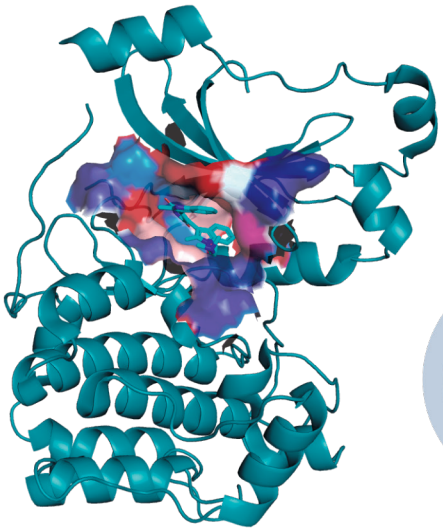
DISSERTATIONES SCHOLAE DOCTORALIS AD SANITATEM
INVESTIGANDAM UNIVERSITATIS HELSINKIENSIS

37/2016

ALEXANDRE BORREL

Development of Computational Methods to Predict Protein Pocket Druggability and Profile Ligands using Structural Data



DIVISION OF PHARMACEUTICAL CHEMISTRY AND TECHNOLOGY
FACULTY OF PHARMACY
DOCTORAL PROGRAMME IN INTEGRATIVE LIFE SCIENCE
UNIVERSITY OF HELSINKI

MOLÉCULES THÉRAPEUTIQUES IN SILICO INSERM UMR-S 973
DOCTORAL PROGRAMME BIO SORBONNE PARIS CITÉ
UNIVERSITY PARIS DIDEROT (PARIS 7) SORBONNE PARIS CITÉ

Division of Pharmaceutical Chemistry and Technology

Faculty of Pharmacy, University of Helsinki, Finland

&

Molécules Thérapeutiques *in Silico* (MT*i*), Inserm UMR-S 973

University Paris Diderot, France

Development of Computational Methods to Predict Protein Pocket Druggability and Profile Ligands using Structural Data

Alexandre Borrel

ACADEMIC DISSERTATION

To be presented, with permission of the Faculty of Pharmacy of the University of Helsinki and the University Paris Diderot, for public examination in Lecture Room 247E, University Paris Diderot Halle aux Farines, on May 26th 2016, at 14:00.

May 2016

Supervisors Anne-Claude Camproux, PhD, Professor
Molécules Thérapeutiques *in silico* (MT*i*), Inserm UMR-S 973
Computational approaches applied to pharmacological profiling
University Paris Diderot, Paris, France

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

Examiners Catherine Etchebest, PhD, Professor
Dynamique des Systèmes et Interactions des Macromolécules Biologiques, Inserm UMR-S 1134
University Paris Diderot, Paris, France

Jari Yli-Kauhaluoma, PhD, Professor
Division of Pharmaceutical Chemistry and Technology
Faculty of Pharmacy, University of Helsinki, Helsinki, Finland

Pre-examiners Bernard Offmann, PhD, Professor
Unité Fonctionnalité et Ingénierie des Protéines CNRS FRE 3478
University of Nantes, Nantes, France

Antti Poso, PhD, Professor
School of Pharmacy
University of Eastern Finland, Kuopio, Finland

Opponent Nathan Brown, PhD
In Silico Medicinal Chemistry group
The Institute of Cancer Research, London, England

Cover: Augmented reality picture to be visualized with the application Augment (http://www.augment.com/; available on the App Store and Google Play, for more information scan the QR Code below). Druggable binding site of the human protein kinase C beta II (PDB code 2I0E). Polar (red) to hydrophobic (blue) gradients are represented. The picture was made using the Pymol (DeLano, 2002) and Blender softwares (Blender Foundation, 2016).

ISBN 978-951-51-2173-8 (paperback)
ISBN 978-951-51-2174-5 (PDF)
ISSN 2342-3161
http:/ethesis.helsinki.fi

Abstract

This thesis presents the development of computational methods and tools using as input three-dimensional structures data of protein-ligand complexes. The tools are useful to mine, profile and predict data from protein-ligand complexes to improve the modeling and the understanding of the protein-ligand recognition. This thesis is divided into five sub-projects. In addition, unpublished results about positioning water molecules in binding pockets are also presented.

I developed a statistical model, PockDrug, which combines three properties (hydrophobicity, geometry and aromaticity) to predict the druggability of protein pockets, with results that are not dependent on the pocket estimation methods. The performance of pockets estimated on apo or holo proteins is better than that previously reported in the literature (Publication I). PockDrug is made available through a web server, PockDrug-Server (http://pockdrug.rpbs.univ-paris-diderot.fr), which additionally includes many tools for protein pocket analysis and characterization (Publication II).

I developed a customizable computational workflow based on the superimposition of homologous proteins to mine the structural replacements of functional groups in the Protein Data Bank (PDB). Applied to phosphate groups, we identified a surprisingly high number of phosphate non-polar replacements as well as some mechanisms allowing positively charged replacements. In addition, we observed that ligands adopted a U-shape conformation at nucleotide binding pockets across phylogenetically unrelated proteins (Publication III).

I investigated the prevalence of salt bridges at protein-ligand complexes in the PDB for five basic functional groups. The prevalence ranges from around 70% for guanidinium to 16% for tertiary ammonium cations, in this latter case appearing to be connected to a smaller volume available for interacting groups. In the absence of strong carboxylate-mediated salt bridges, the environment around the basic functional groups studied appeared enriched in functional groups with acidic properties such as hydroxyl, phenol groups or water molecules (Publication IV).

I developed a tool that allows the analysis of binding poses obtained by docking. The tool compares a set of docked ligands to a reference bound ligand (may be different molecule) and provides a graphic output that plots the shape overlap and a Jaccard score based on comparison of molecular interaction

fingerprints. The tool was applied to analyse the docking poses of active ligands at the orexin-1 and orexin-2 receptors found as a result of a combined virtual and experimental screen (Publication V).

The review of literature focusses on protein-ligand recognition, presenting different concepts and current challenges in drug discovery.

Tiivistelmä

Tässä väitöskirjassa esitetään tietokoneavusteisia menetelmiä ja työkaluja, jotka perustuvat proteiini-ligandikompleksien kolmiulotteisiin rakenteisiin. Ne soveltuvat proteiini-ligandikompleksien rakennetiedon louhimiseen, optimointiin ja ennustamiseen. Tavoitteena on parantaa sekä mallinnusta että käsitystä proteiini-liganditunnistuksesta. Väitöskirjassa työkalut kuvataan viitenä eri alahankkeena. Lisäksi esitetään toistaiseksi julkaisemattomia tuloksia vesimolekyylien asemoinnista proteiinien sitoutumistaskuihin.

Kehitin PockDrugiksi kutsumani tilastollisen mallin, joka yhdistää kolme ominaisuutta – hydrofobisuuden, geometrian ja aromaattisuuden – proteiinitaskujen lääkekehityskohteeksi soveltuvuuden ennustamista varten siten, että tulokset ovat riippumattomia sitoutumistaskun sijoitusmenetelmästä. Apo- ja holoproteiinien taskujen ennustaminen toimii paremmin kuin alan kirjallisuudessa on aiemmin kuvattu (Julkaisu I). PockDrug on vapaasti käyttäjien saatavilla PockDrug-verkkopalvelimelta (http://pockdrug.rpbs.univ-paris-diderot.fr), jossa on lisäksi useita työkaluja proteiinin sitoutumiskohdan analyysiin ja karakterisointiin (Julkaisu II).

Kehitin myös muokattavissa olevan tietokoneavusteisen prosessin, joka perustuu samankaltaisten proteiinien päällekkäin asetteluun, louhiakseni Protein Data Bankista (PDB) toiminnallisten ryhmien rakenteellisia korvikkeita. Tätä fosfaattiryhmiin soveltaessani tunnistin yllättävän paljon poolittomia fosfaattiryhmän korvikkeita ja joitakin positiivisesti varautuneita korvikkeita mahdollistavia mekanismeja. Lisäksi havaitsin, että ligandit omaksuivat U-muotoisen konformaation fylogeneettisesti riippumattomien proteiinien nukleotidien sitoutumistaskuissa (Julkaisu III).

Tutkin PDB:n proteiini-ligandikompleksien suolasiltojen yleisyyttä viidelle emäksiselle toiminnalliselle ryhmälle. Suolasiltojen yleisyys vaihteli guanidinium-ionin 70 prosentista tertiääristen ammoniumkationien 16 prosenttiin. Jälkimmäisessä tapauksessa suolasiltojen vähäisyys vaikuttaa riippuvan siitä, että vuorovaikuttaville ryhmille on vähemmän tilaa. Mikäli tarkastellut emäksiset ryhmät eivät osallistuneet vahvoihin karboksylaattivälitteisiin suolasiltoihin, niiden ympäristössä vaikutti olevan runsaasti happamia toiminnallisia ryhmiä, kuten hydroksi- ja fenoliryhmiä sekä vesimolekyylejä (Julkaisu IV).

Lopuksi kehitin työkalun, joka mahdollistaa telakoinnista saatujen sitoutumisasentojen analyysin. Työkalu vertaa telakoitua ligandisarjaa

sitoutuneeseen vertailuligandiin, joka voi olla eri molekyyli. Graafisena tulosteena saadaan diagrammi ligandien muotojen samankaltaisuudesta ja molekyylivuorovaikutusten sormenjälkiin perustuvasta Jaccard-pistemäärästä. Työkalua sovellettiin oreksiini-1- ja oreksiini-2-reseptoreille yhdistetyllä virtuaalisella ja kokeellisella seulonnalla löydettyjen aktiivisten ligandien sitoutumisasentojen analyysiin (Julkaisu V).

Kirjallisuuskatsaus keskittyy proteiini-liganditunnistukseen sekä esittää niihin liittyviä käsitteitä ja lääkkeenkeksimisen ajankohtaisia haasteita.

Résumé

Cette thèse présente le développement de méthodes et d'outils informatiques basés sur la structure tridimensionnelle des complexes protéine-ligand. Ces différentes méthodes sont utilisées pour extraire, optimiser et prédire des données à partir de la structure des complexes afin d'améliorer la modélisation et la compréhension de la reconnaissance entre une protéine et un ligand. Ce travail de thèse est divisé en cinq projets. En complément, une étude sur le positionnement des molécules d'eau dans les sites de liaisons a aussi été développée et est présentée.

Dans une première partie un modèle statistique, PockDrug, a été mis en place. Il combine trois propriétés de poches protéiques (l'hydrophobicité, la géométrie et l'aromaticité) pour prédire la druggabilité des poches protéiques, si une poche protéique peut lier une molécule *drug-like*. Le modèle est optimisé pour s'affranchir des différentes méthodes d'estimation de poches protéiques. La qualité des prédictions, est meilleure à la fois sur des poches estimées à partir de protéines apo et holo et est supérieure aux autres modèles de la littérature (Publication I). Le modèle PockDrug est disponible sur un serveur web, PockDrug-Server (http://pockdrug.rpbs.univ-paris-diderot.fr) qui inclus d'autres outils pour l'analyse et la caractérisation des poches protéiques.

Dans un second temps un protocole, basé sur la superposition de protéines homologues a été développé pour extraire des replacements structuraux de groupements chimiques fonctionnels à partir de la Protein Data Bank (PDB). Appliqué aux phosphates, un grand nombre de remplacements non-polaires ont été identifié pouvant notamment être chargés positivement. Quelques mécanismes de remplacements ont ainsi pu être analysé. Nous avons, par exemple, observé que le ligand adopte une configuration en forme U dans les sites de liaison des nucléotides indépendamment de la phylogénétique des protéines (Publication III).

Dans une quatrième partie, la prévalence des ponts salins de cinq groupements chimiques basiques a été étudié dans les complexes protéine-ligand. Ainsi le pourcentage de pont salin fluctue de 70% pour le guanidinium à 16% pour l'amine tertiaire qui a le plus faible volume disponible autour de lui pour accueillir un group pouvant interagir. L'absence d'acide fort comme l'acide carboxylique pour former un pont salin est remplacé par un milieu enrichis en groupement chimiques fonctionnels avec des propriétés acides comme l'hydroxyle, le phénol ou encore les molécules d'eau (Publication IV).

Dans un dernier temps un outil permettant l'analyse des poses de ligand obtenues par une méthode d'ancrage moléculaire a été développé. Cet outil compare ces poses à un ligand de référence, qui peut être une molécule différente en combinant l'information du chevauchement de forme de la pose et du ligand de référence et un score de Jaccard basé sur une comparaison des empreintes d'interaction moléculaires du ligand de référence et de la pose. Cette méthode a été utilisé dans l'analyse des résultats d'ancrage moléculaires pour des ligands actifs pour les récepteurs aux orexine 1 et 2. Ces ligands actifs ont été trouvés à partir de résultats combinant un criblage virtuel et expérimental.

La revue de la littérature associée est focalisée sur la reconnaissance moléculaire d'un ligand pour une protéine et présente différent concepts et challenges pour la recherche de nouveaux médicaments.

Acknowledgements

"*This is your last chance. After this, there is no turning back. You take the blue pill - the story ends, you wake up in your bed and believe whatever you want to believe. You take the red pill - you stay in Wonderland and I show you how deep the rabbit hole goes*." Morpheus (Matrix 1999).

My thesis was for me like the red pill for Neo in Matrix. When you start a thesis, your mind is altered considerably, you become more rigorous, less flexible, more critical and develop a harder personality. No come-backs are possible and the side effects are enormous. These transformations are also dependent on the actions of the many people thanked in this section.

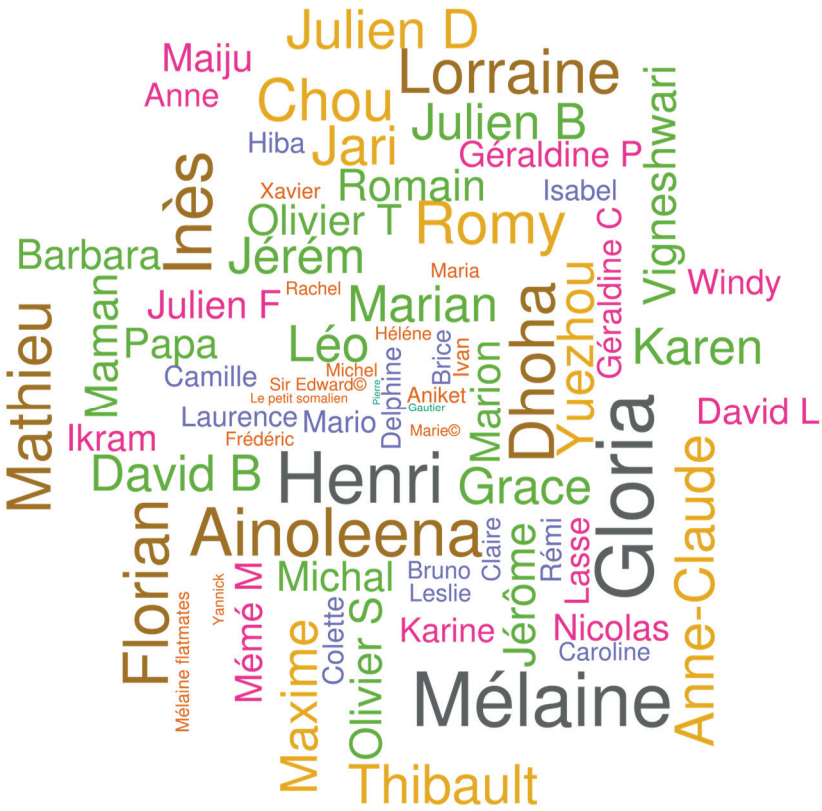
My profound gratitude is owed to my supervisors, Dr. Henri Xhaard and Professor Anne-Claude Camproux. I thank both of you for having given me this opportunity to do this joint thesis and for having organized your schedules to allow me to visit both of your laboratories equally. I hope this joint thesis will have a real impact on my scientific career. Henri, you have been an exceptional mentor over these last four years. I thank you for your ideas, our very interesting scientific (or not so scientific) discussions, your constructive proofreading and your efforts to help me to develop my critical thinking and my English. Anne-Claude, I thank you for proposing the topic of this thesis, for supervising my first publication, the hardest for a PhD student, and for exemplifying scientific perseverance. I thank you also Anne-Claude for giving me the opportunity to be a teacher.

I extend my sincere gratitude to my pre-examiners, Professor Bernard Offmann and Professor Antti Poso, for their observations and comments that improved this manuscript. I also thank my opponent, Dr. Nathan Brown, for accepting the invitation to my defense. I thank my examiners, Professor Catherine Etchebest and Professor Jari Yli-Kauhaluoma, for accepting the invitation to my defense and for representing both Universities. I am very grateful to Jari for commenting on my thesis, improving my English and my style.

During these last four years I have had the opportunity and pleasure to collaborate with numerous individuals and internship students, which

improved my scientific style. Thank you very much all of you for your time, your help and your patience.

Over the last four years, I have met many people, represented in the name cloud below, who were instrumental to the success of this thesis. A special thank-you goes to my colleagues and lab mates in Helsinki and in Paris for the shared time. I especially thank Gloria for her help with Finnish administration, for improving my English, especially when I was a master student, and for animating my life in Finland. I also thank Mélaine, we started together as colleagues, but you are an important friend now. Thank you for the time spent together, the coffee breaks, socializing after work, the parties … but also for the work time. Thank you also for listening to my perpetual criticisms and for considerably improving the thesis years.



Name cloud of thanked people. Colours and sizes are provided randomly; if you have a good colour and a large name, you win my lottery acknowledgements.

I am indebted to the University Paris Diderot and the French Research Minister for having given me an employee position for three years. I also thank the University of Helsinki for having considered me as a PhD student and for having paid my last study year. I would like also express my gratitude to the French Embassy (Kaksin program) for having paid my travel between the France and Finland, to the doctoral programme health of science for having accorded me a travel grand and a completion grant in the end of my thesis, to Magnus Ehrnrooth Foundation for having given me a graduate student fellowship, to the European action CM1207 STSM for having accorded a research grant and to the French Society of Bioinformatics for having paid my trip for the European Conference of Computational Biology 2014. Finally, I would express a special thanks to the National Doctoral Programme in Informational and Structural Biology (ISB) for having given me several travel grants for conferences and courses and for their winter schools in Laplane which are my favorite souvenirs in Finland.

I thank my old friends and members of the association of the young French Bioinformatician, for which I served as president during my first two years in the thesis process, for invigorating my life beyond this thesis.

Finally, I express my heart-felt thanks to my parents and my grandmother for their unconditional support during the last ten years of my college study.

I conclude by thanking everyone who has read or will read this manuscript and who will be supporting me during my defense.

Helsinki, May 2016
Alexandre Borrel

Remerciements

"*C'est là ta dernière chance, après ça tu ne pourras plus faire marche arrière. Choisis la pilule bleue et tout s'arrête : après tu pourras faire de beaux rêves et penser ce que tu veux. Choisis la pilule rouge, tu restes au pays des merveilles, et on descend avec le lapin blanc au fond du gouffre.*" Morpheus (Matrix 1999).

Commencer un doctorat a été un peu pour moi comme prendre la pilule rouge. Aucun retour ne fut possible et cela a considérablement changé mon état d'esprit et mon caractère, plus rigoureux, moins flexible, plus critique et plus endurci. Cette « pilule » a aussi eu de nombreux effets secondaires. Cette transformation est dépendante de nombreuses personnes remerciées ici.

Mes plus profonds remerciements vont à mes directeurs de thèse, le Dr. Henri Xhaard et le Pr. Anne-Claude Camproux. Merci à vous deux de m'avoir donné l'opportunité de faire ce doctorat en cotutelle et de m'avoir permis de passer la moitié de mon temps dans un autre laboratoire que le vôtre. J'espère que ce doctorat en cotutelle aura un réel impact dans ma carrière scientifique et ceci sera en partie grâce à vous. Henri, tu as été un vrai mentor scientifique pour moi durant tout ce doctorat. Merci pour tes idées, nos intéressantes discussions scientifiques et non scientifiques, tes corrections constructives et tes efforts pour m'aider à développer mon esprit critique et mon anglais. Anne-Claude, merci de m'avoir donné cette opportunité de faire ce doctorat, d'avoir supervisé ma première publication scientifique, la plus difficile pour un doctorant et de m'avoir montré ce qu'est la persévérance scientifique. Merci aussi de m'avoir permis d'être un moniteur et ainsi d'avoir une expérience dans l'enseignement.

Mes seconds remerciements vont aux membres du jury de ma soutenance. J'aimerais exprimer toute ma gratitude à mes pré-examinateurs, le Pr. Bernard Offmann et le Pr. Antti Poso pour leurs commentaires et leurs remarques qui ont améliorés mon manuscrit de thèse. Je voudrais exprimer de sincères remerciements à mon président de jury, le Dr. Nathan Brown d'avoir accepté cette invitation. Enfin, j'aimerais remercier mes examinateurs, le Pr. Catherine Etchebest et le Pr. Jari Yli-Kauhaluoma d'avoir accepté d'être mes examinateurs et d'avoir représenté mes deux universités partenaires. J'aimerais

partager mon extrême gratitude pour Jari qui m'a donné beaucoup de son temps pour corriger mon manuscrit afin d'améliorer mon anglais et mon style.

Durant ces années de doctorat j'ai eu l'opportunité de travailler avec de nombreux collaborateurs et étudiants en stage. Ils ont contribué à améliorer mon esprit scientifique et mes capacités de management. Merci beaucoup pour votre temps, votre aide et votre patience.

J'ai eu l'occasion pendant ces quatre années de doctorat de rencontrer beaucoup de monde et cela serait très ennuyeux de citer tout le monde. En figure ci-dessus, sur un nuage de noms est résumé beaucoup, mais pas toutes, les personnes qui ont été importantes durant ce doctorat. Cependant, j'aimerais exprimer des remerciements tout particuliers pour mes collègues de bureau à Helsinki et à Paris pour le temps passé ensemble. J'aimerais remercier Gloria pour son aide avec l'administration finlandaise, d'avoir supporté mon anglais tout particulièrement quand je n'étais qu'un étudiant erasmus et d'avoir animé ma vie en Finlande. Mes plus importants remerciements vont à Mélaine. Nous avons commencé notre doctorat ensemble comme collègue et tu es devenu une grande amie maintenant. Merci pour ce temps passé ensemble, les pauses café, les after-works, les soirées, … mais aussi le temps à travailler. Merci d'avoir supporté mes perpétuelles critiques et d'avoir amélioré ces années de doctorat.

J'aimerais remercier l'université Paris Diderot et le ministère de l'enseignement supérieur et de la recherche française pour m'avoir donné une position de salarié durant trois années et de m'avoir accordé une bourse de voyage pour financer les surcoûts de la cotutelle. Merci aussi à l'université de Helsinki de m'avoir accepté comme doctorant et d'avoir financé ma dernière année de doctorat. J'aimerais remercier l'ambassade de France en Finlande d'avoir payé une partie de mes voyages entre la France et la Finlande (Kaksin programme), l'école doctoral de santé et science de Helsinki pour m'avoir accordé une bourse de voyage et une bourse de fin de doctorat, la fondation Magnus Ehrnrooth pour sa bourse de fin d'étude, l'action européenne CM1207 STSM pour sa bourse de recherche et la Société Française de Bioinformatique (SFBI) pour sa bourse de voyage pour assister à la conférence européenne de bioinformatique en 2014. Enfin, j'aimerais adresser un remerciement tout particulier au programme doctoral national d'information et de biologie

structural (ISB) pour m'avoir donné plusieurs bourses de voyages et pour ses écoles d'hiver en Laponie, qui font partie de mes meilleurs souvenirs de Finlande

J'aimerais remercier mes amis qui ont été là durant ce doctorat et qui ont animé ma vie extra-thèse. J'aimerais aussi exprimer ma gratitude aux membres de l'association des jeunes bioinformaticiens de France dont j'ai été leur président durant mes deux premières années de doctorat.

Finalement, j'aimerais exprimer mes plus fort remerciements à mes parents et ma grand-mère pour leur soutien inconditionnel pendant ces dix dernières années d'études universitaires.

J'aimerais finir par remercier toute les personnes qui liront cette thèse et toute celles qui ont été là et qui m'ont soutenu pendant la soutenance.

Helsinki, Mai 2016
Alexandre Borrel

Table of contents

List of Figures

List of Tables

Abbreviations

3D: three-dimensional
ADME: absorption, distribution, metabolism, and excretion
ADP: adenosine diphosphate
AMP: adenosine monophosphate
ATP: adenosine triphosphate
CoMFA: comparative molecular field analysis
CSD: Cambridge Structural Database
D: druggable
DCD: druggable cavity directory
DNA: deoxyribonucleic acid
FDA: US Food and Drug Administration
FEP: free energy perturbation
H: enthalpy
H-bond: hydrogen bond
IC_{50}: half maximal inhibitory concentration
ITC: isothermal titration calorimetry
IUPAC: International Union of Pure and Applied Chemistry
K_a: association constant
K_d: dissociation constant
K_i: inhibition constant
K_m: Michaelis constant, median concentration
k_{off}: dissociation rate constant
k_{on}: association rate constant
L: ligand
NAMS: non-contiguous atom matching structural similarity
ND: non-druggable
NMR: nuclear magnetic resonance
NRDD: non redundant druggable dataset
NRDLD: non redundant dataset of druggable and less druggable binding sites
MCC: Matthew's coefficient correlation
MCS: maximum common structures
MM/PBSA: molecular mechanics/ Poisson -Boltzmann surface area
PCA: principal component analysis
PDB: protein data bank
POP: pyrophosphate
QSAR: quantitative structure activity relationships
QSPR: quantitative structure property relationships
R: receptor
RMSD: root-mean-square deviation
RNA: ribonucleic acid
S: entropy
SAR: structure activity relationships
SASA: solvent accessible surface area
SMILES: Simplified Molecular-Input Line-Entry System
SPR: surface plasmon resonance

Scientific publications

I. A. Borrel; L. Regad; H. Xhaard; M. Petitjean; A.-C. Camproux. PockDrug: A Model for Predicting Pocket Druggability that Overcomes Pocket Estimation Uncertainties *J. Chem. Inf. Model.* **2015**, *55*, 882–895.

Reproduced by permission from Journal of Chemical Information and Modeling, Copyright (2015). American Chemical Society.

II. A. Borrel*; H. A. Hussein*; C. Geneix; M. Petitjean; L. Regad; A.-C. Camproux. PockDrug-Server: A New Web Server for Predicting Pocket Druggability on Holo and Apo Proteins *Nucleic Acids Res.* **2015**, 1–7.

Reproduced by permission from Nucleic Acids Research, Copyright (2015). Oxford Journals.

III. A. Borrel*; Y. Zhang*; L. Ghemtio; L. Regad; G. Boije af Gennäs; A.-C. Camproux; J. Yli-Kauhaluoma; H. Xhaard. Structural replacements of phosphate groups in the Protein Data Bank *(Manuscript)*

IV. A. Borrel; A.-C. Camproux; H. Xhaard. Interactions of amine, carboxylic acid, imidazole, and guanidinium groups in proteins and protein-ligand complexes *(Manuscript)*

V. A. Turku; **A. Borrel;** T. O. Leino; L. Karhu; J. Kukkonen; H. Xhaard. A pharmacophore model to discover OX1 and OX2 orexin receptor ligands. *J. Med. Chem.* **2016***, (Accepted)*

* Equal contribution.

Personal contributions

Publication I: I developed the automatic scripts for pocket estimation and pocket descriptors calculation, built and validated the statistical models and wrote the manuscript.

Publication II: I adapted the model for a webserver, improved the computational performance, benchmarked the model against other webservers and proofread the manuscript and the html pages.

Publication III: I wrote the fully streamlined computational workflow, prepared the Figures and Supplementary Material and participated in analysing the data and writing the manuscript.

Publication IV: I extracted the data, performed the data analysis and wrote the manuscript.

Publication V: I wrote the computational code for the pose selection protocol. I mapped the library and hit compounds in the chemical space using PCA and descriptors.

Unpublished results

Predicting water molecules in protein-ligand binding sites and their favorable displacements.

Additional publications

VI. S. Ménigaud; L. Mallet; G. Picord; C. Churlaud; **A. Borrel;** P. Deschavanne. GOHTAM: A Website for "Genomic Origin of Horizontal Transfers, Alignment and Metagenomics" *Bioinformatics* **2012**, *28*, 1270–1271.

VII. M. Francescatto; S. M. Hermans; S. Babaei; E. Vicedo; **A. Borrel;** P. Meysman. Highlights from the Third International Society for Computational Biology (ISCB) European Student Council Symposium 2014 *BMC Bioinformatics* **2014**, *16*, A3.

VIII. H. A. Hussein; **A. Borrel;** L. Regad; D. Flatters; A. Badel; C. Geneix; M. Petitjean; A. C. Camproux and O. Taboureau. System Biology: a new paradigm for drug discovery, The Practice of Medicinal Chemistry, Fourth Edition - **2015** - ISBN: 9780124172050 (C.-G. Wermuth, D. Aldous, P. Raboisson and D. Rognan, Academic Press)

IX. H. A. Hussein; C. Geneix; M. Petitjean; **A. Borrel;** D. Flatters and A. C. Camproux. Global vision of druggability issues, applications and perspectives. *Drug Discovery Today* **2016**, *(Accepted)*

All scientific publications are referred to in this thesis by their Roman numerals.

Introduction

Molecular recognition events are central to the biochemistry of life, yet not fully understood. A better understanding of these events would have major applications, for example, towards the discovery of chemical probes or therapeutically active molecules. A long-term vision in the field of computational drug discovery is to use statistical models to predict from any protein, or more precisely, from any protein binding pocket, synthesizable compounds with favourable drug-like properties that are able to bind their target with high affinity.

This thesis was conducted under a cotutelle agreement between the laboratories "Molécules Thérapeutiques *in silico*" at the University Paris Diderot and the Computational Drug Discovery Group, Division of Pharmaceutical Chemistry and Technology at the University of Helsinki. It tackles the development of computational methods and computational tools using three-dimensional protein structure data as input. The tools aim at better understanding of the phenomena associated with molecular recognition.

At the start of the thesis project, I developed a statistical model to predict druggability, i.e. whether a protein pocket can bind drug-like molecules with high affinity. I have identified several limitations of the previous statistical models, (i) they are invariably associated with a pocket estimation method, which limit their applicability; (ii) they have weak accuracy when using pockets from apo proteins; and (iii) they are not usually available to the scientific community. I thus developed a new model for druggability prediction, PockDrug (Publication I). This new model combines information obtained from several different pocket estimation methods. In order to develop the model, I needed to re-implement a set of pocket descriptors representing composition, physicochemical and geometric properties of a pocket. PockDrug is now available for the scientific community through a web server: http://pockdrug.rpbs.univ-paris-diderot.fr/ (Publication II).

The focus of the thesis was then directed towards improving our understanding of the ability of proteins to accommodate different ligands, or more precisely ligand fragments, at a local level. I developed a method to mine protein structures for ligand replacement functional groups. The method is written in the form of a computational workflow that can be fully parameterized. The workflow first finds proteins bound to reference ligands and then superimposes

protein-ligand complexes of the same protein crystallized with other ligands. The method is used to study phosphate replacements from the Protein Data Bank (PDB). I identify especially non-polar replacements as being surprising, and are used to discuss molecular mechanisms accompanying the replacements, such as arrangements of water molecules, ion coordination or protein displacements. In addition, the U-shape of ligands at nucleotide binding sites across phylogenetically unrelated proteins is observed (Publication III).

In observing the diversity of ligands, acidic and basic groups, such as carboxylate or amino and imidazolyl groups, appear to be relatively frequent. Nonetheless, the probability (frequency) of a ligand to form a salt bridge given that it contains a basic or acidic group has not been quantified to date. In contrast, these type of interactions have been relatively well studied for proteins. A limitation is probably the difficulty in mining specific groups in three-dimensional structures of proteins. For five basic and one acidic group, I thus investigated the prevalence of salt bridges. The distribution appeared overall similar to that in proteins for the lysine and arginine side chains (~50% and ~70% of ligand primary amines and guanidinium involved in salt bridges). The lowest proportion (16%) of salt bridges for tertiary amines appears to be connected to a lower volume of the space available around the functional group. In the absence of strong carboxylate-mediated salt bridges, the environment around the functional groups appeared enriched in functional groups with acidic properties such as hydroxyl group, phenol or water molecules (Publication IV).

Encountering difficulties in the docking scoring program to rank pose docking, I decided to write an application to visually analyse sets of poses generated in comparison with a bound reference. The tool provides a visual output that combines the ShaEP score with molecular interaction fingerprints and was applied to analyse induced-fit poses computed by the Glide software for a screening hit compound at the orexin receptors (Publication V).

All of these studies served as a reminder that water molecules play an important role in protein-ligand recognition, e.g. mediating molecular interactions and explaining structural replacements. However, only few computational methods exist to characterize them; these are often not freely available and may require extensive parametrization to, for example, conduct molecular dynamic simulations. Therefore, I started the development of a computational method to position water molecules and to assign for each water

molecule a so-called desirability index. Only preliminary results had been obtained at the time of writing this thesis.

For Publications III and IV, all scripts were released to the scientific community, in line with the need for open and reproducible computational science. The tool in Publication V will be released to the community as an application note.

Statistics available from the platform GitHub (https://github.com) about the number of lines of code produced through the thesis are presented in Figure 1. In total, about 40 000 lines of code have been written over a period of three and half years.

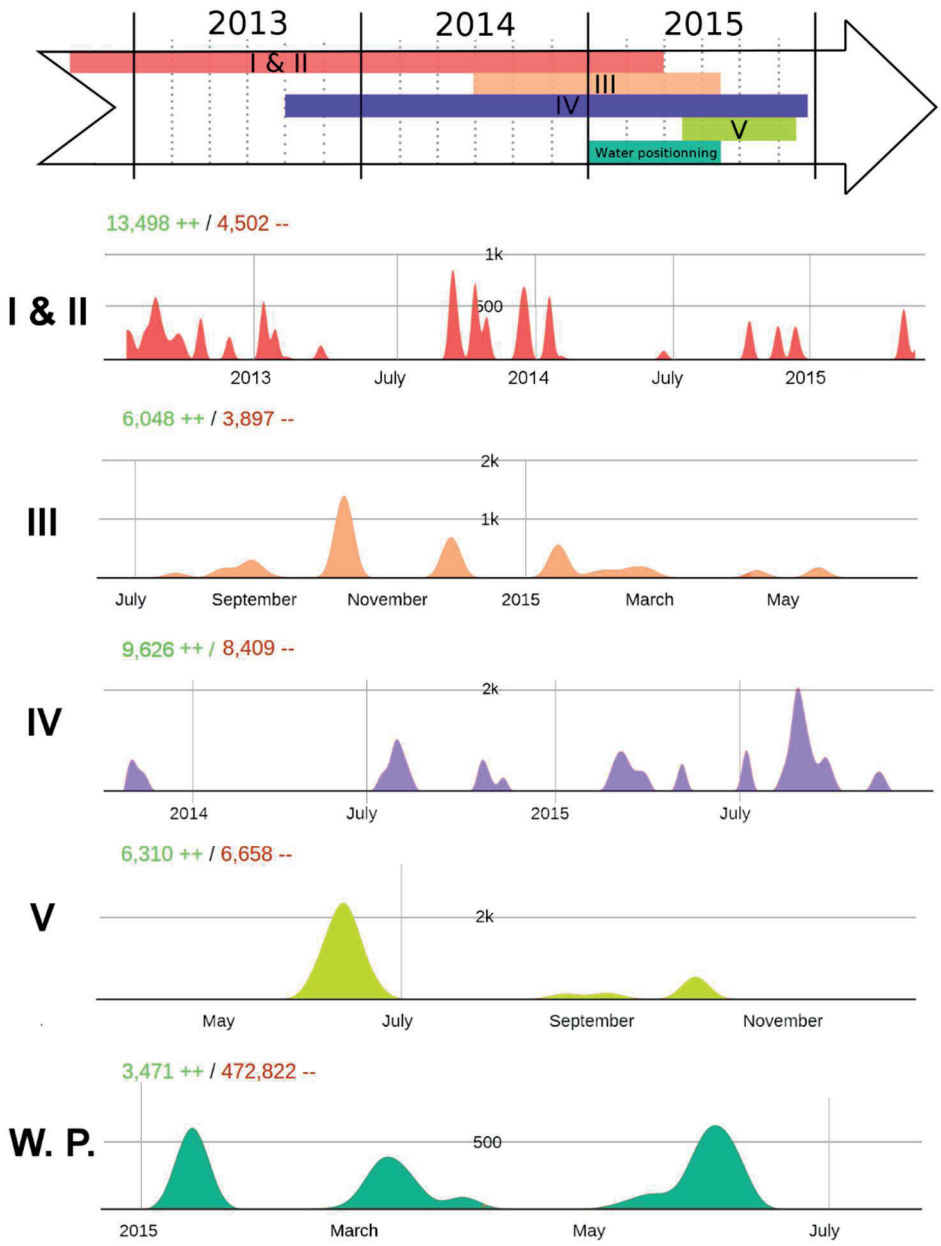


Figure 1: *Representation of programming code production using the representation available in the platform GitHub (https://github.com/). The histograms represent the number of lines function of time. Number of lines contained by project is shown in green, while the number of lines deleted is in red. The high number of deleted lines in the W.P. project is explained by the inadvertent downloading of the dataset into the source code.*

Review of the literature

The process of drug discovery aims at finding compounds that present a disease-modifying phenotype while achieving a concentration-time profile in the body adequate for the desired efficacy and safety. Discovery of new therapeutic molecules is a long, costly and challenging process; the success rate for phase II clinical studies is relatively low, ~15-20% (Arrowsmith, **2011**), and in 2015, only 33 new medical entities were accepted by the Food and Drug Administration (FDA) (Mullard, **2016**). Our incomplete understanding about the biochemistry of life arises from many factors such as complexity of the human body leading to unpredictable metabolic responses, efforts directed towards poorly druggable targets, the relatively limited chemical space explored by medicinal chemistry (Brown and Boström, **2015**; Dahlin et al., **2015**) and a too early focus on potency instead of early ADME-toxicity properties (Absorption, Distribution, Metabolism, Excretion and Toxicity) (Hughes et al., **2011**).

The process of discovering a drug molecule can be divided into three main steps (Kerns and Di, **2003**), presented in Figure 2. The work in this thesis is mostly aligned with the first step, drug discovery of a clinical candidate. It includes the following parts: (i) exploration in order to understand the disease, identify potential targets and discover hit compounds, usually through screening compound collections; (ii) lead selection to identify among hit compounds the most likely compound to be optimized as a successful drug candidate; (iii) lead optimization to find analogues with optimal ADME properties and a low risk of adverse effects through an extensive medicinal chemistry program; and (iv) development to investigate possible formulations. Subsequently, the second and third steps of drug discovery are divided into four clinical phases that involve testing with healthy volunteers and patients.

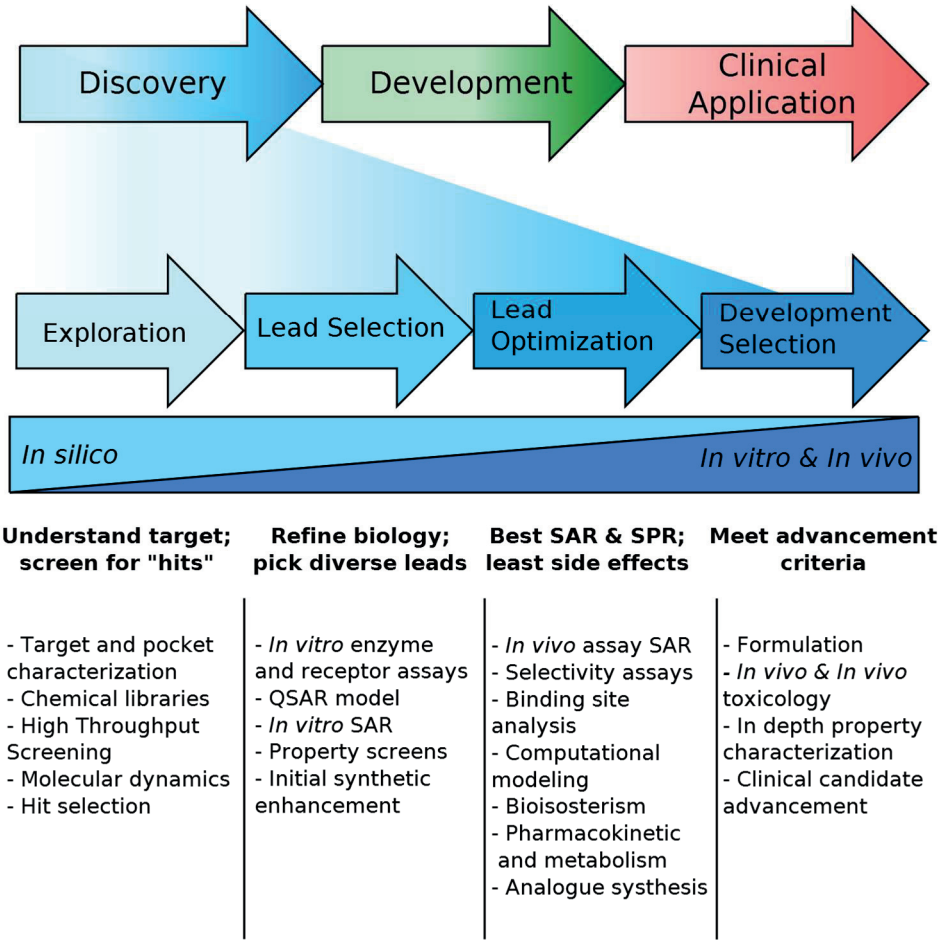


Figure 2: *Overview of drug research protocol, from Kerns (2008) (Kerns and Di, **2008**).*

Computational methods can play a prominent role in drug discovery, as shown in numerous reviews (Schneider and Fechner, **2005**; Ekins et al., **2007**; Roncaglioni et al., **2013**; Carbonell and Trosset, **2014**). The number of success stories in which *in silico* and *in vivo* approaches have been combined to develop a therapeutic molecule is constantly growing (Lambrinidis et al., **2015**; Unzue et al., **2016**). Computational methods for drug discovery include molecular modeling, docking simulations and virtual screening of compound collections, quantitative structure-activity relationships (QSAR) as well as quantitative structure-property relationship (QSPR) modeling. A large portion of these approaches is based on protein and ligand 3D structures, defining the fields of structure-based and ligand-based drug design. For more information, see for example the books entitled *Structure-Based Drug Discovery* edited by

Tari (2012) (Tari, **2012**) and *Drug Design* edited by Merz et al (2010) (Merz et al., **2010**).

In the Review of the literature section, some theoretical background is given to introduce the key concepts used. Section 1 presents structural data and their limitations, Section 2 binding sites that help to introduce the concept of druggability in Publications I and II, and Section 3 thermodynamics of protein-ligand recognition, serving as an introduction to Publications III and IV. Section 4 presents molecular interactions and docking simulations, which are used in Publication V and IV. Section 5 presents ligand optimization from the perspective of QSAR modeling in Publication I and bioisosteric design of analogues in Publication III. More specific literature reviews can be found within the original publications.

1. Structural data

1.1. Structural databases

The Protein Data Bank (PDB) is a freely available database, http://www.rcsb.org/ (last accessed 02-2016) (Berman et al., **2000**). It was established in 1971 at Brookhaven National Laboratory under the leadership of Walter Hamilton. When released, it contained seven structures. The PDB was the first established collaborative structural database. In February 2016, the PDB contained 115 918 biological macromolecular structures, 107 154 of which were proteins.

The Cambridge Structural Database (CSD) (last accessed 02-2016) is a database established by the Department of Chemistry, Cambridge University in 1965 (Allen, **2002**). It is an international repository for small-molecule organic and metal-organic crystal structures. It contains over 800 000 entries of high resolution from X-ray and neutron diffraction analysis.

1.2. Crystallography method

The work conducted in this thesis relies mostly on structural data, i.e. Cartesian atomic coordinates, which reflects the usefulness of structure-based methods in drug discovery (Williams et al., **2005**). X-ray crystallography is the most popular method to obtain this data: the proportion of structure elucidated using X-ray crystallography in the PDB is about 89.3% to date (last accessed 02-2016) (Berman et al., **2000**). The principle of structure elucidation is presented in Figure 3 (Callaway, **2015**). To summarize the method, X-rays pass through a crystallized protein and the resulting waves create a diffraction pattern from which the position of atoms can be deduced.

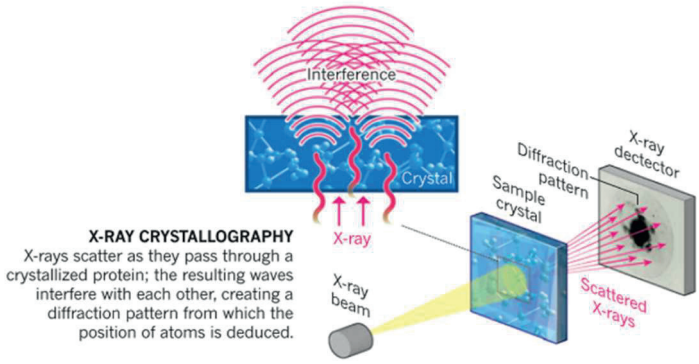


***Figure 3**: Principle of X-ray crystallography, adapted with permission from Callaway et al. (2015) (Callaway, **2015**)*

Since the quantity and quality of the X-ray structures can have a critical impact on the results obtained (Davis et al., **2003**), it is important to review their origin. Protein crystallization is frequently difficult, for example, for membrane proteins, which require highly expressed, purified and often solubilized proteins (Ilari and Savino, **2008**; Bill et al., **2011**). Nevertheless, the number of entries in the PDB is growing continually, improving the coverage from new co-crystallized ligands and conformational intermediate to novel protein folds.

The interferences of the X-rays in the crystal influence the structure resolution, as illustrated in Figure 3. Technical evolutions may emerge to overcome these limitations. Cryo-electron microscopy is a promising evolution of the classic X-ray crystallography. The crystal of the protein is a frozen sample, easier to handle, and the diffraction picture is obtained using an electron beam instead of an X-ray beam (Callaway, **2015**). Femtosecond crystallography (Miller, **2014**) is also an improvement of the crystallography method. X-rays with a wavelength of 10^{-9} m are replaced by X-rays with a wavelength of 10^{-10} m. This wavelength change allows reduction of the interference and elucidation of a protein structure from pseudo crystals (Chapman et al., **2011**).

In data mining studies, there is a fine balance between selecting too low-quality parameters, which may lead to erroneous structures being included in the datasets, and having enough examples to conduct statistical studies. Different criteria of quality can be taken into consideration such as resolution and Free R value (David Blow, **2002**).

1.3. Resolution

Resolution (Å) is a global criteria used to characterize the fuzziness of a crystallographic model. A poor resolution may be explained by different factors, such as X-ray interferences in the crystal, a low diffracting crystal, heterogeneities in the crystal, the mobility of the protein or the presence of multiple protein conformations inside the crystal. A resolution better than (inferior to) 4 Å is required to position individual heavy atoms in the structure. Structures with a resolution from 2 Å are considered of high quality (Lamb et al., **2015**). Figure 4 shows two structures at different resolutions, i.e. 3 Å and 0.48 Å. Hydrogen atoms and water molecules can be reliably positioned in the high-resolution structures, but not at 3 Å.

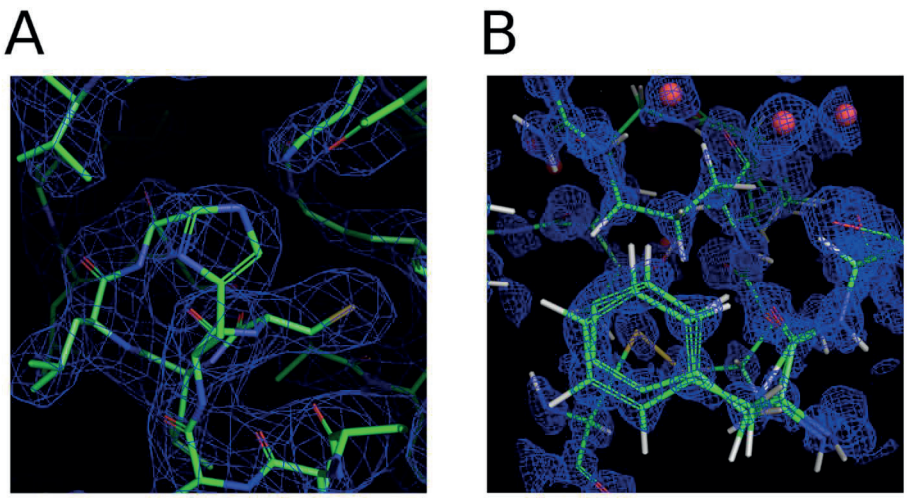


Figure 4: *X-ray map densities. (A) Equus asinus haemoglobin at 3 Å of resolution, PDB code 1S0H. (B) Crambe hispanica crambin protein at 0.48 Å, crystallized using a synchrotron with an intense X-ray source, PDB code 3NIR.*

1.4. Free R value (R-Free)

R- Free, in %, is a measure of the quality of the atomic model obtained from the crystallographic data (Brünger, **1993**). When solving the structure of a protein, the researcher first builds an atomic model and then calculates a simulated diffraction pattern based on that model. The R-Free measures how well the simulated diffraction pattern matches the experimentally observed

diffraction pattern. A completely random set of atoms will give an R-Free of about 63%. An R-Free inferior or equal to 25% is considered good in accordance to the good practice in the field (Brown and Ramaswamy, **2007**; Cooper et al., **2011**; Donald et al., **2011**).

1.5. PDB diversity

The contents of the PDB represent the work that has been conducted to date. It means that the PDB is "enriched" in well-studied proteins. For example, considering enzymes, hydrolases are well studied because they are involved in many diseases. The statistics given on the PDB web server (http://www.rcsb.org/) and reported in Figure 5 shows an over-representation of hydrolases (37.1%) and transferases (29.3%) relative to other functional classes.

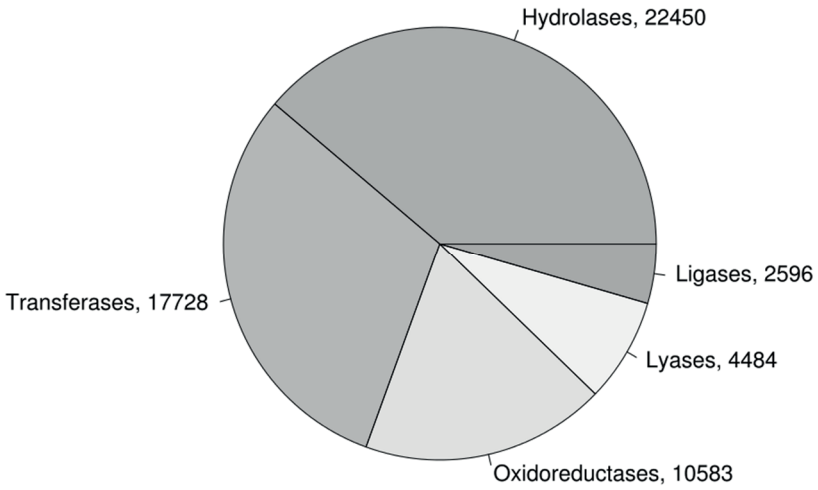


Figure 5: *Distribution of enzyme classes in the PDB from http://www.rcsb.org/, released February-2016.*

2. Protein-ligand recognition

The process of molecular recognition between a ligand and a host molecule is central to all biochemical processes, in particular to signal transduction, regulation of metabolic processes and gene expression. The term ligand (Latin gerundive *ligandum*, "that should be bound") was used for the first time by Alfred Stock in 1916 to define "affinities and valencies" for the ions able to bind analogous hybrids of silicon (Brock et al., **1983**). By the modern definition, a ligand may be a protein, peptide, DNA or RNA, or small molecule. Host molecules are generally proteins. The term "protein" was used for the first time in the correspondences between chemists Berzelius and Mulder in 1836 to define a type of organic matter (Hartley, **1951**). The word originates from the Greek root *proteios* "the first quality".

The "lock-and-key" theory about the recognition of ligands (substrates) by enzymes was first coined by Emil Fischer in 1894 (Fischer, **1894**). This theory demonstrated a complementarity of shape between a ligand (the substrate) and an enzyme. Yet, only with the emergence of crystallographic structure determination could protein-ligand interactions be better characterized; the first protein structure elucidated, myoglobin, was solved by crystallography in 1958(Kendrew et al., **1958**).

In 1958, Daniel Koshland proposed the "induced-fit model" to portray how an enzyme can adapt this conformation to interact with a ligand (Koshland, **1958**). This model is also referred to as the "hand-and-glove" model, to reflect the concept of flexibility of both the ligand and the host protein.

2.1. Binding sites, binding cavities and binding pockets

The vocabulary used in the literature to describe the regions where ligands interact with host molecules is somewhat fuzzy. Words such as "binding site", "binding pocket" or "binding cavity" are often used interchangeably. For example, a binding pocket is defined as being equivalent to a binding site (Kufareva et al., **2012**) whereas to others it implies a cavity on the protein surface (Cammisa et al., **2013**).

In this thesis, a binding site will refer to the atoms of the amino acid at interacting distances (4 to 6 Å) of a bound ligand, and present at the surface of the binding region. Binding sites may be located in cavities, i.e. concave

regions of the protein surface (sometimes occluded cavities), at the hinge between protein domains (sometimes referred to as "Pacman" binding sites) or at a protein-protein interface, which often results in flat and large interacting surfaces (Kuenemann et al., **2015**). Different examples of binding sites are presented in Figure 6.

A binding pocket will describe the region where ligands bind; a common situation is thus that binding sites partially overlap at a given binding pocket. Binding cavity will be restricted to cases of pockets, where a cavity is present in the host protein. It is important to note that some cavities are too small or do not have suitable properties for favourably interacting with a ligand, and as such cannot be *binding* cavities. The term "decoys sites" will be used for pockets not able to bind a ligand, as suggested by Desdouits et al. (2014) (Desdouits et al., **2014**).

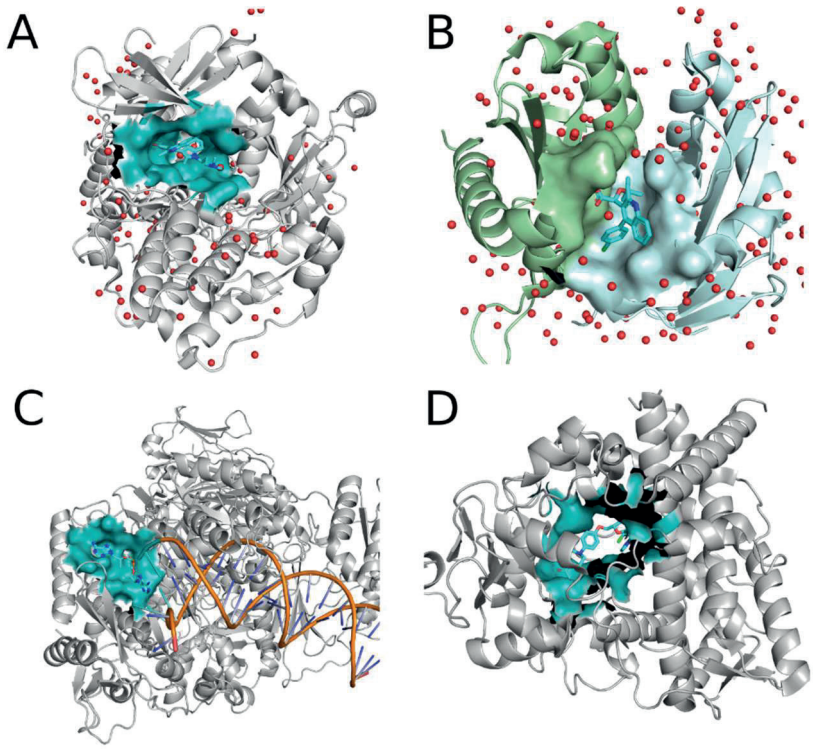


Figure 6: *Examples of protein binding pockets. (A) Binding site of Homo sapiens protein kinase C with bisindolylmaleimide inhibitor (PDB code 2I0E). (B) Binding site of human immunodeficiency virus 1 integrase complexed with (2S)-tert-butoxy[4-(4-chlorophenyl)-2-methylquinolin-3-yl]ethanoic acid in protein-protein interface (PDB code 4NYF). (C) Binding site of human immunodeficiency virus 1b complexed with tenofovir in protein DNA interface (PDB code 1T03). (D) Binding site of Homo*

sapiens microsomal P450 1A2 protein complexed with alpha-naphthoflavone (PDB code 2HI4). Binding sites in A, B and D are shown in blue and water molecules in red. In C, blue and green indicate different monomers.

Ligand activity is important for the function of the host protein, leading to the conservation of binding sites during evolution. Binding sites are usually associated with a lower mutation rate than other protein surfaces. The lower mutation rate of binding sites received considerable attention in the 1990s, and it was demonstrated for the first time by Zvelebil et al. (1987) (Zvelebil et al., **1987**) and Livingstone and Barton (Livingstone and Barton, **1993**). A recent study by Tseng et al. (2009) using 100 enzyme families with thousands of sequences showed that binding sites have a ~20% higher sequence identity than the overall proteins (Tseng et al., **2009**).

In terms of amino acid composition, binding sites are often enriched in the amino acids Arg, His, Trp and Tyr, as shown by Villar and Kauvar (1994) using 50 diverse proteins (Villar and Kauvar, **1994**). The amino acid composition at the binding site reflects the physicochemical properties of co-crystallized ligands (Arenas-Salinas et al., **2014**).

2.2. Environment

2.2.1. Co-factors and metals

Binding sites are not necessarily composed only of protein atoms, but may also contain co-factors such as metal or organic co-factors. Co-factors have a role in chemical reactions as well as in molecular recognition. An example of such a reaction is the dehydrogenation reaction conducted by the dehydrogenase using the nicotinamide adenine dinucleotide as a donor or acceptor of hydrogen. The promiscuity between a metal and the binding site influences reactivity and the presence of reaction intermediates (reviewed by Pordea et al (2015) and Rebilly et al (2015) (Pordea, **2015**; Rebilly et al., **2015**)). Figure 7 presents the binding site of the pyruvate phosphate dikinase bound to a phosphonopyruvate and to a Mg^{2+} atom.

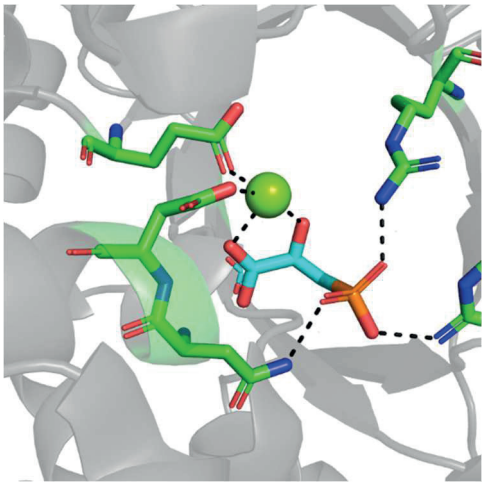


***Figure** 7: Example of pyruvate phosphate dikinase with bound Mg-phosphonopyruvate binding site including an ion Mg^{2+}, represented by a green sphere, coordinated by the ligand carboxyl and hydroxyl oxygen and by the protein carboxylate group (PDB code 1KC7). A hexadentate coordination is likely, perhaps with two missing water molecules.*

2.2.2. Water molecules

Water molecules at binding sites are presented in this section, while thermodynamic considerations about ligand binding are presented later, in Section 3. Hydrogen bonds (H-bonds) are discussed in more detail in Section 4.1.

The occupancy of a water molecule is ~9.2 $Å^3$, and a water molecule can be estimated by a sphere with a radius of 1.35 Å (Nicholls, **2000**). Water molecules ideally form four H-bonds in the solid state; they are composed of two hydrogen atoms, allowing two H-bonds to be donated, and one oxygen atom with two lone pairs that can accept two H-bonds. Water molecules have a permanent dipole moment of 1.8 Debye and are amphoteric, i.e. they have the ability to act as either an acid or a base in chemical reactions. Considering these properties, water molecules can interact with other water molecules, forming a network around biological macromolecules. The arrangement of water molecules inside the network is not fully understood; notably water networks of H-bonds have been reported to "flicker", i.e. to exchange hydrogen atoms by creation and disruption of H-bonds (Sanschagrin and Kuhn, **1998**).

Binding sites are most commonly solvated. Lu et al. (2007) showed that more than 85% of the protein binding sites in a dataset containing 392 high-resolution protein-ligand complexes contain at least one water molecule, with a mean of 4.6 water molecules per binding site (Lu et al., **2007**). In proteins,

water molecules can be classified as different types depending on their positions: (i) bulk, (ii) surface and (iii) buried (Levitt and Park, **1993**; Ladbury, **1996**). Bulk water molecules correspond to the first shell of protein hydration. Surface water molecules correspond to the shell of water further away from the protein. Exchanges between the bulk and the surface are possible and very common. Buried water molecules are at the interface between ligand and protein, and appear to be trapped between these two partners (Kahraman et al., **2007**). These water molecules play an important role in protein-ligand recognition (Bissantz et al., **2010**).

Water molecules have thus a critical role in protein-ligand interaction: (i) a driving force for hydrophobic contacts, (ii) mediation of molecular contacts with protein side-chains and (iii) solvation/desolvation events for the ligand, the protein and the protein-ligand complex (Nicholls, **2000**).

Water molecules play a prominent role in the hydrophobic effect, which is a critical component driving molecular recognition events (Tanford, **1979**; Southall et al., **2002**). This phenomenon is the driving "force" that aggregates lipophilic molecular surfaces to minimize the surface exposed to the solvent.

The International Union of Pure and Applied Chemistry (IUPAC, **2016**) defines the hydrophobic effect as follows:

"The tendency of hydrocarbons (or of lipophilic hydrocarbon-like groups in solutes) to form intermolecular aggregates in an aqueous medium, and analogous intramolecular interactions. The name arises from the attribution of the phenomenon to the apparent repulsion between water and hydrocarbons. However, the phenomenon ought to be attributed to the effect of the hydrocarbon-like groups on the water-water interaction…."

2.3. Induced-fit and "hand-on-glove" models of protein flexibility

The induced-fit model (or "hand-on-glove" if ligand flexibility is considered) refers to the local rearrangements in protein structure occurring concomitantly with the binding of a ligand. It was first used to describe the ability of enzyme to bind ligands in conformations that mimic transition states, but was later generalized to all ligand-binding events. Numerous examples in the literature have demonstrated the flexibility of binding sites, e.g. the glycine-rich P-loop near the ATP site of kinases (Mazanetz et al., **2014**).

Najmanovich et al. (2000) have estimated that 60–70% of binding sites undergo some changes in conformation and orientation of side chains, based on a dataset composed of 980 non-redundant paired apo and holo proteins (Najmanovich et al., **2000**). This percentage is probably a true random sample of ligands and proteins (Cozzini et al., **2008**), since three types of proteins can be defined: (i) "rigid" proteins where ligand-induced changes are limited to relatively small side chain rearrangements; (ii) flexible proteins where relatively large movements around "hinge points" or at active site loops with concomitant side chain motion occur upon ligand binding; and (iii) intrinsically unstable proteins whose conformations are not defined until ligand binding. Databases are, for technical reasons, enriched with rigid proteins, which are easier to crystallize (Tompa, **2003**).

2.4. Computational identification of binding pockets

Computational identification of binding pockets is critical to any computational work about molecular recognition. The easiest way to proceed is to extract the protein atoms (or all atoms to include co-factors, etc.) in the vicinity of a bound ligand. This method is, however, restricted to situations where the three-dimensional coordinates of the complex have been solved or are based on the results of docking simulations, which in turn requires added information to pinpoint the precise location of the ligand binding site.

The problem is more complex if no information about a bound ligand exists. Three types of methods have been devised to estimate the amino acids or atoms that form the pocket surface in such a case: (i) geometry-based, (ii) energy-based and (iii) evolutionary-based (Pérot et al., **2010**).

2.4.1. Geometry-based estimation

Geometry-based estimations derive from an identification of cavities in protein surfaces. Two algorithms are mainly used: (i) a grid and spheres to screen the protein surface or (ii) Delauney triangulation to divide the protein surface into a unique diagram which is then screened with a sphere. For a review, see Zhou and Yan (2012) (Zhou and Yan, **2012**). Numerous examples of geometry-based pocket estimation software have been published in the literature such as POCKET (Levitt and Banaszak, **1992**), LIGSITE (Hendlich et al., **1997**), CASTp (Dundas et al., **2006**), PASS (Brady and Stouten, **2000**), SCREEN (Nayal and Honig, **2006**), PocketPicker (Weisel et al., **2009**), Fpocket (Le Guilloux et al., **2009**), MSdock (Xie and Hwang, **2012**) and Cavitator (Gao and Skolnick, **2013**).

2.4.2. Energy-based estimation

Energy-based pocket estimation methods derive from the mapping of molecular probes on a protein surface (Hall and Enyedy, **2015**). Probes are pseudo atoms, e.g. amino, carbonyl oxygen, carboxy-oxygen, hydroxyl, methyl or water molecules that are "approaching" the protein. Probe-protein atom interactions are usually tested at each point of a grid, including the entire protein. A score is then calculated based on an empirical energy function. The basis of this method was developed by Peter Goodford in 1985 (Goodford, **1985**) and called GRID. From GRID, numerous energy-based estimations have been developed such as FTMap (Brenke et al., **2009**), DoGSite (Volkamer et al., **2010**), DrugSite (An et al., **2004**), QSiteFinder (Laurie and Jackson, **2005**) and PocketFinder (An et al., **2005**).

2.4.3. Evolutionary-based estimation

Evolutionary-based pocket estimation methods derive from the identification of invariant or less-variant amino acids in sequence alignments, as demonstrated by Armon et al (2001) and Lichtarge and Sowa (2002) (Armon et al., **2001**; Lichtarge and Sowa, **2002**). Evolutionary-based pocket estimation methods are less popular than the other pocket estimations methods, as reflected by the low number of methods available in the literature, e.g. ConSurf (Armon et al., **2001**) or Rate4Site (Pupko et al., **2002**). However, this method

may be combined with information about the protein structure to refine the pocket boundaries. For examples, see LIGSITEcsc (Huang and Schroeder, **2006**), SURFNET-ConSurf (Glaser et al., **2006**), SiteMap from the last version (Halgren, **2009**), MetaPocket (Huang, **2009**) and FINDSITE (Skolnick and Brylinski, **2009**).

2.4.4. Limits of pocket estimation methods

Success of pocket estimation methods is tested by addressing the fraction of pockets found in a dataset. Individual protein pockets in test sets can be defined by, for example, the atoms (or amino acids) observed in contact (3.5 Å) with a ligand-bound (holo) form of the protein. The success of prediction of each individual pocket is then estimated by reporting the percentage of binding atoms found.

The overall success rate is usually high, over 80%: 92% for SiteMap using a test set of 538 complexes taken from the PDBbind database (Wang et al., **2005**; Halgren, **2009**), 90% for Q-SiteFinder using a test set of 134 proteins extracted from the PDB (Laurie and Jackson, **2005**) and 83% for Fpocket (Le Guilloux et al., **2009**) using three different datasets containing 48, 63 and 85 protein-ligand complexes, respectively. However, this accuracy depends on the dataset used. A comparison of the different pocket estimation methods has been conducted using a set of 5416 protein-ligand complexes. The authors showed that about 95% of the binding pockets were correctly identified by the software tested (siteFinder, fpocket, pocketFinder and SiteMap), with no difference between energetic and geometric algorithms (Schmidtke et al., **2010b**).

Classical algorithms are, however, often parametrized to identify pocket surfaces able to bind small molecules. For particular pockets, specific pocket estimation methods have been developed. For example, HSpred, based on energetic method, can be used to estimate protein-protein interaction sites (Lise et al., **2011**). CCCPP, based on a geometric approach, can be used for channels (Benkaidali et al., **2014**), and AlloPred, based on an energetic algorithm, has been developed for allosteric pockets (Greener and Sternberg, **2015**).

2.5. Pocket descriptors

A binding pocket can be represented in many ways. It is common to translate the three-dimensional information about atom types and locations into numbers, known as descriptors, in order to compare binding pockets or to build predictive statistical models. Examples of descriptors are the composition in amino acids or atoms (Li et al., **2006**; Milletti and Vulpetti, **2010**; Rao et al., **2011**), their polarity or hydrophobicity (Kyte and Doolittle, **1982**; Hoppe et al., **2006**) or geometric properties of the surface (Perola et al., **2012**; Petitjean, **2014**). Pocket descriptors can also be the number of spheres included in the pocket (for geometry-based methods) or the relative or absolute number of certain probes mapped (for energy-based estimation method). Some examples are shown in Figure 8.

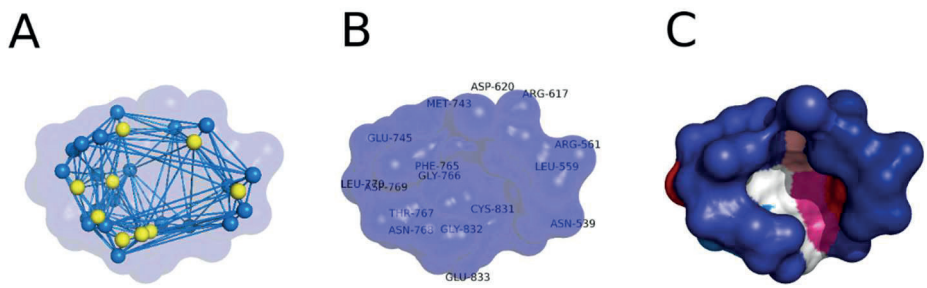


Figure 8: *Example of pocket properties visualized on Clostridium symbiosum phosphopyruvate binding site (PDB code 1KC7) estimated taking protein atoms within 6 Å of the ligand's atoms. (A) Geometric properties from convex hull. (B) Amino acid composition. (C) Hydrophobicity (red being more hydrophobic), based on the Kyte & Doolitle hydrophobicity scale (Kyte and Doolittle,* ***1982****)*

2.6. Pocketome

The pocketome (or pocket space) is defined as a set of pockets sharing similar characteristics. It has received considerable attention recently on the premise that similar pockets are likely to bind similar ligands. The pocketome offers thus the perspective of being able to predict ligands from pockets.

Efforts have been directed to establish pocketomes specific for certain protein families, e.g. the ATP pockets of human protein kinases (Volkamer et al., **2015**). Other studies have collected the ensemble ligandable pockets without

being restricted to certain families (Desaphy et al., **2015**). The Global pocketome, including all binding sites reported, is available in web site (http://pocketome.org) (last accessed January 2016) (Kufareva et al., **2012**). Klebe et al. (2015) have been storing pocket information as pharmacophoric points in order to perform fast comparisons (Krotzky et al., **2015**).

Critical to establishing and exploiting the pocketome is our ability to conduct pairwise comparisons of binding pockets. The methods to compare pockets can be divided into geometry-based and signature-based categories and have been reviewed by Nisius et al (2011) (Nisius et al., **2011**).

2.6.1. Geometry-based pocket comparison

Geometry-based methods define binding pockets by a cloud of points in a 3D space, and the analysis is driven by the comparison of the clouds of points representing each pocket. Typical points are amino acids, atoms, pseudo-atoms or surface points. Characteristic features, such as atom type, residue type or physicochemical properties, can be included in the similarity assessment.

Examples of implementation are eMatchSite using amino acid sequence alignments to compare two pockets (Brylinski, **2014**), PSIM aligning pseudo atoms flanking the pocket borders (Spitzer et al., **2014**), SILIRID aligning a cloud of points defined by close-distance interaction from a ligand (Chupakhin et al., **2014**) and APoc using a carbon-α structural alignment (Gao and Skolnick, **2013**). Other examples are SiteEngine (Shulman-Peleg et al., **2005**), SiteBase (Gold and Jackson, **2006**) and the method suggested by Krotzky et al. (2014) of map clouds of points pondered by physicochemical properties or pharmacophoric features that carry information about hydrogen bonding potential, hydrophobicity or polarity (Krotzky et al., **2014**).

2.6.2. Signature-based pocket comparison

Signature- based approaches define the binding pocket irrespective of its exact 3D coordinates. These methods are generally more robust towards small structural changes within the binding site, but may suffer from information loss. To cite a few recent methods, SiteAlign maps binding site properties into a discretized sphere, placed at the centre of the pocket (Schalon et al., **2008**);

CLIPPERS approximates the pocket using a travel depth algorithm and compares the pocket depth path (Coleman and Sharp, **2010**); FuzCav compares pharmacophoric points using six pharmacophoric properties and fingerprint comparisons (hydrogen bond donor or acceptor, positive ionizable, negative ionizable, aromatic, aliphatic) (Weill and Rognan, **2010**; Ito et al., **2012**); and PatchSurfers compares pockets using three-dimensional Zernike descriptors, a representation of the 3D function of Euclidean space using a 3D Zernike polynomial (Shin et al., **2016**).

2.7. Prediction of pocket druggability

2.7.1. Drug-like molecules

Compounds are not equal in their ability to be used as drug molecules. The term drug-like captures the concept that certain properties are advantageous with respect to compounds becoming successful drug products. The concept of drug-like compounds was coined in 2000 by Lipinski (Lipinski, **2000**):

"Drug-like is defined as those compounds that have sufficiently acceptable ADME properties and sufficiently acceptable toxicity properties to survive through the completion of human Phase I clinical trials".

2.7.1.1. Rules-of-five

Examining structural properties of 2200 compounds extracted from the United States Adopted Names Directory, Lipinski et al. (1997) showed that ninety percent of the compounds with poor absorption or permeation had (i) number of H-bond donor superior to five (expressed as the sum of the OH and NH pairs in the molecule), (ii) molecular weight superior to 500 Da, (iii) logarithm of coefficient partition (log *P*) superior to 5 and (iv) number of H-bond acceptor superior to ten (expressed as the sum of Ns and Os). Lipinski translated this definition into a set of rules, presented for the first time in 1997, which became famous as the "rules-of-five" (Ro5) (Lipinski et al., **2001**) (publication re-edited in 2001). A second paper by Lipinski in 2000 examined a larger set of molecules in phase II clinical trials, 10 000, and enunciated the Ro5 from the perspective of a drug-like compound (Lipinski, **2000**). Although the original definition did not include the concept of oral absorption, specific analyses on

orally absorbed drugs found that they comply well with the Ro5 (Wenlock et al., **2003**; Vieth et al., **2004**; Proudfoot, **2005**).

The rationale for the Ro5 is strongly connected to physicochemical properties (Kerns and Di, **2008**). A large number of H-bonds increases solubility in water (greater ability to interact with water molecules) and reduces partitioning from the aqueous phase into the membrane. A large molecular weight reduces the compound solubility and reduces the passive diffusion through the membrane. Log P also decreases aqueous solubility, which reduces absorption (Abraham et al., **2000**).

2.7.1.2. Consequences of Ro5 on drug development

A recent analysis (2015) show that among the 1543 drugs approved by the FDA and deposited in Drugbank (Law et al., **2014**), 1318 (85.4%) obey the Ro5 (Tian et al., **2015**).

The Ro5 have had a major impact in the community, as seen by the more than 1000 citations in CAS SciFinder by the end of year 2004 (Lipinski, **2004**). Rules are typically used to anticipate drug-like properties and select compounds to be progressed or to prioritize the chemical space used for high-through-put screening and virtual screening. Indeed, the number of theoretically accessible small molecules is in the order of ~$1.2*10^9$ molecules (Hann and Oprea, **2004**; Ursu et al., **2011**), but only a small proportion of them can be drug candidates (Dobson, **2004**). The large impact of the Ro5 can be explained in several ways (see the book of Kerns et al., 2008 for discussion): (i) these rules are easy to understand and intuitive, (ii) their number is limited and easy to remember, (iii) they are easy (fast, no associated cost) to implement and (iv) they are based on a strong physicochemical rationale (Kerns and Di, **2008**).

The Ro5 were originally perceived more as a profiling tool to serve as a guideline for drug discovery than as a strict filter, as discussed by Lipinski himself (Lipinski, **2004**, **2005**). Nonetheless, the impact of the Ro5 on the scientific community was so large that it set pressures on the pharmaceutical industries to include them as hard filters in compound development projects, despite these rules not being universal, for review see Leeson et al (2007) and Abad-Zapatero (2007) (Abad-Zapatero, **2007**; Leeson and Springthorpe, **2007**).

2.7.1.3. Alternatives to the Ro5

Alternative sets of rules and compound profiling methods have been developed. For example, the Veber rules suggest that a compound suitable for oral bioavailability should have (i) number of rotatable bonds less than ten, (ii) polar surface area less than 140 $Å^2$ and (iii) number of hydrogen bond (H-bond; see also Section 4.1) donors/acceptors less than 12 (Veber et al., **2002**). The "rules of three" suggest that fragment hit suitable for optimization should have (i) molecular weight less than 300, (ii) polar surface area less than 60 $Å^2$ and less than three of each of the following: (iii) number H-bond donors, (iv) number of H-bond acceptors, (v) ClogP and (vi) number of rotatble bond (Congreve et al., **2003**).

Any set rules suffer from the use of strict cut-offs (Yusof and Segall, **2013**), which can be alleviated using desirability indices and machine-learning based profiles. The quantitative estimate of drug-likeness index is based on the fit of the distributions of eight properties for 771 marked oral drugs (Bickerton et al., **2012**). The eight properties comprise molecular weight, Log P, number of H-bond donors and acceptors, polar surface area, number of rotatable bonds, number of aromatic ring and number of alerts for undesirable substructures based on Brenk et al. (2008) (Brenk et al., **2008**). This method tolerates more compound as drug-like than strict cutoff rules. Machine learning methods and sets of molecular descriptors are also widely used. Machine-learning methods are built, for example, on support vector machines (Byvatov et al., **2003**; Zernov et al., **2003**; Müller et al., **2005**; Li et al., **2007**), neural networks (Sadowski and Kubinyi, **1998**; Frimurer et al., **2000**; Byvatov et al., **2003**; Takaoka et al., **2003**), genetic algorithms (Gillet et al., **1998**, **2002**) or Bayesian classifiers (Yusof and Segall, **2013**).

A major limitation of any rule is the underlying data, which reflects the current chemical space used for drug discovery. Indeed, any model has an applicability domain, which is always limited by the compounds used for training (Tian et al., **2012**).

2.8. Druggable *versus* non-druggable pockets

Similarly to the concept of drug-like molecules used to rationalize the chemical space of compounds, the concept of druggability defines a protein able to bind a drug-like molecule with relatively high affinity (Hopkins and Groom, **2002**). Assessment of druggability comes from the observation that proteins are unequal in our ability to modulate them with small molecular weight compounds, as exemplified by the very different hit rates obtained in different NMR fragment screening studies (Hajduk et al., **2005**). More precisely, Hajduk et al. (2005) define a druggable pocket as having an affinity constant below 10 μM for a drug-like molecule (Hajduk et al., **2005**).

The assumption underlying the concept of druggability is that the physicochemical properties relevant to drug-like small molecules are reflected in the physicochemical properties of the protein pocket. Druggable pockets should have a volume sufficient to host a drug-like molecule, be strongly hydrophobic and contain a polar component. Perola et al. (2012) defined five properties to characterize druggable pockets, akin to the drug-like Ro5 properties (Perola et al., **2012**): (i) a volume higher than 500 $Å^3$, (ii) a depth higher than 10.4 Å, (iii) an enclosure score higher than 0.28 (score characterizing buried pockets), (iv) proportion of charged residues lower than 26.3% and (v) a hydrophobicity index higher than -1.12 (based on the scale of Kyte and Doolittle (Kyte and Doolittle, **1982**)).

Two examples, one for a druggable pocket, the other for a less druggable pocket, are presented in Figure 9. Both examples are taken from the kinase family and from the Krasowski's dataset (Krasowski et al., **2011**).

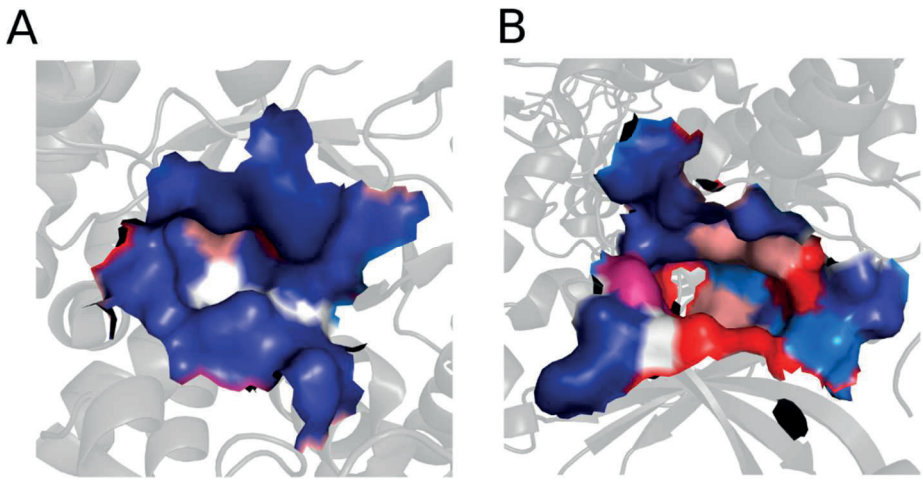


Figure 9: *Representation of druggable and less druggable binding sites. (A) Less-druggable binding site of the ligand phosphopyruvate on Clostridium symbiosum pyrophosphate d kinase, PDB code 1KC7. (B) Druggable binding site of the ligand bisindolylmaleimide inhibitor on Homo sapiens protein kinase C beta II, PDB code 2I0E. Polarity are coloured from hydrophobic gradient, blue (hydrophobe) to red (polar). The proteins share 18.1% of sequence identity when considering a structural alignment of the kinase domain (71/393 aligned positions).*

2.8.1. Computational models to predict binding pocket druggability

Predicting pocket druggability is a challenging procedure with many applications, principally to identify cavities that could be used for binding small molecules or to guide the choice of a target during a drug development programme (Pérot et al., **2010**).

The methods used to predict druggability models are similar to those employed in QSAR modeling (see Section 5.3). Numerical descriptors are used to define the pocket and used to predict an outcome (druggable/non-druggable) that may be associated with a quantity (druggability score) but could also be a probability of classification. Table 1 presents an overview of the different druggability models proposed in the literature. Druggability models usually come as a "package" tied to a pocket estimation method.

.

Table 2: *Overview of druggability models. * models available in a webserver or software. NRDD: Non-Redundant Druggability Dataset. NRDLD: Non-Redundant dataset of Druggable and Less Druggable binding sites. D: druggable pocket. ND: non-druggable pocket.*

Model	Pocket estimation	Dataset	Descriptors
(Hajduk et al., **2005**)	Geometric criteria only, flood-fill algorithm (Insight II, Accelrys)	Hajduk's set 28 D, 29 ND	13 terms representing polar and lipophilic surface areas, surface complexity and dimensions (Linear Regression)
SCREEN (Nayal and Honig, **2006**)	Molecular surface cavity depth detection (GRASP) (Nicholls et al., **1991**)	100 protein–ligand complexes (Perola et al., **2004**).	Only geometric, GRASP score (Random Forest)
MAP_{POD} (Cheng et al., **2007**)	MOE SiteFinder alpha-spheres based estimation (Labute and Santavy, **2010**)	Cheng's set 17 D, 10 ND	Curvature and lipophilic surface area (Biophysical model, free energy estimated)
Dscore-SiteMap * (Halgren, **2009**)	Grid defined on the ligand position	Cheng's set 17 D, 10 ND	Hydrophobicity, size and enclosure (Linear Regression)
SiteScore (Gupta et al., **2009**)	Using SiteMap **(Halgren,** 2009**)**	Gupta's HTS data, 22 proteins and Hajduk's set	Polar, apolar surfaces, shape and volume (Regression using VALSTAT (Gupta et al., **2004**))
DLID (Sheridan et al., **2010**)	icmPocketFinder (Laurie and Jackson, **2005**)	290 000 pockets from PDB and Cheng's set as test set	Volume, buriedness and hydrophobicity (Linear Regression)
(Huang and Jacobson, **2010**)	Probe mapping using Dock version 3.5.54 (Lorber and Shoichet, **1998**)	Hajduk's set and DUD (Huang et al., **2006**) set 35 D, 37 ND	Binding site energy approximation based on OPLS force field (Biophysical model, free energy estimated)

(Schmidtke and Barril, **2010**)*	Fpocket (Le Guilloux et al., **2009**)	Schmidke's set (NRDD) 45 D, 20 ND including Cheng's and Hajduk's sets	Normalized local hydrophobicity density and polarity density, hydrophobicity score (Linear Regression)
DrugPred (Krasowski et al., **2011**)	Probe mapping using Dock version 3.5.54 (Lorber and Shoichet, **1998**)	Krasowski's set (NRDLD) 71 D, 44 ND including NRDD set	Contact area between probes and protein, hydrophobic area, polar surface, occurrence polar amino acids, hydrophobicity index (Partial Least Square Regression)
Volsite (Desaphy et al., **2012**)	Volsite, probe mapping	NRDLD 71 D, 44 ND	Pharmacophores cavities: H-bond acceptor, H-bond donor, H-bond acceptor and donor, negative ionizable, positive ionizable, hydrophobic, aromatic. (Support Vector Machine)
DoGSiteScorer* (Volkamer et al., **2012a**)	DoGSite (Volkamer et al., **2012a**)	NRDD 45 D, 20 ND	Set of 17 descriptors representing, among depth, relative number of amino acid apolar, polar, position and negative, volume, shape, surface lipophilic and solvent exposure (Support Vector Machine)
CAVITY-SCORE (Yuan et al., **2013**)	CAVITY (Chen et al., **2009**)	NRDLD 71 D, 44 ND	Enclosure, number of hydrophobic and H-Bond normalized (Linear Regression)
DrugFEATURE (Liu and Altman, **2014**)	Microenvironment searching FEATURE	Combine Hajduk's and Cheng's sets and few complexes	From comparison score between microenvironment known binding drug-like molecules and microenvironment include in the pocket

		from DrugBank	
FTMap* for druggability (Kozakov et al., **2015**)	FTMap (Brenke et al., **2009**)	Combine Hajduk's and Cheng's sets to validate their score	Hot spot geometric properties, e.g. dimension, number of probes size of probe cluster

2.8.2. Limitations and challenges of druggability models

A major limitation of druggability models is the low number of studies reporting druggability information (i.e. studies using drug-like molecules or fragment screens on multiple targets) that are available to the academic community (Fauman et al., **2011**; Nisius et al., **2011**). Only a few datasets are used in the literature, as shown in Table 1, namely only Cheng's and Hajduk's datasets (Hajduk et al., **2005**; Cheng et al., **2007**). Doak et al. (2015) suggested overcoming this limitation by broadening the drug-like and druggable target definitions, allowing the development of more diversified datasets (Doak et al., **2015**). Another way to loosen the definition is to extend the concept of druggability to include "bindability" or "ligandability", i.e. whereas the ligands of interest are not restricted to drug-like molecules (Sheridan et al., **2010**; Surade and Blundell, **2012**).

A consequence of the limited amount of data is the inability of druggability models to make reliable predictions for cases different from buried or occluded pockets, e.g. binding sites in protein-protein interactions, which are larger and flatter that small molecule binding sites (Jones and Thornton, **1996**; Jin et al., **2014**). To respond to this limitation, druggability models focussing on specific cases have been developed, e.g. for protein-protein interactions (Sugaya and Furuya, **2011**; Johnson and Karanicolas, **2013**), for all mammalian proteins in the PDB (Loving et al., **2014**) and for 565 proteins predicted from the *Pseudomonas aeruginosa* genome (Sarkar and Brenk, **2015**).

The second challenge for druggability prediction models is to consider explicitly protein flexibility, either motions of the binding site or transient pockets in the protein. Different statistical models of druggability taken from snapshots from molecular dynamic simulation trajectories have been developed (Seco et al., **2009**; Yugang et al., **2011**; Bakan et al., **2012**; Cuchillo

et al., **2015**). Molecular dynamic simulations require, however, parametrization, which limits their applicability on a large scale.

Lastly, each model depends on a specific pocket estimation method (Pérot et al., **2010**; Nisius et al., **2011**), making the predictions difficult to compare, especially when the pocket estimation method is not available. This furthermore prevents the estimation of druggability performance from cavities in the protein defined visually. From the 14 models presented in Table 1, only four models are readily available.

3. Principles of molecular recognition

The principles beyond binding pocket druggability are linked to the phenomena of molecular recognition. A rigorous way to describe these phenomena is through thermodynamics.

3.1. Fundamental thermodynamic equations

A binding process can be described as an association or a dissociation between a ligand (L) and a receptor (R):

$$[R] + [L] \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} [RL]$$

[R], [L], [RL]: concentration (mol·L^{-1}) for the ligand, receptor and complex.

The process is dynamic and depends on the association (k_{on}) and dissociation rate constants (k_{off}). The binding affinity constant (K_a) characterizes the affinity between a ligand and a host protein. The equilibrium constant (K_d) is defined by the dissociation constant:

$$K_d = \frac{k_{off}}{k_{on}} = \frac{[R][L]}{[RL]}$$

$$K_a = \frac{1}{K_d}$$

At equilibrium, K_d is connected to the temperature and entropic and enthalpic standard energies using the Van't Hoff equation:

$$\ln(K_d) = \frac{\Delta H^0}{RT} - \frac{\Delta S^0}{R}$$

ΔH^0: standard enthalpy (J·mol^{-1})

ΔS^0: standard entropy (J·mol^{-1})

T: absolute temperature (K)

R: ideal gas constant (8.3144621 J·mol^{-1}·K^{-1})

This relationships is used to estimate the spontaneity of a reaction, in our case the binding (association) of a ligand to a host protein. The standard free energy is connected to standard entropy and standard enthalpy and consequently to the K_d.

$$\Delta G^0 = -RT \ln\left(\frac{C^\theta}{K_d}\right) = \Delta H^0 - T\Delta S^0$$

R: ideal gas constant (8.3144621 J·mol^{-1}·K^{-1})

T: absolute temperature (K)

C^θ: initial ligand concentration (mol·L^{-1})

ΔH^0: standard enthalpy (J)

ΔS^0: standard enthalpy (J)

The spontaneous association takes place only, when the standard Gibbs free energy (in Joules) is negative

3.2. Experimental assessment of thermodynamic parameters

3.2.1. Measuring binding free energy

Affinity constants (K_d, K_i) are classically determined in pharmacological assays and can be used to estimate ΔG. Competitive binding assays are among the most commonly used methods to determine K_i of a compound. This assay uses the displacement (competition) of an unlabelled ligand (passive) against the labelled one (active). It presents the advantage that the passive ligands do not need to be labelled, which is impractical e.g. for screening purposes. A competition assay is usually performed so that the binding site of receptor is first saturated with the labelled ligand, then increasing concentrations of the unlabelled one are added, and displacement of the labelled ligand is measured. IC_{50}, the concentration of unlabelled ligand necessary to displace 50% of the labelled one, can then be estimated. K_i, which is the inhibition constant and as such independent of the labelled ligand binding affinity, can then be estimated from IC_{50} using the Cheng-Prusoff equation (Cheng and Prusoff, **1973**) when a simple mechanism is at hand (e.g. non-competitive inhibition in the case of an enzyme).

$$IC_{50} = K_i(1 + \frac{[L]}{K_m})$$

K_m: Michaelis constant

K_i: inhibition constant

[L]: ligand concentration (mol·L^{-1})

IC_{50}: half maximal inhibitory concentration (mol·L^{-1})

3.2.2. Surface plasmon resonance (SPR)

Surface plasmon resonance (SPR) is an optical technique used for characterization of intermolecular interactions (Helmerhorst et al., **2012**; Olaru et al., **2015**). The technique was published for the first time in 1983 for detection of gases (Liedberg et al., **1983**) and has since then been developed to measure the interaction between biological macromolecules (for review, see Schuck, **1997**). The principle underlying the method is detailed in Figure 10 (Cooper, **2002**). A surface plasmon is a surface charge density wave at a metal surface. The SPR method reposes on the changes of the refractivity index on a surface plasmon. The interactions between the fixed macromolecules and other molecules circulating in a flow channel perturb the wave in the metal film. The evolution of the refractivity index, measured in real-time, gives the kinetic of the interaction (see Figure 10B).

.

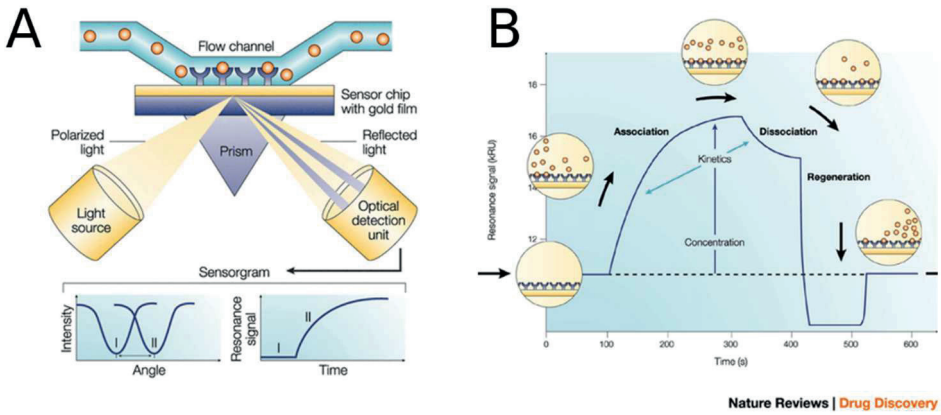


***Figure 10**: SPR principle. (A) Typical set-up for an SPR biosensor. (B) A typical binding cycle observed with an optical biosensor, adapted with permission from Cooper 2002 (Cooper, **2002**).*

The SPR technique is considered one of the most important analytical tools to measure and characterize biological interactions since it is non-invasive and accurate (Yadav et al., **2012**). It is furthermore label-free, i.e. it does not require a radioligand but requires a ligand immobilized.

3.2.3. Isothermal Titration Calorimetry (ITC)

Isothermal titration calorimetry is a biophysical technique that allows a full thermodynamic characterization of a protein-ligand interaction by measuring the heat evolved during molecular association. Indeed, when a protein interacts with a ligand, heat is either released (exothermic) or absorbed (endothermic).

An ITC is composed of two identical cells, a reference and a sample cell, surrounded by an adiabatic environment. The reference cell contains the buffer and the sample cell contains the macromolecules. Series of small aliquots of ligand are injected into the sample cell and, using a temperature differential sensor, the difference temperature between the two cells is measured in real-time. The heat transferred during the injection allows estimation of different thermodynamic variables such as entropy or enthalpy and the protein-ligand interaction kinetic (Peters et al., **2009**; Ladbury, **2010**). An example of ITC material and the data accessed is presented in Figure 11. The first ITC was developed by Laplace and Lavoisier in the 1850s.

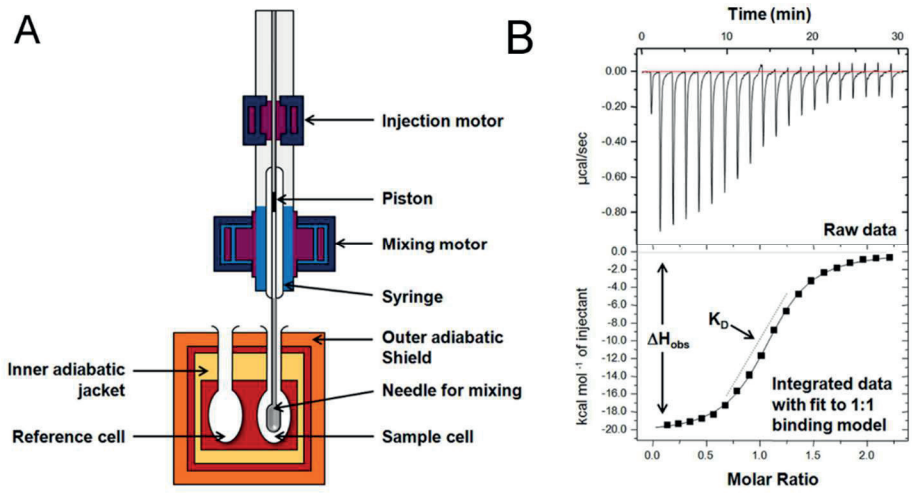


***Figure 11**: (A) Example of experimental ITC instrument. (B) Example of results of a characteristic titration experiment (upper) with the associated data analysis. Reproduced with permission from Geschwindner et al. (2015) (Geschwindner et al., **2015**).*

Sensitivity of ITC is a crucial factor. One of the limits of this method is that it detects the total heat effect in the sample cell upon addition of ligand; the total heat includes non-specific effects such as dilution of ligand in the buffer, dilution of the protein sample, heat of mixing, temperature difference between the cell and the syringe and mixing of buffers of slightly different composition (Böhm and Schneider, **2012**). Today, the quality of the measurements is close to 10^{-9} J using a nanocalorimeter, but requires significant amounts of purified protein (Perozzo et al., **2004**; Roselin et al., **2010**; Torres et al., **2010**). ITC is commonly used in the community, with more than 500 articles reporting its use (citations collected from Web of Science, PubMed, SciDir and OVID databases in 2010 (Ghai et al., **2012**)). Among these, more than 130 citations referenced a study about protein small molecules interactions.

3.3. Factors influencing binding free energy

The free energy of binding is central to understanding protein-ligand interactions, as stated by a pioneer in the field, Professor Peter Kollman:

"Free energy is arguably the most important general concept in physical chemistry" (Kollman, **1993**).

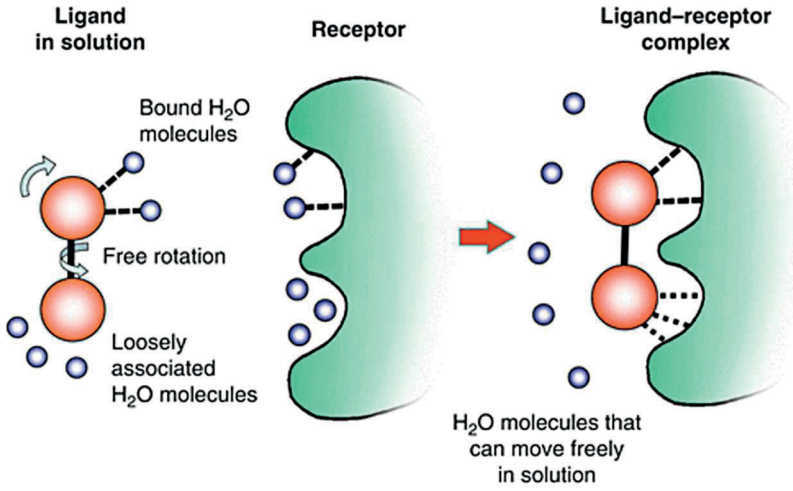


***Figure 12**: Illustration of the phenomena associated with a thermodynamic contribution to the binding free energy. H-bonds are indicated by dashed lines and hydrophobic interactions by dotted lines. Reproduced with permission from Klebe (2013) (Klebe, **2013**).*

Free binding energy is influenced by many factors that depend on the ligand's and receptor's intrinsic properties as well as the environment, especially the solvent. This is illustrated on Figure 12, reproduced with permission from the book "*Drug Design: Methodology, concepts, and mode-of-action"* (Klebe, **2013**). In Figure 12, protein and ligand are shown as dissociated and associated, and three different factors contributing to the binding free energy are reported. The first one is the translational or rotational degrees of freedom, illustrated by the ligand. When the ligand is free, it may be able to adopt different conformations. When the ligand is bound, the degrees of freedom are reduced. This contribution of the free energy is entropic by nature, relative to the order of the system. The second factor influencing the free energy is the interactions formed, which can be illustrated by the H-bonds between the ligand and the water molecules, between the protein and the water molecules and between the ligand and the receptor. These interactions are enthalpic in nature, however, the loss of water molecule-ligand interactions can lead to a

disordered water molecules network and generate an entropic contribution. Surrounding hydrophobic regions of the ligand and the protein, water molecules are also present, forming a "cage" of water molecules. Breaking this ordered network can be unfavourable to the protein-ligand interaction. Finally, when protein and ligand are bound, water molecules in solution are free and form a new water molecules network. The overall contribution of water molecules to the binding free energy is thus very difficult to quantitatively estimate (Tanford, **1979**; Huggins, **2012**; Biela et al., **2013**; Breiten et al., **2013**; Alvarez-Garcia and Barril, **2014**; Jeszenői et al., **2016**) (see also Section 5.4).

3.4. Computational estimation of the binding free energy change of a system

3.4.1. Principle of binding free energy estimation

The theoretical prediction of binding free energy is a "holy grail" of computational drug discovery methods. It has application towards the discovery of novel active compounds, the anticipation of side effects, and more globally about our understanding of the factors involved in the formation of protein-ligand complexes (Gilson and Zhou, **2007**; Steinbrecher and Labahn, **2010**; Ashida and Kikuchi, **2015**). Different methods to estimate the binding free energy have been presented in the literature. The most popular are (i) the free energy perturbation and the thermodynamic integration methods (Beveridge and Dicapua, **1989**; Kollman, **1993**), (ii) the molecular mechanics/ Poisson-Boltzmann surface area (MM/PBSA) (Kollman et al., **2000**) and (iii) the linear interaction energy methods (Aqvist and Marelius, **2001**). For a complete review, see Ashida and Kikuchi (2015) (Ashida and Kikuchi, **2015**).

A fundamental principle underlying binding free energy estimation is that the binding free energy is equivalent to the sum of the contributions of the thermodynamic components (group of atom, receptor, ligand) taken independently (ΔG_{ionic}, ΔG_{H-bond}, ΔG_{vdW}, $\Delta G_{solvation}$, ….) (Cozzini et al., **2004**). To illustrate this principle, the model MM/PBSA is presented below (Kollman et al., **2000**). This model is popular in drug design, with applications in protein design, protein-protein interactions, conformer stability and re-scoring function; it has been described in 100-200 publications a year over the last five years (Genheden and Ryde, **2015**). The model is an end-point model,

i.e. it calculates the binding free energy from molecular dynamics of bound and free ligand only, not for intermediate states.

The binding free energy is given by the equation:

$$\Delta G_{bind} = < G_{RL} > - < G_R > - < G_L >$$

$< G_{RL} >$,$< G_{RL} >$, $< G_{RL} >$: states of the system L ligand only, R receptor only and RL receptor ligand complexed.

For each state:

$$G = E_{bond} + E_{el} + E_{vdW} + G_{pol} + G_{np} - TS$$

E_{bond}, E_{el}, E_{vdW}: Energies of bond (bond, angle and dihedral), electrostatic and van der Waals terms computed using molecular mechanics (i.e. a force-field).

G_{pol}, G_{np}: solvent contributions, polar and non-polar, to the solvation free energies. G_{pol} is obtained by solving the Poisson-Boltzmann equation. G_{np} is estimated from the implicit solvent model, i.e. a linear relation to the solvent accessible surface area (SASA) (Connolly, **1983**). The entropy S is estimated by a normal-mode analysis of the vibrational frequencies based on the data assessed during the molecular dynamic.

To improve the quality of the computation, classically, the estimated binding free energy based on these different equations is averaged from three independent molecular dynamics (Swanson et al., **2004**).

3.4.2. Limitations and challenges for the models of free energy estimation

To date, the main limitation reported is accuracy. Predictions are system- and protocol-dependent (Mikulskis et al., **2014**), and accuracies generally relatively poor, e.g. a mean standard error between 11 and 14 kJ mol^{-1} has been reported (Genheden and Ryde, **2015**). Other limitations are high computational time and need of specific expertise, restricting their routine application (Steinbrecher and Labahn, **2010**; Ashida and Kikuchi, **2015**).

Water molecules represent another important challenge to free energy estimations. MM/PBSA uses an implicit solvent model based on solvent

accessibility surface (Connolly, **1983**). However, different models exist to estimate the contributions to the binding free energy of water molecules; the most popular being TPIP4 (Jorgensen et al., **1983**) and TPIP5 (Mahoney and Jorgensen, **2001**). TPIP4 and TPIP5 approximate water molecules as rigid bodies having four or five interaction points. Energy of the bulk water molecules has been estimated at 94.83 $kJ \cdot mol^{-1}$ for the TPIP4 model and 80.97 $kJ \cdot mol^{-1}$ for the TPIP5 model (Huggins, **2012**). These models consider each hydration site's need sequentially and are therefore extremely computationally intensive.

4. Molecular interactions

The major classes of molecular interactions between a protein and a ligand are the irreversible covalent interactions and the reversible non-covalent interactions.

Covalent interactions represent a particular case in protein-ligand interactions. There are several reasons for this, one being that the process is no longer at equilibrium, or that enzymes covalently linked with an inhibitor need to be degraded and re-synthesized for the enzymatic activity to be continued. In the drug discovery process, irreversible interactions are often associated with adverse effects (Mah et al., **2014**). Nonetheless, it is interesting to note that 30% of marketed drugs acting on enzymes do so through covalent interactions (Robertson, **2005**). A multitude of chemical mechanisms are used for covalent interactions, such as acylation, alkylation, metal-metalloid binding, disulfide-bond formation, hemiketal formation, Michael addition and Pinner reaction reviewed in Postashman and Duggan (2009) (Potashman and Duggan, **2009**).

Weak (non-bonded) interactions between atoms of a ligand and a protein are individually weak but collectively strong. The most common types of non-covalent interactions are presented in Figure 13 and discussed in more detail below: hydrogen bond (H-bond) (Section 4.1), salt bridge (Section 4.2) halogen bond (X-bond) (Section 4.3) and π-systems (Section 4.4). These, of course, do not represent an exhaustive list of all possible molecular interactions (Bissantz et al., **2010**).

In molecular modeling, non-bonded interactions are modelled using two types of forces that may be attractive or repulsive. The electrostatic force can be modelled using the Coulombic law and depends on the charge carried by the interacting atoms as well as the distance r separating them. Electrostatic forces are inversely proportional to r^2. The van der Waals forces (Keesom force, Debye force, London dispersion forces) depend on the presence of permanent or induced dipoles as well as on steric factors and are inversely proportional to r^4 or r^6.

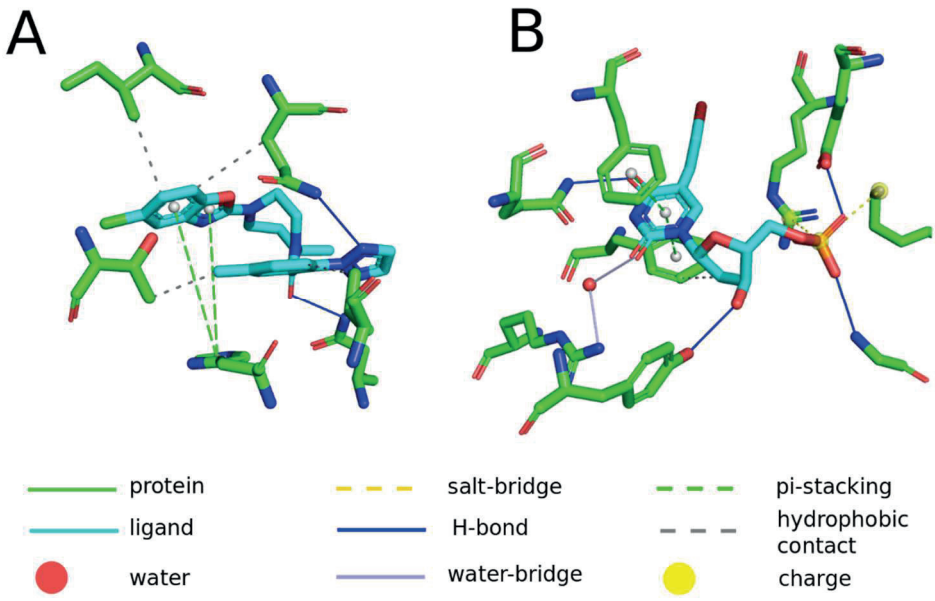


Figure 13: *Examples of non-covalent interactions in protein-ligand complexes. (A) Homo sapiens orexin receptor OX2 complexed with the insomnia drug suvorexant PDB code 4S0V. (B) human herpes virus 3 thymidine kinase in complex with (E)-5-(2-bromovinyl)-2'-deoxyuridine-5'-monophosphate, PDB code 1SON. Interactions are predicted using Protein-Ligand Interaction Profiler (Salentin et al.,* ***2015****).*

4.1. Hydrogen bond (H-bond)

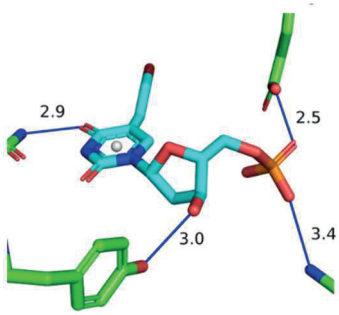


***Figure 14**: Example of H-bond network stabilizing (E) -5-(2-bromovinyl)-2'-deoxyuridine-5'-monophosphate in human herpesvirus 3 thymine kinase; PDB code 1OSN.*

Hydrogen bonds are classical non-bonded interactions, illustrated in Figure 14. A hydrogen bond is characterized by two partners: a hydrogen-bond donor (X-H), i.e. an electronegative atom bonded to a polarized hydrogen atom, e.g. an oxygen atom of a hydroxyl group; and a hydrogen-bond acceptor (Y), which may be an atom (can be an anion) or a group of atoms. An aromatic ring can serve as a hydrogen bond acceptor since its electron-rich π system above and below the aromatic ring hosts a partial negative charge (Du et al., **2013**).

The modern definition of H-bond from the IUPAC (Arunan et al., **2011**) is as follows:

"The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation. A typical hydrogen bond may be depicted as X–H•••Y–Z, where the three dots denote the bond"

The IUPAC defines a list of six criteria for hydrogen bonds (E1-6), and six characteristics (C1-6) particular to hydrogen bonds are summarized in Table 2 for a hydrogen bond X–H•••Y–Z.

***Table 2**: Summary of the criteria and characteristics particular to the H-bond interaction.*

	Criteria		**Characteristics**
E1	Forces involved in formation of the H-bond include those originating from electrostatics, from charge transfer, and from dispersion	C1	The pK_a of X•••H and pK_b of Y-Z in a given solvent correlate strongly with the energy of the H-bond formed between them
E2	The H•••Y bond strength increases with the electronegativity of X.	C2	The proton shared between donor and acceptor may be transferred between both partners (i.e. the case of salt bridge)
E3	The X-H•••Y is usually close to 180°. Closer to 180° indicates a stronger H-bond and shorter H•••Y distances	C3	The network of hydrogen bonds can show co-operativity, leading to deviations from pair-wise additivity in hydrogen bond properties
E4	Greater X-H length indicates stronger H-bonds. Spectroscopic considerations.	C4	H-bond have directional preferences.
E5	NMR considerations.	C5	Interaction energy correlates well with the extent of charge transfer
E6	H-bond should be thermally stable.	C6	Considerations about the electron density topology

A common way to study the characteristics of hydrogen bonds is to mine crystallographic data either from proteins (PDB) or from small molecules (CSD) (Bissantz et al., **2010**). Atomic distances reflect the strength of the H-bonds, weaker H-bonds being longer (Williams and Ladbury, **2003**). Different types of H-bond, i.e. very strong, strong and weak, can be defined as a function of the types of acceptor and donor atoms (see Table 3). In terms of X-H•••Y angle, structure analysis show that although this angle can vary, e.g. in the case of a bifurcated H-bond, it remains greater than 150° (Taylor et al., **1983**;

Lommerse et al., **1997**; Nobeli et al., **1997**). Bifurcated H-bonds are more common for weak H-bonds (Sarkhel and Desiraju, **2004**; Panigrahi and Desiraju, **2007**).

***Table 3**: Potential H-bond donor and acceptor groups classified according to their strength of interaction. X is any atom, Hal is any lighter halogen and M is a transition metal. Adapted from Williams et al. (2003) (Williams and Ladbury, **2003**).*

	Donor	**Acceptor**
Very strong	$N^{+}H3$, X^{+}-H, F-H	COO^{-}, O^{-}, N^{-}, F^{-}
Strong	O-H, N-H, Hal-H	O=C, O-H, N, S=C, F^{-} H, Hal^{-}
Weak	C-H, S-H, P-H, M-H	C=C, Hal-C, π, S-H, M, Hal-M, Hal-H, Se

H-bond energies depend mainly on the properties of the hydrogen-bond donor and acceptor groups. The H-bond contributions to the binding free energy are case-dependent. H-bond-driven contribution to the binding free energy appears to range from 2 to 8 kJ mol^{-1} (Klebe, **2013**). Six chemical "Leitmotifs" have been suggested to be able to describe the diversity of the observed H-bonds (Gilli et al., **1994**, **2009**; Gilli and Gilli, **2000**). In particular, dependencies between the strength of the H-bond interaction, the distance and the pKa value difference between the donor and acceptor atoms were observed (Gilli et al., **2009**).

4.2. Salt bridge

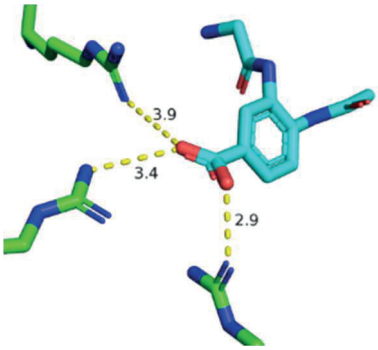


Figure 15: *Example of salt bridges between the carboxylate from the 4-(acetylamino) -3-[(aminoacetyl)amino] benzoic acid and the Influenza A virus neuraminidase, PDB code 1INH.*

Salt bridges are charge-assisted H-bonds formed between a negatively charged (acidic) functional group (H-bond acceptor; such as the side-chain carboxyl group from aspartate or glutamate) and a positively charged (basic) donor (such as side-chain functional groups of lysine, arginine or histidine). This type of H-bond is classified as a very strong H-bond and is characterized by a short distance between the interacting non-hydrogen atoms, median 2.79 Å (Bissantz et al., **2010**).

Salt bridges have been well characterized in proteins since they are key to protein folding, flexibility and thermostability (Hall et al., **2013**; Lee et al., **2014**; Meuzelaar et al., **2014**; Lotze and Bakker, **2015**). For example, salt bridges have been shown to play an important role in the stability of secondary structure elements (Sarakatsannis and Duan, **2005**; Donald et al., **2011**). From a set of 1500 proteins from the PDB (Gvritishvili et al., **2008**), 39.4% of lysine, 60.6% of arginine, 47.1% of aspartate and 52.9% of glutamate were shown to be involved in at least one salt bridge interaction (Gvritishvili et al., **2008**).

The energetic contribution of salt bridges is very difficult to quantify (Debiec et al., **2014**) since not only do they combine van der Waals and electrostatic energetic contributions but also charged H-bond acceptor and donor groups are strongly solvated. Salt bridges may also be mediated by water molecules; 32.8% of histidine, arginine and lysine side chains and 24.6% of asparatic acid and glutamic acid positive side chains are involved in a water-mediated salt bridge (Sabarinathan et al., **2011**).

4.3. Halogen bond (X-bond)

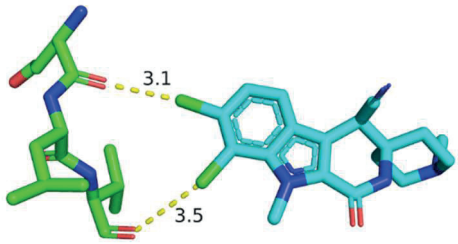


***Figure 16**: Example of X-bond interaction between the chlorines from the (4r)-7,8-dichloro-1', 9-dimethy l-1-oxo-1, 2, 4, 9-tetrahydrospiro [beta-carboline-3, 4'-piperidine]- 4-carbonitrile and the human protein kinase 3, PDB code 3BHY.*

Halogen bonds are comparable to hydrogen bonds, but it is the halogen atom that is shared between a donor and an acceptor. Halogen bonds (X-bond) were discovered in 1986 (Ramasubbu et al., **1986**). In a crystallographic environment, electrophiles such as metal ions tend to approach halogens of C-X (X = F, Cl, Br, I) at an angle of ~ 100° ("side-on"), while nucleophiles, such as oxygen and nitrogen, approach at an angle of ~165° ("head-on") (Ramasubbu et al., **1986**). An example of a halogen bond between chlorine (Cl) atoms and main-chain carbonyl oxygen atoms is presented in Figure 16.

The definition of a halogen bond from the IUPAC (Desiraju et al., **2013**) is given as:

"A halogen bond occurs when there is evidence of a net attractive interaction between an electrophilic region associated with a halogen atom in a molecular entity and a nucleophilic region in another, or the same, molecular entity."

This interaction is based on an electronic depopulation of the valence p_z orbital of the halogen atom, which forms an electropositive "hole" that can function as a halogen bond donor. Two interaction models have been suggested, the quantic σ-hole model and the electrostatic model lump-hole, presented in Figure 17 (Ford and Ho, **2015**).

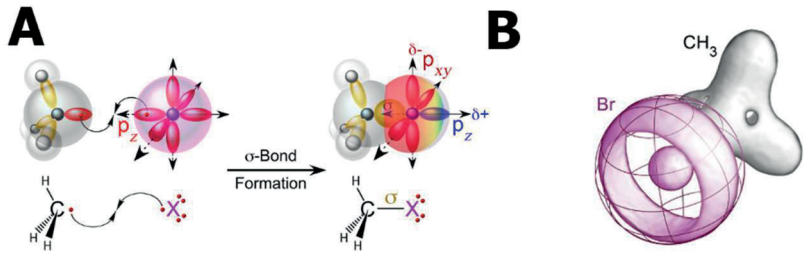


Figure 17: *(A) Example of σ-hole model: The formation of a covalent carbon–halogen bond (a C—X σ-bond) pairs the electrons from the valence orbitals of the two atoms. As a result, the p_z orbital of the halogen opposite the σ-bond becomes depopulated, resulting in an electropositive crown (in blue), whereas the px_y orbitals retain their complement of electrons to account for the overall negative charge of the halogen (reproduced with permission from Scholfield et al (2015); Figure legend from Scholfield et al (2015)) (Scholfield et al., **2015**). (B) Electron distribution of atoms in CH_3Br, as predicted from the lump-hole theory. The distribution of electrons forms a ring around the bromine centre (accounting for the majority of electrons at the atomic surface) and a "hole" at the surface that can interact with the "lump" of electrons from an interacting X-bond acceptor. The standard surface of the bromine atom is outlined as a spherical cage. Reproduced with permission from Ford et al (2015); Figure and legend from Ford et al (2015) (Ford and Ho, **2015**).*

X-bonds have a length of interaction that depends on the type of halogen atom, more specifically, the radius and polarizability of the halogen donor, F > Cl > Br > I (Ford and Ho, **2015**). For example, halogen bonds have been shown to range between 2.57 Å for C-F•••H bonds and 3.42 Å for C-I•••S bonds (Lu et al., **2010**). The angle of a halogen bond was found to be around 160°, but close to 100° in the case where the X-bond acceptor is an H-atom (Lu et al., **2010**)..

Halogen bonds may be found in numerous drug interactions; 50% of the current drugs are halogenated, often in order to increase membrane permeability and decrease metabolic degradation (Xu et al., **2014**). Contrasting with their importance, halogen bonds are poorly implemented in molecular force fields (Ford and Ho, **2015**).

4.4. Π-systems

A π-system is a conjugated system of connected p-orbitals, alternating single and double bonds. π-systems may be linear or cyclic. In proteins, π-systems are carried by the aromatic amino acids Trp, Phe, His and Tyr, as shown in Figure 18.

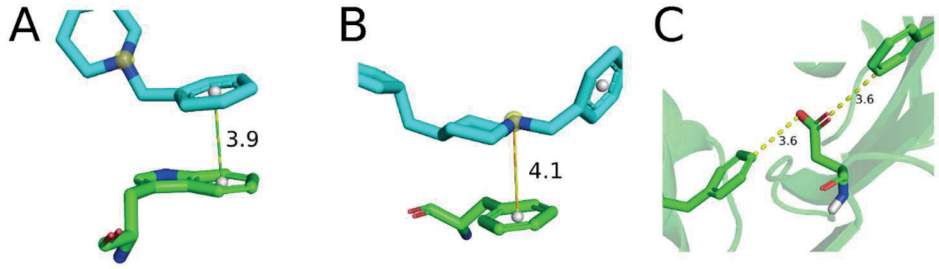


***Figure 18**: Examples of π-systems. (A)-(B) π-stacking and cation-π interaction from Torpedo californica acetylcholinesterase complexed with donepezil, PDB code 1EVE. (C) Intra-protein anion- π interaction from Comamonas testosteroni isomerase, PDB code 8CHO.*

π-stacking – The favourable interaction between two aromatic π-systems can be explained by a particular polarization of π-systems, which creates a quadrupole moment with partial negative and positive charges respectively positioned above both aromatic faces and in the ring periphery (Martinez and Iverson, **2012**). Interaction of polarized π-systems leads to several possible low-energy preferred arrangements such as stacked, T-shape, parallel displaced and reversed polarity (see Figure 19). In terms of energy, it has been estimated that the different configurations may be ordered as stacked < T-shape < parallel < reverse polarity (Tsuzuki et al., **2002**).

In terms of geometry, the distance between the centres of two interacting rings are in the range of 4.5–5 Å for a π-stacked configuration, and below 7.5 Å for the other configurations, as shown from a dataset containing 505 proteins (McGaughey et al., **1998**). In terms of angle, for the T-shape the angle between the orthogonal line from one ring centre to the other ring centre is less than or equal to 30° (Chakrabarti and Bhattacharyya, **2007**).

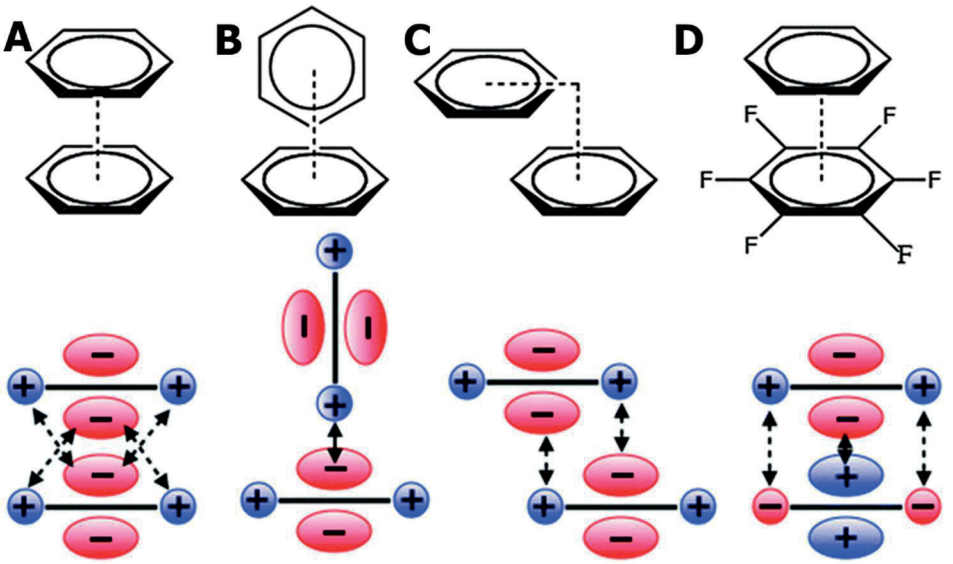


***Figure 19**: Types of π-stacking for benzene rings with the charge distribution around the π systems. (A) stacked. (B) T-shape. (C) paralleled. (D) reverse polarity. Adapted from Matthews et al (2014) (Matthews et al., **2014**).*

Cation-π - Cation-π interactions have been recognized as an interaction between the face of an electron-rich π-system and a cation, and thus, include a component of charge (see Figure 18B). In terms of strength, cation-π interactions are at the same level as strong H-bonds, but below the level of salt bridges (Gallivan and Dougherty, **2000**). The length of cation-π interactions are shorter than the stacking interaction of π-systems, e.g. ~3.5 Å for an interaction between a benzene and a primary amine (Dougherty, **1996**; Gallivan and Dougherty, **2000**). Cation-π interactions are furthermore preferentially exposed to solvent (Gallivan and Dougherty, **2000**). A critical role for cation-π interactions has been shown for many protein-ligand systems (Dougherty, **2013**), e.g. the binding of nicotine to acetylcholinesterase receptors.

Anion-π - Anion-π interactions were first demonstrated by X-ray crystallography using different types of aromatic rings and anions (Quiñonero et al., **2002**). In proteins, anion-π interactions have been observed as an edgewise interaction between Asp or Glu and aromatic amino acids upon analysis of 946 complexes from the PDB (Jackson et al., **2007**). The length of the interaction was found to be between 4.5 and 5 Å (Jackson et al., **2007**). More surprisingly, anions have also been found to position themselves preferably in axial position with respect to the plane of the ring (Giese et al.,

2015), which can be explained theoretically by studying the electronic configuration.

4.5. Molecular docking

Molecular docking is a computational method that aims to predict the binding mode of a ligand to a protein or to rank compound libraries based on the "fit" of compounds for a binding site (Bohm and Schneider, **2000**; Cheng et al., **2012**; Fatumo et al., **2013**). Docking simulations are thus today an essential structure-based drug design strategy (Kroemer, **2007**; Meng et al., **2011**; Ferreira et al., **2015**). Recent examples of structure-based virtual screening applications are many: for the human immunodeficiency virus-1, reverse transcriptase inhibitors screening (Santos et al., **2015**), for the discovery of novel inhibitors against *Mycobacterium* tuberculosis, 3-dehydroquinate dehydratase (Petersen et al., **2015**), and for the discovery of inhibitors of the PyrD protein, proteins allowing antibiotic resistance of *Pseudomonas aeruginosa* (Guo et al., **2016**).

A large number of docking softwares have been and are being developed. A representation of the most commonly used docking software is provided in Figure 20 (Sousa et al., **2013**). The most popular softwares are AutoDock (Morris et al., **1998**), Gold (Jones et al., **1995**, **1997**) Glide (Friesner et al., **2004**) and DOCK (Ewing et al., **2001**).

docking is a computational method that aims to predict the binding mode of a ligand to a protein or to rank compound libraries based of the "fit" of compounds for a binding site (Bohm and Schneider, **2000**; Cheng et al., **2012**; Fatumo et al., **2013**). Docking simulations are thus today an essential structure-based drug design strategy (Kroemer, **2007**; Meng et al., **2011**; Ferreira et al., **2015**). Recent examples of structure-based virtual screening applications are many, for the human immunodeficiency virus-1 reverse transcriptase inhibitors screening (Santos et al., **2015**), the discovery of novel inhibitors against *Mycobacterium* tuberculosis 3-dehydroquinate dehydratase (Petersen et al., **2015**), or the discovery of inhibitors of the PyrD protein, protein allowing antibiotic resistance of *Pseudomonas aeruginosa* (Guo et al., **2016**).

A large number of docking softwares have been and are being developed. A representation of the most used docking software is provided as Figure 20 (Sousa et al., **2013**). The most popular softwares are AutoDock (Morris et al.,

1998), Gold (Jones et al., **1995**, **1997**) Glide (Friesner et al., **2004**) and DOCK (Ewing et al., **2001**).

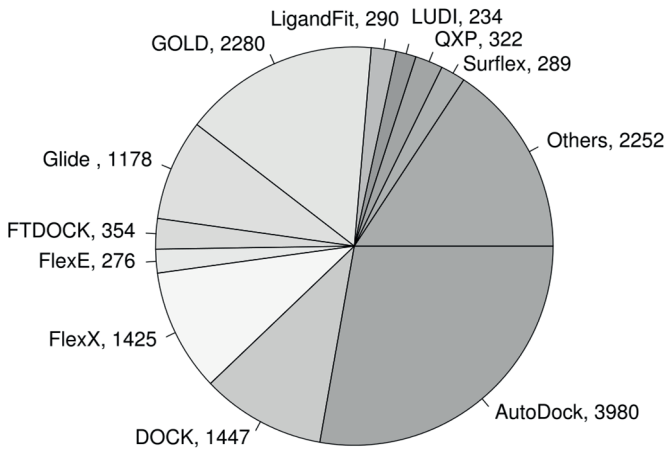


Figure 20: *Number of citations for the most common protein-ligand docking programs in the period 2001-2011. Programs published in 2011 are not included. Adapted from Sousa et al. (2013) (Sousa et al., **2013**).*

As a variety of docking algorithms are available, an understanding of the advantages and limitations of each method is of fundamental importance in the development of effective research strategies. Docking studies are divided into two steps: (i) ligand and receptor preparation, which includes the generation of 3D conformers for a ligand; (ii) pose prediction; and (iii) selection of the best poses using a scoring function.

4.5.1. Preparation

The preparation involves preparation of ligand and receptor, and consists of adding hydrogen atoms, eliminating water molecules, specifying the correct protonation and tautomerization states of the binding site and ligand, and calculating partial charges. This step is docking-software dependent, for the software Glide see e.g. original publication (Friesner et al., **2004**). Assigning for example incorrect ionization can have important consequences on the pose predicted (Jain and Nicholls, **2008**; Kirchmair et al., **2008**).

4.5.2. Pose prediction

The pose prediction step explores the conformational and orientational space accessible to the ligand within the protein binding site (see Figure 21) (Ferreira et al., **2015**). Two strategies are popular, i.e. a systematic search in, for example, DOCK (Ewing et al., **2001**) and Glide (Friesner et al., **2004**), or a random stochastic search in AutoDock (Morris et al., **1998**) and Gold (Jones et al., **1995**, **1997**).

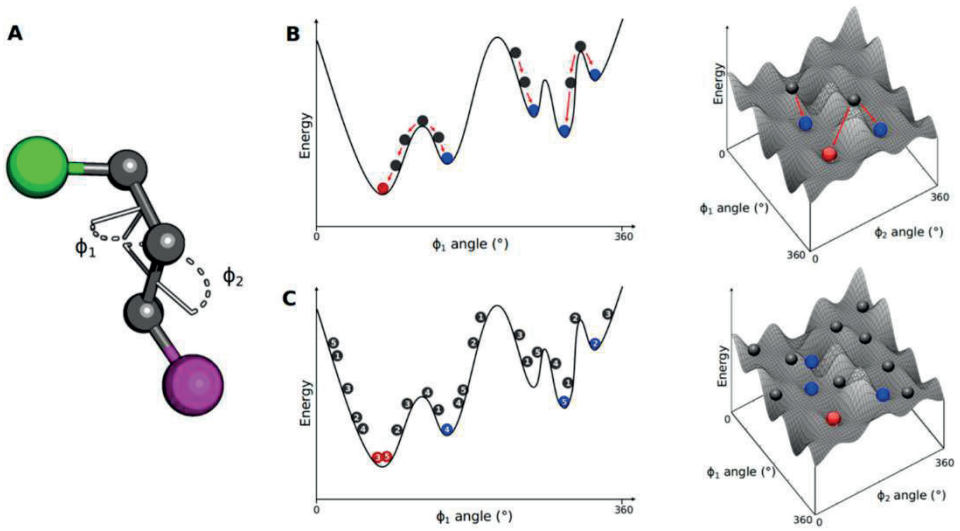


Figure 21*: Small-molecule conformational search methods. (A) A molecule containing two bulky groups (green and purple spheres) has its conformation defined by two internal dihedrals Φ1 and Φ2; (B) Considering Φ2 as a frozen dihedral, the energy variation due to r the energy variation due to rotation of Φ1 is plotted in a 1D energy landscape. The initial structure (grey spheres) is modified by changing Φ1, leading to a decrease in energy. The systematic search algorithm changes all structural parameters until a local (blue spheres) or global (red sphere) energy minimum is reached; (C) The stochastic search explores the conformational space by randomly generating distinct conformations, populating a broad range of the energy landscape. This procedure increases the probability of finding a global energy minimum. Reproduced with permission, Figure and legend, from Ferreira et al. (2015) (Ferreira et al., **2015**).*

Systematic search algorithms – Systematic search algorithms explore all of the degrees of freedom in a molecule. Three approaches are used: (i) conformational search, (ii) fragmentation and (iii) databases of pre-generated conformations. The conformational search is considered the "brute force" solution, e.g. in the software Glide (Friesner et al., **2004**). Ligand conformations are exhaustively sampled by systematically rotating (360°) all

rotatable bonds (Sousa et al., **2006**). Several constraints and restraints on the ligand are employed to reduce the combinatorial explosion that is proportional to the number of rotatable bonds. The fragmentation method consists of breaking down the chemical structure into several fragments, see e.g. in DOCK (Ewing et al., **2001**). One fragment is then selected as an anchor fragment and is docked to the binding site, and the process is continued until the entire ligand is constructed. The database methods used libraries of pre-generated conformations to estimate the ligand flexibility. An example of this method is FLOG, which generates for each compound a small set of conformations, e.g. 25 conformers by ligand (Miller et al., **1994**). Ligand conformations are then treated separately as a rigid docking protocol.

Stochastic algorithms – Random stochastic methods sample the conformational space using random changes to a ligand conformation, rotation or translation. Each change is accepted or rejected based on a probability function. Different algorithms may be used such as Monte Carlo, genetic or Tabu algorithms (Dias and de Azevedo Jr., **2008**). AutoDock (Morris et al., **1998**) and Gold (Jones et al., **1995**, **1997**) use a genetic algorithm, which appears more performant to converge to a global energy minimum in terms of number of search-and-evaluation cycles than the Monte Carlo algorithm (Krovat et al., **2005**). Tabu search algorithms appear to have a high accuracy in finding local energy minima (Baxter et al., **1998**), but they are less popular than other algorithms (Machado et al., **2001**; Dias and de Azevedo Jr., **2008**).

4.5.3. Scoring

Scoring functions are critical to docking simulations. They serve several distinct purposes: (i) to find the more likely poses among a large number of poses that have been sampled, (ii) to rank a set of compound based on the likelihood that they bind to a given protein (virtual screening) and (iii) to predict binding affinities (Kroemer, **2007**; Huang and Zou, **2010**; Meng et al., **2011**; Grinter and Zou, **2014**). Three type of scoring functions have been developed, which are commonly referred to as force-field-based, empirical-based and knowledge-based.

4.5.3.1. Force-field-based scoring function

Force-field-based scoring functions estimate the binding energy by a sum of physical interactions (bonded and non-bonded energy terms) that dominate protein-ligand binding. For example, from DOCK software that uses the AMBER force field (Ewing et al., **2001**)(Grinter and Zou, **2014**; Case et al., **2015**):

$$E_{AMBER} = E_{angle} + E_{bond} + E_{dihedral} + E_{non-bond}$$

E_{angle}: approximations of the bond angle energies

E_{bond}: approximations of strain energies

$E_{dihedral}$: energy term dihedral angles of linearly bonded sets of four atoms

$E_{non-bonded}$: aggregate of non-bonded energies interaction such as Lennard-Jones potential, van der Waals attraction and electrostatic potential terms.

Force-field-based scoring functions generally model well the physical principles that govern binding (Kitchen et al., **2004**; Meng et al., **2011**; Grinter and Zou, **2014**). Nonetheless, there is no entropic energy contribution, and a solvent model needs to be added, e.g. a Poisson-Boltzmann model (Rocchia et al., **2002**) or generalized Born models of solvation (Liu et al., **2009**). Cut-off distances are furthermore used to define non-bonded interactions, which results in decreasing the accuracy of long-range effects. This type of scoring function is also computationally demanding since there is a high number of combinations of atom pairs for which energies need to be computed, especially for large ligands. Force-field-based scoring functions are used, for example, in AutoDOCK, coupled with an empirical function (Morris et al., **1998**).

4.5.3.2. Knowledge-based scoring functions

Knowledge-based scoring functions are based on a sum of atom pairwise energy potentials. Knowledge-based scoring exploits crystallographic information, which reflect the native binding geometry; the potentials are extracted from known receptor-ligand complexes or small molecule crystals, and weighted on their observed frequencies. The pairwise potentials are based on the comparison with reference densities of the frequency densities for paired atom types, torsions and solvent exposition (Li and Liang, **2007**; Huang

et al., **2010**). Examples are the ITScore (Huang and Zou, **2006a**), DrugScore (Gohlke et al., **2000**) or DrugScore eXtended (DSX) (Neudert and Klebe, **2011**). Below we summarize the formalism of knowledge-based scoring, exemplified by the DSX scoring function (Neudert and Klebe, **2011**):

$$score\ (i) = -ln\frac{\rho\ (i)}{\rho_{ref}}$$

$score\ (i)$: score associated in a state i

$\rho\ (i)$: density function in state i

ρ_{ref}: density reference

A state depends on the pair of atom type involved and the distance between them. The score is a function of the density attributed to a given state, compared with the reference density learned previously from protein-ligand complexes. DSX's reference densities (Neudert and Klebe, **2011**) are based on a set of 37 067 X-ray structures with a resolution higher than 2.4 Å and containing at least one ligand from the PDB (Berman et al., **2000**) as well as 345 726 small molecule structures from the CSD (Allen, **2002**).

More precisely:

$$total\ score_{pair} = \sum_{a_p}\sum_{a_l} score(p(a_p), l(a_l), r(a_p a_l))$$

a_p, a_l: atoms in protein and ligand

$r()$, $p()$, $l()$: functions that are dependent on the distance, protein and ligand.

$$score_{pair}(p, l, r) = -ln(\frac{\rho(p, l, r)}{\rho_{ref}})$$

To reduce the combination of atom type paired, DSX grouped similar pairs of atoms based on a clustering of their atom pair densities.

Two additional terms referring to comparison of torsion densities based on four-atom combinations, and solvent exposure densities based on solvent-accessible surface are also included in the scoring function. The final score is given by:

$$score_{total} = w_p total\ score_{pair} + w_t score_{tors} + w_s score_{SR}$$

SR: SAS-ratio for a protein or ligand atom

w_p, w_t and w_s: weighting factors

In DSX, the $score_{total}$ is normalized by the volume available for interaction.

Knowledge-based scoring functions appear well-balanced between their accuracy and performance (Huang et al., **2010**). Limitations of this type of scoring function comes from the difficulty in properly defining the reference densities, including setting up artificial distance cut-offs (Gohlke et al., **2000**; Huang and Zou, **2006a**; Li and Liang, **2007**; Neudert and Klebe, **2011**). A dataset of high quality is furthermore required.

4.5.3.3. Empirical-based scoring functions

Empirical scoring functions are based on statistical models of binding energy prediction trained from protein-ligand complexes with known binding affinities (Kitchen et al., **2004**; Huang and Zou, **2010**). The prediction terms are associated with the physical events involved in the formation of protein-ligand complexes, e.g. H-bond interactions or ionic interactions. Each term is pondered by a factor computed from multiple linear regressions to fit the energy predicted with the real energy of interaction (Ferreira et al., **2015**). The first empirical function was developed by Böhm (1995) using a set of 45 protein-ligand complexes and was called LUDI (Böhm, **1994**). Different empirical functions exist in the literature such as X-CScore (Wang et al., **2002**), ITscore (Huang and Zou, **2006b**), SFCscore (Sotriffer et al., **2008**) and AIScore (Raub et al., **2008**). Other examples of empirical scoring functions are ChemScore (Eldridge et al., **1997**), GlideScore (Friesner et al., **2004**) and XP GlideScore (Friesner et al., **2006**).

For the ChemScore scoring function, the binding free energy is estimated as follows:

$$\Delta G_{bind} = \Delta G_{Hbond} \sum_{Hbond} f(\Delta R, \Delta\alpha) + \Delta G_{metal} \sum_{metal} f(\Delta R, \Delta\alpha) + \Delta G_{Lipo} \sum_{Lipo} f(\Delta R) + \Delta G_{rotor} \sum_{rotor} f(P_{nl}, P'_{nl}) + \Delta G_0$$

ΔG_{Hbond} , ΔG_{metal} , ΔG_{Lipo} , ΔG_{rotor}: linear correlation coefficients corresponding to the energies associated with H-bond, metal interaction, lipophilicity and ligand rotational entropy.

ΔG_0: regression constant

$f(\Delta R, \Delta\alpha)$: free energy term calculated with a function, f, which can depend on an angular ($\Delta\alpha$) and/or a distance term (ΔR).

$f(P_{nl}, P'_{nl})$: function of rotational entropy based on the number of rotatable bonds and the percentages of non-lipophilic heavy atoms on either side of the rotatable bond (P_{nl}, P'_{nl}).

The linear correlation coefficients in ChemScore (ΔG_{Hbond}, ΔG_{metal}, ΔG_{Lipo}, ΔG_{rotor}) are calculated from a regression model on a set of 82 experimentally determined protein-ligand complexes.

Empirical scoring functions have the advantage of predicting a free energy calibrated on real free energy, in contrast to e.g. knowledge-based functions, which predict a rank or force-field based functions that do not include experimental data in their calculation (Huang et al., **2010**). However, empirical scoring functions are limited by the experimental data used to build the underlying statistical models.

4.5.3.4. Comparison and consensus scoring function

Generally, no type of scoring function seems to outperform the others, as demonstrated by Wang et al. (2003) who compared 11 scoring functions and, more recently, Xu et al. (2015) who compared 16 scoring functions (Wang et al., **2003**; Xu et al., **2015**). The performance of the different scoring functions is dependent on the target used. A general test case for comparing scoring functions is the retrospective ability to recreate the poses of a co-crystallized ligand, predicting experimentally measured binding energies (Huang et al., **2010**).

Consensus scoring is the use of several scoring functions to extract results that are robustly found by many functions (Clark et al., **2002**; Wang et al., **2003**). Typical consensus methods are consensus score based on a rank-by-number,

average rank or linear combination of different methods. Consensus scoring allows a balancing of the advantages and limitations of several scoring functions. It is mostly used in virtual screening experiments, but can also be used as a rescoring tool when different poses have been suggested (Feher, **2006**).

4.5.4. Challenges for molecular docking

Scoring functions are one of the main challenges in docking simulations. For example, for three docking scoring functions using a dataset of 164 high-resolution protein-ligand complexes, considering only the top scoring solution, a pose with RSMD within 2 Å of native confirmation are found in only 42.6% of the cases for FlexX, 55.4% for GOLD and 59-63% for Glide (Perola et al., **2007**). Comparison of different docking scoring software to find native ligand conformation can, however, lead to seemingly conflicting results, even using similar protocols or datasets, and higher performances have been reported (Chen et al., **2006**). This is because docking simulations are affected by the complete docking protocols used, which includes the search space (size of the box) or the ligand preparation atom types (Huang et al., **2010**; Meng et al., **2011**; Ferreira et al., **2015**).

Flexibility of both the protein and ligand is very difficult to consider since it results in an exponentially large search space (Lexa and Carlson, **2012**). A common way to alleviate the problem is to complement docking studies with molecular dynamic simulations (Amaro et al., **2008**; Davare et al., **2015**; Kim et al., **2015**). Another strategy has to use a pre-enumerated conformational ensemble of protein conformations that are often generated by molecular dynamic simulations (Totrov and Abagyan, **2008**). There have been technical advances towards truly flexible docking: for example, the use of pseudo-flexible proteins (Huang and Zou, **2010**) or the creation of protein side chains around a positioned ligand. This latter strategy, implemented in the Glide induced-fit protocol (Friesner et al., **2004**), uses a cycle; starting from a preliminary binding pose, side chains are removed and reconstructed around the ligand, followed by docking and so on. More specific examples of methods that account for protein flexibility include soft docking, which allow a small overlap between the ligand and protein (Jiang and Kim, **1991**). The study of protein flexibility is helped by a better characterization of rotamer side chains, for example, in AutoDockFR (Ravindranath et al., **2015**).

Water molecules are an important challenge for the docking and are most often not taken into accountin docking protocols (Kroemer, **2007**; Lee and Seok, **2008**). Phenomena associated with water molecules, such as water networks, bridging protein-ligand interactions or contributing to the hydrophobic effect, are not considered. In contrast, 65% of the crystallographic protein-ligand complexes contain at least one water molecule (Klebe, **2006**). Glide XP (Friesner et al., **2006**) is one of the few docking programs that integrate water molecules, approximated by spheres.

5. Strategies for ligand optimization

A key premise to drug discovery is that structurally similar molecules exhibit similar biological activities, often referred to as the activity-property principle (Wermuth, **2006**). Medicinal chemists apply this concept to synthesize analogue series bearing potentially bioisosteric replacements, i.e. chemical groups that do not have an identified liability, but should keep the potency intact. Other key applications of this principle are modeling of structure-activity relationships and modeling of structure-property relationships.

5.1. Bioisosterism

Analogue series are generally constructed to carry bioisosteres, often utilized in lead optimization process with the aim of improving properties, such as pharmacokinetics, metabolism, solubility or reducing adverse effects, while keeping or improving potency. A comprehensive survey of bioisosteres and their characteristics has been reported recently (Brown, **2012**).

The definition of bioisosteres from the IUPAC is the following:

"*A bioisostere is a compound resulting from the exchange of an atom or group of atom with another, broadly similar, atom or group of atoms*" (IUPAC, **2016**)

Historically, the concept of the bioisosterism was born from that of isosterism; molecules that contain the same number and arrangement of electrons have similar physicochemical properties, introduced by Langmuir in 1919 (Langmuir, **1919**). Since then, the definition of the bioisostere has evolved and bioiosteres have been more recently defined (Burger, **1991**) as:

"*Compounds or groups that possess near-equal molecular shapes and volumes, approximately the same distribution of electrons, and which exhibit similar physicochemical properties...*"

5.1.1. Example: phosphorus replacement

Proteins related to phosphorylation and dephosphorylation events are central to biochemical processes. An estimated 30% of cellular proteins are phosphoproteins (Cohen, **2000**). Consequently, proteins that recognize phosphate-containing ligand moieties are an attractive target for therapeutic development. Phosphate as such is rarely used in drug molecules since it is predominately charged (pK_{a1} = 1.54 and pK_{a2} = 6.31) and is poorly permeable to the membrane (Smith et al., **2003**).

Classical replacements of phosphate groups can generally be divided into the following six classes (Rye and Baell, **2005**; Elliott et al., **2012**): (i) phosphorus-based bioisosteres such as phosphonate-based (phosporothionate and thiophosphonate-based) and boranophosphate-based bioisosteres, (ii) sulphur-based bioisosteres, (iii) carboxylate-based bioisosteres, (iv) heterocyclic-based bioisosteres, (v) squaric and squaramide-based phosphate bioisosteres and (vi) other bioisosteres containing a heteroatom (boronic acid and selenium-based bioisosteres). Some examples of these categories are presented in Figure 22.

Carboxylic acid bioisosteres are the most commonly used non-phosphorus isosteres of phosphate. They are widely represented in drug molecules with about 450 carboxyl-containing drugs in use worldwide (Ballatore et al., **2013**). For example, for S1P inhibitors the replacement of a phosphate to a carboxyl group allows modulating the activity from a IC_{50} of 0.85 nM to 19 nM for S1P receptor antagonists (Högenauer et al., **2010**).

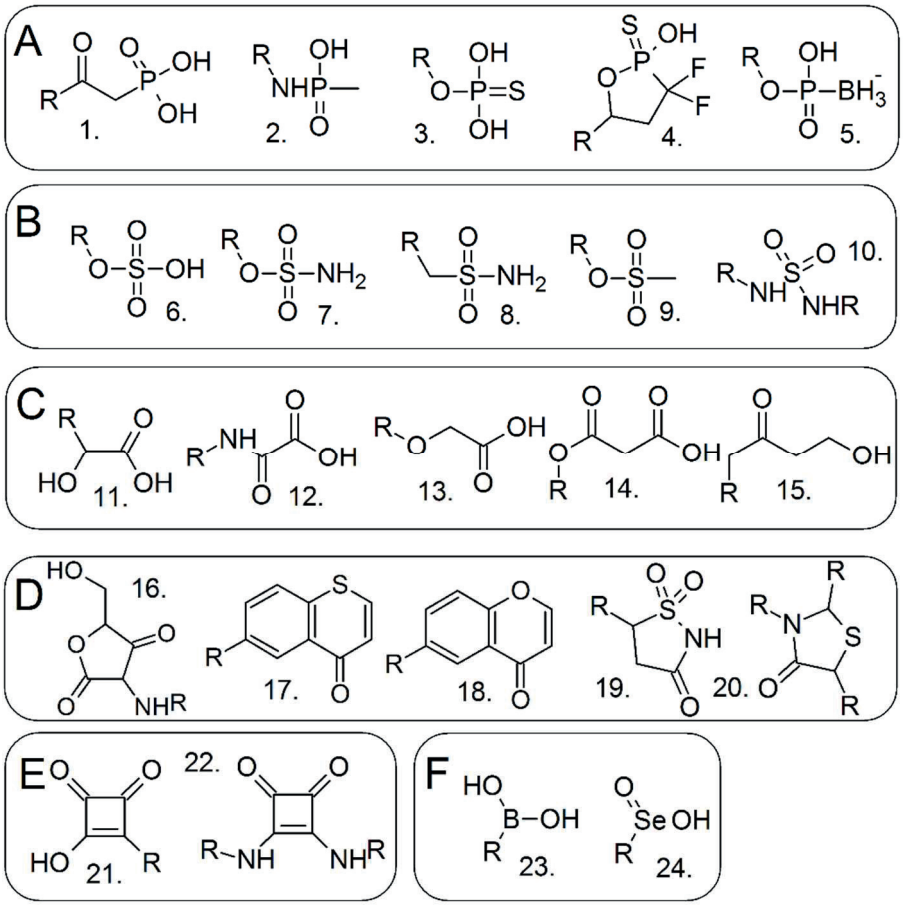
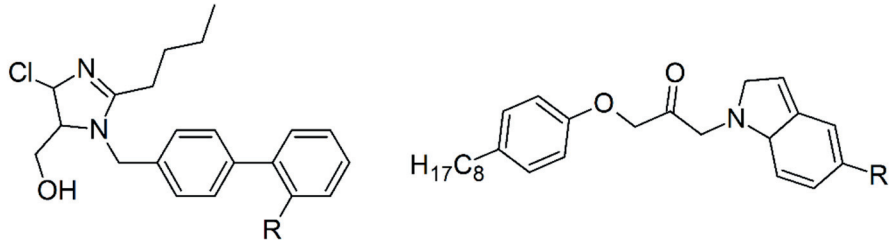


Figure 22: *Examples of classical phosphorus bioisosteres from Elliot et al. (2012) (Elliott et al., **2012**). (A) Phosphorus-based bioisosteres – 1. from PIsY inhibitor, 2. from NagA inhibitor, 3. from lipid analogue, 4. from cyclic phosphatidic acid analogue, 5. from nucleoside derivatives. (B) Sulphur-based bioisosteres – 6.,8.,9. from tyrosine phosphatase inhibitors, 7. $InsP_3$ derivate, 10. from lipid. (C) Carboxylate-based bioisosteres – 11., 12. from PT1B inhibitors, 13., 14. from phosphatase inhibitors, 15. from 6-N'-acetyltransferase inhibitors. (D) Heterocyclic-based bioisosteres – 20. from lysophosphatidic acid and thiazolidinone-derived, 17.,18. from mimic of pTyr, 19.,20. from PT1B inhibitors. (E) Squaric acid and squaramide-based bioisosteres – 21. from nucleotide derivate, 22. from tyrosine phosphatase inhibitors. (F) Other heteroatoms-based bioisosteres – 23. from nucleotide analogues, 24. from pTyr mimic.*

5.1.2. Challenges in identifying bioisosteric replacements

An important consideration regarding bioisosteric replacements is that while they prove effective for one type of target, an efficient replacement in one circumstance does not guarantee efficient replacement in another case (Meanwell, **2011**). Success of bioisosteric replacements is commonly dependent on the target family (Wassermann and Bajorath, **2011**).

For example, for the angiotensin II receptor antagonist (losartan), the tetrazole moiety in losartan offers a 10-fold increase in potency compared with the carboxylic acid analogue (Carini et al., **1991**). However, in a similar replacement used to develop cPLAA2α inhibitors, the novel tetrazole-containing analogue is 31-fold less active than the analogue containing carboxylic acid (Hess et al., **2007**); the structure is given in Figure 23.



1. R = COOH IC_{50} = 200 nM **3. R** = COOH IC_{50} = 35 nM

2. R = 1*H*-Tetrazol-5-yl IC_{50} = 19 nM 4. **R** = 1*H*-Tetrazol-5-yl IC_{50} = 1100 nM

Figure 23*: Limit of bioisosteric replacements on* IC_{50}*. Structures adapted from Meanwell (2011) with corresponding* IC_{50}*. 1. and 2. angiotensin II receptor antagonist analogues. 3. and 4. For cPLAA2α inhibitor analogues (Meanwell,* ***2011****).*

5.1.3. *In silico* bioisosterism identification

In silico methods have been shown to be useful tools, both methods and databases, to suggest and investigate the effects of bioisosteres (Ertl, **2007**). Three types of approaches have been used (Devereux and Popelier, **2010**): (i) rational approaches that define bioisosteres from similar compounds, (ii) literature searching and (iii) chemoinformatics approaches based on investigation of a chemical space or X-ray structures. While literature

searching is specific to a particular case study, the other two approaches have been used to develop databases of bioisostere replacements.

The rational approach has been used, for instance, to develop a large database of bioisosterism replacements. The BIOSTER database (version 15.1) (Ujváry, **1997**; Hayward, **2012**) contains 30 000 bioisosteric transformations, representing over 41 000 bioactive molecules. Another example is the freely available SwissBioisostere database (Wirth et al., **2013**), which contains 4.5 million replacements, automatically extracted as matched molecular pairs from the ChEMBL database (Gaulton et al., **2012**).

Chemoinformatics approaches to identify potential bioisosteres from PDB complexes have usually considered the atoms surrounding molecular fragments in binding sites. For example, the Sc-PDB-Frag database is based on a comparison of protein-ligand interaction fingerprints and contains 12 000 fragments and 15 million pairwise fragment comparisons (Desaphy and Rognan, **2014**). The KRIPO database relies on quantifying similarities of binding site subpockets using pharmacophore fingerprints. Fuzzy 3-point pharmacophore fingerprints were found to have the optimal balance between computational resources and identification of potential replacements (Wood et al., **2012**). An alternative strategy, which is not equivalent but rather complementary, was introduced by Kennewell et al. (2006) (Kennewell et al., **2006**). The three-dimensional structures of protein-ligand complexes are aligned and ligand substructures occupying the same binding region identified. This detects replacements variable in terms of molecular interactions that are occupying the same spot in the binding site.

5.2. Similarity searching

Searching of similar compounds to existing ones (typically active compounds) is a classical task of chemoinformatics. Different strategies are used, covering different molecular representations. The field of compound similarity calculation is very active especially for virtual screening applications (Cereto-Massagué et al., **2015**) or to describe the chemical space and identify the activity cliff (Zhang et al., **2015**)

5.2.1. Maximum common substructure

The maximum common substructure (MCS) is the most intuitive case of similarity searching for ligands. MCS-based methods are based on a pairwise graph matching to find the maximum substructure (Figure 24). The MCS can be used to derive a similarity score, calculated based on the length of the maximal substructure found relative to the complete size of the compounds. Different algorithms can be used to define the MCS, e.g. hyperstructure searching from hashing of small substructures (Teixeira and Falcao, **2013**), count fusion (Ahmed et al., **2014**) or mismatch tolerance (Wang et al., **2013**). MCS searching algorithms are much more computationally intensive than fingerprint similarity searches. For screening large databases, the two methods are usually combined.

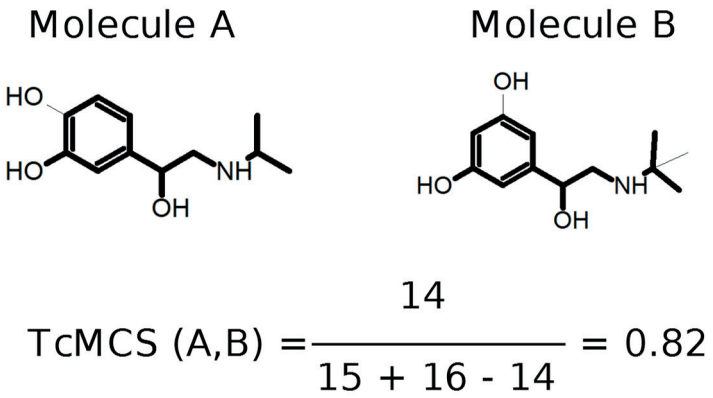


Figure 24: *Principle of MCS, example from Zhang et al. (2015) (Zhang et al., **2015**).*

5.2.2. Fingerprints

Fingerprints methods are based on the representation of molecules as bitstrings, i.e. numerical vectors, where each bit corresponds to a property found (1) or not found (0) in the compound (Figure 25).

At least six types of molecular fingerprints are commonly used (Muegge and Mukherjee, **2015**):

- Topological fingerprints, which capture molecular features, such as the number of bonds or type of atoms used in, for example, Daylight (Daylight Chemical System Information) or atom pairs (Sheridan et al., **1996**).
- Structural keys, which capture structural properties, such as number of ligand configurations used in, for example, BCI (Barnard et al., **2000**) or PubChem (Wang et al., **2009**; PubChem, **2015**).
- Circular fingerprints, which record the radial environments of each atom. The radial environment is first recorded by considering an atom directly connected to the central atom and next widened to increase the number of atoms connected used in, for example, Molprint2D (Bender et al., **2004**) or ECFP (Rogers and Hahn, **2010**).
- Pharmacophore fingerprints, based on the search of key structural properties used in, for example, CAT descriptors (Schneider et al., **1999**), 3pt (McGregor and Muskal, **1999a**, **1999b**) and 4pt 3D fingerprints (Mason Jonatan S., **2000**).
- Hybrid fingerprints, combining the previous categories and used in, for example unit 2D fingerprint (Certara, **2015**).
- Interaction fingerprint which captures the interaction information, such as the PIPLIF method (Radifar et al., **2013**) based on the definition of the pair interactions from Marcou et al. (2007) (Marcou and Rognan, **2007**).

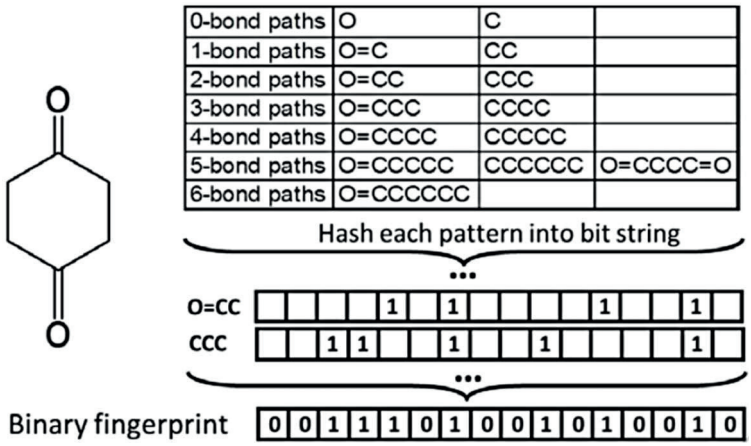


Figure 25: *Generation of topological fingerprint using Daylight (http://www.daylight.com/) fingerprint. Reproduced with permission from Muegge and Mukherjee (2015) (Muegge and Mukherjee, **2015**).*

Fingerprints are very efficiently compared using coefficients such as Tanimoto, Dice or Tversky. Differences between these similarity scoring methods have been presented by Bajusz et al. (2015) (Bajusz et al., **2015**).

5.2.3. Ligand descriptors

Molecular descriptors are sets of characteristics associated with compounds such as molecular weight, geometry, volume, surface areas, ring content, rotatable bonds, interatomic distances, bond distances, atom types, planar and non-planar systems, molecular walk counts, electronegativities, polarizabilities, symmetry, atom distribution, topological charge indices, functional group composition, aromaticity indices, solvation properties and many others (Sliwoski et al., **2014**). Molecular descriptors are widely used for QSAR modeling, sometimes for the pairwise comparison of compounds, especially for representing the chemical space, and almost never for similarity searches of compounds in databases since the fingerprint-based searches are computationally much more efficient (Muegge and Mukherjee, **2015**).

Molecular descriptors have been compiled in the Handbook of Molecular Descriptors from Todeschini and Consonni, with more than 5000 in the latest edition (Todeschini and Consonni, **2010**). Descriptors can be classified according to the "dimensionality" of the chemical representation: 1D for scalar physicochemical, such as molecular weight, 2D for molecular constitution and configuration-derived, such as number of aromatic rings or number of chiral centres and 3D for descriptors conformation-derived (Ekins et al., **2007**). Ligands represented using a set of descriptors can be compared as in, for example, the study of Venkatraman et al. (2009), where ligands are compared using geometric properties based on 3D Zernike descriptors (Venkatraman et al., **2009**).

5.3. Modeling Structure-Activity Relationship (SAR)

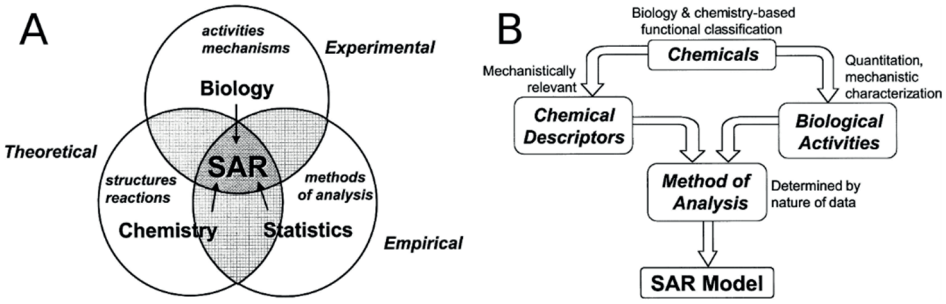


***Figure 26**: Principles of SAR models. (A) SAR models in intersection biology, chemistry and statistics fields. (B) General protocol to develop a SAR model. Adapted with permission from McKinney et al. (2000) (McKinney et al., **2000**).*

Quantitative structure-activity relationship (QSAR) and Quantitative structure-property relationship (QSPR) modeling refers to an ensemble of statistical methods that are used to quantitatively predict biological activities (potency or affinity for a target, biodegradability, diffusion or toxicity) or physicochemical properties. Predictive modeling can also be used to develop non-quantitative classification models, which are not discussed in this Review of the literature.

The concept of structure-activity relationships (SAR) was for the first time demonstrated between the chemical composition of ammonium salts and their physiological action in 1868 (Brown and Fraser, **1868**). The first to establish a mathematical relationship between structural attributes and specific activities of a set of selected chemicals was Hansch and Fujita in the 1960s (Hansch and Fujita, **1964**). To date, QSAR modeling has been used to predict the biological activities of untested and sometimes still unavailable compounds, to optimize an existing lead and to clarify which chemical properties are the most likely determinants for their biological activities (Wang et al., **2015**). The international organization for economic cooperation has also edited regulations for the use of QSAR molecules to test for toxicity of compounds, leading to the QSAR-Toolbox project (http://www.qsartoolbox.org/). The general protocol is presented in Figure 26 (Puzyn et al., **2010**; Verma et al., **2010**).

5.3.1. Development of a QSAR model

Development of a QSAR model may be divided into three main steps. These are (i) selection of a series of active and inactive molecules with known activities, distributed into a training set and a test set; (ii) selection of descriptors; and (iii) construction of the models QSAR and validation (Wang et al., **2015**).

Careful attention is required to develop, validate and exploit a QSAR model. Based on the publication of Tropsha (2010), a generic QSAR workflow is presented in Figure 27 (Tropsha, **2010**; Golbraikh et al., **2012**).

Original dataset and curation – In academic settings, datasets are often extracted from public databases such as ChEMBL (Gaulton et al., **2012**) or PubChem (Wang et al., **2009**) or other commercial databases (for review, see (Oprea and Tropsha, **2006**)). Curation of the dataset is a very important since the error rate in public or commercial databases has been estimated to be around 3.4% (Young et al., **2008**). QSAR models are also influenced by other factors, such as tautomeric forms or structure representation, leading to incorrect descriptor calculation. Fourches et al. (2010) suggested overcoming these limitations with a standardized curation protocol: (i) removal of mixtures; (ii) cleaning structure and removing salt; (iii) normalization of specific chemotypes; (iv) treatment of tautomeric forms; (v) removal of duplicates; and (vi) manual inspection (Fourches et al., **2010**).

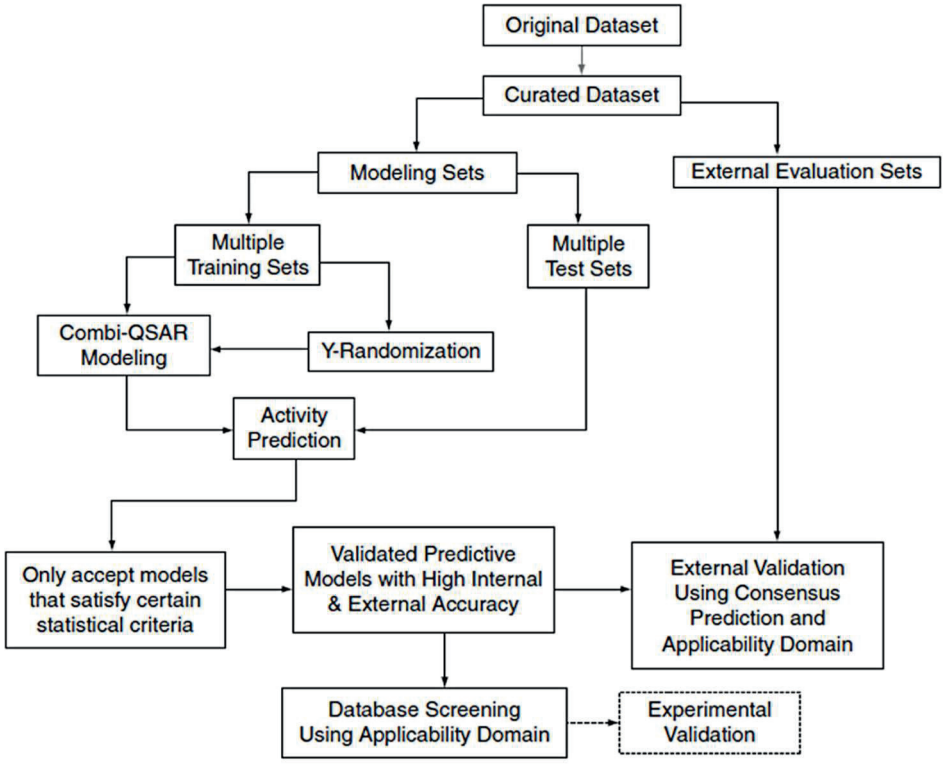


Figure 27: *Predictive QSAR modeling workflow. Reproduced with permission from Golbraikh et al (2012) (Golbraikh et al., **2012**).*

Descriptors selection – Numerous descriptors exist in the literature, as shown in the handbook of molecular descriptors ((Todeschini and Consonni, **2010**) see also (Puzyn et al., **2010**)). Molecular descriptors can be grouped by different dimensionalities, i.e. 0D for constitutional descriptors; 1D for counts of molecular groups or physicochemical properties; 2D for invariants of molecular graphs, e.g. connectivity indices and information indices; and 3D for geometric spacial properties.

Balancing between external test set and training set – This step consists of defining a training set and an external test set (typically 10–20% of data), which will be used to evaluate the QSAR model. The training and test sets need to have the same chemical diversity. Outlier compounds should be deleted to refine the chemical space where the model is trained. Several statistical methods are used to find the activity outliers (Bajorath et al., **2009**; Sisay et al., **2009**). Structural outliers are compounds that are largely dissimilar to all other

compounds in the descriptor space, and they should also be deleted (Puzyn et al., **2010**).

Model construction – QSAR modeling techniques employ various methods of multidimensional data analysis as well as supervised machine learning. The commonly used machine learning methods are multiple linear regression, partial least squares, artificial neural networks or support vector machine (Gertrudes et al., **2012**). Different considerations are employed, depending on the dataset, to select the optimal machine learning method (Sorich et al., **2003**; Louis et al., **2010**; Pourbasheer et al., **2010**; Qin et al., **2011**; Varnek and Baskin, **2012**).

Validation – Three types of validation are possible: (i) internal validation or cross-validation in *n*-fold, which uses a sampling of the training set in *n*-folds to measure the robustness of the predictions and the quality of predicted error; (ii) external validation based on predictions for molecules belonging to an external test set; and (iii) data randomization or Y-scrambling, where the response variables are randomized (Tropsha et al., **2003**; Gramatica, **2007**; Golbraikh et al., **2012**).

Different statistical criteria are commonly used to assess the predictivity of QSAR models (Wang et al., **2015**).

The coefficient Q^2 calculates the predictive power of a model using cross-validation.

$$Q^2 = 1 - \frac{\sum(y_{pred} - y_{obs})^2}{\sum(y_{obs} - y_{mean})^2}$$

$y_{pred}, y_{obs}, y_{mean}$: variable predicted, observed and mean of the prediction in mean from a *n*-fold validation.

The coefficient R^2 is the linear correlation coefficient between observed and predicted variables in the test set.

5.3.2. Dimensionality of QSAR models

Depending on the level of abstraction used to model chemical compounds, different "dimensionalities" of QSAR models have been proposed. Table 4 summarizes the different types of QSAR models.

***Table 4**: Specificity of different generation of QSAR models, from Damale et al. (2014) (Damale et al., **2014**).*

QSAR models	Specificity
1D-QSAR	Use descriptor experimental such as pK_a or log$P$ and topological descriptors without considering the 3D ligand information.
2D-QSAR	Use structural patterns like connectivity indices or 2D-pharmacophores. 3D representations are still not considered.
3D-QSAR	Incorporate 3D descriptors and consider ligand flexibility, for example by using alignment steps.
4D-QSAR and more	Often receptor-dependent model including flexibility, induced fit modeling and solvation modeling.

3D-QSAR is popular as a support method for medicinal chemistry since it allows visualization of the type of replacements that are going to decrease or increase biological activity around a template compound (Verma et al., **2010**). The comparative molecular field analysis (CoMFA) method (Cramer et al., **1988**) is a popular QSAR method in drug design, as shown by the number of publication with the term 'CoMFA', from 50 in 1995 to 160 in 2009 (Zhang et al., **2011**). The CoMFA method is based on an 3D alignment of ligands in an energy grid. For each grid point, a resulting energy is calculated from electrostatic (Coulombic) and steric (van der Waals) energy. These grid point energies are used as descriptors and correlated with the biological activity using a partial least squares regression. The resulting PLS models are combined to define a plot contour for each compound. The principle of

CoMFA method is presented in Figure 28 from Zhang et al. (2011) (Zhang et al., **2011**).

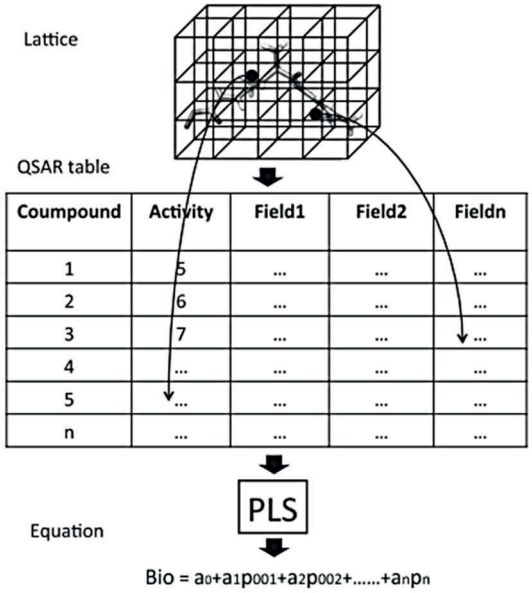


***Figure 28**: Standard CoMFA process, reproduced with permission from Zhang et al (2010) (Zhang et al., **2011**).*

5.3.3. Challenges for QSAR models

A challenge for QSAR is to propose clear workflows to optimize properly a QSAR model. Dearden et al. (2009) identified 21 types of errors perpetrated in QSAR/QSPR models in the literature, e.g. poorly curated datasets with replicate compounds in training and test sets or erroneous descriptors (Dearden et al., **2009**). The computation of the descriptors themselves can have a major impact on the reproducibility of QSAR models (Gramatica, **2007**). Prediction of LogP showed significantly different values using different software (Benfenati et al., **2003**).

Overfitting and overtraining models that have a high theoretical predictivity on training and test sets, but no real external predictivity is another challenge with QSAR modelling (Tropsha, **2010**). As shown by Topliss (1977), the number of descriptors relative to activities may lead to significant unwanted chance correlations, generating a QSAR model that is only apparently predictive (Topliss, **1977**). A rule often considered helpful is to use “at least six or seven

compounds for each descriptor". This rule is nonetheless poorly followed in the literature, as shown from an analysis of 28 QSAR models predicting anticonvulsant activity (Garro Martinez et al., **2015**). Overfitted models are more affected by random variations and irrelevant predictors that reduce their performance and portability (Hawkins, **2004**).

The applicability domain, i.e. the theoretical region of the chemical space where the model was trained and is applicable, has recently received a great deal of attention (Sahigara et al., **2012**), however, it is still often poorly described in the publications (Wang et al., **2015**). Different strategies have been developed to define the applicability domain using molecular descriptors (Sahigara et al., **2012**): (i) range-based or geometric methods that visually represent the applicability domain; (ii) distance-based methods that calculate the relative distance of the compound to the applicability domain based on a transformation of the descriptor matrix to a distance matrix; (iii) density-based methods that use a probability function; and (iv) other methods such as decision trees.

The final issue is availability. Over the last 60 years, many QSAR models have been developed, but only a few are used due to poor visibility in the community. To avoid this problem, databases have been set up such as C-QSAR, which to date contains 18 000 models (Kurup, **2003**). More recently, a collaborative platform, called QSAR-DB, to centralize and classify QSAR models has been developed (Ruusmann et al., **2014**, **2015**).

5.4. Designing analogues and thermodynamic profiles

Analysis of thermodynamic profiles, i.e. entropic and enthalpic contributions for each analogue in a binding site, allows the affinity and specificity for a target to be optimized. During the lead optimization phase the binding constant may be improved by 5–6 orders of magnitude (Klebe, **2013**). Analysis of thermodynamic profiles of different analogues thus provides important information on the biophysical phenomena resulting in protein-ligand binding.

5.4.1. Cooperation between entropy and enthalpy

As discussed in Section 3.3, the binding free energy is directly connected to the changes in enthalpy and entropy. The binding free energy can thus be driven by the entropic term, by the enthalpic term or by both. Understanding the origin of the entropic and enthalpic contributions towards the binding free energy is very important in drug discovery (Geschwindner et al., **2015**) (Homans, **2007**). Entropy-enthalpy cooperation has been a subject of discussion, research and criticism for over half a century (Petersen, **1964**; Lumry and Rajender, **1970**; Sharp, **2001**; Starikov and Nordén, **2007**; Geschwindner et al., **2015**; Pan et al., **2015**).

The phenomenon of compensation between entropic and enthalpic contributions is supported in, for example, 32 protein-ligand complexes of the 102 studied, or in another report 14 protein-ligand complexes of the 171 studied (Olsson et al., **2011**; Reynolds and Holloway, **2011**). This phenomenon is illustrated in Figure 29 for a thermolysin binding site, where different analogues have the same binding free energy, but different thermodynamic contributions (Klebe, **2015**). Nonetheless, the proportion in entropic and enthalpic energy of each modification is clearly system-dependent, as demonstrated by the above-mentioned Olsson study (Olsson et al., **2011**). For some systems, such as pathogen-derived peptides with class II major histocompatibility complex, this phenomenon is considered a epiphenomenon (Ferrante and Gorski, **2012**).

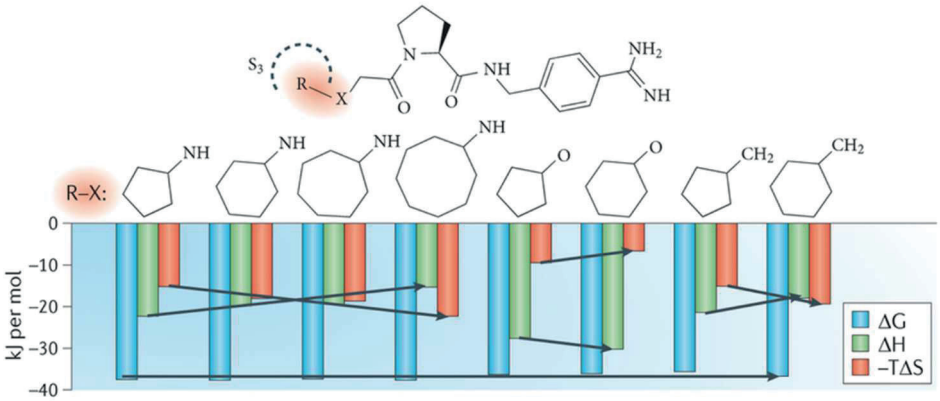


***Figure 29**: Different ligands in a series of modified peptidomimetics showing equipotent binding to trypsin; nevertheless, their affinities factorize differently into enthalpic and entropic components, adapted from Klebe (2015) (Klebe, **2015**).*

The diversity of cooperation between the entropic and enthalpic contributions is thus difficult to analyse; a pragmatic and acceptable theory to increase the affinity of a drug is yet to be developed (Pan et al., **2015**). Generally, water molecules and solvent rearrangement take a prominent place in explaining this phenomenon, as demonstrated in the model of Grunwald and Steel (Grunwald and Steel, **1995**). Hydrophobic effects are as the most favourable contributor to binding free energy, estimated at 80% (Whitesides and Krishnamurthy, **2005**; Garbett and Chaires, **2012**). Amplitude of the hydrophobic contribution is correlated with the compound lipophilicity size (Murray et al., **2012**). This is explained by water molecules adjacent to an apolar group forming a network of H-bonds, and this order is entropically favourable. Desolvation of these hydrophobic groups perturbs the water network, yielding entropy instability with the transfer of water molecules from the network to the bulk solvent (Kyte, **2003**; Starikov, **2013**). Desolvation of polar groups breaks the H-bond between ligand or protein polar groups and water molecules. These water molecules are transferred to the bulk solvent, and this energy influences both the entropic and enthalpic contributions (Olsson et al., **2008**; Klebe, **2013**). Both hydrophobic effect and desolvation of polar groups perturb the water molecule network in the first hydration shell of the ligand, target and protein-ligand complex. Perturbation of the first hydration layer contributes to the entropy-enthalpy compensation phenomenon (Biela et al., **2013**; Betz et al., **2016**).

Conformational rearrangements play also a role in defining the entropic contribution to ligand binding. A local perturbation of the protein and ligand structures has been suggested to flip a switch to a high-entropy conformational state favourable to the interaction. This phenomenon has been named entropy–enthalpy transduction (Fenley et al., **2012**) and can explain the entropic contribution of different analogues, indicating that different high-entropy conformational states exist for different biomolecules with the same binding free energy.

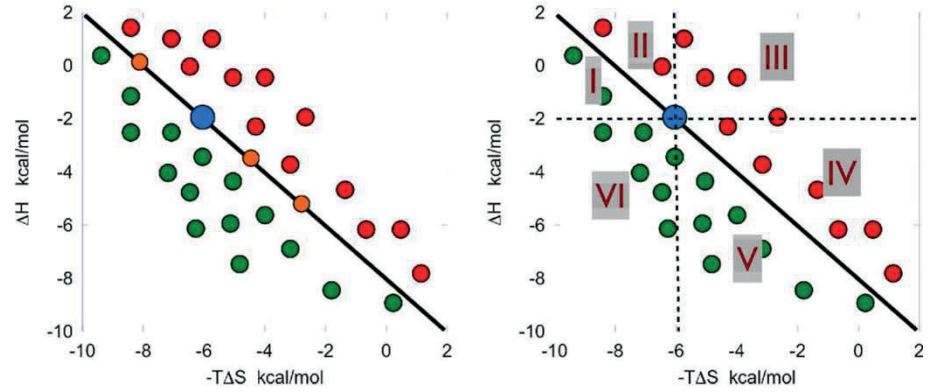
5.4.2. Strategies for drug optimization

Entropy and enthalpy need to be optimized conjointly to overcome the compensation. Entropy seems to be easier to influence by increasing the compound hydrophobicity. Increasing the enthalpic contribution requires well-positioned interactions such an H-bond (Klebe, **2013**).

Five simple rules of optimization have been suggested (Klebe, **2013**): (i) lipophilicity should be limited to maintain good water solubility; (ii) protein polar atoms not solvent exposed must have a ligand partner; (iii) in the case of pockets not optimally solvated, ligands can form more H-bonds with the protein than with water molecules, and thus, the binding affinity of such ligands can be very high; (iv) rigid ligands can bind more strongly than flexible ligands because of the loss of internal degrees of freedom; (v) for most proteins that contain transition metals, maintaining an interaction between the metal and the protein is important.

Another complementary approach has been suggested by Freire et al. (2009), who defined six regions of drug optimization, presented in Figure 30 (Freire, **2009**). *Regions I and II* – entropy is more favourable in the binding free energy that enthalpy. *Region III* – modifications result in both enthalpic and entropic losses. These modifications are not relevant because they are not binding free energy-driven. *Regions IV and V* – an enthalpic gain is compensated by an entropic loss. *Region VI* – modification is enthalpy-dependent, but is not completely compensated by entropy. For the authors, the best optimization region is region VI, where the gain of affinity is due to a gain of enthalpy, enabling an increase also in the selectivity between the protein and target (Freire, **2008**, **2009**). However, positioning an analogue in this region is difficult because to increase the interaction between protein and ligand requires

optimization of H-bonds, which are often in competition for water molecules. A strong H-bond interaction also does not guarantee a binding free energy increase and can be fully compensated, as demonstrated for the HIV-1 protease when comparing different analogues (Lafont et al., **2007**).



.

Figure 30: *After a round of optimization, the (−TΔS, ΔH) points for all compounds are plotted. Compounds with better affinity (shown in green) fall below the optimization line, while compounds with lower affinity (shown in red) appear above the optimization line. Compounds with the same affinity as the parent compound (shown in orange) are situated on the optimization line. By tracing a vertical and horizontal line through the coordinates of the parent compound, six different regions can be defined. These regions define distinct strategies for optimization. Reproduced with permission from Freire et al (2009) (Freire,* ***2009****)..*

5.4.3. A drawback limit of drug optimization: molecular obesity

Recent years have revealed that modern medicinal chemistry programs produce many "obese" compounds (Hann, **2011**) – highly potent but very lipophilic compounds with molecular masses beyond 500–600 Da. These compounds gain in binding free energy using entropy optimization. However, increasing the size of the molecules reduces their bioavailability and results in early fails. A way to avoid creating such compounds is to control early on the ligand efficiency, i.e. the potency with respect to size or molecular weight (Reynolds et al., **2008**; Hann and Keserü, **2012**).

5.5. Computational methods to estimate water molecules contribution in binding sites

Different computational methodologies have been developed to study the contribution to the binding energy of water molecules at binding sites, which can be used to define regions that can be favourably modified during compound optimization (Pearlstein et al., **2010**, **2013**; Brodney et al., **2012**). Based on long molecular dynamics, WaterMap (Abel et al., **2008**) (Schrödinger commercial suite) estimates from a 3D structure the hydration sites, based on investigation of the movement of water molecules during a long molecular dynamic simulation. Alternative freely available tools are WATsite (Hu and Lill, **2014**) and WATCLUST (López et al., **2015**), which are also based on analysis of the trajectory of water molecules during a molecular dynamic.

Another strategy is to investigate the position of water molecules in X-ray structures. WaterScore estimates water molecule displacement using a combination of different factors such as B-factor, solvent-contact surface area, total hydrogen bond energy and number of protein atomic contacts (García-Sosa et al., **2003**). Aqualta reproduces water molecules that bridge polar interactions between ligands and proteins using geometric criteria obtained from extensive searches of the CSD (Rossato et al., **2011**). SZMAP (solvent-Zap-mapping) (Bayden et al., **2015**) estimates the water molecules' binding free energy influence using a sample orientation of the water molecules from water molecules elucidated by crystallography.

Aims of this thesis

The main goals of this thesis were to develop computational methods and tools useful for profiling a ligand for a target (I, II), to mine for ligand structural replacements (III), to mine and analyse the spatial distribution of interacting atoms forming protein-ligand salt bridges (IV) and to visualize and select among a pool of docking poses (V). Preliminary results regarding positioning water molecules in binding sites are also presented (unpublished results).

Particular emphasis was placed on providing the methods to the community free of charge using a web server (II) or providing the source codes through the code development platform GitHub (III, IV).

Specific developments and objectives:

- To build predictive QSAR models that predict protein pocket druggability with high accuracy; and to characterize druggable binding sites using molecular descriptors (I)
- To write a fully automated workflow that extract ligand local structural replacement (akin to bioisosteres) based on a superimposition of homologous proteins. This method is applied in the study of phosphate isosteres (III).
- To analyse the environment of six specific charged ligand groups, highlighting the importance of weakly charged interactions for their recognition, including by water molecules in the absence of strong salt bridges (IV).
- To developed scripts to visualize, compare and select binding poses (obtained by induced-fit docking) by comparison with a bound reference (V).
- Explore methods to position water molecules in binding sites and to characterize them using geometrical considerations (unpublished results).

Materials and methods

1. Databases of structural data

1.1. Protein data bank

The PDB, see Section 1, Review of the literature, is a collaborative structural database containing more than 107 154 protein structures (release 02-2016). This databank is the source of all protein structures used in this doctoral thesis.

1.2. Druggable and non-druggable datasets

Druggable and non-druggable proteins, with their corresponding binding sites have been extracted from the database Druggable Cavities Dataset (DCD) (Schmidtke and Barril, **2010**) (http://fpocket.sourceforge.net/dcd/), containing 1 068 proteins, 159 apo proteins and 909 holo proteins.

1.3. Method for database redundancy

The PDB databank is highly redundant. The same protein can be included several times, crystallized with different ligands, at different resolutions or using different crystallographic techniques or packing. In Publications III and IV, the redundancy is treated using a sequence alignment algorithm and a sequence identity calculation available in the EMBOSS suite (Rice et al., **2000**). The software EMBOSS Needle is used, which implements the Needleman-Wunsch global sequence alignment (Needleman and Wunsch, **1970**). Parameters were conserved by default, i.e. gap opening penalty = 10.0, gap extension penalty = 0.5 and matrix substitution EBLOSUM62. The definition of the sequence identity used for filtering is:

$$Sequence\ identity = \frac{matches}{length\ of\ aligned\ region\ (with\ gap)}$$

It should be noted that this definition is unfavourable to proteins composed of multiple domains. Two identical sequences will have 100% sequence identity, while two random sequences have 5-15% identity; sequences in the 15-25% identity region are in the “twilight zone” in deciding whether or not they are evolutionarily related (Pearson, **2013**).

To avoid composition bias, a subset of the PDB, called PDB50, was used containing only 22 091 different proteins with a cross sequence identity inferior to 50%.

1.4. General protocol extraction

From the PDB, an extraction protocol divided into the following three steps was devised: (i) remove DNA, RNA, NMR structures, (ii) control the R-value and resolution and remove structures with weak quality, (iii) keep only one structure with multiple related structures present using a global sequence alignment.

2. Method for structural analysis

2.1. Protein pocket estimation

Three pocket estimation methods were used: (i) ligand proximity, taking protein atoms at a distance threshold of the ligand atoms (ii) Fpocket (Le Guilloux et al., **2009**) and (iii) DoGSite (Volkamer et al., **2010**). These three methods cover three different pocket estimation types, e.g. (i) using a ligand position, (ii) using a geometric algorithm and (iii) using an energetic algorithm. They are freely available using web servers or source code (Schmidtke et al., **2010a**; Volkamer et al., **2012b**).

2.2. Pocket and ligand descriptors

2.2.1. Ligand descriptors

Ligand descriptors were computed using the Python package (freely available PyDPI) (Cao et al., **2013**). Only descriptors for small molecules were used, 3215 descriptors divided into 12 groups, such as topology, physicochemical properties or composition.

2.2.2. Pocket descriptors

Fifty-two descriptors divided into six types were used. Some needed to be implemented by Python version 2.7 scripts in order to be used. An overview of these pocket descriptors is presented in Table 5.

Table 5: *Pocket descriptors implemented, adapted from Publication I*

Type of descriptors	Descriptors based on	References
Hydrophobicity	atoms and amino acid composition and solvent accessibility	NACCESS (Hubbard, SJ and Thornton, **1992**) (Burgoyne and Jackson, **2006**) (Milletti and Vulpetti, **2010**) (Kyte and Doolittle, **1982**)
Aromaticity	aromatic amino acid frequency	(Milletti and Vulpetti, **2010**)
Polarity	polar amino acid frequency and atoms composition	(Eyrisch and Helms, **2007**)
Physico-chemical properties	atom and amino acid frequency	(Milletti and Vulpetti, **2010**)
Volume	volume of the convex hull computed using atom pocket	(Petitjean, **2014**) (Petitjean, **1992**)
Shape	shape of the convex hull computed using atom pocket	(Petitjean, **2014**) (Petitjean, **1992**)

2.3. Ligand similarity

Ligand similarity is implemented both using fingerprints and maximum common structure (MCS). MCS allows subsequent computation of the RMSD of the atoms included in the MCS.

The MCS searching is realized using the algorithm of non-contiguous atom matching structural similarity (NAMS) (Teixeira and Falcao, **2013**). This algorithm is based on an alignment of a compound and graph comparison, including bounding and atom profiles. The measurement of the similarity is next computed using a Jaccard similarity coefficient (Jaccard, **1901**):

$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{S(A,B)}{S(A,A) + S(B,B) - S(A,B)}$$

where A and B are the compounds compared, $S(A,A)$ is the score matching of the superimposition of A on A, $S(B,B)$ the score of matching of B on B and $S(A,B)$ the score of matching of A on B. The similarity coefficient ranges between 0 and 1, 0 being compounds without similarity and 1 identical compounds. The score matching, based on a graph comparison, included two components, an atom-matching component and a bond-matching component.

2.4. Protein-ligand interaction fingerprint

Protein-ligand fingerprints are used in Publication V to compare docking poses. Interactions are defined using a set of rules based on geometric criteria, e.g. the distance and angle between a ligand atom and a protein atom. The list of contacts was used to define a profile of the interaction (Marcou and Rognan, **2007**).

Comparison of two profiles of interaction, e.g. group of interaction defining a protein-ligand interaction, is conducted by comparing the fingerprints using a Jaccard similarity coefficient (Jaccard, **1901**).

Fingerprint of interaction is computed in Python 2.7 using PyPLIF package (Radifar et al., **2013**).

2.5. Ligand mining

The Simplified Molecular-Input Line-Entry System (SMILES) (Weininger, **1988**) format is used to extract from the ligand simple chemical substructures, e.g. a ligand-containing ring. Ligands, in 3D, are transformed using open Babel software in short ASCII strings (O'Boyle et al., **2011**). The resulting strings were next inspected using regular expression (regex) implemented in Python.

2.6. 3D data mining

For phosphate groups (Publication III), the labels in the PDB are sufficient to retrieve the atoms of interest.

For the chemical substructures of interest (Publication IV), data mining of the PDB is realized using house-script based on a redefinition of connectivity matrix for each atom. Distance criteria, to differentiate different chemical bonds, single, double and triple, together with a distance threshold are fixed empirically based on a statistical analysis using the PDB. For tertiary amines, deviations from planarity of the plane formed by the connected carbons are also tested to avoid nitrogen in resonant systems.

3. 3D superimposition

3.1. Protein superimposition (TM-align)

TM-align is a software that identifies the optimal alignment between the tertiary structure of protein pairs (Zhang, **2005**). TM-align is based on a step of sequences alignment to pair the residues of both proteins. The output of TM align is a rotation/translation matrix.

Superimpositions of paired proteins are subsequently done by using the translation vector and TM-score rotation matrix provided by TM-align.

3.2. Ligand superimposition

For the ligand, with exactly the same number of atom and with a relatively similar geometry, superimpositions were realized using an implementation of Kabsch's algorithm (Kabsch, **1976**) developed by myself in Python 2.7. This algorithm is based on a matrix transformation to find the best rotation and translation matrix between two groups of points in the space.

3.3. Superimposition quality

The quality of the protein superimpositions is measured using the root means square deviation (RMSD) in Å, which characterizes the deviation of paired similar structures in the same referential.

$$RMSD\ (v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + \left(v_{iy} - w_{iy}\right)^2 + (v_{iz} - w_{iz})^2)}$$

where: v and w are a set of n of 3D (x, y, z) coordinates paired.

RMSD is computed using TM-align on Cα paired and has been implemented in Python for the ligand. Lower RMSDs means better superimposition.

Complementary to the RMSD, only for the ligands superimposed, the overlay of the volume- and electrostatic shape was computed using ShaEP software which is based on measuring the overlap between Gaussian functions (Vainio et al., **2009**). ShaEP contains two main functions: (i) a scoring function to score the superimposition of atoms represented by their electrostatic potential and local shape using Gaussian functions; and (ii) a genetic algorithm to find the superimposition that maximizes the score overlap of the molecules. In this thesis, only the scoring was conducted. If two ligands have a good overlay, the scores are maximal.

It is important to note that ShaEP can compare different ligands, in contrast to RMSD calculations.

4. Structure visualization

To visualize the protein-ligand interactions, two tools are used, PyMOL environment (DeLano, **2002**) and LigPlot which allow a visualization of the profile of interaction in two dimensions, using a 3D-2D transformation (Wallace et al., **1995**; Laskowski and Swindells, **2011**). In the late stage of this doctoral thesis the Protein-Ligand Interaction Profiler webserver which uses geometric criteria to find the interaction was also employed (Salentin et al., **2015**). Density maps based on the cloud of the atom are generated using Chimera software (Pettersen et al., **2004**)

5. Statistical analysis and machine learnings

5.1. Descriptors selection

Descriptor pre-filtering is realized by removing (i) uninformative descriptors, e.g. descriptors having a null variance for the pocket set, and (ii) descriptors whose computation returned errors. Another step of descriptor reduction is included in the model learning phase.

5.2. Data visualization

Three types of data transformation are used in this thesis based on different types of data, mostly for data visualization purposes.

5.2.1. Principal Component Analysis (PCA)

PCA is an orthogonal transformation based on a square matrix of covariance computed from a set of descriptors for a dataset, or a profile to reduce the number of dimensions. The projection is realized on a plane that defines a percentage of variability explained by the descriptors in this plane.

5.2.2. Correspondence analysis

Correspondence analysis is an orthogonal transformation based on a contingency table in two entries, individual and classes. The table of contingency is transformed into a table of Chi-square, and individuals and classes are projected on a plane (2D) (Hirschfeld, **1935**). The proximity between two individuals characterizes their similarity and the proximity between a class and an individual characterizes their dependency.

5.2.3. Multidimensional Scaling

Multidimensional Scaling is a method based on the projection in N dimension of a distance matrix. Matrix of values is transformed into distance matrix. Only Euclidian distances are used in this thesis, and this matrix is projected in a plane of N dimensions, generally two, in preserving the relative distance between two individuals by a scaling method, classically the Torgerson–Gower scaling method (Torgerson, **1958**).

5.3. Predicting models

5.3.1. Linear Discriminant Analysis

Linear Discriminant Analysis is a supervised statistical method used to build a statistical classification model based on Fisher's linear discriminant methods (Fisher, **1936**).

5.3.2. Model performance

Classification performances are discussed using the four criteria of quality, accuracy, sensitivity, specificity and Matthew's correlation coefficient (MCC).

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

$$sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$MCC = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TN: true negatives, TP: true positive, FN: false negative and TN: true negative.

6. Programming languages and libraries

All of the scripts were developed using Python version 2.7 with classical libraries: os, re, path, shutil, copy, urllib, gzip; numerical libraries: numpy, sympy, matlib, scipy, math and specific libraries for bioinformatics or chemical data: biopython, openbabel, PyPLIF and PyDPI.

Statistical analysis were conducted using R (Team R Core (R Foundation for Statistical Computing), **2015**) and with the libraries lattice, scatterplot3d, MASS, vrmlgen, FactoMineR, ggplot2, plotrix, rpart, klaR, randomForest, e1071 and kernlab.

The source code has been managed using Git software as source code management system (Hamano et al., **2005**) and using the platform GitHub (https://github.com) for the repository hosting service.

Results and specific discussion

1. PockDrug and PockDrug-Server (Publications I and II)

The prediction of pocket druggability has been presented in Section 2.8 of the Review of the literature.

The main aim of the Publication I is to construct a statistical model capable of predicting pocket druggability, while not being sensitive to the pocket boundaries. In order to do so, a preliminary investigation of the robustness of the pocket estimation method in term of descriptor variability is made. The model is useful in investigating the binding site properties important for druggability.

1.1. Development of the pocket druggability model

The druggability model is similar to a classical QSAR model. The development phases discussed in Section 5.3 of the Review of the literature have been followed.

Original dataset and curation – NRDLD dataset (Krasowski et al., **2011**) is used to train the druggability model. It contains 44 less druggable binding sites and 71 druggable binding sites. Each binding site was estimated using three pocket estimation methods, proximity to the co-crystallized ligand with different distance thresholds, Fpocket and DoGSite. Each pocket was visualized and two estimated by Fpocket are removed because of poor overlay of the ligand position.

Descriptor selection – Most ready-to-use protein pocket descriptors were not available and needed to be implemented. In total, 52 pocket descriptors were used (Table 3). The importance of the descriptors was analysed using PCA, showing that pockets estimated using the three pocket estimation methods are well dispersed in the PCA plan.

Balancing between external test set and training set – The classical division of NRDLD dataset into train and test sets was used in order to allow benchmarking (Krasowski et al., **2011**). In this division, 37 binding sites form the test set, with protein families being well diversified.

Model construction – Linear discriminate analysis has been chosen for machine learning since it is easily interpretable and gives good performance. Descriptor parsimony is controlled using a protocol that consists of testing any combinations of descriptors as input of linear discriminant analysis model and keeping models with the best performances in internal validation and the lowest number of descriptors. The final PockDrug model is a consensus of seven linear discriminant analysis models, each including three descriptors.

Validation – The external validation is performed using the test set from NRDLD set. Another external test set is developed from the DCD database composed of only apo pockets, extracted from protein not complexed. Internal validation is performed using a 10-fold cross-validation and a leave-one-out cross-validation.

1.2. Main results

1.2.1. Binding pockets comparison

While pockets estimated differently were found to exhibit a weak structural overlap in terms of number of common residues (less than 50%), the volume and shape descriptors were noted to be very dependent on the pocket estimation method. In contrast, physicochemical and hydrophobic descriptors were less dependent on the pocket estimation method employed. Example of pockets estimated using the same protein but using different pocket estimation methods are presented in Figure 31, panel A.

PCA was used to analyse globally the pocket spaces occupied by pockets that were differently estimated, and radar plots were used to visualize the linear correlation of the descriptors between the two different pockets sets.

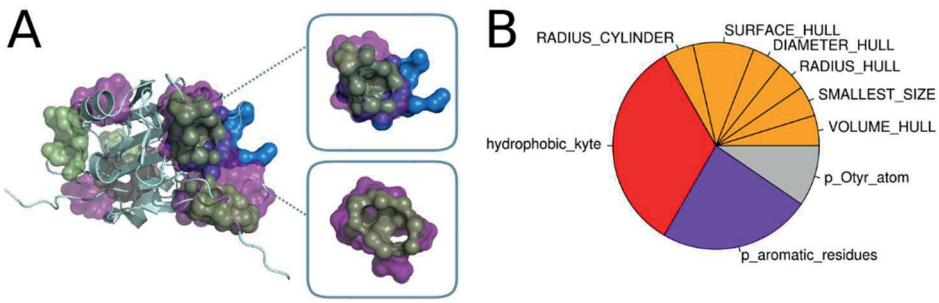


***Figure 31**: (A) Pockets estimated using different pocket estimation methods, in blue using the ligand proximity with a threshold of 4.5 Å, in green using Fpocket and in purple using DoGSite. The protein is the human interleukin-1 β convertase complexed with (3S) - n - methanesulfonyl – 3 - ({ 1- [n - (2 - naphtoyl) -l -valyl] -l - prolyl } amino) –4-oxobutanamide, PDB code 1BMQ (B) Combined descriptors used to build the PockDrug model. In red hydrophobic descriptor, in orange geometric descriptors, in violet aromatic descriptor and in grey atomistic descriptor, adapted from Publication I.*

1.2.2. Druggability model performance

The performance of PockDrug in an external test set was found to be 86.5%. PockDrug was not dependent on the fuzziness of the pocket estimation method. These good performances can be explained by the fact that the most informative descriptors in linear discriminant analysis models are considered as the most robust in pocket estimation methods, such as hydrophobicity or aromatic descriptors, geometric descriptors have a lesser importance. A pie plot presenting the different descriptors used by PockDrug is presented in Figure 31, panel B.

In comparison to other druggability models using the same pocket set and estimated using the same pocket estimation methods, the performance here is more accurate by ~5-10%.

1.2.3. Characteristics for a druggable pocket

The final model is based on a combination of three pocket properties, e.g. (i) hydrophobicity, (ii) geometric properties, shape or volume descriptor and (iii) aromaticity. Hydrophobicity and geometry, but not aromaticity, have previously been reported in the literature.

A dissection of the involvement of the different descriptors in the final model showed that the hydrophobicity property is the key descriptor to explain the druggability. When it was removed, the accuracy decreased from 84% to 63%. It is not surprising that the hydrophobicity is very involved in predicting the druggability, considering its contribution to the binding free energy (see Sections 3.3 and 5.4, Review of the literature)

1.3. PockDrug-Server

The model PockDrug is freely available on the web server http://pockdrug.rpbs.univ-paris-diderot.fr/. This issue is important for the scientific community.

The web server is also able to conduct pocket estimations from protein structures using Fpocket or from a selected distance of a pre-bound ligand. Users can also submit a pocket previously estimated, e.g. visually determined. For each pocket, a set of 17 pocket descriptors is computed and provided in addition to the probability of the selected binding site being druggable.

Problems of applicability domain were encountered during the PockDrug-Server development. Indeed, many less druggable small pockets do not match the applicability domain where PockDrug was developed. A threshold of 14 residues was fixed where the reliability of the druggability prediction is low, i.e. these pockets are outside the applicability domain. This is in agreement with the literature: Perola et al. (2012) reported that only 10% of druggable pockets have between ten and fourteen residues, and Hajduk et al (2005) considered pockets with less than ten residues to be decoy pockets (Hajduk et al., **2005**; Perola et al., **2012**). Furthermore, descriptors computed from a pocket with less than ten residues are also not reliable considering the weak number of atoms included.

1.4. Discussion

1.4.1. Difficulty to define protein druggability

Building a suitable dataset for modeling the druggability of binding pockets is very challenging. The main limitation of the dataset is about being “confident”

that a binding site is non-druggable (see Section 2.8, Review of the literature). We decided to combine two existing widely used datasets.

The challenge in data collection for building a druggability model has been discussed by Crowther et al. (2014):

"*proteins are generally unsuitable for resource-intensive HTS unless they are considered druggable, yet druggability is often difficult to predict in the absence of HTS data.*" (Crowther et al., **2014**)

However, another aspect has to be considered: 52% of the so-called drug-like molecules in the NRDLD dataset have at least one failure with respect to the Lipinski "rules-of-five", as shown in Figure 32.

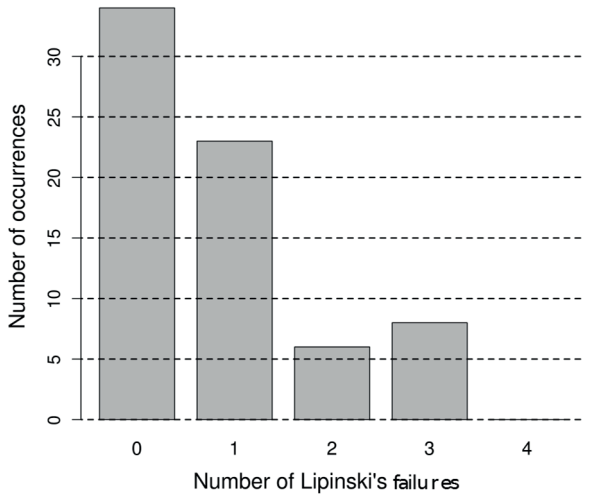


Figure 32: *Number of Lipinski failures for the ligand considered drug-like included in the NRDLD dataset.*

The result can be seen as a paradox: druggability models predict druggable pockets, but in order to improve the druggability models, the definition of the drug-like molecules should be modified.

1.4.2. From druggability models to binding profiles

Druggability models seem to perform well, with accuracies close to 90%. It is therefore logical to ask whether a global model would also be possible to predict (and profile) from pocket descriptors binding of other compounds such as antibiotics or peptides. Even if ligand properties are associated with pocket properties (Pérot et al., **2013**), this type of accurate global model is not a realistic issue today. Challenges to overcome are the diversity in term of ligand and protein chemistries, as well as local target-dependent challenges such as the presence of water molecules at binding sites. In the case of druggability predictions, the chemical space is drastically reduced in a type of ligand that exhibits similar properties, characterized for example by the Ro5 (Section 2.7.1, Review of the literature), which reduces the number of factors introducing predicting errors.

2. Structural replacements of phosphate groups (Publication III)

The concepts of bioisosterism and phosphate replacements have been presented in Section 5.1 of the Review of the literature.

The general aim of this study was to develop a chemoinformatic approach, based on a data-mining using the PDB, useful for suggesting new ligand structural replacements (LSR) for a chemical group. Phosphate group represents an ideal test case since there are many potential reference proteins, such as kinases, phosphatases or ligases, whose endogenous ligands contain a phosphate moiety. Secondly, phosphate bioisosteres is a very active branch of medicinal chemistry since phosphate is generally an unsuitable group in drug molecules.

2.1. Chemoinformatics approach to define structural replacements

Chemoinformatics approach is based on mining the PDB to extract from crystallized homologues complexes superimposed the LSRs. The developed computational workflow is divided into six steps: (i) screening and extraction of proteins containing ligand with at least one phosphate group; (ii) mining of these proteins (or very close homologues or mutants) in complex with other ligands using balstp algorithm; an e-value threshold of blastp to define two homologues fixed at 10^{-100}; (iii) structural 3D superimposition of the reference protein containing the ligand and studied phosphate group with homologue protein; (iv) extraction of the ligand structural replacement based on structural overlap; and (v) clustering of structural replacements based on their SMILES codes composition. An example of structural replacement is presented in Figure 33.

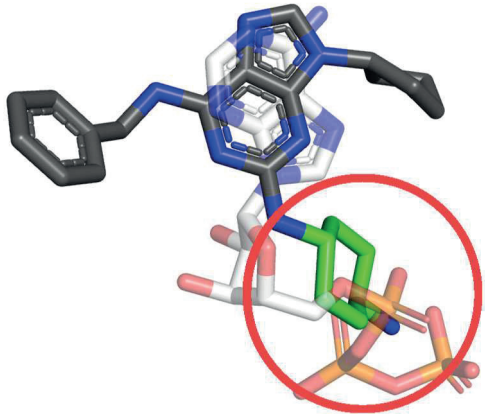


Figure 33: *Example of final ligand superimposition between an ATP, containing phosphate groups, carbon represented in white transparent and the 2-[trans-(4-aminocyclohexyl)amino]-6-(benzyl-amino)-9-cyclopentylpurine, carbon represented in black. The LSR corresponds to the ligand part with the carbon in green encircled in red.*

2.1. Main results

2.1.1. Phosphate replacement

Form the PDB, a set of 15 819 phosphate LSRs were extracted using AMP, ADP, ATP and POP as reference ligands. LSRs are clustered in 16 non-overlapping categories containing phosphorus; boron; fluorine; chlorine; bromine; beryllium; NO2 (generally nitro); SO2; S; CON (generally carbamoyl or amide); COO (ester or carboxylic acid); exclusively C; only C and O; only C and N; only C, O and N; and Other.

2.1.2. Mechanisms of phosphate replacements

In some cases, the proteins replace the phosphate by a 3D rearrangement, i.e. loop displacement, or a side-chain displacement (example in Figure 34, panel A). This is no longer the case in a ligand structural replacement.

The different chemistry of the LSRs is discussed in parallel with the different mechanisms underlying the replacements (presented in Figure 34). (i) Phosphate is often exposed to the solvent; congeners seem to be generally

permissive to chemical substitutions. Water molecule rearrangement may explain the permittivity of the structural replacement. (ii) Metals are present in 55-71% of the phosphate binding sites. Some examples in which a base replaced the metal near the phosphate were found (see e.g. Figure 34, panel B). (iii) Intramolecular ligand interaction, where the ligand adopts a U-shape was noted to transcend protein families. SAR studies suggest that a destabilizing of intramolecular interaction can lead to a drop in the affinity (example in Figure 34, panels C and D).

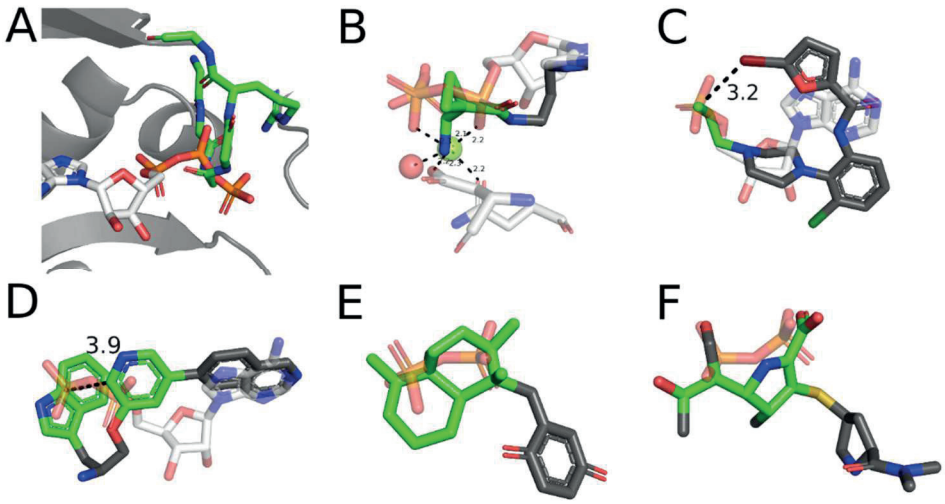


***Figure 34**: Examples of phosphate LSRs, included in Publication III. (A) Phosphate replaced by the protein binding site, Escherichia coli, biotin carboxylase, PDB code 3JZI; ligand JZL (reference 1DV2/ATP). (B) LSR replacing a Mg^{2+} human cyclin dependent-kinase 2, PDB code 3ULI; ligand 1N3 (reference 1DV2/ATP). (C) U-shape replacements in human c-Jun N-terminal kinases, PDB code 3FV8, ligand JK3 (reference 4KK3/AMP). (D) U-shape replacement in Bos Taurus, protein kinase A, PDB code 2F7E, ligand 2EA (reference 1JBP/ADP). (E) Hydrophobic group phosphate replacing, human farnesyl diphosphate synthase, PDB code 4P0W, ligand 1XH (reference 4H5D/POP). (F) Penicillin phosphate replacement; Staphylococcus aureus, sensor domain of BlaR1, PDB code 3Q82, ligand MER (reference 1XA1/POP).*

2.1.3. Previously unrecognized phosphate isosteres

Classical phosphate replacement, such as replacement containing phosphate, carboxyl, esters, sulfones and sulfonamides was identified with this workflow. The approach also highlighted less classical phosphate replacements such as hydrophobic rings, polar rings, nitrile groups and amide groups. Aliphatic apolar groups are also found for phosphate replacement, which is surprising considering that phosphate groups are negatively charged (see example Figure 34, panel E). Some other replacements are interesting in medical chemistry, for example in Figure 34 panel F, penicillin G, which contains a lactam scaffold, replaces phosphate groups. Other miscellaneous replacements, such as the replacement of the phosphate by positively charged groups have been found and discussed in term of affinity. These unrecognized LSR for the phosphate, extracted from a PDB data-mining give a new perspective in medicinal chemistry for phosphate replacement.

2.2. Discussion

2.2.1. Improvement of extracting workflow

The computational workflow has been optimized for phosphate groups. While this approach works for small chemical groups, such as phosphate groups, it is limited when facing more complex substructures; furthermore, it is not flexible. An improvement of the workflow would consist of reconstruction of all connectivity matrices (as is done in order to run ShaEP) and mining them directly for substructure search using a graph-matching algorithm. This type of algorithm has been pioneered by the Ullman algorithm (Ullmann, **1976**). In retrospect, it would have been better to build a more easily generalizable tool from the start.

The computational workflow suffers from some limitations due to the methods used. The extraction of the type of LSR based on a regular expression search on the SMILES strings (generated from the LSR extracted from the pdb file) introduced problems in the LSR hierarchical organization. For example, a current problem is that the conversion of two separate fragments into SMILES connects them as part of a ring system. When the chemical group is incomplete, i.e. the ring is not completely included in the LSR region, these fragments are not properly classified. If the SMILES format is chosen for simplicity and

speed, a substructure chemical search based on the 3D structure should avoid potential problems.

2.2.2. Generalization to other chemical substructures

A main perspective of this work is application of the same protocol as for the phosphate group for new chemical substructures.

This has already been done for ribose, but the data are not yet analysed. Examples of replacements found are presented in Figure 35; ribose was replaced with a sulfonamide, heterocycle or hydrophobic cycle.

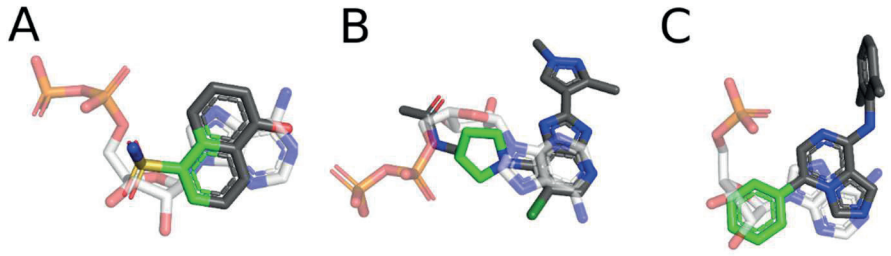


Figure 35: *Example of ribose LSRs extracted using the protocol include in Publication III. (A) Sulfonamide replacement from Spodoptera frugiperda transferase, PDB code 2VTH, ligand LZ2 (reference 4I3Z/ADP). (B) Heterocycle replacement from Escherichia coli aurora kinase, PDB code 4BYI, ligand FH3 (reference 2WQE/ADP). (C) Cycle replacement from human transferase, PDB code 2ZYB, ligand KSL (reference 3DQX/AMP).*

3. Neighbourhood of ionizable groups (IV)

Salt bridges (charge-reinforced H-bonds) comprise one of the strongest interactions in protein-ligand interactions (see Section 4.2, Review of the literature).

Salt bridges are formed between acidic and basic groups and are characterized by the sharing of a proton. Although well studied in proteins, they have been only poorly characterized in the case of protein-ligand complexes. This is probably due to the difficulty in identifying specifically the atoms that belong to the ionizable groups from a ligand in a 3D structure, the lack of ready-to-use datasets and the relative difficulty in operating chemoinformatics data mining tools in the PDB.

A better understanding of salt bridge frequencies would facilitate anticipating them in, for instance, docking simulations, which would enable better assignment of formal charges to the ligand and protein structures. Ligands and proteins are, however, commonly prepared by enumerating ionization states at best. The main goal of this study was thus to dissect and quantify the frequency by which a ligand forms a salt bridge, given that it contains a basic or acidic group.

3.1. Data mining

Six ionizable functional groups are considered. Five basic groups, i.e. primary amine, secondary amine, tertiary amine, imidazole and guanidinium, and one acidic group, carboxylic acid, are considered.

Two main challenges have been faced: (i) the extraction of ionizable functional groups from ligands from the PDB and (ii) controls for the quality of the structures extracted, especially in analysing water molecules, poorly positioned at low resolution.

The extraction of ionizable groups, i.e. the detection of their substructures, would benefit from using a graph-matching algorithm rather that the case-specific code that is used. It would render the study much more flexible to analyse other functional groups. This problem has already been discussed above for detection of phosphate groups (Section 2.2 of Results and specific discussion).

Controls regarding the quality of structures is an important issue. Two datasets of different resolution have been built, containing non-redundant proteins complexes with the ligand's specific substructures (one ligand may contain several substructures). At 1.5 Å resolution, we collected 161 with primary amines, 91 with secondary amines, 64 with tertiary amines, 26 with an imidazole, 11 with guanidinium and 96 with a carboxylic acid. At 3.0 Å resolution, the figures were 1491 primary amines, 1113 secondary amines, 1020 tertiary amines, 251 imidazoles, 134 guanidinium and 1390 carboxylic acids. The R free value is controlled in both datasets to be below 0.25, which is a standard value for mining contact data in PDB complexes. The dataset at 1.5 Å resolution was used to analyse the position of water molecules.

3.2. Main results

3.2.1. Fraction of substructures stabilized by salt bridges

This study quantifies the proportion of ionic interactions for the six chemical groups considered: 54% of the ligand's primary amines are involved in salt bridges, and these numbers are 53% for secondary amines, 16% for tertiary amines, 15% for imidazole, 72% for the guanidinium group and 53% for carboxylic groups. For ligands with guanidinium and primary amino groups, these proportions are similar to those previously reported for proteins. The low proportions found for tertiary amines – which have pK_a of 8-10, similar to that of primary amines – are suggested to be linked to the lower accessible volume for interaction.

Water molecules are analysed using the high-resolution dataset. Salt bridge interactions mediated by water molecules are found for 26% of primary amines, 22% of secondary amines, 5% of tertiary amines and 26% of carboxylates of the protein-ligand interaction.

3.2.2. Environments that stabilize ionizable groups

We developed a method to analyse the neighbourhood of the ionizable groups under scrutiny by focusing on their closest neighbours (closest interacting atoms), collecting types and interaction distances. An example of the investigation of environments is presented in the case of primary amines in

Figure 36. Environments are investigated using different approaches, based on a 3D visualization in panel A, and analysis of neighbour atoms in panel B.

The B panel demonstrates that primary amines are in most cases interacting with no more than three atoms within a distance of 3.5 Å (only non-bonded interactions examined). When considering the six closest neighbours (collected within a distance of 6 Å), the types of interacting atoms are specific for the first four neighbours as well, and thereafter resemble the neighbourhood of any atom. The combination of neighbours is also demonstrated to be enriched in specific preferred neighbourhoods, the most commonly observed being carboxylic acid oxygen, two acyl oxygens and one water molecule. Thus, the interaction properties of a primary amine are well represented by considering only its closest four closest neighbours.

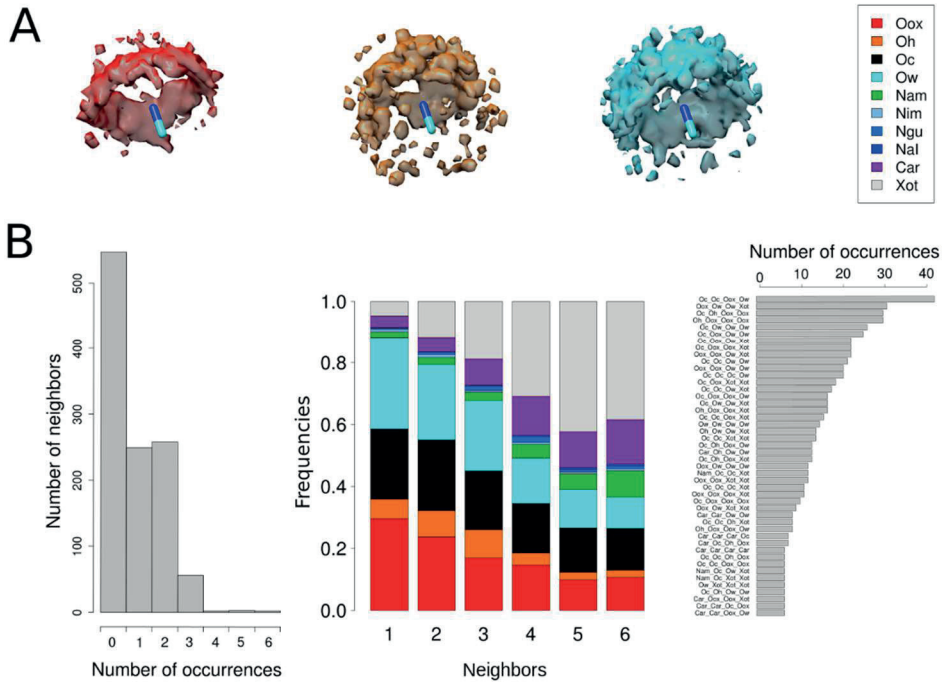


Figure 36: *Environment investigation of primary amine chemical group. (A) Representation of the map density of the position of three types of atom neighbours, in red oxygen atom from carboxylate groups (Oox), in orange oxygen atom from hydroxyl groups (Oh) and in blue water molecules (Ow). (B) From left to right, number of neighbour atoms in the shell of 3.5 Å; proportion of different neighbour atom types for the six first neighbours collected within 6Å; and distribution of the different combination of four atom neighbours; rank 1: Oox + Oc (oxygen from an acyl group) + Oc + Ow; rank 2: Oox + Ow + Ow + Xot (carbon aliphatic and sulfur included in all amino acid); rank 3: Oox + Oox + Oc + Oh. Nam: nitrogen of amine*

group; Nim: nitrogen of imidazole group, Ngu: nitrogen from guanidinium group, NaI: nitrogen from primary amine, Car: carbon in aromatic group.

The method allows demonstration of a clearly different environment in cases where the protein carries a fully charged counter ion and in its absence. See, for example, Figure 37 for primary amine. In the case where no counter ion is present in the vicinity of the ligand ionizable group, the environment that stabilizes the chemical group often includes a weakly ionic hydrogen bonding group such as hydroxyl or water molecules. In addition, the environment of basic groups includes H-bond acceptors such as acyl substituents.

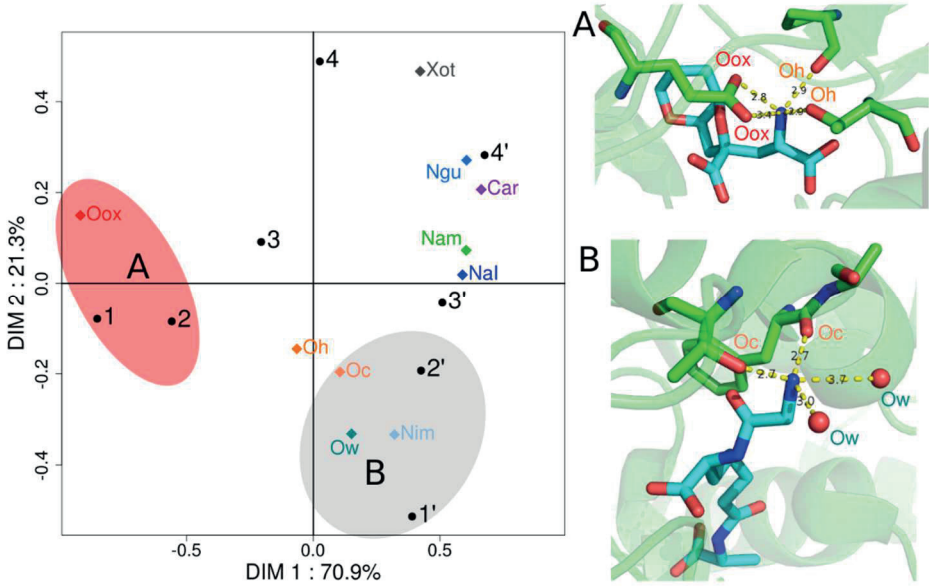


Figure 37: *Correspondence analysis of the contingency table in terms of atom type close to the primary atom, characterizing the neighbouring atoms (1, closest atom, 2, second closest, etc.). Two environments are considered, i.e. salt bridges (A-red, neighbours 1-4) and no salt bridges (B-grey, neighbours 1'-4'). Environment (A) is composed mainly of carboxylate group (Oox) and hydroxyl group (Oh) from the protein. The environment (B) that do not include a fully charged neighbour is composed mainly of water molecules (Ow), acyl groups (Oc), imidazole nitrogen atoms (Nim) and to a lesser extent hydroxyl groups (Oh). Adapted from Publication IV.*

3.3. Discussion

3.3.1. Role of water molecules

The distance distribution of water molecules near basic groups resembles the distance distribution observed for acidic counter ion (carboxylic acid). The study suggests a previously unrecognized role as counter-ions in the vicinity of acidic and basic groups bound to ligands. This phenomenon appears relatively common in the complexes studied. Furthermore, the relatively significant proportion (5-25%) of salt bridges mediated by water molecules is in contrast to the poor consideration that this phenomenon receives in scoring functions and in binding free energy computations.

3.3.2. Considering the closest neighbours to represent the interacting atoms

In Publication IV, we demonstrate that the closest neighbouring atoms from functional groups contain significant information about the interactions that take place. Apolar atoms are also found in the list of close contacts, although hydrophobic contacts are more distant that H-bonding contacts (~4-4.5 Å vs. ~2.7-3.3 Å). The novelty of the method is that all interacting atoms in the close environment are considered at once and not individually, yielding additional information. This study thus paves the way for improvements that would concentrate on only the neighbouring atoms in, for example, score binding poses or local adjustments of binding poses. The number of neighbouring atoms to be considered obviously depends on the functional group at hand. A potential drawback of this type of method is that the sensitivity to small shifts in the binding site would be high.

Publication IV also attempts to bridge the information gap between the interactions (or H-bond networks) considered at the level of a functional group and those considered at the level of isolated atoms.

4. Post-docking selection for a pharmacophore model to discover OX1 and OX2 orexin receptor ligands (Publication V)

The orexin peptide−orexin receptor system is an important regulator of the sleep-wakefulness cycle. Orexin receptors are peptidic G protein-coupled receptors. The goal of this study was to discover agonists of these receptors, none of which have been reported in the literature to date. A protocol combining *in vivo* and computational methods was developed.

This protocol divides into four steps: (i) development of a pharmacophore model based on ~200 antagonists validated by screening a collection of ~137 000 chemically diverse compounds; (ii) pharmacological screening using 395 compounds in the hit list of step (i); (iii) of the 47 most promising compounds, four were validated for low □M orexin agonist activity and seven for high nM antagonist activity; and (iv) docking was used to investigate the molecular mechanism of orexin agonist activity.

4.1. Main results

4.1.1. Discovery of new agonist and antagonist compounds

As the main result of this study, we identified four compounds with promising partial agonist activity and $K_i s$ in the 1−30 μM range (1.5% hit rate), as well as seven antagonists with $K_i s$ in the 0.1−10 μM range for the G protein-coupled receptors OX1 and 1−50 μM for the G protein-coupled receptors OX2 receptor (1.7% hit rate).

4.1.2. Docking simulation and pose selection

Docking simulations were conducted using the Glide-XP induced-fit protocol, including two binding site water molecules, with the experiment performed by my collaborators. A variety of complexes were suggested for each ligand (5-20 poses), each complex being composed of a ligand and a slightly rearranged protein, especially at the level of the side chains. Upon visual examination, the scores and the overlap with the ligand bound in the X-ray (our reference) were found to be independent. This is illustrated in Figure 35, where the pose with rank 1 is also optimal with respect to shape overlap to the reference ligand, as

well as fingerprint similarity to the reference ligand (i.e. the best solution by all three criteria is the "real" best solution). In contrast, in the redocking experiment the second best score shows very limited overlap with the real solution.

A tool to analyse docked poses was developed. The protocol uses three steps: (i) superimposition of the reference protein and the docked complex, (ii) computation of the volume of overlap using ShaEP software between the docked ligand and the reference ligand and (iii) a comparison of the interaction profile between the complex reference and including the pose by using interaction fingerprints and Jaccard score similarity.

The results were useful in identifying the poses that were optimal for shape and fingerprint interactions to the reference. The method is discussed in Figure 38 using a flexible redocking of Suvorexant to the native X-ray structure.

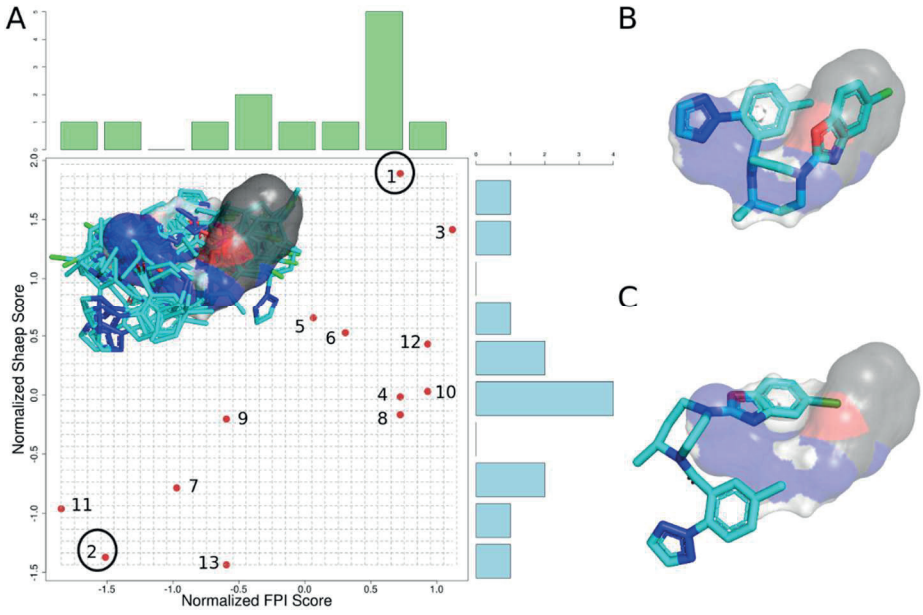


Figure 38: *Example of pose docking selection using crossing information between ligand volume overlap and fingerprint interaction between the ligand in the co-crystallized suvorexant, represented in surface and poses of suvorexant redocked using Glide. (A) Graphic representation of crossing information between Jaccard similarity scores of interaction fingerprint and ShaEP scores, normalized on the distribution, between each pose and the reference co-crystalized ligand. The rank of poses based on GlideScore is also indicated. (B) Representation of the poses ranked*

in the first position and the surface of the co-crystallized ligand in the same referential. (C) Representation of the pose ranked in the second position and the surface of the co-crystallized ligand in the same referential.

4.2. Discussion

4.2.1. Generalization of this protocol for any pose selection: problem of reference ligand

The selection of a reference for a group of poses is a crucial issue for this method. Different co-crystalized ligands may be available for a given protein and will influence the results. The code that we implemented is able to identify for each ligand studied the reference ligand sharing the largest maximum common substructures. However, a single reference is used in Publication V.

4.2.2. Limitation from the interaction fingerprints

Modeling of protein ligand interaction is conducted using interaction fingerprints. However, only seven types of interactions are considered in the fingerprint package PyPLIF (Radifar et al., **2013**) based on fingerprint interaction as defined by Marcou et al. (2007): (i) apolar interaction, (ii) aromatic π-stacking, (iii) aromatic T-shape, (iv) H-bond (protein as donor), (v) H-bond (protein as acceptor), (vi) electrostatic positive interaction (protein positively charged) and (vii) electrostatic negative interaction (protein negatively charged) (Marcou and Rognan, **2007**). The fingerprint does not interpret the chemistry of the ligand; it is thus not capable of assigning "atom types" and hydrogen atoms. This is very problematic, for example, for an oxygen atom, which depending on its hybridization state, can be only an acceptor or a donor/acceptor of a hydrogen bond. Particular interaction types, i.e. halogen bond, are not considered, limiting the overall quality of this method. Furthermore, only protein-ligand contacts, not water molecules, are considered.

Unpublished results - Positioning of water molecules and estimation of their favourable displacement

Water molecules play an important role in protein-ligand recognition in, for example, mediating salt-bridge interaction (Publication IV) or in participating in a mechanism that explains local structure replacements (Publication III). Furthermore, modelling water molecules is a great challenge for numerous computational methods such as free-energy estimation, molecular docking and design of analogues (see Section 4.5, Review of the literature).

Only a few water molecule positioning methods are available to the scientific community, and they are generally associated with an expensive commercial package, e.g. WaterMap in the Schrodinger suite. Methods based on molecular dynamics require long computation and parametrization times and are difficult to generalize to all types of targets (see Section 5.5, Review of the literature).

The aim of this project was to develop a novel method for predicting positions of water molecules in binding sites and assigning for each one an index attributing how favourably or unfavourably replaceable they are. To overcome the limitation of the current model, based on long molecular dynamics, this method is based on positioning of water molecules using known position of water molecules included in crystallographic structures, extracted from the PDB. Currently, geometry models exist to investigate water molecules in the binding site (see Section 2.2, Review of the literature), but none are used both to position and to estimate the favourable displacements of water molecules.

1. Positioning method

A high-quality subset of the PDB, resolution below 1.5 Å and R-free below 25%, was collected. Thresholds of quality criteria were fixed empirically as a balance between the number of water molecules presented by structures and the number of structures available. A set of 1280 structures was collected.

Two methods were tested, a geometric method and a grid-based method. The grid-based method was computationally inefficient and we decided not to pursue it further. Only the geometric approach is described herein (see Figure 39). It is divided into two steps, mapping water molecules around the protein and estimating a desirability index that characterizes their displacements. The second part of the study is ongoing.

The water molecule mapping step is based on the construction of the cartographies of water molecules, characterizing the preferential positions of water molecules around protein amino acids. The 20 amino acids were fragmented, identifying rigid functional groups, e.g. imidazole from histidine or carboxyl group from aspartic acid. The shell around each fragment, including water molecules with 1, 2, 3 or more than 4 interactions, were collected independently. A water molecule contact is defined when the distance between the oxygen atom of a water molecule and a protein's atom is below 3.2 Å.

Water molecule mapping then consists of combining all cartographies and mapping them on the fragmented amino acids of a new protein. Clashes are removed, and the number of contact points with the protein is estimated using geometric methods. Distance clashing between the protein atom and water molecule is fixed at 2 Å.

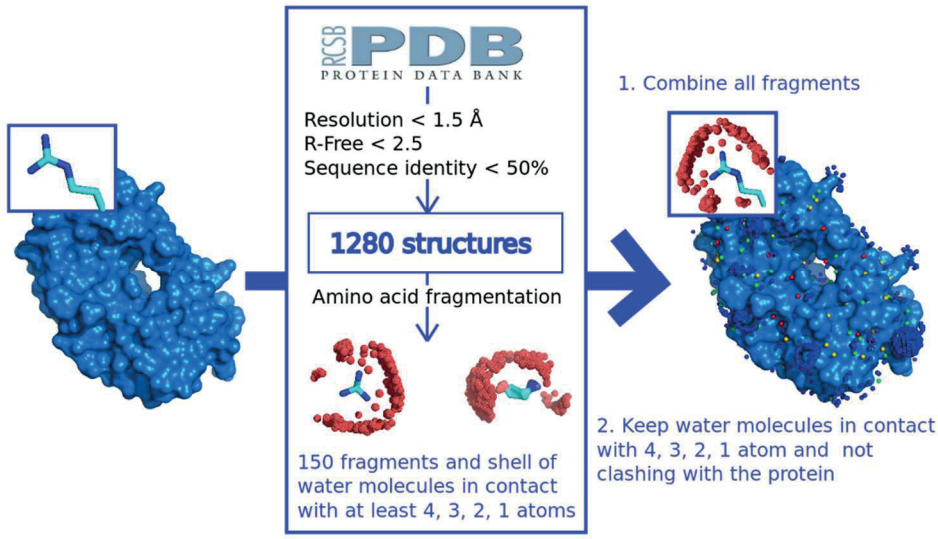


Figure 39: *Method of water positioning based on amino acid fragmentation. Water molecules are presented using a sphere, in red for water molecules with at least 4 contacts, in orange for 3 contacts, in yellow for 2 contacts and in green for one contact with the protein.*

2. Positioning quality

A test set of 100 high-quality proteins, not included and with a sequence identity inferior to 50% with the training set was built. The quality of the water molecules positioning is discussed using the number of water sites that were present in the test set found. Considering a distance of 2 Å between water molecules crystallized and water molecules placed by our algorithm, ~80% of the water sites are found (Figure 40).

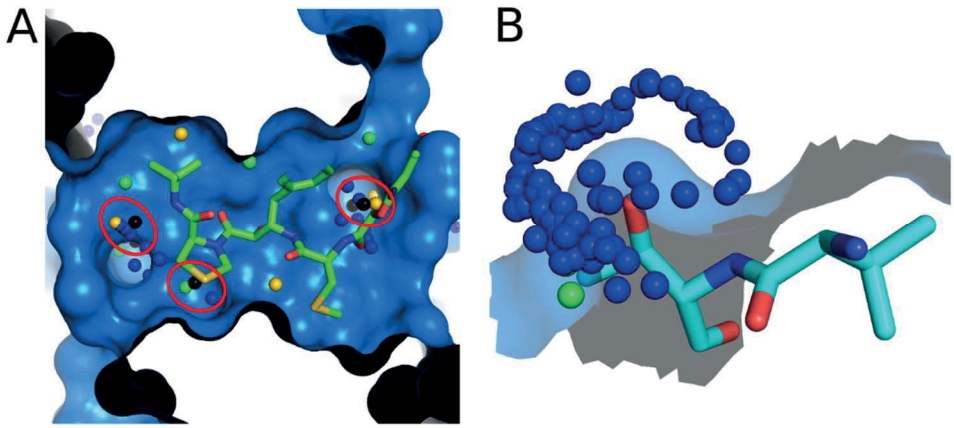


Figure 40: *Example of water positioning (A) from complex of Human immunodeficiency virus 1 protease complexed with the inhibitor KNI-272, PDB code 1HPX previously desolvated and (B) focused on an exposed polar oxygen on the protein surface. Water molecules are presented using spheres, in red for water molecules with at least 4 contacts, in orange for 3 contacts, in yellow for 2 contacts and in green for one contact with the protein. The sites of water correctly found are presented in a red sphere.*

3. Future developments

The current development of this work is divided into three parts: (i) improve water molecules mapping protocol, (ii) develop a desirability index that quantifies the desirability displacement of each water molecule and (iii) benchmark this method with the approaches available in the literature.

For a water molecule mapping protocol, three directions are being investigated: (i) reduce the number of extra water molecules by considering only water molecules with the best geometry in the water network, i.e. by identifying the best water network around the protein and removing redundant water

molecules not included in this network, (ii) refine the cartography. Develop new cartographies for most common ligand fragments, possibly degenerated, that would allow considering water in the protein-ligand interface, (iii) focus the positioning for water molecules in the protein-ligand interface or in protein pockets.

The most challenging part of this development consists of developing an index of water desirability that considers (i) the nature of the direct neighbour's atoms, with or without neighbouring water molecules, (ii) the water network perturbation, (iii) the mobility of water molecules and (iv) the neighbouring atoms of a ligand. This index should be connected to the free energy of the protein-ligand interaction. At the moment, only the number of contacts with a protein for a cluster of water molecules is considered. Finally, the validation of this method needs to be improved by, for example, benchmarking against well-documented examples of HIV-1 protease (Li and Lazaridis, **2003**; Kellogg and Chen, **2004**; Beuming et al., **2012**).

Concluding remarks

1. Program availability

Among the main problems encountered during this thesis was the lack of availability of computational tools developed by the community. For example, for protein pocket estimation methods many articles present new algorithms, but only a few of them are available either as source code or using a web server.

This problem raises the question of scientific reproducibility and reliability of findings, especially for computational sciences; algorithms are not always available and some methods are never publicly released. A well-recognized issue has been, for instance, the retraction of five membrane protein X-ray structures due to an erroneous in-house script (Chang et al., **2006**). The lack of distribution for computation tools has several origins: (i) technical difficulties, (ii) monetization and (iii) protectionism by the institutions to be more competitive, e.g. for grant applications (Morin et al., **2012**; Walters, **2013**).

In contrast, many platforms exist to distribute source code and data such as F1000research for research or GitHub for source code (Black, **2014**; Bajorath, **2015**). The scientific community is also more sensible about the problem of reproducibility for computational work as shown by the editorial recommendation of the American Chemical Society journal to distribute source code (Matters, **2010**).

2. Protein-ligand affinity

Affinity between a ligand and a protein is difficult to take into account in work that statistically addresses molecular interactions. This is due to the relative difficulty in mining binding affinity values when this study was started as well as the low relevance of the values where the bound compounds differ in several places. Nonetheless, it would be interesting to consider the distribution of binding affinities in the complexes used in Publications III compared with the PDB, e.g. extracting data from PDBbind (Wang et al., **2004**).

For Publication III, we decided to read selected original publications in order to discuss relevant affinity changes (see Section 2, Results and specific discussion).

Generally, we should be able to access relatively well-curated binding data from, for instance, the MOAD database, which contains a large number of protein-ligand complexes, 8156 binding affinities for 23 269 complexes extracted from the PDB (Hu et al., **2005**; Ahmed et al., **2015**).

3. Conclusion

Limitations in predicting which types of ligands can bind to which types of proteins are numerous. Computational methods, although complex, tend sometimes to generate trivial predictions that can be summarized as "small ligands bind small pockets and large ligands bind large pockets". In general, we are not able to understand and model all phenomena responsible for protein-ligand recognition.

Taken together, the six studies here investigate different aspects of protein-ligand recognition, considering the pocket description (Publications I and II), ligand replacements tolerated for a binding site (Publication III), specific interaction types (Publication IV) and water molecules (Unpublished result).

The methods developed provide a starting point for further computational developments that should lead to improvements in our understanding of the phenomena involved in protein-ligand recognition.

References

Abad-Zapatero, C. (**2007**). *A Sorcerer's apprentice and The Rule of Five: from rule-of-thumb to commandment and beyond*. Drug Discov. Today *12*: 995–997.

Abel, R., Young, T., Farid, R., Berne, B.J., and Friesner, R.A. (**2008**). *Role of the active-site solvent in the thermodynamics of factor Xa ligand binding*. J. Am. Chem. Soc. *130*: 2817–2831.

Abraham, M.H., Gola, J.M.R., Kumarsingh, R., Cometto-Muniz, J.E., and Cain, W.S. (**2000**). *Connection between chromatographic data and biological data*. J. Chromatogr. B Biomed. Sci. Appl. *745*: 103–115.

Ahmed, A., Saeed, F., Salim, N., and Abdo, A. (**2014**). *Condorcet and borda count fusion method for ligand-based virtual screening*. J. Cheminform. *6*: 1–10.

Ahmed, A., Smith, R.D., Clark, J.J., Dunbar, J.B., and Carlson, H.A. (**2015**). *Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures.* Nucleic Acids Res. *43*: D465–D469.

Allen, F.H. (**2002**). *The Cambridge Structural Database: A quarter of a million crystal structures and rising*. Acta Crystallogr. Sect. B Struct. Sci. *58*: 380–388.

Alvarez-Garcia, D., and Barril, X. (**2014**). *Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites.* J. Med. Chem. *57*: 8530–8539.

Amaro, R.E., Baron, R., and McCammon, J.A. (**2008**). *An improved relaxed complex scheme for receptor flexibility in computer-aided drug design*. J. Comput. Aided. Mol. Des. *22*: 693–705.

An, J., Totrov, M., and Abagyan, R. (**2004**). *Comprehensive identification of 'druggable' protein ligand binding sites.* Genome Inform. *15*: 31–41.

An, J., Totrov, M., and Abagyan, R. (**2005**). *Pocketome via comprehensive identification and classification of ligand binding envelopes.* Mol. Cell. Proteomics *4*: 752–761.

Aqvist, J., and Marelius, J. (**2001**). *The linear interaction energy method for predicting ligand binding free energies.* Comb. Chem. High Throughput Screen. *4*: 613–626.

Arenas-Salinas, M., Ortega-Salazar, S., Gonzales-Nilo, F., Pohl, E., Holmes, D.S., and Quatrini, R. (**2014**). *AFAL: A web service for profiling amino acids surrounding ligands in proteins*. J. Comput. Aided. Mol. Des. *28*: 1069–1076.

Armon, A., Graur, D., and Ben-Tal, N. (**2001**). *ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.* J. Mol. Biol. *307*: 447–463.

Arrowsmith, J. (**2011**). *Trial watch: Phase II failures: 2008–2010*. Nat. Rev. Drug Discov. *10*: 328–329.

Arunan, E., Desiraju, G.R., Klein, R.A., Sadlej, J., Scheiner, S., Alkorta, I., et al. (**2011**). *Defining the hydrogen bond: An account (IUPAC Technical Report)*. Pure Appl. Chem. *83*: 1637–1641.

Ashida, T., and Kikuchi, T. (**2015**). *Overview of Binding Free Energy Calculation Techniques for Elucidation of Biological Processes and for Drug Discovery*. Med. Chem. (Sharjah, United Arab Emirates) *11*: 248–253.

Bajorath, J. (**2015**). *Entering new publication territory in chemoinformatics and chemical information science*. F1000Research *35*: 9–12.

Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M.S., and Drie, J.H. Van (**2009**). *Navigating structure-activity landscapes*. Drug Discov. Today *14*: 698–705.

Bajusz, D., Racz, A., and Héberger, K. (**2015**). *Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?*. J. Cheminform. *7*: 1–13.

Bakan, A., Nevins, N., Lakdawala, A.S., and Bahar, I. (**2012**). *Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules.* J. Chem. Theory Comput. *8*: 2435–2447.

Ballatore, C., Huryn, D.M., and Smith, A.B. (**2013**). *Carboxylic Acid (Bio)Isosteres in Drug Design*. ChemMedChem *8*: 385–395.

Barnard, J.M., Downs, G.M., Scholley-Pfab, A. Von, and Brown, R.D. (**2000**). *Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries*. J. Mol. Graph. Model. *18*: 452–463.

Baxter, C.A., Murray, C.W., Clark, D.E., Westhead, D.R., and Eldridge, M.D. (**1998**). *Flexible docking using Tabu search and an empirical estimate of binding affinity*. Proteins Struct. Funct. Genet. *33*: 367–382.

Bayden, A., Moustakas, D.T., Joseph-McCarthy, D., and Lamb, M.L. (**2015**). *Evaluating free energies of binding and conservation of crystallographic waters using SZMAP*. J. Chem. Inf. Model. *55*: 1552–1565.

Bender, A., Mussa, H.Y., Gill, G.S., and Glen, R.C. (**2004**). *Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT)*. Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern. *5*: 4553–4558.

Benfenati, E., Gini, G., Piclin, N., Roncaglioni, A., and Varì, M.R. (**2003**). *Predicting logP of pesticides using different software*. Chemosphere *53*: 1155–1164.

Benkaidali, L., André, F., Maouche, B., Siregar, P., Benyettou, M., Maurel, F.F., et al. (**2014**). *Computing cavities, channels, pores and pockets in proteins from non-spherical ligands models*. Bioinformatics *30*: 792–800.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al. (**2000**). *The Protein Data Bank.* Nucleic Acids Res. *28*: 235–242.

Betz, M., Wulsdorf, T., Krimmer, S., and Klebe, G. (**2016**). *Impact of Surface Water Layers on Protein-Ligand Binding: How Well Are Experimental Data Reproduced by Molecular Dynamics Simulations in a Thermolysin Test Case*. J. Chem. Inf. Model. *56*: 223–233.

Beuming, T., Che, Y., Abel, R., Kim, B., Shanmugasundaram, V., and Sherman, W. (**2012**). *Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization.* Proteins *80*: 871–883.

Beveridge, D.L., and Dicapua, F.M. (**1989**). *MOLECULAR SIMULATION : Applications to Chemical and Biomolecular Systems*. Annu. Rev. Biophys. Biophys. Chem. *18*: 431–492.

Bickerton, G.R., Paolini, G. V, Besnard, J., Muresan, S., and Hopkins, A.L. (**2012**). *Quantifying the chemical beauty of drugs.* Nat. Chem. *4*: 90–98.

Biela, A., Nasief, N.N., Betz, M., Heine, A., Hangauer, D., and Klebe, G. (**2013**). *Dissecting the Hydrophobic Effect on the Molecular Level: The Role of Water, Enthalpy, and Entropy in Ligand Binding to Thermolysin*. Angew. Chemie Int. Ed. *52*: 1822–1828.

Bill, R.M., Henderson, P.J.F., Iwata, S., Kunji, E.R.S., Michel, H., Neutze, R., et al. (**2011**). *Overcoming barriers to membrane protein structure determination.* Nat. Biotechnol. *29*: 335–340.

Bissantz, C., Kuhn, B., and Stahl, M. (**2010**). *A medicinal chemist's guide to molecular*

interactions. J. Med. Chem. *53*: 5061–5084.

Black, K.J. (**2014**). *F1000Research: Tics welcomes you to 21st century biomedical publishing*. F1000Research 1–6.

Blender Foundation (**2016**). *Blender 2.77, https://www.blender.org/*.

Böhm, H.J. (**1994**). *The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure*. J. Comput. Aided. Mol. Des. *8*: 243–256.

Bohm, H.-J., and Schneider, G. (**2000**). *Virtual Screening for Bioactive Molecules, Volume 10* (Wiley-VCH; 1 edition (November 17, 2000)).

Böhm, H.-J., and Schneider, G. (**2012**). *Protein-Ligand Interactions*. Saudi Med J *33*: 3–8.

Brady, G.P., and Stouten, P.F. (**2000**). *Fast prediction and visualization of protein binding pockets with PASS*. J. Comput. Aided. Mol. Des. *14*: 383–401.

Breiten, B., Lockett, M.R., Sherman, W., Fujita, S., Lange, H., Bowers, C.M., et al. (**2013**). *Water networks contribute to enthalpy/entropy compensation in protein-ligand binding*. J. Am. Chem. Soc. *135*: 15579–15584.

Brenk, R., Schipani, A., James, D., Krasowski, A., Gilbert, I.H., Frearson, J., et al. (**2008**). *Lessons learnt from assembling screening libraries for drug discovery for neglected diseases*. ChemMedChem *3*: 435–444.

Brenke, R., Kozakov, D., Chuang, G.-Y., Beglov, D., Hall, D., Landon, M.R., et al. (**2009**). *Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques*. Bioinformatics *25*: 621–627.

Brock, W.H., Jensen, K. a., Jørgensen, C.K., and Kauffman, G.B. (**1983**). *The origin and dissemination of the term 'ligand' in chemistry*. Polyhedron *2*: 1–7.

Brodney, M. a, Barreiro, G., Ogilvie, K., Hajos-Korcsok, E., Murray, J., Vajdos, F., et al. (**2012**). *Spirocyclic sulfamides as β-secretase 1 (BACE-1) inhibitors for the treatment of Alzheimer's disease: utilization of structure based drug design, WaterMap, and CNS penetration studies to identify centrally efficacious inhibitors*. J. Med. Chem. *55*: 9224–9239.

Brown, A.C., and Fraser, T.R. (**1868**). *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. *2*: 224–242.

Brown, D.G., and Boström, J. (**2015**). *An analysis of past and present synthetic methodologies on medicinal chemistry: Where have all the new reactions gone?*. J. Med. Chem. in press.

Brown, E.N., and Ramaswamy, S. (**2007**). *Quality of protein crystal structures*. Acta Crystallogr. D. Biol. Crystallogr. *63*: 941–950.

Brown, N. (**2012**). Bioisosteres in Medicinal Chemistry, part 1. In Bioisosteres in Medicinal Chemistry, N. Brown, ed. (Wiley-VCH Verlag GmbH & Co. KGaA), pp 1–14.

Brünger, A.T. (**1993**). *Assessment of phase accuracy by cross validation: the free R value. Methods and applications*. Acta Crystallogr. D. Biol. Crystallogr. *49*: 24–36.

Brylinski, M. (**2014**). *eMatchSite: Sequence Order-Independent Structure Alignments of Ligand Binding Pockets in Protein Models*. PLoS Comput. Biol. *10*: e1003829.

Burger, A. (**1991**). *Isosterism and bioisosterism in drug design*. Prog. Drug Res. *37*: 287–371.

Burgoyne, N.J., and Jackson, R.M. (**2006**). *Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces*. Bioinformatics *22*: 1335–1342.

Byvatov, E., Fechner, U., Sadowski, J., and Schneider, G. (**2003**). *Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification*. J. Chem. Inf. Model. *43*: 1882–1889.

Callaway, E. (**2015**). *The revolution will not be crystallized: a new method sweeps through structural biology*. Nature *525*: 172–174.

Cammisa, M., Correra, A., Andreotti, G., and Cubellis, M.V. (**2013**). *Identification and analysis of conserved pockets on protein surfaces.* BMC Bioinformatics *14 Suppl 7*: S9.

Cao, D.-S., Liang, Y., Yan, J., Tan, G., Xu, Q., and Liu, S. (**2013**). *PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies.* J. Chem. Inf. Model. *53*: 3086–3096.

Carbonell, P., and Trosset, J.Y. (**2014**). *Overcoming drug resistance through in silico prediction*. Drug Discov. Today Technol. *11*: 101–107.

Carini, D.J., Duncia, J. V, Aldrich, P.E., Chiu, a T., Johnson, a L., Pierce, M.E., et al. (**1991**). *Nonpeptide angiotensin II receptor antagonists: the discovery of a series of N-(biphenylylmethyl)imidazoles as potent, orally active antihypertensives.* J. Med. Chem. *34*: 2525–2547.

Case, D.A., Berryman, J.T., Betz, R.M., Cerutti, D.S., Cheatham, III, T.E., Darden, T.A., et al. (**2015**). *Amber14*.

Cereto-Massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (**2015**). *Molecular fingerprint similarity search in virtual screening*. Methods *71*: 58–63.

Certara (**2015**). *Unity 2D fingerprints, https://www.certara.com/*.

Chakrabarti, P., and Bhattacharyya, R. (**2007**). *Geometry of nonbonded interactions involving planar groups in proteins*. Prog. Biophys. Mol. Biol. *95*: 83–137.

Chang, G., Roth, C.B.., Reyes, C.L., Pornillos, O., Chen, Y.-J., and Chen, A.P. (**2006**). *Retractation*. Science (80-.). *22*: 1875–1877.

Chapman, H.N., Fromme, P., Barty, A., White, T. a, Kirian, R. a, Aquila, A., et al. (**2011**). *Femtosecond X-ray protein nanocrystallography.* Nature *470*: 73–77.

Chen, H., Duyne, R. Van, Zhang, N., Kashanchi, F., and Zeng, C. (**2009**). *A novel binding pocket of cyclin-dependent kinase 2.* Proteins *74*: 122–132.

Chen, H., Lyne, P.D., Giordanetto, F., Lovell, T., and Li, J. (**2006**). *On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors*. J. Chem. Inf. Model. *46*: 401–415.

Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., et al. (**2007**). *Structure-based maximal affinity model predicts small-molecule druggability.* Nat. Biotechnol. *25*: 71–75.

Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S.H. (**2012**). *Structure-based virtual screening for drug discovery: a problem-centric review.* AAPS J. *14*: 133–41.

Cheng, Y.-C., and Prusoff, W.H. (**1973**). *Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction*. Biochem. Pharmacol. *22*: 3099–3108.

Chupakhin, V., Marcou, G., Gaspar, H., and Varnek, A. (**2014**). *Simple Ligand-Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison*. Comput. Struct. Biotechnol. J. *10*: 33–37.

Clark, R.D., Strizhev, A., Leonard, J.M., Blake, J.F., and Matthew, J.B. (**2002**). *Consensus*

scoring for ligand/protein interactions. J. Mol. Graph. Model. *20*: 281–295.

Cohen, P. (**2000**). *The regulation of protein function by multisite phosphorylation--a 25 year update.* Trends Biochem. Sci. *25*: 596–601.

Coleman, R.G., and Sharp, K. a (**2010**). *Protein pockets: inventory, shape, and comparison.* J. Chem. Inf. Model. *50*: 589–603.

Congreve, M., Carr, R., Murray, C., and Jhoti, H. (**2003**). *A 'Rule of Three' for fragment-based lead discovery?* Drug Discov. Today *8*: 876–877.

Connolly, M.L. (**1983**). *Analytical molecular surface calculation*. J. Appl. Crystallogr. *16*: 548–558.

Cooper, D.R., Porebski, P.J., Chruszcz, M., and Minor, W. (**2011**). *X-ray crystallography: assessment and validation of protein–small molecule complexes for drug discovery*. Expert Opin. Drug Discov. *6*: 771–782.

Cooper, M. a (**2002**). *Optical biosensors in drug discovery.* Nat. Rev. Drug Discov. *1*: 515–528.

Cozzini, P., Fornabaio, M., Marabotti, A., Abraham, D.J., Kellogg, G.E., and Mozzarelli, A. (**2004**). *Free energy of ligand binding to protein: evaluation of the contribution of water molecules by computational methods.* Curr. Med. Chem. *11*: 3093–3118.

Cozzini, P., Kellogg, G.E., Spyrakis, F., Abraham, D.J., Costantino, G., Emerson, A., et al. (**2008**). *Target flexibility: an emerging consideration in drug discovery and design.* J. Med. Chem. *51*: 6237–6255.

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

Crowther, G.J., Booker, M.L., He, M., Li, T., Raverdy, S., Novelli, J.F., et al. (**2014**). *Cofactor-independent phosphoglycerate mutase from nematodes has limited druggability, as revealed by two high-throughput screens.* PLoS Negl. Trop. Dis. *8*: e2628.

Cuchillo, R.R., Pinto-Gil, K., and Michel, J. (**2015**). *A Collective Variable for the Rapid Exploration of Protein Druggability*. J. Chem. Theory Comput. *11*: 1292–1307.

Dahlin, J.L., Inglese, J., and Walters, M. a (**2015**). *Mitigating risk in academic preclinical drug discovery.* Nat. Rev. Drug Discov. *14*: 279–294.

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

Davare, M. a., Vellore, N. a., Wagner, J.P., Eide, C. a., Goodman, J.R., Drilon, A., et al. (**2015**). *Structural insight into selectivity and resistance profiles of ROS1 tyrosine kinase inhibitors*. Proc. Natl. Acad. Sci. *112*: E5381–E5390.

David Blow (**2002**). *Outline of Crystallography for Biologists*.

Davis, A.M., Teague, S.J., and Kleywegt, G.J. (**2003**). *Application and limitations of x-ray crystallographic data in structure-based ligand and drug design*. Angew. Chemie - Int. Ed. *42*: 2718–2736.

Daylight Chemical System Information *Daylight (www.daylight.com)*.

Dearden, J.C., Cronin, M.T.D., and Kaiser, K.L.E. (**2009**). *How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR)*. SAR QSAR Environ. Res. *20*: 241–266.

Debiec, K.T., Gronenborn, A.M., and Chong, L.T. (**2014**). *Evaluating the strength of salt bridges: a comparison of current biomolecular force fields.* J. Phys. Chem. B *118*: 6561–6569.

DeLano, W. (**2002**). *The PyMOL molecular graphics system, http://www.pymol.org/.*

Desaphy, J., Azdimousa, K., Kellenberger, E., and Rognan, D. (**2012**). *Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes.* J. Chem. Inf. Model. *52*: 2287–2299.

Desaphy, J., Bret, G., Rognan, D., and Kellenberger, E. (**2015**). *Sc-PDB: A 3D-database of ligandable binding sites-10 years on*. Nucleic Acids Res. *43*: D399–D404.

Desaphy, J., and Rognan, D. (**2014**). *Sc-PDB-Frag: A database of protein-ligand interaction patterns for bioisosteric replacements*. J. Chem. Inf. Model. *54*: 1908–1918.

Desdouits, N., Nilges, M., and Blondel, A. (**2014**). *Principal Component Analysis Reveals Correlation of Cavities Evolutionand Functional Motions in Proteins*. J. Mol. Graph. Model. *55*: 13–24.

Desiraju, G.R., Ho, P.S., Kloo, L., Legon, A.C., Marquardt, R., Metrangolo, P., et al. (**2013**). *Definition of the halogen bond (IUPAC recommendations 2013).* Pure Appl. Chem. *85*: 1711–1713.

Devereux, M., and Popelier, P.L. a (**2010**). *In silico techniques for the identification of bioisosteric replacements for drug design.* Curr. Top. Med. Chem. *10*: 657–668.

Dias, R., and Azevedo Jr., W. de (**2008**). *Molecular Docking Algorithms*. Curr. Drug Targets *9*: 1040–1047.

Doak, B.C., Zheng, J., Dobritzsch, D., and Kihlberg, J. (**2015**). *How Beyond Rule of 5 Drugs and Clinical Candidates Bind to Their Targets*. J. Med. Chem. in press.

Dobson, C.M. (**2004**). *Chemical space and biology.* Nature *432*: 824–828.

Donald, J.E., Kulp, D.W., and DeGrado, W.F. (**2011**). *Salt bridges: Geometrically specific, designable interactions*. Proteins *79*: 898–915.

Dougherty, D. a. (**2013**). *The cation-π interaction*. Acc. Chem. Res. *46*: 885–893.

Dougherty, D.A. (**1996**). *Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp.* Science *271*: 163–168.

Du, Q.-S., Wang, Q.-Y., Du, L.-Q., Chen, D., and Huang, R.-B. (**2013**). *Theoretical study on the polar hydrogen-π (Hp-π) interactions between protein side chains.* Chem. Cent. J. *7*: 92.

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

Ekins, S., Mestres, J., and Testa, B. (**2007**). *In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling*. Br. J. Pharmacol. *152*: 9–20.

Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G. V, and Mee, R.P. (**1997**). *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes.* J. Comput. Aided. Mol. Des. *11*: 425–445.

Elliott, T.S., Slowey, A., Ye, Y., and Conway, S.J. (**2012**). *The use of phosphate bioisosteres in medicinal chemistry and chemical biology*. Medchemcomm *3*: 735–751.

Ertl, P. (**2007**). *In silico identification of bioisosteric functional groups.* Curr. Opin. Drug Discov. Devel. *10*: 281–288.

Ewing, T.J. a, Makino, S., Skillman, a. G., and Kuntz, I.D. (**2001**). *DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases*. J. Comput. Aided.

Mol. Des. *15*: 411–428.

Eyrisch, S., and Helms, V. (**2007**). *Transient pockets on protein surfaces involved in protein-protein interaction.* J. Med. Chem. *50*: 3457–3464.

Fatumo, S., Adebiyi, M., and Adebiyi, E. (**2013**). *In Silico Models for Drug Discovery* (Totowa, NJ: Humana Press).

Fauman, E.B., Rai, B.K., and Huang, E.S. (**2011**). *Structure-based druggability assessment--identifying suitable targets for small molecule therapeutics.* Curr. Opin. Chem. Biol. *15*: 463–468.

Feher, M. (**2006**). *Consensus scoring for protein-ligand interactions*. Drug Discov. Today *11*: 421–428.

Fenley, A.T., Muddana, H.S., and Gilson, M.K. (**2012**). *Entropy-enthalpy transduction caused by conformational shifts can obscure the forces driving protein-ligand binding.* Proc. Natl. Acad. Sci. U. S. A. *109*: 20006–200011.

Ferrante, A., and Gorski, J. (**2012**). *Enthalpy-entropy compensation and cooperativity as thermodynamic epiphenomena of structural flexibility in ligand-receptor interactions*. J. Mol. Biol. *417*: 454–467.

Ferreira, L., Santos, R. dos, Oliva, G., and Andricopulo, A. (**2015**). *Molecular Docking and Structure-Based Drug Design Strategies*. Molecules *20*: 13384–13421.

Fischer, E. (**1894**). *Einfluss der Configuration auf die Wirkung der Enzyme*. Ber. Dtsch. Chem. Ges. *27*: 2985–2993.

Fisher, R. (**1936**). *The Use of Multiple Measurements in Taxonomic Problems*. Ann. Eugen. *7*: 179–188.

Ford, M.C., and Ho, P.S. (**2015**). *Computational Tools to Model Halogen Bonds in Medicinal Chemistry*. J. Med. Chem. in press.

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

Freire, E. (**2008**). *Do enthalpy and entropy distinguish first in class from best in class?*. Drug Discov. Today *13*: 869–874.

Freire, E. (**2009**). *A thermodynamic approach to the affinity optimization of drug candidates*. Chem. Biol. Drug Des. *74*: 468–472.

Friesner, R. a., Banks, J.L., Murphy, R.B., Halgren, T. a., Klicic, J.J., Mainz, D.T., et al. (**2004**). *Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy*. J. Med. Chem. *47*: 1739–1749.

Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., Greenwood, J.R., Halgren, T.A., et al. (**2006**). *Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes*. J. Med. Chem. *49*: 6177–6196.

Frimurer, T.M., Bywater, R., Naerum, L., Lauritsen, L.N., and Brunak, S. (**2000**). *Improving the odds in discriminating 'drug-like' from 'non drug-like' compounds.* J. Chem. Inf. Comput. Sci. *40*: 1315–1324.

Gallivan, J.P., and Dougherty, D. a. (**2000**). *A computational study of cation-pi interactions vs salt bridges in aqueous media: Implications for protein engineering*. J. Am. Chem. Soc. *122*: 870–874.

Gao, M., and Skolnick, J. (**2013**). *APoc: large-scale identification of similar protein pockets.*

Bioinformatics *29*: 597–604.

Garbett, N.C., and Chaires, J.B. (**2012**). *Thermodynamic studies for drug design and screening.* Expert Opin. Drug Discov. *7*: 299–314.

García-Sosa, A.T., Mancera, R.L., and Dean, P.M. (**2003**). *WaterScore: A novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes*. J. Mol. Model. *9*: 172–182.

Garro Martinez, J.C., Vega-Hissi, E.G., Andrada, M.F., and Estrada, M.R. (**2015**). *QSAR and 3D-QSAR studies applied to compounds with anticonvulsant activity*. Expert Opin. Drug Discov. *10*: 37–51.

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., et al. (**2012**). *ChEMBL: A large-scale bioactivity database for drug discovery*. Nucleic Acids Res. *40*: 1100–1107.

Genheden, S., and Ryde, U. (**2015**). *The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities.* Expert Opin. Drug Discov. *10*: 449–461.

Gertrudes, J.C., Maltarollo, V.G., Silva, R. a, Oliveira, P.R., Honório, K.M., and Silva, a B.F. da (**2012**). *Machine learning techniques and drug design.* Curr. Med. Chem. *19*: 4289–4297.

Geschwindner, S., Ulander, J., and Johansson, P. (**2015**). *Ligand Binding Thermodynamics in Drug Discovery: Still a Hot Tip?*. J. Med. Chem. *58*: 6321–6335.

Ghai, R., Falconer, R.J., and Collins, B.M. (**2012**). *Applications of isothermal titration calorimetry in pure and applied research-survey of the literature from 2010*. J. Mol. Recognit. *25*: 32–52.

Giese, M., Albrecht, M., and Rissanen, K. (**2015**). *Experimental investigation of anion-π interactions - applications and biochemical relevance.* Chem. Commun. (Camb). *52*: 1778–1795.

Gillet, V.J., Khatib, W., Willett, P., Fleming, P.J., and Green, D.V.S. (**2002**). *Identification of biological activity profiles using substructural analysis and genetic algorithms.* J. Chem. Inf. Comput. Sci. *42*: 375–385.

Gillet, V.J., Willett, P., and Bradshaw, J. (**1998**). *Identification of biological activity profiles using substructural analysis and genetic algorithms.* J. Chem. Inf. Comput. Sci. *38*: 165–179.

Gilli, G., and Gilli, P. (**2000**). *Towards an unified hydrogen-bond theory*. J. Mol. Struct. *552*: 1–15.

Gilli, P., Bertolasi, V., Ferretti, V., Gilli, G., and Giui, P. (**1994**). *Covalent Nature of the Strong Homonuclear Hydrogen Bond . Study of the O-H---O System by Crystal Structure Correlation Methods*. J. Am. Chem. Soc. *116*: 909–915.

Gilli, P., Pretto, L., Bertolasi, V., and Gilli, G. (**2009**). *Predicting hydrogen-bond strengths from acid-base molecular properties. The pK(a) slide rule: toward the solution of a long-lasting problem.* Acc. Chem. Res. *42*: 33–44.

Gilson, M.K., and Zhou, H.-X. (**2007**). *Calculation of Protein-Ligand Binding Affinities*. Annu. Rev. Biophys. Biomol. Struct. *36*: 21–42.

Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R. a, and Thornton, J.M. (**2006**). *A method for localizing ligand binding pockets in protein structures.* Proteins *62*: 479–488.

Gohlke, H., Hendlich, M., and Klebe, G. (**2000**). *Knowledge-based scoring function to predict protein-ligand interactions*. J. Mol. Biol. *295*: 337–356.

Golbraikh, A., Wang, X.S., Zhu, H., and Tropsha, A. (**2012**). *Predictive QSAR Modeling:*

Methods and Applications in Drug Discovery andChemical Risk Assessment. Handb. Comput. Chem. 1309–1342.

Gold, N.D., and Jackson, R.M. (**2006**). *Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships*. J. Mol. Biol. *355*: 1112–1124.

Goodford, P.J. (**1985**). *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*. J. Med. Chem. *28*: 849–857.

Gramatica, P. (**2007**). *Principles of QSAR models validation: Internal and external*. QSAR Comb. Sci. *26*: 694–701.

Greener, J.G., and Sternberg, M.J. (**2015**). *AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis.* BMC Bioinformatics *16*: 335.

Grinter, S.Z., and Zou, X. (**2014**). *Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design*. Molecules *19*: 10150–10176.

Grunwald, E., and Steel, C. (**1995**). *Solvent Reorganization and Thermodynamic Enthalpy - Entropy Compensation Ernest*. J. Am. Chem. Soc. *117*: 5687–5692.

Guilloux, V. Le, Schmidtke, P., and Tuffery, P. (**2009**). *Fpocket: an open source platform for ligand pocket detection.* BMC Bioinformatics *10*: 168.

Guo, Q., Wei, Y., Xia, B., Jin, Y., Liu, C., Pan, X., et al. (**2016**). *Identification of a small molecule that simultaneously suppresses virulence and antibiotic resistance of Pseudomonas aeruginosa*. Nat. Publ. Gr. *6*: 1–15.

Gupta, A., Arockia Babu, M., and Kaskhedikar, S. (**2004**). *VALSTAT: validation program for quantitative structure activity relationship studies*. Indian J. Pharm. Sci. *66*: 396–402.

Gupta, A., Gupta, A.K., and Seshadri, K. (**2009**). *Structural models in the assessment of protein druggability based on HTS data.* J. Comput. Mol. Des. *23*: 583–592.

Gvritishvili, A., Gribenko, A., and Makhatadze, G. (**2008**). *Cooperativity of complex salt bridges*. Protein Sci. 1285–1290.

Hajduk, P.J., Huth, J.R., and Fesik, S.W. (**2005**). *Druggability indices for protein targets derived from NMR-based screening data.* J. Med. Chem. *48*: 2518–2525.

Halgren, T. a (**2009**). *Identifying and characterizing binding sites and assessing druggability.* J. Chem. Inf. Model. *49*: 377–389.

Hall, D.R., and Enyedy, I.J. (**2015**). *Computational solvent mapping in structure-based drug design.* Future Med. Chem. *7*: 337–353.

Hall, Z., Hernández, H., Marsh, J. a, Teichmann, S. a, and Robinson, C. V (**2013**). *The role of salt bridges, charge density, and subunit flexibility in determining disassembly routes of protein complexes.* Structure *21*: 1325–1337.

Hamano, J., Torvalds, L., and many contributors (**2005**). *Git (https://git-scm.com/)*.

Hann, M., and Keserü, G. (**2012**). *Finding the sweet spot: the role of nature and nurture in medicinal chemistry*. Nat. Rev. Drug Discov. *11*: 355–365.

Hann, M.M. (**2011**). *Molecular obesity, potency and other addictions in drug discovery*. Medchemcomm *2*: 349–355.

Hann, M.M., and Oprea, T.I. (**2004**). *Pursuing the leadlikeness concept in pharmaceutical research*. Curr. Opin. Chem. Biol. *8*: 255–263.

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

Hartley, H. (**1951**). *Origin of the Word 'Protein'*. Nature *168*: 244.

Hawkins, D.M. (**2004**). *The Problem of Overfitting*. J. Chem. Inf. Comput. Sci. *44*: 1–12.

Hayward, J. (**2012**). BIOSTER: A Database of Bioisosteres and Bioanalogues, part 4. In Bioisosteres in Medicinal Chemistry, N. Brown, ed. (Wiley-VCH Verlag GmbH & Co. KGaA), pp 53–74.

Helmerhorst, E., Chandler, D.J., Nussio, M., and Mamotte, C.D. (**2012**). *Real-time and label-free bio-sensing of molecular interactions by surface plasmon resonance: A laboratory medicine perspective*. Clin. Biochem. Rev. *33*: 161–173.

Hendlich, M., Rippmann, F., and Barnickel, G. (**1997**). *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* J. Mol. Graph. Model. *15*: 359–363.

Hess, M., Schulze Elfringhoff, A., and Lehr, M. (**2007**). *1-(5-Carboxy- and 5-carbamoylindol-1-yl)propan-2-ones as inhibitors of human cytosolic phospholipase A2alpha: bioisosteric replacement of the carboxylic acid and carboxamide moiety.* Bioorg. Med. Chem. *15*: 2883–2891.

Hirschfeld, H.O. (**1935**). *A connection between correlation and contingency*. Proceeding Cambridge Philos. Soc. *31*: 520–524.

Högenauer, K., Hinterding, K., Nussbaumer, P., Hegenauer, K., Hinterding, K., and Nussbaumer, P. (**2010**). *S1P receptor mediated activity of FTY720 phosphate mimics.* Bioorg. Med. Chem. Lett. *20*: 1485–1487.

Homans, S.W. (**2007**). *Water, water everywhere - except where it matters?*. Drug Discov. Today *12*: 534–539.

Hopkins, A.A.L., and Groom, C.R.C. (**2002**). *The druggable genome*. Nat. Rev. Drug Discov. *1*: 727–730.

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

Hu, B., and Lill, M. a (**2014**). *WATsite: hydration site prediction program with PyMOL interface.* J. Comput. Chem. *35*: 1255–1260.

Hu, L., Benson, M.L., Smith, R.D., Lerner, M.G., and Carlson, H. a (**2005**). *Binding MOAD (Mother Of All Databases).* Proteins *60*: 333–340.

Huang, B. (**2009**). *MetaPocket: a meta approach to improve protein ligand binding site prediction.* Omics *13*: 325–330.

Huang, B., and Schroeder, M. (**2006**). *LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.* BMC Struct. Biol. *6*: 19.

Huang, Grinter, S.Z., Zou, X., Huang, S.-Y., Grinter, S.Z., and Zou, X. (**2010**). *Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions.* Phys. Chem. Chem. Phys. *12*: 12899–12908.

Huang, N., and Jacobson, M.P. (**2010**). *Binding-site assessment by virtual fragment screening.* PLoS One *5*: e10109.

Huang, N., Shoichet, B.K., and Irwin, J.J. (**2006**). *Benchmarking Sets for Molecular Docking Benchmarking Sets for Molecular Docking*. J. Med. Chem. *49*: 6789–6801.

Huang, S.Y., and Zou, X. (**2010**). *Advances and challenges in Protein-ligand docking*. Int. J. Mol. Sci. *11*: 3016–3034.

Huang, S.-Y., and Zou, X. (**2006a**). *An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials*. J. Comput. Chem. *27*: 1866–1875.

Huang, S.-Y., and Zou, X. (**2006b**). *An iterative knowledge-based scoring function to predict protein–ligand interactions: II. Validation of the scoring function*. J. Comput. Chem. *27*: 1876–1882.

Hubbard, SJ and Thornton, J. (**1992**). *NACCESS version 2.1.1*.

Huggins, D.J. (**2012**). *Application of inhomogeneous fluid solvation theory to model the distribution and thermodynamics of water molecules around biomolecules.* Phys. Chem. Chem. Phys. *14*: 15106–15117.

Hughes, J.P., Rees, S.S., Kalindjian, S.B., and Philpott, K.L. (**2011**). *Principles of early drug discovery*. Br. J. Pharmacol. *162*: 1239–1249.

Ilari, A., and Savino, C. (**2008**). *Protein Structure Determination by X-Ray Crystallography*. Bioinformatics *452*: 63–87.

Ito, J.I., Tabei, Y., Shimizu, K., Tsuda, K., and Tomii, K. (**2012**). *PoSSuM: A database of similar protein-ligand binding and putative pockets*. Nucleic Acids Res. *40*: 541–548.

IUPAC (**2016**). *IUPAC Gold Book (http://goldbook.iupac.org/)*.

Jaccard, P. (**1901**). *Etude comparative de la distribution florale dans une portion des Alpes et des Jura.* Bull. La Société Vaudoise Des Sci. Nat. *37.*:

Jackson, M.R., Beahm, R., Duvvuru, S., Narasimhan, C., Wu, J., Wang, H., et al. (**2007**). *A preference for edgewise interactions between aromatic rings and carboxylate anions: the biological relevance of anion-quadrupole interactions.* J. Phys. Chem. B *111*: 8242–8249.

Jain, A.N., and Nicholls, A. (**2008**). *Recommendations for evaluation of computational methods*. J. Comput. Aided. Mol. Des. *22*: 133–139.

Jeszenői, N., Bálint, M., Horváth, I., Spoel, D. van der, and Hetényi, C. (**2016**). *Exploration of Interfacial Hydration Networks of Target–Ligand Complexes*. J. Chem. Inf. Model. *56*: 148–158.

Jiang, F., and Kim, S.H. (**1991**). *'Soft docking': Matching of molecular surface cubes*. J. Mol. Biol. *219*: 79–102.

Jin, L., Wang, W., and Fang, G. (**2014**). *Targeting protein-protein interaction by small molecules.* Annu. Rev. Pharmacol. Toxicol. *54*: 435–456.

Johnson, D.K., and Karanicolas, J. (**2013**). *Druggable Protein Interaction Sites Are More Predisposed to Surface Pocket Formation than the Rest of the Protein Surface*. PLoS Comput. Biol. *9*: e1002951.

Jones, G., Willett, P., and Glen, R.C. (**1995**). *Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation.* J. Mol. Biol. *245*: 43–53.

Jones, G., Willett, P., Glen, R.C., Leach, a R., and Taylor, R. (**1997**). *Development and validation of a genetic algorithm for flexible docking.* J. Mol. Biol. *267*: 727–748.

Jones, S., and Thornton, J.M. (**1996**). *Principles of protein-protein interactions.* Proc. Natl. Acad. Sci. U. S. A. *93*: 13–20.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (**1983**). *Comparison of simple potential functions for simulating liquid water*. J. Chem. Phys. *79*: 926–934.

Kabsch, W. (**1976**). *A solution for the best rotation to relate two sets of vectors*. Acta

Crystallogr. Sect. A *32*: 922–923.

Kahraman, A., Morris, R.J., Laskowski, R. a, and Thornton, J.M. (**2007**). *Shape variation in protein binding pockets and their ligands*. J. Mol. Biol. *368*: 283–301.

Kellogg, G.E., and Chen, D.L. (**2004**). *The importance of being exhaustive. Optimization of bridging structural water molecules and water networks in models of biological systems.* Chem. Biodivers. *1*: 98–105.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. (**1958**). *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature *181*: 662–666.

Kennewell, E.A., Willett, P., Ducrot, P., and Luttmann, C. (**2006**). *Identification of target-specific bioisosteric fragments from ligand-protein crystallographic data*. J. Comput. Aided. Mol. Des. *20*: 385–394.

Kerns, E., and Di, L. (**2008**). *Drug-like Properties: Concepts, Structure Design and Methods* (Elsevier Inc.).

Kerns, E.H., and Di, L. (**2003**). *Pharmaceutical profiling in drug discovery*. Drug Discov. Today *8*: 316–323.

Kim, M.O., Feng, X., Feixas, F., Zhu, W., Lindert, S., Bogue, S., et al. (**2015**). *A molecular dynamics investigation of Mycobacterium tuberculosis prenyl synthases: Conformational flexibility and implications for computer-aided drug discovery*. Chem. Biol. Drug Des. *85*: 756–769.

Kirchmair, J., Markt, P., Distinto, S., Wolber, G., and Langer, T. (**2008**). *Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection - What can we learn from earlier mistakes?*. J. Comput. Aided. Mol. Des. *22*: 213–228.

Kitchen, D., Decornez, H., Furr, J., and Bajorath, J. (**2004**). *Docking and scoring in virtual screening for drug discovery: methods and applications*. Nat. Rev. Drug Discov. *3*: 935–949.

Klebe, G. (**2006**). *Virtual ligand screening: strategies, perspectives and limitations*. Drug Discov. Today *11*: 580–594.

Klebe, G. (**2013**). *Drug Design: Methodology, concepts, and mode-of-action* (Berlin, Heidelberg: Springer Heidelberg New York Dordrecht London).

Klebe, G. (**2015**). *Applying thermodynamic profiling in lead finding and optimization*. Nat. Rev. Drug Discov. *14*: 95–110.

Kollman, P. (**1993**). *Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena*. Chem. Rev. *93*: 2395–2417.

Kollman, P. a., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., et al. (**2000**). *Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models*. Acc. Chem. Res. *33*: 889–897.

Koshland, D.E. (**1958**). *Application of a Theory of Enzyme Specificity to Protein Synthesis.* Proc. Natl. Acad. Sci. U. S. A. *44*: 98–104.

Kozakov, D., Hall, D.R., Napoleon, R.L., Yueh, C., Whitty, A., and Vajda, S. (**2015**). *New Frontiers in Druggability*. J. Med. Chem. *58*: 9063–9088.

Krasowski, A., Muthas, D., Sarkar, A., Schmitt, S., and Brenk, R. (**2011**). *DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set.* J. Chem. Inf. Model. *51*: 2829–2842.

Kroemer, R.T. (**2007**). *Structure-based drug design: docking and scoring.* Curr. Protein Pept. Sci. *8*: 312–328.

Krotzky, T., Grunwald, C., Egerland, U., and Klebe, G. (**2015**). *Large-scale mining for similar protein binding pockets: With RAPMAD retrieval on the fly becomes real*. J. Chem. Inf. Model. *55*: 165–179.

Krotzky, T., Rickmeyer, T.T., Fober, T., and Klebe, G. (**2014**). *Extraction of Protein Binding Pockets in Close Neighborhood of Bound Ligands Makes Comparisons Simple due to Inherent Shape Similarity*. J. Chem. Inf. Model. *54*: 3229–3237.

Krovat, E.M., Steindl, T., and Langer, T. (**2005**). *Recent Advances in Docking and Scoring*. Curr. Comput. - Aided Drug Des. *1*: 93–102.

Kuenemann, M. a., Sperandio, O., Labbé, C.M., Lagorce, D., Miteva, M. a., and Villoutreix, B.O. (**2015**). *In silico design of low molecular weight protein–protein interaction inhibitors: Overall concept and recent advances*. Prog. Biophys. Mol. Biol. *119*: 20–32.

Kufareva, I., Ilatovskiy, A. V., and Abagyan, R. (**2012**). *Pocketome: An encyclopedia of small-molecule binding sites in 4D*. Nucleic Acids Res. *40*: 535–540.

Kurup, a. (**2003**). *C-QSAR: A database of 18,000 QSARS and associated biological and physical data*. J. Comput. Aided. Mol. Des. *17*: 187–196.

Kyte, J. (**2003**). *The basis of the hydrophobic effect*. Biophys. Chem. *100*: 193–203.

Kyte, J., and Doolittle, R.F. (**1982**). *A simple method for displaying the hydropathic character of a protein.* J. Mol. Biol. *157*: 105–132.

Labute, P., and Santavy, M. (**2010**). *SiteFinder-Locating Binding Sites in Protein Structures*.

Ladbury, J.E. (**1996**). *Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design.* Chem. Biol. *3*: 973–980.

Ladbury, J.E. (**2010**). *Calorimetry as a tool for understanding biomolecular interactions and an aid to drug design.* Biochem. Soc. Trans. *38*: 888–893.

Lafont, V., Armstrong, A. a, Ohtaka, H., Kiso, Y., Mario Amzel, L., and Freire, E. (**2007**). *Compensating enthalpic and entropic changes hinder binding affinity optimization.* Chem. Biol. Drug Des. *69*: 413–422.

Lamb, A.L., Kappock, T.J., and Silvaggi, N.R. (**2015**). *You are lost without a map: Navigating the sea of protein structures*. Biochim. Biophys. Acta - Proteins Proteomics *1854*: 258–268.

Lambrinidis, G., Vallianatou, T., and Tsantili-Kakoulidou, A. (**2015**). *In vitro, in silico and integrated strategies for the estimation of plasma protein binding. A review*. Adv. Drug Deliv. Rev. *86*: 27–45.

Langmuir, I. (**1919**). *Isomorphism, isosterism and covalence*. J. Am. Chem. Soc. *41*: 1543–1559.

Laskowski, R.A., and Swindells, M.B. (**2011**). *LigPlot + : Multiple Ligand À Protein Interaction Diagrams for Drug Discovery*. J. Chem. Inf. Model *51*: 2778–2786.

Laurie, A.T.R., and Jackson, R.M. (**2005**). *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics *21*: 1908–1916.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., et al. (**2014**). *DrugBank 4.0: Shedding new light on drug metabolism*. Nucleic Acids Res. *42*: 1091–1097.

Lee, C.-W., Wang, H.-J., Hwang, J.-K., and Tseng, C.-P. (**2014**). *Protein Thermal Stability Enhancement by Designing Salt Bridges: A Combined Computational and Experimental Study*. PLoS One *9*: e112751.

Lee, J., and Seok, C. (**2008**). *A statistical rescoring scheme for protein–ligand docking: Consideration of entropic effect*. Proteins *70*: 1074–1083.

Leeson, P.D., and Springthorpe, B. (**2007**). *The influence of drug-like concepts on decision-making in medicinal chemistry*. Nat. Rev. Drug Discov. *6*: 881–890.

Levitt, D.G., and Banaszak, L.J. (**1992**). *POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids*. J. Mol. Graph. *10*: 229–234.

Levitt, M., and Park, B.H. (**1993**). *Water: now you see it, now you don't.* Structure *1*: 223–226.

Lexa, K.W., and Carlson, H.A. (**2012**). *Protein flexibility in docking and surface mapping*. Q. Rev. Biophys. *45*: 301–343.

Li, Q., Bender, A., Pei, J., and Lai, L. (**2007**). *A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification*. J. Chem. Inf. Model. *47*: 1776–1786.

Li, X., and Liang, J. (**2007**). Knowledge-Based Energy Functions for Computational Studies of Proteins. In Computational Methods for Protein Structure Prediction and Modeling, (New York, NY: Springer New York), pp 71–123.

Li, Z., and Lazaridis, T. (**2003**). *Thermodynamic contributions of the ordered water molecule in HIV-1 protease.* J. Am. Chem. Soc *125*: 6636–6637.

Li, Z.R., Lin, H.H., Han, L.Y., Jiang, L., Chen, X., and Chen, Y.Z. (**2006**). *PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence*. Nucleic Acids Res. *34*: W32–W37.

Lichtarge, O., and Sowa, M.E. (**2002**). *Evolutionary predictions of binding surfaces and interactions.* Curr. Opin. Struct. Biol. *12*: 21–27.

Liedberg, B., Nylander, C., and Lunström, I. (**1983**). *Surface plasmon resonance for gas detection and biosensing*. Sensors and Actuators *4*: 299–304.

Lipinski, C. a. (**2000**). *Drug-like properties and the causes of poor solubility and poor permeability*. J. Pharmacol. Toxicol. Methods *44*: 235–249.

Lipinski, C. a. (**2004**). *Lead- and drug-like compounds: the rule-of-five revolution*. Drug Discov. Today Technol. *1*: 337–341.

Lipinski, C. a. (**2005**). *Filtering in drug discovery.* Annu. Reports Comp Chem 155–168.

Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. (**2001**). *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv. Drug Deliv. Rev. *46*: 3–26.

Lise, S., Buchan, D., Pontil, M., and Jones, D.T. (**2011**). *Predictions of hot spot residues at protein-protein interfaces using support vector machines*. PLoS One *6*: e16774.

Liu, H.Y., Grinter, S.Z., and Zou, X. (**2009**). *Multiscale generalized born modeling of ligand binding energies for virtual database screening*. J. Phys. Chem. B *113*: 11793–11799.

Liu, T., and Altman, R.B. (**2014**). *Identifying druggable targets by protein microenvironments matching: application to transcription factors.* CPT Pharmacometrics Syst. Pharmacol. *3*: e93.

Livingstone, C.D., and Barton, G.J. (**1993**). *Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.* Comput. Appl. Biosci. *9*: 745–756.

Lommerse, J.O.S.P.M., Price, S.L., and Taylor, R. (**1997**). *Hydrogen Bonding of Carbonyl, Ether, and Ester Oxygen Atoms with Alkanol Hydroxyl Groups*. J. Comput. Chem. *18*: 757–774.

López, E.D., Arcon, J.P., Gauto, D.F., Petruk, A.A., Modenutti, C.P., Dumas, V.G., et al. (**2015**). *WATCLUST: a tool for improving the design of drugs based on protein-water interactions*. Bioinformatics 1–3.

Lorber, D.M., and Shoichet, B.K. (**1998**). *Flexible ligand docking using conformational ensembles*. Protein Sci. *7*: 938–950.

Lotze, S., and Bakker, H.J. (**2015**). *Structure and dynamics of a salt-bridge model system in water and DMSO*. J. Chem. Phys. *142*: 212436.

Louis, B., Agrawal, V.K., and Khadikar, P. V. (**2010**). *Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses*. Eur. J. Med. Chem. *45*: 4018–4025.

Loving, K. a, Lin, A., and Cheng, A.C. (**2014**). *Structure-based druggability assessment of the mammalian structural proteome with inclusion of light protein flexibility.* PLoS Comput. Biol. *10*: e1003741.

Lu, Y., Wang, R., Yang, C.-Y., and Wang, S. (**2007**). *Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes.* J. Chem. Inf. Model. *47*: 668–675.

Lu, Y., Wang, Y., and Zhu, W. (**2010**). *Nonbonding interactions of organic halogens in biological systems: implications for drug discovery and biomolecular design.* Phys. Chem. Chem. Phys. *12*: 4543–4551.

Lumry, R., and Rajender, S. (**1970**). *Enthalpy-entropy compensation phenomena in water solutions of proteins and small molecules: a ubiquitous property of water.* Biopolymers *9*: 1125–1227.

Machado, J.M., Shiyou, Y., Ho, S.L., and Peihong, N. (**2001**). *A common Tabu search algorithm for the global optimization of engineering problems*. Comput. Methods Appl. Mech. Eng. *190*: 3501–3510.

Mah, R., Thomas, J.R., and Shafer, C.M. (**2014**). *Drug discovery considerations in the development of covalent inhibitors*. Bioorganic Med. Chem. Lett. *24*: 33–39.

Mahoney, M.W., and Jorgensen, W.L. (**2001**). *Quantum, intramolecular flexibility, and polarizability effects on the reproduction of the density anomaly of liquid water by simple potential functions*. J. Chem. Phys. *115*: 10758–10768.

Marcou, G., and Rognan, D. (**2007**). *Optimizing fragment and scaffold docking by use of molecular interaction fingerprints*. J. Chem. Inf. Model. *47*: 195–207.

Martinez, C.R., and Iverson, B.L. (**2012**). *Rethinking the term 'pi-stacking'*. Chem. Sci. *3*: 2191.

Mason Jonatan S., C.D.L. (**2000**). *Library Design and Virtual Screening Using Multibple 4-Point Pharmacophore Fingerprints*. Pacific Symp. Biocomput. 5573-584 *584*: 573–584.

Matters, W.I. (**2010**). *Reproducible Research*. Comput. Sci. Eng. 8–13.

Matthews, R.P., Welton, T., and Hunt, P.A. (**2014**). *Competitive pi interactions and hydrogen bonding within imidazolium ionic liquids.* Phys. Chem. Chem. Phys. *16*: 3238–53.

Mazanetz, M.P., Laughton, C.A., and Fischer, P.M. (**2014**). *Investigation of the flexibility of protein kinases implicated in the pathology of Alzheimer's disease*. Molecules *19*: 9134–9159.

McGaughey, G.B., Gagné, M., and Rappé, A.K. (**1998**). *pi-Stacking Interactions Alive and well in proteins*. J. Biol. Chem. *273*: 15458–15463.

McGregor, M.J., and Muskal, S.M. (**1999a**). *Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design*. J. Chem. Inf. Comput. Sci. *40*: 117–125.

McGregor, M.J., and Muskal, S.M. (**1999b**). *Pharmacophore fingerprinting. 2. Application to primary library design.* J. Chem. Inf. Comput. Sci. *40*: 117–125.

McKinney, J.D., Richard, a, Waller, C., Newman, M.C., and Gerberick, F. (**2000**). *The practice of structure activity relationships (SAR) in toxicology.* Toxicol. Sci. *56*: 8–17.

Meanwell, N.A.N.N. a (**2011**). *Synopsis of some recent tactical application of bioisosteres in drug design.* J. Med. Chem. *54*: 2529–2591.

Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. (**2011**). *Molecular docking: a powerful approach for structure-based drug discovery.* Curr. Comput. Aided. Drug Des. *7*: 146–157.

Merz, K.M., Ringe, D., and Reynolds, C.H. (**2010**). *Drug Design* (Cambridge: Cambridge University Press).

Meuzelaar, H., Tros, M., Huerta-Viga, A., Dijk, C.N. van, Vreede, J., and Woutersen, S. (**2014**). *Solvent-Exposed Salt Bridges Influence the Kinetics of α-Helix Folding and Unfolding.* J. Phys. Chem. Lett. *5*: 900–904.

Mikulskis, P., Genheden, S., and Ryde, U. (**2014**). *A large-scale test of free-energy simulation estimates of protein-Ligand binding affinities*. J. Chem. Inf. Model. *54*: 2794–2806.

Miller, M.D., Kearsley, S.K., Underwood, D.J., and Sheridan, R.P. (**1994**). *FLOG: A system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure*. J. Comput. Aided. Mol. Des. *8*: 153–174.

Miller, R.J.D. (**2014**). *Femtosecond Crystallography with Ultrabight Electron and X-rays: Capturing Chemistry in Action*. Science (80-.). *343*: 1108–1116.

Milletti, F., and Vulpetti, A. (**2010**). *Predicting polypharmacology by binding site similarity: from kinases to the protein universe.* J. Chem. Inf. Model. *50*: 1418–1431.

Morin, a., Urban, J., Adams, P.D., Foster, I., Sali, a., Baker, D., et al. (**2012**). *Shining Light into Black Boxes*. Science (80-.). *336*: 159–160.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., et al. (**1998**). *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function*. J. Comput. Chem. *19*: 1639–1662.

Muegge, I., and Mukherjee, P. (**2015**). *An overview of molecular fingerprint similarity search in virtual screening.* Expert Opin. Drug Discov. *0441*: 1–12.

Mullard, A. (**2016**). *2015 FDA drug approvals*. Nat. Publ. Gr. *15*: 73–76.

Müller, K.R., Rätsch, G., Sonnenburg, S., Mika, S., Grimm, M., and Heinrich, N. (**2005**). *Classifying 'drug-likeness' with kernel-based learning methods*. J. Chem. Inf. Model. *45*: 249–253.

Murray, C.W., Verdonk, M.L., and Rees, D.C. (**2012**). *Experiences in fragment-based drug discovery.* Trends Pharmacol. Sci. *33*: 224–232.

Najmanovich, R., Kuttner, J., Sobolev, V., and Edelman, M. (**2000**). *Side-chain flexibility in proteins upon ligand binding*. Proteins Struct. Funct. Genet. *39*: 261–268.

Nayal, M., and Honig, B. (**2006**). *On the Nature of Cavities on Protein Surfaces : Application to the Identification of Drug-Binding Sites*. Proteins *906*: 892–906.

Needleman, S.B., and Wunsch, C.D. (**1970**). *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. J. Mol. Biol. *48*: 443–453.

Neudert, G., and Klebe, G. (**2011**). *DSX: A Knowledge-Based Sscoring Function for the Assessment of Protein-Ligand Complexes.* J. Chem. Inf. Model. *51*: 2731–2745.

Nicholls, A., Sharp, K.A., and Honig, B. (**1991**). *Protein folding and association: insights from*

the interfacial and thermodynamic properties of hydrocarbons. Proteins *11*: 281–296.

Nicholls, P. (**2000**). *Introduction: the biology of the water molecule.* Cell. Mol. Life Sci. *57*: 987–992.

Nisius, B., Sha, F., and Gohlke, H. (**2011**). *Structure-based computational analysis of protein binding sites for function and druggability prediction.* J. Biotechnol. *159*: 123–134.

Nobeli, I., Price, S.L., Lommerse, J.P.M., and Taylor, R. (**1997**). *Hydrogen Bonding Properties of Oxygen and Nitrogen Acceptors in Aromatic Heterocycles*. J Comput Chem *18*: 2060–2074.

O'Boyle, N.M., Banck, M., James, C. a., Morley, C., Vandermeersch, T., and Hutchison, G.R. (**2011**). *Open Babel: An open chemical toolbox.* J. Cheminform. *3*: 33.

Olaru, A., Bala, C., Jaffrezic-Renault, N., and Aboul-Enein, H.Y. (**2015**). *Surface plasmon resonance (SPR) biosensors in pharmaceutical analysis.* Crit. Rev. Anal. Chem. / CRC *45*: 97–105.

Olsson, T.S.G., Ladbury, J.E., Pitt, W.R., and Williams, M. a. (**2011**). *Extent of enthalpy-entropy compensation in protein-ligand interactions*. Protein Sci. *20*: 1607–1618.

Olsson, T.S.G., Williams, M.A., Pitt, W.R., and Ladbury, J.E. (**2008**). *The Thermodynamics of Protein-Ligand Interaction and Solvation: Insights for Ligand Design*. J. Mol. Biol. *384*: 1002–1017.

Oprea, T.I., and Tropsha, A. (**2006**). *Target, chemical and bioactivity databases - integration is key*. Drug Discov. Today Technol. *3*: 357–365.

Pan, A., Biswas, T., Rakshit, A.K., and Moulik, S.P. (**2015**). *Enthalpy-Entropy Compensation (EEC) Effect: A Revisit*. J. Phys. Chem. B *119*: 15876–15884.

Panigrahi, S.K., and Desiraju, G.R. (**2007**). *Strong and Weak Hydrogen Bonds in the Protein – Ligand Interface*. PROTEINS Struct. Funct. Bioinforma. *67*: 128–141.

Pearlstein, R. a, Hu, Q.-Y., Zhou, J., Yowe, D., Levell, J., Dale, B., et al. (**2010**). *New hypotheses about the structure-function of proprotein convertase subtilisin/kexin type 9: analysis of the epidermal growth factor-like repeat A docking site using WaterMap.* Proteins *78*: 2571–2586.

Pearlstein, R. a., Sherman, W., and Abel, R. (**2013**). *Contributions of water transfer energy to protein-ligand association and dissociation barriers: Watermap analysis of a series of p38 MAP kinase inhibitors*. Proteins Struct. Funct. Bioinforma. *81*: 1509–1526.

Pearson, W.R. (**2013**). An Introduction to Sequence Similarity ('Homology') Searching. In Current Protocols in Bioinformatics, (Hoboken, NJ, USA: John Wiley & Sons, Inc.), pp 1286 – 1292.

Perola, E., Herman, L., and Weiss, J. (**2012**). *Development of a Rule-Based Method for the Assessment of Protein Druggability.* J. Chem. Inf. Model. *52*: 1027–1038.

Perola, E., Walters, W.P., and Charifson, P. (**2007**). *Comments on the article 'on evaluating molecular-docking methods for pose prediction and enrichment factors'.* J. Chem. Inf. Model. *47*: 251–253.

Perola, E., Walters, W.P., and Charifson, P.S. (**2004**). *A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance*. Proteins Struct. Funct. Bioinforma. *56*: 235–249.

Pérot, S., Regad, L., Reynès, C., Spérandio, O., Miteva, M. a., Villoutreix, B.O., et al. (**2013**). *Insights into an Original Pocket-Ligand Pair Classification: A Promising Tool for Ligand Profile Prediction*. PLoS One *8*: e63730.

Pérot, S., Sperandio, O., Miteva, M.A., Camproux, A.-C., and Villoutreix, B.O. (**2010**). *Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery*. Drug Discov. Today *15*: 656–667.

Perozzo, R., Folkers, G., and Scapozza, L. (**2004**). *Thermodynamics of protein-ligand interactions: history, presence, and future aspects*. J. Recept. Signal Transduct. Res. *24*: 1–52.

Peters, W.B., Frasca, V., and Brown, R.K. (**2009**). *Recent developments in isothermal titration calorimetry label free screening*. Comb. Chem. High Throughput Screen. *12*: 772–790.

Petersen, G.O., Saxena, S., Renuka, J., Soni, V., Yogeeswari, P., Santos, D.S., et al. (**2015**). *Structure-based virtual screening as a tool for the identification of novel inhibitors against Mycobacterium tuberculosis 3-dehydroquinate dehydratase*. J. Mol. Graph. Model. *60*: 124–131.

Petersen, R.C. (**1964**). *The linear relationship between enthalpy and entropy of activation*. J. Org. Chem. *29*: 3133–3135.

Petitjean, M. (**1992**). *Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds*. J. Chem. Inf. Model. *32*: 331–337.

Petitjean, M. (**2014**). *RADI version 4.0, http://petitjeanmichel.free.fr/*.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., et al. (**2004**). *UCSF Chimera--a visualization system for exploratory research and analysis*. J. Comput. Chem. *25*: 1605–1612.

Pordea, A. (**2015**). *Metal-binding promiscuity in artificial metalloenzyme design*. Curr. Opin. Chem. Biol. *25*: 124–132.

Potashman, M.H., and Duggan, M.E. (**2009**). *Covalent modifiers: An orthogonal approach to drug design*. J. Med. Chem. *52*: 1231–1246.

Pourbasheer, E., Riahi, S., Ganjali, M.R., and Norouzi, P. (**2010**). *QSAR study on melanocortin-4 receptors by support vector machine*. Eur. J. Med. Chem. *45*: 1087–1093.

Proudfoot, J.R. (**2005**). *The evolution of synthetic oral drug properties*. Bioorganic Med. Chem. Lett. *15*: 1087–1090.

PubChem (**2015**). *PubChem Substructure Fingerprints: ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt*.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (**2002**). *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues*. Bioinformatics *18 Suppl 1*: S71–S77.

Puzyn, T., Leszczyński, J., and Cronin, M. (**2010**). *Recent Advances in QSAR Studies: Methods and Applications*.

Qin, Y., Deng, H., Yan, H., and Zhong, R. (**2011**). *An accurate nonlinear QSAR model for the antitumor activities of chloroethylnitrosoureas using neural networks*. J. Mol. Graph. Model. *29*: 826–833.

Quiñonero, D., Garau, C., Rotger, C., Frontera, A., Ballester, P., Costa, A., et al. (**2002**). *Anion-π interactions: Do they exist?*. Angew. Chemie - Int. Ed. *41*: 3389–3392.

Radifar, M., Yuniarti, N., and Istyastono, E.P. (**2013**). *PyPLIF: Python-based Protein-Ligand Interaction Fingerprinting*. Bioinformation *9*: 325–328.

Ramasubbu, N., Parthasarathy, R., and Murray-Rust, P. (**1986**). *Angular preferences of intermolecular forces around halogen centers: preferred directions of approach of*

electrophiles and nucleophiles around carbon-halogen bond. J. Am. Chem. Soc. *108*: 4308–4314.

Rao, H.B., Zhu, F., Yang, G.B., Li, Z.R., and Chen, Y.Z. (**2011**). *Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence*. Nucleic Acids Res. *39*: 385–390.

Raub, S., Steffen, A., Kämper, A., and Marian, C.M. (**2008**). *AIScore - Chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes*. J. Chem. Inf. Model. *48*: 1492–1510.

Ravindranath, P.A., Forli, S., Goodsell, D.S., Olson, A.J., and Sanner, M.F. (**2015**). *AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility*. PLoS Comput. Biol. *11*: 1–28.

Rebilly, J.-N., Colasson, B., Bistri, O., Over, D., and Reinaud, O. (**2015**). *Biomimetic cavity-based metal complexes*. Chem. Soc. Rev. Chem. Soc. Rev *44*: 467–489.

Reynolds, C.H., and Holloway, M.K. (**2011**). *Thermodynamics of ligand binding and efficiency*. ACS Med. Chem. Lett. *2*: 433–437.

Reynolds, C.H., Tounge, B. a, and Bembenek, S.D. (**2008**). *Ligand binding efficiency: trends, physical basis, and implications*. J. Med. Chem. *51*: 2432–2438.

Rice, P., Longden, I., and Bleasby, A. (**2000**). *EMBOSS: The European Molecular Biology Open Software Suite*. Trends Genet. *16*: 276–277.

Robertson, J.G. (**2005**). *Current Topics / Perspectives Mechanistic Basis of Enzyme-Targeted Drugs Current Topics Mechanistic Basis of Enzyme-Targeted Drugs*. Society *44*: 5561–5571.

Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., and Honig, B. (**2002**). *Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects*. J. Comput. Chem. *23*: 128–137.

Rogers, D., and Hahn, M. (**2010**). *Extended-connectivity fingerprints*. J. Chem. Inf. Model. *50*: 742–754.

Roncaglioni, A., Toropov, A.A., Toropova, A.P., and Benfenati, E. (**2013**). *In silico methods to predict drug toxicity*. Curr. Opin. Pharmacol. *13*: 802–806.

Roselin, L.S., Lin, M.S., Lin, P.H., Chang, Y., and Chen, W.Y. (**2010**). *Recent trends and some applications of isothermal titration calorimetry in biotechnology*. Biotechnol. J. *5*: 85–98.

Rossato, G., Ernst, B., Vedani, A., and Smiesko, M. (**2011**). *AcquaAlta: a directional approach to the solvation of ligand-protein complexes*. J. Chem. Inf. Model. *51*: 1867–1881.

Ruusmann, V., Sild, S., and Maran, U. (**2014**). *QSAR DataBank - An approach for the digital organization and archiving of QSAR model information*. J. Cheminform. *6*: 1–17.

Ruusmann, V., Sild, S., and Maran, U. (**2015**). *QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models*. J. Cheminform. *7*: 32.

Rye, C.S., and Baell, J.B. (**2005**). *Phosphate isosteres in medicinal chemistry*. Curr. Med. Chem. *12*: 3127–3141.

Sabarinathan, R., Aishwarya, K., Sarani, R., Vaishnavi, M.K., and Sekar, K. (**2011**). *Water-mediated ionic interactions in protein structures*. J. Biosci. *36*: 253–263.

Sadowski, J., and Kubinyi, H. (**1998**). *A scoring scheme for discriminating between drugs and nondrugs*. J. Med. Chem. *41*: 3325–3329.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., and Todeschini, R. (**2012**).

Comparison of different approaches to define the applicability domain of QSAR models. Molecules *17*: 4791–4810.

Salentin, S., Schreiber, S., Haupt, V.J., Adasme, M.F., and Schroeder, M. (**2015**). *PLIP: fully automated protein-ligand interaction profiler*. Nucleic Acids Res. *43*: W443–W447.

Sanschagrin, P.C., and Kuhn, L. a (**1998**). *Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity.* Protein Sci. *7*: 2054–2064.

Santos, L.H., Ferreira, R.S., and Caffarena, E.R. (**2015**). *Computational drug design strategies applied to the modelling of human immunodeficiency virus-1 reverse transcriptase inhibitors*. Mem. Inst. Oswaldo Cruz *110*: 847–864.

Sarakatsannis, J.N., and Duan, Y. (**2005**). *Statistical characterization of salt bridges in proteins.* Proteins *60*: 732–9.

Sarkar, A., and Brenk, R. (**2015**). *To Hit or Not to Hit, That Is the Question – Genome-wide Structure-Based Druggability Predictions for Pseudomonas aeruginosa Proteins*. PLoS One *10*: e0137279.

Sarkhel, S., and Desiraju, G.R. (**2004**). *NOH…O, OOH…O, and COH…OHydrogen Bonds in Protein–Ligand Complexes: Strong and Weak Interactions in Molecular Recognition*. *259*: 247–259.

Schalon, C., Surgand, J.-S., Kellenberger, E., and Rognan, D. (**2008**). *A simple and fuzzy method to align and compare druggable ligand-binding sites.* Proteins *71*: 1755–1778.

Schmidtke, P., and Barril, X. (**2010**). *Understanding and predicting druggability. A high-throughput method for detection of drug binding sites.* J. Med. Chem. *53*: 5858–5867.

Schmidtke, P., Guilloux, V. Le, Maupetit, J., and Tufféry, P. (**2010a**). *Fpocket: Online Tools for Protein Ensemble Pocket Detection and Tracking.* Nucleic Acids Res. *38*: W582–W589.

Schmidtke, P., Souaille, C., Estienne, F., Baurin, N., and Kroemer, R.T. (**2010b**). *Large-scale comparison of four binding site detection algorithms.* J. Chem. Inf. Model. *50*: 2191–200.

Schneider, G., and Fechner, U. (**2005**). *Computer-based de novo design of drug-like molecules*. Nat. Rev. Drug Discov. *4*: 649–663.

Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (**1999**). *'Scaffold-Hopping' by topological pharmacophore search: A contribution to virtual screening*. Angew. Chemie - Int. Ed. *38*: 2894–2896.

Scholfield, M.R., Ford, M.C., Zanden, C.M. Vander, Billman, M.M., Ho, P.S., and Rappé, A.K. (**2015**). *Force Field Model of Periodic Trends in Biomolecular Halogen Bonds*. J. Phys. Chem. B *119*: 9140–9149.

Schuck, P. (**1997**). *Use of surface plasmon resonance to probe the equilibrium and dynamic aspects of interactions between biological macromolecules.* Annu. Rev. Biophys. Biomol. Struct. *26*: 541–566.

Seco, J., Luque, F.J., and Barril, X. (**2009**). *Binding site detection and druggability index from first principles.* J. Med. Chem. *52*: 2363–2371.

Sharp, K. (**2001**). *Entropy – enthalpy compensation : Fact or artifact ?*. Protein Sci. *10*: 661–667.

Sheridan, R.P., Maiorov, V.N., Holloway, M.K., Cornell, W.D., and Gao, Y.-D. (**2010**). *Drug-like density: a method of quantifying the 'bindability' of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank.* J. Chem. Inf. Model. *50*: 2029–2040.

Sheridan, R.P., Miller, M.D., Underwood, D.J., and Kearsley, S.K. (**1996**). *Chemical Similarity Using Geometric Atom Pair Descriptors*. J. Chem. Inf. Model. *36*: 128–136.

Shin, W.H., Bures, M.G., and Kihara, D. (**2016**). *PatchSurfers: Two methods for local molecular property-based binding ligand prediction*. Methods *93*: 41–50.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H.J. (**2005**). *SiteEngines: Recognition and comparison of binding sites and protein-protein interfaces*. Nucleic Acids Res. *33*: W337–W341.

Sisay, M.T., Peltason, L., and Bajorath, J. (**2009**). *Structural interpretation of activity cliffs revealed by systematic analysis of structure-activity relationships in analog series*. J. Chem. Inf. Model. *49*: 2179–2189.

Skolnick, J., and Brylinski, M. (**2009**). *FINDSITE: A combined evolution/structure-based approach to protein function prediction*. Brief. Bioinform. *10*: 378–391.

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E.W. (**2014**). *Computational methods in drug discovery*. Pharmacol. Rev. *66*: 334–395.

Smith, F.W., Mudge, S.R., Rae, A.L., and Glassop, D. (**2003**). *Phosphate transport in plants*. Plant Soil *248*: 71–83.

Sorich, M.J., Miners, J.O., McKinnon, R.A., Winkler, D.A., Burden, F.R., and Smith, P.A. (**2003**). *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. *43*: 2019–2024.

Sotriffer, C.A., Sanschagrin, P., Matter, H., and Klebe, G. (**2008**). *SFCscore: Scoring functions for affinity prediction of protein-ligand complexes*. Proteins Struct. Funct. Genet. *73*: 395–419.

Sousa, S.F., Fernandes, P.A., and Ramos, M.J. (**2006**). *Protein-ligand docking: Current status and future challenges*. Proteins Struct. Funct. Bioinforma. *65*: 15–26.

Sousa, S.F., Ribeiro, a J.M., Coimbra, J.T.S., Neves, R.P.P., Martins, S. a, Moorthy, N.S.H.N., et al. (**2013**). *Protein-ligand docking in the new millennium--a retrospective of 10 years in the field.* Curr. Med. Chem. *20*: 2296–2314.

Southall, N.T., Dill, K.A., and Haymet, a D.J. (**2002**). *A View of the Hydrophobic Effect*. J. Phys. Chem. *106*: 521–533.

Spitzer, R., Cleves, A.E., Varela, R., and Jain, A.N. (**2014**). *Protein function annotation by local binding site surface similarity*. Proteins Struct. Funct. Bioinforma. *82*: 679–694.

Starikov, E.B. (**2013**). *Valid entropy-enthalpy compensation: Fine mechanisms at microscopic level*. Chem. Phys. Lett. *564*: 88–92.

Starikov, E.B., and Nordén, B. (**2007**). *Enthalpy-entropy compensation: A phantom or something useful?*. J. Phys. Chem. B *111*: 14431–14435.

Steinbrecher, T., and Labahn, A. (**2010**). *Towards accurate free energy calculations in ligand protein-binding studies.* Curr. Med. Chem. *17*: 767–785.

Sugaya, N., and Furuya, T. (**2011**). *Dr. PIAS: an integrative system for assessing the druggability of protein-protein interactions.* BMC Bioinformatics *12*: 50.

Surade, S., and Blundell, T.L. (**2012**). *Structural biology and drug discovery of difficult targets: the limits of ligandability.* Chem. Biol. *19*: 42–50.

Swanson, J.M.J., Henchman, R.H., and McCammon, J.A. (**2004**). *Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy*. Biophys. J. *86*: 67–74.

Takaoka, Y., Endo, Y., Yamanobe, S., Kakinuma, H., Okubo, T., Shimazaki, Y., et al. (**2003**). *Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition*. J. Chem. Inf. Comput. Sci. *43*: 1269–1275.

Tanford, C. (**1979**). *Interfacial free energy and the hydrophobic effect*. Proc. Natl. Acad. Sci. U. S. A. *76*: 4175–4176.

Tari, L.W. (**2012**). *Structure-Based Drug Discovery* (Totowa, NJ: Humana Press).

Taylor, R., Kennard, O., and Versichel, W. (**1983**). *Geometry of the imino-carbonyl (N-H...O:C) hydrogen bond. 1. Lone-pair directionality*. J. Am. Chem. Soc. *105*: 5761–5766.

Team R Core (R Foundation for Statistical Computing) (**2015**). *R: A Language and Environment for Satistical Computing*.

Teixeira, A.L., and Falcao, A.O. (**2013**). *Noncontiguous Atom Matching Structural Similarity Function*. J. Chem. Inf. Model. *53*: 2511–2524.

Tian, S., Wang, J., Li, Y., Li, D., Xu, L., and Hou, T. (**2015**). *The application of in silico drug-likeness predictions in pharmaceutical research*. Adv. Drug Deliv. Rev. *86*: 2–10.

Tian, S., Wang, J., Li, Y., Xu, X., and Hou, T. (**2012**). *Drug-likeness analysis of traditional chinese medicines: Prediction of drug-likeness using machine learning approaches*. Mol. Pharm. *9*: 2875–2886.

Todeschini, R., and Consonni, V. (**2010**). *Handbook of Molecular Descriptors* (Wiley-VCH Verlag GmbH & Co. KGaA).

Tompa, P. (**2003**). *Intrinsically unstructured proteins evolve by repeat expansion*. BioEssays *25*: 847–855.

Topliss, J.G. (**1977**). *A manual method for applying the Hansch approach to drug design.* J. Med. Chem. *20*: 463–469.

Torgerson, W.S. (**1958**). *Theory & Methods of Scaling*.

Torres, F.E., Recht, M.I., Coyle, J.E., Bruce, R.H., and Williams, G. (**2010**). *Higher throughput calorimetry: opportunities, approaches and challenges*. Curr. Opin. Struct. Biol. *20*: 598–605.

Totrov, M., and Abagyan, R. (**2008**). *Flexible ligand docking to multiple receptor conformations: a practical alternative*. *18*: 178–184.

Tropsha, A. (**2010**). *Best practices for QSAR model development, validation, and exploitation*. Mol. Inform. *29*: 476–488.

Tropsha, A., Gramatica, P., and Gombar, V.K. (**2003**). *The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models*. Qsar Comb. Sci. *22*: 69–77.

Tseng, Y.Y., Dundas, J., and Liang, J. (**2009**). *Predicting Protein Function and Binding Profile via Matching of Local Evolutionary and Geometric Surface Patterns*. J. Mol. Biol. *387*: 451–464.

Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M., and Tanabe, K. (**2002**). *Origin of attraction and directionality of the pi-pi interaction: Model chemistry calculations of benzene dimer interaction*. J. Am. Chem. Soc. *124*: 104–112.

Ujváry, I. (**1997**). *BIOSTER-a database of structurally analogous compounds*. Pestic. Sci. *51*: 92–95.

Ullmann, J.R. (**1976**). *An Algorithm for Subgraph Isomorphism*. J. ACM *23*: 31–42.

Unzue, A., Lafleur, K., Zhao, H., Zhou, T., Dong, J., Kolb, P., et al. (**2016**). *Three stories on Eph kinase inhibitors: From in silico discovery to in vivo validation*. Eur. J. Med. Chem. *112*: 347–366.

Ursu, O., Rayan, A., Goldblum, A., and Oprea, T.I. (**2011**). *Understanding drug-likeness*. Wiley Interdiscip. Rev. Comput. Mol. Sci. *1*: 760–781.

Vainio, M.J., Puranen, J.S., and Johnson, M.S. (**2009**). *ShaEP: molecular overlay based on shape and electrostatic potential.* J. Chem. Inf. Model. *49*: 492–502.

Varnek, A., and Baskin, I. (**2012**). *Machine learning methods for property prediction in chemoinformatics: Quo Vadis?*. J. Chem. Inf. Model. *52*: 1413–1437.

Veber, D.F., Johnson, S.R., Cheng, H., Smith, B.R., Ward, K.W., and Kopple, K.D. (**2002**). *Molecular Properties That Influence the Oral Bioavailability of Drug Candidates*. J. Med. Chem. *45*: 2615–2623.

Venkatraman, V., Chakravarthy, P.R., and Kihara, D. (**2009**). *Application of 3D zernike descriptors to shape-based ligand similarity searching*. J. Cheminform. *1*: 1–20.

Verma, J., Khedkar, V.M., and Coutinho, E.C. (**2010**). *3D-QSAR in drug design--a review*. Curr Top Med Chem *10*: 95–115.

Vieth, M., Siegel, M.G., Higgs, R.E., Watson, I. a, Robertson, D.H., Savin, K. a, et al. (**2004**). *Characteristic physical properties and structural fragments of marketed oral drugs.* J. Med. Chem. *47*: 224–232.

Villar, H.O., and Kauvar, L.M. (**1994**). *Amino acid preferences at protein binding sites.* FEBS Lett. *349*: 125–130.

Volkamer, A., Eid, S., Turk, S., Jaeger, S., Rippmann, F., and Fulle, S. (**2015**). *Pocketome of Human Kinases: Prioritizing the ATP Binding Sites of (Yet) Untapped Protein Kinases for Drug Discovery.* J. Chem. Inf. Model. *55*: 538–549.

Volkamer, A., Griewel, A., Grombacher, T., and Rarey, M. (**2010**). *Analyzing the topology of active sites: on the prediction of pockets and subpockets.* J. Chem. Inf. Model. *50*: 2041–2052.

Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F., and Rarey, M. (**2012a**). *Combining global and local measures for structure-based druggability predictions.* J. Chem. Inf. Model. *52*: 360–372.

Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (**2012b**). *DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment*. Bioinformatics *28*: 2074–2075.

Wallace, a C., Laskowski, R. a, and Thornton, J.M. (**1995**). *Ligplot - a Program To Generate Schematic Diagrams of Protein Ligand Interactions*. Protein Eng. *8*: 127–134.

Walters, W.P. (**2013**). *Modeling, informatics, and the quest for reproducibility*. J. Chem. Inf. Model. *53*: 1529–1530.

Wang, R., Fang, X., Lu, Y., and Wang, S. (**2004**). *The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.* J. Med. Chem. *47*: 2977–2980.

Wang, R., Fang, X., Lu, Y., Yang, C.-Y., and Wang, S. (**2005**). *The PDBbind database: methodologies and updates.* J. Med. Chem. *48*: 4111–4119.

Wang, R., Lai, L., and Wang, S. (**2002**). *Further development and validation of empirical scoring functions for structure-based binding affinity prediction*. J. Comput. Aided. Mol. Des. *16*: 11–26.

Wang, R., Lu, Y., and Wang, S. (**2003**). *Comparative evaluation of 11 scoring functions for molecular docking*. J. Med. Chem. *46*: 2287–2303.

Wang, T., Wu, M.-B., Lin, J.-P., and Yang, L.-R. (**2015**). *Quantitative structure–activity relationship: promising advances in drug discovery platforms*. Expert Opin. Drug Discov. *0441*: 1–18.

Wang, Y., Backman, T.W.H., Horan, K., and Girke, T. (**2013**). *FmcsR: Mismatch tolerant maximum common substructure searching in R*. Bioinformatics *29*: 2792–2794.

Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., and Bryant, S.H. (**2009**). *PubChem: A public information system for analyzing bioactivities of small molecules*. Nucleic Acids Res. *37*: 623–633.

Wassermann, A.M., and Bajorath, J. (**2011**). *Identification of target family directed bioisosteric replacements*. Medchemcomm *2*: 601.

Weill, N., and Rognan, D. (**2010**). *Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites.* J. Chem. Inf. Model. *50*: 123–135.

Weininger, D. (**1988**). *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*. J. Chem. Inf. Comput. Sci. *28*: 31–36.

Weisel, M., Proschak, E., Kriegl, J.M., and Schneider, G. (**2009**). *Form follows function: shape analysis of protein cavities for receptor-based drug design.* Proteomics *9*: 451–459.

Wenlock, M.C., Austin, R.P., Barton, P., Davis, A.M., and Leeson, P.D. (**2003**). *A comparison of physiochemical property profiles of development and marketed oral drugs*. J. Med. Chem. *46*: 1250–1256.

Wermuth, C.G. (**2006**). *Similarity in drugs: reflections on analogue design*. Drug Discov. Today *11*: 348–354.

Whitesides, G.M., and Krishnamurthy, V.M. (**2005**). *Designing ligands to bind proteins*. Q. Rev. Biophys. *38*: 385.

Williams, M.A., and Ladbury, J.E. (**2003**). Hydrogen Bonds in Protein-Ligand Complexes. In Protein-Ligand Interactions: From Molecular Recognition to Drug Design, H. Bohm, and G. Schneider, eds. (Wiley-VCH Verlag GmbH & Co. KGaA), pp 137–161.

Williams, S.P., Kuyper, L.F., and Pearce, K.H. (**2005**). *Recent applications of protein crystallography and structure-guided drug design*. Curr. Opin. Chem. Biol. *9*: 371–380.

Wirth, M., Zoete, V., Michielin, O., and Sauer, W.H.B. (**2013**). *SwissBioisostere: A database of molecular replacements for ligand design*. Nucleic Acids Res. *41*: 1137–1143.

Wood, D.J., Vlieg, J. De, Wagener, M., and Ritschel, T. (**2012**). *Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement*. J. Chem. Inf. Model. *52*: 2031–2043.

Xie, Z.-R., and Hwang, M.-J. (**2012**). *Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles.* Bioinformatics *28*: 1579–1585.

Xu, W., Lucke, A.J., and Fairlie, D.P. (**2015**). *Comparing sixteen scoring functions for predicting biological activities of ligands for protein targets*. J. Mol. Graph. Model. *57*: 76–88.

Xu, Z., Yang, Z., Liu, Y., Lu, Y., Chen, K., and Zhu, W. (**2014**). *Halogen bond: Its role beyond drug-target binding affinity for drug discovery and development*. J. Chem. Inf. Model. *54*: 69–78.

Yadav, S.P., Bergqvist, S., Doyle, M.L., Neubert, T.A., and Yamniuk, A.P. (**2012**). *MIRG*

survey 2011: Snapshot of rapidly evolving label-free technologies used for characterizing molecular interactions. J. Biomol. Tech. *23*: 94–100.

Young, D., Martin, T., Venkatapathy, R., and Harten, P. (**2008**). *Are the chemical structures in your QSAR correct?*. QSAR Comb. Sci. *27*: 1337–1345.

Yuan, Y., Pei, J., and Lai, L. (**2013**). *Binding site detection and druggability prediction of protein targets for structure-based drug design.* Curr. Pharm. Des. *19*: 2326–2333.

Yugang, Z., Huaiqing, C., and Zhirong, L. (**2011**). *Binding Cavities and Druggability of Intrinsically Disordered Proteins*. Protein Sci. 1–50.

Yusof, I., and Segall, M.D. (**2013**). *Considering the impact drug-like properties have on the chance of success*. Drug Discov. Today *18*: 659–666.

Zernov, V. V., Balakin, K. V., Ivaschenko, A.A., Savchuk, N.P., and Pletnev, I. V. (**2003**). *Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions*. J. Chem. Inf. Comput. Sci. *43*: 2048–2056.

Zhang, B., Vogt, M., Maggiora, G.M., and Bajorath, J. (**2015**). *Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures*. J. Comput. Aided. Mol. Des. *29*: 937–950.

Zhang, L., Tsai, K.-C., Du, L., Fang, H., Li, M., and Xu, W. (**2011**). *How to generate reliable and predictive CoMFA models.* Curr. Med. Chem. *18*: 923–930.

Zhang, Y. (**2005**). *TM-align: a protein structure alignment algorithm based on the TM-score*. Nucleic Acids Res. *33*: 2302–2309.

Zhou, W., and Yan, H. (**2012**). *Alpha shape and Delaunay triangulation in studies of protein-related interactions.* Brief. Bioinform. *15*: 54–64.

Zvelebil, M.J., Barton, G.J., Taylor, W.R., and Sternberg, M.J.E. (**1987**). *Prediction of protein secondary structure and active sites using the alignment of homologous sequences*. J. Mol. Biol. *195*: 957–961.

Recent Publications in this Series

17/2016 Mpindi John Patrick
Bioinformatic Tools for Analysis, Mining and Modelling Large-Scale Gene Expression and Drug Testing Datasets
18/2016 Hilla Sumanen
Work Disability among Young Employees Changes over Time and Socioeconomic Differences
19/2016 Oyediran Olulana Akinrinade
Bioinformatic and Genomic Approaches to Study Cardiovascular Diseases
20/2016 Prasanna Sakha
Development of Microfluidic Applications to Study the Role of Kainate Receptors in Synaptogenesis
21/2016 Neha Shrestha
Mesoporous Silicon Systems for Oral Protein/Peptide-Based Diabetes Mellitus Therapy
22/2016 Tanja Holopainen
Targeting Endothelial Tyrosine Kinase Pathways in Tumor Growth and Metastasis
23/2016 Jussi Leppilahti
Variability of Gingival Crevicular Fluid Matrix Metalloproteinase -8 Levels in Respect to Point-of-Care Diagnostics in Periodontal Diseases
24/2016 Niina Markkula
Prevalence, Predictors and Prognosis of Depressive Disorders in the General Population
25/2016 Katri Kallio
The Roles of Template RNA and Replication Proteins in the Formation of Semliki Forest Virus Replication Spherules
26/2015 Hanna Paatela
Role of Dehydroepiandrosterone in High-Density Lipoprotein-Mediated Vasodilation and in Adipose Tissue Steroid Biosynthesis
27/2016 Johanna Mäkelä
Neuroprotective Effects of PGC-1α Activators in Dopaminergic Neurons
28/2016 Sandra Söderholm
Phosphoproteomic Characterization of Viral Infection
29/2016 Mariann Lassenius
Bacterial Endotoxins in Type 1 Diabetes
30/2016 Mette Ilander
T and NK Cell Mediated Immunity in Chronic Myeloid Leukaemia
31/2016 Ninja Karikoski
The Prevalence and Histopathology of Endocrinopathic Laminitis in Horses
32/2016 Michael Backlund
Regulation of Angiotensin II Type 1 Receptor by Its Messenger RNA-Binding Proteins
33/2016 Stanislav Rozov
Circadian and Histaminergic Regulation of the Sleep-Wakefulness Cycle
34/2016 Bárbara Herranz Blanco
Multi-Approach Design and Fabrication of Hybrid Composites for Drug Delivery and Cancer Therapy
35/2016 Siri Tähtinen
Combining Oncolytic Immunotherapies to Break Tumor Resistance
36/2016 Katri Sääksjärvi
Diet, Lifestyle Factors, Metabolic Health and Risk of Parkinson's Disease – A Prospective Cohort Study

Helsinki 2016 ISSN 2342-3161 ISBN 978-951-51-2173-8