31G*, HF/3-21G, AM1 and CHARMM are reported in the ESI (Fig. S1†). The MP2/6-31+G** energies correlate well with those at M06-2X/6-31G* ($r^2 = 0.74$) and HF/6-31G* ($r^2 = 0.83$). Methods such as HF/3-21G and AM1, relying on smaller basis sets, do not correlate as well, with r^2 's of 0.54 and 0.36, respectively. The CHARMM

based energies show no correlation with the MP2 based results, or any other QM measure.

Also investigated was the effect of BSSE, a common artifact in QM calculations that can lead to inaccurate interaction energies. BSSE arises due to orbitals in the combined complex, which have negligible overlap, and can in fact lead to a lowering of the overall energy in the combined complex compared to the isolated components. This effect can be removed in the QM calculation of each individual component by including the ghost orbitals of the other component. The results from BSSE calculations at the M06-2X/6-31G* and HF/3-21G models are reported in the ESI (Fig. S2†). The correlation between the BSSE corrected energy and the uncorrected value for HF/3-21G displayed an r^2 of 0.74, while the value for the calculation at M06-2X/6-31G* level was 0.97. The effect of a common solvent model (PCM) was also investigated for both the M06-2X/6-31G* and HF/3-21G models, and these results are also reported in Fig. S2.† The M06-2X results including a PCM solvent model of water correlates only moderately well with the gas phase energies ($r^2 = 0.57$) while those at the HF/3-21G level display an r^2 of just 0.27. These results also highlight the dramatic effect the inclusion of solvent can have on the rank ordering for a given method.

The overall correlation between the different energies can be appreciated more clearly using principal components analysis (PCA). PCA is a method for identifying small numbers of correlated, orthogonal components for a dataset containing many descriptors. The QM energies (descriptors), and the kinase QM models (observations), that show a high degree of inter correlation will be located in the same region of component space on the combined scores/loadings bi-plot. In this case, a two component model can describe over 80% of the total variation in the dataset of 10 descriptors and 8 observations (Fig. 5). The

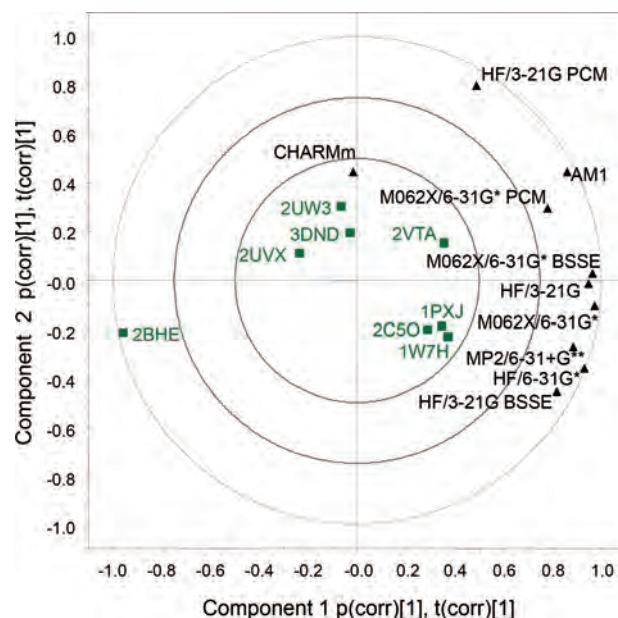


Fig. 5 PCA loadings bi-plot highlighting the inter-correlation between the different computed energies. The 2 component model describes 82% of the total variation (66 and 16% respectively for components 1 and 2) in the 10 energies computed for the 8 different model proteins.

combined loading/score plots show that all of the QM energies display a significant degree of correlation on component 1, as can be seen by their large positive loading. The CHARMM based result correlates poorly with the QM results, since it is located close to the origin on component 1. On component 2, the HF/3-21G PCM, AM1, M06/6-31G* PCM and the CHARMM model deviate more significantly from the other computed energies as can be discerned from their more positive loadings.

These results appear to suggest that the use of a moderately sized basis set, such as 6-31G* is preferred, especially as the effect of BSSE is minimal. The fact that the M06-2X method gives energies close to those of MP2/6-31+G**, and also good geometries (unlike HF/6-31G* for example which also correlates well), suggests it may be a preferred method to compute interactions energies. However, given the impact of the implicit solvent correction on the energies, it may also be beneficial to evaluate this term given its possible impact on rank ordering.

4. Conclusions

A number of methods are available to assess the H-bonding potential of different heterocycles or functional groups.^{41,43,44,49–51} The advantage of such methods is that substituents or frameworks that have the potential to more effectively interact, will presumably lead to greater binding efficiency with a receptor with an opposing feature (assuming it does not interact with water to a greater extent). However, in most receptors multiple features are present and these have distinct directionality meaning simple models may not be so ideal to compute interaction energies.

The use of more representative cluster models, more closely mimicking the polar features of a specific active site, might prove advantageous in assessing substituent alterations to a template, or different templates altogether. Understanding the strength of the polar interactions formed between a ligand and the active site is important if we wish to improve the formers enthalpic binding efficiency.^{13–15} Such an understanding would be beneficial in our attempts to increase the ligand efficiency of molecules in development and concomitantly improve their ADMET characteristics.^{16,67}

In this study we have assessed the effect of using a number of different theoretical methods to optimize QM active site models of protein kinase–ligand complexes. We employed MP2/6-31+G**, M06-2X/6-31G*, HF/6-31G*, HF/3-21G, AM1 and CHARMM methods, and considered the effect of BSSE and the inclusion of an implicit solvation model. We are interested in the effect these different choices have on the structures and energetics obtained for the systems in question. The results reported here on small, active site models, indicate that the structures obtained can vary noticeably depending on the method used. Validation of the method for the system under investigation should be undertaken to ensure that the method can reliably account for the interactions present within the system in question.

These results show that relatively rapid methods could be used to assess interactions of non-charged heterocycles, using the well validated M06-2X method with a modestly sized basis set,

giving optimized structures with both the lowest RMSDs, and interaction distances closest to benchmark values, at relatively modest computational expense. Analysis of the computed energies shows that a significant degree of correlation exists between the methods. The effect of BSSE on the rank ordering of the ligands in this study is negligible with a moderately sized basis set such as 6-31G*. The effect of PCM was shown to be more significant and may warrant consideration. The observation that the M06-2X method gives energies close to those of MP2/6-31+G**, and also reasonable optimized geometries, suggests it is the preferred method here for computing interactions energies.

The information derived from such models could be used to guide the ranking and selection of substituents or heterocyclic templates to improve their ligand efficiency by maximizing polar interactions. Alternately small QM models (or more descriptive QM/MM models⁴⁷) could be employed to benchmark ligand conformation and active site interactions which could be used to guide the refinement of X-ray structures, particularly of low to moderate resolution. We believe that while such calculations certainly have limitations, they have a place in SBD applications, alongside methods such as experimental X-ray structure, CSD structural analyses, QM/MM calculations of full protein–ligand complexes, with each offering a different insight into the interactions found within biological complexes.

Acknowledgements

This work was supported by the Thailand Research Fund grant RSA5480016 (MPG) and RMU5180032 (DG) and the Faculty of Science, Kasetsart University (KU).

Notes and references

- 1 Y. N. Imai, Y. Inoue and Y. Yamamoto, *J. Med. Chem.*, 2007, **50**, 1189–1196.
- 2 T. Zhou, D. Z. Huang and A. Caffisch, *Curr. Top. Med. Chem.*, 2010, **10**, 33–45.
- 3 S. Grimme, *Angew. Chem., Int. Ed.*, 2008, **47**, 3430–3434.
- 4 L. Estevez, N. Otero and R. A. Mosquera, *J. Phys. Chem. A*, 2009, **113**, 11051–11058.
- 5 E. A. Meyer, R. K. Castellano and F. Diederich, *Angew. Chem., Int. Ed.*, 2003, **42**, 1210–1250.
- 6 L. M. Salonen, M. Ellermann and F. Diederich, *Angew. Chem., Int. Ed.*, 2011, **50**, 4808–4842.
- 7 R. Wu and T. B. McMahon, *J. Am. Chem. Soc.*, 2008, **130**, 12554–12555.
- 8 Y. Zhao, H. T. Ng and E. Hanson, *J. Chem. Theory Comput.*, 2009, **5**, 2726–2733.
- 9 P. Zhou, J. W. Zou, F. F. Tian and Z. C. Shang, *J. Chem. Inf. Model.*, 2009, **49**, 2344–2355.
- 10 Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang and W. Zhu, *J. Med. Chem.*, 2009, **52**, 2854–2862.
- 11 K. A. Brameld, B. Kuhn, D. C. Reuter and M. Stahl, *J. Chem. Inf. Model.*, 2008, **48**, 1–24.
- 12 M.-H. Hao, O. Haq and I. Muegge, *J. Chem. Inf. Model.*, 2007, **47**, 2242–2252.
- 13 G. G. Ferenczy and G. M. Keseru, *J. Chem. Inf. Model.*, 2010, **50**, 1536–1541.
- 14 G. M. Keseru and G. G. Ferenczy, *Drug Discovery Today*, 2010, **15**, 919–932.
- 15 G. M. Keseru and G. M. Makara, *Nat. Rev. Drug Discovery*, 2009, **8**, 203–212.
- 16 M. P. Gleeson, A. Hersey, D. Montanari and J. Overington, *Nat. Rev. Drug Discovery*, 2011, **10**, 197–208.
- 17 M. M. Hann, *Med. Chem. Commun.*, 2011, **2**, 349–355.

- 18 C. Abad-Zapatero and J. T. Metz, *Drug Discovery Today*, 2005, **10**, 464–469.
- 19 S. D. Bembenek, B. A. Tounge and C. H. Reynolds, *Drug Discovery Today*, 2009, **14**, 278–283.
- 20 A. L. Hopkins, C. R. Groom and A. Alex, *Drug Discovery Today*, 2004, **9**, 430–431.
- 21 E. Perola, *J. Med. Chem.*, 2010, **53**, 2986–2997.
- 22 C. H. Reynolds, S. D. Bembenek and B. A. Tounge, *Bioorg. Med. Chem. Lett.*, 2007, **17**, 4258–4261.
- 23 C. H. Reynolds and M. K. Holloway, *ACS Med. Chem. Lett.*, 2011, **2**, 433–437.
- 24 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.
- 25 M. P. Gleeson, *J. Med. Chem.*, 2008, **51**, 817–834.
- 26 P. Gleeson, G. Bravi, S. Modi and D. Lowe, *Bioorg. Med. Chem.*, 2009, **17**, 5906–5919.
- 27 A. R. Leach, *Molecular Modelling. Principles and Applications*, Longman limited, Harlow, 1996.
- 28 A. de Ruiter and C. Oostenbrink, *Curr. Opin. Chem. Biol.*, 2011, **15**, 547–552.
- 29 S. L. Dixon and K. M. J. Merz, *J. Chem. Phys.*, 1996, **104**, 6643–6649.
- 30 K. Raha, A. J. van der Vaart, K. E. Riley, M. B. Peters, L. M. Westerhoff, H. Kim and K. M. Merz, *J. Am. Chem. Soc.*, 2005, **127**, 6583–6594.
- 31 K. M. Merz and M. B. Peters, *J. Chem. Theory Comput.*, 2006, **2**, 383–399.
- 32 P. Dobeš, J. Řezáč, J. i. Fanfrlík, M. Otyepka and P. Hobza, *J. Phys. Chem. B*, 2011, **115**, 8581–8589.
- 33 U. C. Singh and P. A. Kollman, *J. Comput. Chem.*, 1986, **7**, 718–730.
- 34 A. Warshel, *Annu. Rev. Biophys. Biomol. Struct.*, 2003, **32**, 425–443.
- 35 T. Vreven, K. S. Byun, I. Komáromi, S. Dapprich, J. A. Montgomery, K. Morokuma and M. J. Frisch, *J. Chem. Theory Comput.*, 2006, **2**, 815–826.
- 36 X. Li, S. A. Hayik and K. M. Merz, *J. Inorg. Biochem.*, 2010, **104**, 512–522.
- 37 L. C. Menikarachchi and J. A. Gascon, *Curr. Top. Med. Chem.*, 2010, **10**, 46–54.
- 38 H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.*, 2009, **48**, 1198–1229.
- 39 H. Lin and D. G. Truhlar, *Theor. Chem. Acc.*, 2007, **117**, 185–199.
- 40 S. Shaik, S. Cohen, Y. Wang, H. Chen, D. Kumar and W. Thiel, *Chem. Rev.*, 2010, **110**, 949–1017.
- 41 C. Gueto-Tettay, J. Drosos and R. Vivas-Reyes, *J. Comput.-Aided Mol. Des.*, 2011, **25**, 583–597.
- 42 M. B. Peters, K. Raha and K. M. Merz, *Curr. Opin. Drug Discovery*, 2006, **9**, 370–379.
- 43 R. Villar, M. J. Gil, J. I. Garcia and V. Martinez-Merino, *J. Comput. Chem.*, 2005, **26**, 1347–1358.
- 44 J. Schwöbel, R.-U. Ebert, R. Kühne and G. Schüürmann, *J. Comput. Chem.*, 2009, **30**, 1454–1464.
- 45 R. S. Paton and J. M. Goodman, *J. Chem. Inf. Model.*, 2009, **49**, 944–955.
- 46 M. Korth, *J. Chem. Theory Comput.*, 2010, **6**, 3808–3816.
- 47 U. Ryde, *Dalton Trans.*, 2007, 607–625.
- 48 H. Hu and W. Yang, *Annu. Rev. Phys. Chem.*, 2008, **59**, 573–601.
- 49 J. Schwöbel, R.-U. Ebert, R. Kühne and G. Schüürmann, *J. Phys. Chem. A*, 2009, **113**, 10104–10112.
- 50 M. Nocker, S. Handschuh, C. Tautermann and K. R. Liedl, *J. Chem. Inf. Model.*, 2009, **49**, 2067–2076.
- 51 J. A. H. Schwöbel, R.-U. Ebert, R. Kühne and G. Schüürmann, *J. Phys. Org. Chem.*, 2011, **24**, 1072–1080.
- 52 P. E. M. Siegbahn and F. Himo, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 323–336.
- 53 P. Siegbahn and F. Himo, *JBIC, J. Biol. Inorg. Chem.*, 2009, **14**, 643–651.
- 54 U. Ryde and K. Nilsson, *J. Mol. Struct. (THEOCHEM)*, 2003, **632**, 259–275.
- 55 M. P. Gleeson and D. Gleeson, *J. Chem. Inf. Model.*, 2009, **49**, 1437–1448.
- 56 M. P. Gleeson and D. Gleeson, *J. Chem. Inf. Model.*, 2009, **49**, 670–677.
- 57 M. P. Gleeson, S. Hannongbua and D. Gleeson, *J. Mol. Graphics Modell.*, 2010, **29**, 507–517.
- 58 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. M. Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian Inc., Wallingford CT, 2009.
- 59 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
- 60 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. C*, 2008, **112**, 6860–6868.
- 61 Y. Zhao and D. G. Truhlar, *Chem. Phys. Lett.*, 2011, **502**, 1–13.
- 62 Discovery Studio 2.5, Accelrys Inc, San Diego, CA, United States <http://accelrys.com/products/discovery-studio/>.
- 63 F. B. van Duijneveldt, J. G. C. M. van Duijneveldt-van de Rijdt and J. H. van Lenthe, *Chem. Rev.*, 1994, **94**, 1873–1885.
- 64 A. M. Davis, S. J. Teague and G. J. Kleywegt, *Angew. Chem., Int. Ed.*, 2003, **42**, 2718–2736.
- 65 D. Yusuf, A. M. Davis, G. J. Kleywegt and S. Schmitt, *J. Chem. Inf. Model.*, 2008, **48**, 1411–1422.
- 66 A. Malde and A. Mark, *J. Comput.-Aided Mol. Des.*, 2011, **25**, 1–12.
- 67 P. D. Leeson and B. Springthorpe, *Nat. Rev. Drug Discovery*, 2007, **6**, 881–890.
- 68 S. Wang, C. Meades, G. Wood, A. Osnowski, S. Anderson, R. Yuill, M. Thomas, M. Mezna, W. Jackson, C. Midgley, G. Griffiths, I. Fleming, S. Green, I. McNae, S.-Y. Wu, C. McInnes, D. Zheleva, M. D. Walkinshaw and P. M. Fischer, *J. Med. Chem.*, 2004, **47**, 1662–1675.
- 69 G. Kontopidis, C. McInnes, S. R. Pandalaneni, I. McNae, D. Gibson, M. Mezna, M. Thomas, G. Wood, S. Wang, M. D. Walkinshaw and P. M. Fischer, *Chem. Biol.*, 2006, **13**, 201–211.
- 70 M. J. Hartshorn, C. W. Murray, A. Cleasby, M. Frederickson, I. J. Tickle and H. Jhoti, *J. Med. Chem.*, 2004, **48**, 403–413.
- 71 A. L. Gill, M. Frederickson, A. Cleasby, S. J. Woodhead, M. G. Carr, A. J. Woodhead, M. T. Walker, M. S. Congreve, L. A. Devine, D. Tisi, M. O'Reilly, L. C. A. Seavers, D. J. Davis, J. Curry, R. Anthony, A. Padova, C. W. Murray, R. A. E. Carr and H. Jhoti, *J. Med. Chem.*, 2004, **48**, 414–426.
- 72 R. Jautelat, T. Brumby, M. Schäfer, H. Briem, G. Eisenbrand, S. Schwahn, M. Krüger, U. Lücking, O. Prien and G. Siemeister, *Chem-BioChem*, 2005, **6**, 531–540.
- 73 A. Donald, T. McHardy, M. G. Rowlands, L.-J. K. Hunter, T. G. Davies, V. Berdini, R. G. Boyle, G. W. Aherne, M. D. Garrett and I. Collins, *J. Med. Chem.*, 2007, **50**, 2289–2292.
- 74 G. Saxty, S. J. Woodhead, V. Berdini, T. G. Davies, M. L. Verdonk, P. G. Wyatt, R. G. Boyle, D. Barford, R. Downham, M. D. Garrett and R. A. Carr, *J. Med. Chem.*, 2007, **50**, 2293–2296.
- 75 P. G. Wyatt, A. J. Woodhead, V. Berdini, J. A. Boulstridge, M. G. Carr, D. M. Cross, D. J. Davis, L. A. Devine, T. R. Early, R. E. Feltell, E. J. Lewis, R. L. McMenamin, E. F. Navarro, M. A. O'Brien, M. O'Reilly, M. Reule, G. Saxty, L. C. A. Seavers, D.-M. Smith, M. S. Squires, G. Trewartha, M. T. Walker and A. J. A. Woolford, *J. Med. Chem.*, 2008, **51**, 4986–4999.
- 76 J. Orts, J. Tuma, M. Reese, S. K. Grimm, P. Monecke, S. Bartoschek, A. Schiffer, K. U. Wendt, C. Griesinger and T. Carlomagno, *Angew. Chem., Int. Ed.*, 2008, **47**, 7736–7740.
- 77 J. J. L. Liao, *J. Med. Chem.*, 2007, **50**, 409–424.
- 78 F. H. Allen and R. Taylor, *Chem. Soc. Rev.*, 2004, **33**, 463–475.

CONCISE ARTICLE

Ring opening polymerization of mannosyl tricyclic orthoesters: rationalising the stereo and regioselectivity of glycosidic bond formation using quantum chemical calculations†

Cite this: *Med. Chem. Commun.*, 2013, **4**, 265

Siwarutt Boonyarattanakalin,^a Somsak Ruchirawat^{bc} and M. Paul Gleeson^{*d}

Received 2nd July 2012

Accepted 19th October 2012

DOI: 10.1039/c2md20178j

www.rsc.org/medchemcomm

Quantum chemical calculations have been used to assess the physico-chemical origin of the stereo and regio-selectivity of polymerisation reactions of glycosyl tricyclic orthoesters. From the theoretical reaction pathway we find that subtle modulation of steric and electronic effects at the initiation event can dramatically influence the nature of the polymer products.

Infectious diseases such as tuberculosis (TB) remain a globally life-threatening health problem.¹ TB is a particular problem in developing countries as the long term treatment of the disease using antibiotics is financially unviable.² Further research is undoubtedly needed therefore, to allow the development of more, rapid and cost effective treatments.

Lipomannan (LM) is one of the key glycolipids that comprise the unique cell envelope of *Mycobacterium tuberculosis* (Mtb). Consisting of a α (1–6) mannopyranan backbone,³ LMs have been implicated during the infectious, virulent, and survival events in host mammalian cells.^{4,5} Thus, improved understanding of LM interactions with the host immune system should help in the development of improved treatments for TB.⁶ Unfortunately, many experimental studies on this glycolipid are impeded by the limited amount of naturally occurring oligosaccharides.

Yongyat *et al.*⁷ have previously reported a synthetic approach utilizing mannosyl tricyclic orthoesters as monomers for regio- and stereocontrolled polymerizations to generate α (1–6) mannopyranan. The Lewis acids, trimethylsilyl trifluoromethanesulfonate (TMSOTf) and boron trifluoride etherate (BF₃·Et₂O) were used as catalysts to promote the

polymerizations of two of the monomers including 3,4-*O*-benzyl- β -D-mannopyranose 1,2,6-orthobenzoate (**1**) and 3,4-*O*-benzyl- β -D-mannopyranose 1,2,6-orthopivalate (**2**) (Fig. 1). From a single chemical transformation step, polymers of different lengths and differing degrees of regio- and stereo-selectivity were obtained. It was found that under the same conditions; (a) TMSOTf leads to longer and more selective α (1–6) mannan chains when compared to BF₃ and (b) that the monomer 3,4-*O*-benzyl- β -D-mannopyranose 1,2,6-orthobenzoate (**1**) gives rise to longer, and more selective α (1–6) chains when compared to 3,4-*O*-benzyl- β -D-mannopyranose 1,2,6-orthopivalate (**2**).

To facilitate the design of alternate methods to control regio- and stereo-selectivity, and to help improve reaction yields, quantum chemical (QC) calculations have been undertaken, the goal is to try and understand the physico-chemical origin of the control in polymerisation results described above. To this end, we explore the structures and energies associated with the critical activation step (Fig. 2).

Simulations were performed using complete molecular models of monomers **1** and **2**, the active form of TMSOTf (*i.e.* Si(CH₃)₃⁺), and BF₃. The initiation pathway (Fig. 1) was characterised using density functional theory (DFT) calculations at the M062X/6-31+G** level with a polarizable continuum model (PCM) consisting of dichloromethane. M062X is an increasingly popular DFT method that has performed better in recent benchmarking studies compared to the more common B3LYP method.⁸ Frequency analyses were used to fully characterise all

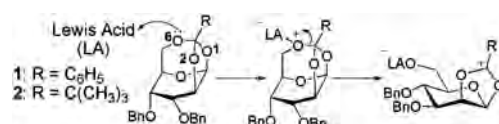


Fig. 1 The initiation step of monomer **1** and **2**, prior to the formation of α (1–6) mannopyranan.

^aSchool of Bio-Chemical Engineering and Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani 12121, Thailand

^bLaboratory of Medicinal Chemistry, Chulabhorn Research Institute, 54 Moo 4, Vibhavadee-Rangsit Highway, Laksi, Bangkok 10210, Thailand

^cChemical Biology Program, Chulabhorn Graduate Institute, and the Center for Environmental Health, Toxicology and Management of Chemicals, 54 Moo 4, Vibhavadee-Rangsit Highway, Laksi, Bangkok 10210, Thailand

^dDepartment of Chemistry, Faculty of Science, Kasetsart University, 50 Phaholyothin Rd, Chatuchak, Bangkok 10900, Thailand. E-mail: paul.gleeson@ku.ac.th; Fax: +66-2-5793955; Tel: +66-2-562-5555 ext. 2210

† Electronic supplementary information (ESI) available: The 3D coordinates of the optimized structures (mol2 format) and the tables of the enthalpies and free energies are provided. See DOI: 10.1039/c2md20178j

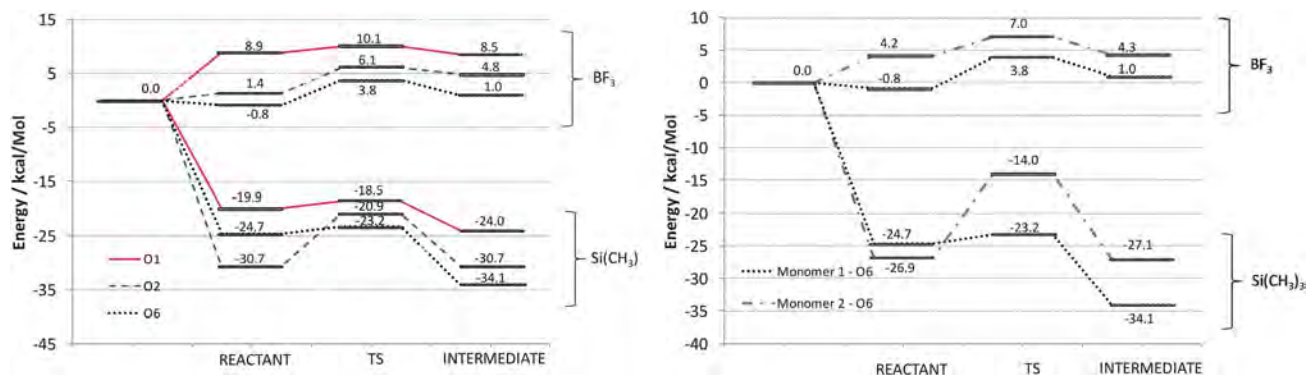


Fig. 2 Reaction pathways obtained for monomer **1**, at the O1, O2 and O6 positions, with two different initiators (left) and pathways obtained for monomers **1** and **2** using two different initiators, at the O6 position (right).

minima and transition states and to compute thermochemical information. The potential energy surface was corrected for vibrational, rotational and translational motion.⁹ All minima and transition state energies are reported relative to the isolated, solvated reactants. All calculations were performed in Gaussian 09.¹⁰ Basis set superposition errors are known to be less significant for DFT based methods so were not evaluated.¹¹ The energetic pathways associated with the initiation step of the 2 different glycosyl tricyclic orthoesters (GTO) are reported in Fig. 2. We have obtained the reactant (the non-bonded complex between monomer and initiator), the initiated intermediate, and the transition state that connects them. We begin our discussion with the preferred position of activation, followed by the initiator and finally the monomer.

GTOs contain three suitable Lewis bases on the three oxygen atoms (labelled as O1, O2, and O6 in Fig. 1) that are connected to the orthoester carbon (C_{orth}). Upon monomer activation by a Lewis acid, the bond between the C_{orth} and the activated oxygen atom is broken. The resulting carbocation ion intermediate is then capable of propagating the polymerisation process (Fig. 1). The first step in the process involves the formation of the initial non-bonded complex, or reactant as termed here.

The non-bonded reactant structures obtained with BF₃ display B–O interaction distances of ~1.6 Å while the larger (CH₃)₃Si⁺ based complexes display Si–O interaction distance of ~1.85 Å. These distances drop to ~1.5 Å and ~1.7 Å respectively on reaction with Lewis acid to form the intermediate. We find that irrespective of the position of attachment, the (CH₃)₃Si⁺ based complexes formed with monomer **1** are considerably lower in energy than the corresponding BF₃ based structures. Indeed, it should be noted that the large difference in binding energy observed (~30 kcal mol⁻¹) is considerably larger than the expected BSSE errors obtained for a system of comparable size using similar DFT methods (<5 kcal mol⁻¹).¹²

Of the three possible oxygen atoms that can be activated (O1, O2, and O6, Fig. 1), activation at the O6 position is preferred thermodynamically irrespective of the Lewis acid (Fig. 2). This is consistent with the fact that the breaking of the O6–C_{orth} bond leads to the least sterically hindered intermediate. The intermediate formed by the breaking of the O2–C_{orth} bond is found to be lower in energy than that formed by the cleavage of the

O1–C bond. Both the O1 and O2 intermediates form fused eight and six membered ring systems. However the O1 intermediate is the least stable as the resultant cyclic ring must span a longer distance between the C2 and C6 atoms. In contrast, activation at the O6 position results in a more stable bicyclic intermediate composed of a fused five and six membered ring system which spans the adjacent C1 and C2 atoms. The relative energy of the intermediates are also found to correlate reasonably well with the cleaved C⁺–O distance. We find that the shorter this distance on average over all the intermediates observed, the lower the energy of the intermediate. For the O1 position, the C⁺–O1 distances are ~2.65 and 2.65 Å for (CH₃)₃Si⁺ and BF₃ respectively, 2.44 Å and 2.34 Å for the O2 position respectively, and 2.43 Å and 2.35 Å for the C⁺–O6 position respectively (Fig. 3). Energetically, we find that BF₃ results in less favourable intermediates than (CH₃)₃Si⁺ due to the fact that the positively charged (CH₃)₃Si⁺ can interact more effectively with the resultant monomer on cleavage. In addition, while the order of the stability of the 3 possible intermediates is the same, the BF₃ intermediates are considerably higher in energy since they are zwitterionic structures in the non-polar solvent, rather than

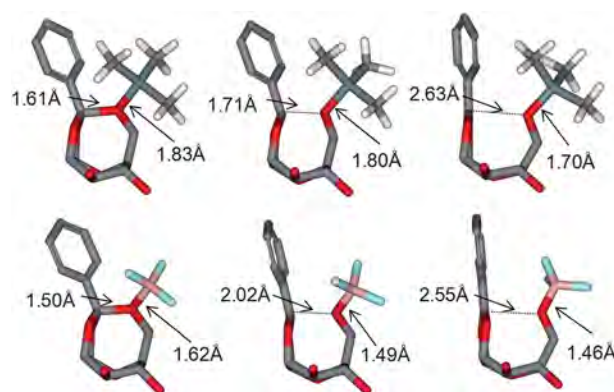


Fig. 3 Optimized structures for monomer **1** pathway corresponding to the reactants (left), transition states (middle) and intermediates (right) for (CH₃)₃Si⁺ (top) and BF₃ (bottom). Benzyl groups and monomer H atoms have been removed for clarity. Key distances are given. The BF₃ based initiator results in a later transition state, closer in structure to the intermediate.

cationic for $(\text{CH}_3)_3\text{Si}^+$. The latter Lewis acid forms the most stable intermediate, this being a pre-requisite for further polymerization. This helps to explain why the experimental yields with this Lewis acid are greater.⁷

It is also necessary to explain why BF_3 is experimentally observed to form less selective polymers compared to $(\text{CH}_3)_3\text{Si}^+$. Based on the calculated data obtained here, this appears to be a function of both the kinetics and thermodynamics of the initiation step (Fig. 2). For the BF_3 initiated process for monomer **1** the energies of the initial non-bonded complexes for positions 2 and 6 are similar and the corresponding barriers to reaction are rather close at 4.7 and 4.6 kcal mol⁻¹, respectively. While the O1 barrier is the lowest of all, it is the least stable of all the BF_3 based reactants, and also gives rise to the highest energy intermediate.

For $(\text{CH}_3)_3\text{Si}^+$, we observe that the initial non-bonded complex formed at the O2 position is preferred over the O6 position due to the unfavourable steric interaction of the initiator with the C6 methylene. This effect is absent in case of the smaller BF_3 Lewis acid. As is the case for BF_3 , the O1 position with $(\text{CH}_3)_3\text{Si}^+$ also displays a low barrier to reaction, but the reactant and intermediates are also of high energy. It can therefore be concluded that the rate determining barrier associated with the O6 position is dramatically lower than that for the next preferable position (O2), being 1.5 and 9.8 kcal mol⁻¹, respectively.

We postulate that the reaction to form the O6 intermediate is critical because it leads to the formation of the most preferable intermediate, which also displays a very high barrier to re-form the non-bonded complex (10.9 kcal mol⁻¹). These results therefore suggest that $(\text{CH}_3)_3\text{Si}^+$ will selectively form comparatively large quantities of the O6 intermediate, while BF_3 can presumably activate at both the O2 and O6, and possibly O1. The subsequent nucleophilic attack of the carbocation intermediate by the sterically more accessible O6 nucleophile of an additional monomer is further complicated due to steric interference from the bridged ring system and the OBn substituents, as well as electronic effects from the ring oxygens. Polymerization with initiators at the O1 and O2 positions is expected to be much more challenging compared to the O6 position as the former two cases contain more sterically hindered points of attack. Attack at the C1 position of the O6-intermediate (Fig. 1) is preferential since it is both less sterically hindered and because the two adjacent oxygen atoms can better stabilize the resulting transition state with their lone pairs. Thus, it is clear that the monomer/initiator combination that give rise to the energetically most favourable and accessible intermediate, and which displays a large energy gap to the other possible intermediates, will result in the most selective, longer chained polymers.

From an analysis of the optimized transition state structures we indeed find evidence showing that the better the stabilizing effect on the C^+ by the leaving group, the lower the energy of the transition state. We find that the C^+-O distances in the transition state do indeed correlate with the relative energy. The longest interaction distance is observed for O1 position, being 1.86 and 2.19 Å for BF_3 and $(\text{CH}_3)_3\text{Si}^+$, respectively, compared to

1.73 and 2.08 Å, respectively, for O2, and 1.69 and 2.02 Å, respectively for O6. The results also show that the $(\text{CH}_3)_3\text{Si}^+$ transition states lie closer to the reactant state than BF_3 , consistent with the Hammonds postulate.

Finally, we investigated why monomer **2** is found to have experimentally lower reaction yields and poorer selectivity under the same experimental conditions as monomer **1**. We therefore investigated how the replacement of the phenyl ring on the C_{orth} of monomer **1** with *tert*-butyl affected the reaction profile at the preferred O6 position. We find that polymerization reactions involving monomer **2** with both BF_3 and $(\text{CH}_3)_3\text{Si}^+$ result in higher energy barriers to reaction, and higher intermediate energies (Fig. 2). We find that the O6-intermediate of monomer **2** obtained with BF_3 is 3.3 kcal mol⁻¹ higher in energy than the corresponding value for monomer **1**, while that for $(\text{CH}_3)_3\text{Si}^+$ is 7.1 kcal mol⁻¹. This is due to the generally longer $\text{C}^+-\text{O6}$ interaction distances, which is indicative of reduced electronic stabilization of the carbocation. This helps to explain why monomer **2** is found to have experimentally lower reactivity compared to monomer **1**.⁷

Conclusions

The results presented here show that the initiation step is the critical step in the polymerisation process reported by Yongyat *et al.*⁷ Formation of $\alpha(1-6)$ bonds requires the selective activation of the first monomer at the O6 position, with favourable energies (increased quantities), leading to longer, more regio- and stereoselective linear chains. This can be achieved with the bulky initiator, TMSOTf and a stabilizing capping group (*i.e.* phenyl ring) that can effectively stabilise the resulting carbocation ion of the intermediate *via* resonance. It is also clear that the positively charged Lewis acid $\text{Si}(\text{CH}_3)_3^+$ results in lower energy intermediates than does neutral BF_3 by providing more effective stabilization of the carbocation center. This is particularly important since the reaction is performed in non-polar solvent which cannot provide effective stabilization of the high energy species formed over the course of the reaction.

The dimerization step of the O6-intermediate leads to the more selective formation of $\alpha(1-6)$ bond since the C1 position in this intermediate is the preferential point of attack from the most sterically unhindered O6 atom of another monomer. This is because the C1 position is less sterically hindered and the resulting transition state can be more effectively stabilized by the two adjacent oxygen atoms. With less sterically hindered initiators (*i.e.* BF_3), or monomers with more poorly stabilizing, and bulky capping groups (*i.e.* *tert*-butyl), the relative preference for the $\alpha(1-6)$ glycosidic bond decreases. As a result, the polymerisation process proceeds in a disorderly fashion, *via* a sterically hindered and less energetically favourable pathway, resulting in shorter, more disorderly chains.

The information derived from such models could be used to guide the selection of the more optimal substituent at the orthoester position (R group in Fig. 1), and the initiators. Such calculations are currently in active use to design further experiments. QC methods are often seen as rather unapproachable, yet as seen here, they can provide a means to

post-rationalise complex results and provide a method to quickly simulate alternative reagents in a matter of days, where a comparable synthetic approach may take days, weeks or even months.

Acknowledgements

Financial support for this work from the Chulabhorn Research Institute and the Thailand Research Fund (RSA5480016 for MPG and RSA5580059 for SB) are gratefully acknowledged.

Notes and references

- 1 M. W. Borgdorff, K. Floyd and J. F. Broekmans, *Bull. W. H. O.*, 2002, **80**, 217–227.
- 2 S. H. Kaufmann, *Nat. Rev. Immunol.*, 2006, **6**, 699–704.
- 3 P. Andersen, *Nat. Rev. Microbiol.*, 2007, **5**, 484–487.
- 4 V. Briken, S. A. Porcelli, G. S. Besra and L. Kremer, *Mol. Microbiol.*, 2004, **53**, 391–403.
- 5 N. N. Driessen, E. J. M. Stoop, R. Ummels, S. S. Gurcha, A. K. Mishra, G. Larrouy-Maumus, J. Nigou, M. Gilleron, G. Puzo, J. J. Maaskant, M. Sparrius, G. S. Besra, W. Bitter, C. M. J. E. Vandenbroucke-Grauls and B. J. Appelmelk, *Microbiology*, 2010, **156**, 3492–3502.
- 6 P. Hutacharoen, S. Ruchirawat and S. Boonyarattanakalin, *J. Carbohydr. Chem.*, 2011, **30**, 415–437.
- 7 C. Yongyat, S. Ruchirawat and S. Boonyarattanakalin, *Bioorg. Med. Chem.*, 2010, **18**, 3726–3734.
- 8 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
- 9 J. B. Foresman and A. Frisch, *Exploring Chemistry With Electronic Structure Methods: A Guide to Using Gaussian*, Gaussian Inc., Wallingford CT, ISBN0963676938, 1996.
- 10 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09 Rev C01*, Gaussian Inc., Wallingford CT, 2009.
- 11 H. Valdés, V. Klusák, M. Pitoňák, O. Exner, I. Starý, P. Hobza and L. Rulíšek, *J. Comput. Chem.*, 2008, **29**, 861–870.
- 12 D. Gleeson, B. Tehan, M. P. Gleeson and J. Limtrakul, *Org. Biomol. Chem.*, 2012, **10**, 7053–7061.

Computational study of EGFR inhibition: molecular dynamics studies on the active and inactive protein conformations

Napat Songtawee · M. Paul Gleeson ·
Kiattawee Choowongkamon

Received: 24 November 2011 / Accepted: 2 August 2012
© Springer-Verlag 2012

Abstract The structural diversity observed across protein kinases, resulting in subtly different active site cavities, is highly desirable in the pursuit of selective inhibitors, yet it can also be a hindrance from a structure-based design perspective. An important challenge in structure-based design is to better understand the dynamic nature of protein kinases and the underlying reasons for specific conformational preferences in the presence of different inhibitors. To investigate this issue, we performed molecular dynamics simulation on both the active and inactive wild type epidermal growth factor receptor (EGFR) protein with both type-I and type-II inhibitors. Our goal is to better understand the origin of the two distinct EGFR protein conformations, their dynamic differences, and their relative preference for Type-I inhibitors such as gefitinib and Type-II inhibitors such as lapatinib. We discuss the implications of protein dynamics from a structure-based design perspective.

Keywords EGFR · Kinase · Molecular dynamics · AMBER force field · GROMACS · Gefitinib · Lapatinib

Abbreviations

EGFR	Epidermal growth factor receptor
G-loop	Glycine-rich loop
A-loop	Activation loop
R-spine	Regulatory spine
H-cluster	Hydrophobic cluster
DFG motif	Asp-Phe-Gly conserved motif
HRD motif	His-Arg-Asp conserved motif
PDB	Protein data bank
MD	Molecular dynamics
RMSD	Root average square deviation
RMSF	Root average square fluctuation
SD	Standard deviation

Electronic supplementary material The online version of this article (doi:10.1007/s00894-012-1559-0) contains supplementary material, which is available to authorized users.

N. Songtawee · K. Choowongkamon
Department of Biochemistry, Faculty of Science,
Kasetsart University,
50 Phaholyothin Rd, Chatuchak,
Bangkok 10900, Thailand

M. P. Gleeson (✉)
Department of Chemistry, Faculty of Science,
Kasetsart University,
50 Phaholyothin Rd, Chatuchak,
Bangkok 10900, Thailand
e-mail: paul.gleeson@ku.ac.th

K. Choowongkamon (✉)
Center for Advanced Studies in Tropical Natural Resources,
NRU-KU, Kasetsart University,
50 Phaholyothin Rd, Chatuchak,
Bangkok 10900, Thailand
e-mail: fsciktc@ku.ac.th

Introduction

Protein kinases are an important class of therapeutic targets in drug discovery. At present, eight kinase inhibitors are currently marketed as anti-cancer treatments [1], and it has been estimated that approximately one-third of all pharmaceutical research projects are dedicated to such targets [2]. Three of the eight marketed kinase drugs target the epidermal growth factor receptor (EGFR), also known as ErbB1 kinase. A wealth of biochemical and structural information has been generated on this target, offering us considerable insight into the structure, function and inhibition of this important therapeutic target class [3–5].

Over 160 unique protein kinase X-ray structures have been deposited in the RCSB (<http://www.rcsb.org/resources/kinases>), offering a great deal of information to aid in the design of new or improved kinase-directed therapies. The protein kinase domain of EGFR is comprised of two lobes: a

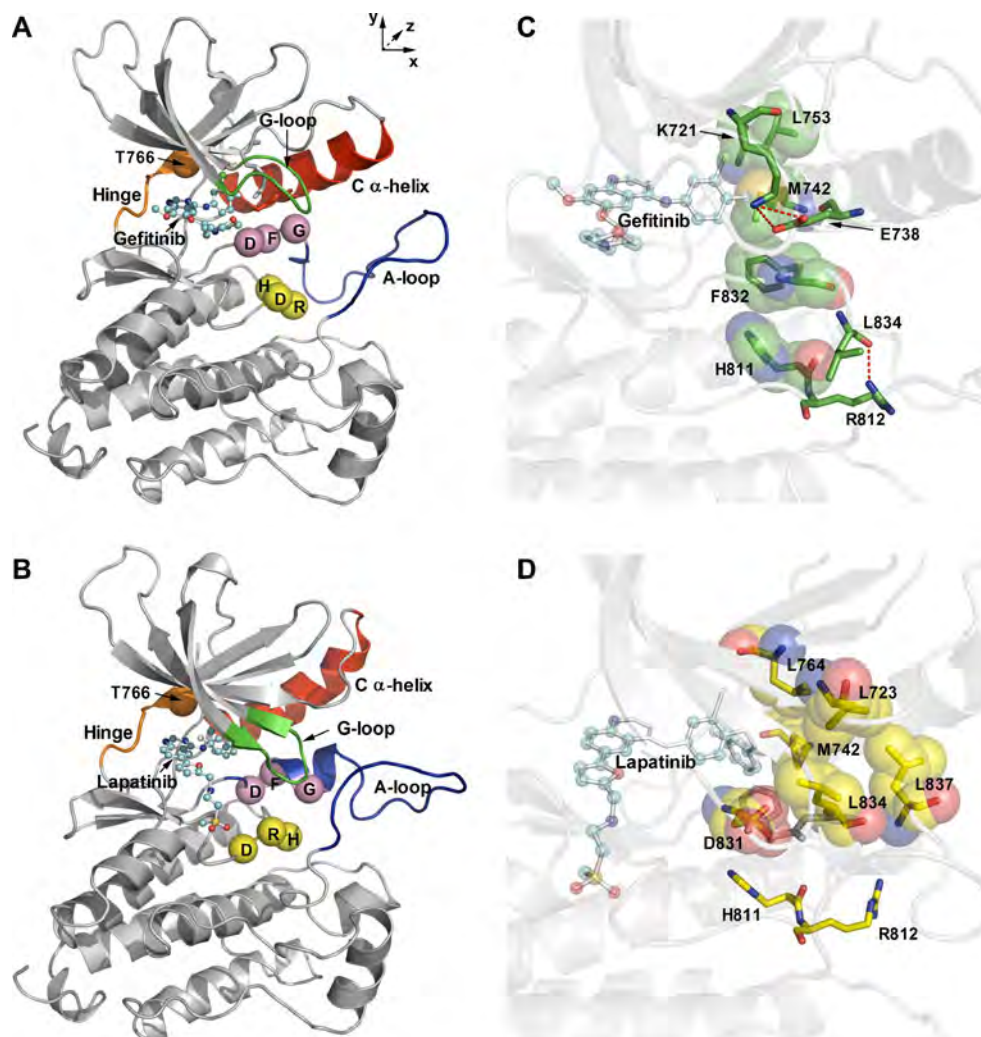
smaller N-terminal lobe consisting mainly of β -strands and a single large α -helix; and a larger C-terminal lobe, which is almost exclusively α -helical. The ATP-binding site is located at the hinge region between the lobes, meaning the active site is dynamic in size and shape. The structure of EGFR kinase can be further divided into a number of structural regions, as highlighted in Fig. 1a,b. These include the glycine-rich loop (G-loop), the C α -helix on the N-lobe, the activation loop (A-loop), and conserved DFG and HRD motifs on the C-lobe. Several important features for EGFR activation include: (1) reorientation of the C α -helix closer to the ATP-binding site, resulting in the formation of a salt bridge between E738 on the helix and the conserved K721 residue on the β 5-strand. The latter also interacts with the α - and β -phosphates of ATP (see supporting information Fig. S5A). (2) The positioning of DFG-D831 and HRD-D813 residues to interact with the ATP phosphate groups and the peptide substrate, respectively; and (3) extension of the A-loop, and translation away from the active site; (4) the formation of the regulatory spine (R-spine) by three hydrophobic (M742, L753, F832) and one polar residue (H811), leading to a H-bond between the HRD-

R812 and the DFG+1-L834, which is proposed to help maintain the active kinase conformation (Fig. 1c) [6, 7]. Several other residues, including L723, M742, L764, D831, L834 and L837, are proposed to form a small hydrophobic cluster (H-cluster) between the C α -helix and the A-loop, which is also believed to be important for stabilizing the inactive conformation of EGFR kinase (Fig. 1d) [7, 8].

The structural diversity observed across protein kinases, resulting in subtly different active site cavities, as well as the often distinctly different protein conformations, is highly desirable in the pursuit of selective inhibitors, yet it is also can be a hindrance from a structure-based design perspective. For example, analysis of the active EGFR-gefitinib crystal structure (PDB accession code: 2ITY) would suggest that the addition of the substituent 1-methoxy,3-F-phenyl to the quinazoline template would not be tolerated. However, not only is this substituent tolerated, it is believed that the resultant complex, along with increased ErbB2 activity, give lapatinib its improved efficacy (PDB accession code: 1XKK).

Kinase inhibitors can be classified into two to three distinct categories [1, 9, 10]. Type-I inhibitors target the

Fig. 1 Ribbon representations of active (a) and inactive (b) epidermal growth factor receptor (EGFR) kinase structures (PDB codes 2ITY and 1XKK, respectively). Key secondary structural elements are colored (green glycine-rich loop; red C α -helix; blue activation loop). The ligands are shown in ball and stick notation (C-atoms in cyan for both gefitinib and lapatinib). Gatekeeper (T766) and DFG and HRD motifs are shown in space-filled balls. Conserved interactions and residue clusters differentiating the active (c) and inactive (d) EGFR conformational states are indicated. Regulatory spine and hydrophobic cluster are represented as transparent spheres (C-atom in green and yellow, respectively). The salt bridge of K721-E738 and the H-bond of R812-L834 are shown in red dashed lines. The figure was made using PYMOL (DeLano Scientific, San Carlos, CA)

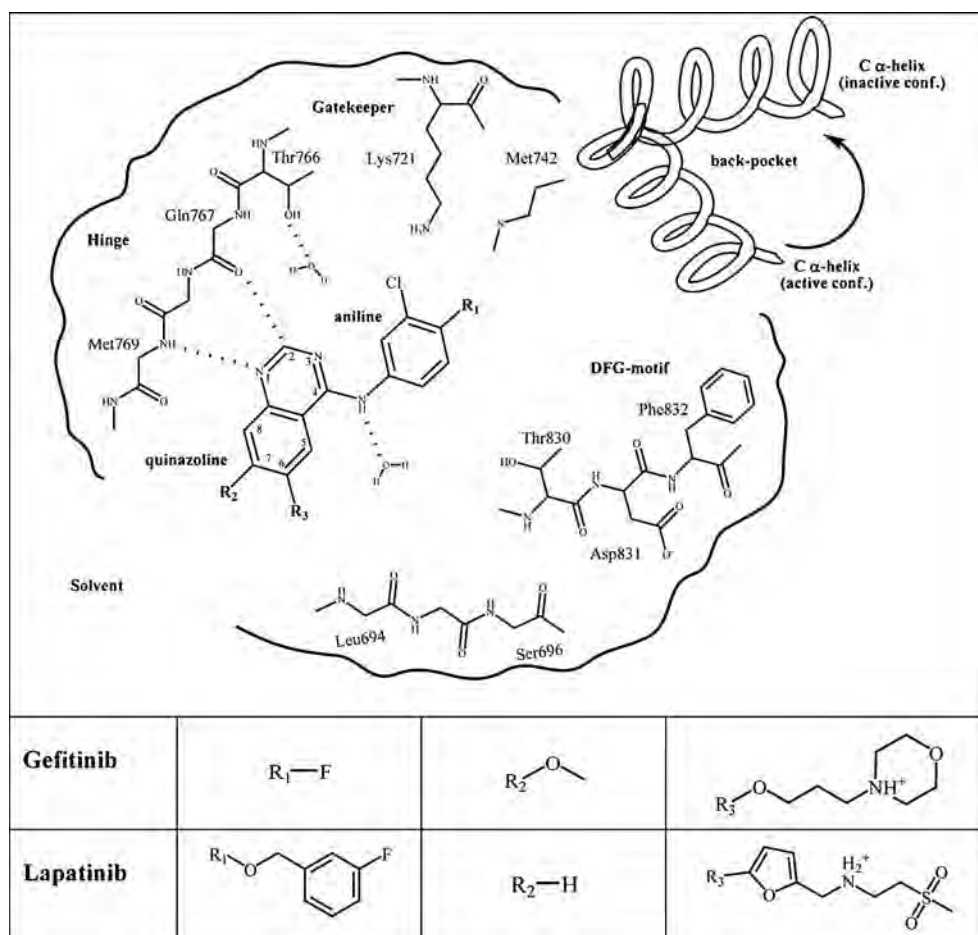


ATP-binding site of the active or inactive protein. As these inhibitors target the more generic, evolutionary conserved, ATP-binding pocket, undesirable activity towards other members of the approximately 500 strong protein kinase families is frequently observed [11]. Type-II inhibitors target both the ATP-binding and allosteric pockets formed within an inactivated protein. This includes the DFG-out conformation in c-Abl [12] and the allosteric binding pocket of MEK that lies adjacent to the ATP-binding site [13]. Type-I½ inhibitors can be considered a hybrid of the previous two, targeting the ATP-binding site of the DFG-in conformation of the inactive protein, as well as a rather large back-pocket as exemplified in the EGFR-lapatinib complex (Fig. 2) [4]. There is particular interest in Type-I½ and Type-II inhibitors from the point of view of selectivity as these regions will be under reduced evolutionary pressure to remain constant. These differences are therefore more likely to be exploited to produce a selective kinase inhibitor. However, problems pursuing this type of inhibitor also exist. Mutations at or around allosteric pockets are more likely to occur than at evolutionary conserved regions, potentially leading to problems associated with drug-resistance [14].

Molecular dynamics (MD) studies have in the past been used to elucidate dynamic aspects associated with protein

kinases, including EGFR [7, 15–18]. Such simulations offer additional insight beyond the static, but nonetheless critical, snapshot as represented by an X-ray crystal structure. MD has proved particularly insightful for EGFR from a drug resistance perspective, as a dynamic assessment of the effect of mutations, including L834R, G695S and L834R and T766M, on protein structure can be assessed [7, 15–17]. Liu et al. [15] studied the origin of resistance for the Type-I EGFR inhibitor gefitinib (Iressa®), noting the implications each mutation had on the ATP-binding pocket and on inhibitor binding. Balias et al. [16] studied the effect of EGFR mutations on Type-I inhibitors; erlotinib (Tarceva®), gefitinib and AEE788, using models of the wild type (WT) and three different active EGFR mutants. They were able to explain the majority of the fold resistance changes in the different mutants from the calculated binding free energy, as well as giving an explanation for their physical origin. Recently, Wan et al. [17] also studied the changes in drug-binding affinities due to the cancer-related mutations of EGFR using multiple short MD simulations, which provide significantly enhanced conformational sampling. Also worth mentioning is a study by Papakyriakou et al. [7], who investigated EGFR protein dynamics in the absence of inhibitors. The focus of their study was understanding the

Fig. 2 Illustration of the EGFR kinase binding site for potential kinase inhibitors (PDB codes 2ITY and 1XKK). Important amino acid residues located in the binding site and the chemical structure of ligands (gefitinib and lapatinib) are shown. The main interactions of the EGFR kinase hinge with the quinazoline moiety are indicated. The LIGPLOT diagram [40] for all hydrophobic and H-bond interactions from the PBD files are shown in Supporting Information Fig. S6



dynamic differences between the active and inactive protein conformations of EGFR, and the transition between them. Using 5 ns of targeted MD to drive the transition between the two different states, they concluded that the timescales needed for the formation the back-pocket in inactive EGFR protein are beyond the timescales of conventional MD.

An important challenge in structure-based design is to better understand the dynamic nature of protein kinases, and the underlying reasons for the different protein conformational preferences observed with different inhibitors. In this novel study, we perform MD simulations on both active and inactive protein complexes of wild type EGFR, with both type-I and type-II inhibitors. Our goal is to try to understand the origin of the distinct EGFR conformations, and the relative preference of these protein conformations for the Type-I inhibitor gefitinib and Type-I $\frac{1}{2}$ inhibitor lapatinib. An improved understanding of inhibitor binding to the inactive conformation is highly desirable given that inhibitors of this conformation, rather than the active form, appear to be more efficacious [9]. To this end, we employed MD simulations using the AMBER force field within GROMACS to simulate the active and inactive forms of the protein. We assessed the drug molecules gefitinib and lapatinib to try and decipher the relative contribution of the inhibitor to the stability of the two protein conformations. We also simulate the APO forms of both protein conformations, and the case where gefitinib bound to the inactive conformation. The value of this information is that the contribution of the various structural elements, or individual residues, to inhibitor binding and protein stability can be better understood, potentially allowing the more focused direction of chemistry resources to target the area most likely to give rise to higher affinity, tighter binding inhibitors.

Methods and materials

Protein preparation

The EGFR protein coordinates for the active and inactive conformations were obtained from the Protein Data Bank (<http://www.pdb.org>). The active and inactive coordinates used in this study correspond to the PDB structures with accession codes 2ITY [5] and 1XKK [4], respectively. EGFR-ligand protein structure models were created by removing all ions, and all water molecules except those found within the binding site (three molecules in 2ITY and eleven in 1XKK). We retained the water molecules found in the active sites of each protein for all the simulations used here as they have been shown to be critical in protein–ligand simulations in the past [19, 20]. Water molecules have been shown to be important for rationalizing dynamic phenomena from MD simulations [21] as well as docking and scoring results [22, 23].

The amino acid sequences used for our simulations begin with A678 and finish at G959, based on the 2ITY numbering (or A702 to G983 using the alternative EGFR numbering system). Missing loops in both PDB structures (E842 to K851 in 2ITY and E710 to K713, A726 to S728, E844 to K851 in 1XKK) were built using MODELER 9v4 [24]. The stereochemical quality of the resultant models was assessed using PROCHECK v3.5.4 [25]. The N- and C-terminal ends of both models were capped with acetyl (ACE) and methyl amino (NME) groups, respectively. The 2ITY and 1XKK based models are henceforth referred to as gefitinib-EGFR_(active) and lapatinib-EGFR_(inactive), respectively.

A model of the inactive EGFR protein with gefitinib bound was also generated for the purpose of comparison. This model was created by replacing lapatinib with gefitinib in the inactive 1XKK based EGFR model. Superposition of the ligands was performed based on the 2ITY-1XKK C α -atom alignment. Due to differences in the G-loop conformations between active and inactive proteins, the orientation of the 7-methoxy-6-(3-morpholin-4-ylpropoxy) moiety of gefitinib was altered subtly to avoid steric clashes, adopting a conformation resembling that of the 5-{[2-(methylsulfonyl)ethyl] amino} methyl group of lapatinib. This structure is termed gefitinib-EGFR_(inactive). No model of lapatinib bound to the active EGFR conformation was generated as the absence of the back-pocket II prevents the inhibitor from binding. APO structures of the active and inactive EGFR protein were created by removing the ligands from the gefitinib-EGFR_(active) and lapatinib-EGFR_(inactive) models. These models are termed the APO-EGFR_(active) and APO-EGFR_(inactive) models, respectively.

The AMBER-99SB force field was used to simulate all protein structures and the ionization state of amino acid residues was set according to the standard protocol [26]. All models were solvated in a triclinic box of TIP3P water, keeping a distance of 10 Å between the protein and the sides of the solvent box. Chloride ions were added to neutralize the charge of the system.

Ligand preparation

Gefitinib and lapatinib were simulated in their protonated state in line with their predicted basic pKa according to Marvin-View v5.3.1 (ChemAxon, Budapest, Hungary) at pH 7.0 (Fig. 2). This protonation state is further favored due to the close proximity of a number of H-bond acceptors (either the negatively charge side chain of D776 or the carbonyl oxygen atom of L694) at the entrance to the ATP-binding pocket. GROMACS topology files were generated using ACPYPE script [27]. GAFF force field parameters [28] were used for both inhibitors. Partial charges were calculated using the AM1-BCC method [29] as implemented in QUACPAC 1.3.1 (OpenEye Scientific Software, Santa Fe, NM).

Simulation conditions

Simulations were carried out using GROMACS v4.0.2 [30, 31] with the AMBER force field ports [32, 33]. All simulations used isobaric-isothermal (NPT) conditions at standard temperature (300 K) and pressure (1 bar), using the Berendsen coupling method [34]. The linear constraint (LINCS) algorithm was applied to fix all hydrogen related bond lengths, facilitating the use of a 2-fs time step [35]. A short-range nonbonded interaction cut-off distance of 10 Å was used. The particle mesh Ewald (PME) method with a 0.12 nm Fourier grid spacing was used to account for long-range electrostatics [36, 37].

A three-step procedure was used for MD simulations. First, each of the EGFR models was energy-minimized using the steepest descent method (until the maximum force was less than 100 kJ mol⁻¹ nm⁻¹ on any atom) to reduce undesirable van der Waals contacts, and to optimize H-bond interactions present. In the second step, each model was subjected to 500 ps of a position-restrained MD in which heavy atom positions of each protein were restrained harmonically using a force constant of 1000 kJ mol⁻¹ nm⁻². Water molecules, counterions and inhibitors, if present, were not restrained. The systems were then heated from 0 K to 300 K over the first 50 ps, followed by 450 ps of equilibration. The third step involved unrestrained MD for a period of 20 ns. Coordinates were archived every 1 ps. The simulations for the APO-EGFR_(inactive) and gefitinib-EGFR_(inactive) models were subsequently extended to 50 ns to assess the conformational characteristics of their C α -helices.

Analyses

All MD analyses were performed using tools available within the GROMACS suite. The tool “*g_rms*” was used to evaluate the root mean square deviation (RMSD) of heavy atoms in MD trajectories from those of original structures obtained before energy minimizations. The tool “*g_rmsf*” was used to compute the root mean square fluctuation (RMSF) of heavy-atom positions with respect to their time-averaged position and was used to calculate a theoretically derived B-factor (temperature factor).

The statistical significance of any reported differences in either the means or standard deviations (SD) in the RMSD or RMSF have been confirmed using an unpaired Student's *T*-test or F-test, respectively. All reported differences are significance above the commonly used 95 % confidence level unless otherwise stated. All statistics were computed in Microsoft Excel 2007.

Results and discussion

MD simulations were performed on five separate EGFR models that differ in terms of the bound inhibitor [gefitinib,

lapatinib and no inhibitor (i.e. APO)], or the protein conformation (active or inactive). Analysis of the RMSD of the protein heavy-atoms (i.e., compared to the initial X-ray structures obtained before energy minimizations) showed that all simulations had reached equilibrium well before *t*=10 ns. The protein structures remained stable throughout the simulation; with the overall heavy-atom RMSD remaining within 3.0 Å of the original X-ray coordinates (Fig. 3a–b). In addition, the total energy of each model remained essentially constant over the course of the simulation, giving further confirmation of its stability (Supporting Information Fig. S1). We report all structural parameter analyses between *t*=10 to 20 ns unless otherwise stated.

The reliability of such simulations can be assessed qualitatively by comparing the experimental C α -atom B-factor values to those computed from the MD simulation (Fig. 3c, d). As shown in Fig. 3d, the MD-predicted B-factors of the gefitinib-EGFR_(active) and APO-EGFR_(active) models are in good qualitative agreement with the corresponding experimental data. Deviation to some degree is expected since the X-ray data is obtained in a dynamically restricted crystalline phase. This suggests the MD results are physically representative of the protein in general.

Dynamic characteristics of the active and inactive protein-inhibitor complexes

In the following sections, we consider the dynamic characteristics of the gefitinib-EGFR_(active) and lapatinib-EGFR_(inactive) models, and whether the structural differences observed between the models have arisen due to (1) the effect of the ligand, (2) the protein, or (3) a combination of both. To ascertain their origin, we contrast the results to simulation data obtained using models of gefitinib bound to the inactive EGFR conformation and models of both active and inactive APO protein conformations.

From Fig. 3a,b and Table 1, it can be seen that the average RMSD and SD of the gefitinib-EGFR_(inactive) model is greater than that of the lapatinib-EGFR_(inactive) model over the course of the simulation. The differences were found to be statistically highly significant (*P*<0.0001). From the APO simulations, we can see that the inactive protein conformation (APO-EGFR_(inactive)) behaves similarly to the lapatinib-EGFR_(inactive) model in that the average RMSD is roughly comparable (the SD of the RMSD is only moderately lower by 0.01, Table 1). It is somewhat different for the active EGFR model in that the RMSD increases by 0.39 Å in the APO structure compared to the ligand bound structure. This suggests that the ligand plays a more important role in stabilizing this conformation (Table 1).

To investigate these differences further, we computed the average energy of the proteins over the course of the simulation as the sum of the bonded (internal) and

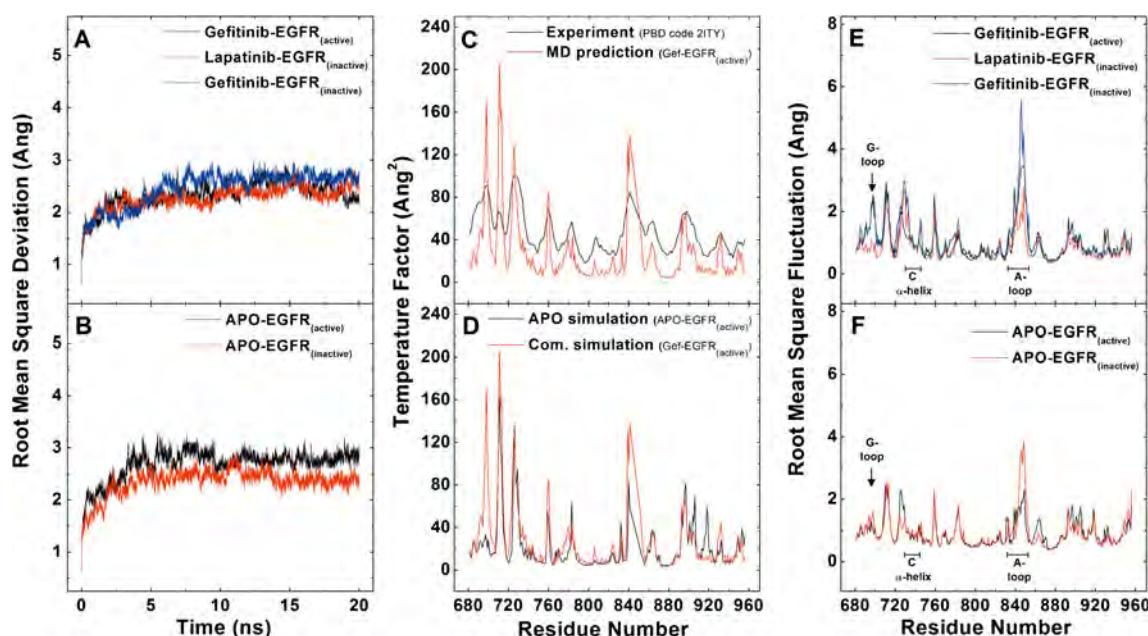


Fig. 3 Heavy-atom protein RMSD plots of the complex (a) and apo (b) simulations. Comparisons of experimental versus predicted B-factor values (c) and predicted B-factors of complex versus apo

simulations (d). Heavy-atom RMSF/residue plots of the complex (e) and apo (f) EGFR kinase structures. Key secondary structural elements (glycine-rich loop, C α -helix and activation loop are indicated

nonbonded (electrostatic and van der Waals) energies. This energy neglects energetic components of the solvent and the inhibitor (apart from the interactions with the protein, which are considered) as a means of comparing, albeit approximately, their energy profile over the course of the simulations. We observed the following statistically significant trends in the computed energy: $\text{APO-EGFR}_{(\text{active})} < \text{lapatinib-EGFR}_{(\text{inactive})} < \text{APO-EGFR}_{(\text{inactive})} < \text{gefitinib-EGFR}_{(\text{inactive})} < \text{gefitinib-EGFR}_{(\text{active})}$. It is possible to compare these energies qualitatively since all models have an identical number of atoms. The most energetically favorable protein conformation over the course of the simulation was the APO protein in the active conformation. Surprisingly, the binding of gefitinib to the active protein conformation resulted in the protein with the highest overall energy, more so even than when bound to the inactive EGFR conformation. The APO

form of the inactive protein is found to be less energetically favourable than the active form but binding of lapatinib stabilizes this conformation. Gefitinib was found to destabilize the inactive protein conformation. One might therefore conclude that the active protein is destabilized to a degree by inhibitor binding, whereas the inactive protein conformation is stabilized, at least by inhibitors such as lapatinib, which possess a back-pocket binding group.

These observations do not represent the complete picture since many of the individual structural elements found within the five different protein models will be dynamically dissimilar over the course of the simulation. To assess these differences, the heavy-atom RMSF value of each residue (that which is related to the crystallographic B-factor or thermal motion) was calculated to understand how the different structural elements behave between $t=10$ to 20 ns

Table 1 The heavy-atom root mean square deviation (RMSD) average and standard deviation (SD) values (Å) (in parenthesis) for the overall protein structure, key secondary structural elements conserved

	gefitinib-EGFR _(active)	lapatinib-EGFR _(inactive)	gefitinib-EGFR _(inactive)	APO-EGFR _(active)	APO-EGFR _(inactive)
Overall structure	2.39 (0.17)	2.34 (0.15)	2.55 (0.24)	2.78 (0.15)	2.42 (0.14)
Glycine-rich loop	1.73 (0.21)	0.66 (0.11)	0.81 (0.15)	1.68 (0.14)	1.52 (0.39)
C α -helix	1.84 (0.28)	2.99 (0.61)	2.72 (0.54)	1.80 (0.19)	2.55 (0.30)
Activation loop	3.76 (0.72)	3.14 (0.23)	4.01 (0.74)	3.92 (0.20)	2.82 (0.30)
Hydrophobic cluster	1.72 (0.19)	1.66 (0.15)	2.15 (0.26)	1.84 (0.13)	1.77 (0.27)
Regulatory spine	1.58 (0.27)	0.97 (0.14)	1.43 (0.20)	1.24 (0.19)	1.07 (0.26)
Ligand	1.26 (0.28)	1.08 (0.11)	1.14 (0.26)	—	—

interactions and ligands bound over the course of the simulation (values calculated from $t=10$ to 20 ns)

(Fig. 3e–f). We have also assessed the heavy-atom RMSD (compared to the initial structures obtained before energy minimizations) of three important conserved secondary structural elements, C α -helix (P729 to A743), glycine-rich loop (G-loop; G695 to T701) and activation loop (A-loop; D831 to V852) and the bound ligands. In addition, we have assessed the differences in the EGFR specific regions, termed the regulatory spine (R-spine) and hydrophobic cluster (H-cluster), which are believed to be important in differentiating the active and inactive protein conformations, respectively. The data are summarized in Table 1.

Ligand binding

The average RMSD of the inhibitors in the gefitinib-EGFR_(active) and lapatinib-EGFR_(inactive) complexes were comparable over the course of the simulation, yet gefitinib was found to fluctuate to a much greater extent than lapatinib since the RMSD SD of the ligand in the former is 0.17 Å greater than in the latter (Table 1, Fig. 4). The differences are due primarily to the solvent-exposed tails of the two inhibitors, which is evident given that the RMSD of the quinazoline and anilino-based substituents are rather small (Supporting

Information Fig. S4). The results from the gefitinib-EGFR_(active) simulation are consistent with a recent study in that the solvent-exposed tail of the inhibitors exhibits greater movement than the central scaffold [17]. This is expected given the importance of the hinge interaction between the quinazoline acceptor and the M769 donor (Fig. 2 and Supporting Information Fig. S6). The differences in flexibility at the solvent-exposed region could be a result of differences in the intrinsic flexibility of adjacent protein structural elements, or a reflection of the differing binding strengths/characteristics of the substituents in question. Analysis of the interactions present in Fig. 5 and Supporting Information Table S1 indicates that the basic nitrogen of gefitinib interacts strongly with the carbonyl oxygen of L694 located on the G-loop, considerably more so compared to lapatinib. In contrast, lapatinib forms stronger interactions with the C-lobe via residues such as C773 and its furan ring, D776 and its basic nitrogen, and R817 and its sulfone group. The three strong interactions formed by lapatinib contrast with just one strong interaction of gefitinib, and help to explain the latter's larger RMSD. Analysis of the results from the gefitinib-EGFR_(inactive) simulation, where gefitinib was placed in the active site of the inactive protein conformation, reveals that the ligand also

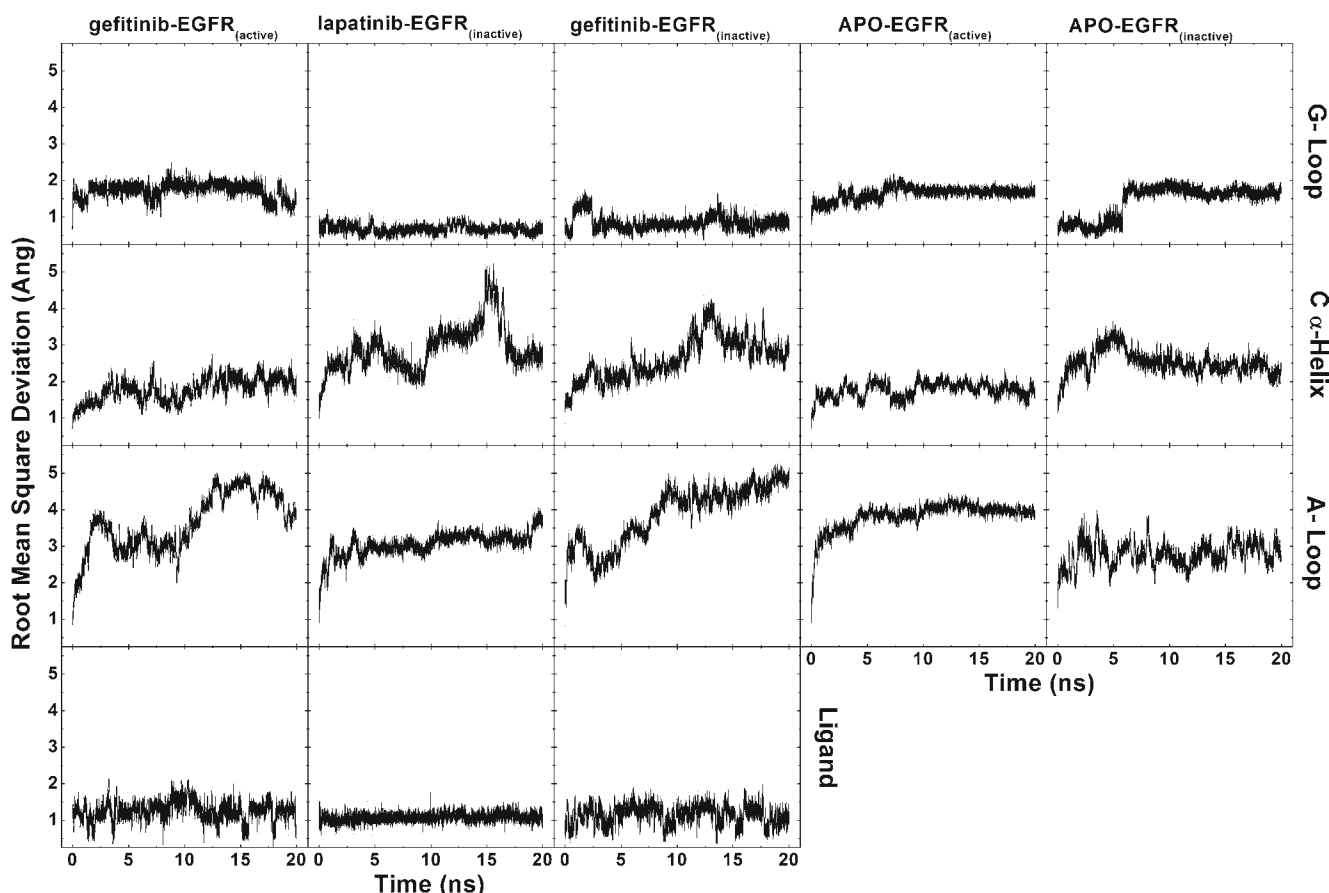
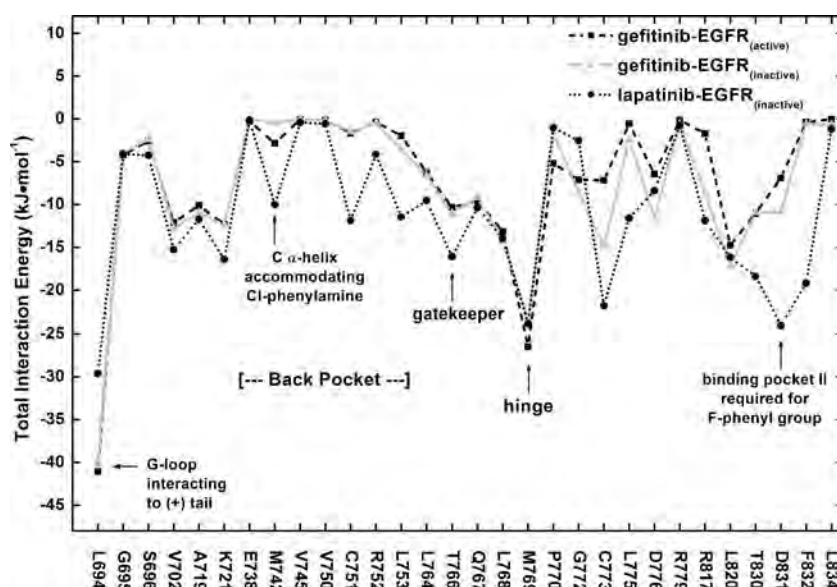


Fig. 4 Heavy-atom RMSD plots of the residues in the key secondary structural elements; glycine-rich loop, C α -helix, activation loop and ligands for all molecular dynamics (MD) simulations

Fig. 5 Plots for the interaction energy of selected residues located in the binding site for the three complex simulations. The important residues are indicated



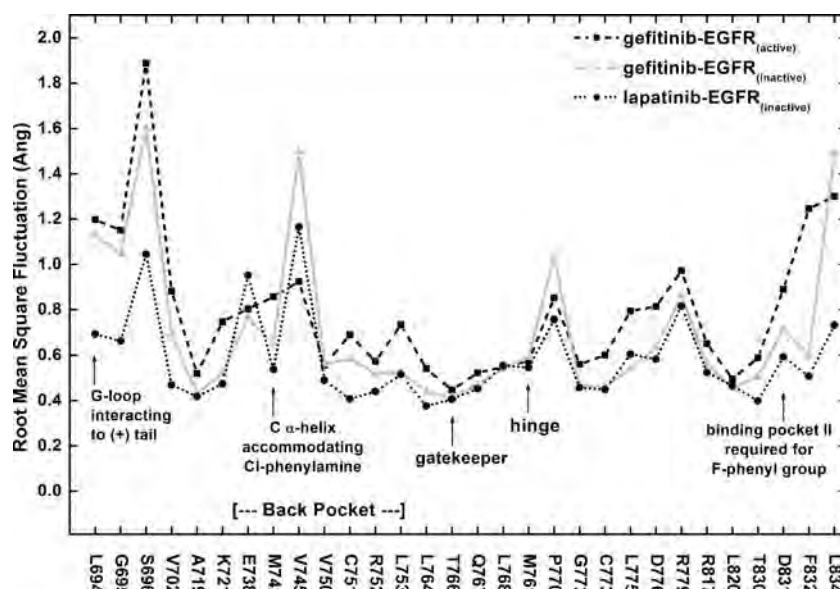
fluctuates to a larger degree also in this cavity (Table 1, Fig. 4). This suggests that the behavior of gefitinib is a characteristic of the molecule itself and not an effect of the protein conformation.

Analysis of the interaction energies (the sum of electrostatic and van der Waals interactions) between the inhibitors and adjacent active site residues was subsequently considered (Fig. 5 and Supporting Information Table S1). A key difference between the two inhibitors is the presence of a large back-pocket binding group of lapatinib, which can make a number of strong interactions not made by gefitinib. These include significant interactions by the methoxy-3-F-phenyl substituent with T830, L753, M742, D831, and F832 (see also Supporting Information Fig. S6). Common interactions made by both inhibitors are those between the 3-Cl-phenylamine portion and T830 and K721, although these

are noticeably stronger in the case of lapatinib. Strong interactions are made with the hinge M769 residues in all cases as can be seen in Fig. 5 and Supporting Information Table S1.

It is also possible to look at the atomic fluctuation of individual active site residues as this can shed light on the nature of the binding site interactions (Fig. 6). An analysis of the RMSF of these residues shows noticeable differences between the different EGFR complexes. For example, L694 and S696 of the glycine-rich loop fluctuate considerably more in the gefitinib-based complexes. Gefitinib forms a favorable interaction with the G-loop residue L696 and, over the course of the simulation, this interaction is maintained even as the protein undergoes significant fluctuation. Residues E738 to V745 are located in the back-pocket and fluctuate along with the inhibitor back-pocket substituents

Fig. 6 Plots for the RMSF for the selected residues located in the binding site for the three complex simulations. The important residues are indicated



(or lack of the extended back-pocket binding group of gefitinib). The hinge binding region, along with the gatekeeper (T766 to M769), show rather low RMSF values as might be expected given the importance of the interaction made by the inhibitors and M769 (although for gefitinib the values are generally slightly larger). Residues D776–L834 in Fig. 6 correspond to the floor of the ATP-binding site, as well as part of the back-pocket. It appears that the RMSF values of these are also slightly larger for gefitinib-EGFR_(active) than lapatinib-EGFR_(inactive), which would suggest that lapatinib forms the more tightly bound complex of the two inhibitors.

We assessed the binding free energies of the inhibitors (ΔG_{bind}) using the linear interaction energy (LIE) method [38, 39]. We found that the gefitinib ΔG_{bind} to the active and inactive proteins were comparable, but that these were $\approx 7 \text{ kJ mol}^{-1}$ higher in energy than that observed for lapatinib binding to the inactive protein. Although these energies are not considered as precise as alternate methods such as FEP or even MM-PBSA, the results obtained are consistent with the reports from Woods et al. [4] who found that lapatinib has a much slower off rate than other EGFR inhibitors such as gefitinib. That said, the results are not in agreement with the experimental IC_{50} values. IC_{50} values are generally determined to the active protein conformation in biochemical assays, and not the inactive form (requiring longer equilibration times), which might help explain the discrepancy.

C α -helix

The C α -helix of EGFR kinase is the principle region that differs between the active and inactive protein conformations. The formation of the inactive EGFR conformation requires translation of the C α -helix in the z-direction relative to the rest of the protein (Fig. 1). This movement then leads to the formation of an additional hydrophobic pocket (back-pocket II), which is occupied by the 3- F-phenyl group of lapatinib.

The C α -helix in the lapatinib-EGFR_(inactive) structure was found to have a larger average RMSD and RMSF over the course of the simulation than those observed in the gefitinib-EGFR_(active) simulation (Fig. 3e, Fig. 4 and Table 1). This appears to be due, in part, to helix-coil transitions at the N-terminus of the helix (P729–I735) in simulations of the EGFR_(inactive) structures, even though those residues tended to be helical at the end of simulation. The higher degree of flexibility and conformational transitions are in fact consistent with the ambiguous electron density found for residues A726–P729 in the 1XKK structures [4] and are consistent with the simulation results reported by others [7]. This observation was assumed initially to be an artifact of the modeled loop, consisting of

three amino acid residues at the top of the C α -helix in 1XKK. However, a similar effect is observed in 2ITY, where these residues have been resolved experimentally. Additionally, upon simulation of the APO protein derived from 2ITY (APO-EGFR_(active)), we found the helix remained essentially intact over the course of the simulation, suggesting the transitioning was, at least in part, ligand induced. We also observed a dramatic drop in the flexibility of the C α -helix going from the inactive ligand-bound structure (lapatinib-EGFR_(inactive)) to the APO structure (APO-EGFR_(inactive)). The RMSD SD dropped from (0.61 to 0.30 Å) indicating that the ligand plays an important role in inducing this instability. Additional evidence for this is the drop in flexibility, albeit smaller, going from lapatinib-EGFR_(inactive) to gefitinib-EGFR_(inactive). This is because the latter lacks the back-pocket II binding group created by the C α -helix movement. As noted by Papakyriakou et al. [7], it is also likely that differences between crystal stacking forces and those of the simulated water will be responsible for some of these differences.

Of additional interest to us was whether the active and inactive structures had converged to any degree, particularly when the simulations were extended to 50 ns for the APO-EGFR_(inactive) and gefitinib-EGFR_(inactive) models. The N-terminus of the C α -helix (P729 to I735) of the EGFR_(inactive) structures is unstable and shows the helix to coil transitioning. In the APO-EGFR_(inactive) simulation, we also found helical bending leading to the N-terminal segment of the helix drifting toward that of the active conformation (Supporting Information Fig. S5C and S5E). In contrast, the C-terminal segment of the helix does not change its position, consistent with the long distance between K721 and E738 (Supporting Information Fig. S2). This helical bending was seldom observed in the gefitinib-EGFR_(inactive) and lapatinib-EGFR_(inactive) complexes (Supporting Information Fig. S5B, S5D and S5F).

Comparing the C α -helices from structures obtained from the five different simulation structures, to that of the original active X-ray structure (2ITY), reveals some interesting trends (Table 2). The APO-EGFR_(active) structure displays the RMSD of the helix of ≈ 3.7 Å, compared to 3.3 Å for the gefitinib-EGFR_(active). RMSD analysis of the EGFR_(inactive) simulations using the initial active conformation as the reference structure showed that the C α -helix of the EGFR_(inactive) complexes have the RMSD of ≈ 7 Å to the active conformation and increase over time to 8.6 Å for gefitinib-EGFR_(inactive) after 50 ns of simulation; however, in simulation of the EGFR_(inactive) protein in the APO form, the RMSD decreases from 7.03 Å to 5.44 Å after 20 ns and to 5.01 after 50 ns (Table 2 and Supporting Information Fig. S7). This suggests that the C α -helix of both the active and inactive APO structure start to converge towards a similar minimum (RMSD decreasing from 7.0 to 3.3 Å at $t=20$ ns). However,

Table 2 Heavy-atom RMSD (Å) of the C α -helices after overall superimposition of the EGFR_(active) and EGFR_(inactive) structures at t=20 ns and t=50 ns when compared to the EGFR_(active) structure at t=0 ns

	APO-EGFR _(active)	APO-EGFR _(inactive)	gefitinib-EGFR _(active)	lapatinib-EGFR _(inactive)	gefitinib-EGFR _(inactive)
t=20 ns	3.69	5.44	3.31	7.33	7.15
t=50 ns	—	5.01	—	—	8.64

while the C α -helix of the inactive protein did appear to move towards a more active-like conformation, we did not observe the formation of the salt bridge between K721 and E738, the most critical element associated with the EGFR activation [3].

Analysis of the regulatory (R-) spine (residues, M742, L753, H811 and F832) [6] shows that the conformation adopted by these residues in the gefitinib-EGFR_(inactive) and APO-EGFR_(inactive) structures are not close to those found in the active conformation. (Table 1 and Supporting Information Fig. S3). The H-bond between HRD R812 and DFG+1 L834, which is an important feature in the active conformation, was measured over the course of simulation (Supporting Information Fig. S2). Although at the end of the simulations, the R812–L834 distance is shorter in the gefitinib-EGFR_(inactive) structure, when compared to the lapatinib-EGFR_(inactive) and APO-EGFR_(inactive) structures, this key H-bond is still unlikely to form during the gefitinib-EGFR_(inactive) simulation since it is still found to be >4.0 Å.

Glycine-rich loop

The G-loop consists of a set of flexible residues that are located on the N-lobe of protein kinases. These residues are important in defining the size and shape of the ATP-binding pocket, as well as its dynamic characteristics, and should have major implications for inhibitor binding. Indeed, the opening of the active site pocket to solvent in the inactive protein containing lapatinib is slightly smaller than in the active protein with gefitinib due to the conformation adopted by the G-loop.

Over the course of the 20 ns simulation, it can be seen that the G-loop of gefitinib-EGFR_(active) model deviates further than that of lapatinib-EGFR_(inactive), and fluctuates to a greater extent (Fig. 4, Table 1). Analysis of the atomic coordinates reveals that this movement is primarily in the y-dimension as defined in Fig. 1, corresponding to the expansion and contraction of the entrance into the ATP-binding site.

The dynamics characteristics of the G-loop in gefitinib-EGFR_(active) and lapatinib-EGFR_(inactive) appear to correlate with that of the bound inhibitors (Table 1). For example, the average RMSD and SD of the inhibitor and G-loop in gefitinib-EGFR_(active) simulation are almost double those obtained from the lapatinib-EGFR_(inactive) simulation. This

is expected given that gefitinib interacts more strongly with the G-loop than lapatinib, as discussed previously. From the results of the separate gefitinib-EGFR_(inactive) simulation, we observed that the G-loop and inhibitor have an average RMSD between the gefitinib-EGFR_(active) and lapatinib-EGFR_(inactive) values. However, while the RMSD SD of the G-loop is also intermediate in value, that of gefitinib is the same as the original gefitinib-EGFR_(active) simulation. From a consideration of the APO simulation results (Fig. 4), it appears that the presence of an inhibitor significantly stabilizes the G-loop of the inactive protein. Removal of the inhibitor leads to a dramatic increase in the RMSD and RMSF compared to the gefitinib-EGFR_(inactive) and lapatinib-EGFR_(inactive) structures (Table 1, Fig. 3e–f and Fig. 4). In contrast, the RMSD and RMSF values of the APO-EGFR_(active) structure do not deviate dramatically from that of the gefitinib-EGFR_(active) structure, suggesting the G-loop conformation in the active protein is intrinsically more stable.

Activation loop

The A-loop is an important structural element found in protein kinases. It contains amino acid residues that are critical to achieve their catalytic function of phosphorylation. The key portion of the A-loop is the so called DFG motif, which is found in the “in” conformation in known EGFR structures. A key interaction between residues in the C and N-lobes are mediated through the DFG D831 residue of the A-loop, and K721 of the β 3-strand. There also exists a short α -helical segment towards the N-terminus of the A-loop (L834 to L837) in the inactive EGFR structure, but is not present in the active structure. This may also affect the conformational characteristics in this region.

The A-loop showed the largest RMSD and RMSF values among the three key structural elements over the course of the simulation for both EGFR conformations (Table 1, Figs. 3e, f; 4). The largest movement of the A-loop is in agreement with the work of others [17] and this is in line with the experimental X-ray data in that the residues E844 to K851 are disordered in both 2ITY and 1XKK structures [4, 5].

The A-loop found in the lapatinib-EGFR_(inactive) structure was found to fluctuate considerably less than that found in gefitinib-EGFR_(active) (RMSD s.d. of 0.23 Å vs 0.72 Å, respectively) (Table 1). It appears that the short α -helical

segment towards the N-terminus of the A-loop in the inactive protein limits the degree of flexibility. In addition, the interaction between D831 and K721 over the course of the MD simulations was somewhat weaker in gefitinib-EGFR_(active) than lapatinib-EGFR_(inactive) simulations, consistent with the original X-ray structure (Supporting Information Fig. S2).

The simulation of gefitinib bound to the inactive EGFR protein structure (gefitinib-EGFR_(inactive)) helps to shed light on the characteristics of the A-loop. The average RMSD and RMSF from this simulation are roughly comparable to that of the gefitinib-EGFR_(active) simulation, suggesting that the effect of gefitinib on the A-loop is similar to that observed for the G-loop (Table 1, Figs. 3e,f; 4).

From the simulations of the APO-EGFR_(active) and APO-EGFR_(inactive) structures, it appears that the A-loop in the active protein deviates to a greater extent (RMSD average of 3.9 Å vs 2.8 Å, respectively), but fluctuates to a lesser degree (RMSD SD of 0.20 Å vs 0.30 Å, respectively). However, the binding of lapatinib to the inactive protein leads to a drop in the RMSD and RMSF suggesting that it help stabilize the A-loop conformation. In contrast, gefitinib binding to the active protein conformation leads to dramatically increased RMSD SD values in particular (0.20 Å vs 0.72 Å, respectively) suggesting it has the opposite effect. The binding of gefitinib to the inactive protein conformation also results in the larger RMSF and RMSD SD, suggesting that the destabilizing effect is due to the fact it makes no interactions with residues on the C-lobe (in contrast to the three interactions made by lapatinib) (see also Table 1, Figs. 3e,f; 4; 5).

Hydrophobic cluster

A network of several residues (L723, M742, L764 and D831), including two on the activation loop (L834 and L837), form a small hydrophobic (H-) cluster, and are reported to be important for the stabilization of the inactive EGFR conformation [7, 8]. Indeed, the cancer-related mutation L834R, has been known to be involved in either the disruption of the hydrophobic packing of the inactive EGFR kinase structure [8] or the introduction of an intermediate state in the active-inactive transformation pathway, adjusting the relative stability of both states [18], subsequently inducing EGFR activation.

From our simulation data, we find that the H-cluster does not vary dramatically over the course of the simulation. Although the H-cluster does not exist in the active conformation, we show the RMSD values for the purpose of comparison (Table 1 and Supporting Information Fig. S3). The average RMSD of the H-cluster in both APO proteins was roughly comparable; however, the RMSD SD in the APO-EGFR_(active) protein is considerably lower than that in the APO-EGFR_(inactive) (0.13 vs 0.27 Å), suggesting that it

does play some form of stabilizing role in the active conformation. The observation that the lapatinib-EGFR_(inactive) simulation displays a very low RMSD (both average and SD values) is not surprising given that these residues can reorientate, with interactions with lapatinib in the back-pocket region being made. The larger RMSD of the H-cluster in the gefitinib-EGFR_(inactive) structure is consistent with the partial unfolding of the short α helix located close to L834 and L837. This helical transition was not observed in the APO-EGFR_(inactive) and lapatinib-EGFR_(inactive) structures.

Conclusions and future directions

In this study, we performed MD simulations on both the active and inactive protein complexes of the wild type EGFR, with both type-I and type-II inhibitors. Our goal was to better understand the origin for the two distinct protein conformations and their relative preference for the Type-I inhibitor gefitinib and Type-II inhibitor lapatinib. We also simulated both APO forms and gefitinib bound to the inactive conformation to decipher the relative contribution of the inhibitor to stability of the two EGFR conformations.

We find that binding of gefitinib to the active protein appears somewhat destabilizing when compared to the APO simulation of the same conformation. The major cause of destabilization is increased fluctuation of the G-loop, A-loop and, to a lesser extent, the C α -helix. In contrast, the binding of lapatinib to the inactive conformation helps to lower the energy of the protein. Lapatinib binding leads to lower fluctuation in the G-loop and A-loop but a dramatic increase for the C α -helix.

Calculation of binding free energies suggests that lapatinib also binds more strongly to the inactive protein than gefitinib does to the active protein. While this would appear to contradict their reported experimental pIC₅₀ values, it is in agreement with the experimentally determined slower off rate displayed by lapatinib [4], which is one of the proposed reasons for its better efficacy (along with its ErbB2 inhibition). The more favorable binding energy can be explained by the more extensive π type interactions it can make in the EGFR back-pocket by virtue of its additional 3-Cl-phenyl substituent. In addition, lapatinib makes three moderately strong interactions with the C-lobe (via its sulfone, furan and basic nitrogen) in contrast to gefitinib, which makes a single strong interaction with the G-loop (via its basic nitrogen). In addition, an analysis of the RMSF shows that the G-loop in the gefitinib-EGFR_(active) structure fluctuates dramatically, but in the process also maintains the interaction already present. The lack of any interaction between gefitinib and the C-lobe also explains the large RMSF values of the A-loop. In contrast, the three relatively strong interactions that lapatinib makes with the C-lobe appears to restrain

the A-loop, while the weaker interaction with the G-loop leads to a higher RMSF in this region.

The C α -helix of the inactive EGFR conformation is intrinsically more mobile than that of the active conformation. This is unsurprising given its more extended position based on the known X-ray structure. It is found that the presence of an inhibitor in either protein conformation increases the flexibility in this region compared to the equivalent APO structure as a result of interactions mediated with residues in the back-pocket. Interestingly, from an analysis of the APO simulations, it appears that the C α -helix conformation in the active and inactive proteins begins to converge to a similar minimum after 50 ns of MD. However, we did not observe the formation of a salt bridge between K721 and E738—a critical element associated with EGFR activation [3].

The principle value of inhibitor binding information from MD is that the contribution of the various structural regions or individual residues to inhibitor binding and protein stability can be better understood. This potentially allows the more focused direction of chemistry resources to target regions in the active site most likely to give rise to higher affinity, tighter binding inhibitors. For example, rather than searching for any H-bond interactions to increase inhibitor affinity, an analysis of inhibitor binding, and the resultant change this has on receptor flexibility, may help to determine which H-bonds will contribute best to inhibitor binding, since the formation of such interactions may induce either greater stability or instability in the protein.

Acknowledgments This work was supported by the Higher Education Research Promotion and National Research University Project of Thailand, Office of the Higher Education Commission and the Thailand Research Fund grants RMU5180032 (KC) and RSA5480016 (MPG). We wish to express our gratitude for the use of Laboratory for Computational and Applied Chemistry (LCAC) research facilities at Kasetsart University provided by the National Center of Excellence in Petroleum, Petrochemical Technology and Advanced Materials. M.P.G. is grateful for the support of the Faculty of Science at KU and Associate Professor Supa Hannongbua in particular. N.S. was supported by a grant under the program Strategic Scholarships for Frontier Research Network for the Joint PhD Program Thai Doctoral degree from the Office of the Higher Education Commission, Ministry of Education, Thailand.

References

- Zuccotto F, Ardini E, Casale E, Angiolini M (2009) Through the “Gatekeeper Door”: exploiting the active kinase conformation. *J Med Chem* 53:2681–2694
- Weinmann H, Metternich R (2005) Editorial: drug discovery process for kinase inhibitors. *Chem Bio Chem* 6:455–459
- Stamos J, Sliwkowski MX, Eigenbrot C (2002) Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J Biol Chem* 277:46265–46272
- Wood ER, Truesdale AT, McDonald OB, Yuan D, Hassell A, Dickerson SH, Ellis B, Pennisi C, Horne E, Lackey K, Alligood KJ, Rusnak DW, Gilmer TM, Shewchuk L (2004) A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib). *Cancer Res* 64:6652–6659
- Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, Eck MJ (2007) Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* 11:217–227
- Kornev AP, Haste NM, Taylor SS, Ten Eyck LF (2006) Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci USA* 103:17783–17788
- Papakyriakou A, Vourloumis D, Tzortzatou-Stathopoulou F, Karpusas M (2009) Conformational dynamics of the EGFR kinase domain reveals structural features involved in activation. *Proteins* 76:375–386
- Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J (2006) An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* 125:1137–1149
- Liu Y, Gray NS (2006) Rational design of inhibitors that bind to inactive kinase conformations. *Nat Chem Biol* 2:358–364
- Liao JLL (2007) Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors. *J Med Chem* 50:409–424
- Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298:1912–1934
- Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, Kuriyan J (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science* 289:1938–1942
- Tong L, Pav S, White DM, Rogers S, Crane KM, Cywin CL, Brown ML, Pargellis CA (1997) A highly specific inhibitor of human p38 MAP kinase binds in the ATP pocket. *Nat Struct Biol* 4:311–316
- Bikker JA, Brooijmans N, Wissner A, Mansour TS (2009) Kinase domain mutations in cancer: implications for small molecule drug design strategies. *J Med Chem* 52:1493–1509
- Liu B, Bernard B, Wu JH (2006) Impact of EGFR point mutations on the sensitivity to gefitinib: insights from comparative structural analyses and molecular dynamics simulations. *Proteins* 65:331–346
- Balius TE, Rizzo RC (2009) Quantitative prediction of fold resistance for inhibitors of EGFR. *Biochemistry* 48:8435–8448
- Wan S, Coveney PV (2011) Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs. *J R Soc Interface* 8:1114–1127
- Wan S, Coveney PV (2011) Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor. *J Comput Chem* 32:2843–2852
- Amadasi A, Surface JA, Spyraakis F, Cozzini P, Mozzarelli A, Kellogg GE (2008) Robust classification of “relevant” water molecules in putative protein binding sites. *J Med Chem* 51:1063–1067
- Barillari C, Taylor J, Viner R, Essex JW (2007) Classification of water molecules in protein binding sites. *J Am Chem Soc* 129:2577–2587
- Treesuwan W, Hannongbua S (2009) Bridge water mediates nevirapine binding to wild type and Y181C HIV-1 reverse transcriptase—evidence from molecular dynamics simulations and MM-PBSA calculations. *J Mol Graph Model* 27:921–929
- Santos R, Hritz J, Oostenbrink C (2009) Role of water in molecular docking simulations of cytochrome P450 2D6. *J Chem Info Model* 50:146–154
- Huang N, Shoichet BK (2008) Exploiting ordered waters in molecular docking. *J Med Chem* 51:4862–4865

24. Šali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815
25. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 26:283–291
26. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65:712–725
27. Sousa da Silva AW, Vranken WF (2012) ACPYPE—AnteChamber PYthon Parser interfAcE. *BMC Res Notes* 5:367
28. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *J Comput Chem* 25:1157–1174
29. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 25:247–260
30. Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: fast, flexible, and free. *J Comput Chem* 26:1701–1718
31. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theor Comput* 4:435–447
32. Sorin EJ, Pande VS (2005) Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys J* 88:2472–2493
33. DePaul AJ, Thompson EJ, Patel SS, Haldeman K, Sorin EJ (2010) Equilibrium conformational dynamics in an RNA tetraloop from massively parallel molecular dynamics. *Nucleic Acids Res* 38:4856–4867
34. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684–3690
35. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. *J Comput Chem* 18:1463–1472
36. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J Chem Phys* 98:10089–10092
37. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–8593
38. Åqvist J, Medina C, Samuelsson JE (1994) A new method for predicting binding affinity in computer-aided drug design. *Protein Eng* 7:385–391
39. Hansson T, Marelus J, Åqvist J (1998) Ligand binding affinity prediction by linear interaction energy methods. *J Comput Aided Mol Des* 12:27–35
40. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng* 8:127–134

Skin Sensitization Prediction Using Quantum Chemical Calculations: A Theoretical Model for the S_NAr Domain

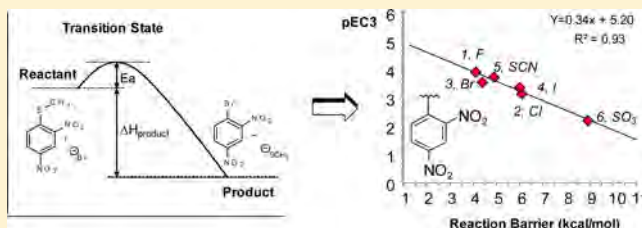
Malinee Promkatkaew,[†] Duangkamol Gleeson,[‡] Supa Hannongbua,[†] and M. Paul Gleeson^{*,†}

[†]Department of Chemistry, Faculty of Science, Kasetsart University, 50 Phaholyothin Road, Chatuchak, Bangkok 10900, Thailand

[‡]Department of Chemistry, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

S Supporting Information

ABSTRACT: It is widely accepted that skin sensitization begins with the sensitizer in question forming a covalent adduct with a protein electrophile or nucleophile. We investigate the use of quantum chemical methods in an attempt to rationalize the sensitization potential of chemicals of the S_NAr reaction domain. We calculate the full reaction profile for 23 chemicals with experimental sensitization data. For all quantitative measurements, we find that there is a good correlation between the reported pEC3 and the calculated barrier to formation of the low energy product or intermediate ($r^2 = 0.64$, $N = 12$) and a stronger one when broken down by specific subtype ($r^2 > 0.9$). Using a barrier cutoff of ~ 10 kcal/mol allows us to categorize 100% ($N = 12$) of the sensitizers from the nonsensitizers ($N = 11$), with just 1 nonsensitizer being mispredicted as a weak sensitizer (9%). This model has an accuracy of $\sim 96\%$, with a sensitivity of 100% and a specificity of $\sim 91\%$. We find that the kinetic and thermodynamic information provided by the complete profile can help in the rationalization process, giving additional insight into a chemical's potential for skin sensitization.



1.0. INTRODUCTION

Contact dermatitis is a common environmental and occupational health concern that arises from exposure to certain chemical substances. Contact dermatitis can be caused by the physical effects of chemical irritants on tissue directly (irritant contact dermatitis, ICD), which includes solvents, acids, or bases. An irritation may also result from a more extreme allergic response (allergic contact dermatitis, ACD), a complex phased response of the immune system to an allergen.¹ Experimental methods for the detection of sensitizers include the guinea pig maximization test (GPT) and the more recent murine local lymph node assay (LLNA).² The LLNA assay is now the method of choice following extensive validation and has been adopted by the OECD as a standard protocol.³ The assay works by identifying compounds with the capacity to provoke a T lymphocyte proliferative response within the lymph nodes. Chemicals are classified as sensitizers if they show a 3-fold or greater proliferative response in the induced draining lymph nodes compared with controls.³ While the EC3 is not an absolute response, it can be used to rank order compounds in terms of their relative toxicity. EC3 can be subclassified into strong, weak, and moderate sensitizers as shown in Table 1. According to the European Union's Registration, Evaluation, Authorization and Restriction of Chemical Substances Regulations (REACH), greater effort is needed to reduce the numbers of animals and the costs associated with toxicity testing. This requires the greater use of chemical assay surrogates^{4,5} and theoretical methods such as QSAR models and read-across methods.^{6,7}

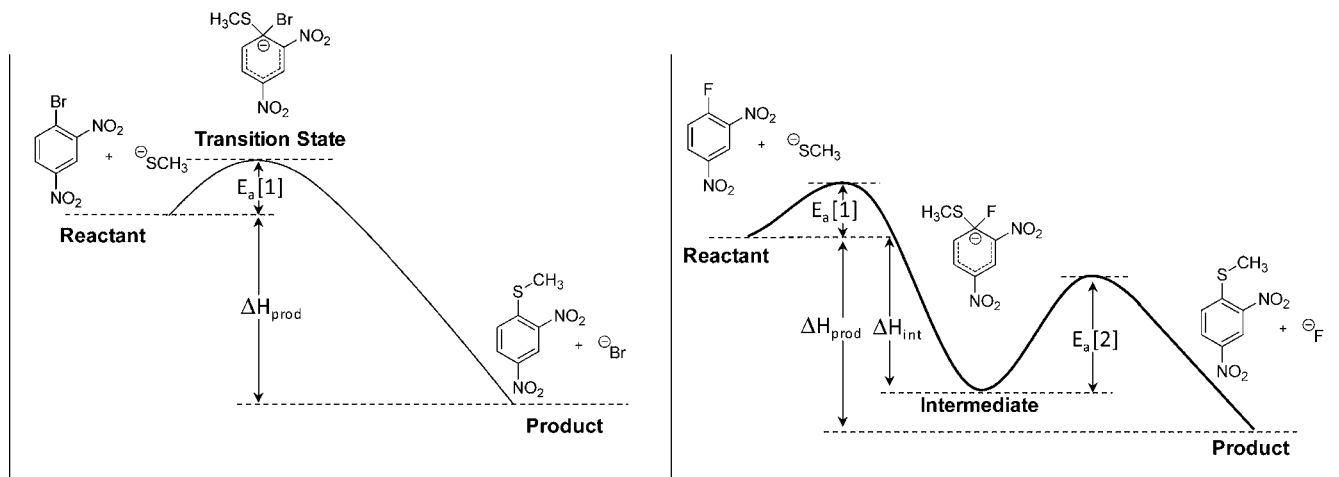
Table 1. EC3 Cut-Offs Used to Classify the Sensitization Potential of Chemicals in the LLNA Assay

potency classification	EC3 value (% weight)
nonsensitizer	NR
weak	≥ 10 to ≤ 100
moderate	≥ 1 to < 10
strong	≥ 0.1 to < 1
extreme	≤ 0.1

Skin sensitization begins with the sensitizer in question forming a covalent adduct with a protein electrophile or nucleophile. From the pioneering work in this field by Roberts and Aptula, skin sensitizing chemicals can be assigned to 5 separate chemical classes (or domains) capable of causing protein adducts: aromatic nucleophilic substitution (S_NAr), Schiff base formation (SB), Michael-type addition (MA), aliphatic nucleophilic substitution (S_N2), and acylation (Acy).^{8,9} The presence of structural or reactive features alone are not reliable indicators of toxicity,^{10,11} which is perhaps unsurprising given that a classification scheme neglects the overall molecular and local electronic characteristics of a molecule and the fact that a degree of target recognition may be present.

Attempts to develop truly global (i.e., covering a wide diversity of sensitizer types) quantitative structure–activity

Received: September 5, 2013

Scheme 1. Two Possible Nucleophilic Aromatic Substitution Reaction (S_NAr) Profiles for Chemicals in This Study^a

^aAddition–elimination via concerted (right) and stepwise processes.

relationships (QSAR), either by relative alkylation index (RAI) approaches¹² or by theoretical descriptor-based QSAR approaches,^{13,14} have not yet met with sufficient success to conform with the complete set of OECD QSAR guidelines.^{6,15} These guidelines are (a) a defined end point and (b) an unambiguous QSAR model, which is (c) mechanistically interpretable. In addition, the model must have (d) predictivity that is fit for the purpose and (e) a defined domain of applicability for which the model can be used. QSAR models that currently best fit the OECD principles are termed quantitative mechanistic models (QMM). These are restricted to chemicals from an individual reaction domain and thus resemble the simple but very functional QSARs first reported by Hansch and Fujita in the 1960s.¹⁶ These QSAR methods typically make use of experimentally derived physicochemical descriptors and are generally accurate for the particular chemical series under investigation.^{5,8,9,17,18} However, given that these methods rely on experimentally derived descriptors (i.e., σ electronic and π steric parameters), novel compounds cannot be predicted without first determining these parameters directly if they are not already known. Thus, a QMM-like approach based on purely theoretical methods would therefore be desirable if it could match the performance of that obtained with experimentally derived descriptors.

In a recent paper, Enoch and Roberts reported the development of a theoretically based QMM. This method relied on quantum chemical (QC) and an empirically based molecular descriptor to derive an LLNA QSAR for Michael acceptors.¹⁹ The authors approximated the rate determining barrier to reaction by using the energy of the high energy intermediate formed following the attack of a substrate by the negatively charged nucleophile (i.e., relying on the Hammond postulate to estimate the barrier),²⁰ and included an additional solvent accessible surface area term in their equation. This QC based protocol appears to be a significant improvement over the HOMO-LUMO estimate often used as a surrogate for reactivity.¹⁴ The model led to good discrimination between sensitizers and nonsensitizers for 26 compounds, with only 4 outliers.

In this article, we investigate the use of QC methods to rationalize the sensitization potential of chemicals. We start by collating the LLNA data in the literature to assess the

prevalence of skin sensitizers within the different reaction domains. On the basis of this analysis, we then focus on the most problematic domain. The electrophilic reactivity of the S_NAr domain, which was identified as the most toxic of all the five domains, is determined by a combination of the effects of the leaving group X and the activating groups Y. The reaction can occur when X is any halogen or pseudohalogen or a range of other groups which are not usually considered as good leaving groups (NO_2 , SO_2Ph , $SOPh$, and SO_3^-). We calculate the full reaction profile for 23 chemicals reported in the QMM study of Roberts et al.,²¹ providing complete details of the reaction kinetics and thermodynamics using a model sulfur nucleophile (Scheme 1). This is because it is not clear how the reactivity of the chemicals are influenced by kinetic and thermodynamic factors. Thus, computing the complete energy profile is preferable to estimate reactivity. We then use the kinetic and thermodynamic data to try and rationalize the experimentally reported sensitization results for the compounds in question. We also compare the discriminating potential of an approximated barrier to the S_NAr domain, as used by Enoch and Roberts for the MA domain.¹⁹ In addition, we assess the performance of the commonly used HOMO–LUMO band gap descriptor. We are particularly interested in determining whether the extra cost of the detailed profile is in any way beneficial over the latter two more approximate representations of reactivity.

2.0. COMPUTATIONAL PROCEDURES

Three different data sets were extracted in order to determine whether the chemical application affects the prevalence of chemical sensitizers. Topical drugs were obtained from ChEMBL,²² and the top 200 drugs (primarily oral) were taken from Stepan et al.¹⁰ The LLNA data set was created from 3 sources. Four hundred forty-three unique chemicals with LLNA test information were obtained from (1) ICCVAM (Interagency Coordination Committee on the Validation of Alternative Methods),²³ (2) Kern et al.,²⁴ and (3) Enoch et al.²⁵ The data was merged and cleaned using the following protocol: CAS numbers and/or smiles were rechecked; LLNA data for smiles duplicates were averaged; and compounds with contradictory measurements were excluded. The EC3 values of the 296 compounds with quantitative EC3 values were converted to the molar logarithmic parameter pEC3 ($-\log(MWT/EC3)$). Compounds were assigned to a reaction domain using SMARTs rules created by Enoch et al.²⁵ recoded in Pipeline Pilot 6.1,²⁶ as well as manually for the purpose of

comparison. Manual assignment was only performed in cases where it was not previously reported.²⁵ Compounds that were unambiguously assigned to a single reaction class were used for categorization analyses (~83%).

The S_NAr domain chemicals under QC investigation consist of the 23 halo- and pseudohalobenzenes from the publication of Roberts et al.²¹ Of these, 12 are reported skin sensitizers and 11 are nonsensitizers. An ⁻SCH₃ model of a typical cysteine nucleophile was used to simulate the reaction profile for these 23 compounds in line with others.¹⁹ We evaluate the following states for each chemical: (a) the isolated reactants, (b) the nonbonded reactant complex, (c) the bonded intermediate complex (if stable), (d) the nonbonded product complex, and (e) the isolated products, as well as (f) the transitions state(s) connecting the different states. The reaction coordinates of each chemical were modeled using a density functional theory (DFT) based QC model in Gaussian 09, revision C01.²⁷ In this case, we use the extensively validated M062X functional developed by Truhlar and co-workers,^{28–30} in conjunction with the 6-31+G(d,p) basis set. Stationary points were confirmed as such using vibrational frequency analysis. Transition states were confirmed as having a single negative frequency, while minima were confirmed to have none. Calculated properties (MWT and clogP) were obtained using the Chemaxon JChem software package.³¹

3.0. RESULTS AND DISCUSSION

We were interested in assessing whether chemicals used in a particular application (consumer vs pharmaceutical for example) are more likely to display differences in reactivity alerts. For example, it might be expected that drugs will be more carefully screened for reactive features (e.g., potentially leading to drug–drug interactions³²) than chemicals used in manufacturing or consumer products due to the high systemic concentration generally achieved.³³ All compounds were therefore assigned to their chemical domains using the SMARTs developed by Enoch et al.²⁵ We first assessed the concordance between manual and in-silico assignment of all chemicals in the LLNA data set, which had been assigned using both methods. For cases where a single, unambiguous assignment can be made by both methods (~70% of all 443 compounds), it was observed that the in-silico assignment was correct ~82% of the time. While lower than the reported statistics for the original 210 compounds used to generate the SMARTs patterns,²⁵ the performance of the method has still remained high.

Compounds in the oral drug data set are expected to achieve higher systemic exposure than the topically applied drug data set. The LLNA data set differs from the former two as it contains primarily nondrug-like compounds used in consumer products. Indeed, Figure 1 shows that the three different data sets are subtly different in terms of the proportion of nonreactive chemicals present (NR = 55%, 50%, and 39% for oral drugs, topical drugs, and consumer chemicals, respectively). This might have been expected given the extensive development testing of the latter set and also because the LLNA data set is likely artificially enriched with sensitizers. For example, many known sensitizers identified from other assays will have been used to validate the LLNA assay. Nevertheless, almost 50% of oral and topical drugs are members of one or more reactive domains. This confirms that the presence of a reactive structural alert in a molecule should not simply be taken as meaning a compound is high risk and should only be used in a weight of evidence approach.

Analysis of the compiled LLNA measurements broken down by the reaction domain reveals significant differences. The mean pEC₃ (quantitative measures only) of each reaction

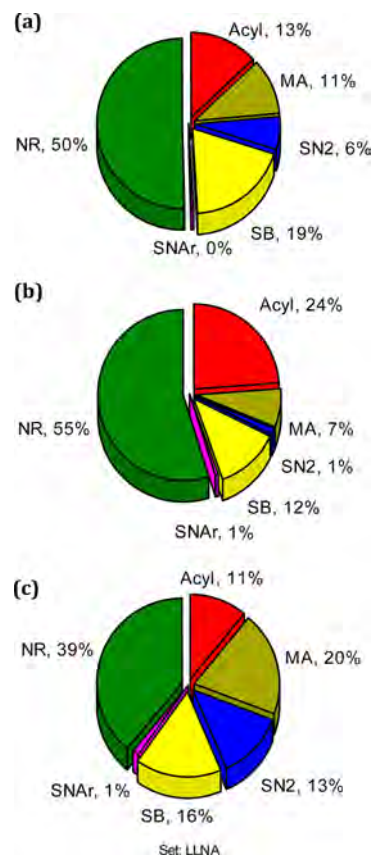


Figure 1. Pie chart showing the distribution of chemical domains within three different compound sets: (a) topical oral drugs, (b) the top-200 reported oral drugs, and (c) all compounds reported with LLNA measurements.

domain clearly shows that some are more likely to lead to severe skin sensitization than others (Figure 2). It is notable that the least represented type of compound in Figure 1 is that of the S_NAr domain. This class on average leads to the highest sensitization response of all of the chemical classes. MA, Acyl, and S_N2 are shown to display a comparable risk, with SB being

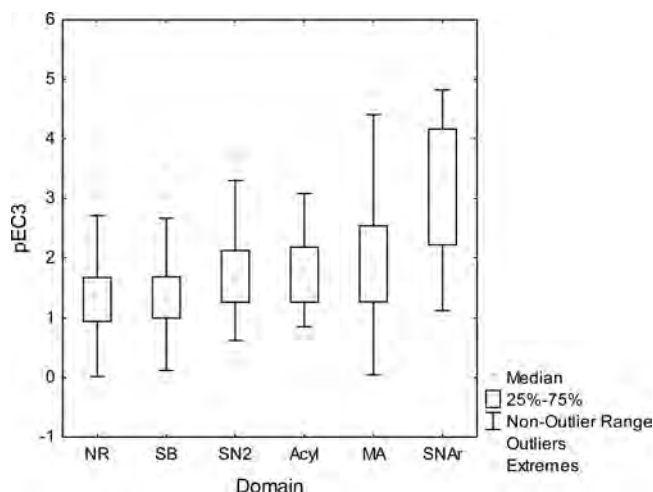


Figure 2. Relationship between skin sensitization potential (pEC₃) and chemical domain for 275 chemicals with both absolute pEC₃ values and a single defined chemical domain: SB, S_NAr, S_N2, Acyl, MA, and nonsensitizer (NR).

Table 2. Predicted QC Reaction Profiles for 23 Chemicals of the S_NAr Domain^a

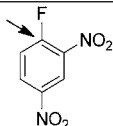
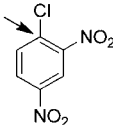
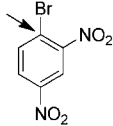
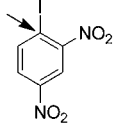
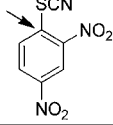
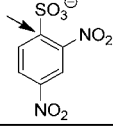
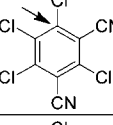
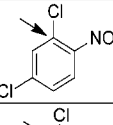
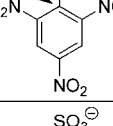
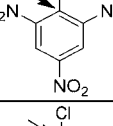
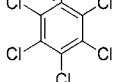
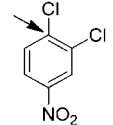
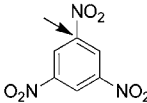
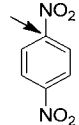
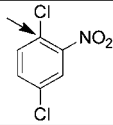
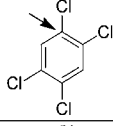
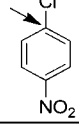
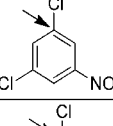
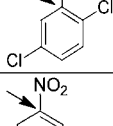
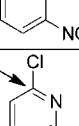
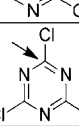
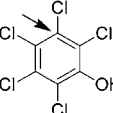
Structure		X	Class	<i>pEC3</i>	Relative energy (kcal/mol)						Homo-Lumo E	clogP
					Isolated Reactant	TS1	Inter	TS2	Product	Isolated Product		
1		F	Extreme	3.76	4.80	4.27	-14.04	-3.67	-12.29	-11.85	3.99	1.81
2		Cl	Extreme	3.42	5.25	5.84	-11.84	15.97	-42.36	-39.76	3.85	2.33
3		Br	Extreme	3.46	5.66	4.41	-	-	-41.75	-31.76	3.81	2.50
4		I	Strong	3.24	3.24	5.99	-	-	-48.60	-50.31	4.02	2.81
5		SCN	Extreme	3.68	5.26	4.90	-12.29	-12.28	-37.02	-33.44	3.70	2.13
6		SO ₃ [⊖]	Moderate	2.12	2.17	8.91	-9.37	9.13	-1.35	98.75	3.74	0.97
7		Cl	Extreme	4.88	5.37	5.37	-	-	-37.35	-34.52	4.38	3.88
	4.36					-37.92			-35.00			
	12.29					-37.25			-34.45			
8		Cl	Weak	0.98	4.13	9.93	-	-	-41.74	-39.68	4.11	3.07
9		Cl	Extreme	3.69	8.99	1.88	-27.36	-26.12	-41.52	-36.99	3.55	2.20
10		SO ₃ [⊖]	Strong	2.99	3.32	5.89	-24.55	1.56	0.44	103.98	3.70	0.85
11		Cl	NR	NR	4.28	15.29	-	-	-36.67	-34.61	5.41	5.68

Table 2. continued

Structure		X	Class	pEC3	Relative energy (kcal/mol)						Homo-Lumo E	clogP
					Isolated Reactant	TS1	Inter	TS2	Product	Isolated Product		
12		Cl	NR	NR	3.07	11.12	-	-	-39.48	-36.98	4.08	3.07
13		NO ₂	NR	NR	5.52	10.77	-	-	-23.06	-18.46	3.69	1.58
14		NO ₂	NR	NR	3.64	7.31	-0.67	0.90	-21.71	-18.57	3.59	1.71
15		Cl	NR	NR	4.41	10.59	-	-	-40.59	-39.08	4.09	3.07
16		Cl	NR	NR	2.86	19.03	-	-	-38.76	-37.47	5.61	4.44
17		Cl	NR	NR	2.11	13.06	-	-	-38.31	-37.15	4.19	2.45
18		Cl	NR	NR	3.07	18.96	-	-	-37.63	-36.37	4.06	3.07
19		Cl	NR	NR	2.51	20.43	-	-	-38.57	-37.39	5.81	3.82
20		NO ₂	NR	NR	3.93	18.03	-	-	-19.71	-16.31	3.97	1.71
21		Cl	Strong	2.33	3.16	8.69	-	-	-38.34	-37.43	5.18	2.20
22		Cl	Extreme	3.31	4.93	2.35	-	-	-42.88	-39.48	4.75	3.53
23		Cl	NR	NR	3.82	16.83	-	-	-36.41	-34.68	5.54	4.68

^aAlso reported are the measured skin sensitization pEC3 values (or corresponding NR results from the GPA assay). Also shown are the predicted HOMO-LUMO and clogP.

the least problematic on average. It is also worth noting that some chemicals classified as containing no reactive functionality also show sensitization potential. This, however, is over-estimated since most NR compounds do not have quantitative pEC3 measures so are excluded from the analysis (i.e., class based result only).

Attempts to develop trends with simple molecule properties were unsuccessful, which is consistent with reports by others.²¹ This may be due to the relatively small size and limited chemical diversity of the LLNA data sets currently available.³⁴ Nevertheless, Figures 1 and 2 clearly show the need for additional methods, on top of the reaction domain scheme, to help discriminate sensitizers from nonsensitizers more

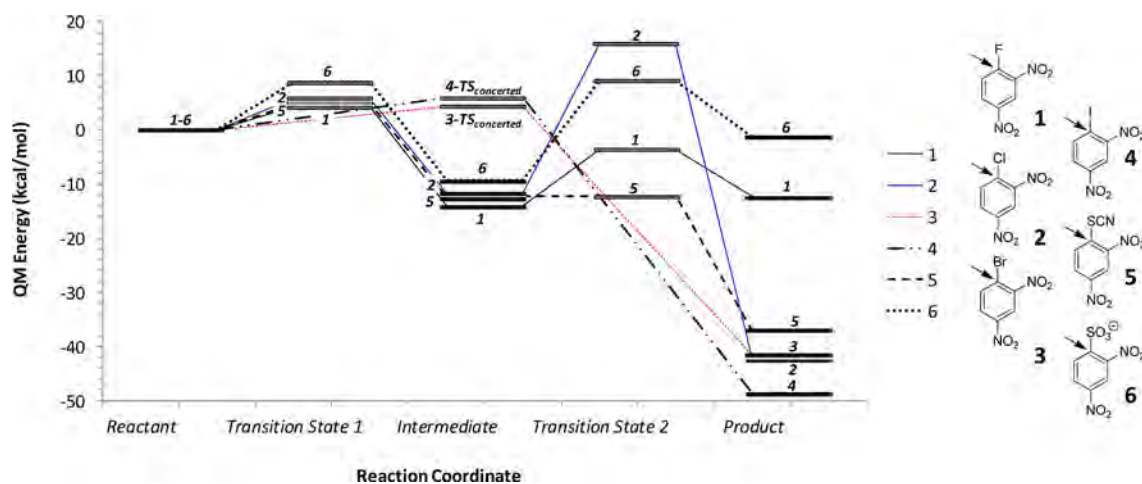


Figure 3. Reaction profiles obtained for compounds 1–6. Compounds 1, 2, 5, and 6 show a stepwise reaction profile, while for compounds 3 and 4, it is concerted.

effectively. It is also notable that the S_NAr domain represents a particularly significant threat. Nevertheless, such compounds make up 13% of oral drugs, 2% of topical drugs, and 2% of all compounds tested in the LLNA assay. In the next section, we discuss a purely theoretical QC based method suitable for use in the ranking of the sensitization potential of chemicals.

3.1. QC Model of Skin Sensitization. The addition–elimination reaction of the S_NAr domain chemicals modeled here are summarized in Scheme 1. On the basis of the gas-phase QC calculations, the displacement of the halogen or pseudohalogen group was found to proceed in either one or two steps. Addition of the nucleophile to the aromatic ring leads to the expected resonance-stabilized carbanion intermediate³⁵ in 7 out of the 23 cases. These intermediates are primarily found where the leaving group is less bulky (F, SCN, SO_3^- , NO_2 , and Cl in 2 out of 14 cases). For the Cl leaving group, it is also found that the ring system and substituents also play an important role. The majority of reactions are predicted to proceed via an S_N2 -like process that lacks the resonance stabilized intermediate. In this case, the resonance stabilized structure is found to be the transition state. From these calculations, it is clear that the nature of the leaving group (bulk and electronics), the stabilizing effect of the ring substituents, and the resonance effects can lead to dramatically different barrier heights, as well as the profile.²¹ We note that these calculations lack solvent effects, which may lead to the reaction profile changing from the expected S_NAr process to an S_N2 process. Nevertheless, we still expect the computed barriers to be a reasonable reflection of the relative reactivity of each of the chemicals under investigation here.

The complete reaction profile for the 23 chemicals studied here are reported in Table 2. The first step in the reaction is expected to be the rate determining step due to the loss of aromaticity. If the reactivity of the sensitizing chemicals is under kinetic control, we would expect the RDS to correlate well with the experimental pEC3s. Alternatively, should the process depend on the overall thermodynamics, we would expect the exothermicity of the products to be important. The reaction profiles of compounds 1–6 are shown pictorially in Figure 3. It can be seen for those chemicals with halogen or pseudohalogen leaving groups and common core (i.e., 2,4-dinitrobenzene), that the profile can vary substantially. It is apparent that compounds with more effective leaving groups, Br- (3) and I- (4), display

the expected stepwise S_NAr reaction and result in highly exothermic products. In contrast, compounds 1, 2, 5, and 6 follow a concerted pathway. For compounds 1 and 6, the products are equivalent or somewhat higher in energy than the corresponding intermediates or reactants, and the second step is rate determining for the full addition–elimination process. Compound 2 displays a higher barrier for the second step. While this result might suggest that the intermediate can also lead to the sensitization response, rather than the elimination product, it may be a subtle artifact of the method due to the lack of solvent stabilization.

For the 2,4-dinitrobenzene series of compounds, analysis of the correlation between the quantitative pEC3 value and the computed barrier to intermediate/product shows a rather strong correlation ($r^2 = 0.93$) (Figure 4). Plotting the rate determining step to product formation (i.e., for the stepwise process it may correspond to transition state 2) does not improve the correlation. As mentioned above, the lack of solvent in the simulation may in some cases make the carbanion intermediate appear more stable than it is in reality (see Figure 2). Analysis of a more diverse set of chemicals that contain a common Cl leaving group, shows a moderate correlation between the pEC3 and the barrier ($r^2 = 0.58$). More importantly, the line of best fit for both relationships in Figure 4 are remarkably similar suggesting that a single QC descriptor is needed to explain the sensitization potential, irrespective of ring substituents or leaving group.

Compound 7 appears to be an outlier. Like compounds 11, 16, and 23, compound 7 has substituents at all six phenyl positions. However, compound 7 has both more strongly electron withdrawing CN substituents and three distinct positions for nucleophilic attack. To try and understand whether this was the cause of this compound being an outlier, we investigated the reaction profile associated with the two other distinct addition–elimination positions (Figure 5). However, the other two positions of attack, not initially considered, were as expected considerably higher in energy and were thus not the reason for compound 7 being an outlier. We note that the r^2 obtained from a plot of the observed pEC3 vs QC barrier for the chemicals with Cl leaving groups (Figure 4) would increase from 0.58 to 0.76 were compound 7 to be removed. However, no good reason exists for excluding this compound, and we do not consider it prudent to include an

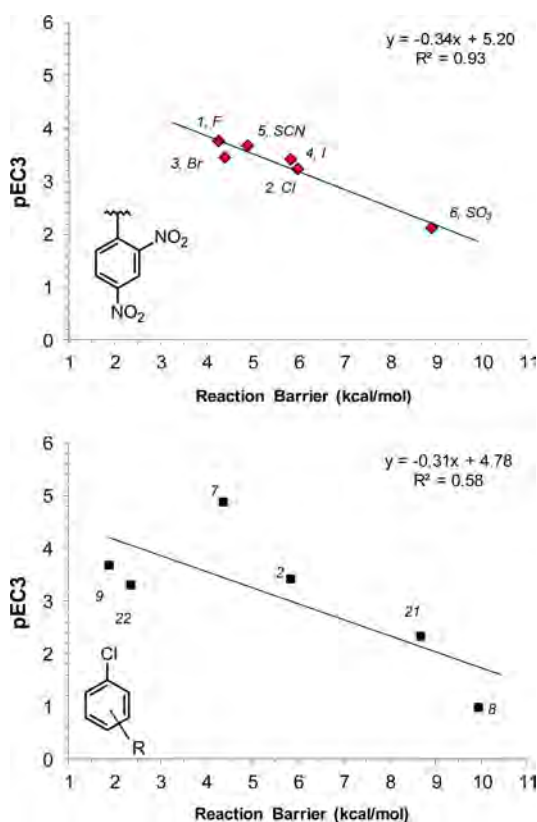


Figure 4. Plot of the pEC3 vs the predicted barrier to reaction for chemicals with a common dinitro-phenyl core but different halogen or pseudohalogen leaving groups (top). A more diverse set of chemicals with a common chloro leaving group (bottom).

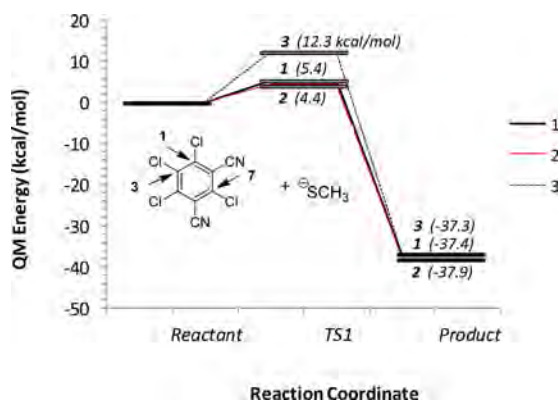


Figure 5. Three distinct concerted reaction profiles observed for compound 7.

additional term in this model to account for lipophilicity or other factors that might also contribute. This is because the small number of data points available are not sufficient to reliably test any hypothesis or to fit a multiparameter QSAR equation which can conform to the OECD QSAR guidelines. We cannot of course rule out contributory factors including (a) the use of a suboptimal QC description in this study, (b) the fact that confounding factors associated with molecular recognition between a protein and sensitizer are lacking, or (c) that a lack of concordance is due to experimental error.³

In Figure 6, we plot the predicted QC barrier to formation of the stable product or intermediate vs the pEC3 (or class) for all 23 chemicals investigated here. The correlation of pEC3 vs the

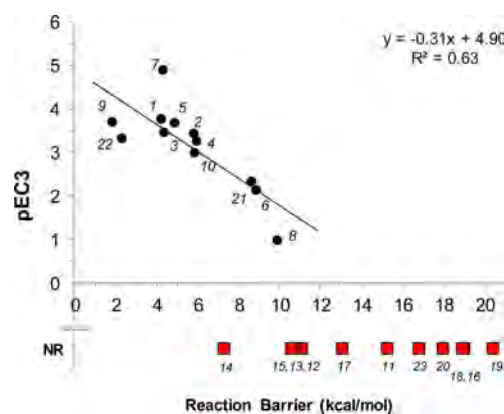


Figure 6. Plot of the LLNA pEC3 vs the predicted barrier for all chemicals investigated in this study. NR compounds are those determined as nonreactive in the GPA assay and are included for the purpose of comparison.

barrier for the 12 chemicals with quantitative pEC3s is reasonably strong ($r^2 = 0.63$). Barriers to reaction below ~ 10 kcal/mol indicate a sensitizer, and the absolute value can be related directly to the strength of sensitization response. Compound 7 still appears to be an outlier as discussed above, being more potent than predicted. This could be in part due to its higher than average clogP.¹⁷ Again, discarding this outlier would result in an $r^2 = 0.74$.

The least potent sensitizer has a predicted barrier of 9.93 kcal/mol. A cutoff of ~ 10 kcal/mol could be used to separate the sensitizers from nonsensitizers with 100% accuracy ($N = 12$), with just 1 nonsensitizer out of 11 being mispredicted as a sensitizer (9%). This corresponds to an overall accuracy of $\sim 96\%$, with a sensitivity of 100% and a specificity of $\sim 91\%$. For quantitative measurements, we find that there is a high correlation between the experimental pEC3 and the calculated barrier to formation of the low energy product or intermediate. The nonsensitizer mispredicted is compound 14. However, analysis of the QC data shows that compound 14 has a rather unstable product compared to that of the others chemicals (only 1, 6, 10, and 20 have higher product energies). The relatively low barrier to reaction but also the only moderately stable product appears to act as a counter balance.

The S_NAr QMM model reported by Roberts et al.²¹ consists of a QSAR equation that takes into account both the electron withdrawing effect of the ring substituents (i.e., electronegative inductive effects of substituents and resonance effects) and the strength of the leaving group (i.e., electronegative inductive effects). Their 2 descriptor model was also effective at discriminating between chemical sensitizers and nonsensitizers as can be seen in Supporting Information, Figure S1 and Table S1. They obtained an r^2 of 0.41 for the 12 compounds with quantitative pEC3 values, although it should be noted that some of these compounds were also used to fit the model. The correlation is lower than the 0.63 observed for the QC model developed here. In addition, using the suggested cutoff of ~ 1 for their model, 100% of the sensitizers are correctly classified, but 5 of the nonsensitizers would also be misclassified (i.e., 45%). This suggests the 2 descriptor QMM is somewhat inferior to the 1 descriptor QC model developed here. The key difference between the two methods is that the QC based model accounts for the same electronic terms used in the QMM but also accounts for interaction terms (steric and electrostatic) between the ring, substituents, and leaving and

attacking groups implicitly. The QC model is also unambiguous since QMMs require fitting and can also require user modification of standard physicochemical substituent parameters to give optimal results.²¹ Nevertheless, an advantage of the QMM approach is that estimations take seconds to minutes when substituent parameters are available. In contrast, simulations to obtain the complete reaction profile can take 1 day per compound per computer (i.e., Intel i7).

The recent study by Enoch and Roberts¹⁹ on the MA reaction domain is worth comparing and contrasting to the approaches used here. The authors in the former study approximated the barrier by calculating the energy of the intermediate, minus the isolated energies of the reactants (i.e., Michael acceptor and $^-SCH_3$), which we term E_{int} . This approximation assumes that the barrier heights are directly proportional to the energy of the intermediates. This approach will not work in situations where no stable intermediate is formed. The authors found that the correlation between the approximated barrier and the pEC3 for 26 compounds (4 outliers removed) was sizable at $r^2 = 0.43$. Addition of a single additional descriptor (solvent accessible surface area) and removal of an additional outlier led to a much better correlation ($r^2 = 0.79$). In our study of the S_NAr domain, only 6 of the compounds react via a stepwise process, thus using the E_{int} measure is limited, leading to a rather poor correlation with the pEC3 ($r^2 = 0.08$, $N = 6$ for compounds 1, 2, 5, 6, 9, and 10). It is also apparent from Figure 3 that in our case the barrier heights do not necessarily correlate well with either the intermediate or product energy, helping to explain the poor correlation between pEC3 with E_{int} . Finally, it is also worth noting that the correlation between the HOMO–LUMO energy and pEC3 for all 12 compounds with quantitative pEC3 measurements is negligible ($r^2 = 0.02$). This suggests that the latter parameter is not such a good surrogate for the barrier height of the S_NAr domain compounds assessed here.

4.0. CONCLUSIONS

In this article, we have reported the use of a QC based approach to assess skin sensitization potential. We have used a model $^-SCH_3$ nucleophile to predict kinetic and thermodynamic parameters associated with the addition–elimination reaction for a set of 23 chemicals from the S_NAr domain. We find that calculating the full reaction profile for the chemicals is important since, as highlighted in Scheme 1, the reactions can proceed by either concerted or stepwise addition–elimination processes depending on the activating substituents, ring resonance effects, and the nature of the leaving group. It does not appear to be suitable to approximate the transition state with either the HOMO–LUMO energy or the energy of the high energy intermediate.

We find that the use of a single computed descriptor, namely, the barrier to formation of the stable product or intermediate can help us to separate sensitizers and nonsensitizers. Barriers to reaction below ~ 10 kcal/mol indicate a sensitizer, and the absolute value can be related directly to the strength of sensitization. The use of a cutoff of ~ 10 kcal/mol allows us to categorize 100% ($N = 12$) of the sensitizers from the nonsensitizers ($N = 11$), with just 1 nonsensitizer being mispredicted as a weak sensitizer (i.e., 9%). This corresponds to a sensitivity of 100% and a specificity of $\sim 91\%$. For quantitative measurements, we find that there is a high correlation between the experimental pEC3 and the calculated barrier to formation of the low energy product or intermediate. We find an $r^2 = 0.64$

for all 23 chemicals, compared to $r^2 = 0.41$ for the comparable QSAR based approach reported elsewhere.²¹ The one non-sensitizer found to be an outlier can be rationalized by a consideration of the reaction thermodynamics. In the case of compound 14, it has a low barrier to reaction but forms a less stable product than most other sensitizers and nonsensitizers..

Physical chemistry approaches such as QSARs^{5,8,9,17,18} based on physicochemical parameters and substituent constants, or QC calculations¹⁹ have proved useful in helping discriminating sensitizers from nonsensitizers. The physical insight and understanding that can be garnered from QC methods could prove useful for skin sensitization assessment, especially when combined in the so-called weight of evidence approach with other methods. QC calculations by necessity must employ surrogate nucleophiles for what is a complex biological process, and that is a key limitation. However, we postulate that the experimental identification of the most prevalent nucleophiles or indeed the precise proteins that cause skin sensitization would provide an additional means to help improve the performance of such atomic simulations.

■ ASSOCIATED CONTENT

Supporting Information

Optimized coordinates and the full list of LLNA data compiled for this study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: paul.gleeson@ku.ac.th.

Funding

M.P.G. would like to acknowledge the financial support from the Thailand Research Fund (TRF) (RSA5480016). M.P. is grateful to the TRF for her Royal Golden Jubilee (RGJ) Ph.D. scholarship (TRF RTA5380010, PhD/0009/2552).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the help of Dr. Sandeep Modi and Dr. Aynur Aptula who helped with the classification of chemicals into reaction domains. We also thank Professor David Roberts for providing insight into the QMM developed for the S_NAr domain.

■ ABBREVIATIONS

ChEMBL, Chemistry at European Bioinformatics Institute (database); clogP, calculated logarithm of the octanol–water partition coefficient; DFT, density functional theory; EC3, concentration needed to produce a 3-fold increase response compared to that of vehicle-treated controls; GPT, guinea pig maximization test; HOMO, highest occupied molecular orbital; ICCVAM, Interagency Coordination Committee on the Validation of Alternative Methods; ICD, irritant contact dermatitis; LLNA, local lymph node assay; LUMO, lowest unoccupied molecular orbital; M062X, a DFT methodology; MA, Michael acceptor; MWT, molecular weight; NR, non-reactive; OECD, Organization for Economic Cooperation and Development; PCM, polarizable continuum model; pEC3, the log of the molar EC3; QC, quantum chemical; QMM, quantitative mechanistic model; QSAR, quantitative structure–activity relationship; RAI, relative alkylation index;

REACH, registration, evaluation, authorization and restriction of chemical substances regulations; SB, Schiff base; S_N1 , substitution nucleophilic 1 (unimolecular); S_N2 , substitution nucleophilic 2 (bimolecular); S_NAr , aromatic nucleophilic substitution

REFERENCES

- (1) Kimber, I., Basketter, D. A., Gerberick, G. F., and Dearman, R. J. (2002) Allergic contact dermatitis. *Int. Immunopharmacol.* 2, 201–211.
- (2) Kimber, I., and Weisenberger, C. (1989) A murine local lymph node assay for the identification of contact allergens. Assay development and results of an initial validation study. *Arch. Toxicol.* 63, 274–282.
- (3) Anderson, S. E., Siegel, P. D., and Meade, B. J. (2011) The LLNA: a brief review of recent advances and limitations. *J. Allergy* 2011, doi: 10.1155/2011/424203.
- (4) Gerberick, F., Aleksic, M., Basketter, D., Casati, S., Karlberg, A. T., Kern, P., Kimber, I., Lepoittevin, J. P., Natsch, A., Ovigne, J., Rovida, C., Sakaguchi, H., and Schultz, T. (2008) Chemical reactivity measurement and the predictive identification of skin sensitizers. *Altern. Lab. Anim.* 36, 215–242.
- (5) Roberts, D. W., Schultz, T. W., Wolf, E. M., and Aptula, A. O. (2009) Experimental reactivity parameters for toxicity modeling: application to the acute aquatic toxicity of $SN2$ electrophiles to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* 23, 228–234.
- (6) Patlewicz, G., Chen, M. W., and Bellin, C. A. (2011) Non-testing approaches under REACH – help or hindrance? Perspectives from a practitioner within industry. *SAR QSAR Environ. Res.* 22, 67–88.
- (7) Schaafsma, G., Kroese, E. D., Tielemans, E. L. J. P., Van de Sandt, J. J. M., and Van Leeuwen, C. J. (2009) REACH, non-testing approaches and the urgent need for a change in mind set. *Regul. Toxicol. Pharmacol.* 53, 70–80.
- (8) Aptula, A. O., and Roberts, D. W. (2006) Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity. *Chem. Res. Toxicol.* 19, 1097–1105.
- (9) Aptula, A. O., Patlewicz, G., and Roberts, D. W. (2005) Skin sensitization: reaction mechanistic applicability domains for structure–activity relationships. *Chem. Res. Toxicol.* 18, 1420–1426.
- (10) Stepan, A. F., Walker, D. P., Bauman, J., Price, D. A., Baillie, T. A., Kalgutkar, A. S., and Aleo, M. D. (2011) Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol.* 24, 1345–1410.
- (11) Park, B. K., Boobis, A., Clarke, S., Goldring, C. E. P., Jones, D., Kenna, J. G., Lambert, C., Lavery, H. G., Naisbitt, D. J., Nelson, S., Nicoll-Griffith, D. A., Obach, R. S., Routledge, P., Smith, D. A., Tweedie, D. J., Vermeulen, N., Williams, D. P., Wilson, I. D., and Baillie, T. A. (2011) Managing the challenge of chemically reactive metabolites in drug development. *Nat. Rev. Drug Discovery* 10, 292–306.
- (12) Roberts, D. W., Fraguinals, R., Lepoittevin, J. P., and Benezra, C. (1991) Refinement of the relative alkylation index (RAI) model for skin sensitization and application to mouse and guinea-pig test data for alkyl alkanesulphonates. *Arch. Dermatol. Res.* 286, 387–394.
- (13) Li, Y., Tseng, Y. J., Pan, D., Liu, J., Kern, P. S., Gerberick, G. F., and Hopfinger, A. J. (2007) 4D-Fingerprint categorical QSAR models for skin sensitization based on the classification of local lymph node assay measures. *Chem. Res. Toxicol.* 20, 114–128.
- (14) Miller, M. D., Yourtee, D. M., Glaros, A. G., Chappelow, C. C., Eick, J. D., and Holder, A. J. (2005) Quantum mechanical structure–activity relationship analyses for skin sensitization. *J. Chem. Inf. Model.* 45, 924–929.
- (15) Roberts, D. W., Patlewicz, G., Kern, P. S., Gerberick, F., Kimber, I., Dearman, R. J., Ryan, C. A., Basketter, D. A., and Aptula, A. O. (2007) Mechanistic applicability domain classification of a local lymph node assay dataset for skin sensitization. *Chem. Res. Toxicol.* 20, 1019–1030.
- (16) Hansch, C., and Fujita, T. (1964) p - σ - π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* 86, 1616–1626.
- (17) Roberts, D. W., Aptula, A. O., and Patlewicz, G. (2006) Mechanistic applicability domains for non-animal based prediction of toxicological endpoints. QSAR analysis of the Schiff base applicability domain for skin sensitization. *Chem. Res. Toxicol.* 19, 1228–1233.
- (18) Roberts, D. W., Aptula, A. O., and Patlewicz, G. (2007) Electrophilic chemistry related to skin sensitization. reaction mechanistic applicability domain classification for a published data set of 106 chemicals tested in the mouse local lymph node assay. *Chem. Res. Toxicol.* 20, 44–60.
- (19) Enoch, S. J., and Roberts, D. W. (2013) Predicting skin sensitization potency for michael acceptors in the LLNA using quantum mechanics calculations. *Chem. Res. Toxicol.* 26, 767–774.
- (20) Hammond, G. S. (1955) A correlation of reaction rates. *J. Am. Chem. Soc.* 77, 334–338.
- (21) Roberts, D. W., Aptula, A. O., and Patlewicz, G. Y. (2011) Chemistry-based risk assessment for skin sensitization: quantitative mechanistic modeling for the $S(N)Ar$ domain. *Chem. Res. Toxicol.* 24, 1003–1011.
- (22) ChEMBL <https://www.ebi.ac.uk/chembl/> (last accessed Nov 14, 2013).
- (23) ICCVAM, LLNA database, <http://iccvam.niehs.nih.gov/methods/immunotox/LLNA.htm> (last accessed Nov 14, 2013).
- (24) Kern, P. S., Gerberick, F., Ryan, C. A., Kimber, I., Aptula, A., and Basketter, D. (2010) Local lymph node data for the evaluation of skin sensitization alternatives: a second compilation. *Dermatitis* 21, 8–32.
- (25) Enoch, S. J., Madden, J. C., and Cronin, M. T. (2008) Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach. *SAR QSAR Environ. Res.* 19, 555–578.
- (26) Pipeline Pilot www.accelrys.com (last accessed Nov 14, 2013).
- (27) Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Montgomery, J. J. A., Vreven, T., Kudin, K. N., Burant, J. C., Millam, J. M., Iyengar, S. S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G. A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J. E., Hratchian, H. P., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Ayala, P. Y., Morokuma, K., Voth, G. A., Salvador, P., Dannenberg, J. J., Zakrzewski, V. G., Dapprich, S., Daniels, A. D., Strain, M. C., Farkas, O., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Ortiz, J. V., Cui, Q., Baboul, A. G., Clifford, S., Cioslowski, J., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Gonzalez, C., and Pople, J. A. (2004) Gaussian 03, revision C.02, Gaussian, Inc., Wallingford, CT.
- (28) Zhao, Y., and Truhlar, D. G. (2011) Applications and validations of the Minnesota density functionals. *Chem. Phys. Lett.* 502, 1–13.
- (29) Zhao, Y., and Truhlar, D. G. (2008) Exploring the limit of accuracy of the global hybrid meta density functional for main-group thermochemistry, kinetics, and noncovalent interactions. *J. Chem. Theory Comput.* 4, 1849–1868.
- (30) Zhao, Y., and Truhlar, D. G. (2008) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* 120, 215–241.
- (31) ChemAxon JChem, <http://www.chemaxon.com/jchem/> (last accessed Nov 14, 2013).
- (32) Kalgutkar, A. S., and Soglia, J. R. (2005) Minimising the potential for metabolic activation in drug discovery. *Expert Opin. Drug. Metab. Toxicol.* 1, 91–142.

- (33) Stumpf, W. E. (2006) The dose makes the medicine. *Drug Discovery Today* 11, 550–555.
- (34) Weaver, S., and Gleeson, M. P. (2008) The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* 26, 1315–1326.
- (35) Schwöbel, J. A. H., Koleva, Y. K., Enoch, S. J., Bajot, F., Hewitt, M., Madden, J. C., Roberts, D. W., Schultz, T. W., and Cronin, M. T. D. (2011) Measurement and estimation of electrophilic reactivity for predictive toxicology. *Chem. Rev.* 111, 2562–2596.

9th Annual Conference

INTELLIGENT COMPOUND LIBRARIES 2013

Achieving real productivity in drug discovery – Collaborative Working-Structures in new screening centres – Predicting adverse effects with innovative computational informatics

28 – 29 October 2013 | Steigenberger Hotel Berlin, Germany

The only Compound Libraries conference in Europe!

Don't miss our expert speaker panel:

- AstraZeneca plc
- Sanofi-Aventis Deutschland GmbH
- The European Lead Factory
- Janssen Research & Development, LLC
- Centre de Recherche en Cancérologie de Marseille (CRCM)
- Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI
- Asclepia MedChem Solutions
- Mercachem b.v
- Institute of Pharmaceutical Sciences
- Université Bordeaux Ségalen
- Leibniz Institute of Molecular Pharmacology FMP
- Kasetsart University
- F. Hoffmann-La Roche Ltd.
- Actelion Pharmaceuticals Ltd
- Boehringer Ingelheim Pharma GmbH & Co. KG

Highlights of this year's Compound Libraries conference:

- Understand the impact of the **paradigm shift** in the pharma industry to **push innovation** in your compound selection
- Optimize screening in **Early ADME** to enhance the prediction of **adverse effects**
- Assess the impact of **polypharmacology**, **signalling pathways** and **systems biology** on compound validation
- Discover new **computational methods of screening mixtures** to increase output and hit-to-lead
- Listen to the AstraZeneca-Bayer Alliance to improve **collaborative working** in new **screening centres** and to cut costs and expedite deadlines

Learn from these expert speakers amongst others:



Dr. Peter Ten Holte,
Global Business Office Liaison,
Janssen Research & Development, LLC, USA



Urs Luethi,
Senior Lab Head HTS,
Actelion Pharmaceuticals Ltd, Switzerland



Werngard Czechtizky,
Section Head, Medicinal Chemistry,
Sanofi-Aventis Deutschland GmbH, Germany



Philippe Roche,
PhD, Laboratory of 'integrative Structural & Chemical Biology (iSCB)' Cancer Research Center of Marseille (CRCM),
Université Aix-Marseille Institut Paoli Calmettes, France

Interactive Evening Workshops

- A** | New business models for successful compound library acquisition – How do I best spend my compound library budget?
- B** | The challenge of reducing compound library size without reducing the quality and success rate for lead generation

Sponsors



Researched and developed by



a division of IQPC



SAVE
up to € 320,- with our
Early Birds if you book
and pay by
02 August 2013!

INTELLIGENT COMPOUND LIBRARIES 2013

28 – 29 October 2013 | Steigenberger Hotel Berlin, Germany

Dear Colleagues,

The world of compound libraries is changing time and again. We perform **screenings** and **validations** in order to provide better predictions. We may ask ourselves every day important questions such as:

- What are the properties of the **"ideal" chemical library**? (number of compounds, diversity, chemical space coverage, fragments or small molecules...)
- Which global strategy to adopt for the **acquisition** of new compounds?
- How to design **chemical libraries** for specific targets? (Design of focused chemical libraries)
- How to give new adventures to old drugs? (importance of drug **repositioning**)
- How to assess the **quality** of chemical collections? (storage, quality control, ADME/Toxicity...)
- Which **strategy** to screen the compounds? (target selection, screening bioassays, integration of system biology, collaboration between pharma industry and academics,...)

All of these are very important questions, which every colleague and every manager within a pharma company, ranging from **small-size up to large international companies**, should be concerned about.

In an atmosphere of structural changes and constant search for **innovative tools**, combining a **successful collaboration strategy** with **cutting-edge new scientific approaches** is a must. Be sure to update yourself on how others have successfully accomplished progress and network with potential partners.

The '9th Intelligent Compound Libraries' conference is offering you a great opportunity to do just that.

You will meet a great team of global experts in Berlin, who truly understand the value of knowledge-exchange and networking.

Seize this opportunity to network and listen to some of the most innovative ideas, for you to implement in your own company as well.

Yours sincerley,

Philippe Roche
*PhD, Laboratory of 'integrative Structural & Chemical Biology (iSCB)' Cancer Research Center of Marseille (CRCM), **Université Aix-Marseille Institut Paoli Calmettes, France***

Sponsors



Vitas-M Laboratory
Hodyski Blvd. 15 Moscow
125252, Russia
Phone: +7 495 939 4831
Fax: +7 495 939 5737
Email: irina@vitasmlab.com
www.vitasmlab.com



WuXi AppTec
288 Fute Zhong Road
Waigaoqiao Free Trade Zone
Shanghai 200131, China
Phone: +86 (21) 5046-1111
Fax: +86 (21) 5046-1000
Email: info@wuxiapptec.com
www.wuxiapptec.com

Sponsorship

We have a variety of packages available to suit your requirements. For all Sponsorship and Exhibition opportunities call our Sponsorship-Team: **+49 (0)30 20 91 32 75** or email sponsorship@iqpc.de

Previous attendees at our conference include:

Actelion Pharmaceuticals Ltd. AstraZeneca R&D Mölndal Astellas Pharma Inc
Bayer Pharma AG Centre National de la Recherche Scientifique Beiersdorf AG
Vipergen A/S Eli Lilly and Company Leibniz Institut für Molekularische Pharmakologie- FMP
Grünenthal Pharma GmbH & Co.KG F. Hoffmann-La Roche AG Pfizer Ltd.
GlaxoSmithKline UK Limited Vernalis R&D Ltd Janssen Research & Development
WuXi Apptec (Tianjin) Co. Ltd. Boehringer Ingelheim Pharma GmbH & Co. KG
Laboratoire de Biologie Moléculaire et Cellulaire du Cancer (LBMCC) Saarland University
Novartis International AG Janssen Pharmaceutica NV Ranbaxy Laboratories Ltd.
WuXi PharmaTech. Ltd. Shionogi & Co. Ltd. Sanofi-Aventis Deutschland GmbH
Takeda Pharmaceutical Company Ltd. Universität Karlsruhe University of Aberdeen
DainipponSumitomoPharmaCo.Ltd. Wisdom Pharmaceutical Daiichisankyo Co.Ltd. UCBPharmaSANTV

08:15 Registration & coffee



Who is Who?

Learn about your peers. Discover who else is participating in the conference and gain an understanding of the core challenges your peers are facing right now. The match-making wall will also help you identify the delegates you want to meet at the conference.

08:55 Chairman's welcome & opening address



Dr. Peter Ten Holte, Global Business Office Liaison,
Janssen Research & Development, LLC, USA

Compound collections: Updates, new trends and paradigm shift

09:00 **The paradigm shift in the pharma industry and its impact on in-house innovation and especially compound selection**

- Changes in the structure of the industry's way of doing business
- New scientific trends
- Partnerships, outsourcing and strategic alliances
- What is yet to come? How do we pursue innovation?



Philippe Roche, PhD, Laboratory of 'integrative Structural & Chemical Biology (iSCB)' Cancer Research Center of Marseille (CRCM),
Université Aix-Marseille Institut Paoli Calmettes, France



Dr. Peter Ten Holte, Global Business Office Liaison,
Janssen Research & Development, LLC, USA

New scientific directions for enhancing target validation

09:30 **Cutting costs and efforts by the combinatorial de novo design – The new adventures of old drugs**

- Computer-based de novo design
- Virtual synthesis
- Scaffold-hopping
- Fragment evolution and grafting



Prof. Gisbert Schneider, Professor Chem- & Bioinformatics,
Institute of Pharmaceutical Sciences, Switzerland

10:05 Refreshment break & **speed networking**



Get in touch with the other conference guests for a quick exchange of views (and business cards)

11:05 **How to make most of in-house compound libraries – The impact of library quality and variety of screening scenarios on the screen to lead phase**

- How to make most of in-house compound collections: the traditional single target HTS approach vs. alternative screening scenarios
- Intelligent assembly of screening collections: how to enrich internal compound collections with public/private external collections
- Novel lead finding/optimization libraries: added value to drug discovery programs and compound collections through high throughput biological, physchem and eADME profiling



Werngard Czechtizky, Section Head, Medicinal Chemistry,
Sanofi-Aventis Deutschland GmbH, Germany

11:40 **Development of focused chemical libraries dedicated to protein-protein interactions: an academic perspective**

- Biological and chemical spaces of PPI with known orthosteric inhibitors
- Druggability assessment of protein-protein interactions
- Design and validation of chemical libraries dedicated to PPI
- Experimental validation with an academic screening platform



Philippe Roche, PhD, Laboratory of 'integrative Structural & Chemical Biology (iSCB)' Cancer Research Center of Marseille (CRCM),
Université Aix-Marseille Institut Paoli Calmettes, France

Future Collaboration Models for the pharma industry

12:15 **The European Lead Factory, the new pan-European drug discovery platform for innovative medicine**

- Presentation on the Innovative Medicine Initiative (IMI) – The EU lead Factory, an "industry-like discovery platform to translate cutting edge academic research into high-quality drug lead molecules.
- Architecture of the EU Lead Factory; EFPIA Library, Public Library, Screening Centre.
- Expected outcome and end goals.
- Interaction between EU Lead Factory and contributing third parties from Public Domain.



Dr. Sylviane Boucharens, Director of Discovery Operations,
'The European Lead Factory', UK

12:40 Networking lunch

14:10 **EU-OPENSREEN: A European Infrastructure of Open Screening Platforms for Chemical Biology**

- Foundation by European academic Chemical Biologists from 14 countries
- Supporting you the development of novel 'tool' compounds (i.e. chemical inhibitors or activators) for target-of-interest
- Vital part of the European Commission Roadmap of research infrastructures
- Transforming your assay into an HTS-ready format – Screening assays against EU-OPENSREEN's unique compound collection & developing initial 'hits' into a valuable (i.e. potent, selective) research 'tool' compound
- Who should apply and how?



Dr. Edgar Specker, Head of Compound Management,
Leibniz Institute of Molecular Pharmacology FMP, Germany

14:45 **Public-Private-Partnerships at National and EU-Level and their potential impact on the discovery of new bioactive compounds**

- Public-private partnerships (PPP) to leverage the creativity of academic research
- Examples for PPP: Neuroalliance and AETIONOMY (IMI project; 8th call; in preparation)
- Strategies for integrative approaches combining systems biology and pharmacological information (the Human Brain Pharmacome)
- Strengths and weaknesses of public-private-partnerships and potential for sustainable collaborative efforts



Prof. Martin Hofmann-Apitius, Head of the Department of Bioinformatics,
Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI, Germany

15:20 Refreshment break & networking

15:50 **Business models for intelligent compound acquisition**

- New business models and new chemical space
- IP transfer of acquired compounds
- Compound library acquisition as part of the discovery partnership
- The Asia factor



Dr. Peter Ten Holte, Global Business Office Liaison,
Janssen Research & Development, LLC, USA

Highlight

Highlight

16:25 **Idea Factory: Round Table Groups - Practical Experience – Join one of the groups and take active part in the discussions:**



Securing IP and patentability of acquired compounds
Peter Ten Holte

New screening approaches

Werngard Czechtizky

CADD and computational methods
Prof. Gisbert Schneider

17:00 Presentation and summary of group discussion

17:15 Closing remarks from the chairman Peter Ten Holte & end of conference day one



Join us for an informal evening get-together!

This is an excellent opportunity for you to meet the other attendees and to make new business contacts in a relaxed atmosphere.

17:30 – 19:30 Evening Workshop A:

New business models for successful compound library acquisition – How do I best spend my compound library budget?

- Subscription models and consortia
- The “Get Compounds for Free Model”
- Integrated models
- Blind screening models
- Membership/partnership
- Exclusive vs. non-exclusive and patentability



Dr. Peter Ten Holte, Global Business Office Liaison,
Janssen Research & Development, LLC, USA

Conference Day 2 | Tuesday, 29 October 2013

08:30 Registration & coffee

08:55 Chairman's welcome & opening address



Werngard Czechtizky, Section Head, Medicinal Chemistry,
Sanofi-Aventis Deutschland GmbH, Germany

09:00 **TrendTracking: Compounds and early-discovery collaborations**



Live-survey to the industry's recent trends. Please prepare your laptops, smartphones or tablets and take part!

Future Collaboration Models for the pharma industry

09:15 **How to develop a mutual beneficial relationship with your CRO and create win-win partnerships?**

- CRO – do not only evaluate what they do, but how can they can truly help us discovering new drugs
- The change in collaborating with CROs – the paradigm shift
- How much knowledge and experience do you need to retain, to intellectually assess and also challenge the CRO's
- What is the role of emerging markets (China, Brazil, India and Russia)?



Dr. Frederik Deroose, CEO,
Asclepia MedChem Solutions, Belgium

09:50 **Charting new bioactive space for drug discovery – CROs as innovation source for pharma and biotech companies**

- An exclusive route towards efficient lead identification: quality beats quantity
- Design and synthesis of Fsp3-enriched scaffolds: leaving flatland behind
- In pursuit of most modern design guidelines: pre-engineering kinetic signatures
- Target family-centric libraries for epigenetic targets and protein-protein interactions: IP-able privileged structures



Dr. Gerhard Müller, Senior Vice President Medicinal Chemistry,
Mercachem b.v., The Netherlands

Highlight

10:25 Refreshment break & networking

New screening experiences and approaches – Optimizing hit-to-lead

11:00 **Detection of specific gaps in screening collections and strategies to fill those via combinatorial library design: e.g. search for potential PPI inhibitors**

- Comparison of screening collections
- Search for interesting property combinations missing in a collection
- Design of combinatorial libraries enriched with compounds showing the desired property profile
- Application example: PPI inhibitors



Dr. Uta Lessel, Principal Scientist,
Boehringer Ingelheim Pharma GmbH & Co. KG, Germany

11:35 **The benefits of targeting optimal ADMET space in Lead Generation and Optimization: Findings based on a systematic analysis of the physicochemical parameters of oral drugs and preclinical compounds**

- The importance of ADMET in drug discovery. Are the lessons being taken on board?
- What are physico-chemical properties influencing almost all ADMET liabilities?
- Are we searching in the optimal region of physical property and biological space?
- How can we use this information to increase the success rate in drug discovery



Dr. M. Paul Gleeson, Lecturer,
Kasetsart University, Thailand

12:10 Networking lunch

13:40 **The role and significance of Polypharmacology for the early drug discovery stage**

- Intro: Polypharmacology can cause adverse effects, but can also confer superior efficacy
- Current trends: How to address polypharmacology in early drug discovery
- How to reduce undesired polypharmacology in the hit-to-lead stage
- How to find leads with a desirable polypharmacological profile



Jens-Uwe Peters, Principal Scientist,
F. Hoffmann-La Roche Ltd., Switzerland

14:15 **Strategy to select a screening library subset with a bias towards bioactivity**

- In-house development of Flexophor, a versatile 3D pharmacophore descriptor
- Selection and validation of a biodiverse screening sub-library based on Flexophor
- Hit success of biodiverse vs chemical diverse sub-library
- Hit success of generic biodiverse sub-library vs customized compound selection



Urs Luethi, Senior Lab Head HTS,
Actelion Pharmaceuticals Ltd, Switzerland

14:50 **Panel discussion: Assessing new methods and approaches for an early and optimized lead generation**



What is the most promising way to enhance hit-to-lead in future? Do we always need to learn it the hard way by using trial and error? Or is there a best of breed solution that colleagues can recommend? Discuss with your peers some of the most talked-about approaches and how they might support next generation lead finding:

- Epigenetics
- Early ADME
- The role of biotech
- Systems biology



Urs Luethi, Senior Lab Head HTS,
Actelion Pharmaceuticals Ltd, Switzerland



Dr. M. Paul Gleeson, Lecturer,
Kasetsart University, Thailand

15:20 Refreshment break & networking

15:50 **The AstraZeneca-Bayer Alliance – Evolution of a Strategic Partnership Between Pharma for a Sustainable Long-term Relationship**

Highlight

- Expansion of a successful collaborative partnership
- Opportunities associated with open innovation
- Fingerprint comparison of large pharma screening collections



Kirsty Rich, High Throughput Screening Team Leader,
AstraZeneca programme director for the AstraZeneca-Bayer alliance,
AstraZeneca plc

Thierry Kogej, Associate Principal Scientist Computational Chemistry,
AstraZeneca Sweden Operations

Co-authors: Bernd Kalthof and Jens Schamberger, **Bayer**

Computational methods for chemistry and biology to predict adverse effects earlier

16:25 **Modulating signalling pathways through the perturbation of protein-protein interactions can achieve better compound validation for cancer therapeutics**

- Targeting signalling pathways
- Small molecules and peptides
- Case studies
- Protein-protein interactions



Dr. Eric Chevet, DR2 INSERM, Team Leader,
Université Bordeaux Ségalen, France

17:00 Closing remarks from the chairman Werngard Czechtizky & end of the conference

17:15 – 19:15 **Workshop B:**

The challenge of reducing compound library size without reducing the quality and success rate for lead generation

- How do we quality design our libraries and new screening campaigns for new targets
- How do we design small molecules that might interact with the protein-protein interaction process
- The quality of the samples, storage & logistics for cost effectiveness
- How do we maximally increase diversity in the screening deck
- What are the strategies of large pharma versus SMEs and academic centres



Dr. Frederik Deroose, CEO,
Asclepia MedChem Solutions, Belgium

For further information

please visit our website www.compound-libraries.com/MM

or contact us on **+49 (0)30 20 91 33 88** or email eq@iqpc.de

Media Partners



Since 72 years, pharmind® has been the periodical to be read by decision-makers and multipliers of the pharma scene. Taking into consideration the national and international (in particular EU and FDA) regulatory environment, the entire range of issues involving the development, manufacture, and marketing/sales of pharmaceutical products is covered in depth. **www.ecv.de**



PharmaVOICE magazine, reaching more than 27,000 BPA-qualified life-sciences executives, is the forum that allows business leaders to engage in a candid dialogue on the challenges and trends impacting the industry. PharmaVOICE provides readers with insightful and thought-provoking commentary in a multiple-perspective format through forums, topics, and articles covering a range of issues from molecule through market. PharmaVOICE subscribers are also kept abreast of the latest trends and information through additional media resources, including WebLinx Interactive WebSeminars, Podcasts, Videocasts, White Papers, E-Surveys and e-Alerts. Additionally, PharmaVOICEMarketplace.com provides a comprehensive directory of products, services, and solutions for the life-sciences industry. To Raise Your VOICE, contact feedback@pharmavoice.com.



www.PharmCast.com is the world leading website designed specifically for pharmaceutical, clinical and biotechnology professionals. **www.PharmCast.com** brings up-to-date information on pharmaceutical patents, FDA, news, jobs and Buyer's Guide to our visitors. It was created and is maintained by pharmaceutical and biotechnology professionals. Visit www.PharmCast.com and discover for yourself why it is so popular among professionals.



Technology Networks has been established for over 10 years and is now the leading provider of free information services for life science professionals. This includes news, events, posters, videos, webcasts, application notes, new products, a literature search engine and product directories. The information is specifically tailored to over 30 individual communities with specific interests, all of which can be accessed through **TechnologyNetworks.com**.

Overcoming the Bottlenecks in Drug Discovery and Development

Thursday 20 March 2014

08:00	Registration
09:00	Inaugural ceremony and welcoming remarks
Session 1 A: Pharmaceutical research Session Chair: Hidemi Minami, <i>RLL, Gurgaon</i>	
09:20 O01	Gut instincts: explorations in intestinal physiology and drug delivery Abdul Basit, <i>University College London, UK</i>
10:00 O02	Chemo-enzymatic route to bicyclic nucleosides and sugar-PEG polymeric architectures or drug delivery applications Ashok K Prasad, <i>University of Delhi, India</i>
10:40	Tea and coffee break
Session 1 B: Pharmaceutical research Session Chair: Mohan Prasad, <i>RLL, Gurgaon</i>	
11:10 O03	Crystal engineering in the pharmaceutical context: some recent results Gautam Desiraju, <i>IISc, Bangalore, India</i>
11:50 O04	Challenges in establishing bioequivalence throughout the product lifecycle Ravi Shanker, <i>Pfizer, Groton, USA</i>
12:30	Lunch and poster session
Session 2 A: Chemical hit to lead optimisation Session Chair: Ramesh Bambal, <i>DSIN, Gurgaon</i>	
13:50 O05	The benefits of targeting optimal ADMET space in lead generation and optimization M Paul Gleeson, <i>Kasetsart University, Bangkok, Thailand</i>
14:30 O06	Applications of matched molecular pair analysis in drug discovery Andrew G Leach, <i>Liverpool John Moores University, Liverpool, UK</i>
15:10	Tea and coffee break
Session 2 B: Chemical hit to lead optimisation Session Chair: Andy Davis, <i>AstraZeneca, Sweden</i>	
15:40 O07	Computational challenges in deciphering drug receptor interactions G Narahari Sastry, <i>CSIR-Indian Institute of Chemical Technology, Hyderabad, India</i>
16:20 O08	Using multifactorial data to improve compound survival in drug discovery William D Pennie, <i>Pfizer, Groton, USA</i>
17:15	Open Forum Discussion on open innovation in drug discovery Session Chair: Alejandra Palermo, <i>Royal Society of Chemistry</i> Panelists: Jeremy Burrows, <i>Medicines for Malaria Venture, Switzerland</i> T K Chandrashekar, <i>Science and Engineering Research Board, India</i> Ahmed Kamal, <i>IICT, Hyderabad</i> Anshu Bhardwaj, <i>Open Source Drug Discovery Unit, India</i> Professor KVSP Rao, <i>Department of Scientific and Industrial Research.</i>
18:30	Close
18:30	Conference dinner

Overcoming the Bottlenecks in Drug Discovery and Development

Friday 21 March 2014

Session 3 A: Chemical hit identification	
Session Chair: Biswajit Das, <i>DSIN, Gurgaon</i>	
09:10 O09	Open data for drug discovery: what is it and how can it be used Anne Hersey, <i>EMBL-EBI, Cambridge, UK</i>
09:50 O10	Building a natural product-inspired, chemical toolbox for protein-protein interactions Prabhat Arya, <i>Dr. Reddy's Institute of Life Sciences, Hyderabad, India</i>
10:30	Tea and coffee break
Session 3 B: Chemical Hit identification	
Session Chair: Simon Ward, <i>University of Sussex, UK</i>	
11:00 O11	Have confidence in your fragments - avoiding the "unknown knowns" of FBLD Ben Davis, <i>Vernalis, Cambridge, UK</i>
11:40 O12	Targeted anticancer drug discovery: from hits to chemical tools and candidates Ian Collins, <i>Institute of Cancer Research, London, UK</i>
Session 3 C: Flash poster presentations	
Session Chairs: Andy Davis, <i>AstraZeneca, Sweden</i> and Simon Ward, <i>University of Sussex, UK</i>	
12:20	Short Oral Presentations
13:10	Lunch and poster session
Session 4 A: Biological target identification	
Session Chair: Dilip Upadhyay, <i>DSIN, Gurgaon</i>	
14:40 O13	Discovery of orally active antitubercular agents through cell-based optimization of HTS hits Fumiaki Yokokawa, <i>Senior Research Investigator, Novartis Institute for Tropical Diseases, Singapore</i>
15:20 O14	Overcoming the bottleneck through traditional knowledge-inspired drug discovery Kalpana Joshi, <i>Sinhgad Institutes, India</i>
16.00	Tea and coffee break
Session 4 B: Biological target identification	
Session Chair: Jitendra Sattigeri, <i>DSIN, Gurgaon</i>	
16.30 O15	A chemical approach to understanding cell division Ulrike Eggert, <i>King's College London, UK</i>
Session 4 C: Case history	
Session Chair: Jitendra Sattigeri, <i>DSIN, Gurgaon</i>	
17:10 O16	Discovery of the oral Factor Xa inhibitor, Edoxaban Masatoshi Nagamochi, <i>Daiichi Sankyo, Tokyo, Japan</i>
17:50	Closing remarks
18:10	Close
