Division of Pharmaceutical Chemistry and Technology
Faculty of Pharmacy, University of Helsinki, Finland

Field-based Proteochemometric Models Derived from 3D Protein Structures: A Novel Approach to Visualize Affinity and Selectivity Features

Vigneshwari Subramanian

ACADEMIC DISSERTATION

To be presented, with permission of the Faculty of Pharmacy of the University of Helsinki, for public examination in Auditorium 1041, Biocenter 2, Viikinkaari 5 on December 21st, 2016 at 12 noon.

Helsinki, 2016

Supervisors Henri Xhaard, PhD
Division of Pharmaceutical Chemistry and Technology
Faculty of Pharmacy
University of Helsinki, Finland

Gerd Wohlfahrt, PhD
Computer-Aided Drug Design
Orion Pharma
Finland

Peteris Prusis, PhD
Computer-Aided Drug Design
Orion Pharma
Finland

Reviewers Aleksejs Kontijevskis, PhD
Nuevolution A/S
Denmark

Outi Salo-Ahen, PhD
Faculty of Science and Engineering, Pharmacy
Åbo Akademi University
Finland

Opponent Olivier Taboureau, PhD
Molécules Thérapeutiques In Silico, Inserm UMR-S 973
Université Paris Diderot, France

ISBN 978-951-51-2812-6 (paperback)
ISBN 978-951-51-2813-3 (PDF)
ISSN 2342-3161 (print)
ISSN 2342-317X (online)
https://ethesis.helsinki.fi

Cover: Field points representing the affinity and selectivity related features of ABL1 kinase with the reference ligand dasatinib.

Abstract

Designing drugs that are selective is crucial in pharmaceutical research to avoid unwanted side effects. To decipher selectivity of drug targets, computational approaches that utilize the sequence and structural information of the protein binding pockets are frequently exploited. In addition to methods that rely only on protein information, quantitative approaches such as proteochemometrics (PCM) use the combination of protein and ligand descriptions to derive quantitative relationships with binding affinity. PCM aims to explain cross-interactions between the different proteins and ligands, hence facilitating our understanding of selectivity.

The main goal of this dissertation is to develop and apply field-based PCM to improve the understanding of relevant molecular interactions through visual illustrations. Field-based description that depends on the 3D structural information of proteins enhances visual interpretability of PCM models relative to the frequently used sequence-based descriptors for proteins. In these field-based PCM studies, knowledge-based fields that explain polarity and lipophilicity of the binding pockets and WaterMap-derived fields that elucidate the positions and energetics of water molecules are used together with the various 2D / 3D ligand descriptors to investigate the selectivity profiles of kinases and serine proteases.

Field-based PCM is first applied to protein kinases, for which designing selective inhibitors has always been a challenge, owing to their highly similar ATP binding pockets. Our studies show that the method could be successfully applied to pinpoint the regions influencing the binding affinity and selectivity of kinases. As an extension of the initial studies conducted on a set of 50

kinases and 80 inhibitors, field-based PCM was used to build classification models on a large dataset (95 kinases and 1572 inhibitors) to distinguish active from inactive ligands. The prediction of the bioactivities of external test set compounds or kinases with accuracies over 80% (Matthews correlation coefficient, MCC: ~0.50) and area under the ROC curve (AUC) above 0.8 together with the visual inspection of the regions promoting activity demonstrates the ability of field-based PCM to generate both predictive and visually interpretable models. Further, the application of this method to serine proteases provides an overview of the sub-pocket specificities, which is crucial for inhibitor design. Additionally, alignment-independent Zernike descriptors derived from fields were used in PCM models to study the influence of protein superimpositions on field comparisons and subsequent PCM modelling.

Tiivistelmä

Lääketutkimuksessa selektiivisten lääkeaineiden suunnittelu on ratkaisevan tärkeää haittavaikutusten välttämiseksi. Kohdeselektiivisyyden selvittämiseen käytetään usein tietokoneavusteisia menetelmiä, jotka hyödyntävät proteiinien sitoutumiskohtien sekvenssi- ja rakennetietoja. Proteiinilähtöisten menetelmien lisäksi kvantitatiiviset menetelmät kuten proteokemometria (proteochemometrics, PCM) yhdistävät sekä proteiinin että ligandin tietoja muodostaessaan kvantitatiivisen suhteen sitoutumisaffiniteettiin. PCM pyrkii selittämään eri proteiinien ja ligandien vuorovaikutuksia ja näin auttaa ymmärtämään selektiivisyyttä.

Väitöstutkimuksen tavoitteena oli kehittää ja hyödyntää kenttäpohjaista proteokemometriaa, joka auttaa ymmärtämään relevantteja molekyylitasoisia vuorovaikutuksia visuaalisen esitystavan kautta. Proteiinin kolmiulotteisesta rakenteesta riippuva kenttäpohjainen kuvaus helpottaa PCM-mallien tulkintaa, etenkin usein käytettyihin sekvenssipohjaisiin kuvauksiin verrattuna. Näissä kenttäpohjaisissa PCM-mallinnuksissa käytettiin tietoperustaisia sitoutumistaskun polaarisuutta ja lipofiilisyyttä kuvaavia kenttiä ja WaterMap-ohjelman tuottamia vesimolekyylien sijaintia ja energiaa havainnollistavia kenttiä yhdessä lukuisten ligandia kuvaavien 2D- ja 3D-deskriptorien kanssa. Malleja sovellettiin kinaasien ja seriiniproteaasien selektiivisyysprofiilien tutkimukseen.

Tutkimuksen ensimmäisessä osassa kenttäpohjaista PCM-mallinnusta sovellettiin proteiinikinaaseihin, joille selektiivisten inhibiittorien suunnittelu on haastavaa samankaltaisten ATP-sitoutumistaskujen takia. Tutkimuksemme osoitti menetelmän soveltuvan kinaasien sitoutumisaffiniteettia ja

selektiivisyyttä ohjaavien alueiden osoittamiseen. Jatkona 50 kinaasia ja 80 inhibiittoria käsittäneelle alkuperäiselle tutkimukselle rakensimme kenttäpohjaisia PCM-luokittelumalleja suuremmalle joukolle kinaaseja (95) ja inhibiittoreita (1572) erotellaksemme aktiiviset ja inaktiiviset ligandit toisistaan. Ulkoisen testiyhdiste- tai testikinaasijoukon bioaktiivisuuksien ennustaminen yli 80 % tarkkuudella (Matthews korrelaatiokerroin, MCC noin 0,50) ja ROC-käyrän alle jäävä ala (AUC) yli 0,8 yhdessä aktiivisuutta tukevien alueiden visuaalisen tarkastelun kanssa osoittivat kenttäpohjaisen PCM:n pystyvän tuottamaan sekä ennustavia että visuaalisesti ymmärrettäviä malleja. Tutkimuksen toisessa osassa metodin soveltaminen seriiniproteaaseihin tuotti yleisnäkemyksen sitoutumistaskun eri osien spesifisyyksistä, mikä on ensiarvoisen tärkeää inhibiittorien suunnittelulle. Lisäksi kentistä johdettuja, proteiinien päällekkäinasettelusta riippumattomia Zernike-deskriptoreita hyödynnettiin PCM-malleissa arvioidaksemme proteiinien päällekkäinasettelun vaikutusta kenttien vertailuun ja sen jälkeiseen PCM-mallinnukseen.

Acknowledgements

My journey to reach the destination of acquiring a Doctorate degree wouldn't have been possible without the efforts and support of many people, who have been thanked in this thesis.

I would like to express my sincere gratitude to my supervisors Dr. Henri Xhaard, Dr. Gerd Wohlfahrt and Dr. Peteris Prusis. I have always been inspired by their passion for science, critical thinking and their attitude towards the accomplishment of various tasks. Scientific questions raised by them during the discussions have been a motivating factor for me to explore more in the field of Computational Drug discovery. Their mentorship during all stages of the project and article writing has motivated me to make progress during my PhD. Apart from the scientific advice, their friendly nature and support to deal with the practical issues have helped me to face even the most challenging situations with ease.

I would like to thank Professor Jari-Yli Kauhaluoma for his constant encouragement and support during my doctoral studies. The experience and knowledge, I gained by working on collaborative projects together with Jari is immense.

My sincere thanks goes to my reviewers Dr. Aleksejs Kontijevskis and Dr. Outi Salo-Ahen for their insightful comments. I respect the time, they have dedicated for my thesis and making it more precise. I also thank Professor Olivier Taboureau for accepting the invitation to be an opponent.

I have had the opportunity to collaborate with many of my colleagues in the Faculty of Pharmacy. It was really a great pleasure to work with all of them, for the scientific input and the friendly atmosphere, they provided. A special thanks goes to the members of 3i project, in particular, Professor Heikki Ruskoaho, Professor Risto Kostiainen, Dr. Gustav Boije af Gennäs, Dr. Virpi Talman,

Mr. Mikael Jumppanen, Ms. Päivi Pöhö and Mr. Jaakko Teppo for the thought-provoking discussions.

I would also like to thank our collaborators Dr. Andreas Bender and Dr. Qurrat Ul Ain for their guidance and critical comments during the Serine protease project work. This collaborative work has been a great learning experience, which helped me to stay tuned with the recent developments in the field of proteochemometrics.

My deepest gratitude goes to all my friends and colleagues in Computational Drug Discovery group. I thank Dr. Aniket Magarkar, Dr. Gloria Wissel, Dr. Alexandre Borrel, Ms. Ainoleena Turku, Mr. Lasse Karhu, Ms. Maiju Rinne, Dr. Leo Ghemtio, Dr. Yue Zhou Zhang, Dr. Michal Stepniewski, Dr. Zia Ur Rehman, Mr. Ashenafi Legehar and Mr. Evgeni Grazhdankin for all the interesting scientific discussions and the fun times, we have had together. I wholeheartedly thank each and every one for their emotional support and advice during the stressful thesis writing period. I have always enjoyed your company during the conferences, monthly beers and annual retreats to sauna and avantouinti. My memories with you in Helsinki are going to stay for the rest of my life.

I would like to extend my thanks to all my colleagues in Orion Pharma. I thank Dr. Lars-Olof Pietilä, Dr. Martti Ovaska, Mr. Julius Sipilä, Dr. Heikki Käsnänen, Dr. Josef Messinger, Dr. Ari Hietanen and Dr. Kai Penttillä for their motivation and guidance during my PhD. A special thanks goes to Lasse for his technical support in writing scripts meant for field calculations. This work wouldn't have been possible without his contributions.

I would like to acknowledge the National Doctoral Programme in Informational and Structural Biology (ISB) and the Doctoral Programme in Integrative Life Sciences (ILS) for organizing the graduate studies and providing travel grants

to attend many international conferences. I am indebted to the Helsinki University Research Foundation, Orion pharma and 3i project (TEKES) for funding my doctoral studies.

Finally, I would like to thank all my friends for their constant encouragement and emotional support. Thanks to my friends in Finland for inviting me to many events and keeping my mind off the work. A special thanks goes to Ms. Elizabeth Cole and Ms. Nusrat Jung for providing me valuable advices at times of need.

Helsinki, December 2016

Vigneshwari Subramanian

Dedicated to my parents, Mr. Subramanian Ramanathan
and Ms. Manonmani Subramanian

Table of Contents

List of Original Publications

This dissertation is based on the following publications, referred to hereafter with Roman numerals.

I. Subramanian, V.; Prusis, P.; Pietilä, L. O.; Xhaard, H.; Wohlfahrt, G. Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics. *J. Chem. Inf. Model.* **2013**, *53*, 3021–3030. doi: 10.1021/ci400369z

II. Subramanian, V.; Prusis, P.; Xhaard, H.; Wohlfahrt, G. Predictive Proteochemometric Models for Kinases Derived from 3D Protein Field-Based Descriptors. *Med. Chem. Commun.* **2016**, *7*, 1007–1015. doi: 10.1039/C5MD00556F

III. Subramanian, V. *; Ain, Q.U.*; Henno, H.; Pietilä, L.O.; Fuchs, J.E.; Prusis, P.; Bender, A.; Wohlfahrt, G. Implementation of 3D Proteochemometrics: Using Three-dimensional Information of Ligands and Proteins to Address the Ligand Selectivity for Serine Proteases *(Manuscript)*

* Equal contribution

Unpublished results

Proteochemometric models based on protein fields and Zernike descriptors: A comparative study

Additional publications

IV. Cortés-Ciriano, I.; Ain, Q. U.; **Subramanian, V.**; Lenselink, E. B.; Méndez-Lucio, O.; IJzerman, A. P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T. E.; van Westen, G. J. P.; Bender, A. Polypharmacology Modelling Using Proteochemometrics (PCM): Recent Methodological Developments, Applications to Target Families, and Future Prospects. *Med. Chem. Commun.* **2015**, *6*, 24–50. doi: 10.1039/C4MD00216D

V. Subramanian, V.; Kuenemann, M.; Ghemtio, L.; Xhaard, H. Difficulties in Retrieving Agonists in GPCR virtual screening *(Manuscript)*

VI. Subramanian, V.; Poho, P.; Teppo, J.; Kostiainen, R.; Xhaard, H. Identification of Metabolites and Pathways through KNIME Workflow Automation *(Manuscript in preparation)*

Abbreviations

2D	two-dimensional
3D	three-dimensional
4-PFP	4-point pharmacophoric fingerprints
AD	Applicability Domain
ATP	Adenosine Triphosphate
AUC	Area Under the Curve
CoMFA	Comparative Molecular Field Analysis
CV	Cross-Validation
DFG	Asp-Phe-Gly
DHFR	Dihydro Folate Reductase
GBM	Gradient Boosting Machine
GFP	Green Fluorescent Protein
GP	Gaussian Process
GPCR	G protein-coupled receptor
GRIND	Grid-INdependent Descriptor
HRD	His-Arg-Asp
IC_{50}	Half-maximal inhibitory concentration
K_d	Dissociation constant
K_i	Inhibition constant
KNIME	Konstanz Information Miner
KNN	K-Nearest Neighbors
LOCCO	Leave One Compound Cluster Out
LOCO	Leave One Compound Out
LOO	Leave One Out
LOTO	Leave One Target Out
MAPK	Mitogen-activated protein kinase

MCC	Matthews Correlation Coefficient
MOE	Molecular Operating Environment
OECD	Organization for Economic Co-operation and Development
PARP	Poly (ADP-Ribose) Polymerase
PC	Principal Component
PCA	Principal Component Analysis
PCM	Proteochemometrics
PDB	Protein Data Bank
pIC_{50}	Negative logarithm of half-maximal inhibitory concentration
pK_d	Negative logarithm of dissociation constant
pK_i	Negative logarithm of inhibition constant
PLS	Partial Least Squares Regression
QSAR	Quantitative Structure-Activity Relationship
RF	Random Forests
RMSEE	Root Mean Square Error of Estimation
RMSEP	Root Mean Square Error of Prediction
ROC	Receiver Operating Characteristic
SDF	Structure Data Format
SMILES	Simplified Molecular Input Line Entry System
SVM	Support Vector Machine
UV	Unit Variance

Introduction

Pharmaceutical companies often aim at "magic bullet" drug molecules that act exclusively on a singe target, thus minimizing side effects. However, complexity of the biological systems poses a major challenge for designing selective molecules. Studies conducted to analyse drug-target interaction networks have revealed that the majority of the drugs available on the market have the potential to bind to multiple targets (Mestres et al., 2008) and are likely to induce unwanted side effects (Smith et al., 2006; Peters, 2013). The low selectivity profiles of drug molecules highlight the importance of understanding the polypharmacological profiles in the initial stages of drug design. Proteochemometrics (PCM) (Prusis et al., 2001), a statistical modelling approach that aims to quantify the structure-activity relationships by considering both protein and ligand features, has been developed to study polypharmacology. PCM has been used in this dissertation with the aim of elucidating the features that promote selectivity and identifying the potential off-targets (proteins not intended as targets) across a protein family.

This dissertation was conducted as a joint collaborative project between the Division of Pharmaceutical Chemistry and Technology at the University of Helsinki and the Computer-Aided Drug Design group of Orion Pharma. The study mainly focuses on the development of field-based PCM approaches and their application to understand the selectivity profiles of ligands that bind to kinases and serine proteases. The methods developed not only help to predict off-targets, but also provide an illustration of the regions in protein binding sites and ligands that are likely to influence binding affinity and selectivity.

In the initial stages of the project, field-based PCM relying on the fields derived from 3D protein structures was established. The concept of using 3D structural information in PCM mainly evolved from the application of molecular interaction fields to decipher ligand selectivity (Hoppe et al., 2006; Wohlfahrt et al., 2009). Field-based PCM was developed to overcome the drawbacks of sequence-based PCM. Models that utilize the amino acid sequence descriptor information fail to consider the spatial arrangement of amino acids in the binding site and have limited visual interpretability. Nevertheless, the orientation or directionality of amino acids is critical for ligand design. Therefore, building visually interpretable PCM models that take advantage of the highly informative protein field descriptors could benefit the day-to-day inhibitor design.

The suitability of employing knowledge-based and WaterMap fields derived from protein binding sites in PCM modelling was first tested on kinases, resulting in Publications I and II. In Publication I, field-based PCM models with a set of 50 kinases and 80 inhibitors were built using the linear Partial Least Squares (PLS) regression approach to enable easy interpretation and visualization of the kinase and ligand features relevant for selectivity. In Publication II, the kinase dataset was expanded to build a global field-based PCM model using non-linear machine learning approaches to classify active and inactive ligands. Further, the applicability of the field-based methods to other protein families was demonstrated by building PCM models with serine proteases (Publication III).

Field-based methods are sensitive to shifts in protein superimposition. Building PCM models based on the fields calculated from poorly aligned structures

could have a significant impact on the model quality. To assess the influence of alignment errors in PCM modelling, a systematic comparison was made between the models built on protein fields of superimposed proteins and alignment-independent Zernike descriptors (Novotni and Klein, 2003) (unpublished results). Models based on superimposed protein fields and Zernike descriptors have virtually the same performance. For the kinase dataset, Matthews correlation coefficient for field-based models is 0.52 and for Zernike descriptor based models 0.48. Only preliminary results were obtained at the time of writing the thesis.

Review of the literature

The target families used for PCM modelling and overviews of computational approaches used for selectivity design are presented in this section.

1 Target families

1.1 Kinases

1.1.1 Kinases as enzymes

Protein kinases are one of the largest families of drug targets, encoded by more than 518 genes in humans (Manning et al., 2002). Kinases are enzymes involved in phosphorylation, a post-translational modification that catalyses the transfer of a phosphate group from adenosine triphosphate (ATP) to the substrate proteins (Zuccotto et al., 2010). For a kinase to promote catalysis of phosphate transfer, it should remain in the activated state, which can be triggered by various signalling events (Johnson, 2007). Following phosphorylation, the substrates can alter the recognition properties of enzymes through conformational changes. The best example of this phenomenon is the sequential activation of kinases in the Mitogen-activated protein kinase (MAPK) pathway, where phosphorylation of MAP3K activates MAP2K, which in turn triggers MAPK activity (Cargnello and Roux, 2011). Kinases play a key role in mediating various cellular, metabolic and signalling pathways via phosphorylation (Johnson, 2007; Melnikova and Golden, 2008). Among the substrates phosphorylated by kinases are the G protein-coupled receptors (GPCRs) that participate in a wide array of signal transduction pathways. The phosphorylated GPCRs can in turn initiate other signalling events and cellular responses (Tobin, 2008).

1.1.2 Kinase families

All eukaryotic kinases share a catalytic domain that is homologous among the kinase subfamilies (Hanks et al., 2013). Apart from the kinase domain, there are 83 additional domains observed in 258 kinases (Manning et al., 2002). These additional domains are unique to certain groups of kinases. They mediate various signalling activities and control the phosphorylation states of the kinases that act in a network (Manning et al., 2002). Based on phylogenetic classification, kinases can be categorized into 11 major groups, which are further classified into several families and subfamilies (http://kinase.com/wiki/index.php/Kinase_classification; Manning et al., 2002).

1. AGC – Adenine Guanine Cytosine kinase
2. CAMK – Calcium / Calmodulin-dependent protein Kinase
3. CK I – Casein Kinase I
4. CMGC (CDK, MAPK, GSK and CLK families)
5. STE – Serine / Threonine Protein kinase
6. TK – Tyrosine Kinase
7. TKL – Tyrosine Kinase-Like
8. RGC – Receptor Guanylate Cyclases
9. OPK – Other Protein kinases
10. PKL – Protein kinase-like / Pseudokinases
11. Atypical kinases

1.1.3 Structural organization of kinases

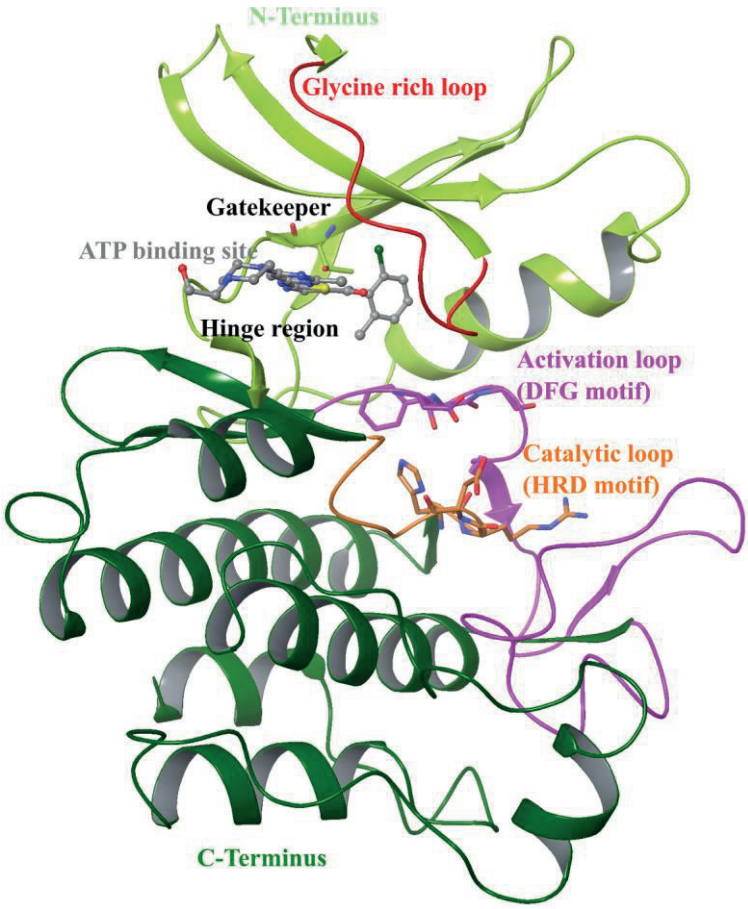

Figure 1. Structural organization of kinase domains. The crystal structure of ABL1 kinase with the bound inhibitor dasatinib (PDB: 2GQG) is shown here. The kinase structure is represented as a cartoon and the ligand is shown in ball and stick style. Different colours in the figure correspond to different structural elements. The gatekeeper residue, DFG (Asp-Phe-Gly) motif and HRD (His-Arg-Asp) motif are shown as thin tubes.

The kinase domain (Figure 1) includes two lobes, namely the N-terminal lobe with five β strands and one α helix and the C-terminal lobe mainly with α

helices (Huse and Kuriyan, 2002). The ATP molecule binds in a cleft located between the two lobes and participates in hydrogen bonding interactions with the residues in the hinge region (Zuccotto et al., 2010). A glycine-rich loop present in the N-terminal lobe facilitates phosphoryl transfer by interacting with ATP. The catalytic loop found in the C-terminal lobe contains the highly conserved HRD (His-Arg-Asp) motif that stabilizes the conformation of the activation loop (Kornev et al., 2006). The activation loop that is centrally located and contains about 20-30 residues controls the activation states of the kinases (Huse and Kuriyan, 2002; Kornev et al., 2006). The orientation of the DFG (Asp-Phe-Gly) motif present in the activation loop serves as a deciding factor for whether the kinases are in active or inactive state (Kornev et al., 2006).

1.1.4 Active and inactive forms of kinases

Kinases are known to exist in two different conformations, the DFG-in / active and DFG-out / inactive conformation. In the active form, the phenylalanine of the DFG motif points away from the ATP binding pocket (Figure 2, left panel) and one of the residues, serine, threonine or tyrosine, frequently remains phosphorylated in the activation loop (Huse and Kuriyan, 2002; Kornev et al., 2006). The phosphorylation states of these residues can either be influenced by additional domains present in kinases or by autophosphorylation, which occurs due to receptor dimerization (Nolen et al., 2004; Johnson, 2009). Nevertheless, phosphorylation is not always necessary for kinase activation. There are examples of kinases that can be activated without phosphorylation, e.g. CDKs that are activated by cyclins (Nolen et al., 2004). In case of inactive conformation, the phenylalanine of the DFG motif is oriented towards the ATP binding site (Figure 2, right panel), which can sometimes block the ATP

binding pocket and direct the ligands towards allosteric pocket (Nolen et al., 2004; Kornev et al., 2006).

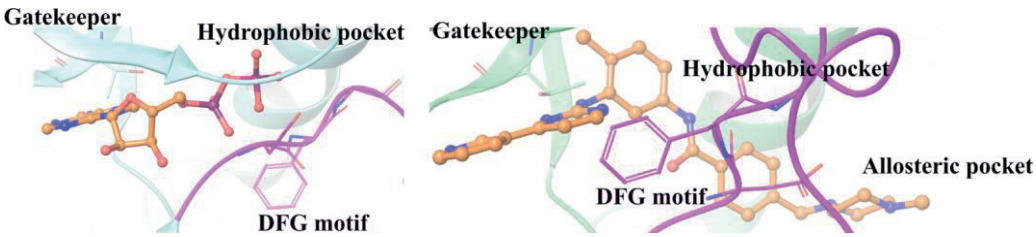

Figure 2. Detailed view of the different activation states of kinases. Left panel corresponds to the DFG-in conformation of KIT kinase with bound adenosine diphosphate (PDB: 1PKG). Right panel represents the DFG-out conformation of KIT kinase with the bound inhibitor Imatinib (PDB: 1T46). The X-ray ligands are shown in orange and activation loops are shown in magenta in both conformations.

1.1.5 Kinases and associated diseases

Genetic abnormalities can lead to aberrant activation of kinases (Tsatsanis and Spandidos, 2000; Fleuren et al., 2016) that are involved in modulating the cell growth and differentiation processes. Their atypical activation can have a significant effect on the phosphorylation patterns of the substrates, which might lead to uncontrolled proliferation of cells, resulting in cancer (Zhang et al., 2009; Fleuren et al., 2016). Therefore, kinases are the promising drug targets for cancer, alongside heart diseases, diabetes and inflammatory disorders (Melnikova and Golden, 2008; Fleuren et al., 2016).

1.1.6 Kinase inhibitors: an overview

Of the 33 clinically approved kinase inhibitors currently available on the market (http://www.brimr.org/PKI/PKIs.htm), many are known to interact with multiple kinases and are likely to induce side effects. The vast majority of these

inhibitors target the ATP binding site of kinases, whose high similarities often pose a major challenge for designing selective inhibitors (Huang et al., 2010). Another known issue in kinase inhibitor design is the drug resistance arising from gene amplifications and mutations in kinase domains, which frequently limits the efficacy of kinase inhibitors (Bikker et al., 2009). A typical example for drug resistance is the ineffectiveness of imatinib in targeting BCR-ABL1 kinase, which has either a mutated gatekeeper residue (T315I), resulting in unfavourable interactions, or includes multiple copies of the gene, leading to reactivation of signalling pathways (Gorre et al., 2001; Daub et al., 2004; Barouch-Bentov, 2012).

1.1.7 Selectivity of kinase inhibitors

Generally, type I inhibitors compete with ATP and bind to rigid active conformations (Liu and Gray, 2006; Muller et al., 2015). These inhibitors mainly interact with the residues in hinge region by forming hydrogen bonds (Zuccotto et al., 2010). The hinge interactions are well conserved among the kinase families. Therefore, the gatekeeper residue located close to the hinge region acts as a determinant for the selectivity of type I inhibitors by influencing the size/accessibility of the hydrophobic back pockets (Zuccotto et al., 2010, Muller et al., 2015). The presence of smaller and medium gatekeeper residues such as Leu, Val, Thr and Met offers more room in the back cavity, thereby promoting the efficacy and selectivity of kinase inhibitors. On the other hand, bulky residues such as Phe and Tyr often contribute to steric hindrance and restrict inhibitor binding (Zuccotto et al., 2010). An example of a highly potent inhibitor, where the small size of the gatekeeper residue is exploited for selectivity, is skepinone-L, which targets the p38α/β MAPK kinase (Koeberle et al., 2011).

Type II inhibitors bind to inactive conformations that have dynamic and extended binding pockets (Liu and Gray, 2006; Muller et al., 2015). These inhibitors also participate in hinge interactions that mimic ATP. Additionally, they interact with the hydrophobic back pockets, enhancing their selectivity (Muller et al., 2015). BCR-ABL inhibitors imatinib and nilotinib serve as examples for achieving selectivity through interactions with the hydrophobic back pocket (Barouch-Bentov, 2012; Muller et al., 2015).

Non-ATP competitive inhibitors known as type III inhibitors bind to the allosteric pocket that is adjacent to the hydrophobic pocket in DFG-out kinases (Muller et al., 2015). Allosteric pockets are not well conserved across kinases (Treiber and Shah, 2013) and this contributes to the high selectivity of type III inhibitors. An example of an allosteric inhibitor that adopts type III binding mode is trametinib, which targets MEK1 and MEK2 kinases (Zhao et al., 2014).

Recently, the cavity between the folded p-loop / glycine-rich loop and helix αC has been targeted for designing the selective inhibitor SCH772984 against ERK1/2 kinase. This selectivity strategy could be exploited for kinases that have folded p-loop conformations (Muller et al., 2015).

1.2 Serine proteases

1.2.1 Proteases as enzymes

More than 550 genes in humans encode proteases (Puente et al., 2003; Cera et al., 2009). Proteases are enzymes that catalyse the cleavage of peptide bonds of their substrates (Cera et al., 2009). Information concerning all proteases and their probable cleavage sites is included in the MEROPS database (http://merops.sanger.ac.uk), a comprehensive resource (Rawlings, 2016). Proteases have a pivotal role in regulating various physiological processes including cell-cycle regulation, digestion, blood coagulation, wound healing and immune response (Hedstrom et al., 2002; Cera et al., 2009). They also act as key regulators of signalling pathways, as proteolysis (digestion of peptides to amino acids) can modulate the activities of many kinases and transcription factors (Ehrmann and Clausen, 2004).

1.2.2 Protease families

Proteases are broadly classified into 7 families, depending on the catalytic residues or metal ions involved in protein degradation: threonine proteases, aspartic proteases, serine proteases, cysteine proteases, metalloproteases, glutamic acid proteases and asparagine peptide lyases (Oda, 2012). Nearly one-third of the human proteases belong to the serine protease family (Puente et al., 2003; Cera et al., 2009). Serine proteases can further be classified into several subfamilies based on their substrate specificities. Trypsin-like and chymotrypsin-like proteases that have roles in digestion, thrombin-like proteases that are involved in blood coagulation and elastase-like proteases that trigger immune response comprise the serine proteases found in eukaryotes. Another class of serine proteases known as subtilisin-like proteases is mainly found in prokaryotes and it differs from other subfamilies based on the

arrangement of catalytic residues in protein scaffolds (α/β in subtilisin and β/β in other subfamilies) (Siezen and Leunissen, 1997). Only the eukaryotic serine proteases are discussed in detail in the following sections.

1.2.3 Structural organization of serine proteases

Serine proteases have two six-stranded β barrels with the active site enclosed between the two (Hedstrom et al., 2002) (Figure 3). A catalytic triad formed by aspartate, serine and histidine residues in the active site acts as a charge relay system (Hedstrom et al., 2002; Cera et al., 2009). These catalytic triads form a part of the extensive intramolecular hydrogen-bonding network and regulate the activities of serine proteases (Hedstrom et al., 2002). Another component linked to the catalytic triad is the oxyanion hole, a pocket formed between the positively charged backbone NH groups of Gly193 and Ser195 residues in the active site and the negatively charged carbonyl groups of substrate peptides. The oxyanion hole is involved in the stabilization of transition state intermediates formed during peptide hydrolysis (Hedstrom et al., 2002). Further, the active site is divided into sub-pockets characterized by specific amino acid residues that cleave different regions of the substrate peptide (Figure 4).

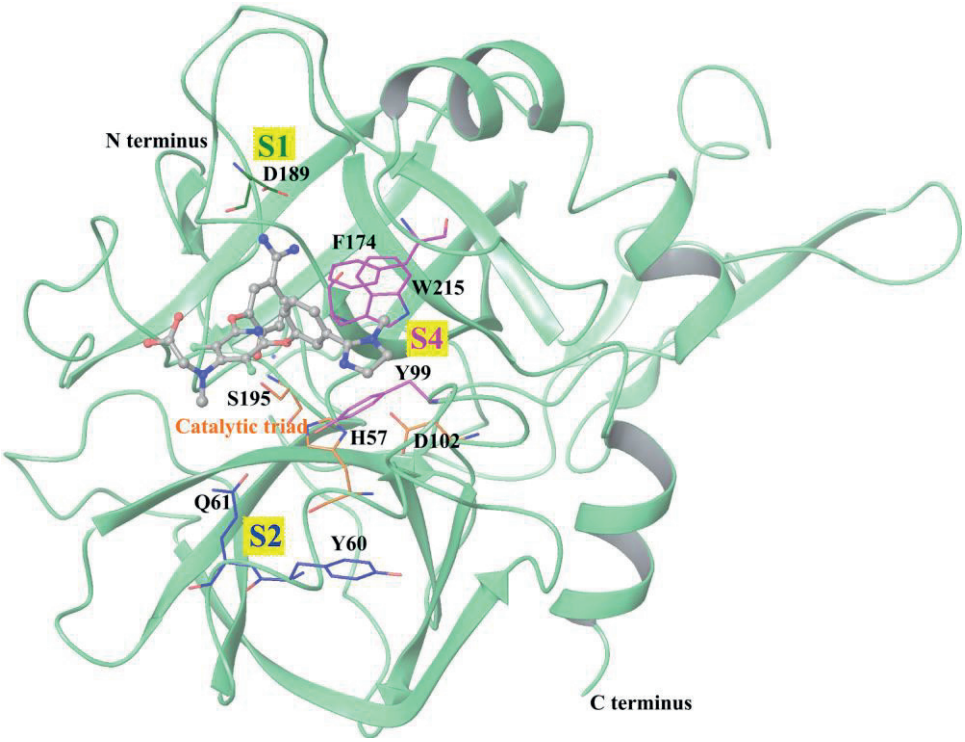

Figure 3. Structural organization of serine proteases. The crystal structure of coagulation factor Xa with the bound inhibitor ZK-807834 (PDB: 1FJS) is shown here. The protease structure is represented as a cartoon and the ligand is shown in ball and stick style. The catalytic triad Asp102-His57-Ser195 is shown in orange as thin tubes. S1, S2 and S4 correspond to the different sub-pockets. The residues in S1 (Asp189), S2 (Tyr60, Gln61) and S4 (Tyr99, Phe174, Trp215) are shown as thin tubes, highlighted in green, blue and magenta, respectively.

1.2.4 Activation of serine proteases

Serine proteases or in general proteases exist as inactive precursors termed 'zymogens' to avoid unwanted protein degradation. These zymogens have a distorted active site and are converted into active enzymes upon initiation of peptide-bond cleavage in the N-terminus region (Neurath et al., 1967; Khan and James, 1998). Zymogen activation can also be influenced by a drop in pH

levels, an autocatalytic mechanism (Khan and James, 1998). Examples of zymogens that are later converted to active serine proteases include chymotrypsinogen, trypsinogen and proelastase. Upon activation, they are converted into chymotrypsin, trypsin and elastase, respectively (Khan and James, 1998).

1.2.5 Sub-pocket specificity of serine proteases

The interactions occurring at the protein-protein interface between substrates and enzymes act as the determinants of sub-pocket specificity. The active sites of the proteases that bind to the substrate and catalyse the proteolysis reaction are usually represented as Sn…S4-S3-S2-S1—S1'-S2'-S3'-S4'…. Sn' from N terminus to C terminus (Schechter, 2012). The corresponding substrate peptide residues on which these proteases act are denoted as Pn…P4-P3-P2-P1—P1'-P2'-P3'-P4'…. Pn'. The cleavage occurs in the region between prime and non-prime sites (P1'—P1) (Figure 4). The cleavage sites are unique for specific proteases.

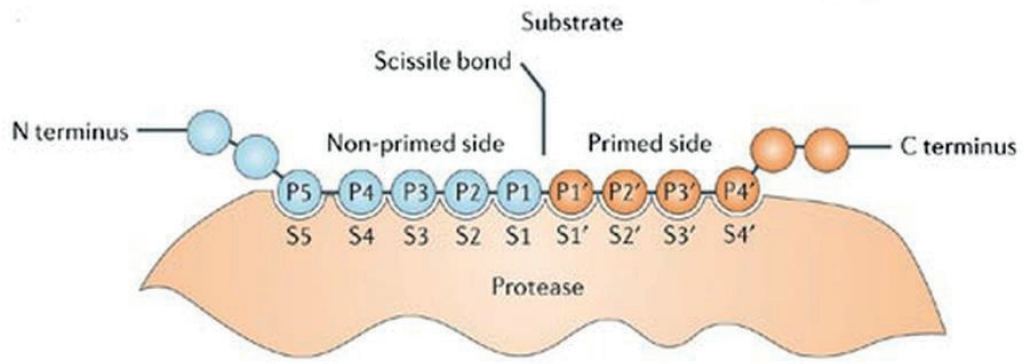

Figure 4. Illustration of sub-pockets and peptide sites of proteases. *Reproduced in part by permission from Macmillan Publishers Ltd: Nature Reviews Drug Discovery (Turk, B. Targeting Proteases: Successes, Failures and Future Prospects. Nat. Rev. Drug Discov.* ***2006****, 5 (9), 785–799.), copyright (2006).*

Generally, analysis of sub-pocket specificities in proteases is focused on S1/P1 interactions (Hedstrom et al., 2002). Considering the S1 specificities of serine protease subfamilies, the trypsin-like and thrombin-like proteases have a negatively charged Asp, which prefers substrates with positively charged Lys / Arg at P1. On the other hand, the hydrophobic Phe and Val residues of chymotrypsin-like and elastase-like proteases have a preference for substrates that contain aromatic or small aliphatic residues at P1. S1-S4 sub-pocket specificities of the four serine protease subfamilies are shown in Figure 5.

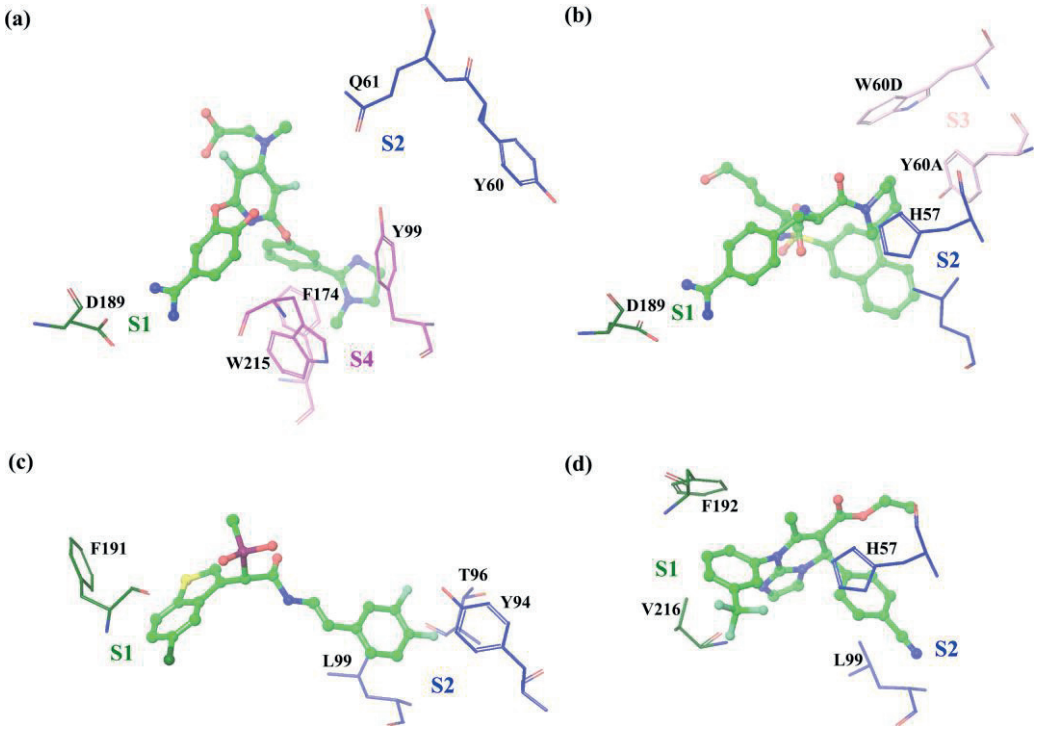

Figure 5. Sub-pocket specificities of serine protease families. The ligands are shown in green with the ball and stick style. Key residues in the different sub pockets S1 (green), S2 (blue) and S4 (magenta) are shown as thin tubes. **(a)** Trypsin-like (PDB: 2XBW) **(b)** Thrombin-like (PDB: 1QUR) **(c)** Chymotrypsin-like (PDB: 3N7O) **(d)** Elastase-like (PDB: 5A8X)

1.2.6 Proteases and associated diseases

Diminished proteolysis or excessive proteolysis resulting from genetic irregularities affects many signalling pathways that have predominant roles in causing cancer, inflammation, cardiovascular diseases and viral infection (Drag and Salvesan, 2010). The role of proteases in regulating a multitude of biological processes makes them promising drug targets.

1.2.7 Protease inhibitors: an overview

Thirty-two protease inhibitors targeting various protease classes, such as metallo (14), aspartic (8), serine (9) and threonine (1), have been approved so far to treat hypertension, thrombosis, respiratory diseases, pancreatitis and cancer (Turk, 2006). Protease inhibitors can be small molecules (e.g. Angiotensin Converting Enzyme (ACE) inhibitor, captopril) or peptides (e.g. Factor X inhibitor, bivalirudin) or peptidomimetics (e.g. HIV protease inhibitor, saquinavir). A majority of the protease inhibitors currently available on the market target ACE, which regulates blood pressure (Turk, 2006).

Based on their mechanism of action, peptidic inhibitors targeting serine proteases can be grouped into three categories, namely canonical, non-canonical and serpins.

Canonical inhibitors are reversible protein inhibitors whose binding is influenced by the presence of a protease-binding loop that remains complementary to the active site (Krowarsch et al., 2003). Their interactions mimic enzyme-substrate complexes (Turk, 2006). Hirustasin, which inhibits trypsin, chymotrypsin and kallikrein, is an example of a canonical inhibitor.

Non-canonical inhibitors are peptides that bind through their N-terminus to the active site, forming a strong parallel β sheet, which is further strengthened by additional interactions with the regions outside the active site (Krowarsch et al., 2003). Hirudin, a natural peptide inhibitor, interacts with the thrombin active site in a similar fashion.

Serpins (serine proteinase inhibitors) are globular proteins that act as natural irreversible inhibitors of serine proteases (Gettins, 2002). Serpins tend to adopt multiple conformations and modulate the activity of serine proteases through a complete blockage of the active sites (Janciauskiene, 2001). Typical examples of serpin inhibitors include Alpha-1 antitrypsin, which acts on neutrophil elastases, and antithrombin, which regulates various coagulation factors. Mutations in serpins can affect their inhibitory properties, leading to several disease states, including inflammation, bleeding disorders and neurodegenerative diseases (Janciauskiene, 2001).

Besides peptidic inhibitors, there are many competitive small molecule inhibitors. Some synthetic covalent inhibitors such as halomethyl ketones and β-lactams bind irreversibly to serine proteases (Sanderson, 1999; Powers et al., 2002). The irreversible inhibitors are usually not favoured owing to selectivity issues arising from their tendency to block many proteases (Turk, 2006). Therefore, designing reversible inhibitors that resemble the transition state intermediates of substrate hydrolysis is an ideal strategy in the protease field (Turk, 2006; Drag and Salvesan, 2010). Examples of reversible inhibitors include non-covalent thrombin inhibitors such as argatroban and napsagatran (Sanderson, 1999).

Selective targeting of proteases can be achieved by designing inhibitors that bind non-competitively through exosite and allosteric interactions. Exosite inhibitors have direct effects on the active site and influence the catalytic rates by binding to secondary sites that remain far from the active site (Turk, 2006; Drag and Salvesan, 2010). The thrombin inhibitor desirudin, which binds to the fibrinogen-binding site, is an example of an exosite inhibitor (Warkentin, 2004). Allosteric inhibitors have indirect effects on substrate recognition by inducing conformational changes in the enzymes and are highly selective. The designed ankyrin repeat proteins (motifs with 33 residues) that target caspase-2 protease serve as an example of allosteric inhibition (Drag and Salvesan, 2010).

2 Selectivity of drug targets

The term ‘selectivity’ refers to the potential of a ligand to bind to a specific drug target or affect a particular cell population (Mecher et al., 2005). A drug that hits many targets and pathways besides the desired ones can have harmful side effects. As of September 2015, altogether 270 drugs have been withdrawn from the market due to adverse effects resulting from their binding to off-targets (http://cheminfo.charite.de/withdrawn). Non-steroidal anti-inflammatory drugs like rofecoxib and valdecoxib, which have the potential to increase the risk of heart attack and stroke, serve as typical examples of drug withdrawals that are initiated for safety reasons (Qureshi et al., 2011). It is also very common that candidate drugs have to be pulled from clinical trials because of adverse drug reactions (Kola and Landis, 2004; Peters, 2013). Since drug development is expensive and time-consuming, screening for potential off-targets and analysing the selectivity profiles of ligands in the early stages of drug design are likely to reduce failure rates at a later point (Peters, 2013).

Nevertheless, selectivity should not be gained at the expense of efficacy (Mencher and Wang, 2005). It is highly probable that the effectiveness of a drug might be reduced by directing it to a single target. Disease states are often influenced by multiple targets or in many cases the involvement of biological pathways rather than individual targets (Mencher and Wang, 2005; Mestres et al., 2008). Therefore, considering selectivity in a broader sense by designing promiscuous ligands that have targeted polypharmacology by acting on targets associated with specific biochemical pathways would help to establish a balance between efficacy and side effects induced by non-specific binding (Peters, 2013). An example for targeted polypharmacology is the kinase inhibitor sorafenib, which is highly effective in controlling tumor progression

and angiogenesis by acting on vascular endothelial growth factor receptor (VEGFR) and platelet-derived growth factor receptor (PDGFR). Apart from being efficacious, its limited side effects make sorafenib a promising inhibitor for treating renal cell and hepatocellular carcinomas (Adnane et al., 2005).

2.1 Computational approaches to address selectivity

Pharmaceutical companies usually conduct safety screening against a panel of targets to test for potential off-target effects (Peters, 2013). Despite the availability of a multitude of experimental approaches (Graczyk, 2007; Karaman et al., 2008; Cheng et al., 2010), conducting an exhaustive screening is often not possible. Computational approaches that consider the binding characteristics, such as shape, electrostatics, flexibility, hydration and allostery are frequently exploited to understand selectivity across protein families, taking advantage of the target's structural information (Huggins et al., 2012). Commonly used computational methods for selectivity design are described in the following section, with a specific focus on examples related to kinases and proteases.

2.1.1 Shape complementarity

Designing a ligand whose shape remains complementary to the binding pocket helps to gain selectivity by optimizing the interactions with the binding site residues (Huggins et al., 2012). Shape complementarity can be analysed through ligand-based approaches such as Phase Shape (Sastry et al., 2011). Examples of achieving selectivity through shape complementarity include the design of ROCKI (Rho-associated protein kinase) inhibitors, whose ATP binding site shape is influenced by the unique arrangement of five key residues not found in other kinases (Breitenlechner et al., 2003), and the development of

HIV protease inhibitors, in which the mutant I84V affects the binding site shape, and hence, the affinity of the inhibitors (Kovalevsky et al., 2006).

2.1.2 Charge complementarity

Charge complementarity is a key concept in molecular recognition. Charged ligands form salt bridges (combination of electrostatic and hydrogen bonding interactions), which enhance their selectivity profiles against lipophilic ligands (Huggins et al., 2012). Variations in charges across the binding pockets can be analysed through the calculation of molecular electrostatic potentials with the help of software packages such as Adaptive Poisson Boltzmann Solver (Baker et al., 2001). Differences in the electrostatics of the S1 sub-pocket of factor Xa (Gln192) and thrombin (Glu192) have been exploited to design a selective inhibitor DX-9065, which is ~20 times more potent on factor Xa than thrombin (Pinto et al., 2010).

2.1.3 Conformational flexibility

Accounting for conformational flexibilities through molecular dynamics simulations distinguishes the desired target from an off-target, thereby improving selectivity (Huggins et al., 2012). This is true for kinases that switch between DFG-in and DFG-out conformations based on the movements of the activation loop. The tendency to adopt the DFG-out conformations is not observed in many kinases, which provides an opportunity for designing inhibitors that selectively target DFG-out states (Huggins et al., 2012). Typical examples include imatinib and BIRB796, which inhibit the DFG-out states of ABL and p38 MAP kinase, respectively.

2.1.4 Water molecules in the binding site

Position and energetics of water molecules in the binding site, analysed through approaches such as WaterMap, have been shown to influence selectivity profiles of drug molecules (Abet et al., 2008; Robinson et al., 2010; Beuming et al., 2012). A ligand that binds by displacing unfavourable water molecules in the target can have many fold higher affinity than an off-target (Huggins et al., 2012). Analysing the thermodynamic properties of water molecules in the factor Xa binding site revealed that displacing the entropically structured water molecules in the S4 sub-pocket enhanced ligand binding by contributing to energetically favourable interactions with the hydrophobic residues of the S4 sub-pocket (Abel et al., 2008). In another study, the presence of high-energy hydration sites displaced during ligand binding in Src kinase has been suggested to increase the binding affinity by 15-fold relative to GSK-3β, which lacked this hydration site (Robinson et al., 2010).

2.1.5 Allosteric binding

Targeting binding sites other than the primary active sites could enhance selectivity with respect to off-targets (Huggins et al., 2012). Although computational approaches such as molecular dynamics can support the identification of allosteric sites, experimental confirmation is required in most cases. Kinase inhibitors that bind to the allosteric pocket of DFG-out conformations have been shown to be selective, compared with the DFG-in inhibitors targeting ATP binding sites (Zuccotto et al., 2010; Treiber and Shah, 2013).

2.2 Systematic comparison of protein binding sites: a strategy to elucidate ligand selectivity

Comparing proteins based on binding site properties extracted from sequence and structure information has been shown to be valuable for understanding the selectivity of ligands. Selected studies focusing on the comparison of binding sites in kinases and serine proteases are presented here.

2.2.1 Kinases

- Identification of energetically favourable binding site residues that influence a specific kinase-ligand interaction generates a binding site signature. Subsequent mapping of these signatures to the multiple sequence alignment of all protein kinases could recognize potential off-targets (Sheinerman et al., 2005).
- Hierarchical clustering of 75 kinases utilizing knowledge-based interaction fields derived from polar and lipophilic probes grouped the kinases distinctly based on their ligand binding modes and different conformations of the activation loops. The possibility to compute similarity and difference fields provides a way to visualize the regions that can be exploited for selectivity design (Hoppe et al., 2006).
- FLAP (Fingerprints for Ligands and Proteins) approach allows exploration of the protein-ligand interaction space by defining 4-point pharmacophoric fingerprints based on molecular interaction fields for proteins and the features complementary to the binding site for ligands. FLAP analysis that accounts for shape complementarity and flexibility when applied to kinases could distinguish the similarities and differences among binding sites and contribute to selective inhibitor design (Baroni et al., 2007).

- Binding site similarity analysis based on a geometric hash approach, which accounts for atom-atom similarity, and CavBase, which characterizes the properties of protein binding sites, explained probable cross-relationships among kinase subfamilies (Kuhn et al., 2007; Kinnings and Jackson, 2009).
- Pharmacophoric fingerprints extracted from Cα atoms in the binding cavity classified the ATP binding sites of 522 kinases with an AUC (Area Under the ROC Curve; For details, see Methods) of 0.89. The distinct classifications of kinase sub-groups generated by this alignment-free approach could provide a way to analyse the ligand binding preferences of various kinase families (Figure 6) (Weill and Rognan, 2010).
- Alignment-independent Zernike descriptors computed from DrugScore potential fields enabled identification of distant kinases that are likely to be hit by similar ligands, thereby providing a way to predict off-targets and hence selectivity (Nisius and Gohlke, 2012).
- Exploration of the key binding site interactions from knowledge-rich databases like KLIFS (Kinase-Ligand Interaction Fingerprints) could provide insight into the affinity and selectivity promoting regions of kinase families and subfamilies (http://klifs.vu-compmedchem.nl; Van Linden et al., 2014).
- Knowledge on differences in ATP binding pockets computed from multiple target and off-target structures, using a grid-based pocket detection approach, enables visualization of kinase sub-pockets relevant for designing selective inhibitors (Volkamer et al., 2016).

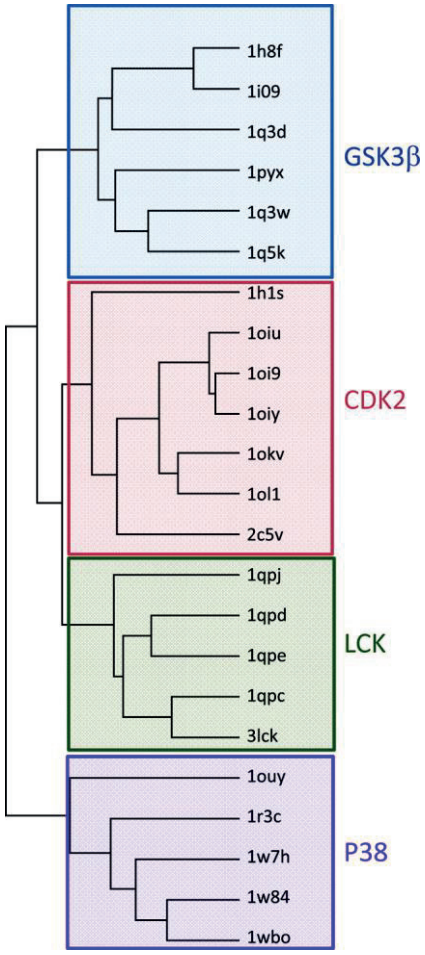

Figure 6. ATP binding sites of multiple PDB structures representing 4 kinase subtypes, clustered based on the pharmacophoric fingerprints of binding cavities. *Figure reproduced with permission from (Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. J. Chem. Inf. Model. **2010**, 50 (1), 123–135). Copyright (2010) American Chemical Society.*

2.2.2 Serine proteases

- Clustering the water sites found in the crystal structures of thrombin and trypsin revealed the presence of 22 conserved water sites in thrombin that contribute to substrate specificity and could be readily exploited for thrombin inhibitor design (Sanschagrin et al., 1998).
- GRID molecular interaction fields of 3D protein structures computed with ten different probes and subsequent principal component analysis identified the selectivity promoting regions of factor Xa, trypsin and thrombin (Kastenholz et al., 2000).
- Hierarchical clustering of knowledge-based interaction fields derived with polar and lipophilic probes highlighted differences in ligand binding specificities for trypsin, thrombin and factor Xa (Hoppe et al., 2005).
- Cluster analysis of the serine protease families based on the properties of binding cavities detected by CavBase explained cross-reactivity (Figure 7) (Glinca and Klebe, 2013).
- Cα distance calculations of amino acid residues and subsequent multivariate analysis identified the deviations in distances of sub-pocket residues between the coagulation factors (II, VII, IX, X and XI) (Uzelac et al., 2015).
- Ensemble clustering of the propagated motifs that lie in close proximity to the binding cavity captured the conformational flexibilities of binding sites and classified the serine protease families correctly based on their ligand binding preferences (Guo and Chen, 2015).

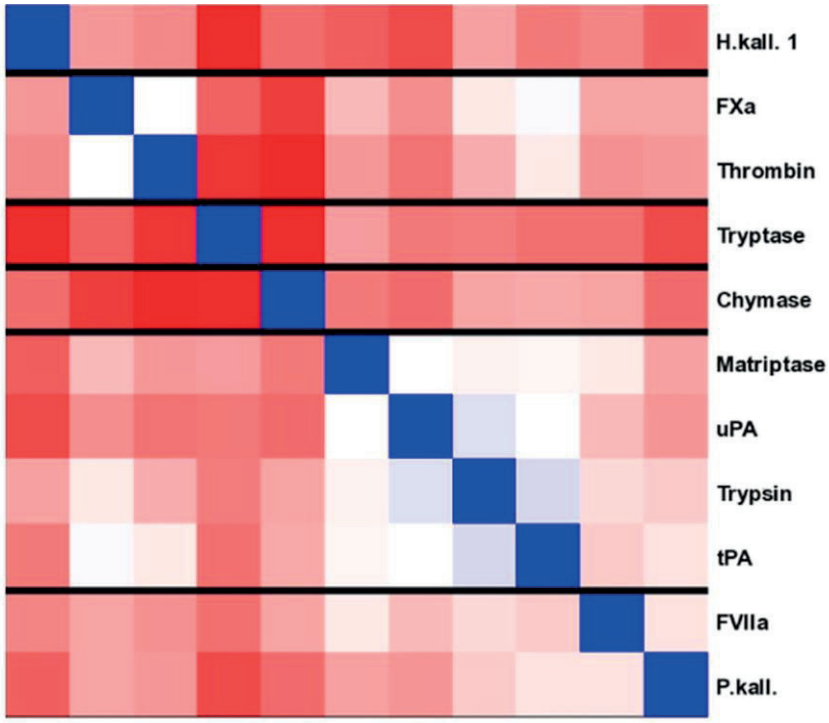

Figure 7. Heatmap showing the clusters of serine proteases generated based on the properties of binding cavities. Deep blue and deep red colours correspond to maximum similarity and dissimilarity, respectively. Black lines in the heatmap separate the different clusters. *Figure reproduced in part with permission from (Glinca, S.; Klebe, G. Cavities Tell More than Sequences: Exploring Functional Relationships of Proteases via Binding Pockets. J. Chem. Inf. Model.* ***2013****, 53 (8), 2082–2092). Copyright (2013) American Chemical Society.*

3 Deciphering ligand selectivity through predictive modelling

Apart from the computational approaches presented above, ligand selectivity can also be investigated through Structure Activity Relationship (SAR) analysis. The concept of SAR first evolved from the analysis of the relation between chemical composition of ammonium salts and their physiological action (Brown and Fraser, 1868). An attempt to estimate this relationship quantitatively was introduced by Hansch and Fujita in 1964 (Hansch and Fujita, 1964). Quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) are the statistical methods commonly employed in drug discovery to elucidate relationships between chemical structures and biological activities / physicochemical properties.

3.1 QSAR: an overview

QSAR models can highlight ligand features that have the potential to modulate the ligands' activities at drug targets, hence providing a way to propose suitable chemical modifications relevant for enhancing the efficacy of the ligand. Since its inception in 1964, there has been a growing trend for applying QSAR modelling in various fields of science. A simple search in PubMed with the term "QSAR" results in 14587 hits with more than 500 publications in 2016, revealing the popularity of these methods. QSAR models should be validated based on a set of principles published by the Organization for Economic Co-operation and Development (OECD), with a specific focus on robustness, applicability domain and mechanistic interpretation (http:// www.oecd.org). Apart from their applications in drug discovery to predict binding affinities and toxicities, QSAR techniques are also used in other fields such as environmental research, chemical mixtures modelling and nanomedicine (for review, see Cherkasov et al., 2014).

3.1.1 Classical versus non-classical QSAR

QSAR techniques fall into two categories, namely classical and non-classical QSAR depending on the compound series, descriptors and machine learning approaches used for modelling (Table 1) (Fujita and Winkler, 2016).

Table 1. Differences between classical QSAR and non-classical QSAR models (adapted from Fujita and Winkler, 2016).

	Classical QSAR	**Non-classical QSAR**
Type of compounds	Congeneric series	Large and diverse datasets
Descriptors	Empirical descriptors like Hammett substituent parameters and log P that reflect the compound's electrostatic, hydrophobic and steric properties	Non-empirical descriptors that cover a wide range of properties including physicochemical properties, molecular connectivity and stereochemistry
Machine learning approach	Simple linear regression techniques like PLS	Both linear and non-linear techniques (PLS, RF, SVM, etc.)
Applicability domain	Small; local models	Reliable predictions for a large set of new compounds; global models
Interpretation	Easy to interpret and gain clear insights into relevant molecular features	Limited interpretability due to multiple modes of action resulting from heterogeneous datasets

PLS – Partial Least Squares Regression; RF-Random Forests; SVM-Support Vector Machines

3.1.2 QSAR modelling

QSAR modelling involves a series of steps: data collection, data curation, descriptor calculation, model building and model validation. A typical QSAR modelling workflow is presented in Figure 8 (Golbraikh et al., 2012).

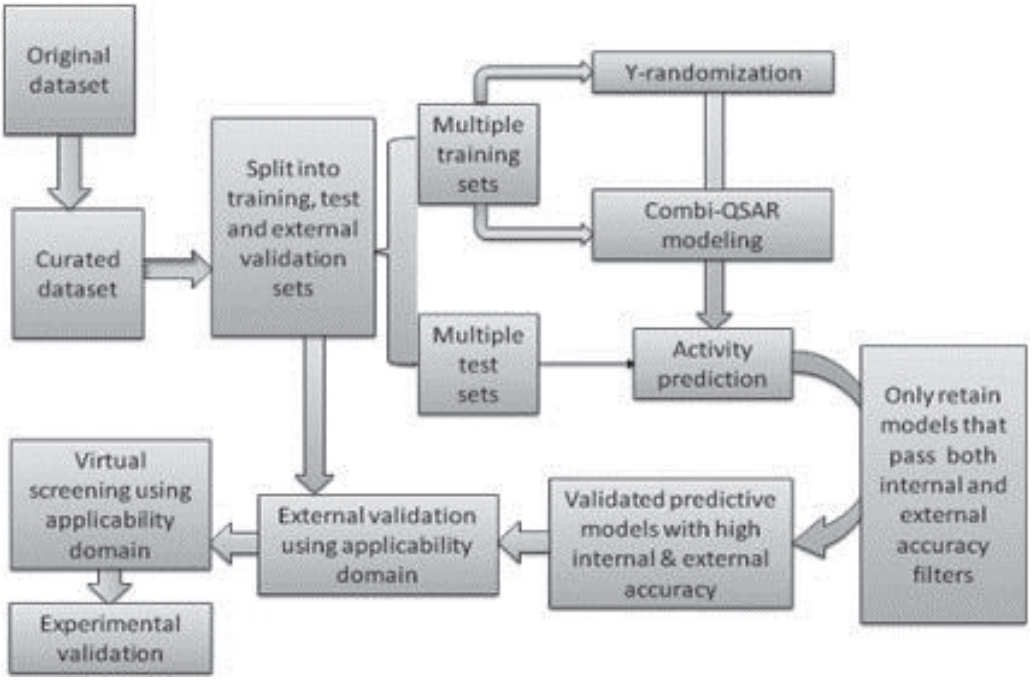

Figure 8. Predictive QSAR modeling workflow. *Figure reproduced by permission from John Wiley & Sons, Inc.: Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol. Inform.* ***2010****, 29 (6–7), 476–488., copyright (2010).*

Data collection

The most common resources for extracting structural and activity data suitable for QSAR modelling are public databases like ChEMBL (Bento et al., 2014), ZINC (Irwin et al., 2012) and PubChem (Kim et al., 2016). Also, commercial

databases such as WOMBAT and Merck Index are used (for review, see Oprea and Tropsha, 2006).

Data curation

Following data collection, data need to be curated to avoid errors resulting from incorrect chemical structures. Adopting a curation protocol that involves removal of inorganic compounds and salts, removal of duplicates and standardization of chemical structures would allow generation of reliable chemical structures ideal for descriptor calculations (Fourches et al., 2010). In addition to the chemical structures, the quality of biological data also influences QSAR modelling (Williams and Ekins, 2011). Care should be taken when utilizing the data generated by experiments conducted under similar assay conditions.

Descriptor calculations

QSAR models can be derived from 1D, 2D or 3D descriptors that differ by the level of information encoded (Damale et al., 2014). Seldom, QSAR modelling is also extended to 4D or more, with the inclusion of advanced descriptors that takes into account e.g. ligand and receptor flexibility (Kuz'min et al., 2005). A detailed account of the descriptors used in QSAR modelling is provided in Table 2.

Model building

QSAR models are generally built by employing supervised machine learning algorithms (Wikberg et al., 2011). Supervised approaches allow the model to learn from the training set that includes both input data and output variables.

The knowledge acquired is then used to predict an external set. Supervised machine learning approaches can be further classified into regression and classification techniques based on the output variables involved, either real values or class labels (Wikberg et al., 2011). Examples of supervised machine learning algorithms include linear regression, random forests, support vector machines and artificial neural networks. The choice of the machine learning algorithm to be used for QSAR modelling depends on many factors such as the dataset, model training time, predictive performance on external test sets and ease of interpretation (Sorich et al., 2003; Louis et al., 2010; Wikberg et al., 2011; Varnek and Baskin, 2012).

Model validation

QSAR models can be evaluated by internal cross-validation involving repeated exclusion of subsets of compounds from model training and using them for predictions, external prediction of compounds not used for training and Y-scrambling, which involves randomization of response variables (Grammatica et al., 2007; Tropsha, 2010; Golbraikh et al., 2012). Correlation between the observed and predicted variables (R^2), predictions from cross-validation (Q^2) and errors from cross-validation / external prediction (RMSEP) are some of the measures commonly used to assess prediction performances of QSAR models (For details on performance measures, see Methods).

Machine learning algorithms used for model building and model validation used in this thesis are explained in detail in the Methods section.

Table 2. Summary of QSAR model dimensions and the descriptors used (adapted from Kuz'min et al., 2005 and Damale et al., 2014).

QSAR model dimensions	Type of description	Examples of descriptors
1D	Macroscopic properties of molecules that rely on chemical composition	pKa, log P, electronic properties, molecular volume
2D	Molecular connectivity	Constitutional descriptors – Molecular weight, number of hydrogen bond donors, acceptors, aromatic rings Topological descriptors – Connectivity indices that account for polarizability, valency, charge transfer Quantum chemical descriptors like atomic charges, dipole moments, orbital densities 2D Molecular fingerprints – Hashed and unhashed fingerprints that define the molecular geometry
3D	Sterochemistry of molecules	3D Molecular fingerprints – 3-point and 4-point pharmacophoric fingerprints Alignment dependent descriptors – CoMFA, CoMSIA Alignment independent descriptors – GRIND
4D	Conformational flexibility of ligands	Receptor independent QSAR – Pharmacophoric features based on an ensemble of ligand conformations Receptor dependent QSAR – Weighted pharmacophores based on receptor-ligand complexes
5D	Receptor and ligand flexibility based on induced fit modelling	Receptor properties that account for steric and electrostatic effects together with the ligand properties derived from multiple protonation and stereoisomeric states
6D	Solvation effects	Solvent properties that contribute to free binding energy

3.1.3 Comparative molecular field analysis in medicinal chemistry

Comparative molecular field analysis (CoMFA) (Cramer et al., 1988), a 3D QSAR technique that relies on a template with the native binding mode has long been used in medicinal chemistry for its ability to visually illustrate the features of ligands that affect biological activity (Verma et al., 2010). CoMFA models are built based on a grid defined around the ligands that are superimposed on the template. The interaction energies calculated from steric and electrostatic fields are assigned to the grid points and are later used as descriptors to study their correlation with biological activities using Partial Least Squares (PLS) regression approach. The output from PLS models is then used to generate contour maps to gain visual understanding (Cramer et al., 1988; Zhang et al., 2011).

3.1.4 QSAR and selectivity

Assessing selectivity through QSAR has often been shown to be tedious, as it involves generation and comparison of multiple QSAR models. Studies on serine proteases demonstrated that multiple CoMFA models had to be generated for different classes of serine proteases (factor Xa, thrombin, tissue plasminogen activator, trypsin and plasmin) to understand the selectivity of a series of indole/benzimidazole-5-carboxamidines (Bhongade et al., 2005).

An alternative approach to circumvent the limitations of generating multiple QSAR models for selectivity analysis is comparative binding energy analysis (COMBINE). In COMBINE, the interaction energies calculated from ligand-receptor complexes are used to predict bioactivities (Ortiz et al., 1995). COMBINE analysis of ligands bound to several structurally related receptors could guide selectivity design. Studies on a series of 3-amidinophenylalanines

bound to various serine proteases showed that COMBINE could be successfully used to analyse the sub-pocket specificities of thrombin, trypsin and factor Xa (Murcia et al., 2006).

3.2 Proteochemometrics

Although QSAR modelling is frequently used to understand SAR, its dependency on ligand descriptors limits its usefulness in selectivity-related studies. Despite the availability of approaches like COMBINE for selectivity analysis, the need to generate ligand-receptor complexes through docking is a drawback. As the quality of the docked poses can often be questionable, using them to calculate interaction energies poses a major challenge for generating reliable QSAR models based on COMBINE. To deal with all of these limitations, Peteris and his co-workers developed proteochemometrics (Prusis et al., 2001), a method that accounts for selectivity by combining both protein and ligand description with experimentally measured data. In proteochemometric models, protein and ligand descriptors are generated independently and do not require ligand-receptor complexes. A PCM modeling workflow describing the approaches used in this thesis is shown in Figure 13. The advantage of using protein descriptors in PCM models comes mainly from their ability to extrapolate to novel chemical and target space (Van Westen et al., 2011; Cortés-Ciriano et al., 2015).

Apart from the large-scale applications of PCM in modelling the bioactivities of many drug targets (Prusis et al., 2001; Lapins et al., 2002; Strömbergsson et al., 2006; Kontijevskis et al., 2008; Lapins et al., 2008; Prusis et al., 2008; Kontijevskis, A. et al., 2009; Lapins et al., 2010; Fernandez et al., 2010; Bruyn et al., 2013; Ain et al., 2014; Cortes et al., 2015; Paricharak et al., 2015; De

Rasti et al., 2016; Simeon et al., 2016), its suitability to model antigen-antibody interactions (Mandrika et al., 2007; Dimitrov et al., 2010; Dimitrov et al., 2015), predict ligand binding free energies (Kramer et al., 2011), investigate spectral properties of fluorescent proteins (Nantasenamat et al., 2014) and utilize omics data to predict drug sensitivities against cancer cell lines (Cortés-Ciriano et al., 2015) makes proteochemometrics a promising approach. Reviews by Van Westen et al. and Cortes-Ciriano et al. summarize all PCM studies conducted up to 2013, together with the commonly employed machine learning approaches in PCM (Van Westen et al., 2011; Cortés-Ciriano et al., 2015). Some of the recently conducted studies (beyond 2013) are reported in Table 3.

Table 3. List of PCM studies conducted on various datasets since 2014.

Dataset	Ligand descriptors	Protein descriptors	Machine learning approach	*In silico* model validation	References
K_i, K_d and IC_{50} data for 2860 inhibitors and 361 kinases	Physiochemical properties, FCFP6 circular fingerprints	Z-scale descriptors	RF	Out Of Bag validation; EV	Christmann-Franck et al., 2016
K_i data for 33 inhibitors and 6 carbonic anhydrase isoforms retrieved from Carta et al., 2013	GRIND	Z-scale descriptors	PLS	Venetian Blinds CV, EV, Y-Sc	Rasti et al., 2016
Interaction data for 3063 sc-PDB complexes	Morgan fingerprints	PROFEAT sequence descriptors and 3D structural descriptors from FuzCav	RF, SVM, KNN, Naïve Bayes	10-fold CV, EV	Shaik et al., 2016
pIC_{50} data for 10 inhibitors and 8 aromatase mutants collected from Kao et al. 1996 and Auvray et al., 2002	Substructure fingerprints	Z-scale descriptors	RF, PLS, SVM, RR, BLR	10-fold CV, LOO, Y-Sc, EV	Simeon et al., 2016

Dataset	Ligand descriptors	Protein descriptors	Machine learning approach	*In silico* Model validation	References
pIC_{50} data for 3284 peptides and 5 HLA-DQ proteins from Immune Epitope database	Z-scale descriptors	Z-scale descriptors	Self-consistent iterative PLS	Y-Sc, EV	Dimitrov et al., 2015
ΔT_m for 181 compounds and 12 PARPs from Wahlberg et al., 2012	Morgan fingerprints, PaDEL descriptors	Amino acid sequence descriptors from camb package	RF	LOTO, LOCO, EV	Cortés-Ciriano et al., 2015
pIC_{50} for 1,505 inhibitors and 20 DHFRs from ChEMBL	Morgan fingerprints	Z-scale descriptors	SVM, RF, GBM, GP	5-fold CV, EV	Paricharak et al., 2015
pIC50 values from 3228 compounds and 11 mammalian cyclooxygenases from ChEMBL	Morgan fingerprints, PaDEL descriptors	Z-scale descriptors	SVM, RF, GBM	5-fold CV, EV	Cortés-Ciriano et al., 2015
Excitation and emission maxima for 18 GFP chromophores and GFP variants from Nantasenamat et al., 2007	Quantum chemical descriptors	Z-scale descriptors	PLS	LOO, EV, Y-Sc	Nantasenamat et al., 2014

Dataset	Ligand descriptors	Protein descriptors	Machine learning approach	*In silico* Model validation	References
3 different case studies (i) pK_i for 91 aminergic receptors and 11,121 ligands from ChEMBL (ii) K_{cat} for 4 dengue virus NS3 proteases and 56 ligands from Prusis et al., 2008 (iii) pK_i for 8 adenosine receptors and 4419 ligands from ChEMBL	Scitegic circular fingerprints (ECFP6)	Z-scale descriptors	GP, SVM	EV, Y-Sc	Cortés-Ciriano et al., 2014

PLS - Partial Least Squares; RF - Random Forest; SVM - Support Vector Machine; GBM - Gradient Boosting Machine; GP - Gaussian Process; Y-Sc – Y-Scrambling; EV – External Validation; LOTO – Leave One Target Out; LOCO – Leave One Compound Out; CV – Cross Validation; LOO – Leave One Out

Although the usefulness of PCM is usually shown theoretically based on the model's predictabilities and interpretabilities, there are also instances where follow-up experimental validations (Table 4) have been done to demonstrate the successful applications of PCM in compound design.

Table 4. Examples of prospective validations of PCM models.

PCM study	**Experimental validation / Inference**	**Reference**
SVM modelling of GPCRs and kinases	Novel scaffolds were identified by PCM: 3 agonists and 6 antagonists for ADRB2 and 5 inhibitors for EGFR	Yabuuchi et al., 2011
SVM modelling of HIV reverse transcriptase mutants	Experimental measurements of EC_{50} for 317 protein-ligand pairs shows a correlation of 0.69 with the EC_{50} predicted by PCM models; PCM models outperform both QSAR and KNN, with an increase in R_0^2 by 40-60%	Van Westen et al., 2011
RF modelling of oxytocin receptor	Experimental testing of 128 compounds resulted in 10 hits with >20% inhibition at 10 μM concentration; 6 hits retrieved from chemogenomics screening, including 2 potent antagonists (87 and 38% inhibition at 10 μM concentration)	Weill et al., 2011

PCM study	Experimental validation / Inference	Reference
RF modelling of OATP1B1 and OATP1B3 receptors	Comparison of model predictions and experimental validations of 54 compounds shows that OATP1B1 and OATP1B3 inhibitors can be correctly classified as actives and inactives with 80% and 74% accuracy, respectively.	De Bruyn et al., 2013

RF - Random Forest; SVM - Support Vector Machine; KNN - K-Nearest Neighbour

3.2.1 Protein descriptors in PCM modelling

Protein descriptors that explain the characteristics of a target's binding site can be derived from either the amino acid sequences or from the 3D structures. Some of the commonly used sequence-based descriptors include Z-scales (Sandberg et al., 1998), FASGAI (Liang et al., 2008) descriptors that account for physicochemical properties, T-scales (Tian et al., 2007) and ST-scales (Yang et al., 2010) that describe topological properties and BLOSUM matrix-derived amino acid descriptors (Georgiev, 2009). A benchmarking study by Van Westen et al. (2013) shows that Z-scale descriptors generate the best-performing models, and this is in line with the increased use of Z-scale descriptors in sequence-based PCM studies conducted to date (Van Westen et al., 2011; Cortés-Ciriano et al., 2015).

Although structure-based descriptors are not frequently used in PCM modelling, the additional information captured by the 3D descriptors and visual interpretability open up new opportunities for future developments in 3D PCM. Some typical examples of the 3D descriptors used in PCM studies include local substructures of proteins (Strömbergsson et al., 2006; Strömbergsson et al.,

2008) and pharmacophoric properties of binding cavities (Weill and Rognan, 2009; Meslamani et al., 2012; Shaikh et al., 2016).

None of the descriptors discussed in this section are used in our studies. 3D field-based descriptors used in this thesis are explained in detail in the Methods section.

3.2.2 Ligand descriptors in PCM modelling

Ligand descriptors can either be 2D descriptors, comprising the ligand's physicochemical properties and molecular connectivity, or 3D descriptors, allowing the conformational space and stereochemistry of the ligands to be explored (Damale et al., 2014). A wide array of descriptors available for explaining the ligand space is compiled in the "Molecular Descriptors for Chemoinformatics" book by Todeschini and Consonni (Todeschini and Consonni, 2009). The choice of the ligand descriptors to be used for PCM modelling is purely subjective and depends on the dataset (similarity or dissimilarity between ligands), flexibility of the ligands and interpretability.

In PCM studies, the most commonly used ligand descriptors are circular fingerprints that take into account the molecule's connectivity and chemical features by considering neighbouring atoms within a certain diameter (http://www.rdkit.org). These descriptors are mainly used for their simplicity and good model predictabilities (Van Westen et al., 2011; Cortés-Ciriano et al., 2015). Apart from circular fingerprints, 3D Grid INdependent Descriptors (GRIND) is also frequently used to provide spatial representation of molecules and enhance interpretability. GRIND descriptors are alignment-independent variables, resulting from the transformation of molecular interaction fields

derived from several probes that describe the hydrogen bonding and hydrophobic properties of a molecule (Pastor et al., 2000). Despite the frequent usage of certain ligand descriptors, several PCM studies have used multiple ligand descriptors and compared the model performances based on the predictabilities during internal and external validations (Weill and Rognan, 2009; Huang et al., 2012; Gao et al., 2013; Shiraishi et al., 2013; Cortés-Ciriano et al., 2015). Ligand descriptors used in this thesis are summarized in Table 5 of the Methods section.

3.3 Perspectives on QSAR / PCM modelling

Despite the popularity and usefulness of QSAR and PCM modelling approaches in drug discovery, there are some limitations inherent to these predictive modelling techniques. Experimental errors and errors during data curation can have a significant impact on the model quality (Dearden et al., 2009). One of the commonly encountered problems in QSAR / PCM modelling is the error in descriptor calculations resulting from incorrect chemical structures and the discrepancies in values calculated by different software (Dearden et al., 2009). Other sources of error arise from insufficient training data and the use of incorrect statistics for model evaluation. The pitfalls of using q^2 based on LOO validation and R^2 based on training data as the only measures to evaluate model performances have been discussed in Tropsha et al. (2001) and Alexander et al. (2015), respectively. Another well-known problem is model overfitting, owing to the use of excessive numbers of descriptors in model training (Topliss and Costello, 1972).

QSAR models can be useful only if the descriptors reflect the actual phenomena (Johnson, 2008). The choice of descriptors is crucial to acquire a

meaningful interpretation of the models. Even though 2D descriptors are widely used in QSAR / PCM modelling, failure to account for the spatial representation of the molecules limits their usefulness. On the other hand, using 3D descriptors also entails limitations such as incorrect 3D conformations (Guimarães et al., 2016) used for descriptor calculations, limited exploration of the ligand conformational space (Cappel et al., 2015) and need for bioactive conformations for CoMFA models (Cramer et al., 1988).

Aims of this thesis

The main objective of this thesis was to deal with the limitations inherent to sequence-based PCM models that lack visual interpretability. To this end, we developed field-based proteochemometrics and applied this method to model the bioactivities of kinase and serine protease families. We used field-based protein descriptors and several 2D / 3D ligand descriptors to generate proteochemometric models that are visually interpretable. Extensive validations were conducted to support the credibility of the models and their usefulness for real-time purposes.

Specific aims were as follows:

1. To demonstrate the possibilities of applying field-based descriptors in proteochemometric modelling using kinases as a specific example (Publication I)
    - To build models on continuous bioactivity data using knowledge-based and WaterMap fields derived from kinase binding sites and 2D ligand descriptors
    - To visualize the features identified as important for binding affinity and selectivity
2. To investigate the prediction capabilities of field-based proteochemometric models on kinases (Publication II)
    - To build global classification models to predict active and inactive ligands
3. To use information-rich 3D descriptors for both kinases and ligands and to extend the application of field-based proteochemometric approaches to study the features relevant for the selectivity of serine proteases (Publication III)

- To build predictive and visually interpretable models on a set of 24 serine proteases and 5863 inhibitors
- To perform extensive validations, such as Leave One Target Out (LOTO) and Leave One Compound Cluster Out (LOCCO), to investigate the extrapolative power of the models in terms of target and chemical space

4. To investigate the influence of protein superimposition on field calculations
    - To use alignment-independent Zernike descriptors for proteins in proteochemometric modelling
    - To evaluate the prediction performances of the models based on protein fields and Zernike descriptors (unpublished results)

Materials and Methods

This section includes a short description of the materials and methods used in this thesis. Detailed explanations concerning the methods are available in the original publications (I – III).

1 Data collection

1.1 Interaction data

The experimental data used in proteochemometric modelling were mainly extracted from scientific literature and public databases such as ChEMBL and Kinase SARfari. ChEMBL (Bento et al., 2014) is an open source database that includes the bioactivity data of ~2 million compounds (small molecules, peptides and antibodies) and about 11000 targets. Additional information about assay conditions, patents and literature references make this database a useful resource for conducting large-scale virtual screening and chemogenomics studies (Bento et al., 2014). Kinase SARfari is a resource dedicated to support chemogenomics research on kinases by integrating the information about sequences, 3D structures and bioactivities (https://www.ebi.ac.uk/ChEMBL/sarfari/kinasesarfari). In Publication I, the interaction data (K_d / K_i) for 50 kinases and 80 inhibitors were extracted from three publications (Karaman et al., 2008; Davis et al., 2011; Metz et al., 2011). In Publication II, activity data (K_d, K_i, inhibition% and residual activity) for 95 kinases and 1572 inhibitors were compiled from a variety of sources, including the literature used in Publication I, Kinase SARfari and ChEMBL 18 (GSK and Millipore screening data). In Publication III, bioactivity data (K_i) for 24 proteases and 5863 inhibitors were extracted from ChEMBL 20. ChEMBL 18 and 20 mentioned here correspond to the different versions of ChEMBL.

1.2 Criteria for ligand selection

In Publication I, we applied some criteria to choose the kinase inhibitors that have the potential to bind to either DFG-in/active or DFG-out/inactive conformations of kinases. Since our models were mainly focused on the DFG-in conformations, we carefully extracted the DFG-in-like inhibitors based on the literature and by performing a fingerprint-based similarity search to known DFG-in inhibitors. Similarity searches based on Extended-connectivity fingerprints such as ECFP4 hardly found any DFG-in like inhibitors, owing to the diverse nature of compounds in the kinase dataset. Therefore, the commonly used MACCS keys (Durant et al., 2002) that rely on the predefined SMARTS patterns are an ideal choice for conducting fingerprint-based searches. In addition to filtering the DFG-in-like inhibitors, we removed compounds that had activity data for less than three kinases. In Publications II and III, no specific compound selection criteria were applied. All data available in the literature and public databases were used to build global predictive proteochemometric models for kinases and serine proteases.

1.3 Ligand structures

Ligands collected from scientific literature were downloaded from PubChem database in SDF format, whereas the structures of ligands from ChEMBL and Kinase SARfari were generated in Maestro (Schrödinger, 2011) based on their SMILES notation. Schrödinger's LigPrep module was then used to convert 2D structures to 3D and generate possible ionization states at pH 7.0 ± 2.0. To explore the conformational space further, ConfGen's (Watts et al., 2010) conformation generation module was employed. The force fields used in ConfGen were OPLS-2001 for initial structure generation and OPLS-2005 for energy minimization. From the multitude of conformations generated for each

ligand, the lowest energy conformation was chosen for 3D descriptor calculations.

1.4 Protein structures

X-ray structures of kinases (95) and proteases (24) used in this study were downloaded from PDB. The completeness of the structures (no missing residues within 5Å of the crystal ligands) together with the resolution (< 3 Å) was used as the main criterion for choosing structures. Initially, the protein structures were cleaned by removing water molecules, additional chains and ligands. A standard protocol defined by the nodes in a KNIME workflow was used for protein preparation. KNIME (Berthold et al., 2007) is an open source workflow tool with a wide range of applications, including data mining, text processing, sequence analysis and statistical data analysis. An advantage of using KNIME is the possibility to integrate the modules provided by commercial vendors like Schrödinger and MOE. The workflow used for protein preparation included the following steps: (1) Correct residues with missing atoms, (2) Add hydrogen atoms, (3) Assign protonation states to ionizable residues based on the pH values determined from pK_a predictions of PROPKA, (4) Optimize the geometry of hydrogen atoms, keeping heavy atoms fixed.

2 Protein superimposition

Protein superimposition is a procedure used to align protein structures to enable easy comparison. Following superimposition, the orthosteric binding pockets fit on top of each other, which makes them suitable for field comparisons. We used Schrödinger's protein structure alignment tool that relies on dynamic programming to provide a best fit based on the sequence and secondary structural elements and to minimize the RMSDs of C∝ atoms. The common

reference structures used for superimposing kinases and proteases were c-Met kinase (PDB: 3A4P) and Matriptase (PDB: 1EAX), respectively.

3 Methods for protein descriptor calculations

3.1 Knowledge-based fields

The ligand binding sites of kinases and serine proteases were described by fields derived from the knowledge-based contact potentials calculated by MOE. Knowledge-based contact potentials (Figure 9) are the joint probability densities derived from interatomic distance, lone-pair interaction angle and out-of-plane angle (MOE, 2011). The joint probability densities are expressed by

$$\Pr(Ligand = l \mid Position = x, Protein = p)$$

The hydrophilic and hydrophobic contact probabilities are calculated by considering the conditional probability of observing a ligand atom l at position x, provided the ligand atom l is in contact with the protein structure p. In knowledge-based field calculations, preferences for hydrophilic and hydrophobic contacts are determined based on the protein crystallographic data available in PDB (MOE, 2011). An advantage of using the knowledge-based contact potentials is that the directional preferences are taken into account and additional information is provided for describing the target-ligand interaction space.

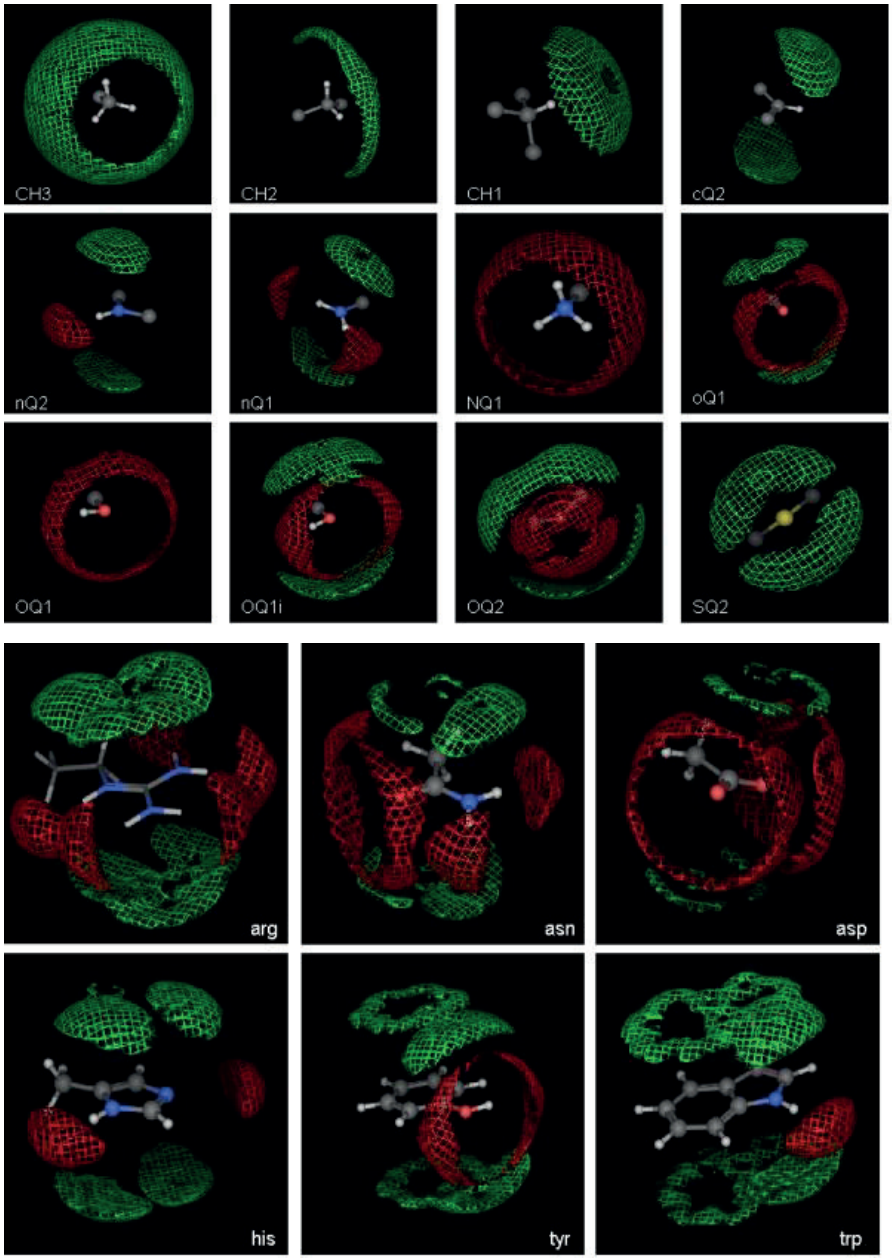

Figure 9. Contact maps of selected ligand atoms and amino acid residues. Hydrophilic and hydrophobic contact maps are shown in red and green, respectively. *Figure created using MOE (Molecular Operating Environment) is reproduced with permission from Chemical Computing Group (https://www.chemcomp.com/journal/f_surfmap.htm).*

In our studies, we calculated the knowledge-based fields by spanning a grid around the binding site. The crystal ligands extracted from PDB structures determined the size of the grid. For kinases, the dimensions of the grid were defined by 41 X-ray ligands (Publications I and II). For serine proteases, the peptide-like inhibitor that extends across the S1-S4 sub-pockets of Activated Protein C crystal structure (PDB: 1AUT) was used as the reference for grid size definition (Publication III). The space between the grid points was set to 0.5 Å. The hydrophilic and hydrophobic contact probabilities were calculated at every grid point and were influenced by the neighbouring atoms (Figure 10). The contact probabilities that exceeded the 0.9 thresholds were used as descriptors in proteochemometric modelling (Figure 11).

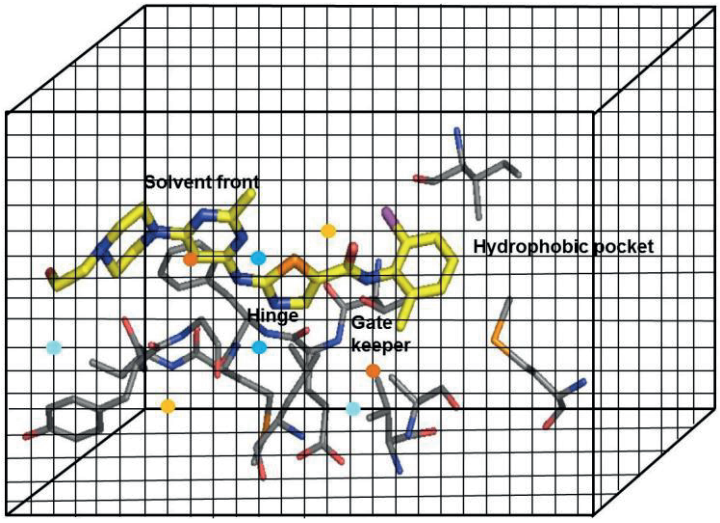

Figure 10. Schematic representation of the grid enclosing the ligand-binding site of ABL1 kinase. Blue and orange spheres with varying colour intensities correspond to the polar and lipophilic field points with different contact probabilities. *Figure adapted with permission from (Subramanian, V.; Prusis, P.; Pietilä, L. O.; Xhaard, H.; Wohlfahrt, G. Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics. J. Chem. Inf. Model.* ***2013****, 53, 3021–3030). Copyright (2013) American Chemical Society.*

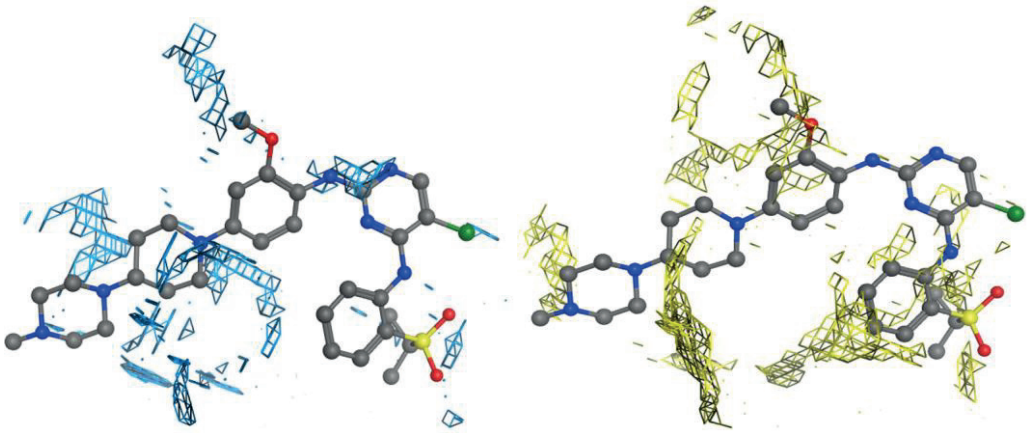

Figure 11. Knowledge-based fields calculated from the ligand-binding site of ALK kinase. The inhibitor TAE-684 extracted from the X-ray structure 2XB7 is shown as reference. Only the fields that lie in close proximity to the inhibitor are shown for clarity. **Left panel:** Polar protein fields with probability density > 0.9. **Right panel:** Lipophilic protein fields with probability density > 0.9.

3.2 WaterMap-derived fields

Schrödinger's WaterMap (WaterMap, 2012; Abel et al., 2008) can be used to predict the position of water molecules in the binding site. WaterMap calculations are based on molecular dynamics simulations that involve explicit water molecules. The predicted water sites are assigned statistical thermodynamic properties such as enthalpy, entropy and Gibb's free energy that are likely to influence ligand binding. Displacement of entropically unfavourable water molecules in the binding site facilitates ligand binding and maximizes affinity (Abel et al., 2008; Michel et al., 2009). Therefore, analysing the position and energetics of the water molecules is crucial in drug design.

We have calculated the WaterMaps for the kinase and serine protease binding sites (Figure 12a) and projected them onto the knowledge-based grids to derive WaterMap fields (Figure 12b, c) (Publications I-III). Water density values were

initially assigned to the grid points to extract the high-density regions (density > 0.06). Gibb's free energy value was subsequently assigned to classify these grid points as stable (ΔG < -1 kcal/mol) or unstable water fields (ΔG > 3 kcal/mol), which were later used as descriptors in proteochemometric modelling.

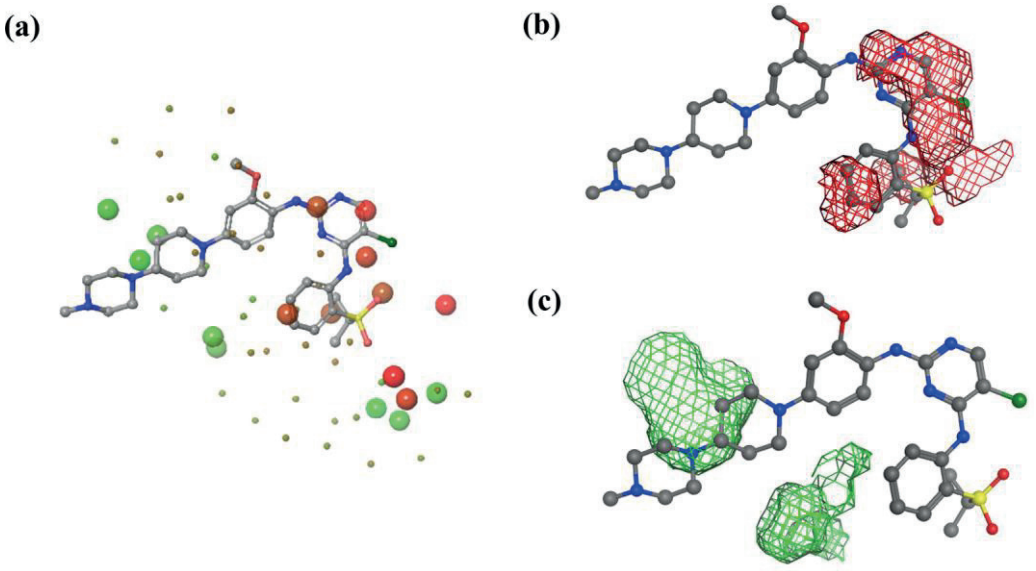

Figure 12. (a) WaterMaps calculated from the ligand-binding site of ALK kinase. The inhibitor TAE-684 extracted from the X-ray structure 2XB7 is shown as a reference. Water molecules with ΔG < -1 kcal/mol (stable - green) and ΔG > 3 kcal/mol (unstable - red) are highlighted with larger spheres. **(b)** Unstable water fields with ΔG > 3 kcal/mol. **(c)** Stable water fields with ΔG < -1 kcal/mol.

3.3 Zernike descriptors

Zernike descriptors (Novotni and Klein, 2003) are alignment-independent descriptors commonly used for shape retrieval and comparison of ligand binding sites (Nisius and Gohlke, 2012). These descriptors are a vector of coefficients of terms in the Zernike polynomial expansion series and are

insensitive to misalignments of binding sites. We have calculated Zernike descriptors using our in-house scripts by transforming the knowledge-based and WaterMap-derived fields through a series expansion in 3D Zernike polynomials. Zernike function can be represented as

$$f(\overrightarrow{x}) = \sum_{n=0}^{N} \sum_{l=0}^{n} \sum_{m=-l}^{l} \Omega_{nl}^{m} . Z_{nl}^{m} \ (\overrightarrow{x}) \qquad \text{Eq.1}$$

where $(\overrightarrow{x})$ is the field vector, Ω_{nl}^{m} is the Zernike moment, Z_{nl}^{m} is the Zernike polynomial and N is the maximum number of expansion terms. n, l and m in Eq.1 correspond to the principal, azimuthal and magnetic quantum numbers, respectively.

Zernike descriptors were computed with varying orders of N (3, 5, 10, 20, 30, 40, 50) and used as protein descriptors in PCM modelling.

4 Ligand descriptors

Both 2D and 3D ligand descriptors were used for proteochemometric modelling (Table 5). This includes Open Babel's (Boyle et al., 2011) FP4 fingerprints (Publications I and II), Mold^2 (Hong et al., 2009) descriptors (Publications I and II), Volsurf (Cruciani et al., 2000) descriptors (Publication I), 4-point pharmacophoric fingerprints (4-PFP) from Canvas (Canvas, 2014; Duan et al., 2010) (Publications II and III), MOE (MOE, 2015) descriptors (Publication III), RDKit fingerprints (http://rdkit.org) (Publication III) and Pentacle's GRIND descriptors (Pastor et al., 2000) (Publication III).

Table 5. Ligand descriptors for PCM modeling.

Descriptor	Information encoded
	2D descriptors
Open Babel FP4 fingerprints	Atom and bond properties assigned based on predefined SMARTS patterns
Mold[2]	Counts of atoms and bonds, physicochemical properties and topological descriptors
MOE descriptors	Counts of atoms and bonds, physicochemical properties and descriptors that account for molecule's topology, pharmacophore features and partial charges
RDKit fingerprints	Circular fingerprints that describe molecule's connectivity and chemical features
	3D descriptors
Volsurf	Physiochemically relevant numerical descriptors derived from GRID molecular interaction fields calculated based on water, hydrophobic and hydrogen bond acceptor probes
4-PFP	Pharmacophoric fingerprints based on the hydrogen bond donor (D), hydrogen bond acceptor (A), hydrophobic (H) and aromatic ring features (R)
GRIND	Alignment-independent variables obtained from the transformation of molecular interaction fields calculated using hydrogen bond donor (O), hydrogen bond acceptor (N1) and hydrophobic (DRY) probes

5 Data pre-processing

Descriptors with near-zero variance were removed to reduce random noise. Presence of multiple classes of descriptors could introduce bias in the modelling process, which makes it necessary to apply the scaling and centring techniques. All of the descriptors were centred and scaled to unit variance (UV). In UV scaling, the descriptors are multiplied by the base weight, which is the inverse of the standard deviation calculated for each descriptor column. Additionally, block scaling (SIMCA, 2011) was applied for the descriptors in PLS models. Each descriptor class was considered as a separate entity called a block. In case of block scaling, each variable was multiplied by the block weight $1/\sqrt{b}$, where b is the number of variables in each block.

6 Principal Component Analysis

Principal Component Analysis (PCA) (Wold et al., 1987) is a dimensionality reduction technique that involves orthogonal transformation of variables. The projection of variables to a lower dimensional space transforms them into linearly uncorrelated variables and they constitute the principal components (PCs). The transformed values for each data point constitute the PC scores, and the weights that explain the contributions of original variables towards the PC score calculations represent the loadings. PCA enables one to visualize the variation in data. We have applied PCA to both protein and ligand descriptors (Publications I-III).

7 Cross-term descriptors in proteochemometric modelling

Cross-terms (Wikberg et al., 2004) are commonly used to introduce non-linearity in models that use linear approaches like PLS. Cross-terms can be computed as a product of protein-protein or ligand-ligand or protein-ligand

descriptors. Protein-protein and ligand-ligand cross-terms account for the intramolecular interactions in target and ligand space, respectively, whereas, protein-ligand cross-terms that explain the intermolecular interactions between proteins and ligands facilitate understanding of selectivity. Only the protein-ligand cross-terms are used in our studies (Publications I and III).

8 Generation of training and test sets

The reliability of the PCM models can be best assessed by training the models on a dataset (training set) and using them to predict the bioactivities of a new set (test set) that has not been used in training the model. The training and test sets in PCM models can either be generated randomly (Publication III) or selected carefully after performing a diversity analysis on the dataset (Publication II). Training and test sets used for PCM modelling in different publications are presented in Table 6.

Table 6. Training and test sets used in PCM modelling

Publication	Dataset	Methods used for training set generation	Training set	Test set
I	50 kinases and 80 inhibitors	-	951 observations	25 inhibitors and 50 kinases (309 observations)
II	95 kinases and 1572 inhibitors	**Ligand prediction set (80:20):** RDkit diversity picker node in KNIME used to pick diverse compounds based on MACCS fingerprints	1257 inhibitors and 95 kinases (51042 observations)	315 inhibitors and 95 kinases (12145 observations)
		Target prediction set (80:20): One kinase selected from each cluster generated based on the knowledge-based and WaterMap-derived fields	1572 inhibitors and 75 kinases (51302 observations)	1572 inhibitors and 20 kinases (11885 observations)
III	24 serine proteases and 5863 inhibitors	Random split (70:30) of observations	4350 inhibitors and 24 serine proteases (6199 observations)	1513 inhibitors and 24 serine proteases (1709 observations)

Observations / data points: Protein-ligand combinations

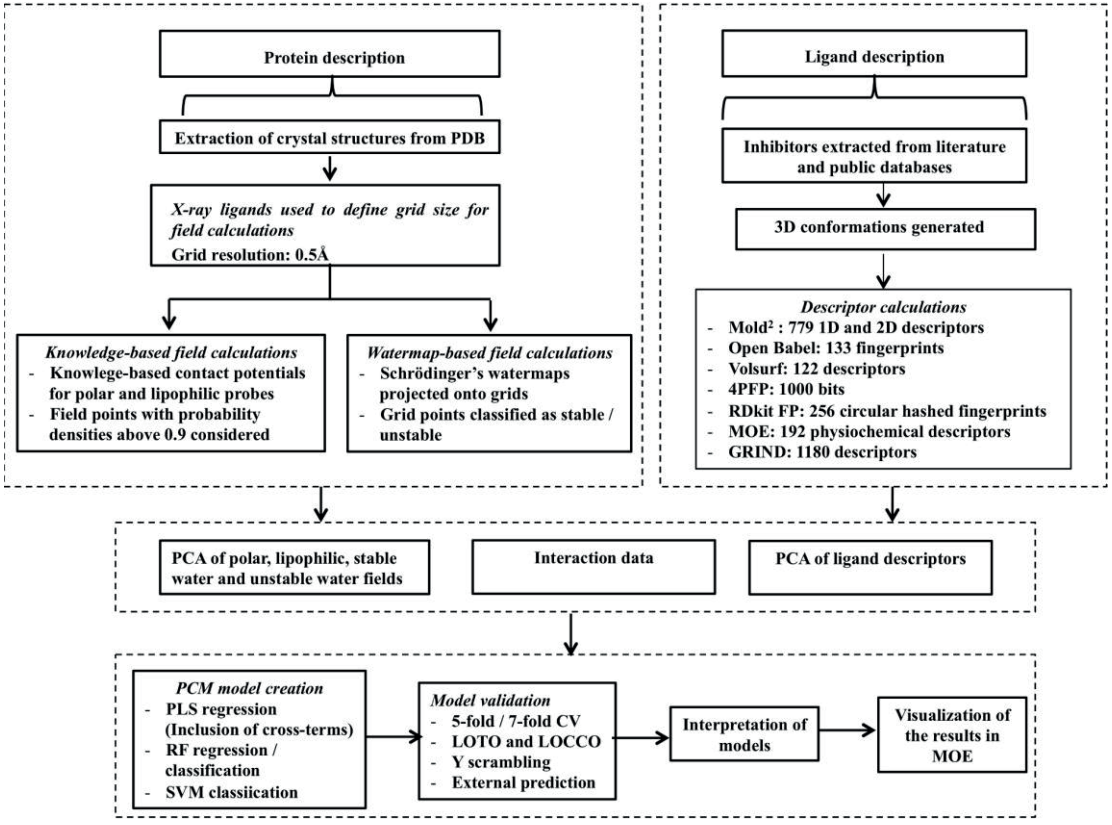

Figure 13. Flow chart of the steps involved in field-based PCM modelling (Publications I-III). *Figure adapted with permission from (Subramanian, V.; Prusis, P.; Pietilä, L. O.; Xhaard, H.; Wohlfahrt, G. Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics. J. Chem. Inf. Model.* ***2013****, 53, 3021–3030). Copyright (2013) American Chemical Society.*

9 Machine learning approaches

9.1 Partial Least Squares Regression

Partial Least Squares (PLS) (Geladi et al., 1986; Wold et al., 2001) regression is a linear modelling technique used to study the correlation between a set of independent/predictor variables (X) and one or more dependent/response variables (Y). The PLS components are extracted by projecting both X and Y variables into new spaces to explain the maximum covariation between X and Y. In Publications I and III, we used the PLS approach to model the correlation between protein/ligand descriptors (X) and experimental binding affinities (Y). The equation describing the protein-ligand interactions in PLS models can be expressed as follows (Lapinsh et al., 2005):

$$Y_c = Y_m + \sum_{l=1}^{L} coeff_l * x_l + \sum_{p=1}^{P} coeff_p * x_p + \sum_{l=1,p=1}^{L*P} coeff_{l,p} * x_l * x_p \quad \text{Eq.2}$$

where Y_c is the computed Y value, Y_m is the mean Y value, x_l is the ligand descriptor matrix, x_p is the protein descriptor matrix and $x_l * x_p$ is the cross-term; $coeff_l$, $coeff_p$ and $coeff_{l,p}$ are the regression coefficients of ligands, proteins and cross-terms, respectively.

9.2 Random Forests

Random Forests (RF) (Breiman, 2001) are non-linear machine learning approaches dependent on an ensemble of decision trees to generate predictive models. A decision tree is a tree-like model that includes a series of decisions and possible outcomes. Using a single decision tree could lead to biased modelling and affect the prediction accuracies. Growing a random forest of decision trees by selecting random subsets of attributes from the feature space could significantly boost the model performances. The RF approach employs a

bagging algorithm where subsets of samples are randomly excluded to estimate the prediction errors. RF models are robust, as they are not strongly dependent on data pre-processing strategies and are less sensitive to outliers and noise. RF was used for training classification models in Publication II and regression models in Publication III.

9.3 Support Vector Machines

Support Vector Machine (SVM) (Cortes and Vapnik, 1995) is a machine learning approach that aims to construct hyperplanes to maximize the separation between different classes of data. For data that are not linearly separable, SVMs project the data to a high-dimensional feature space by employing kernel tricks. SVMs rely on many different kernel functions, some of which include radial basis function kernel, polynomial kernel, linear kernel and string kernels. The kernel functions differ by the parameters used and the way in which the feature mapping is performed. Optimizing the kernel parameters is critical in SVM modelling to find the optimal classifier. We have used SVM to train classification models on a kinase dataset to separate actives and inactives (Publication II).

10 Model validation

10.1 Cross-validation

Validating the models is crucial to assess their robustness. One of the commonly used internal validation procedures is cross-validation (CV), where a subset of data is excluded from the modelling process and used as an external test set. The different variants of cross-validation include

(i) *K-fold CV*: Data are split into k subsets. The model is trained on k-1 subsets and tested on the omitted set. This procedure is repeated,

until each of these subsets is tested once. The K-fold CV approach is frequently used in model validation.

(ii) *Leave One Out validation (LOO)*: It is an exhaustive validation procedure, where each and every observation is excluded for testing and the model is trained on the remaining observations. This approach is not so robust and the studies have shown that using LOO as the only validation approach is not optimal, but using it in combination with other validation techniques could be useful (Golbraikh et al., 2001).

(iii) *Leave One Target Out Validation (LOTO)*: The observations corresponding to the targets used in PCM modelling are excluded one at a time to evaluate the model's extrapolation capabilities in terms of target space.

(iv) *Leave One Compound Cluster Out validation (LOCCO)*: The observations corresponding to a compound cluster (compounds grouped together based on their descriptor space) are excluded to assess the model's prediction performances on a new compound space.

(v) *Double CV*: It is a nested CV approach, where the validations are conducted by considering outer and inner loops. In the outer loop, the dataset is split randomly into training and test sets. The training set is validated by dividing it into k subsets in the inner loop to choose the optimal model for external test set predictions. Double CV provides a robust way to validate the models.

(vi) *Repeated random subsampling*: Also known as the Monte Carlo method, where the validations are performed multiple times on different random splits of training and test sets. This method has the

disadvantage of choosing the same observation numerous times and excluding some observations completely from the validation cycles.

K-fold CV (7-fold in Publication I, 5-fold in Publications II and III), LOTO (Publication I and III) and LOCCO (Publication III) are the validation procedures used in our studies.

10.2 External prediction

A model that performs well in internal cross-validation does not ensure its predictability on a completely external test set (Golbraikh et al., 2002). So, there is a need to assess the external predictive power of the models by either acquiring a new set of observations (Publication I) or by dividing the existing dataset into training and test sets (Publications II and III).

10.3 Permutation validation

Permutation validation / Y scrambling (Eriksson et al., 1997) is a procedure generally used to evaluate model overfitting. Models dependent on a large number of descriptor variables can often result in spurious correlations (Topliss and Costello, 1972; Eriksson et al., 1997) and perform poorly when applied to an external test set. Therefore, it is necessary to validate the models on random data. We have conducted permutation testing 20 times by fitting the models to random data generated by reordering experimental affinity values (Publication I) or activity classes (Publication II).

11 Model performance

In the assessment of the continuous model performances (Publications I and III) based on internal cross-validations, external predictions and permutation validations, the following measures were used:

Correlation coefficient (R^2) and Predictability (Q^2): R^2 is a measure used to estimate the agreement between the observed and calculated values of the training data.

$$R^2 = 1 - \sum_{i=1}^{N} \frac{(Y_i - Y_{calculated})^2}{(Y_i - \bar{Y})^2} \quad \text{Eq.3}$$

Here, Y_i refers to the observed measurements and $\bar{Y}$ is the mean value of all of the observed measurements.

Q^2 is an estimate of the correlation between the observed and predicted values during CV rounds.

$$Q^2 = 1 - \sum_{i=1}^{N} \frac{(Y_i - Y_{PredCV})^2}{(Y_i - \bar{Y})^2} \quad \text{Eq.4}$$

Here, Y_i refers to the observed measurements of the subset excluded during CV; Y_{PredCV} corresponds to the predicted values during cross-validation and $\bar{Y}$ is the mean value of all of the observed measurements.

Root Mean Square Error of Estimation (RMSEE): Prediction errors computed by comparing the calculated values ($Y_{calculated}$) of all of the observations (N) used for modelling with the experimentally measured values (Y_i)

$$RMSEE = \sum_{i=1}^{N} \sqrt{\frac{(Y_i - Y_{calculated})^2}{N}} \quad \text{Eq.5}$$

Root Mean Square Error of Prediction ($RMSEP_{CV}$): Prediction errors computed by comparing the values of all of the observations (N) predicted during CV rounds (Y_{PredCV}) with that of the experimentally measured values (Y_i)

$$RMSEP_{CV} = \sum_{i=1}^{N} \sqrt{\frac{(Y_i - Y_{PredCV})^2}{N}} \quad \text{Eq.6}$$

Root Mean Square Error of Prediction ($RMSEP_{test}$): Prediction errors computed by comparing the values of all the observations (N) predicted for an external test set ($Y_{Predicted}$) against the experimentally measured values (Y_i)

$$RMSEP_{test} = \sum_{i=1}^{N} \sqrt{\frac{(Y_i - Y_{Predicted})^2}{N}} \quad \text{Eq.7}$$

Assessment of Permutation validation results

The intercepts obtained by plotting the correlation coefficients of the original and permutated values against the correlation (R^2) and predictability (Q^2) values were used as the basis of assessing permutation validation (Figure 14). R^2 intercepts below 0.3 and negative Q^2 intercepts imply that a model is valid (Eriksson et al., 1999) enough for further predictions and interpretations.

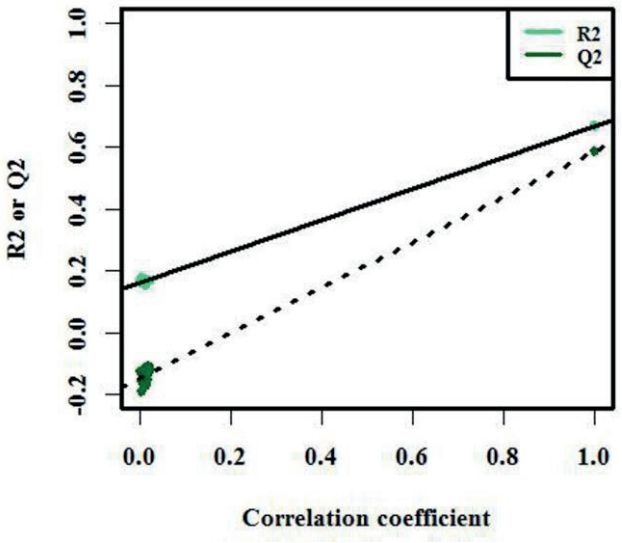

Figure 14. An example of permutation validation conducted on serine protease dataset. Permutation plots shown here correspond to the PLS models based on protein fields and RDkit fingerprints. Colored dots in the figure correspond to the R^2 and Q^2 values of the 20 models built with randomly permuted Y values.

For classification models, the performances were evaluated by computing several measures (accuracy, sensitivity/true positive rate, specificity/1-false positive rate, Matthews correlation coefficient (MCC) (Matthew, 1975), kappa coefficient (Cohen, 1960) and area under the ROC curve (AUC) (Linden, 2006)) dependent on the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

$$Accuracy(\%) = \frac{TP+TN}{TP+FP+FN+TN} \quad \text{Eq.8}$$

$$Sensitivity = \frac{TP}{TP+FN} \quad \text{Eq.9}$$

$$Specificity = \frac{TN}{FP+TN} \quad \text{Eq.10}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad \text{Eq.11}$$

$$Kappa\ coefficent = \frac{Accuracy-Random\ accuracy}{1-Random\ accuracy} \quad \text{Eq.12}$$

where $Random\ Accuracy = \frac{(TP+FN)*(TP+FP)+(TN+FP)*(TN+FN)}{N^2}$; N is the total number of observations.

AUC: AUC refers to Area Under the Curve and it is a measure of a classifier's potential to rank true positives higher than false positives; AUC values are estimated based on the ROC curve, which is a plot of the false-positive rate (FPR) against the true-positive rate (TPR) (Figure 15). AUC of 1 implies that the classification is perfect, and AUC of 0.5 indicates randomness.

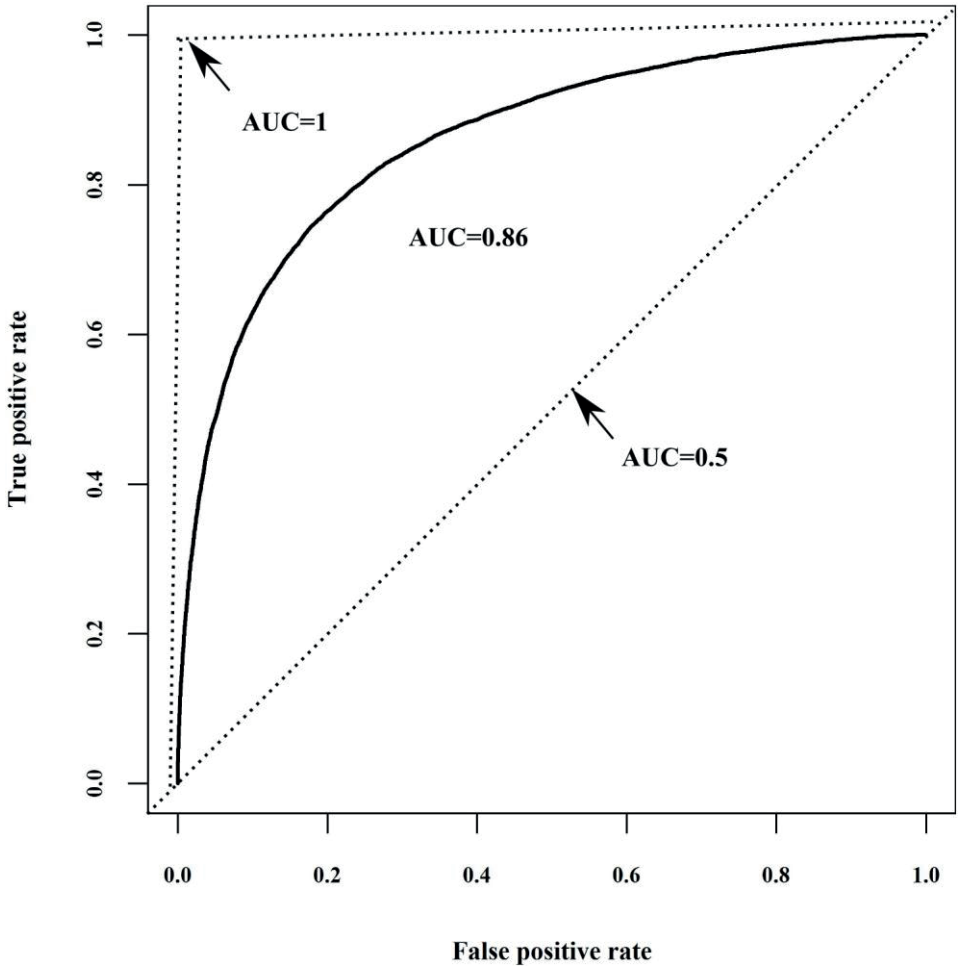

Figure 15. ROC curves of the SVM classification models of the kinase dataset shown relative to the ROC curves with random and maximum AUCs.

12 Model interpretation

12.1 Interpretation of PLS models

PLS models were interpreted by analysing the features that have a positive influence on binding affinity (Publications I and III). Protein and ligand descriptors related to affinity of ligands towards kinases/proteases were identified based on the positive PLS coefficients ($coeff_l$ and $coeff_p$ as described

in Eq. 2). Further, to interpret the features related to selectivity, we used cross-term coefficients ($coeff_{i,p}$ in Eq. 2). The descriptors used in PCM modelling being the PC scores, the original protein field points and ligand functional groups / fingerprints were traced back by examining the loadings of the PCs. Since interpretation of the protein field points is a laborious process, we restricted the interpretation to positive PLS coefficients and analysed only the top 10 loadings of PCs.

12.2 Interpretation of RF models

In RF models, Gini index and the descriptor's correlation to active class were used as the basis for interpretation (Publication II). Gini indices are a measure of the homogeneity of the nodes and depend on the variables used for splitting in decision trees. The higher the decrease in Gini index, the more the descriptor has relevance for classification (Liaw, 2002). However, the Gini indices fail to account for the descriptor's relevance for active or inactive class. Subsequently, the descriptors that had high correlation values for the active class were selected for interpretation. The protein field points and ligand's 4-PFPs that make a ligand active or inactive towards a specific kinase were identified by analysing the loadings of PCs, as described above.

13 Applicability Domain (AD) analysis

A model's usefulness can be evaluated by its capability to predict new targets and ligands that have not been used in the modelling process. The scope and accuracy of predictions can be ascertained based on the similarity of the external ligand or target to its training space (Jaworska et al., 2005).

13.1 Ligand space

In Publications II and III, the AD analysis was conducted by inspecting the Tanimoto similarities of the test set ligands based on their fingerprints against the training set compounds.

$$Tanimoto\ coefficient = \frac{N_{Te.Tr}}{N_{Te}+N_{Tr}-N_{Te.Tr}} \quad \text{Eq.13}$$

Here, $N_{Te.Tr}$ is the number of bits found in both training and test set ligands; N_{Tr} is the number of bits found only in the training set ligand and N_{Te} is the number of bits found only in the test set ligand.

The similarity thresholds ideal for reliable predictions of the test set ligands was determined by considering their prediction accuracies (> 80%) in Publication II and RMSEPs in Publication III ($RMSEP_{test} < 1$).

13.2 Target space

The extent to which the models can be applied to understand polypharmacology depends on their extrapolative power. Extrapolation to novel targets was assessed by computing the Euclidean distance between the training and test set kinases, based on the PC scores of protein field descriptors (Publication II). Euclidean distances were calculated by using the following formula:

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{N}(Kin_{Te} - Kin_{Tr})^2} \quad \text{Eq.14}$$

where Kin_{Tr} and Kin_{Te} refer to the protein descriptors of training and test set kinases, respectively.

Summary of main results

1 Characterization of datasets

The ligand, target and bioactivity space of the kinase and serine protease datasets were analysed by considering the description of the molecules, knowledge-based and WaterMap-derived fields of the protein's binding pockets and the distribution of bioactivities, respectively. Ligand space was analysed only by considering the descriptors that gave the best performance in PCM modelling (for details, see Table 8 and 10)

1.1 Kinase dataset

Ligand space

In Publication I, clustering 80 ligands based on the Euclidean distances computed from the PC scores of Open Babel fingerprints shows that there are many compounds that overlap in terms of chemical space, which in turn limits the diversity of the dataset. Ligands clustered together based on chemical space have different propensities towards kinases (Figure 16), which makes it rather difficult to predict the bioactivities of new compounds. Among the 80 ligands for which activity data are known for all 50 kinases, 29 ligands are highly selective, interacting with less than 5 kinases. Nearly 20% of the ligands are non-selective, the most promiscuous being staurosporine, which interacts with 47 out of 50 kinases in the dataset. Considering the grouping patterns of ligands based on their selectivity, no distinct clusters are observed with respect to ligands that act on 5 kinases or 40 kinases.

In Publication II, the dataset was extended to include 1572 inhibitors. Principal component analysis based on 4-PFPs shows the uniform distribution of the compounds in the PC space, which in turn reflects the diversity of the dataset.

Sixty-two PCs that explain 80% of the total variation in data were extracted to be used in PCM modelling. For simplicity, only two PCs that explain 33% of the variance are shown here (Figure 17). Further, comparing the distributions of Tanimoto similarities of the training and test ligands (Figure 18) shows that nearly 80% of the test set ligands have Tanimoto similarities above 0.7 with the training set ligands. The overlap in chemical space between the training and test set compounds is comparatively higher than the similarities among the training set compounds, supporting the reliability of the test set predictions.

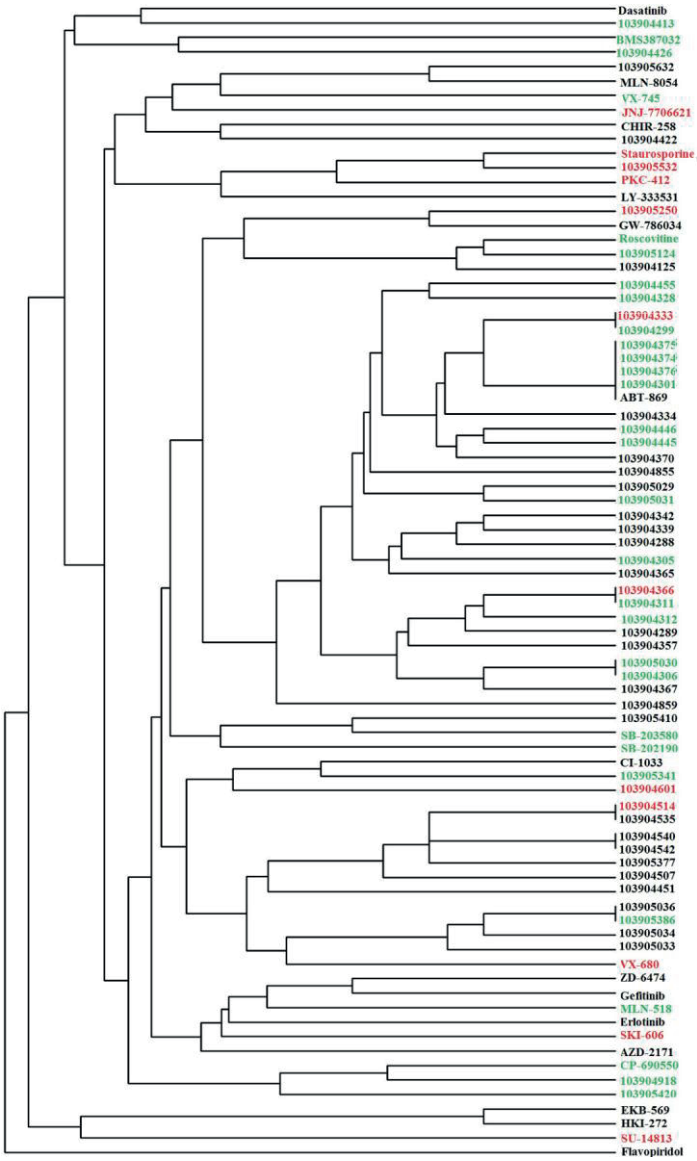

Figure 16. Hierarchical average linkage clustering of 80 kinase inhibitors based on their Open Babel fingerprints (Publication I). Different coloured text in the figure corresponds to selectivities of the compounds, categorized by number of kinases, with which they are active (Most selective (green): ≤ 5 kinases; Moderately selective (black): 6-20 kinases; Least selective (red): ≥ 20 kinases).

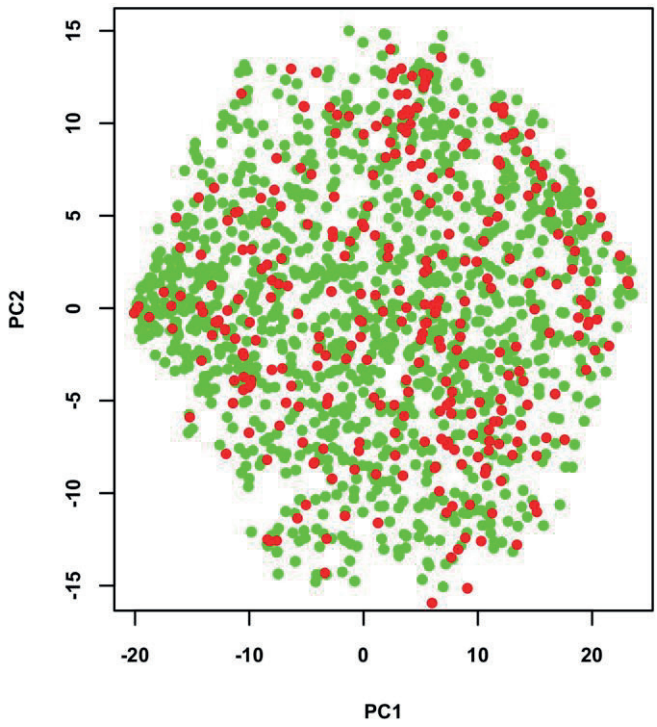

Figure 17. Scatter plots of the first two principal components of the 4-PFP descriptor space of 1572 kinase inhibitors (Publication II). Green and red dots correspond to inhibitors in training and test set, respectively.

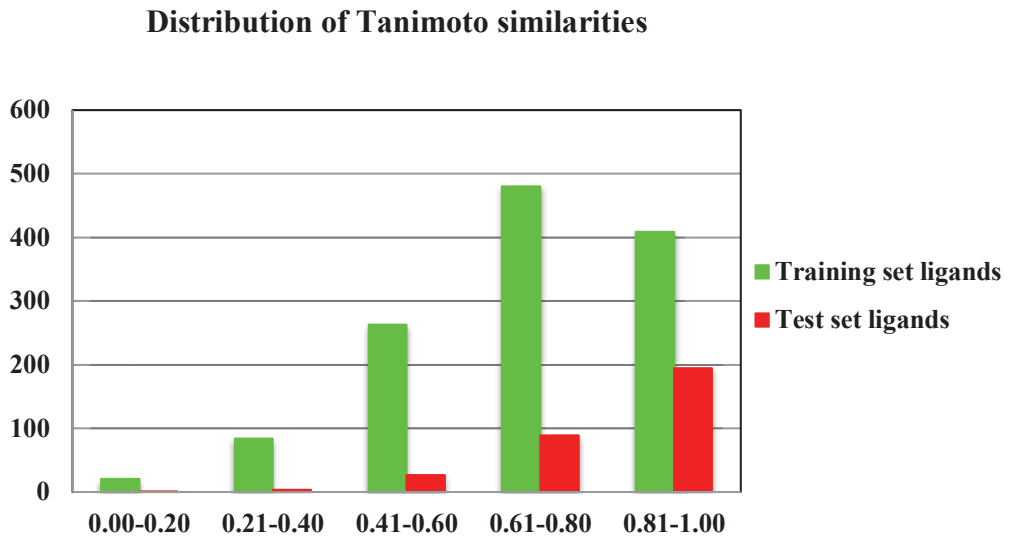

Figure 18. Distribution of Tanimoto similarities computed based on the 4-PFPs of kinase inhibitors in Publication II. Green bars correspond to the Tanimoto similarities among the 1257 ligands in training set; Red bars correspond to the Tanimoto similarities of the 315 test set ligands against the training set ligands.

Target space

As far as the kinase space is concerned (Publications I and II), it includes multiple representatives from all of the major kinase families. The knowledge-based and WaterMap fields seem to be rather similar for various kinase subgroups, which is reflected in their clustering patterns (Figures 19 and Figure 20). For most of the kinase families, the subtypes are grouped together in the same cluster. However, there are a few exceptions. For instance, the CDK (CDK2, CDK5, CDK9) and PAK (PAK1, PAK6, PAK7) subtypes are placed apart in different clusters, despite their high sequence similarity. The even spread of different kinase families along the cluster tree together with the subgroup similarities enables extrapolation to novel kinases and provides a dataset suitable for predicting the activity and selectivity profiles of inhibitors

that bind to different kinase subgroups. The 50 kinases in Publication I are a subset of the 95 kinases included in Publication II.

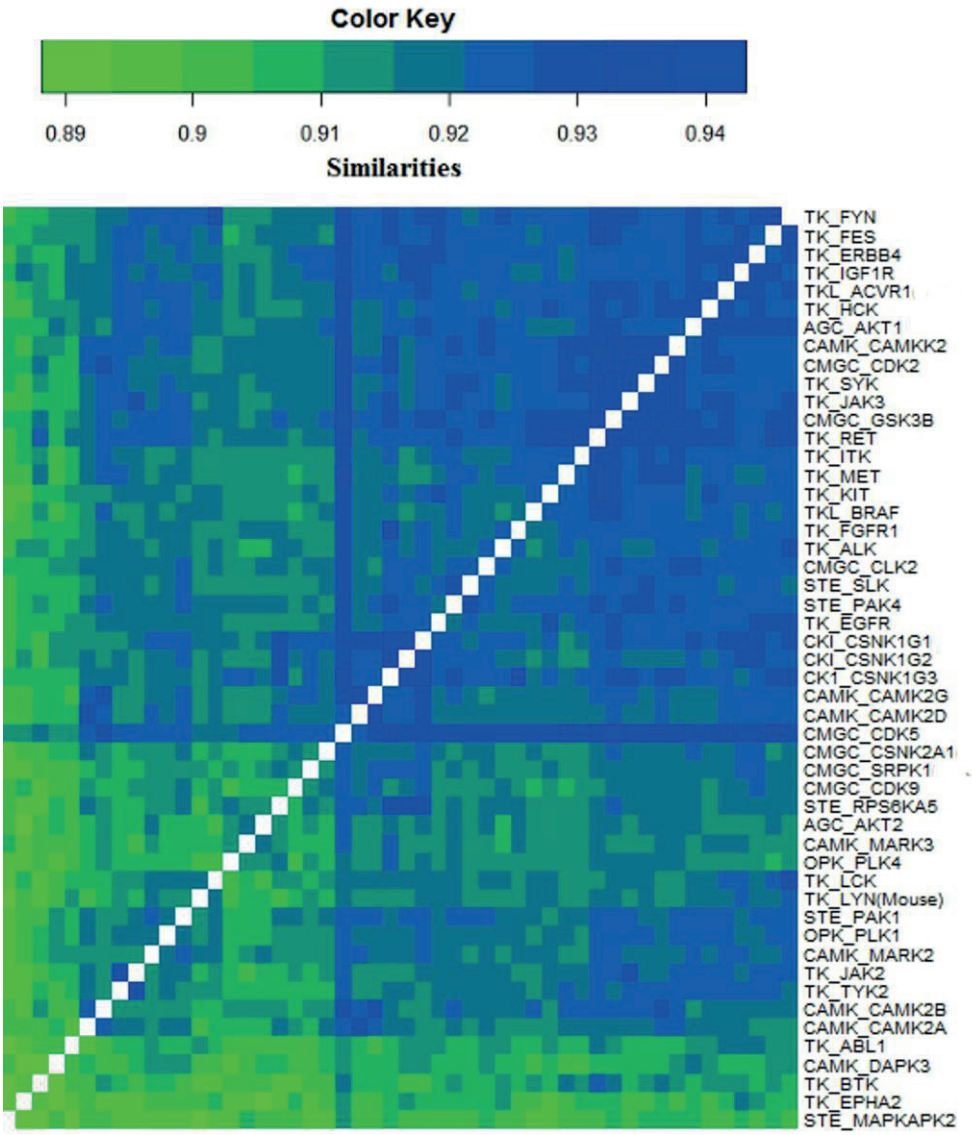

Figure 19. Similarity heatmaps of 50 kinases derived from the combined knowledge-based and WaterMap fields (Publication I). The row labels correspond to the various kinase subtypes with the family names preceding them.

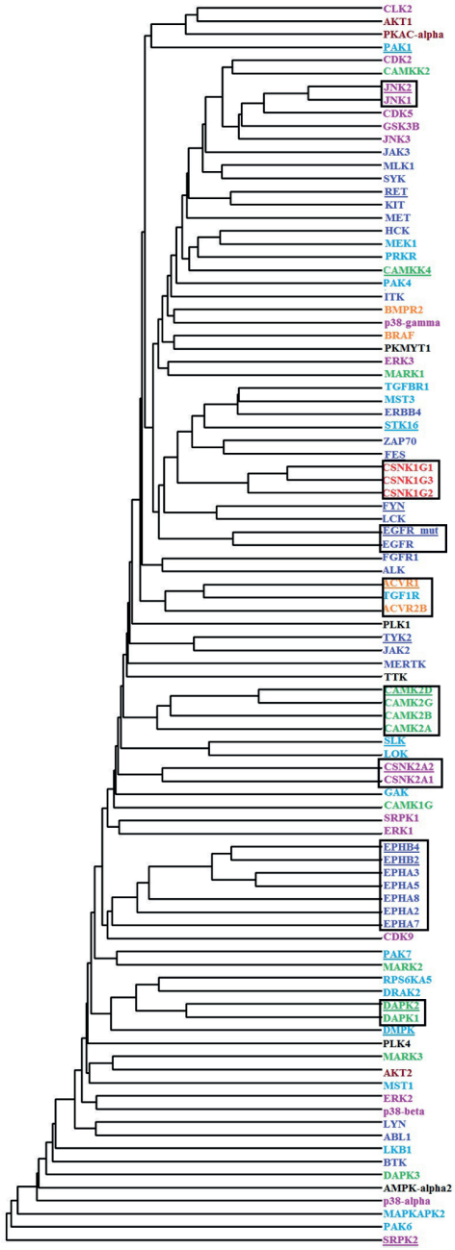

Figure 20. Hierarchical average linkage clustering of 95 kinases derived from the combined knowledge-based and WaterMap fields (Publication II). Different colours correspond to kinase subfamilies (AGC, CAMK, CK1, CMGC, OPK, STK, TK, TKL). Kinases included in the target prediction test set are underlined. Boxes in the cluster represent the kinase subtypes grouped together.

Activity space

In Publication I, the bioactivity spectrum of kinases covers a wide range of continuous values with pK_d / pK_is varying from 5 to 11 (Figure 21). Of the total of 951 observations considered for modelling, nearly 85% of the interaction data falls within the range of 5-8. With only a few highly potent compounds, the dataset is imbalanced in terms of pK_d / pK_i ranges, and it might create a challenge for future predictions of compounds with high potencies.

In Publication II, the dataset is compiled from multiple sources and includes different types of bioactivities. Therefore, classifying the data points as actives and inactives based on certain cut-offs seems to be an optimal choice for modelling (Table 7). Analysing the distributions of bioactivity ranges clearly shows that nearly 75% of the interaction data belongs to the inactive class (Figure 22). This data imbalance is likely to have an impact on the bioactivity predictions.

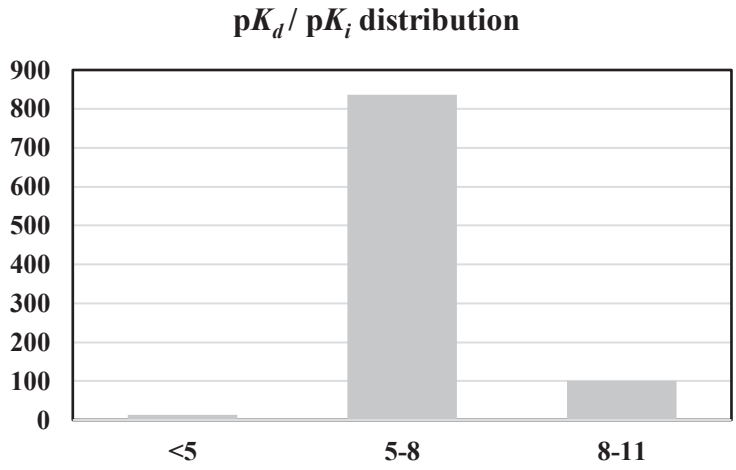

Figure 21. Distribution of pK_d / pK_i values of 80 inhibitors against 50 kinases in Publication I

Table 7. Distribution of actives and inactives in four different datasets used in Publication II

Dataset	Interaction data	Number of data points	Actives (cut-off)	Inactives (cut-off)
Ambit	pK_d	5491	1474 (>= 5)	4017 (<5)
Metz	pK_i	31667	9151 (>= 5)	22516(<5)
GSK	Inhibition% at 1 μM	17629	3288 (>=10%)	14341 (<10%)
Millipore	Residual activity	8400	1816 (<=50%)	6584 (>50%)

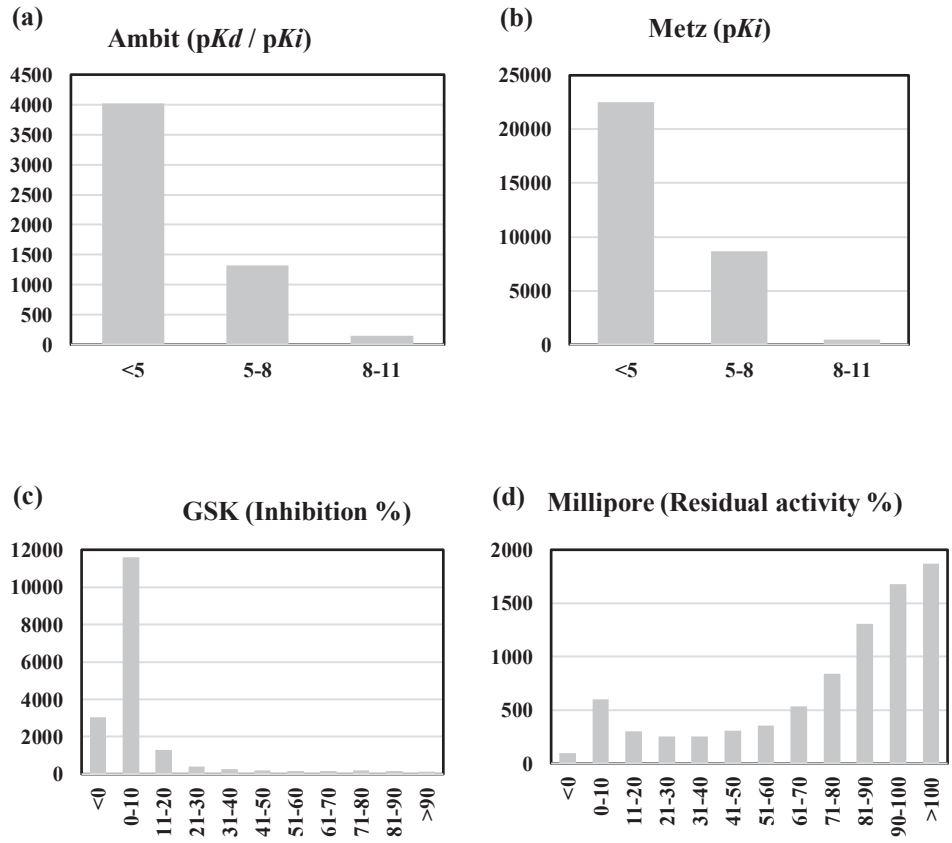

Figure 22. Distribution of data points (1572 inhibitors and 95 kinases) in four different datasets used in Publication II. **(a)** Ambit **(B)** Metz **(C)** GSK **(D)** Millipore

1.2 Serine protease dataset

Ligand space

The serine protease dataset of 5863 inhibitors is quite sparse in terms of activity. For many inhibitors, interaction data are available only for 1 or 2 proteases. K-means clustering of the inhibitors based on RDkit circular fingerprints resulted in 20 clusters (Figure 23) that had significant overlap in chemical space (for details, see Publication III). However, there are a few exceptions, with three clusters containing compounds with polycyclic ring systems linked to chlorine or fluorine and compounds with pyrazopyrimidines remaining distant from the rest.

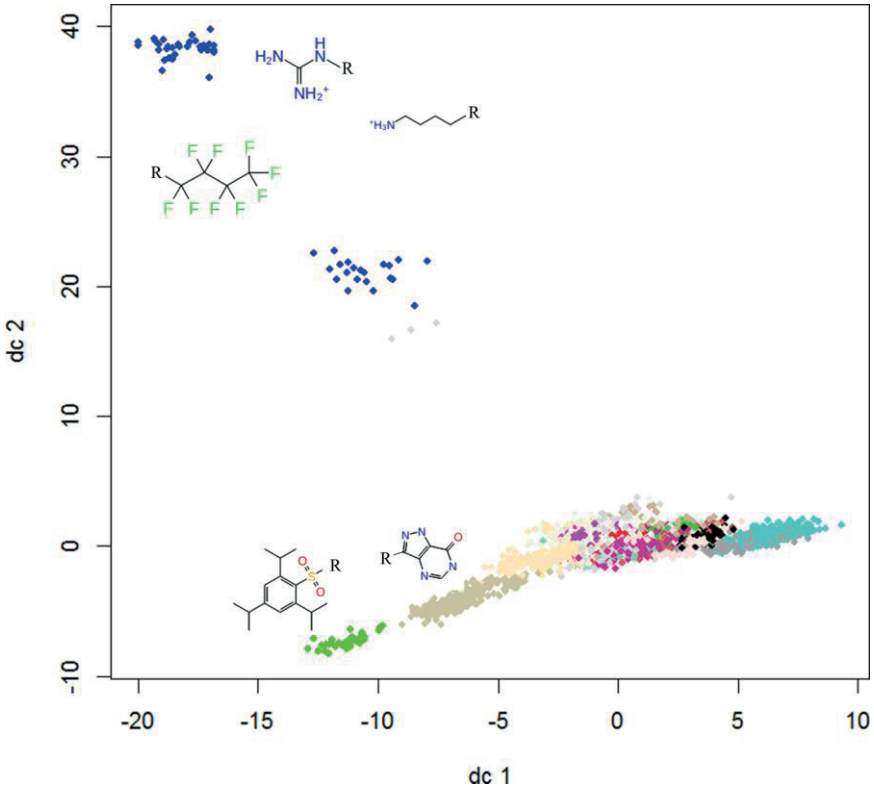

Figure 23. Distribution of the 20 compound clusters in two dimensions. Different colours correspond to different clusters. dc1 and dc2 refer to discriminant coordinates. *Figure reproduced from Supplementary Material of Publication III.*

Target space

Owing to the limited availability of interaction data, the number of serine proteases was restricted to 24 in this dataset. Knowledge-based and WaterMap fields of the serine proteases resulted in clusters where even some of the subgroups are placed apart, e.g. kallikrein 1, 3, 5 and 7 (Figure 24). The low similarity between the different proteases and the uneven distribution of data points with 70% representing either coagulation factor Xa or thrombin make it rather difficult to extrapolate in terms of target space.

Activity space

Unlike the kinase dataset, the bioactivities of serine proteases are more uniformly distributed with the presence of many highly potent and moderately potent compounds (Figure 25). Of the 7908 data points used for modelling, 10% have pK_is less than 5, with the majority of the interaction data representing kallikrein 7 and coagulation factor XII. Of the data points, 26% have pK_is ranging from 8 to 11. Most of these highly potent compounds target coagulation factor Xa, thrombin, plasma kallikrein and granzyme B.

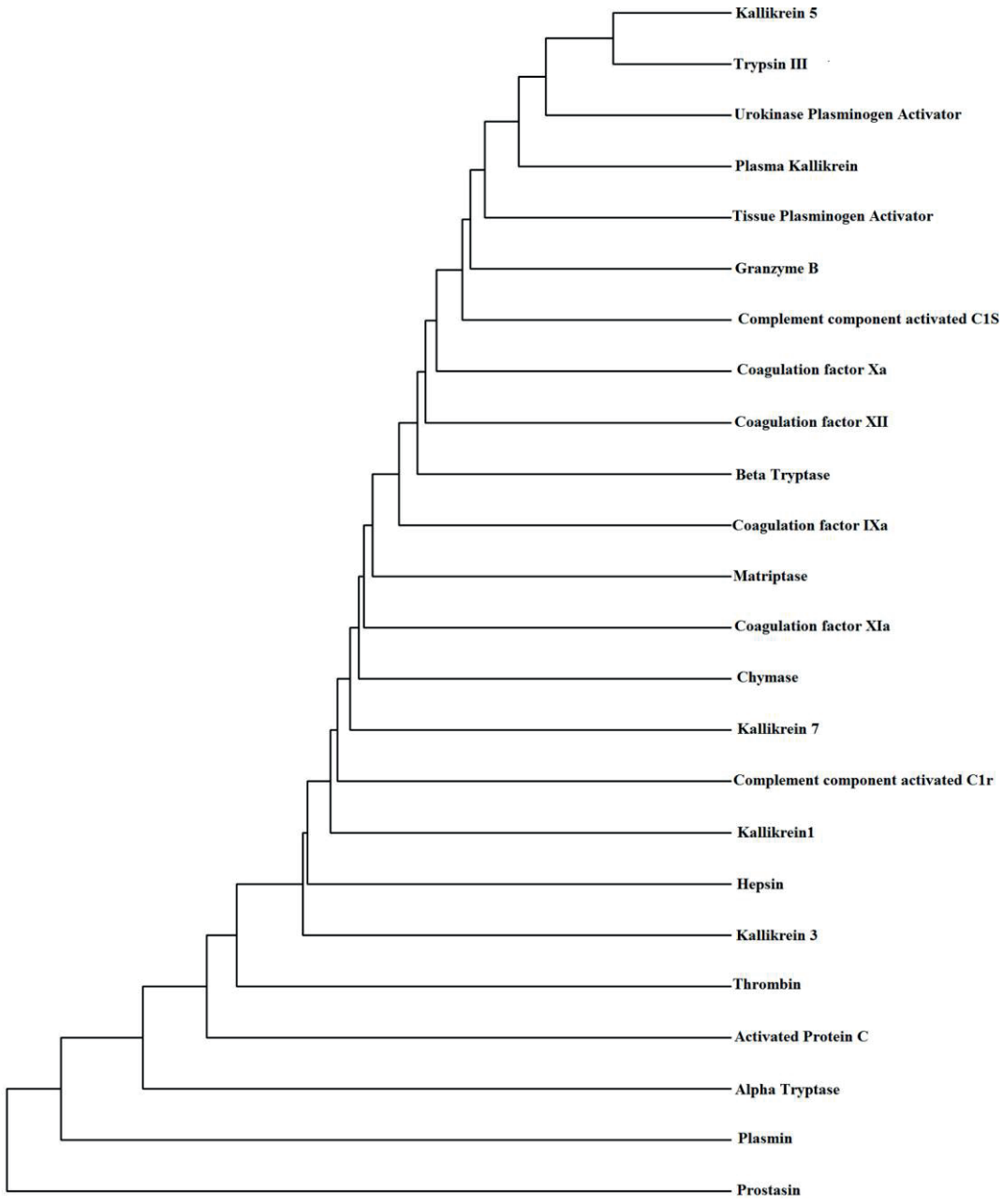

Figure 24. Hierarchical average linkage clustering of 24 serine proteases derived from Knowledge-based and WaterMap fields (Publication III).

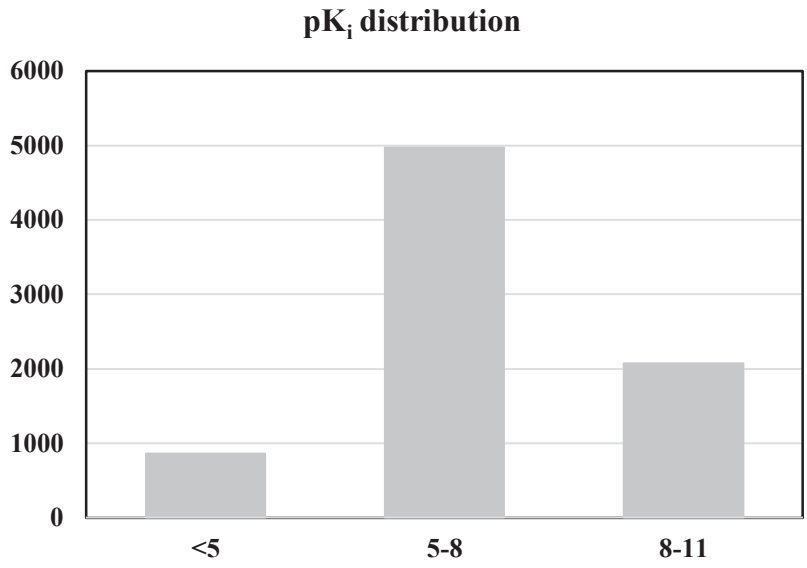

Figure 25. Distribution of pK_i values of 5863 inhibitors against 24 serine proteases in Publication III

2 Field-based PCM on continuous data

In Publication I, PCM models were built on a dataset of 50 kinases and 80 inhibitors with 951 known K_d or K_i values, retrieved from well-curated sources. Polar, lipophilic and WaterMap fields were used as protein descriptors together with the different 2D (Open Babel, Mold^2) and 3D (Volsurf) ligand descriptors. PLS models based on ligands' Open Babel fingerprints showed the best performance during both internal cross-validation (Q^2: 0.465; $RMSEP_{cv}$: 0.796) and external prediction ($RMSEP_{test}$: 0.800). An interesting observation is that the cross-terms, introduced to account for non-linearity in PLS models improved the model's overall performance by about 30%, with R^2 increasing from 0.336 in models without cross-terms to 0.662 in models with cross-terms. Also, the predictabilities (Q^2) of the models were improved with the inclusion of the cross-terms (Table 8).

In Publication III, PCM modelling was conducted on a dataset with pK_i values extracted for 24 serine proteases and 5863 inhibitors. Field-based descriptors were used for proteins, as in Publication I, whereas, ligands were described by RDkit fingerprints, MOE descriptors, 4-PFPs and GRIND descriptors. Both PLS and RF approaches were used for model training. Considering the model performances with respect to different ligand descriptors (Table 8), RF models based on RDkit fingerprints had the best performance, with R^2 and Q^2 as high as 0.957 and 0.737, respectively. Further, the RMSEP values from cross-validation and external test set predictions were below 1, making these models more robust. Overall, PLS models had slightly poorer performances than the RF models, regardless of the ligand descriptors used. However, the PLS models based on RDkit fingerprints performed reasonably well during the model training (R^2:0.670; Q^2:0.588). Their $RMSEP_{test}$ values were close to 1 and the

prediction errors were within the experimental error ranges, which is typically 1 log unit. Therefore, the models are sufficiently valid to be considered for further predictions and interpretations.

Table 8. Performances of field-based PCM models on continuous data.

Ligand descriptors	Method	Correlation (R^2)	Predictability (Q^2)	$RMSEP_{cv}$ [a]	$RMSEP_{test}$ [b]
		Kinase dataset			
Open Babel[c]	PLS	0.336	0.250	0.954	0.865
Open Babel	PLS	0.662	0.465	0.796	0.800
Mold[2]	PLS	0.539	0.445	0.811	0.716
Volsurf	PLS	0.520	0.400	0.842	0.947
		Serine protease dataset			
RDkit	PLS	0.670	0.588	1.024	1.006
	RF	0.957	0.737	0.799	0.810
MOE	PLS	0.505	0.428	1.219	1.129
	RF	0.961	0.703	0.857	0.840
4-PFP	PLS	0.543	0.438	1.229	1.136
	RF	0.928	0.566	1.025	0.990
GRIND	PLS	0.273	0.238	1.360	1.300
	RF	0.951	0.430	1.175	1.150

[a] Root mean square error of prediction resulting from cross-validation

[b] Root mean square error of prediction resulting from external test set predictions

[c] PCM models without cross-terms

3 Influence of protein descriptors in PCM modelling

Protein descriptors used in PCM models are likely to have an impact on the model's prediction performances. To verify this, performances of PLS models based on ligand's Open Babel fingerprints and different combinations of protein fields / sequence-based descriptors were compared in Publication I. Building PLS models using knowledge-based fields (polar and lipophilic) or WaterMap-derived fields separately did not result in significant variations regarding performances. Nevertheless, using these two descriptors in combination boosts the model's overall performance and lowers the prediction errors. On comparing the performances of field-based and sequence-based PCM models, models derived from sequence information had consistently lower predictabilities (Q^2: 0.32-0.37) and higher RMSEPs (0.84-0.90) than the field-based model (Q^2: 0.47; RMSEP: 0.80). Similar trends were observed regardless of the sequence descriptors used (Table 9).

In Publication III, the effect of including protein descriptors in PCM modelling was assessed by building RDkit fingerprint-based models on serine protease dataset, where the protein fields were excluded completely. Eliminating protein descriptors led to a significant drop in prediction performances (Table 9), with R^2 and Q^2 going as low as 0.5, relative to the models trained with protein fields (R^2 :0.957 ; Q^2: 0.737).

Unlike kinases in Publication I, a slightly different trend is observed with the sequence-based PCM models on serine proteases (Publication III). Sequence-based models perform nearly as well as field-based models in terms of internal and external validations (Table 9). However, their limited visual interpretability makes them less favourable than field-based models.

Table 9. Performances of PCM models based on protein fields and sequence-based descriptors.

Protein descriptors	Ligand descriptors	Method	R^{2a}	Q^{2b}	$RMSEP_{cv}$ [c]	$RMSEP_{test}$ [d]
Kinase dataset						
Polar, Lipophilic	Open Babel	PLS	0.614	0.437	0.816	0.823
WaterMap	Open Babel	PLS	0.599	0.395	0.846	0.836
Amino acid and dipeptide composition	Open Babel	PLS	0.517	0.365	0.867	0.836
Composition, transition and distribution	Open Babel	PLS	0.513	0.343	0.882	0.874
Pseudo amino acid composition and distribution	Open Babel	PLS	0.484	0.318	0.898	0.884
Serine protease dataset						
-	RDkit	RF	0.585	0.429	1.188	1.110
Amino acid and dipeptide composition	RDkit	RF	0.956	0.740	0.795	0.810
Composition, transition and distribution	RDkit	RF	0.956	0.742	0.792	0.810
Pseudo amino acid composition and distribution	RDkit	RF	0.955	0.739	0.795	0.810
Autocorrelation descriptors	RDkit	RF	0.956	0.739	0.796	0.810

[a] Correlation

[b] Predictability

[c] Root mean square error of prediction resulting from cross-validation

[d] Root mean square error of prediction resulting from external test set predictions

4 Field-based PCM on classification data

In Publication II, attempts were made to build a global classification model for kinases to test the robustness of the field-based PCM approaches, in terms of predictions. RF and SVM models were built on ligand and target prediction sets to classify the actives and inactives with Open Babel fingerprints, Mold^2 and 4-PFPs as ligand descriptors. Results of the best-performing RF models are reported in Table 10.

Table 10. Performance of field-based PCM models on classification data generated by using the Random Forest approach.

Ligand descriptors	Cross-validation			External prediction		
	Accuracy	MCC[a]	AUC[b]	Accuracy	MCC[a]	AUC[b]
Ligand Prediction dataset						
Open Babel	0.82	0.48	0.86	0.78	0.25	0.73
Mold^2	0.83	0.50	0.88	0.83	0.47	0.85
4-PFP[c]	0.83	0.49	0.87	0.81	0.42	0.83
Target Prediction dataset						
Open Babel	0.83	0.49	0.86	0.80	0.39	0.82
Mold^2	0.83	0.49	0.87	0.81	0.42	0.83
4-PFP[c]	0.83	0.49	0.87	0.81	0.41	0.82

[a] Matthews correlation coefficient

[b] Area Under the ROC curve

[c] 4-Point pharmacophoric fingerprints

Considering the internal cross-validation performances, the ligand and target prediction models have nearly the same performance, with AUCs of over 0.85 and MCC ranging from 0.48 to 0.50. With regard to external test set predictions, models based on Mold^2 and 4-PFPs are more efficient in predicting external ligands and targets (Table 10) than Open Babel fingerprint models. The low performance of the Open Babel models can be attributed to the simplistic functional groups description provided by the Open Babel fingerprints. These descriptors are less informative than the more complex Mold^2 and 4-PFPs, which capture additional information relevant for making good predictions.

Predicting the activities of test set kinases using the models based on target prediction datasets with AUCs above 0.87 (Table 10) shows that the target prediction PCM models can be used to estimate the polypharmacological profiles of the kinase inhibitors with reasonable accuracy.

Efforts to compare the classification performances based on different kinase families revealed no significant differences in AUCs, except that the AGC and OPK families have low sensitivities owing to the sparse distribution of activity points and the presence of few closer homologues. Furthermore, analysis of the prediction performances of data retrieved from different sources (Ambit, Metz, GSK and Millipore) suggested that prediction accuracies are independent of the data source and activity type (pK_d, inhibition%, residual activity). Prediction accuracies are reasonable provided that sufficient data points are available for each activity category.

5 Visual interpretation of PCM models

An important aspect of the field-based PCM modelling is the ability to acquire visual interpretation of both the protein and ligand features important for binding affinity and selectivity, simultaneously. Interpreting the PLS continuous models on kinase dataset (Publication I) revealed that the presence of polar, lipophilic and unstable water field points in close proximity to the well-conserved hinge motifs is generally important for the affinity of ABL1 kinase towards any ligand (Figure 26 a, b). Likewise, for Dasatinib to interact with any kinase, the presence of "hetero-N-nonbasic", "isothiourea" and "hetero-S" functional groups is relevant. To elucidate the features that contribute to the selective binding of dasatinib towards ABL1, the protein and ligand features should be considered as a combination and not as separate entities (Figure 26 c, d). Existence of lipophilic field points near the aryl chloride of dasatinib influences selectivity. Further, the unstable water field points in this region are expected to promote strong binding, as they are likely to be displaced during ligand binding. Additionally, the model suggests that the stable water field points near the hydroxy group of dasatinib might influence selectivity, probably by mediating the interactions of dasatinib with ABL1 binding site residues.

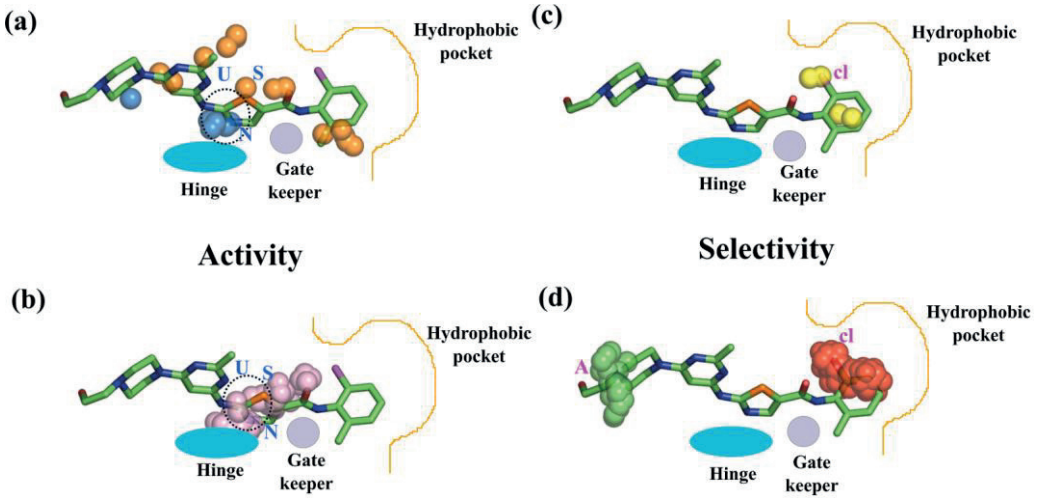

Figure 26. Protein field points and ligand functional groups relevant for the interactions of the inhibitor dasatinib with ABL1 kinase. (a) & (b) Features that influence binding affinity. Polar, lipophilic and unstable water field points are shown as slate, orange and pink spheres, respectively. (c) & (d) Features related to selectivity. Lipophilic, unstable and stable water field points are represented as yellow, red and green spheres, respectively. Ligand functional groups (from Open Babel) relevant for affinity (N: hetero-N-nonbasic, S: hetero-S, U: isothiourea) and selectivity (Cl: aryl chloride, A: primary alcohol) are indicated in blue and magenta, respectively. *Figure adapted with permission from (Subramanian, V.; Prusis, P.; Pietilä, L. O.; Xhaard, H.; Wohlfahrt, G. Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics. J. Chem. Inf. Model.* ***2013****, 53, 3021–3030). Copyright (2013) American Chemical Society.*

In Publication II, interpretation of the RF classification models provided an illustration of the protein and ligand features that make a compound active or inactive towards a specific kinase. Features relevant for the interactions of TAE-684 towards ALK and AKT2 kinase are presented in Figure 27. Polar and unstable water field points near the hinge motif serve as the affinity-promoting regions, together with the 4-PFP AAAR. In contrast to the unstable water field

points seen in the hydrophobic pocket of ALK kinase, the stable water field points in AKT2 kinase are expected to weaken ligand binding, thereby contributing to the low affinity of TAE-684 towards AKT2 kinase. Further, the interactions between the piperazine moiety of TAE-684 and E1210 in ALK kinase is most likely mediated by the stable water present in this region. On the other hand, in AKT2 kinase, a phenylalanine (F239) is present in the same position as glutamate (E1210) in ALK kinase, which leads to unfavourable interactions with piperazine, lowering affinity.

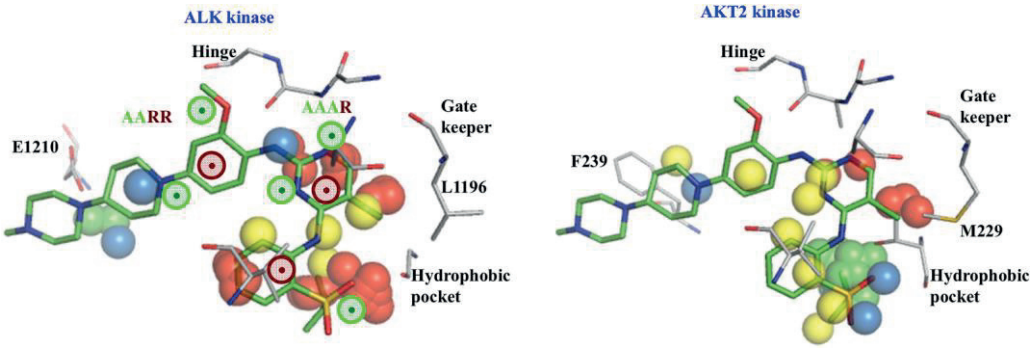

Figure 27. Protein field points and ligand pharmacophoric groups relevant for the strong binding of TAE-684 towards ALK kinase and weak affinity of TAE-684 towards AKT2 kinase. Polar, lipophilic, unstable and stable water field points that are likely to influence affinity are shown as blue, yellow, red and green spheres, respectively. 4-PFPs (AARR, AAAR) identified to be important for the interactions of TAE-684 and ALK kinase are represented as colored circles (A = H-acceptor (green); R = Aromatic ring (brown)). *Figure reproduced by permission of The Royal Society of Chemistry (Subramanian, V.; Prusis, P.; Xhaard, H.; Wohlfahrt, G. Predictive Proteochemometric Models for Kinases Derived from 3D Protein Field-Based Descriptors. Med. Chem. Commun.* ***2016****, 7, 1007–1015). http://pubs.rsc.org/en/content/articlelanding/2016/md/c5md00556f*

6 Applicability domain analysis

Models are considered useful provided that they can be successfully applied to predict a new compound or target space. In Publication I, the models had limited applicability because the training space includes only active representatives of 80 inhibitors and inactives are excluded from the modelling domain. Despite these limitations, when applied to a small test set of 25 kinase inhibitors, the $RMSEP_{test}$ obtained was close to 0.8. In Publication II, 64% of the test set compounds, whose 4-PFP-based Tanimoto similarities with the training set compounds exceeded 0.8, were predicted with more than 80% accuracy. Additional efforts to predict external targets revealed that the Euclidean distances calculated based on the protein field descriptors should be below 0.992 for reliable prediction of test set kinases. In Publication III, prediction errors of 84% of the test set compounds were below 1 provided that their RDkit-based Tanimoto similarities were above 0.7. Overall, in kinase and protease models, a few test set compounds were poorly predicted, despite their high similarities with the training set.

In addition to external test set predictions, the applicability domain was further explored by LOTO and LOCCO validations. LOTO validations on kinase (Publication I) and protease datasets (Publication III) resulted in average RMSEPs of 0.820 ± 0.022 and 1.302 ± 0.443, respectively. Even though the LOTO RMSEPs seem to be higher than the model's global RMSEPs, assessing the predictions of individual targets showed that the models can be extrapolated to novel targets provided that there are some close homologues whose activity space overlaps with the target to be predicted. In Publication III, LOCCO models with excluded compound clusters had a significant drop in performance. Further, the high RMSEPs (1.550 ± 0.269) resulting from LOCCO validations

revealed that it is more demanding to extrapolate to novel compound space for the serine protease dataset.

Additional Unpublished Results

1 PCM models on superimposed protein fields and Zernike descriptors: a comparative study

As protein fields are sensitive to errors in protein structural alignments, there is a need to investigate the extent to which these errors can influence a model's prediction power. PCM models based on protein fields and alignment-independent Zernike descriptors (N=10) were built using RDkit fingerprints as ligand descriptors (Table 11). RF classification and RF regression techniques were used to model the bioactivities of kinases and serine proteases, respectively.

Table 11. RF-based PCM models on superimposed protein fields and Zernike descriptors**.**

Protein descriptors	Ligand descriptors	Cross-validation (MCC^{a} / Q^{2b})	External Prediction (MCC / R^{2c}_{test})
Kinase dataset (classification) [a]			
Protein fields	RDkit	0.52	0.48
Zernike (N=10)		0.48	0.48
Serine protease dataset (Continuous) [b]			
Protein fields	RDkit	0.67	0.71
Zernike (N=10)		0.67	0.71

[a] Matthews correlation coefficient (MCC) values reported for classification models

[b] Predictabilities (Q^2) resulting from cross-validation

[c] Correlation (R^2_{test}) resulting from external test set predictions reported for continuous models

Comparing the model performances based on protein fields and Zernike descriptors, the predictions seem to be nearly the same during both internal cross-validation and external predictions (Table 11). Similar trends were observed for both kinase and serine protease datasets, suggesting that the structural alignments used for field calculations are reasonably good and the global performances of the models are not affected.

An in-depth analysis was conducted to examine the prediction differences of the individual kinases and serine proteases. When the predictions made by field-based and Zernike descriptor-based PCM models for the individual observations were compared (Table 12), the correlation of 0.77 for kinases and 0.92 for serine proteases confirms the robustness of field-based and Zernike descriptor models, in terms of predictions, with a few exceptions. On analysis of the individual predictions of kinases, predictions based on Zernike descriptors were found to improve by at least 10% for five kinases. It is probable that these kinases had poor superimposition, which in turn contributed to shifts in field positions and hence the predictions were affected in field-based models. On the other hand, predictions based on Zernike descriptors deteriorated by more than 10% for nearly 15 kinases. This drop in predictions could be attributed to the information loss that occurs during the transformation of fields to alignment-independent descriptors. A slightly different scenario was observed in serine protease models, as the predictions based on Zernike descriptors improved by 10% for only one protease and reduced by 10% for three proteases. Overall, the protease models are more stable with respect to predictions.

Nevertheless, the results obtained for Zernike descriptors are preliminary. A detailed analysis concerning the shifts in protein structure alignments and subsequently field point positions is necessary to draw further conclusions.

Table 12. Comparison of prediction performances of protein fields and Zernike descriptor-based (N=10) PCM models.

Protein target	Performance measure	$R^2_{(fields,Zernike)}$ [a]	Predictions of individual proteins	
			Improved by atleast 10% [d]	Reduced by 10% or more [e]
Kinases	MCC [b]	0.77	DAPK1, GAK, EPHA7, CAMKK2, ERK1	MAPKAPK2, GSK3B, DRAK2, ALK, EPHA3, PAK7, EPHB4, PRKR, JNK1, CSNK1G3, CAMK4, RET, MEK1, CSNK1G1, EFGR_mut
Serine proteases	R^2_{test} [c]	0.92	APC	FXIIa, KLK5, FIXa

[a] Correlation between predictions obtained from field-based and Zernike descriptor-based PCM models

[b] Matthews correlation coefficient

[c] Correlation resulting from external test set predictions

[d] Kinases whose MCC values and serine proteases whose R^2_{test} values improved by atleast 10%

[e] Kinases whose MCC values and serine proteases whose R^2_{test} values reduced by 10% or more

2 Impact of expansion order in Zernike descriptor-based PCM modelling

It is often a challenge to decide the extent to which Zernike polynomials should be expanded. Expanding the Zernike polynomials increases the level of description and captures more information from the protein fields. As there are no standard methods, the order of expansion is decided on a trial and error basis. Therefore, the Zernike descriptors were calculated by assigning a series of N values (N=3, 5, 10, 20, 30, 40, 50). Performances of PCM models based on Zernike descriptors (Table 13) remained nearly the same up to the order of 40 for both internal and external validations (MCC: 0.46 - 0.50). Expanding the Zernike polynomials further by setting N to 50 led to a decrease in both internal and external MCCs, suggesting that these descriptors tend to add random noise to the models. The model performances neither increase nor decrease significantly by expanding Zernike polynomials beyond the order of 10. As inclusion of irrelevant descriptors is likely to result in overfitted models, the optimal value of N can be chosen as 10.

Table 13. Impact of expansion order in Zernike descriptors based PCM modelling on kinase dataset.

Order of expansion (N)	No. of Zernike descriptors	Cross-validation (MCC)[a]	External Prediction (MCC)[a]
3	6	0.49	0.47
5	12	0.49	0.48
10	36	0.48	0.48
20	121	0.50	0.48
30	256	0.49	0.47
40	441	0.49	0.46
50	676	0.47	0.44

[a]Matthews correlation coefficient

Discussion

1 Impact of data quality and coverage in PCM modelling

Information-rich databases like ChEMBL (Bento et al., 2014), BindingDB (Liu et al., 2007), PDBbind (Wang et al., 2005) and PubChem (Kim et al., 2016) offer the possibility to conduct large-scale data analysis and generate QSAR / PCM models to elucidate structure activity relationships of specific targets or target families. Nevertheless, the inconsistencies found in these databases resulting from multiple measurements by different assays for the same protein-ligand pair, incorrect structures, erroneous measurement units and incorrect values raise concern about the data quality and their use for generating empirical models (Williams and Ekins, 2011, Tiikkainen et al., 2012).

The reliability of the models based on experimental data collected from various sources and different assay conditions is often questionable. Yet, another concern is the reproducibility. Experimental measurements made for a specific protein-ligand pair by two different laboratories can have significant variations, as reported in Publication I. Lack of standard operating protocols often makes it difficult to compare the experimental results. The reproducibility issues can impose limitations on model quality and their usefulness for making further predictions. Despite these limitations, the wealth of data available in public databases is frequently exploited in QSAR and PCM modelling.

Data collection suitable for model generation is a crucial step and it should be done with caution. Building models exclusively by utilizing the data from well-curated sources as in Publication I, can limit the applicability domain of the models. Extracting data from various sources after applying certain filters (Kramer et al., 2012) and ensuring checks for data quality (Tiikkainen et al.,

2012) would help to establish predictive modelling. An alternative option would be to use manually curated datasets, which has been thoroughly investigated in terms of data quality and coverage. Such large datasets are frequently not available for many targets. However, the recently published proteochemometric models on kinases with 356908 data points (Christmann-Franck et al., 2016) and the ligand-based activity prediction models generated for a set of 280 kinases (Merget et al., 2016) serve as examples of large-scale predictive modelling.

Data coverage is yet another important aspect in PCM modelling. It is often demanding to generate datasets with a complete bioactivity matrix. In Publications II and III, the sparse activity matrix led to a rise in prediction errors for a few test set compounds that had good overlap with the training set descriptor space. The model's robustness increases with data coverage, which in turn limits the application of proteochemometric approaches for some of the less extensively studied targets, with meager activity data.

2 Availability of structures for field-based PCM modelling

Protein fields used for PCM modelling are highly dependent on X-ray structures. It is often challenging to find crystal structures with high resolution and completeness for all of the targets included in PCM modelling. Lack of crystal structures limits the use of field-based PCM for these targets. This is true for GPCRs with only a few solved crystal structures. Using homology models for field calculations could be an alternative. However, homology models have their own limitations, as the model quality depends on sequence identity with the template and correctness of the alignment. Homology models are frequently mere reflections of template structures and fail to account for

protein flexibility and structurally more different areas, which in turn adds to errors during field calculations and subsequent PCM modelling.

3 DFG-in (active) and DFG-out (inactive) conformations of kinases

As protein fields are influenced by conformational flexibilities of the target protein, it is necessary to separate the active and inactive conformations of kinases prior to field calculations. It is easy to distinguish between DFG-in and DFG-out kinase structures by means of manual inspection, considering the orientation of phenylalanine in the DFG motif. However, the experimental data lack clear-cut information regarding the binding modes of inhibitors, which makes it difficult to discriminate DFG-in and DFG-out inhibitors. PCM models in Publications I and II are based on the DFG-in conformations of kinases, owing to the presence of abundant structural data for active conformations. The selection of DFG-in like inhibitors in Publication I is based on the information available in the literature (Karaman et al., 2008; Uitdehaag and Zaman, 2011) and a similarity search conducted against the known references. It is probable that there are more DFG-out-like inhibitors than the ones described in the literature. Combining protein fields generated from DFG-in conformations together with the experimental data of DFG-out inhibitors could lead to additional sources of error in PCM modelling.

4 Influence of ligand conformations in PCM modelling

In Publication II, the influence of 3D conformation generation in 4-PFP descriptor calculations and subsequent PCM modelling was investigated. The results show that neither the 4-PFP calculations nor the model's overall performances are affected by the ligand conformations used. Nevertheless, the

probability of predicting active ligands correctly improved by using the lowest energy conformation, when compared to the other higher energy conformations.

In Publication III, models based on 3D GRIND descriptors had the worst performance, despite using the lowest energy conformation. However, the 4-PFP models on serine proteases performed better than the GRIND models, further confirming that 4-PFPs are less sensitive to the conformations used. In case of GRIND models, the starting conformations used for descriptor calculations have an impact on the model performance (Caron et al., 2007). Therefore, it is not straightforward to generalize that PCM models are not influenced by 3D ligand conformations used for descriptor calculations. It should rather be considered on a case-by-case basis, depending on the descriptor type and the flexibility of the ligands.

5 QSAR versus Field-based PCM

In Publications I and III, field-based PCM is shown to be clearly advantageous over traditional QSAR methods. In Publication I, reliable QSAR models were obtained for only 44% of the kinase targets with a wide range of activity values, which serves as evidence that ligand descriptors alone cannot capture all of the features relevant for binding. Moreover, to understand selectivity with respect to different targets, comparison of multiple QSAR models is required, which is frequently not feasible considering the data availability for individual targets. Even though some targets have limited data, field-based PCM models that take advantage of the protein structural information and ligand description of multiple targets and ligands allow investigation of the target-ligand interaction space in greater depth. However, the field-based PCM models have their own

limitations such as issues with data coverage, availability of crystal structures and poor protein superimposition affecting field calculations.

6 Predicting mutants in field-based PCM studies

Previously conducted studies on kinases and proteases have shown that mutants can cause drug resistance, thereby revealing the need for the design of inhibitors targeting these mutants (Gorre et al., 2001; Kovalevsky et al., 2006). Sequence-based PCM studies conducted on HIV protease mutants serve as an example for using PCM to predict the bioactivities of mutants (Lapins et al., 2008; Van Westen et al., 2013). However, the field-based PCM models reported in our studies have limitations in predicting the activities of mutant types. The limited structural and activity data for mutant structures have restricted our studies mostly to wild types. In kinase PCM modelling, EGFR mutant T719S was included in the external test set. Prediction accuracy of this mutant was quite limited due to the absence of similar mutant representatives in the training set. Nevertheless, availability of more structural and activity data would support field-based PCM modelling of mutants in the future.

7 Complementarity of docking and field-based PCM approaches

Docking, a commonly used structure-based approach in drug design to study the interactions between proteins and ligands provides 3D illustrations of the features that affect binding. Although, docking aims to provide visual interpretation as that of field-based PCM, there is a need to conduct several docking experiments to estimate selectivity. The problems inherent to docking, such as difficulties in generating the right binding pose and poor abilities to rank binding affinities, can sum up to large errors in identifying the compounds that bind selectively. Nevertheless, docking does not require large-scale

experimental data for a set of compounds measured against a panel of targets, unlike PCM modelling. Overall, docking and field-based methods complement each other, with neither superior to the other.

Conclusions and Perspectives

The current study is dedicated to the development of field-based proteochemometrics, tackling limitations in visual interpretability, a frequent issue encountered in sequence-based PCM models. In Publication I, PCM studies on kinases demonstrated the first successful application of field-based PCM to generate visually interpretable models. Possibilities to visually inspect the features that affect the selective binding of a drug towards a target are highly beneficial in suggesting suitable chemical modifications and designing compounds with improved efficacy. Subsequent studies on kinases in Publication II proved that the protein field-based descriptors also have the potential to generate predictive PCM models. The highlight of this study is the estimation of potential polypharmacology, which could be valuable for the design of new kinase inhibitors. Further, studies on serine proteases in Publication III provided an example that field-based PCM can be applied to any target family with well-characterized 3D structures and adequate experimental data.

In summary, PCM models can be used to investigate the target-ligand interaction space in greater depth, hence being more advantageous than the traditional QSAR approaches. Field-based PCM models are either more predictive, as in kinases, or as predictive as sequence-based models in serine proteases with the benefit of visual interpretation. The ability to illustrate molecular interactions similar to structure-based approaches like docking together with the possibilities to extrapolate to novel chemical and target space makes field-based PCM a promising approach.

Field-based PCM studies conducted so far provide clear evidence for the usefulness of the method in drug design, and new avenues for further development are likely to emerge. The protein fields used for PCM modelling are generated from a single protein structure. The different binding modes of the ligands, their flexibilities and the conformational changes of the protein induced during ligand binding influence the field calculations, and this might have an impact on the predictions of novel ligands in PCM modelling. Generating fields based on an ensemble of protein conformations could solve this problem to some extent. Currently, the model interpretation procedure is highly demanding, as it involves extensive manual work. Selectivity interpretation is restricted to the top 5 or 10 cross-terms. Automating the field interpretations would enable interpretation of all cross-terms, thereby providing a more thorough understanding of the selectivity-related features. Also, there is a need to investigate PCM modelling based on alignment-independent Zernike descriptors in greater depth. Conducting further comprehensive studies by introducing artificial shifts in alignments and analysing their effects on field calculations and Zernike descriptors could shed light on the prediction issues in PCM modelling, probably caused by the fields calculated from misaligned structures.

References

Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, 2817–2831.

Adnane, L.; Trail, P. A.; Taylor, I.; Wilhelm, S. M. Sorafenib (BAY 43-9006, Nexavar), a Dual-Action Inhibitor That Targets RAF/MEK/ERK Pathway in Tumor Cells and Tyrosine Kinases VEGFR/PDGFR in Tumor Vasculature. *Methods Enzymol.* **2005**, *407* (5), 597–612.

Ain, Q. U.; Méndez-Lucio, O.; Ciriano, I. C.; Malliavin, T.; van Westen, G. J. P.; Bender, A. Modelling Ligand Selectivity of Serine Proteases Using Integrative Proteochemometric Approaches Improves Model Performance and Allows the Multi-Target Dependent Interpretation of Features. *Integr. Biol.* **2014**, *6* (11), 1023–1033.

Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55* (7), 1316–1322.

Auvray, P.; Nativelle, C.; Bureau, R.; Dallemagne, P.; Séralini, G. E.; Sourdaine, P. Study of Substrate Specificity of Human Aromatase by Site Directed Mutagenesis. *Eur J Biochem* **2002**, *269* (5), 1393–1405.

Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (18), 10037–10041.

Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/comparing Proteins and Ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47* (2), 279–294.

Barouch-Bentov, R. Mechanisms of Drug-Resistance in Kinases. *Expert Opin. Investig. Drugs* **2012**, *20* (2), 153–208.

Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res* **2014**, *42* (D1), D1083–D1090.

Berthold, M.R.; Cebron, N.; Dill, F.; Fatta, G.D.; Gabriel, T.R.; Georg,F.; Meinl, T.; Ohl, P.; Sieb, C.; Wiswedel, B. KNIME: The Konstanz Information Miner, In Studies in Classification, Data Analysis, and Knowledge Organization; Springer: Germany, 2007.

Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W. Thermodynamic Analysis of Water Molecules at the Surface of Proteins and Applications to Binding Site Prediction and Characterization. *Proteins: Structure, Function and Bioinformatics* **2012**, *80* (3), 871–883.

Bhongade, B. A.; Gouripur, V. V.; Gadad, A. K. 3D-QSAR CoMFA Studies on Trypsin-like Serine Protease Inhibitors: A Comparative Selectivity Analysis. *Bioorg. Med. Chem.* **2005**, *13* (8), 2773–2782.

Bikker, J. a.; Brooijmans, N.; Brooijmans, N.; Wissner, A.; Wissner, A.; Mansour, T. S.; Mansour, T. S. Kinase Domain Mutations in Cancer: Implications for Small Molecule Drug Design Strategies. *J. Med. Chem.* **2009**, *52* (6), 1493–1509.

Boyle, N.M.O.; et al. Open Babel: An open chemical toolbox. *J. Chem. Inf.* **2011**, 3 – 33.

Breiman, L. E. O. Random Forests. *Machine Learning* **2001**, 45, 5–32.

Breitenlechner, C.; Gassel, M.; Hidaka, H.; Kinzel, V.; Huber, R.; Engh, R. A.; Bossemeyer, D. Protein Kinase a in Complex with Rho-Kinase Inhibitors Y-27632, Fasudil, and H-1152p: Structural Basis of Selectivity. *Structure* **2003**, *11*, 1595−1607.

Brown, A. C.; Fraser, T. R. On the Connection between Chemical Constitution and Physiological Action; with Special Reference to the Physiological Action of the Salts of the Ammonium Bases Derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *J. Anat. Physiol.* **1868**, *2* (2), 224–242.

Bruyn, T. De; Westen, G. J. P. Van; Ijzerman, A. P.; Stieger, B.; Witte, P. De; Augustijns, P. F.; Annaert, P. P. Structure-Based Identification of OATP1B1 / 3 Inhibitors. *Mol. Pharmacol.* **2013**, *83*, 1257–1267.

Canvas, version 2.0, Schrödinger, LLC, New York, NY, 2014.

Cappel, D.; Dixon, S. L.; Sherman, W.; Duan, J. Exploring Conformational Search Protocols for Ligand-Based Virtual Screening and 3-D QSAR Modeling. *J. Comput. Aided Mol. Des.* **2015**, *29* (2), 165–182.

Cargnello, M.; Roux, P. P. Activation and Function of the MAPKs and Their Substrates, the MAPK-Activated Protein Kinases. *Microbiol. Mol. Biol. Rev.* **2011**, *75* (1), 50–83.

Caron, G.; Ermondi, G. Influence of Conformation on GRIND-Based Three-Dimensional Quantitative Structure - Activity Relationship (3D-QSAR). **2007**, *3*, 5039–5042.

Carta, F.; Vullo, D.; Maresca, A.; Scozzafava, A.; Supuran, C. T. Mono-/dihydroxybenzoic Acid Esters and Phenol Pyridinium Derivatives as Inhibitors of the Mammalian Carbonic Anhydrase Isoforms I, II, VII, IX, XII and XIV. *Bioorg. Med. Chem.* **2013**, *21* (6), 1564–1569.

Cera, E. Di. Serine Proteases. *IUBMB Life* **2009**, *61* (5), 510–515.

Cheng, A. C.; Eksterowicz, J.; Geuns-Meyer, S.; Sun, Y. Analysis of Kinase Inhibitor Selectivity Using a Thermodynamics-Based Partition Index. *J. Med. Chem.* **2010**, *53* (11), 4502–4510.

Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'Min, V.

E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010.

Christmann-Franck, S.; van Westen, G. J. P.; Papadatos, G.; Beltran Escudie, F.; Roberts, A.; Overington, J. P.; Domine, D. Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* **2016**, 56 (9), 1654-1675.

Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960,** 20 (1), 37–46.

Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, 20, 273–297.

Cortes Ciriano, I.; van Westen, G. J. P.; Lenselink, E. B.; Murrell, D. S.; Bender, A.; Malliavin, T. Proteochemometrics Modeling in a Bayesian Framework. *J. Cheminf.* **2014**, 1–16.

Cortés-Ciriano, I.; Ain, Q. U.; Subramanian, V.; Lenselink, E. B.; Méndez-Lucio, O.; IJzerman, A. P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T. E.; van Westen, G. J. P.; Bender, A. Polypharmacology Modelling Using Proteochemometrics (PCM): Recent Methodological Developments, Applications to Target Families, and Future Prospects. *Med. Chem. Commun.* **2015**, *6* (1), 24–50.

Cortés-Ciriano, I.; Van Westen, G. J. P.; Bouvier, G.; Nilges, M.; Overington, J. P.; Bender, A.; Malliavin, T. E. Improved Large-Scale Prediction of Growth Inhibition Patterns Using the NCI60 Cancer Cell Line Panel. *Bioinformatics* **2015**, *32* (1), 85–95.

Cortes-Ciriano, I.; Murrell, D. S.; Van Westen, G. J.; Bender, A.; Malliavin, T. E. Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling. *J. Cheminform.* **2015**, *7* (1), 1–18.

Cortés-Ciriano, I.; Bender, A.; Malliavin, T. Prediction of PARP Inhibition with Proteochemometric Modelling and Conformal Prediction. *Mol. Inform.* **2015**, *34*.

Cramer, R.D.; Patterson, D.E.; Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988** ,*110* (18), 5959-5967.

Cruciani, G.; Crivori, P.; Carrupt, P.A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: the Volsurf approach. *Theochem - J. Mol. Struc.* **2000**, 503, 17–30.

Damale, M.; Harke, S.; Kalam Khan, F.; Shinde, D.; Sangshetti, J. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. *Mini-Rev. Med. Chem.* **2014**, *14* (1), 35–55.

Daub, H.; Specht, K.; Ullrich, a. Strategies to Overcome Resistance to Targeted Protein Kinase Inhibitors. *Nat Rev Drug Discov.* **2004**, *3* (12), 1001–1010.

Davis, M.I.; Hunt, J.P.; Herrgard, S.; Ciceri, P.; Wodicka, L.M.; Pallares, G.; Hocker, M.; Treiber, D.K.; Zarrinkar, P.P. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnol.* **2011**, *29* (11), 1046 – 1051.

Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How Not to Develop a Quantitative Structure-Activity or Structure-Property Relationship (QSAR/QSPR). *SAR QSAR Environ Res* **2009**, *20*, 241–266.

De Bruyn, T.; van Westen, G.J.; Ijzerman, A.P.; Stieger, B.; de Witte, P.; Augustijns, P.F.; Annaert, P.P. Structure-based identification of OATP1B1/3 inhibitors. *Mol Pharmacol.* **2013**, *83*(6), 1257-1267.

Dimitrov, I.; Garnev, P.; Flower, D.R.; Doytchinova, I. Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis. .*Eur J Med Chem.* **2010**, *45* (1), 236-243.

Dimitrov, I.; Garnev, P.; Flower, D.R.; Doytchinova, I. EpiTOP – a proteochemometric tool for MHC class II binding prediction. *Bioinformatics* **2010**, *26* (16), 2066-2068.

Dimitrov, I.; Doytchinova, I. Peptide Binding Prediction to Five Most Frequent HLA-DQ Proteins - a Proteochemometric Approach. *Mol. Inform.* **2015**, *34* (6–7), 467–476.

Drag, M.; Salvesen, G. S. Emerging Principles in Protease-Based Drug Discovery. *Nat. Rev. Drug Discov.* **2010**, *9* (9), 690–701.

Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods. *J. Mol. Graph. Model.* **2010**, *29* (2), 157–170.

Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inform. Comput. Sci.* **2002**, *42* (6), 1273–1280.

Ehrmann, M.; Clausen, T. Proteolysis as a Regulatory Mechanism. *Annu. Rev. Genet.* **2004**, *38*, 709–724.

Eriksson, L.; Johansson, E.; Wold, S. Quantitative structure-activity relationship model validation, In Quantitative Structure-Activity Relationships in Environmental Sciences – VII. Chen, F.; Scuurmann, G., Eds.; SETAC: Pensacola, **1997**, pp 381- 397.

Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS). Umetrics AB.Umea.**1999**.

Fernandez, M.; Ahmad, S.; Sarai, A. Proteochemometric Recognition of Stable Kinase Inhibition Complexes Using Topological Autocorrelation and Support Vector Machines. *J. Chem. Inf. Model*. **2010**, *50* (6), 1179 – 1188.

Fleuren, E. D. G.; Zhang, L.; Wu, J.; Daly, R. J. The Kinome "at Large" in Cancer. *Nat. Rev. Cancer* **2016**, *16* (2), 83–98.

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), 1189–1204.

Fujita, T.; Winkler, D. A. Understanding the Roles of The "two QSARs." *J. Chem. Inf. Model.* **2016**, *56* (2), 269–274.

Gao, J.; Huang, Q.; Wu, D.; Zhang, Q.; Zhang, Y.; Chen, T.; Liu, Q.; Zhu, R.; Cao, Z.; He, Y. Study on Human GPCR-Inhibitor Interactions by Proteochemometric Modeling. *Gene* **2013**, *518* (1), 124–131.

Geladi, P.; Kowalski, B. R. Partial Least Squares Regression: A Tutorial. *Analytica Chimica Acta* **1986**, 185, 1 -17.

Georgiev, A. G. Interpretable Numerical Descriptors of Amino Acid Space. *J. Comp. Biol.* **2009**, *16* (5), 703–723.

Gettins, P. G. W. Serpin Structure, Mechanism, and Function. *Chem. Rev.* **2002**, *102* (12), 4751–4803.

Glinca, S.; Klebe, G. Cavities Tell More than Sequences: Exploring Functional Relationships of Proteases via Binding Pockets. *J. Chem. Inf. Model.* **2013**, *53* (8), 2082–2092.

Golbraikh, A.; Tropsha, A. Beware of Q 2 ! *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

Golbraikh, A.; Wang, X.S.; Zhu, H.; Tropsha, A. Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment, In Handbook of Computational Chemistry; Springer-Verlag, **2012**; pp.1-36.

Gorre, M. E.; Mohammed, M.; Ellwood, K.; Hsu, N.; Paquette, R.; Rao, P. N.; Sawyers, C. L. Clinical Resistance to STI-571 Cancer Therapy Caused by BCR-ABL Gene Mutation or Amplification. *Science* **2001**, *293* (5531), 876–880.

Graczyk, P. P. Gini Coefficient: A New Way to Express Selectivity of Kinase Inhibitors against a Family of Kinases. *J. Med. Chem.* **2007**, *50* (23), 5773–5779.

Guimarães, M. C.; Duarte, M. H.; Silla, J. M.; Freitas, M. P. Is Conformation a Fundamental Descriptor in QSAR? A Case for Halogenated Anesthetics. *Beilstein J. Org. Chem.* **2016**, *12*, 760–768.

Guo, Z.; Chen, B. Y. Predicting Protein-Ligand Binding Specificity Based on Ensemble Clustering. *Proceedings (IEEE Int Conf Bioinformatics Biomed)* **2015**, 1239–1245.

Hanks, S. K.; Quinn, A. M.; Hunter, T. The Kinase Family : Conserved Protein Phylogeny Features and Deduced Domains of the Catalytic. *Science* **2013**, *241* (4861), 42–52.

Hansch, C.; Fujita, T. *p*-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.*, **1964**, *86* (8), 1616–1626.

Hedstrom, L. Serine Protease Mechanism and Specificity Serine Protease Mechanism and Specificity. *Chem. Rev.* **2002**, *102* (12), 4501–4524.

Hong, H.; et al. Mold2, Molecular Descriptors from 2D structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2009**, *48* (7), 1337 – 1344.

Hoppe, C.; Steinbeck, C.; Wohlfahrt, G. Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials. *J. Mol. Graph. Model* **2006**, *24* (5), 328 – 340.

Huang, D.; Zhou, T.; Lafleur, K.; Nevado, C.; Caflisch, A. Kinase Selectivity Potential for Inhibitors Targeting the ATP Binding Site: A Network Analysis. *Bioinformatics* **2010**, *26* (2), 198–204.

Huang, Q.; Jin, H.; Liu, Q.; Wu, Q.; Kang, H.; Cao, Z.; Zhu, R. Proteochemometric Modeling of the Bioactivity Spectra of HIV-1 Protease Inhibitors by Introducing Protein-Ligand Interaction Fingerprint. *PLoS ONE* **2012**, *7* (7), 1–8.

Huggins, D. J.; Sherman, W.; Tidor, B. Rational Approaches to Improving Selectivity in Drug Design. *J. Med. Chem.* **2012**, *55* (4), 1424–1444.

Huse, M.; Kuriyan, J. The Conformational Plasticity of Protein Kinases. *Cell* **2002**, *109* (3), 275–282.

Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52* (7), 1757–1768.

Janciauskiene, S. Conformational Properties of Serine Proteinase Inhibitors (Serpins) Confer Multiple Pathophysiological Roles. *BBA-Mol. Basis Dis.* **2001**, *1535* (3), 221–235.

Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ALTA Altern. To Lab. Anim.* **2005**, *33* (5), 445–459.

Johnson, L. Protein Kinases and Their Therapeutic Exploitation. *Biochem. Soc. Trans*. **2007**, *35*, 7–11.

Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008**, *48*, 25–26.

Johnson, L. N. The Regulation of Protein Phosphorylation. *Biochem. Soc. Trans*. **2009**, *37*, 627–641.

Kao, Y.C.; Cam, L.; Laughton, C. A.; Zhou, D.; Chen, S. Binding Characteristics of Seven Inhibitors of Human Aromatase : A Site-Directed Mutagenesis Study1. Cancer Res. **1996**, *56* (15), 3451-3460.

Karaman, M.W.; Herrgard, S.; Treiber, D.K.; Gallant, P.; Atteridge, C.E.; Campbell, B.T.; Chan, K.W.; Ciceri, P.; Davis, M.I.; Edeen, P.T.; Faraoni, R.; Floyd, M.; Hunt, J.P.; Lockhart, D.J.; Milanov, Z.V.; Morrison, M.J.; Pallares, G.; Patel, H.K.; Pritchard, S.; Wodicka, L.M.; Zarrinkar, P.P. A quantitative analysis of kinase inhibitor selectivity. *Nature Biotechnol.* **2008**, 26 (1), 127 – 132.

Kastenholz, M.A.; Pastor, M.; Cruciani, G.; Haaksma, E. J.; Fox, T. GRID/CPCA: A New Computational Tool To Design Selective Ligands. *J. Med. Chem.* **2000**, *43*, 3033-3044.

Khan, A. R.; James, M. N. G. Molecular Mechanisms for the Conversion of Zymogens to Active Proteolytic Enzymes. *Protein sci.* **1998**, *7* (4), 815–836.

Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem

Substance and Compound Databases. *Nucleic Acids Res* **2016**, *44* (D1), D1202–D1213.

Kinnings, S.; Jackson, R. Binding Site Similarity Analysis for the Functional Classification of the Protein Kinase Family. *J. Chem. Inf. Model.* **2009**, *49*, 318–329.

Koeberle, S. C.; Romir, J.; Fischer, S.; Koeberle, A.; Schattel, V.; Albrecht, W.; Grütter, C.; Werz, O.; Rauh, D.; Stehle, T.; Laufer, S. a. Skepinone-L Is a Selective p38 Mitogen-Activated Protein Kinase Inhibitor. *Nat. Chem. Biol.* **2011**, *8* (2), 141–143.

Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat. Rev. Drug Discov.* **2004**, *3*, 1–5.

Kontijevskis, A.; Komorowski, J.; Wikberg, J.E. Generalized proteochemometric model of multiple cytochrome p450 enzymes and their inhibitors. *J Chem Inf Model.* **2008**, 48(9), 1840-1850.

Kontijevskis, A.; Petrovska, R.; Yahorava, S.; Komorowski, J.; Wikberg, J. E. S. Proteochemometrics Mapping of the Interaction Space for Retroviral Proteases and Their Substrates. *Bioorg. Med. Chem.* **2009**, *17* (14), 5229–5237.

Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Eyck, L. F. Ten. Surface Comparison of Active and Inactive Protein Kinases Identifies a Conserved Activation Mechanism. *Proc. Natl. Acad. Sci. USA* **2006**, *103* (47), 17783–17788.

Kovalevsky, A. Y.; Tie, Y. F.; Liu, F. L.; Boross, P. I.; Wang, Y. F.; Leshchenko, S.; Ghosh, A. K.; Harrison, R. W.; Weber, I. T. Effectiveness of Nonpeptide Clinical Inhibitor TMC-114 on HIV-1 Protease with Highly Drug Resistant Mutations D30N, I50V, and L90M. *J. Med. Chem.* **2006**, *49*, 1379−1387.

Kramer, C.; Gedeck, P. Global Free Energy Scoring Functions Based on Distance-Dependent Atom-Type Pair Descriptors. *J. Chem. Inf. Model.* **2011**, *51* (3), 707–720.

Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K I Data. *J. Med. Chem.* **2012**, *55* (11), 5165–5173.

Krowarsch, D.; Cierpicki, T.; Jelen, F.; Otlewski, J. Canonical Protein Inhibitors of Serine Proteases. *Cell. Mol. Life Sci.* **2003**, *60* (11), 2427–2444.

Kuhn, D.; Weskamp, N.; Hüllermeier, E.; Klebe, G. Functional Classification of Protein Kinase Binding Sites Using Cavbase. *ChemMedChem* **2007**, *2* (10), 1432–1447.

Kuz'min, V. E.; Artemenko, A. G.; Polischuk, P. G.; Muratov, E. N.; Hromov, A. I.; Liahovskiy, A. V.; Andronati, S. A.; Makan, S. Y. Hierarchic System of QSAR Models (1D-4D) on the Base of Simplex Representation of Molecular Structure. *J. Mol. Model.* **2005**, *11* (6), 457–467.

Lapins, M.; Prusis, P.; Lundstedt, T..; Wikberg, J.E.S. Proteochemometrics modelling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol Pharmacol.* **2002**, *61* (6), 1465-1475.

Lapinsh, M.; Prusis, P.; Uhle, S. Improved Approach for Proteochemometrics Modeling : Application to Organic Compound — Amine G Protein-Coupled Receptor Interactions. *Bioinformatics* **2005**, *21* (23), 4289–4296.

Lapins, M.; Eklund, M.; Spjuth, O.; Prusis, P.; Wikberg, J. E. S. Proteochemometric Modeling of HIV Protease Susceptibility. *BMC Bioinf.* **2008**, *9* (1), 181–192.

Lapins, M.; Wikberg, J.E.S.; Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. *BMC Bioinf.* **2010**, *11*, 339.

Liang, G.; Chen, G.; Niu, W.; Li, Z. Factor Analysis Scales of Generalized Amino Acid Information as Applied in Predicting Interactions between the Human Amphiphysin-1 SH3 Domains and Their Peptide Ligands. *Chem. Biol. Drug Des.* **2008**, *71* (4), 345–351.

Liaw, A; Wiener, M. Classification and Regression by randomForest. *R news* **2002**, *2*, 18–22.

Linden, A. Measuring Diagnostic and Predictive Accuracy in Disease Management : An Introduction to Receiver Operating Characteristic (ROC) Analysis. *J. Eval. Clin. Pract.* **2006**, *12* (2), 132–139.

Liu, Y.; Gray, N. S. Rational Design of Inhibitors That Bind to Inactive Kinase Conformations. *Nat. Chem. Biol.* **2006**, *2* (7), 358–364.

Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res* **2007**, *35* (SUPPL. 1), 198–201.

Louis, B.; Agrawal, V. K.; Khadikar, P. V. Prediction of Intrinsic Solubility of Generic Drugs Using MLR, ANN and SVM Analyses. *Eur. J. Med. Chem.* **2010**, *45* (9), 4018–4025.

Mandrika, I.; Prusis, P.; Yahorava, S.; Shikhagaie, M.; Wikberg, J.E. Proteochemometric modelling of antibody-antigen interactions using SPOT synthesised peptide arrays. *Protein Eng Des Sel.* **2007**, *20* (6):301-307.

Manning, G.; Whyte, D.B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298* (5600), 1912–1934.

Melnikova, Irena and Golden, J. Targeting Protein Kinases. *Nat. Rev. Drug Discov.* **2008**, *3*, 993–994.

Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **1975**, 405, 442–451.

Mencher, S. K.; Wang, L. G. Promiscuous Drugs Compared to Selective Drugs (Promiscuity Can Be a Virtue). *BMC Clin. Pharmacol* . **2005**, *5*, 3.

Merget, B.; Turk, S.; Eid, S.; Rippmann, F.; Fulle, S. Profiling prediction of kinase inhibitors – towards the virtual assay. *J. Med. Chem.* **2016** Just Accepted Manuscript. DOI: 10.1021/acs.jmedchem.6b0161.

Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H. O.; Rognan, D. Protein-Ligand-Based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling. *J. Chem. Inf. Model.* **2012**, *52* (4), 943–955.

Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. Data Completeness—the Achilles Heel of Drug-Target Networks. *Nat. Biotechnol.* **2008**, *26* (9), 983–984.

Metz, J.T.; Johnson, E.F.; Soni, N.B.; Merta, P.J.; Kifle, L.; Hajduk, P.J. Navigating the kinome. *Nat. Chem. Biol.* **2011**, 7(4), 200 – 202.

Michel, J.; Tirado-rives, J.; Jorgensen, W. L. Energetics of Displacing Water Molecules from Protein Binding Sites : Consequences for Ligand Optimization. *J. Am. Chem. Soc.* **2009**, *9* (9), 15403–15411.

Molecular Operating Environment (MOE), 2011.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2011.

Molecular Operating Environment (MOE), 2015.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2015.

Müller, S.; Chaikuad, A.; Gray, N. S.; Knapp, S. The Ins and Outs of Selective Kinase Inhibitor Development. *Nat. Chem. Biol.* **2015**, *11* (11), 818–821.

Murcia, M.; Morreale, A.; Ortiz, A. R. Comparative Binding Energy Analysis Considering Multiple Receptors: A Step toward 3D-QSAR Models for Multiple Targets. *J. Med. Chem.* **2006**, *49* (21), 6241–6253.

Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Tansila, N.; Naenna, T.; Prachayasittikul, V. Prediction of GFP Spectral Properties Using Artificial Neural Network. *J. Comput. Chem.* **2007**, 28, 1275-1289.

Nantasenamat, C.; Simeon, S.; Owasirikul, W.; Songtawee, N.; Lapins, M.; Prachayasittikul, V.; Wikberg, J. E. S. Illuminating the Origins of Spectral Properties of Green Fluorescent Proteins via Proteochemometric and Molecular Modeling. *J. Comput. Chem.* **2014**, *35* (27), 1951–1966.

Neurath, H.; Walsh, K. a; Winter, W. P. Evolution of Structure and Function of Proteases. *Science* **1967**, *158* (809), 1638–1644.

Nisius, B.; Gohlke, H. Alignment-Independent Comparison of Binding Sites Based on DrugScore Potential Fields Encoded by 3D Zernike Descriptors. *J. Chem. Inf. Model.* **2012**, *52* (9), 2339–2347.

Nolen, B.; Taylor, S.; Ghosh, G. Regulation of Protein Kinases: Controlling Activity through Activation Segment Conformation. *Mol. Cell* **2004**, *15* (5), 661–675.

Novotni, M.; Klein, R. 3D Zernike Descriptors for Content Based Shape Retrieval. *Proceedings of the eighth ACM Symposium on Solid Modeling and Applications* **2003**, 216–225.

Oda, K. New Families of Carboxyl Peptidases: Serine-Carboxyl Peptidases and Glutamic Peptidases. *J. Biochem.* **2012**, *151* (1), 13–25.

Oprea, T. I.; Tropsha, A. Target, Chemical and Bioactivity Databases - Integration Is Key. *Drug Discov. Today Technol.* **2006**, *3* (4), 357–365.

Organisation for Economic Co-operation and Development OECD principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship models. http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf

Ortiz, a R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681–2691.

Paricharak, S.; Cortés-Ciriano, I.; Ijzerman, A. P.; Malliavin, T. E.; Bender, A. Proteochemometric Modelling Coupled to in Silico Target Prediction: An Integrated Approach for the Simultaneous Prediction of Polypharmacology and Binding Affinity/potency of Small Molecules. *J. Cheminform.* **2015**, *7* (1), 1–11.

Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43* (17), 3233–3243.

Peters, J. U. Polypharmacology - Foe or Friend? *J. Med. Chem.* **2013**, *56* (22), 8955–8971.

Pinto, D. J. P.; Smallheer, J. M.; Cheney, D. L.; Knabb, R. M.; Wexler, R. R. Factor Xa Inhibitors: Next-Generation Antithrombotic Agents. *J. Med. Chem.*. **2010**, *53*, 6243−6274.

Powers, J. C.; Asgian, J. L.; James, K. E. Irreversible Inhibitors of Serine , Cysteine , and Threonine Proteases. *Chem. Rev.* **2002**, 102, 4639−4750.

Prusis, P.; Muceniece, R.; Andersson, P.; Post, C.; Lundstedt, T.; Wikberg, J.E.S. PLS modelling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. *Biochim. Biophys. Acta*. **2001**, 1544, 350 - 357.

Prusis, P.; Lapins, M.; Yahorava, S.; Petrovska, R.; Niyomrattanakit, P.; Katzenmeier, G.; Wikberg, J. E. S. Proteochemometrics Analysis of Substrate Interactions with Dengue Virus NS3 Proteases. *Bioorg. Med. Chem.* **2008**, *16* (20), 9369–9377.

Puente, X. S.; Sanchez, L. M.; Overall, C. M.; Lopez-Otin, C. Human and Mouse Proteases: A Comparative Genomic Approach. *Nat. Rev. Genet.* **2003**, *4* (7), 544–58.

Qureshi, Z.P. ; Seoane- Vazquez, E.; Rodriguez- Monguio, R.; Stevenson, K.B.; Szeinbach, S.L. Market Withdrawal of New Molecular Entities Approved in the United States from 1980 to 2009. *Pharmacoepidemiol. Drug Saf.* **2011**, *20*, 772–777.

Rasti, B.; Karimi-jafari, M. H. Quantitative Characterization of the Interaction Space of the Mammalian Carbonic Anhydrase Isoforms I , II , VII , IX , XII , and XIV and Their Inhibitors , Using the Proteochemometric Approach. *Chem. Biol. Drug. Des.* **2016**, 341–353.

Rawlings, N. D. Peptidase Specificity from the Substrate Cleavage Collection in the MEROPS Database and a Tool to Measure Cleavage Site Conservation. *Biochimie* **2016**, *122*, 5–30.

RDKit: Open-source cheminformatics; http://www.rdkit.org

Robinson, D. D.; Sherman, W.; Farid, R. Understanding kinase selectivity through energetic analysis of binding Site waters. *ChemMedChem*, **2010**, 5(4), 618 – 627.

Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41* (14), 2481–2491.

Sanderson, P. E. J. Small, Noncovalent Serine Protease Inhibitors. *Med. Res. Rev.* **1999**, *19* (2), 179–197.

Sanschagrin, P. C.; Kuhn, L. a. Cluster Analysis of Consensus Water Sites in Thrombin and Trypsin Shows Conservation between Serine Proteases and Contributions to Ligand Specificity. *Protein Sci.* **1998**, *7* (10), 2054–2064.

Sastry, M.; Dixon, S.; Sherman, W. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature- Pair Similarities and Volume Overlap Scoring. *J. Chem. Inf. Model.* **2011**, *51*, 2455−2466.

Schechter, I. Reprint of "On the Size of the Active Site in Proteases. I. Papain." *Biochem. Biophys. Res. Commun.* **2012**, *425* (3), 497–502.

Schrödinger Suite 2011 Protein Preparation Wizard; Epik version 2.2, Schrödinger, LLC, New York, NY, 2011; Impact version 5.7, Schrödinger, LLC, New York, NY, 2011; Prime version 3.0, Schrödinger, LLC, New York, NY, 2011.

Shaikh, N.; Sharma, M.; Garg, P. Molecular BioSystems An Improved Approach for Predicting Drug – Target Interaction : Proteochemometrics to Molecular. *Mol. BioSyst.* **2016**, *12*, 1006–1014.

Sheinerman, F. B.; Giraud, E.; Laoui, A. High Affinity Targets of Protein Kinase Inhibitors Have Similar Residues at the Positions Energetically Important for Binding. *J. Mol. Biol.* **2005**, *352* (5), 1134–1156.

Shiraishi, A.; Niijima, S.; Brown, J. B.; Nakatsui, M.; Okuno, Y. Chemical Genomics Approach for GPCR-Ligand Interaction Prediction and Extraction of Ligand Binding Determinants. *J. Chem. Inf. Model.* **2013**, *53* (6), 1253–1262.

Siezen, R. J.; Leunissen, J. A. Subtilases: The Superfamily of Subtilisin-like Serine Proteases. *Protein sci* **1997**, *6* (3), 501–523.

SIMCA-P version 12, Umetrix AB, Box 7960, SE -907, 19 Umea, Sweden, 2011.

Simeon, S.; Spjuth, O.; Lapins, M.; Nabu, S.; Anuwongcharoen, N.; Prachayasittikul, V.; Wikberg, J. E. S.; Nantasenamat, C. Origin of Aromatase Inhibitory Activity via Proteochemometric Modeling. *PeerJ*. **2016**, *4*, e1979.

Smith D.A.; Schmid, E.F. Drug withdrawals and the lessons within. *Curr Opin Drug Discov Devel.* **2006**, *9* (1), 38-46.

Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms. *J Chem Inf Comput. Sci.* **2003**, *43* (6), 2019–2024.

Strömbergsson, H.; Kryshtafovych, A.; Prusis, P.; Fidelis, K..; Wikberg, J.E.S.; Komorowski, J.; Hvidsten, T.R. Generalized modelling of enzyme-ligand interactions using proteochemometrics and local protein substructures. *Proteins* **2006**, *65* (3), 568-579.

Strömbergsson, H.; Daniluk, P.; Kryshtafovych, A.; Wikberg, J. E. S.; Kleywegt, G. J.; Hvidsten, T. R. Interaction Model Based on Local Protein Substructures Generalizes to the Entire Structural Enzyme-Ligand Space Interaction Model Based on Local Protein Substructures Generalizes to the Entire Structural Enzyme-Ligand Space. *J. Chem. Inf. Model.* **2008**, *48*, 2278–2288.

The Open Babel Package, 2.3.0 http://openbabel.org (accessed Sep 2011).

Tian, F.; Zhou, P.; Li, Z. T-Scale as a Novel Vector of Topological Descriptors for Amino Acids and Its Application in QSARs of Peptides. *J. Mol. Struct.* **2007**, *830* (1–3), 106–115.

Tiikkainen, P.; Franke, L. Analysis of Commercial and Public Bioactivity Databases. *J. Chem. Inf. Model.* **2012**, *52* (2), 319–326.

Tobin, A. B. G-Protein-Coupled Receptor Phosphorylation: Where, When and by Whom. *Br. J. Pharmacol.* **2008**, *153*, S167-S176.

Todeschini, R.; Consonni, V. Molecular Descriptors for Chemoinformatics. Mannhold, R.; Kubinyi, H.; Folkers, G., editors. Weinheim: Wiley-VCH; 2009. p. 1257

Topliss, J. G.; Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, 15(10), 1066 -1068.

Treiber, D. K.; Shah, N. P. Ins and Outs of Kinase DFG Motifs. *Chem. Biol.* **2013**, *20* (6), 745–746.

Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29* (6–7), 476–488.

Tsatsanis, C.; Spandidos, D. A. The role of oncogenic kinases in human cancer (Review). *Int. J. Mol. Med.* **2000**, *5*(6), 583 – 590.

Turk, B. Targeting Proteases: Successes, Failures and Future Prospects. *Nat. Rev. Drug Discov.* **2006**, *5* (9), 785–799.

Uitdehaag, J.CM; Zaman, G.JR. A theoretical entropy score as a single value to express inhibitor selectivity. *BMC Bioinformatics* **2011**, 12, 94.

Uzelac, I.; Olsson, T.; Eriksson, L. A.; Gottfries, J. A Structural Comparison Approach for Identifying Small Variations in Binding Sites of Homologous Proteins. *Comput Mol Biosci.* **2015**, *5*, 45–55.

Van Linden, O. P. J.; Kooistra, A. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Knowledge-Based Structural Database To Navigate Kinase–Ligand Interaction Space. *J. Med. Chem.* **2014**, *57* (2), 249–277.

Van Westen, G. J. P.; Wegner, J.K.; IJzerman, A.P.; van Vlijmen, H.W.T.; Bender, A. Proteochemometric modelling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, **2011**, 2, 16 – 30.

Van Westen, G. J. P.; Wegner, J. K.; Geluykens, P.; Kwanten, L.; Vereycken, I.; Peeters, A.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Which Compound to

Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. *PLoS ONE* **2011**, *6* (11).

Van Westen, G. J. P.; Van Den Hoven, O. O.; Van Der Pijl, R.; Mulder-Krieger, T.; De Vries, H.; Wegner, J. K.; IJzerman, A. P.; Van Vlijmen, H. W. T.; Bender, A. Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. *J. Med. Chem.* **2012**, *55* (16), 7010–7020.

Van Westen, G. J. P.; Swier, R. F.; Cortes-Ciriano, I.; Wegner, J. K.; Overington, J. P.; Jzerman, A. P. I.; Van Vlijmen, H. W. T.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. *J. Cheminform.* **2013**, *5*, 42.

Van Westen, G. J. P.; Hendriks, A.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Significantly Improved HIV Inhibitor Efficacy Prediction Employing Proteochemometric Models Generated From Antivirogram Data. *PLoS Comput. Biol.* **2013**, *9* (2).

Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics: *Quo Vadis* ? *J. Chem. Inf. Model.* **2012**, *52* (6), 1413–1437.

Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **2010**, *10* (1), 95–115.

Volkamer, A.; Eid, S.; Turk, S.; Rippmann, F.; Fulle, S. Identification and Visualization of Kinase-Specific Subpockets. *J. Chem. Inf. Model.* **2016**, *56* (2), 335–346.

Wahlberg, E.; Karlberg, T.; Kouznetsova, E.; Markova, N.; Macchiarulo, A.; Thorsell, A.-G.; Pol, E.; Frostell, A.; Ekblad, T.; Oncu, D.; Kull, B.; Robertson, G. M.; Pellicciari, R.; Schuler, H.; Weigelt, J. Family-Wide Chemical Profiling and Structural Analysis of PARP and Tankyrase Inhibitors. *Nat. biotechnol.* **2012**, *30* (3), 283–288.

Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48* (12), 4111–4119.

WaterMap, version 1.4, Schrödinger, LLC, New York, NY, 2012.

Watts, K.S.; Dalal, P.; Murphy, R.B.; Sherman, W.; Friesner, R.A.; Shelley, J.C., "ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers," *J. Chem. Inf. Model.* **2010**, 50, 534-546.

Weill, N.; Rognan, D. Development and Validation of a Novel Protein-Ligand Fingerprint To Mine Chemogenomic Space : Application to G Protein-Coupled Receptors and Their Ligands. *J. Chem. Inf. Model.* **2009**, 49, 1049–1062.

Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50* (1), 123–135.

Weill, N.; Valencia, C.; Gioria, S.; Villa, P.; Hibert, M.; Rognan, D. Identification of Nonpeptide Oxytocin Receptor Ligands by Receptor-Ligand Fingerprint Similarity Search. *Mol. Inform.* **2011**, *30* (6–7), 521–526.

Wikberg, J. E. S.; Lapinsh, M.; Prusis, P. Proteochemometrics: A Tool for Modeling the Molecular Interaction Space, In Chemogenomics in Drug Discovery: A Medicinal ChemistryPerspective; Kubinyi, H. and Muller, G., Eds.; Wiley-VCH: Weinheim, 2004; pp 289 – 309.

Wikberg, J.E.S.; Eklund, M.; Willighagen, E.L.; Spjuth, O.; Lapins, M.; Engkvist, O.; Alvarsson, J. Proteochemometrics, In Introduction to Pharmaceutical Bioinformatics; Oakleaf Academic: Stockholm, 2011; pp 289-344.

Williams, A. J.; Ekins, S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discov. Today* **2011**, *16* (17–18), 747–750.

Wohlfahrt, G.; Sipila, J.; Pietila, L. O. Field-based comparison of ligand and coactivator binding sites of nuclear receptors. *Biopolymers* **2009**, 91(10), 884 – 894.

Wold; Svante; Esbensen, K.; Geladi. P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, 2(1), 37–52.

Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemometr. Intell. Lab.* **2001**, *58* (2), 109–130.

Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T. Analysis of Multiple Compound – Protein Interactions Reveals Novel Bioactive Molecules. *Mol. Syst. Biol.* **2011**, *7* (472), 1–12.

Yang, L.; Shu, M.; Ma, K.; Mei, H.; Jiang, Y.; Li, Z. ST-Scale as a Novel Amino Acid Descriptor and Its Application in QSAM of Peptides and Analogues. *Amino Acids* **2010**, *38* (3), 805–816.

Zhang, J.; Yang, P. L.; Gray, N. S. Targeting Cancer with Small Molecule Kinase Inhibitors. *Nat. Rev. Cancer* **2009**, *9* (1), 28–39.

Zhang, L.; Tsai, K.-C.; Du, L.; Fang, H.; Li, M.; Xu, W. How to Generate Reliable and Predictive CoMFA Models. *Curr. Med. Chem.* **2011**, *18* (6), 923–930.

Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol.* **2014**, *9* (6), 1230–1241.

Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M. Through The "gatekeeper Door": Exploiting the Active Kinase Conformation. *J. Med. Chem.* **2010**, *53* (7), 2681–2694.