Chémoinformatique intégrative pour guider la conception des médicaments : application à la re-conception d'un inhibiteur clinique de kinase protéinique

Melanie Schneider

▶ **To cite this version:**

Melanie Schneider. Chémoinformatique intégrative pour guider la conception des médicaments : application à la re-conception d'un inhibiteur clinique de kinase protéinique. Agricultural sciences. Université Montpellier, 2019. English. NNT : 2019MONTT057 . tel-02485659

HAL Id: tel-02485659
https://theses.hal.science/tel-02485659

Submitted on 20 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INTEGRATIVE CHEMOINFORMATICS TO GUIDE DRUG DESIGN:
Application to re-design a clinical protein kinase inhibitor

CHÉMOINFORMATIQUE INTÉGRATIVE POUR GUIDER LA CONCEPTION DES MÉDICAMENTS:
Application à la re-conception d'un inhibiteur clinique de kinase protéinique

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
THÈSE POUR OBTENIR LE GRADE DE DOCTEUR

Melanie Schneider

University of Montpellier

Doctoral School: Chemical & Biological Sciences for Health (CBS2)
Disciplinary field: Medical Biology

Academic advisor:
Dr. Gilles Labesse[1]
[1]Centre de Biochimie Structurale (CBS), CNRS, INSERM, Univ Montpellier, 34090 Montpellier, France

Date of defence: 06/12/2019

Assessment committee:

Prof. Annick Dejaegere , IGBMC, University of Strasbourg, France	Reviewer
Prof. Michael Nilges , Structural Biology and Chemistry Department, Institut Pasteur - Paris, France	Reviewer
Dr. Dominique Douguet , IPMC, University of Nice Sophia Antipolis, France	Examiner
Dr. Dragos Horvath , Chemoinformatics Laboratory, University of Strasbourg, France	Examiner

INTEGRATIVE CHEMOINFORMATICS TO GUIDE DRUG DESIGN: Application to re-design a clinical protein kinase inhibitor

Short abstract:

The main objective of this thesis is to improve the design of dabrafenib, a drug that is in clinical use against cancer, but is rapidly metabolized and shows undesired adverse effects potentially triggered through binding to a recently discovered secondary target. Therefore, the goal is an improved drug that still binds its primary target, the oncogenic protein kinase mutant BRAF-V600E, with high affinity, but not any more the secondary target, the nuclear receptor PXR. In order to achieve this, computational tools are developed and applied, such as virtual compound synthesis, virtual screening, machine learning, modelling and molecular dynamics coupled with MM-PBSA. A central aim of the thesis is to obtain accurate affinity predictions, which are crucial for subsequent design and development steps. The presented project also contains an experimental part, where the binding of the synthesized molecule is verified by solving the protein-complexed structure via X-ray crystallography.

Keywords: drug design, cancer, kinase, nuclear receptor, virtual screening, affinity prediction, machine learning, molecular dynamics, MM-PBSA, X-ray crystallography

Acknowledgements

It is a pleasure for me to express my gratitude to the many people who have helped me or have been an important part of my life during my time as PhD student.

First of all, I am grateful to my supervisor Gilles Labesse, for supporting me all the way from my Masters thesis work to pursuing the PhD degree. Gilles, I am incredibly thankful for the responsibility you trusted me with all this time that gave me the opportunity to explore new scientific territories and pursue my own ideas with all the projects I had in mind. I admire your knowledge and skills in all the diverse scientific fields and I thank you for the great amount of freedom and trust.

A particular thank you goes to the members of my PhD examination committee, for accepting to evaluate my work, for reading the manuscript, and for their invested time and effort.

I am thankful to the whole CBS for providing me with such a nice working environment, including all the pleasant lunches on the "always-sunny" institute's terrace. Of all the people involved in my thesis, I am particularly grateful to the whole ABCIS team. Jean-Luc, thank you for your constant availability and help concerning all issues with the @TOME server; Muriel, thank you for all your help with the crystallogenesis; Corinne and Rahila, thank you for making our office an enjoyable workplace; Jeff, thank you so much for all the nice, funny and enriching conversations, for your explanations of French culture, local customs and particularities, and above all, for all your help with the French administrative and juridical system. You were an extraordinary help and support. I am also thankful to collaborators, in particular to Martin and Laurène for their help with the protein purification. A special and warm thank you goes to Alessandro and his two postdocs Rémy and Matteo for "adopting" me in your team, for your care and constant support, in particular when I felt lonely, for including me in group meetings, for your constant and extraordinary help with computational issues and resources. Matteo, thank you so much for always being available no matter which issue, for all your advice and support concerning professional and personal questions, during and outside working hours. You have been an anchor during my PhD and I am very happy for the friendship we have built up during the years.

A very special thanks goes to, Sarah, Annika, Ashley, Peter, Robert and Pedro for their supportive friendship and for making me feel welcome in the CBS. Thank you for all the fun moments we shared, especially during relaxed weekend brunches, coffee breaks and *apéros* in the city center, and certainly, not to forget, all the great adventures and discovery trips to different places. Pedro, we didn't see each other a lot in the CBS, but the more I am grateful for the great time we had together outside the CBS. Thank you for all the inspiring and enriching discussions about science and different aspects of life, your friendship and amazing encouragement. I am equally thankful to all my friends outside the CBS for having had you at my side. A particular thanks goes to Carola, my lovely flat mate. You were a great companion during the last years and I am very happy for our mutual support during hard times and for all our shared happy moments.

Finally, I am forever thankful to my family for their constant support, love and unshakable belief that I will do the "right" thing, no matter where my path was taking me, and to Rémi for all of your love, support, understanding, appreciation, extraordinary care, and for simply being at my side.

Thank you all for this incredible PhD journey, full of emotions and great memories.

Melanie Schneider

Montpellier, France, October 2019

Contents

Acknowledgments 1

Abstract 9

Preface 10

List of Abbreviations 12

1 Background 13
1.1 Drug design . . . 14
1.1.1 An overview of drug discovery and development . . . 14
1.1.2 The drug's life-cycle in the organism and associated effects . . . 16
1.1.2.1 Pharmacokinetics / ADME . . . 16
1.1.2.2 Selectivity . . . 17
1.1.3 Drug-target interactions . . . 18
1.1.3.1 Physicochemical basis of drug-target interactions . . . 19
1.1.3.2 Mechanisms of action - how drugs interfere with organisms . . . 21
1.1.4 Experimental methods for drug design - studying the target . . . 23
1.1.4.1 X-ray crystallography . . . 23
1.1.4.2 Other structural techniques: MicroED, CryoEM, NMR & MS . . . 25
1.1.4.3 Affinity measurements: ITC, TSA, SPR, fluorescence & reporter assays 27
1.2 Computational methods for screening and affinity estimation . . . 30
1.2.1 Ligand-Based Virtual Screening (LBVS) . . . 30
1.2.1.1 Molecular descriptors . . . 31
1.2.1.2 Molecular fingerprints . . . 32
1.2.1.3 Similarity coefficients . . . 33
1.2.2 Quantitative Structure-Activity Relationship (QSAR) modelling . . . 33
1.2.2.1 Important considerations for data gathering and preparation . . . 33
1.2.2.2 Model building and validation . . . 35
1.2.2.3 Applicability domain . . . 35
1.2.3 Machine learning algorithms . . . 36
1.2.3.1 Random Forest (RF) and other tree-based algorithms . . . 36
1.2.3.2 Support Vector Machine (SVM) . . . 37
1.2.4 Structure-Based Virtual Screening (SBVS) . . . 38
1.2.4.1 Inverse virtual screening . . . 39
1.2.4.2 Ligand Flexibility in SBVS . . . 40
1.2.4.3 Target Flexibility in SBVS . . . 40
1.2.5 MM-PBSA - a Molecular Dynamics based method . . . 42
1.2.5.1 Introduction to Molecular Dynamics (MD) simulations . . . 42
1.2.5.2 Advantages and limitations of MM-PBSA . . . 44
1.2.6 Methodological overview with a special look on endocrine disruptors . . . 45
1.3 Fighting cancer: drug targets, antitargets and resistance . . . 54
1.3.1 Oncogenic protein kinase BRAF . . . 54
1.3.1.1 Structural basis of BRAF activation . . . 56
1.3.1.2 BRAF inhibitors . . . 57
1.3.2 Nuclear receptors . . . 59
1.3.2.1 The Pregnane X Receptor (PXR) . . . 60
1.3.2.2 The Estrogen Receptor alpha (ERα) . . . 61

2 ERα - a well studied target **63**
2.1 Affinity prediction method development on ERα . . . 64
2.2 The flexibility universe of ERα . . . 77

3 The drug design project **93**
3.1 Motivation . . . 94
3.2 Protein structure flexibility - a global view on BRAF & PXR . . . 95
3.2.1 Analysis of X-ray structures . . . 95
3.2.1.1 PXR - global analysis . . . 96
3.2.1.2 PXR - binding pocket analysis . . . 99
3.2.1.3 PXR - ensemble refinement . . . 101
3.2.1.4 BRAF - global analysis . . . 103
3.2.1.5 BRAF - binding pocket analysis . . . 107
3.2.1.6 BRAF - ensemble refinement . . . 108
3.3 Where to modify the drug? . . . 110
3.3.1 Drug binding in target (BRAF) and anti-target (PXR) . . . 110
3.3.2 BRAF structures with similar ligands . . . 112
3.3.3 Interfering with drug metabolism . . . 113
3.4 *In silico* synthesis of drug candidates . . . 115
3.5 Computational approaches investigating targets and ligands . . . 117
3.5.1 Molecular modelling and molecular dynamics . . . 117
3.5.1.1 Molecular modelling - BRAF and its loops . . . 117
3.5.1.2 Molecular dynamics simulations . . . 119
3.5.2 Machine learning methods for drug design from two perspectives . . . 121
3.6 Drug design synthesis rounds - a chronological overview . . . 135
3.7 MM-PBSA affinity calculations for designed molecules . . . 138
3.7.1 MM-PBSA approaches with dabrafenib and its metabolites . . . 138
3.7.2 MM-PBSA approaches on BRAF and PXR with designed drugs . . . 147
3.7.2.1 The impact of protein structure loop-model on MM-PBSA results . . . 147
3.7.2.2 The effect of different drug scaffolds . . . 148
3.7.3 MM-PBSA approaches on ensemble-refined BRAF structures . . . 152
3.8 Experimental work . . . 158
3.8.1 Activity tests by collaborators . . . 158
3.8.1.1 Results of the drug design rounds . . . 159
3.8.2 BRAF crystallogenesis . . . 161
3.8.3 Crystallographic structure determination . . . 162
3.9 Crystal structure analysis . . . 163

4 Conclusions and Perspectives **173**
4.1 Summary, discussion and conclusions . . . 174
4.1.1 The biological systems . . . 174
4.1.2 Molecular modelling . . . 176
4.1.3 Affinity prediction . . . 179
4.1.4 Selected and tested drug candidates . . . 180
4.2 General current issues and trends . . . 182
4.2.1 Molecular dynamics in drug development . . . 182
4.2.2 Machine learning in drug development . . . 183
4.2.3 The trend to target biological networks . . . 185
4.2.4 Target tractability & druggability . . . 185

Bibliography **187**

List of Figures **199**

List of Tables 203

Abstract

Despite years of intensive research and development, cancer remains one of the leading causes of death worldwide. Chemotherapy is the most commonly used treatment for cancer, as surgery and radiation therapy are often not effective in treating cancer at every location where it spreads. However, drug resistance of cancer cells to chemotherapeutic agents and/or reduction in effectiveness of a drug is the leading cause of failure of chemotherapy. Drugs are developed to bind efficiently to a given therapeutic target, called the primary target. Unfortunately, drug treatments can suffer from binding to a secondary target that perturbs drug activity and/or impacts its metabolism. The main aim of the PhD project was to develop an integrative chemoinformatics approach to optimize drug design by studying not only the primary target but also putative secondary effects at the atomic level in order to compute more accurate binding modes and to derive better affinity estimates.

The presented drug design project aims for an improved inhibitor of the serine/threonine kinase mutant BRAFV600E with simultaneous loss of binding to the secondary target PXR. Focus is on the study of both, protein kinase BRAF and nuclear receptor PXR, which is involved in regulation of xenobiotic metabolism. A machine learning model is first developed on the well studied nuclear receptor ERα due to large amounts of experimental data, and subsequently similarly generated for BRAFV600E. Despite its recognized importance in drug metabolism, we are still lacking sufficient structural information and affinity measurements to develop machine learning models for PXR. So, an alternative approach that relies on molecular dynamics combined with the Molecular Mechanics Poisson-Boltzmann Surface Area method is employed in order to obtain a precise estimation of ligand affinities. Finally, diverse computational tools are applied to design new derivatives of the initial drug, which is too rapidly metabolized in many patients resulting in resistance and cancer relapse. The properties of the new compounds prevent activation of metabolizing enzymes that are degrading the original drug. This is expected to provide a new drug-candidate with much better pharmacokintics properties and enhanced efficacy.

This thesis comprises a complete drug design pipeline and presents an integrated strategy that includes modeling, *in silico* design and synthesis, virtual screening, affinity predictions, *in vitro* tests and X-ray crystallography. The main focus is on the computational part that comprises complementary approaches from the drug's and from the proteins' point of view.

RÉSUMÉ

Malgré des années de recherche et de développement intensifs, le cancer reste l'une des principales causes de décès dans le monde. La chimiothérapie est le traitement le plus couramment utilisé contre le cancer, car la chirurgie et la radiothérapie ne sont souvent pas efficaces pour traiter le cancer à tous les endroits où il se propage. Cependant, la pharmacorésistance des cellules cancéreuses aux agents chimiothérapeutiques et / ou la réduction de l'efficacité d'un médicament est la principale cause d'échec de la chimiothérapie. Les médicaments sont développés pour se lier efficacement à une cible thérapeutique donnée, appelée cible primaire. Malheureusement, les traitements médicamenteux peuvent souffrir de la liaison à une cible secondaire qui perturbe l'activité du médicament et / ou a un impact sur son métabolisme. L'objectif principal du projet de thèse était de développer une approche chémo-informatique intégrative pour optimiser la conception de médicaments en étudiant non seulement la cible primaire, mais également les effets secondaires putatifs au niveau atomique afin de calculer des modes de liaison plus précis et d'obtenir de meilleures estimations d'affinité.

Le projet de conception de médicament présenté vise un inhibiteur amélioré du mutant de la sérine/thréonine kinase BRAFV600E avec perte simultanée de la liaison à la cible secondaire, le récepteur PXR. L'accent est mis sur l'étude à la fois de la protéine kinase BRAF et du récepteur nucléaire PXR, impliqué dans la régulation du métabolisme xénobiotique. Un modèle d'apprentissage automatique est d'abord développé sur le récepteur nucléaire bien étudié ERα en raison de grandes quantités de données expérimentales, puis généré de manière similaire pour BRAFV600E. Malgré son importance reconnue dans le métabolisme des médicaments, nous manquons toujours d'informations structurelles et de mesures d'affinité suffisantes pour développer l'apprentissage automatique sur PXR. Aussi, une approche alternative reposant sur la dynamique moléculaire associée à la méthode «Molecular Mechanics Poisson-Boltzmann Surface Area» est utilisée afin d'obtenir une estimation précise des affinités des ligands. Enfin, divers outils informatiques sont utilisés pour concevoir de nouveaux dérivés du médicament initial, métabolisé trop rapidement chez de nombreux patients, entraînant une résistance et une rechute du cancer. Les propriétés des nouveaux composés empêchent l'activation des enzymes métabolisantes qui dégradent le médicament initial. Ceci devrait fournir un nouveau médicament candidat aux propriétés pharmacocinétiques bien meilleures et à une efficacité accrue.

Cette thèse comprend un pipeline complet de conception de médicaments et présente une stratégie intégrée comprenant la modélisation, la conception et la synthèse *in silico*, le criblage virtuel, les prédictions d'affinité, les tests *in vitro* et la cristallographie de rayon X. L'accent est mis principalement sur la partie informatique qui comprend des approches complémentaires du point de vue du médicament et des protéines.

PREFACE

This thesis is submitted to the Faculty of Science, University of Montpellier, as a partial fulfillment of the requirements to obtain the PhD degree. The work presented was carried out in the years 2016-2019 in the group of Gilles Labesse at the Centre de Biochimie Structurale (CBS) of Montpellier, which is affiliated to the CNRS, INSERM, and the University of Montpellier.

THESIS OBJECTIVES

The main aim of this thesis was to develop an integrative drug design strategy that makes use of diverse computational and also experimental techniques to be able to take into account the primary and the undesired secondary target. The work may be divided into interconnected sub-topics: 1) *in silico* synthesis of producible compounds, 2) investigation on the proteins' conformational flexibility, 3) development and calibration of computational tools for improved affinity predictions, and 4) experimental confirmation of the drug binding mode by X-ray crystallography.

THESIS OUTLINE

The first chapter of the thesis provides a general introduction to drug design, introduces the protein targets ERα and PXR (two nuclear receptors), and BRAF (an oncogenic protein kinase), as well as all methods that are used within the scope of this thesis.

The second chapter is focused on the well studied nuclear receptor ERα. Here, advantage was taken of the large amount of data that is available in different databases to develop and test a machine learning method, which combines structure-based and ligand-based approaches. The method relies on the random forest algorithm calibrated on 1500 known ligands of ERα, uses a combination of structure-based virtual screening (with docking) and chemometrics and exploits ensembles on different levels. This work is published and included in the Thesis. The developed tool should help detecting potential endocrine disruptors and guide their modifications. In general, it can also be used to check potential binding of any drug candidate, as this nuclear receptor can be an unwanted secondary target. A Web-server gives access to the tool within a user-friendly pipeline which allows for a quick evaluation of putative binders of the estrogen receptors (ERα and ERβ) and the peroxisome proliferator-activated receptor (PPARγ). Although, primarily focus on the theoretical and fundamental aspect of ligand docking on ERα, the approach can be generalized straightforwardly to other nuclear receptors thanks to extensive characterization by means of X-ray crystallography and affinity measurements.

The third chapter describes the drug design project that aims for an improved inhibitor of the serine/threonine kinase mutant BRAFV600E with simultaneous loss of binding to the secondary target PXR. The unexpected and unwanted binding to PXR was recently characterized by *in vitro* cell assay and also by X-ray crystallography. These results explain the observed pharmacokinetics of this drug at the atomic level. A new chemical series is derived from the initial drug currently used against cancer, with the drug design procedures comprising several rounds of iterative improvement. The crystal structure with the primary target BRAF confirmed the mode of binding of one designed drug candidate and highlighted a sub-domain swapping previously undescribed for protein kinases. The properties of the new compounds are expected to prevent activation of metabolizing enzymes

that are degrading the original drug while maintaining the activity at the molecular and cellular levels. This should provide a new drug-candidate with much better pharmacokinetics properties and enhanced efficacy.

Finally, in chapter four the thesis is rounded up with a summary, discussion and conclusions on the performed work and obtained results. As perspective, general current issues and trends in the different fields are additionally presented.

List of abbreviations

Experimental methods

FBDD = Fragment-Based Drug Discovery

MicroED = Micro-crystal Electron Diffraction

CryoEM = Cryo Electron Microscopy

NMR = Nuclear Magnetic Resonance spectroscopy

MS = Mass Spectrometry

TSA = Thermal Shift Assay

ITC = Isothermal Titration Calorimetry

SPR = Surface Plasmon Resonance

FRET = Förster/Fluorescence Resonance Energy Transfer

Computational methods & tools

VS = Virtual Screening

LBVS = Ligand-Based Virtual Screening

SBVS = Structure-Based Virtual Screening

QSAR = Quantitative Structure-Activity Relationship

MD = Molecular Dynamics

MM-PBSA = Molecular Mechanics Poisson–Boltzmann Surface Area

MM-GBSA = Molecular Mechanics Generalized Born Surface Area

ML = Machine Learning

RF = Random Forest

SVM = Support Vector Machine

CV = Cross Validation

PCA = Principal Component Analysis

Proteins & domains

BRAF = Serine/threonine-protein kinase B-Raf

NR = Nuclear Receptor

PXR = Pregnane X Receptor

ER = Estrogen Receptor

DBD = DNA-binding domain

LBD = Ligand binding domain

Chemical compounds / drugs

DB = Dabrafenib

CDB = Carboxy-Dabrafenib

DDB = Desmethyl-Dabrafenib

HDB = Hydroxy-Dabrafenib

Databases

PDB = Protein Data Bank

BDB = Binding Database (BindingDB)

Evaluation metrics

DPI = Diffraction Precision Index

R^2 = Coefficient of determination

RMSE = Root Mean Square Error

RMSD = Root Mean Square Deviation

RMSF = Root Mean Square Fluctuation

1

Background

The first chapter introduces underlying concepts and highlights the motivation of the research conducted in the thesis

1.1 Drug design

Drug design, also known as rational drug design or simply rational design, is an umbrella term for the ensemble of processes necessary for the development of a drug. It describes the inventive procedure aimed at finding a new medication for a defined biological and medical relevant target.

In general, target-based drug design follows three main steps to create a new drug: First, a receptor or enzyme that is relevant to the targeted disease needs to be identified. For full efficiency in drug design this target needs to fulfill several tractability criteria. Often, in a second step, the structure and function of this receptor or enzyme is elucidated. Subsequently, this information can be used to design a drug molecule that interacts with the receptor or enzyme in a therapeutically beneficial way, which is often performed from the key-lock model point of view for molecular interactions. During all steps a lot of different factors have to be taken into account. For example, whether or not the drug will bind to a specific target or more than one (single-target versus multi-target drugs and polypharmacology). Polypharmacology may bring higher efficiency and lower the emergence of resistance but binding to several targets might also involve more adverse effects. As a drug may also have natural competitors addressing the same binding site, the minimal required affinity to the target molecule needs to be clarified. Usually the chemical composition of the molecule(s), including shape and charge, is investigated to determine the way it binds to the target. Subsequently, the ligand's properties have also to be investigated with respect to how they might affect absorption, distribution, metabolization, and excretion by the body, shortly called ADME or pharmacokinetics. In a nutshell, geometric and thermodynamic aspects need to be determined or predicted to rationalize drug design.

1.1.1 An overview of drug discovery and development

Pharmaceutical sciences target the development of medical health products with a major emphasis on the development of drugs, which are supposed to interact in a specific manner with the processes of life. A drug's purpose is to interfere with the biological processes in a favorable manner so that an improvement or cure of a disease is achieved. This is not a trivial task with respect to all possible interfering interactions that can occur and are potentially unfavorable or even toxic in nature, e.g. off target effects through unwanted binding to secondary targets, drug-drug interactions, or metabolic effects. Therefore, the process of drug development involves many steps in conceptual design, refinement, and testing that are repeated in a cyclic fashion and accompanied by security checks.

Currently, the drug discovery process follows a general pathway of target validation, experimental assay development, small molecule library screening, Hit-to-Lead, Lead optimization, pre-clinical drug development, and clinical drug development. So, the whole process begins with the identification of a target macromolecule. This should be a molecule (usually a protein or a protein complex, rarely also DNA or RNA) whose function is essential for the development or expression of a disease and therefore a promising point of attack to fight the disease. The second step is the development of specific assays for small molecule screening experiments, which are used to test the activity of a large set of possible drug candidates. These assays are "artificial" measurements, but supposed to

partially mimic "real life" (at the molecular level at least) by approximating biological interactions in a minimalist way. Upon identification of good binders, which are termed as hit compounds, the Hit-to-Lead optimization cycle starts as an iterative optimization cycle to further improve the affinity. The most promising compounds are then used in pre-clinical and, if successful, in clinical tests. This process is based on an enormous amount of molecules that have to be synthesized and tested in order to (possibly) obtain one single drug.

A slightly different setting applies to Fragment-Based Drug Discovery (FBDD), which aims to find low molecular weight compounds (fragments) that should serve as chemical starting points for drug discovery. FBDD has become an established technique in industry and academia.[1] In contrast to the conventional high-throughput screening involving the screening of large numbers (from tens of thousands to millions) of higher molecular weight compounds (MW 300–500 Da) usually via *in vitro* bioassays, FBDD is based on the biophysical screening of a smaller number (thousands) of low molecular weight compounds (MW < 250 Da) against target proteins. In FBDD typically higher hit rates are attained,[2] as with smaller sizes of the compounds the probability of matching within the binding site increases and the likelihood of clashes decreases. Simultaneously, fragments sample the chemical space more efficiently than larger drug-sized molecules at a similar library size.[3] As fragments usually bind to their targets with much lower affinity (>1 mM)[4] (due to the limited numbers of potential interactions that can be formed), screening methods need to be sufficiently sensitive to detect weak interactions, which is the case for biophysical methods, such as nuclear magnetic resonance (NMR), thermal shift assay (TSA), surface plasmon resonance (SPR) and X-ray crystallography. Vemurafenib, an inhibitor of the protein kinase BRAF for the treatment of late-stage melanoma, was the first FDA-approved drug (in 2011) originating from FBDD,[5] and was followed by further approved drugs.

Despite the fact that the efforts and money pharmaceutical companies are investing in drug development constantly increased during the last decades, the number of drug approvals stay almost constant.[6] Furthermore, about 81% of all new drug candidates fail,[7] which is mainly caused by a lack of drug efficiency and side effects associated with off-target binding. Therefore, it is very cost and time intensive, consuming billions of dollars and taking up to 15 years.[8] Considering these high failure rates limitations in the current methods of drug development can be assumed.[6] This reflects the need for improvement of drug development methods.

In general, drug discovery and development require the integration of multiple scientific and technological disciplines, such as chemistry, biology, pharmacology and extensive use of information technology that rely on mathematical and physical concepts. Pharmaceutical or medical chemistry is a highly interdisciplinary field dedicated to drug design, optimization of pharmacokinetics and pharmacodynamics, and synthesis of new drug molecules and shows an increasing need for computational developments in order to cope with all the recent advances in data generation and availability (Big Data in healthcare). In particular pharmacoinformatics is considered as a rather new discipline combining different informatics branches, such as bioinformatics and chemoinformatics, into a single platform that aims for a systematic approach in drug discovery and development in order to increase efficiency and safety.

1.1.2 The drug's life-cycle in the organism and associated effects

Pharmacology, a discipline that deals with the origin, nature, chemistry, effects, and uses of drugs, is dedicated to the question how to deliver a drug to the living organism, which cures or ameliorates a disease, while at the same time harmful effects should be kept minimal and eliminate in due time. As chemical compounds promote multiple effects (beneficial ones and harmful ones) the issue is highly complex. Further, the organism affects the drug itself, influencing its distribution, its metabolism and its excretion from the body.

1.1.2.1 Pharmacokinetics / ADME

Pharmacokinetics, also termed ADME, refers to the four major aspects absorption, distribution, metabolism and elimination.

Drug absorption and distribution

Absorption questions the first requirement of a drug, the fact that it has to enter the body. Often blood circulation is a main destination making a further distribution possible. In order to reach this destination a drug must overcome multiple biological barriers, with the cell membrane being the most important. The cell membrane separates the cell's intracellular space from the surrounding extracellular space and is composed of a bilayer of phospholipids and proteins. The phospholipid bilayer creates a barrier that hinders large hydrophilic drugs from entering the cell, while small hydrophobic ones pass more easily, as they can solubilize in the cell membrane and are therefore able to diffuse over it, although they do sometimes accumulate there when hydrophobicity is too high, as they cannot solubilize in the aqueous environment within the cell. Due to the high importance of lipophilicity in drug discovery, numerous log P (partition coefficient) calculators (e.g AlogP,[9] XlogP,[10] etc) have been developed that can be used to filter drug candidates. For very small hydrophilic compounds (< 100-200 Da) there is still the possibility to pass via water filled pores and channels, situated within the membrane. Another possibility to pass the membrane (besides passive diffusion) are transporters. Transporters are dedicated to facilitate the translocation of large and/or hydrophilic compounds over the cell membrane that could otherwise not pass, fulfilling several important roles, e.g. supplying the cell with water soluble nutrients, such as sugars and amino acids, or exporting waste products from the cell formed during cell metabolism. Therefore, transporters are mayor modifiers of the drugs distribution in the body. It may happen that a drug is hindered to access an organ because a transporter pumps it actively out of the cells, in a faster rate than it could diffuse passively inside. Among the ways of acquiring drug resistance in cancer cells is also the upregulation of such transporters expelling the anti-cancer drug. Furthermore, a drug has also the possibility to be taken up or excreted by a cell via vesicular transport, called endocytosis/exocytosis, but this plays rather a minor role for the distribution of drugs. Depending on the drugs destination it may have to pass other biological barriers on its way to the final destination. Such barriers, composed of cells forming a densely packed structure that restricts the passage in between the cells are present in certain areas of the body with the major ones being the gastrointestinal mucosa, the blood brain

barrier, the epidermis and the placenta. All the barriers a drug has to pass constrain its desired properties and are therefore affecting its conception. For example, if a swallowed drug shows too low solubility, it may just pass through the intestine and be excreted without any uptake, as it is not able to be dissolved in the gastrointestinal fluid upon release from the tablet. Moreover, drug transport is majorly impacted by the drug's metabolism within the organism.

Drug metabolism, elimination and toxicity

Metabolism can rapidly inactivate or eliminate a drug, for example by transforming a hydrophobic compound into a more water soluble one in order to be easily excreted by the kidney into the urine. Metabolism is a vital function of the organism and not only affecting drugs. It is also responsible for treating xenobiotics and endogenous substances, such as hydrophobic compounds taken up via the food, or hydrophobic metabolic side products produced by the body. The primary enzymes that introduce polar groups (e.g. a hydroxy group) into a compound are the Cytochrome P450 monooxygenases (CYPs). Metabolism does not necessarily make a drug inactive. A modified drug may be equally active on its target, or a drug may be even designed to become only active after a first step of metabolization, converting a so called pro-drug to the active compound. A drug molecule may also become toxic upon metabolization and/or subsequent drug interactions, such as enzymatic inhibition or induction, which may lead to adverse effects.

It is equally important for a drug to be eliminated from the body, as accumulation of the drug will eventually lead to long-term toxic effects. Excretion of substances happens primarily via the kidney, as one of its functions is to clean the blood from "suspicious" components (xenobiotics absorbed from the environment) and excrete them to the urine. For this to happen effectively the compounds to be excreted need to be polar, as non-polar compounds will be passively reabsorbed through the cell walls of the kidney tubuli cells and subsequently passed back to the blood. Thus, highly hydrophobic compounds inert to metabolism (e.g. certain pesticides) cannot easily be removed by renal excretion and are retained in the body with very long half lives, eventually leading to adverse effects.

The specific modelling of pharmacocinetics within an organism is represented as a field of research by its own, termed systems biology, and falls beyond the scope of this thesis.

1.1.2.2 Selectivity

Besides all the behaviours and effects a drug can show during the ADME processes, it is essential to investigate possible molecular interactions. Selectivity is one of the crucial issues in pharmacology. A drug that binds to many different targets may trigger multiple pharmacological effects, including undesired adverse ones. Therefore, binding to a single or only a few targets is being striven for when designing a drug. Such a drug is highly selective. However, if the drug's concentration is high enough it may interact with quasi any protein. Therefore, selectivity of a given drug is not an absolute value, but rather expressed as ratio of K_d values, the fold difference between two targets. Finding highly selective molecules can be a difficult task, especially when a given target has many structurally similar class members, which is the case for protein kinases with its more than 500 class members encoded within the human genome. Thus, highly selective kinase inhibitors often still bind several kinases.[11–13] This reflects the great challenge of attaining selectivity among distinct members

of protein superfamilies. Additionally, many functional protein domains are widely spread even across different protein families.
Nonetheless, conventional chemotherapy is often not specific at all addressing cell division/metabolism in general, but also comes along with severe adverse affects and damage of normal tissues.[14] In contrast, targeted drugs are supposed to act in a more specific manner on dedicated subpopulations of cells, but are unfortunately often more prone to the development of resistance.

1.1.3 Drug-target interactions

The molecular mechanisms and mode of action of drugs is an important information that enables further activity modulation and improvement. In many cases those are fairly well understood, but in other cases they are still unknown. The conceptual design of a drug is based on the knowledge about the mode of action and more precisely the molecular mode of interaction with the target. At the same time, it is important to evaluate a drug's effect not only with respect to its interaction with one single target, but also by taking into account the other essential requirements for a drug, for instance, the pharmacokinetic (ADME) and toxicological properties. Since biological organisms are extremely complicated systems, the effect of a drug within an organism and the effect of the organism's response to the drug are multifaceted. Already a minor structural change aimed to optimize one particular property of the drug can heavily impact another characteristic. Therefore, the simultaneous fine tuning of many different drug characteristics is mandatory and makes drug development so difficult. Keeping this in mind, it may be similarly beneficial to start from a licensed drug and improve certain properties to achieve a higher efficacy, or reduced adverse side effects. Another aspect is that it may take time to find out about adverse effects of drugs and how they are caused, thus requiring later modifications of the drug. Moreover, completely different strategies exist that are based on permanent inactivation via covalent inhibitors or on the removal of the target protein via degradation induced by proteolysis targeting chimeras (PROTACs).

As drugs mediate their actions by interacting with their dedicated macromolecular target, the way of interaction is of major concern. The biological macromolecules serving as drug targets are primarily proteins, sometimes also DNA or RNA, but proteins with their higher structural variability are thought to present much more versatile points of attack to engineer low molecular weight compounds with very specific pharmacological effects. In general, proteins have either a structural or functional role, whereas the functional role can be very diverse. Examples are enzymes, transporter proteins, ion channels, or signalling proteins. One medically prominent example for enzymes are kinases that phosphorylate their substrate and are involved in many cellular pathways, and for signalling proteins such an example would be hormone receptors. Depending on the proteins function, drugs can be designed to interact in different ways to exert their function.

1.1.3.1 Physicochemical basis of drug-target interactions

All types of proteins obey very similar chemical principles for exerting their pharmacological actions and for establishing a drug-target or, more general, a ligand-receptor interaction.
Macroscopically, association or dissociation of the ligand-receptor complex can be seen as a reaction in a chemical equilibrium:

$$L + R \rightleftharpoons LR \tag{1.1}$$

Consequently, the *dissociation constant* K_d for the complex is defined as:

$$K_d = \frac{[L][R]}{[LR]} \tag{1.2}$$

where [L], [R], and [LR] are the equilibrium concentrations of ligand L, receptor R, and the complex LR, respectively. Generally, the smaller the dissociation constant, the stronger is the interaction between ligand and receptor, and the higher is the affinity between them. Note that this is a simplistic description of a ligand-receptor binding event, without taking into account more complicated scenarios, such as different stoichiometry of ligand and receptor, or the occurrence of allosteric effects, intermediate transition steps, or solvent effects.

Thermodynamics, Gibbs Energy & Entropy

The occurrence of interactions between ligand and target can be expressed in energetic terms and can be related to the equilibrium constant for association (K_a) or dissociation (K_d) by the equation:

$$\Delta G = -RT \ln K_a = RT \ln K_d \tag{1.3}$$

where ΔG is the Gibbs free energy, R is the gas constant and T is the absolute temperature in Kelvin. For example, a decrease in free energy of 10 kcal/mol ($\approx$ 42 kJ/mol) relates to an approximate K_d of 10^{-7} mol/L. Following to the discovery that many developed drugs have the most favorable binding enthalpy,[15] the consideration of thermodynamic data is nowadays often included in the drug development process.[16]

The formation of bonds between ligand and target is not only represented by the enthalpy. Also the entropy has an impact on the change in Gibbs free energy. The Gibbs free energy (ΔG) and the change in the entropy (ΔS) upon binding can be calculated using the relationship

$$\Delta G = \Delta H - T\Delta S \tag{1.4}$$

where ΔH is the enthalpy. If the Gibbs free energy decreases, binding occurs spontaneously and energy is freed in form of heat. This is the case either when enthalpy decreases and/or when entropy increases. Drug binding can be primarily enthalpy-driven or rather entropy-driven. Different classes of ligands can have different enthalpy-entropy impact partitioning even on the same receptor[17] and enthalpy-entropy compensation is a commonly observed effect in Lead optimization,[18] which is usually measured by Isothermal Titration Calorimetry (ITC). During optimization, while the structure of the molecule is optimized to form more or better contacts with the target, which makes ΔH more negative, the introduced interactions can force the system into a more ordered state, which

in turn changes ΔS unfavorably. Hence, drug design approaches tend to introduce constraints to the movement of the molecule (e.g. through macro-cycles) simultaneously when adding new functional groups with increased interactions (favorable enthalpy). This ligand restraining strategy aims to reduce the entropic effect when the ligand is transferred from the free state in solution into the more restraining binding pocket. Other approaches to reduce the entropic effect are focused on the binding pocket. For example hydrogen bonds can be directed to already structured regions of the protein or multiple hydrogen bonding interactions can be used from a single group to strengthen the interaction, whereby the entropic penalty has already been paid.[19] When thinking about the transfer of the molecule from the solvated state into the binding pocket, there is another important factor to be taken into account - the dehydration and with it the hydrophobic effect. The hydrophobic effect is the preference of hydrophobic, non-polar molecules to aggregate in aqueous solution in order to minimize the solvent exposed surface area towards the polar water molecules. It involves the displacement of water molecules arranged around the hydrophobic surfaces of both the protein and ligand. This effect can equally be described by the enthalpy-entropy compensation. For instance, most water molecules that are located in the binding pocket prior to binding are relocated during the binding event to leave space for the ligand. The water molecules displaced into the bulk water have more freedom to form hydrogen bonds with the surrounding, leading to an increase in translational and rotational entropy of the water molecules, resulting in a favorable entropy of water release. Not only water molecules that are replaced from the binding pocket can contribute. Even subtle changes in the water network surrounding a ligand can have a compensatory thermodynamic effect.[20,21] Wienen-Schmidt et al.[20] demonstrate that an entropically more favored binding can even be mainly caused by shedding the ligand's hydration shell upon leaving the bulk water. In this case, the termodynamic signature is affected by the ligand's water trapping capabilities in aqueous solution prior to binding and has nothing to do with the binding to the protein. Breiten et al.[21] identify the water molecules on the surface that contact the ligand from the outside (when the ligand is bound in the binding site) as important contributors to the enthalpy-entropy compensation. Additionally, the organization of the whole ligand-receptor complex has to be taken into account, as it could be that the bound complex attains more flexibility compared to the unbound state, resulting in an overall decreased order of the bound system. This highlights the complexity of dynamic processes involved in drug binding, which cannot easily be analyzed by only taking into account single rigid conformations of a given drug and target (as proposed by the rather simplistic key-lock model).

It has also been suggested that thermodynamic profiles could be of help for identifying inhibitors that are optimized with respect to flexibility, water solubility and specificity. Optimizing flexibility can be beneficial in case of rapid mutation of the target binding site to minimize drug resistance.[22] Solubility in water may be optimized to maximize the ligand's efficiency with respect to polar interactions.[23,24] Specificity/selectivity is optimized to reduce side effects caused by binding to unwanted secondary targets.[25–27] All these molecular features are results from the atomic nature of the targets and ligands which provide a wide range of possible interactions.

Intermolecular forces

For drug binding to occur intermolecular forces/interactions need to be established on a microscopic/atomic level between the macromolecular target (and potentially its prosthetic groups) and one (or several) ligand molecule(s). Some drugs react chemically with their targets and form covalent bonds, which make the attachment generally irreversible (40-140 kcal/mol). In general, interactions depend

on the atom type and therefore its properties. For many drugs the interactions are non-covalent in nature and include electrostatic interactions, such as ionic bonds (5 kcal/mol), hydrogen bonds (1-10 kcal/mol), and halogen bonds (1-40 kcal/mol), Van der Waals forces (0.5-1 kcal/mol), such as dipole-induced dipole interactions (Debye forces), and London dispersion forces. The latter are the weakest type of interaction, but the high number of occurring contacts in organic molecules can sum up to large contributions. The direct environment of atoms modifies the formation capability and strength of the particular interaction. Moreover, π-effects - molecular interactions with the π-systems of conjugated molecules (π-π interactions, cation/anion-π interactions, and polar-π interactions) - are often crucial for protein-ligand recognition, but difficult to model. Additionally, chelation with metal ions within the binding site can occur impacting ligand binding.

1.1.3.2 Mechanisms of action - how drugs interfere with organisms

It is important to distinguish between the action and the effect of a drug. Drug action refers to the initial consequence of a drug-target combination, the mechanisms by which the chemical produces a response in an organism, whereas the drug effect refers to the biochemical and physiological changes that occur as a consequence of drug action.

Depending on the actual site of binding, the interaction of the ligand/drug with the target can be competitive or allosteric. **Competitive binding** takes place in the orthosteric binding site of the target (or active site in case of enzymes) and the competition occurs between the drug and the natural ligand (or substrate) that exerts a specific effect in the organism. **Allostery** refers to the binding of an effector molecule at a site other than the orthosteric binding site, whereupon usually through a flexibility/conformational change the activity of the target is modulated. Both ways of interaction, competitive and allosteric, can be described for receptors and enzymes, but the term allostery is more often used when describing enzyme modulation.

Another way of distinguishing mechanisms of action is based on the initial consequence of the drug-target interaction, which can be positive or negative (activating or inhibiting) for receptors and enzymes.

Receptor modulation - principle of agonism, inverse agonism and antagonism

Ligand-receptor interactions can be very precise modulators and result in different levels of activity. Therefore, in the field of pharmacology the ligands/drugs are classified according to their effect on the receptor. The term 'pharmacological receptor theory' explains ligand behavior and classifies the ligand-receptor activity. The main classes are agonist, inverse agonist and antagonist.
An ***agonist*** is a ligand that binds to a receptor and causes the increase of a biological response by a direct activation of the receptor. Depending on the strength of the exerted stimulation, agonists can further be distinguished between *partial agonists* and *full agonists*. An ***inverse agonist***, causes the opposing effect by directly reducing the receptor's basal activity. Therefore, for the effect of inverse agonism a prerequisite is that the receptor must have a basal level of activity in the absence of any ligand in order to enable a reduction of activity. An ***antagonist*** blocks the action of the agonist through competitive binding to the same binding site, but does not lower the basal activity. An additional class are ***selective receptor modulators***. They display an agonist response in some tissues

and an antagonistic response in other tissues. To exert such a behavior, they are expected to promote a conformation of the receptor that is closely balanced between agonism and antagonism and the resulting effect in the tissue is depending on the concentrations of coactivators and corepressors.

The effect of agonism, inverse agonism, antagonism and intermediate effects can be explained by a shift of equilibrium between conformational states. Overall, ligands have the ability to shift the equilibrium between the conformational states of a protein, which in turn can be described by distinct binding affinities of the ligand for the different states. The conformation with the highest binding affinity for the ligand will be the most stabilized and therefore the most occurring in complex with the ligand, but as energetic differences between several states may be small, a large conformational diversity may occur in solution.

Enzyme modulation

The ***active site*** of an enzyme, also known as catalytic site, is the part of an enzyme at which catalysis of a (or several) substrate(s) into a (or several) product(s) occurs. When drugs are designed to bind within the active site (sometimes mimicking the natural substrate), they may block the catalytic activity of the enzyme and therefore serve as enzyme inhibitors. This inhibition results in a reduced availability of the product, which may cause an altered physiology and or an accumulation of substrate. When inhibitors resemble the transition state of a catalyzed reaction, they are called *transition state analogs/inhibitors*, and when the inhibitor during the catalysis reaction covalently attaches to the target enzyme by forming an atomic bond, with a probability of detachment becoming so small (quasi irreversible), it is called *suicide inhibition*.
When the drug target is an enzyme, the mechanism of action is usually inhibition. Activation of an enzyme is usually more complicated and can generally only be achieved through allosteric binding.

Key points

⇒ Globally, a drug's effect on an entire organism is difficult to predict, as a multitude of effects come into play.

⇒ Locally, drug-target binding can be a complex event involving enthalpic and entropic effects originating from the target, the ligand and the solution/environment.

⇒ An atomic structure is the basis for rational target-based drug design.

1.1.4 Experimental methods for drug design - studying the target

The increased understanding of the molecular mechanisms of diseases has allowed the identification of many biological macromolecules implicated in disease. Nonetheless, the identification of promising targets is not sufficient to successfully develop medication. The target molecules, together with their mode of action, need to be characterized, before they can be used as a starting point for target-based drug design. The protein structures themselves are used in target identification and selection (the assessment of the druggability or tractability of a target), as well as in the identification of hits by virtual screening and in the screening of fragments. This highlights the key role of structural biology within the process of drug development. Currently, the required atomic structures are mainly delivered by X-ray crystallography and sometimes also by other structural techniques.

1.1.4.1 X-ray crystallography

X-ray crystallography is one of the most common techniques used to determine the three-dimensional structure of biological macromolecules and plays an important role in structure-based drug design, as it provides atomic details of the macromolecular structures. There has been an explosion in the number of macromolecular structures that are available (mainly deposited in the PDB). The increased pace for structure determination is probably due to several technical advances: the automation of protein production and crystallogenesis including the establishing of crystallographic screening platforms that opened the way to high throughput screening of small molecules crystallized within their targets; the availability of powerful synchrotron radiation; as well as new algorithms and software that automate many processes within data collection, structure solution and refinement. X-ray crystallography is particularly well suited for drug discovery, as it can be used to determine the structure for rather large heteromeric complexes, a very high (potentially atomic) resolution can be attained, and it often reveals detailed experimental information about the binding mode of ligands found in the crystal. The possibility to precisely visualize the architecture and the interactions between ligand and target facilitates the understanding of mechanisms and drug activity at a molecular level, but lacks information on the thermodynamics of the system.

As the crystallographic form implies a rather compact packing of the protein, one may wonder whether the protein is totally rigid or distorted in a crystal. The answer to this question is ambiguous and depends largely on the protein itself: On one side, it is not totally rigid, because the solvent content within a crystal is generally about 50%,[28] different protein conformations can co-exist within a crystal,[29,30] a ligand can freely diffuse into a crystallized protein (soaking), and enzymatic activity has even been observed within crystals.[31] Furthermore, many proteins are crystallized in different conformations, and even as differing independent assemblies within the same asymmetric unit of a single crystal. On the other side, artifacts may occur, because the protein structure may be affected by the crystallization conditions or the crystal packing, as conformational changes may occur, or be restricted in some flexible regions of the protein due to the crystal packing.[32,33] Moreover, it may not always be possible to "trap" important conformational states that occur in solution (especially those requiring large structural rearrangements).

Crystallogenesis - a long road

Crystallogenesis - the production of the crystal(s) - can be a long and tedious process, as it is not always easy and straightforward to find the right crystallization conditions. First of all, the process relies on a stable sample. Then, the process of crystal growth has to be initialized by trying to find the nucleation zone. Generally, the aim during crystallogenesis is to moderately decrease the protein's solubility, while not affecting its stability. There are different crystallization methods, with one of the most prominent ones being based on vapour diffusion. Nowadays, for screening many different conditions in a reasonable time, automation is established in form of commercial pre-made crystallization kits (with up to 2000 conditions) and robots for pipetting and monitoring. After a first hit the conditions usually have to be further optimized in order to obtain exploitable crystals. This is usually done by screening a matrix around the initial condition with slightly varied parameters, such as pH, protein concentration, salt concentration and type, precipitant type, or temperature. Seeding can be performed with previously obtained micro-crystals to favor larger crystals. Once conditions are known, crystallogenesis is often highly reproducible and can be employed for screening and drug design, even with low affinity fragments (either by soaking or co-crystallization).

A brief introduction to X-ray diffraction

Having finally obtained a crystal of the macromolecule or complex whose structure is to be determined, it is transferred to the X-ray radiation source, traditionally as frozen sample to better resist radiation damage during the experiment when the X-ray beam is passed through the crystal. The crystal is rotated within the X-ray beam and diffraction patterns (visible as spots) are recorded on a detector behind the crystal. X-rays are scattered by interaction with the electrons of the material.There are different types of interactions of X-rays: elastic scattering that is the main interaction and occurs without loss of energy, inelastic scattering (Compton scattering) that contributes to the noise in diffraction experiments, and absorbtion. When the X-ray waves pass through the crystal, the intrinsic properties of a crystal are crucial for obtaining a detectable diffraction pattern. A crystal is a periodic arrangement of molecules, perfectly repeated in a regular lattice, and all atoms are (or should be) in the same relative position compared to any other atom, which results in the amplification of the diffraction signal. Diffraction, a special case of elastic scattering, occurs when a wave meets an ordered object (e.g. a crystal) causing constructive and destructive interference, which is described by Braggs' Law. The interference potency of diffracted rays in each direction and therefore the intensity of each reflection (diffracted beam) depends on the constellation of all atoms within the smallest repeating unit (unit cell) within the crystal. Thus, the information about the positions of all the atoms in the crystal (real space) is encoded in the diffraction pattern (reciprocal space) and the position of each atom in the crystal influences the intensities of all the reflections. To relate the points in the diffraction pattern to the planes in the crystal lattice, a mathematical operation, the Fourier transformation is employed. To obtain a map of electron density, where the peaks in the electron density map correspond to the atomic positions, the intensities of the spots measured by the detector are used for the Fourier transformation calculations. Nonetheless, each reflection is characterized by its amplitude and phase. The amplitude can be obtained from the measured intensities, but the the diffraction pattern does not provide direct information about reflection phases, which constitutes the famous "phase problem", a mayor hurdle in structural crystallography.

Phasing diffraction data and refining models

There are three types of methods to solve the phase problem: 1) direct methods, which use probabilistic relations between certain groups of reflections to estimate the phases, 2) special-atom methods, which are experimental phasing methods, and 3) molecular replacement, which uses a structural model to infer the phases.

Direct methods require the diffraction data to extend to atomic resolution ($\sim$ 1Å) and are usually employed in small-molecule crystallography, and sometimes also on protein molecules. If a suitable atomic model of the unknown crystal structure is available, molecular replacement is the method of choice, being also the most commonly used method for solving protein structures (this success can be attributed to the enormous growth of available structures in the PDB and to improvements of computational tools). It exploits the Fourier transform of reflection intensities, wherein model derived interatomic vectors compared to the experimental data reveal the orientation and location of the model molecule in the unit cell.[34] Improvements in the algorithms used in structure refinement software, such as the implementation of maximum entropy methods, contributed to the great success of molecular replacement. Moreover, there are software pipelines that automatically select useful structures for molecular replacement from the PDB. Furthermore, molecular replacement can now be combined with ab-initio structure prediction algorithms, as done by ROSETTA,[35] to solve structures that could not be solved otherwise.[36] X-ray diffraction data can also be exploited simultaneously, or combined with other experimental techniques (whereas the link between them is usually formed by computational tools).
In general, model refinement is considered as successful if both crystallographic R values, R_{work} and R_{free}, decrease during refinement. The R values R_{work} and R_{free} are used for validation of the agreement between model and measured data. They are crystallographic quality measures of the model that is obtained from the crystallographic data. Based on a built atomic model, R values measure how well the simulated diffraction pattern matches the experimentally observed diffraction pattern. Since refinement is aimed in improving the atomic model to make it fit better to the experimental data and improve the R value, this can lead to a over-fitting of the data. Therefore, before refinement, about 5% of the experimental observations are removed from the dataset, refinement is only performed on the remaining data. The data used in refinement is subsequently used to calculate R_{work} and the previously removed 5% are used to calculate R_{free}. Therefore, R_{free} can be seen as a relatively unbiased fitness function to examine model-data agreement.

One major limitation of X-ray crystallography is the fact that decently sized, homogeneous crystals are required to perform a diffraction experiment. Therefore, difficulties in protein crystallization due to limited solubility, elevated intrinsic flexibility, or other reasons, may even prevent any X-ray experiment and thus require other structural techniques.

1.1.4.2 Other structural techniques: MicroED, CryoEM, NMR & MS

If single well-ordered crystals of sufficient size (> 1000μm^3) cannot be obtained, a rather recent technique called **Micro-crystal Electron Diffraction (MicroED)**[37] can be the solution, as it works on micro-crystals with sizes of less than 300 nm and it has been shown that crystals of minuscule size compared to the size needed for X-ray crystallography can yield atomic-resolution structures.[38] It is

also a diffraction based method, but instead of an X-ray beam an electron beam is used to produce the diffraction pattern. The experiment is performed using a transmission electron microscope with a slightly adapted setup. Therefore, sample preparation, after obtaining the micro or nano-crystals, is the same as for all other CryoEM techniques, including advantages and limitations. Instead of using the imaging mode one uses the diffraction mode with an extremely low electron dose. Data processing, in turn, is performed using standard X-ray crystallographic software.

If crystallogenesis is not an option, **Cryo Electron Microscopy (CryoEM)** has become a popular option with increasing numbers in solved structures down to near-atomic resolution. CryoEM has several advantages compared to X-ray crystallography: First, it allows structural study of proteins that cannot be crystallized due to different reasons, such as intrinsic flexibility and solubility or stability issues. It enables as well the visualization of very large macromolecules and complexes, such as the nuclear pore complex, or the ribosome. Moreover, there are no crystallization effects that may affect the protein's structure, and it is possible to observe structures in different conformational states (even within one experiment), giving more insights on structural heterogeneity. This can include also post-translation modifications, such as methylation, acetylation, phosphorylation, ubiquitination, and glycosylation, that are de-homogenizing the sample and would therefore not be detectable within a crystal that requires a compact packing of a more homogeneous sample. Nonetheless, the typical resolution of CryoEM structures is still much lower as compared to X-ray structures, and CryoEM is still a low throughput technique and therefore, not employable for ligand screens. Despite this, CryoEM can be used in a complementary way to X-ray crystallography: Phases derived from CryoEM density maps may be used to solve X-ray structures by molecular replacement. CryoEM experiments can provide help for finding and selecting the optimal conditions for macromolecular stability (e.g. to guide optimal crystallization), and CryoEM data may also be used to verify results and confirm, for example the biological assembly of a macromolecule.

Nuclear Magnetic Resonance (NMR) spectroscopy is a classical 3D protein structure solution technique that also allows to study of the interaction of ligands with macromolecular targets. It is capable of determining flexible elements of macromolecules that cannot be revealed in X-ray experiments. Despite the fact that NMR is a rather low throughput technique compared to X-ray crystallography and other methods, and the sample requirements are large (labelling is often required), NMR experiments can simultaneously provide details about the structure, and the thermodynamics and kinetics of binding processes. They allow for measuring affinity (K_d) and for detecting flexibility changes, but on a low resolution level compared to X-ray crystallography. Ligand-protein interactions can be detected by analyzing changes of the ligand's signals in the presence of the target (ligand-observed), or by analyzing changes of the target's signals in the presence of the ligand (target-observed). Ligand-observed NMR is rather fast and simple to employ and very useful for fragment-based screening, because the high sensitivity enables detection of very weak binders. Another advantage is the ability to identify a single binder from a mixture of compounds. Target-observed NMR provides a higher quantity of information on ligand-target interactions and is rather used to validate screening hits (to eliminate false-positives). The main advantage is the ability to detect where and how the ligand binds to the target, in solution, and how this affects the dynamics of the molecules.[39,40]

Mass Spectrometry (MS) is a useful tool to test hypotheses in the case of ambiguous structural

data. MS is an often label-free analytical technique that gives the accurate mass of molecules and enables chemical identification. It has a wide range of applications and can be combined with various other techniques. In general, MS requires that the proteins in solution or solid state are turned into an ionized form in the gas phase. Fragmentation of the sample is usually part of the process and allows for identification within the analysis. MS has a very low 3D resolution, but requires only tiny amounts of protein. Therefore, it is used for molecular characterization and for evaluating the molecule's success chances for crystallogenesis.

In **native MS** the protein is not fragmented and therefore intact. It allows for ligand screening and provides information on various aspects, such as the overall shape and the stoichiometry of the protein-ligand complex (by measuring the mass of bound versus free target), binding reversibility (by inducing dissociation), binding-site specificity (by competition assays), and complex affinity as K_d (by titration).[41,42] Nonetheless, native MS has also drawbacks, such as experimental difficulties with binding assay conditions, gas-phase dissociation and non-specific binding.[43]

Chemical cross-linking coupled with Mass Spectrometry (XL-MS) is a technique to get sparse information about the overall structure of a protein or a complex by introducing covalent links between two amino acids that are close in space (but can be far in sequence or belong to different proteins). After a subsequent digestion of the protein, where the chemical links are remaining intact, the linked parts can be identified by MS. This provides information about relative arrangements between domains, subunits, or multi-protein complexes.[44]

Hydrogen-Deuterium exchange coupled with Mass Spectrometry (HDX-MS) can determine the overall deuterium content of molecules that have undergone H/D exchange. The experiment gives information about the solvent accessibility of various parts of the molecule. Thus, it sheds light not only onto the overall tertiary structure of the protein, but also on the degree of accessibility (as revealed by the speed of exchange) and therewith involved dynamic aspects, such as protein conformation and stability shifts. HDX-MS can also be used to examine protein-protein and protein small molecule interactions and has therefore potential in drug development.[45]

For structure determination of small molecules, X-ray crystallography is the most widely employed technique (sometimes also NMR), but the rather new MicroED technique is becoming more and more popular, especially when submicrometer-sized crystals are a limiting factor.[46,47]

Not only the atomic molecular structures, but also the characterization of the dynamic interaction between compound and target (often in form of affinity measurements) represents the basis for structure-based drug design on which further strategies can be built on and which is therefore of paramount importance in the process of drug development.

1.1.4.3 Affinity measurements: ITC, TSA, SPR, fluorescence & reporter assays

One very accurate and commonly used method for measuring affinity is **Isothermal Titration Calorimetry (ITC)**. Besides binding affinity (in form of a K_d), it also provides information on reaction stoichiometry. ITC measures the direct heat that is exchanged between interacting molecules

and the solvent. One disadvantage is the high amount of (target) protein that is required that may become limiting when protein production is not very abundant.[19,48]

Another technique that has become popular for ligand screening is fluorescence-based **Thermal Shift Assay (TSA)**. This method assumes that protein stability will increase upon ligand binding. The thermal stability of the target protein is tested by measuring the critical melting temperature (unfolding occurs). The unfolding is reported by environmentally sensitive dyes that emit fluorescence upon binding to exposed hydrophobic parts of a protein. Not only different ligands, but also different buffer conditions can be screened, which may come handy for crystallization. TSA is a versatile technique with a broad range of applications and mayor advantages are low cost, easy employment, and high-throughput capabilities.[49,50] TSA already gives insights in thermodynamic aspects of the protein structure, but an accurate quantitative analysis of protein-ligand interactions, which is essential for drug design, cannot be performed. As indirect measure, it only provides a K_d proxy.

Surface Plasmon Resonance (SPR) is an optical sensor technique that measures changes in optical reflectivity when molecules bind to a functionalized surface. It is a label-free method and allows for monitoring interactions between a large variety of different molecules and can be applied to drug discovery in many ways, such as ligand screening (especially on membrane proteins), or monitoring reaction rates. It does not only provide an affinity measure (K_d), but also the rate of association (k_{on}) and dissociation (k_{off}), which are highly valuable in drug development.[51]

Other approaches for ligand binding affinity measurements are fluorescence-based methods, such as **Tryptophan fluorescence Quenching (TQ)** and **Förster/Fluorescence Resonance Energy Transfer (FRET)**. In TQ binding affinity is calculated based on decreasing tryptophan fluorescence while increasing ligand concentration, and therefore depends on the presence of a tryptophan close by the ligand binding site. FRET is based on energy transfer between two light-sensitive molecules (chromophores) and extremely sensitive to small changes in distance, but requires labeling of protein and ligand with those chromophores, one being the donor that is excited by a laser, and the other the acceptor that emits light when in close proximity (<10nm) to the donor. Moreover, Time-Resolved FRET (TR-FRET) kinase activity assays have also been developed that use a labeled peptide substrate, which upon kinase activity gets phoshorylated, and can then bind the second fluorophore required for the FRET signal to occur.

Cell-based reporter assays are widely used tools in drug discovery applications and particularly useful for evaluating a drug's inhibition strength on a targeted enzyme within the cellular context. Enzyme activity is linked to the expression of a reporter, which can then be detected. The inhibition or activation strength of a drug can be evaluated by measuring the altered reporter signal compared to a reference ligand. Reporter gene assays, such as the luciferase reporter assay, have been increasingly used in high throughput screening to identify small molecule inhibitors and activators of protein targets.[52,53]

Commonly reported **affinity measurements** for enzymes are the *inhibition constant* K_i and *IC50*. The K_i is the dissociation constant of the inhibitor and indicates its binding affinity. It is usually measured

directly (by various methods) as the concentration of inhibitor needed to occupy 50% of receptors or to determine the rate of an enzyme-catalyzed reaction it usually requires multiple measurements (independently varying the concentration of substrate, and the concentration of inhibitor).[54] IC50 represents the functional strength of the inhibitor. It can be measured using a functional assay or a competition binding assay. In a functional assay it is measured as the concentration needed to inhibit 50% of the maximum biological response of a reference ligand using a dose-response curve. In a competition binding assay it is the concentration needed to displace 50% of a reference ligand (e.g. radioligand) and determined with a competition curve. In case of a competitive inhibiting compound the relationship of IC50 and K_i is stated by

$$IC50 = K_i \left(1 + \frac{[S]}{K_m}\right), \tag{1.5}$$

where $[S]$ is the substrate concentration and K_m the Michaelis constant of the substrate (S).[54,55] Obtaining an IC50 value requires less effort than a K_i, since it is determined at only one concentration of substrate over a range of inhibitor concentrations. Thus, IC50 is the most commonly used target activity metric and the IC50 datasets for a given target are usually much larger than the respective K_i dataset. Nonetheless, it has to be kept in mind that IC50 is a relative value, whose magnitude depends on the concentration of reference ligand/substrate used in the assay, whereas K_i is usually a constant value for a given compound with a target.

For the development of prediction models special care has to be taken about data quality. Therefore, more direct measures, such as the K_i, with supposedly lower measurement noise are often preferentially chosen when the aim is to find out the most informative features for predicting the given property. Nevertheless, a larger dataset size could be beneficial for extensive testing of developed prediction models.

Key points

⇒ All experimental techniques have their limitations, but they can already provide a wealth of information that can be used for drug design.

⇒ The limitations and the complementarity of the techniques highlight the benefit of combining the resulting data, but the heterogeneity may be a major challenge.

⇒ As conclusion, the development of further computational methods for integrative molecular modelling may be advantageous.

1.2 Computational methods for screening and affinity estimation

In order to reduce time and cost consumption within the drug development process computer aided methods have been implemented, which are often used for Hit identification and also for Lead optimization. These include virtual screening (VS) campaigns, target modelling at different levels of complexity and in later stages also computationally more expensive methods such as, molecular dynamics (MD) based methods. Additionally, there is the need to check for chemical feasibility of designed molecules, since real and virtual molecules are simultaneously under study for drug design.

Nowadays, virtual screening plays a major role in the process of drug development, as it enables a fast evaluation of hundreds or thousands of small molecules to select a smaller number for biological testing. Virtual screening can be used to select compounds for screening from in-house databases, to choose compounds to purchase from external suppliers, or to decide which compounds to synthesize next. Structure-based and ligand-based virtual screening are two main techniques, which are widely used.[56,57] Depending on the available information about the target or existing ligands and the aim of research the applicable method is chosen.[58,59] To employ computational techniques, a description of the compounds (and possibly the target) is required.

Molecular representations

To understand the properties of molecules, it is important to find common patterns, and in order to find these patterns, the molecules must be represented in such a way that a degree of similarity can be calculated. There are many different ways to represent molecules, e.g. by chemical graphs (graph-based representations), using connection matrices or tables, or with line notations systems, such as SMILES strings. SMILES stands for Simplified Molecular Input Line Entry Specification and provides a simple syntax, containing each (non-hydrogen) atom with its element symbol and additional constructs to specify charges, bond orders and stereochemistry. Starting from the SMILES sting already a lot of molecular descriptors can be calculated, and in order to be able to compare those among different molecules, or perform further calculations or analysis, fixed length descriptors are often required. Thus, common descriptor representations of molecules are in the form of vectors (see Figure 2 in Section 1.2.6).

1.2.1 Ligand-Based Virtual Screening (LBVS)

The basis of medicinal chemistry efforts and of all ligand-based virtual screening (LBVS) methods is founded on the similar property principle, which states that structurally similar molecules tend to have similar properties - despite the existence of "activity cliffs".[60,61] LBVS methods are based on analyzing features of substructures and chemical properties related to activity of the ligand.

They are used to extract similar compounds from libraries. Different approaches are similarity and substructure searching[62] that may be based on pharmacophores, shape-matching,[63] or fingerprints.

Starting from one or several active molecule(s) **similarity search** may be a fast and relatively straight forward approach to get first ideas about promising molecules (that may already have been synthesized). Similarity-based virtual screening requires a way to evaluate the similarity of a pair of compounds and is dedicated to rank order a database of compounds on similarity to a given active reference structure. The top ranking compounds can then be selected for biological testing. One major issue is that similarity is inherently subjective. Consequently, many different similarity evaluation metrics have been proposed to provide a quantitative basis for structure ranking. One limitation is that as "similar" considered molecules do not always share identical pharmacological profiles.

A **pharmacophore** is defined by the IUPAC as an "ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response".[64] Such a model can be used to compare the compatibility of candidate ligands with it in order to decide whether the ligand could be a potential binder.
Shape-matching approaches are often combined with pharmacophore features and are based on the superposition and comparison of the 3D shapes of a known binding molecule and a set of ligands in question. Therefore, the selection of the initial query ligand represents one of the crucial and most challenging tasks in the shape-matching approach.[65] Usually, the conformation of a ligand being present in the target's crystal structure is used. If no structural data is available, computational methods, such as docking, are needed to obtain a ligand pose. Moreover, when dealing with structurally diverse or highly flexible compounds the structural alignment of ligands can be particularly challenging.[66] To overcome the limitations of employing a single model in LBVS, ensemble methods could be a choice.

1.2.1.1 Molecular descriptors

Molecular descriptors are basically numerical values assigned to molecular structures or substructures. Many different molecular descriptors have been proposed with the aim to find the most information rich and and best suited descriptor to model a certain property. A classification of the large amount of diverse descriptors can be done based on dimensionality (number of geometrical dimensions the captured information takes into account) and/or based on the type of information they contain. Thus they are usually classified as 0D, 1D, 2D, 3D, and 4D descriptors, or as constitutional, topological, geometric, electronic, physicochemical, or quantum-mechanical descriptors. 0D descriptors take into account the molecular formula and are sometimes referred to as count descriptors, including for example atom or bond counts and molecular weight. 1D descriptors include fragment information (groups of atoms within a molecule). 2D descriptors are based on the chemical graph and use information from the atomic connectivity tables/matrices, e.g. topological radius or diameter. 3D descriptors include information about the 3D geometry of a molecule, such as shape descriptors, volume descriptors, or descriptors that require 3D coordinates (positions in space). As molecules are flexible, 3D descriptors have been extended to 4D descriptors, adding for example different conformations of a molecule as fourth dimension.

1.2.1.2 Molecular fingerprints

Molecular fingerprints are vector representations of molecules. They can be fixed in length, or not, and their way of construction may differ. In a bit format, either each bit corresponds to one pre-defined feature (also termed structural keys), or instead several substructures may correspond to the same bit. They can be binary (containing only zeros and ones), or may be extended to a count format, where the vector of bits is extended to a vector of integers, accounting for the amount of occurrences of the features. The common carateristic for all fingerprints is that the positions in the fingerprint sequence are used to refer to specific features of the molecule.
A representative for **substructure fingerprints** that map pre-defined substructures to a certain bit and have therefore a pre-defined fixed length are the commonly used MACCS (Molecular ACCess System) keys fingerprints.[67]

For many other fingerprints the setting of the bits is performed by a hashing algorithm that translates a feature to a bit position. Thus, they are often referred to as **hashed fingerprints**. As a result, a given feature will always set the same bit, but multiple paths may also simultaneously set the same bit. Therefore, similar molecules will have similar fingerprints, but similar (hashed) fingerprints do not necessarily mean that molecules are similar.
Fingerprints can be constructed directly from the molecular graph, where a bit corresponds to one or more paths in the graph. They are termed **path-based fingerprints**. For example, a path of length two consists of three atoms that are connected by two bonds. The bits are usually set by a hashing function.
Circular fingerprints, represent the molecules as sub parts, starting from a atom and looking at surrounding atoms, and are also constructed using hashing functions. The Extended-Connectivity Fingerprints (ECFPs)[68] are the most well known and most commonly used circular fingerprints, with e.g. ECFP4, where the number 4 corresponds to the diameter of the atom environments considered.

In general, 2D fingerprints are very good at identifying close analogues, but they also have their limitations. One major limitation of traditional 2D descriptors is that they are usually not suited for "scaffold hopping" - the identification of structurally novel compounds by modifying the central core structure of the molecule. Scaffold hopping can be desired or required in different situations, for example due to patent reasons, to move away from competitor compounds, or to provide alternate Lead series if problems arise due to difficult chemistry or poor ADME properties. Descriptors that can be used for scaffold hopping are for example reduced graphs, topological pharmacophore keys, or 3D descriptors. Another limitation of descriptors is that they describe the molecules as static model, while many chemical phenomena are not static and may depend on the local environment. Examples are changes in geometry, in ionization state, or protonation state.

1.2.1.3 Similarity coefficients

Similarity coefficients are used to calculate the score of similarity between two molecules representations, which is often an inverse of a measure of distance in descriptor space, so that the greater the degree of similarity between two molecular representations the smaller the value of the coefficient. The Tanimoto coefficient[69] is the most commonly used coefficient to quantify similarity between two sets of binary fingerprints and ranges from 0 to 1. It calculates the similarity of two bit strings as the size of the intersection divided by the size of the union:

$$\text{similarity}(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1.6}$$

where $|A \cap B|$ are the number of bits in common for both fingerprints, and $|A|$ and $|B|$ are the number of bits set in each fingerprint individually. Many other types of similarity coefficient exist that can be applied, e.g. cosine coefficient, Euclidean distance, Manhattan distance, or Tversky index, but they are rather rarely used. In case of non-binary data more complex forms of similarity measures are used, such as physicochemical property vectors.

1.2.2 Quantitative Structure-Activity Relationship (QSAR) modelling

If actives and inactive molecules for a given target are known, Quantitative Structure-Activity Relationship (QSAR) modelling can be performed. In a nutshell, QSAR modelling uses knowledge of known active and known inactive compounds to build a predictive model based on quantitative "activity" data while employing machine learning methods. QSAR modelling can be used with data consisting of diverse structural classes and multiple binding modes. In a standard QSAR approach the biological activity of several compounds are studied using a biological assay whereupon a mathematical relationship is created between the measured activity and the compound's structure. To establish such a relationship, also called model, data analysis and machine learning is performed.

1.2.2.1 Important considerations for data gathering and preparation

The standard procedure for building QSAR models starts with the assembly of a library of chemical compounds. Already when selecting the compounds several considerations have to be taken into account, such as availability of the molecules, or price and effort for molecule synthesis and testing, the required chemical space and diversity of the molecules, and quality of data if activity measures are provided (e.g. from an existing database). All these aspects have to be taken into account carefully from the beginning, because the data setup heavily influences the predictability of the final model. QSAR models can also be built in an iterative way, first covering a rather large chemical space and then new compounds are designed based on results from previous designs,

often narrowing down the chemical diversity.
If activity data is not already available, the compounds are tested for their activity using a biological assay. Concerning the choice of assay, one has to consider the assay's variability and the response space. For example, there is a huge variability between assays performed in different laboratories, but even in the same laboratory performed measures on different days or different time points have a certain variability, and assays on large plates can further show inter-plate drifts so that the position of the sample on the plate affects the result. The variability may be caused by differences in assay conditions and the experimental protocol, in particular buffer composition, and by the stability of used reagents and proteins. It is also worth checking the response space of an assay, because if all compounds are about equally active due to high structural similarity, it will be difficult to construct a meaningful model, and therefore, a range of different responses among the compounds (including "good" and "bad" ones) is much more beneficial for a well performing model. The response of an assay can be a continuous scale, such as the binding affinity for a target, or categorical, such as "active" and "inactive". Therefore developed QSAR models are either regression models (for continuous data), or classification models (for discrete classes).
Next, chemical descriptors need to be selected, calculated, and (if required) pre-processed for the compounds. The selection of descriptors depends principally on the desired usage of the resulting model. The type of available descriptors ranges from very simple, such as molecular weight or bond counts, to very complex, which take into account the 3D structure and environment. Simple descriptors are usually preferred when the goal is to develop very large and global QSAR models based on thousands of compounds, as they are more apt to uncover general trends common to compound subsets than uncovering detailed information about compound interactions (that they would not be capable of doing anyway). In these cases, complex descriptors would only require more computational effort without much benefit. In contrast, when the goal is to develop a local QSAR model focused on a smaller set of compounds, the usage of sophisticated descriptors may enable the study of finer details. This is for example the case when compound properties are to be improved. Here descriptors are needed that refer to details of the chemical structure to suggest molecular changes. Different pre-processing steps may be required depending on the nature of the data (descriptors and responses) and the algorithm to be used for modelling afterwards. This includes treating potentially missing data, transforming data, centering and scaling, and can even extend to data engineering (creating new descriptors, e.g. by combining several ones). For example, logarithmic transformations are often performed when values span several orders of magnitude, and centering and scaling is used to avoid the "artificial" over/under-weighting of the importance of a certain descriptor that may simply occur due to different value ranges.
Before passing to the actual training of the QSAR models the data needs to be split into training and testing set, whereas the testing data is not used for model building. Setting away data for external validation is important in order to estimate the models ability to predict the activity of new molecules correctly. The fraction of the data that is set away depends on the size and the homogeneity of the dataset and is often about 15-30%. Additionally, when activity is sampled an even distribution of activities is desired in the training and the test set. This is obtained by stratified sampling, which first attributes the compounds to activity bins before randomly distributing them from each bin to training and testing set.

1.2.2.2 Model building and validation

QSAR model building aims to provide an equation describing the relation between the molecular descriptors and the biological activity. Many different algorithms are available via different programming languages (common ones being R and Python) and also implemented in different tools. Popular linear algorithms are for example partial least squares and ridge regression, and among the non-linear algorithms support vector machines, random forest, and neural networks are commonly used. In general, models of low complexity will only be able to explain partially the response, whereas too complex models will result in overfitting. On top of that, all algorithms have parameters that need to be adjusted to control model performance. This is usually done with the help of internal cross-validation, where the data is split into groups, the model is trained on on the data leaving out one group, and the left out group is used for validation. This process is repeated until all groups have been left out once. For each round the parameters can be modified, which finally gives the most suitable set of parameters for modelling the data. Moreover, the model's performance can be estimated using different metrics, the most frequently used ones are the dimensionless coefficient of determination (R^2), which represents the fraction of the explained variance of the predicted variable, and the root mean square error (RMSE), which is in the units of the measured activity and therefore useful to compare the accuracy of the model in relation to the measurement error.
Upon careful validation a QSAR model can be used for predictions and/or interpretations and serve as ligand-based virtual screening tool, as it is possible to assess large chemical libraries due to the high computation speed. However, a QSAR model can also be used to uncover particular properties that are important for a biological effect.
A QSAR model that is built from a well-designed/balanced dataset is supposed to cover the required chemical space well and has higher chances to be predictive within the covered descriptor range. Nevertheless, it is important to keep in mind that QSAR model is only suited to make predictions within (or very close to) the covered descriptor space, which is generally known as applicability domain.

1.2.2.3 Applicability domain

When training a QSAR model, validation is performed internally, by evaluating its performance for predictions within the dataset. When the descriptor values of the molecules to be predicted are within the descriptor space covered by the training set, also called interpolation, predictions are rather reliable. However, new compounds can have descriptor values that are outside the descriptor space covered by the training set. In this case extrapolation is needed. While the developed model may still perform well for compounds that are relatively similar to the training set, it is likely to fail for compounds that are more different. Therefore, it is important to know the limits of each model and to define whether the model's assumptions are met. This is attempted by analyzing the applicability domain, which should finally help to decide whether a QSAR model can be used for a given set of compounds.[70] While the term is conceptually easy to grasp, it is difficult to define. The characterization of interpolation space is important for defining the applicability domain. Thus, it is often defined using either the chemical similarity of compounds (the chemical space) or a similarity

measure that is based on descriptors (the descriptor space). The restrictions imposed by the applicability domain (primarily impacting pure ligand-based approaches), can be partially circumvented by approaches that are based on the target structure, such as structure-based virtual screening.

1.2.3 Machine learning algorithms

Nowadays, machine learning has many applications in virtual screening and ligand-based approaches, as well as structure-based docking have benefited from machine learning algorithms.[57,71,72] Here, the focus is on supervised techniques that require a training set for model development. The choice of a machine learning algorithm depends on many factors, such as the envisioned goal, the complexity of the problem, the nature of the data, the required accuracy, and the degree of required interpretability. Many non-linear methods, such as neural networks and support vector machines are very performant and high accuracy of predictions can be obtained, however they are rather suited for a "black box" employment mode, as they hardly allow for simple interpretations on how a certain result was obtained. Linear models are more straightforward to interpret, but they usually show sub-optimal performance if the relationship between the descriptor and the predicted activity is not linear. Decision trees are also very useful for determining variable importance. When combining single decision trees to larger ensembles, as done by the random forest algorithm, the predictive power can be largely increased, with a trade-off in interpretability. Choosing the most suitable algorithm among the multitude of available ones is therefore not an easy task and often several ones are tested. The random forest algorithm and support vector machines are popular representatives and are employed within this thesis work, and thus explained in more details below.

1.2.3.1 Random Forest (RF) and other tree-based algorithms

The Random Forest (RF) algorithm is based on decision trees. The goal is to predict the value of a target variable by learning simple decision rules inferred from the data features. The big advantage of decision trees are that they are simple to understand and to interpret and that they can be visualized.[73] A decision tree represents the conjunction of a series of "rules". Decisions are taken at each interior node to follow one branch or another. When visualized, its branching path structure resembles an inverted tree. Each interior node corresponds to one of the input variables, for each of the possible values of an input variable there are edges to children, and each leaf represents a value of the target variable. Difference between classification and regression trees are found in the procedure used to determine where to split.
A RF is an ensemble method, more precisely an ensemble of decision trees. Each decision tree is constructed by using a random subset of the training data. Furthermore, during the construction, at each node of a tree, a small group of input variables is selected at random to split on. The variable that provides the best split, according to an objective function, is used to do the split on that node. At the next node, another set of variables is chosen at random from all variables and the process is continued. The single trees are grown to maximum size and they are not pruned. In terms of

the ensemble method, the single trees are considered as the weak learners and the RF (comprising all trees) is the combined strong learner. Therefore a RF has no visualization and is not as easy to interpret as a single decision tree. Finally, when a new input is tested on the constructed RF, it is run down all of the trees and the final result is the average of all of the terminal nodes that are reached in case of regression, and the majority vote in case of classification.[73]
For the RF machine learning model implemented in the R package *caret* two parameters can be varied to improve performance. The varied parameters are the amount of trees that are produced (*ntree*), and the *tuneLength*. Upon generation of a candidate set of parameter values (by the train function), the *tuneLength* argument controls how many of them are evaluated. Additionally, the *random seed* is used to control the randomness in order to assure reproducible results. Since RF is a stochastic algorithm, the seed is used by the random number generator. It utilizes random numbers during the phase where parameters are estimated and also for choosing the resampling indices.

Examples for other tree-based ensemble algorithms (with more tunable hyperparameters) that are used within this Thesis work are regularized Random Forest (rRF), global regularized Random Forest (rRFglobal), Extreme Gradient Boosted Trees (xgbTree), and Extreme Gradient Boosted Trees with dropout (xgbDART).
In order to find the best set of hyperparameters, different techniques can be applied: exhaustive grid search, random search, or Bayesian optimization. In the presented work, Bayesian optimization was employed to select the best hyperparameters (5 to 7 depending on the method), which demands a substantially increase in computational expense compared to the one-variable optimization required for the RF algorithm.

1.2.3.2 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are considered to be among the best supervised learning algorithms. The basic idea of support vector machines is to find an optimal hyperplane for linearly separable patterns. In order to do so, the margine of the hyperplane, which is the distance to the nearest training-data point of any class, is maximized. Support vectors are the coordinates of individual observations.
For patterns that are initially not linearly separable a transformation of original data is used to map the data to a higher dimensional space in order to gain a linear separation. The transformations are performed using kernel functions, more precisely, every dot product is replaced by the kernel function in the resulting algorithm. Therefore, the classifier is a hyperplane in the transformed feature space, even though it may be nonlinear in the original input space.[73]

- SVM with Linear Kernel (SVM_L): (method = "svmLinear") This is the basic SVM form ($x^T x_i$, where x and x_i are vectors in the input space), also considered without kernel, since no transformation is performed.

- SVM with Polynomial Kernel (SVM_P): (method = "svmPoly") The polynomial kernel has the form: $(x^T x_i + c)^p$, where $c \geq 0$ is a parameter trading off the influence of higher-order versus lower-order terms in the polynomial, and power p, a tunable parameter (often set to 2).

- SVM with Radial Basis Function Kernel (SVM_R):] (method = "svmRadial") also called Gaussian, has the form $exp(||x - x_i||^2/(2\sigma^2))$ as kernel, with σ as tunable parameter. It is one of the most popular kernel functions.

1.2.4 Structure-Based Virtual Screening (SBVS)

Structure-based virtual screening (SBVS) can be performed if the structure of the macromolecular target of interest and also a chemical library is available. Its aim is to predict whether (or not) a molecule binds to a protein using 3D information. For probing the potential interactions of ligands towards the target receptor *in silico* docking of these ligands into the macromolecule is performed. SBVS can be used to predict the binding mode of drugs, to define the important specific interactions between ligand and target and finally also to discover a way to improve the drug by guiding further derivatization to optimize specificity and/or affinity.

SBVS includes the docking of candidate ligands into a protein target and a following evaluation of the likelihood of binding in this pose using a scoring function. Therefore, it consists of 2 parts: The first part, the search algorithm, is supposed to generate "poses" (comprising conformation, position and orientation) of the ligand within the active site.The second part, the scoring function, is supposed to identify the most likely pose for an individual ligand and to assign a priority order to a set of ligands. In order to do so, a pseudo energy score is calculated that estimates the binding affinity between protein and ligand.

The major problem in this approach is the fact that it involves many degrees of freedom (rotation, conformation) and also solvent effects that should be taken into consideration. On the other hand, to be able to perform VS on hundreds of ligands each docking needs to be very fast. In order to reach the required performance the complexity and the computational cost to calculate protein-ligand binding needs to be reduced significantly. This results in a trade-off between speed and accuracy.[74] Since accuracy is critical for successful VS,[75] one of the biggest limitations in molecular docking lies in the scoring function and in particular in its accuracy, because it relies on several assumptions and simplifications.[76]

One of the biggest limitations in the applicability of structure-based methods is the lack of an experimentally determined target structure.[77] Moreover, often several structures are required in different biologically relevant conformations, to sample sufficiently the protein's conformational space. Structures for many therapeutically relevant target receptors remain unavailable despite major advances in structure solving techniques (such as X-ray crystallography or NMR). Therefore, if comparative (homology) crystal structures of the protein target (structures of homologous proteins) are available, they are used to model the actual target and are also useful to complete or refine partial structures (such as loops).

3D structure modelling for targets

Homology modelling relies on the observation that the structural conformation of a naturally folded protein is more highly conserved than its amino acid sequence, and that small changes in sequence

usually result in only small changes in the 3D structure. In cases where no homologous structure is available *ab initio* modelling has to be performed. Nonetheless, the model needs a degree of accuracy that is sufficient to obtain reliable results for the subsequent docking. In general, if the sequence identity is greater than 70%, the predicted model is very accurate producing good and highly reliable results in VS. In the range between 35-70% of sequence identity VS is still feasible, but a more detailed analysis of the binding pocket is recommended. Schafferhans and Klebe[78] and Oshiro et al.[79] for example found that a 40% sequence identity was needed in order to obtain reliable results in VS. Below 35% of sequence identity VS results are not very reliable and moreover, the 20-30% homology range is also called "twilight zone", because the quality of those homology models may vary widely.

3D structure prediction for ligands

On the ligand side, 3D structures need to be predicted based on the information provided by the databases containing the small molecule libraries. They provide valuable information to understand physical, chemical, and biological properties of small molecules, including how they interact with other molecules. Moreover, low-energy conformations for small molecules are important for many molecular modelling and drug design methods. Concerning SBVS methods the initial conformation submitted to a docking program can have an impact on the docking result (on the pose generation), since docking programs attempt to sample efficiently the conformational space, but cannot perform an exhaustive conformational search. There are various tools available (free ones and commercial ones) that are able to calculate 3D structures based on the often provided molecular string format "SMILES", such as OpenBabel,[80] Frog2,[81] RDKit,[82,83] Balloon,[84] and COSMOS[85] as free examples, or CORINA[86] and OMEGA[87] as commercial ones. Additionally, small molecule databases may provide their own 3D generator. For example BindingDB offers a download option of 3D conformations generated by VConf (and partial charges generated by VCharge).[88] To perform 3D structure prediction for ligands in a high-throughput mode basic generators are often data-driven tools that use libraries of fragment and torsion angle parameters. As this approach may lack accuracy in some cases (e.g. for very complex molecules), dedicated optimization tools for small molecule 3D structures have been developed and are sometimes already implemented in the generators. Still, the generation of conformations for small molecules represents a problem of continuing interest and tools are under constant development for sampling the conformational space and to score conformational stability.[89]

1.2.4.1 Inverse virtual screening

Inverse (or reverse) VS involves the docking of a ligand (or a few ligands) against an array of protein structures. In contrast to classical SBVS, which aims to find specific ligands, for inverse VS the target space serves as filter. Here, the ligand(s) is/(are) used to select the ligand-specific protein targets.[59] Docking a ligand to many proteins can lead to the identification of targets with a shared and/or specific activity. Therefore, by browsing the target space inverse VS can be regarded as a complementary tool to VS, which can reveal information about common features of the target structures. Consequently, inverse docking is a useful approach to investigate the underlying molecular mechanism of a biological effect.[90] Inverse docking also involves challenges. First of all, a

panel of target structures needs to be available (or modeled), into which the ligands can be docked. Furthermore, proteins often exist in several closely related isoforms. To rank these small differences might be a challenging task for the scoring functions, since they are mainly trained to rank different protein-ligand complexes and many ligands against a smaller number of proteins.[90] Additionally, the panel of structures still represent an incomplete description of most targets, as flexibility (as coverage of the conformational space) is often not considered in an extensive way.

1.2.4.2 Ligand Flexibility in SBVS

Since the binding process can also involve intrinsic conformational changes of the ligand and the receptor, sampling is a fundamental challenge for protein-ligand docking methods. How to best model flexibility is still an open question: whether the bound conformation of the ligand should be sampled and therefore predicted prior to or during docking. Nevertheless, the way and also the extend to which docking algorithms explore this conformational space differs between different docking software.[91]
In molecular docking the scoring function has two tasks: The first is to assist the docking program to efficiently explore the binding space of a ligand, which partially mirrors the ligand flexibility, and the second is the evaluation of the binding affinity once the correct binding pose is identified. In general, scoring functions can be roughly classified into three types: (i) Force field-based scoring functions, which employ a classic force field to compute the noncovalent ligand-target interactions; (ii) Empirical scoring functions, in which regression or machine learning methods are used to compute for example the binding affinity and (iii) Knowledge-based scoring functions, which evaluate the interactions between the ligand and the target as a sum of distance-dependent statistical potentials by assigning energy-like terms to the structural features of protein-ligand interactions depending on their occurrence frequency.
Concerning the coverage of the conformational space, docking programs are able to generate ligand conformations very similar to the crystallographic one and correctly identify active molecules.[92] Scoring functions are less successful at ranking first the correct binding mode, success rates are mainly target-dependent and moreover, useful ligand binding affinity predictions represent a main challenge, since there is often weak correlation found between docking scores and measured ligand affinity.[92] The latter limitation is most likely due to the large number of approximations used by docking scores to improve computation efficiency. For instance, there is usually no term that accounts for the full entropic contribution on the binding event.

1.2.4.3 Target Flexibility in SBVS

In the physiological cellular environment proteins are dynamic, which is often crucial for their function and activity. The protein binding pocket often adapts to accommodate an entering ligand or a certain conformation of the conformational space of the protein is stabilized by the bound ligand. Erickson et al[91] showed that docking accuracy falls off dramatically if an "average" or apo structure is used instead of an experimental crystal structure with a bound ligand, suggesting that

the binding event itself introduces important movements. Those conformational changes can range from minor movements of single side-chains to large shifts of whole secondary structures or even domains. Therefore, the experimentally obtained crystal structures can currently be regarded as static snapshots of the dynamic conformational space of the protein. Nonetheless, this static view can be extended by ensemble refinement, by subsequent molecular dynamics simulations, or by combining multiple experimental structures.

Docking involves a trade-off between the speed of the docking algorithm and its accuracy. Therefore, by adding flexibility to the protein structure used for docking higher accuracy can be achieved, but this also adds noise to the computation and increases the computational cost intensively. Especially in large-scale virtual database screening this comes into play, since due to the high number of compounds to be screened there is a practical limit of available computational time per compound.[90] Unfortunately, the degree of required flexibility is not known beforehand for new ligand types. All this underlines the fact that target flexibility represents one of the greatest challenges for docking programs.

One way to circumvent the problem of small rearrangements is to perform "**soft docking**". Implemented in several docking programs, it allows a small overlap of the ligand and the receptor by reducing the actual volume of the atom spheres and therefore avoiding VdW clashes.[93] Unfortunately, this could introduce errors like the detection of false positives, as affinity cliffs may be ignored, and it also does not account even for slightly larger conformational changes like side-chain rotations.
But nowadays, the ability to include **side-chain flexibility** (for a limited amount of side-chains) by using libraries of preferred conformational states (e.g. sets of torsion angles) is fortunately implemented in several docking programs, such as PLANTS,[94] which is used in this study. As for ligand flexibility, the same question, whether the conformational space should be sampled prior to or during the docking process, persists with respect to target flexibility.
A way to take into account the flexibility of a macromolecular structure prior to docking is to built an ensemble of static models, called "**ensemble docking**". Such an ensemble of structures can be composed of several available crystallographic structures, and since more recently also of a crystallographic structure refined as ensemble.[95] Conformation ensembles can be generated computationally, for example, by molecular dynamics (MD) simulation.[96,97]This should avoid a bias towards one protein conformation while implicitly including protein flexibility. Therefore, especially long MD simulations (or advanced sampling methods) could help to sample the conformational space of the receptor prior to docking.[58] The difficulty in ensemble docking lies in the selection of appropriate target structures e.g. from a MD trajectory. Since it is not possible yet to simulate large macromolecules on a long enough time scale that is relevant for domain movement, new attempts are needed to obtain a set of structures which represents the important conformational space a protein adopts during activation or activity. One attempt to select relevant target conformations is normal mode analysis, which has been demonstrated to be an effective tool,[98] depending on the amplitude of the observed movement. Unfortunately, additional noise is introduced by each extra conformation added to an ensemble, which may mask the beneficial information it provides. Generally, docking results are very sensitive to small variations (in protein side-chain and ligand positions), which is demonstrated by the fact that re-docking in the same structure from which ligands were extracted gives usually better results than cross-docking in different conformations of the same target.[99,100] Therefore, the choice of the most appropriate receptor conformations is key for the success of the VS experiments and for the results to be representative. This problem highlights

the need for clear guidelines to select the structures that should compose an ensemble.

1.2.5 MM-PBSA - a Molecular Dynamics based method

Molecular mechanics energies combined with the Poisson-Boltzmann or generalized Born and surface area continuum solvation (MM-PBSA and MM-GBSA) are widely used techniques for binding affinity estimation, which are not based on any training dataset. They can be applied even on single conformations, but are usually applied on sets of structural conformations (that simultaneously provide error estimations). To produce these conformations Molecular Dynamics simulations are commonly employed.

1.2.5.1 Introduction to Molecular Dynamics (MD) simulations

Molecular Dynamics (MD) simulations are increasingly employed to study biological molecules of biomedical interest, in particular in the drug discovery field, where they are being more and more used.[101–103] MD simulations have been combined with a wide variety of different approximations to study mobility related effects, such as the impact of protein motions on catalytic activity[104] and binding of ligands.[105–108]

The central ideas in MD simulations are that biological activity is the result of time dependent interactions between molecules, that macroscopic observables (as observed in laboratory experiments) are related to microscopic behavior on the atomic level, and that the microscopic behavior of a molecule can be calculated by MD simulations. Major mistakes that can be made by performing a computational experiment, such as a MD simulation, are very similar to the ones when performing a wet-lab experiment - e.g sample preparation has not been performed correctly, the measurement is not long enough, the system undergoes an irreversible change, or the measured quantities do not correspond to what one thinks. Advantages of MD simulations are that they allow the prediction of static and dynamic properties of molecules directly from the underling interactions between the molecules and they permit to gain insight into situations that are impossible or difficult to study experimentally, especially with atomic resolution and on short time scales (pico to micro seconds).

In standard MD simulations a molecule is described as a series of charged points (atoms) linked by springs (bonds) and potential energy functions model the basic interactions. MD simulations solve Newton's equations of motion (for a system of N interacting atoms), which implies the use of classical mechanics to describe the motion of atoms:

$$F_i = m_i \cdot a_i = m_i \cdot \frac{\delta v_i}{\delta t} = m \cdot \frac{\delta^2 r_i}{\delta t^2}, \text{ with } i = 1...N \tag{1.7}$$

MD simulations calculate the motion of the atoms in a molecular assembly using Newtonian dynamics to determine the net force and acceleration experienced by each atom at a given time t. Each atom i at position r_i, is treated as a point with a mass m_i and a fixed charge q_i. The integration of the equations of motion gives the initial structure with an initial distribution of velocities $v(t_0)$.

Starting from the initial coordinates a trajectory is calculated with positions as function of time. The potential energy is a function of the positions, so the acceleration, and since the positions vary as a function of time, so does the acceleration. Temperature is related to the microscopic description of simulations through the kinetic energy and the kinetic energy is calculated from the atomic velocities. Different algorithms can be used to integrate Newton's equations of motion. The most common one is the Verlet algorithm (or slight variations of it) that solves the differential equations numerically at discrete time steps (usually 2 fs) to determine the trajectory of each atom.[109]

A force field is built up from the set of equations (called the potential functions) used to calculate the potential energies and their derivatives, the forces, and the parameters used in this set of equations. The potential functions that are used to model atomic interactions can be subdivided into three parts: 1. non-bonded (Lennard-Jones or Buckingham, and Coulomb or modified Coulomb); 2. bonded (covalent bond-stretching, angle-bending, improper dihedrals, and proper dihedrals); and 3. restraints (position restraints, angle restraints, distance restraints, orientation restraints and dihedral restraints). Several force fields are commonly used in MD simulations, including AMBER,[110–112] CHARMM,[113,114] and GROMOS,[97,115] which differ principally in the way they are parameterized. Besides the choice of force field a user also has to choose the representation of water molecules, the so called water model, since most biological processes occur in aqueous solution. Moreover, solvation effects play a crucial role in determining molecular conformation, electronic properties, binding energies, etc. When explicitly treating the solvent, the actual solvent molecules are added to the molecular system. An alternative to the explicit water models is to use an implicit solvation model, also termed a continuum model, where the solvent is modeled as a continuum dielectric.

Ensembles in MD simulations

Systems can be described by statistical ensembles that depend on a few macroscopically observable parameters, which are in statistical equilibrium. In the microcanonical ensemble, also called (NVE) ensemble, the system is isolated and the total energy is conserved (E), the number of basic particles is conserved (N), and there is a boundary/volume limit (V). When the simulated system is embedded in an infinite heat bath, but does not have particle exchange with this bath, it forms a canonical ensemble. In the canonical ensemble the system temperature is conserved (not absolutely constant) (T), the number of basic particles is conserved (N), and there is either a boundary limit (V) or a constant pressure (p). They are also called (NVT) ensemble or (NpT) ensemble. The isothermal–isobaric (NpT) ensemble corresponds most closely to laboratory or cellular conditions with constant temperature and pressure and is therefore frequently used.

Steps in a typical MD simulation are:

1. Preparation of the molecule(s)/system under investigation;
2. Minimization - to reconcile the system with the force field used;
3. Equilibration - to reach the desired quantities of the system (temperature, pressure, etc) and to ensure that it is stable;
4. Production dynamics - the actual simulation under desired conditions (NVE, NpT, etc) and collection of data;
5. Analysis - includes the collection of data and evaluation of observables (macroscopic level properties), or a comparison to single molecule experiments.

Intrinsic limitations of MD simulations

As classical mechanics is used in MD simulations, the force field is a function of the positions of atoms only. This means that the electronic motions are not considered (Born-Oppenheimer approximation) and electron transfer processes, electronically excited states and chemical reactions cannot be treated. Most force fields cannot incorporate polarizabilities, and do not contain fine-tuning of bonded interactions. The omission of polarizability also means that electrons in atoms do not provide a dielectric constant as they should. The subsequent overestimation of long-range electrostatic interactions is slightly compensated by the fact that long-range interactions are cut off, but this introduces its own artifacts. The classical way to minimize edge effects in a finite system (e.g. to avoid real phase boundaries) is to apply periodic boundary conditions. In order to do so, the atoms of the system to be simulated are put into a space-filling box, which is surrounded by translated copies of itself. Unfortunately, for small systems the periodic boundaries may enhance internal correlation and introduce errors.

1.2.5.2 Advantages and limitations of MM-PBSA

Free energy calculations have been shown to be useful for drug optimization, as it enables the prediction of inhibitor activity and gives insights into the drugs thermodynamic signature.[116] MM-PBSA and MM-GBSA are continuum-solvation methods to estimate the binding free energy between a ligand and a receptor to form a complex.[117–125] The GB approach is a computationally more efficient approximation to the PB theory, and thus usually less accurate. MM-PBSA has a lower computational costs, compared to free energy pathway methods, and has a more sophisticated computation of the free energy components, compared to common scoring functions. This makes MM-PBSA an attractive method for drug design. The binding free energy (ΔG_{bind}) is calculated with the following equations:

$$\Delta G_{bind} = \Delta H - T\Delta S = \Delta E_{MM} + \Delta G_{sol} - T\Delta S \tag{1.8}$$

$$\Delta E_{MM} = \Delta E_{internal} + \Delta E_{electrostatic} + \Delta E_{vdW} \tag{1.9}$$

$$\Delta G_{sol} = \Delta G_{PB/GB} + \Delta G_{SA} \tag{1.10}$$

where ΔE_{MM}, ΔG_{sol} and $-T\Delta S$ are the changes of the gas phase molecular mechanics energy, the solvation free energy, and the conformational entropy upon binding, respectively. ΔE_{MM} is the sum of $\Delta E_{internal}$ (bond, angle, and dihedral energies), $\Delta E_{electrostatic}$ (electrostatic), and ΔE_{vdW} (van der Waals) energies. ΔG_{sol} contains $\Delta G_{PB/GB}$, the electrostatic solvation energy (polar contribution), and ΔG_{SA}, the nonelectrostatic solvation energy (nonpolar contribution). The polar contribution is calculated using either the PB or GB implicit solvent model, which estimates the change in the free energy upon transfer of a charged molecule from gas-phase (modeled as a homogeneous medium with a dielectric constant ε often set to 1 or 2) to solvent (modeled as a homogeneous medium with ε=80), while the nonpolar energy is estimated by solvent accessible surface area (SASA). The configurational entropy (entropic contribution $T\Delta S$) is either estimated using a rigid-rotor harmonic oscillator approximation, applying normal mode analysis or quasi-harmonic analysis, or completely neglected if only the relative binding free energy of similar ligands shall be analyzed.

Finally, the binding free energy can be calculated as the difference between the free energy of a complex and the sum of the free energies of its components by applying a thermodynamic cycle. This difference between the free energies of the complex and its components is calculated either from a single trajectory of the complex ("single-trajectory approach") or from separate trajectories of complex, receptor, and ligand ("three-trajectory approach"). Although the singe-trajectory approach neglects the conformational flexibility of the unbound components, it is the most commonly applied approach, especially when no large structural changes upon binding are expected. It benefits from the reduced computational cost (one instead of three trajectories) and from a reduction of noise due to the cancellation of intramolecular contributions and thus, the MM-PBSA analyses can be based on shorter simulations as well.
It has been previously reported that using MM-PB/GBSA long MD simulations seem not to result in better predictions and short MD simulations can be adequate in calculating binding affinities,[126,127] and moreover that MM-PBSA can be applied to single-minimized structures instead of MD trajectories.[121,128] In order to achieve a higher precision it has been suggested to run many short independent simulations (produced by e.g. replicate sampling) instead of a single long one, which should avoid underestimation of the uncertainty.[129] All together, this enforces the approach of using a rather small set of uncorrelated structures, which should balance out the trade off between efficient VS and statistical significance. In analogy, the usage of NMR ensembles instead of MD simulations has been proposed,[130] and X-ray structure ensembles still need to be tested. Nevertheless, MM-PB/GBSA is a technique mainly used for predicting relative binding energies and not absolute ones, since several effects such as hydration/dehydration, entropy and binding pathway contributions can hardly be taken into account.

1.2.6 Methodological overview with a special look on endocrine disruptors

The following Mini-Review (published in the journal *Endocrinology*) provides a methodological overview of various computer-assisted approaches that use ligands and targets properties to predict binding and focuses on endocrine disruptor activities.

In Silico Predictions of Endocrine Disruptors Properties

Melanie Schneider,[1] Jean-Luc Pons,[1] Gilles Labesse,[1] and William Bourguet[1]

[1]Centre de Biochimie Structurale, CNRS, INSERM, Université de Montpellier, 34090 Montpellier, France

ORCiD numbers: 0000-0002-7085-6731 (M. Schneider); 0000-0002-6861-3300 (G. Labesse); 0000-0002-0643-7719 (W. Bourguet).

Endocrine-disrupting chemicals (EDCs) are a broad class of molecules present in our environment that are suspected to cause adverse effects in the endocrine system by interfering with the synthesis, transport, degradation, or action of endogenous ligands. The characterization of the harmful interaction between environmental compounds and their potential cellular targets and the development of robust *in vivo*, *in vitro*, and *in silico* screening methods are important for assessment of the toxic potential of large numbers of chemicals. In this context, computer-aided technologies that will allow for activity prediction of endocrine disruptors and environmental risk assessments are being developed. These technologies must be able to cope with diverse data and connect chemistry at the atomic level with the biological activity at the cellular, organ, and organism levels. Quantitative structure–activity relationship methods became popular for toxicity issues. They correlate the chemical structure of compounds with biological activity through a number of molecular descriptors (*e.g.*, molecular weight and parameters to account for hydrophobicity, topology, or electronic properties). Chemical structure analysis is a first step; however, modeling intermolecular interactions and cellular behavior will also be essential. The increasing number of three-dimensional crystal structures of EDCs' targets has provided a wealth of structural information that can be used to predict their interactions with EDCs using docking and scoring procedures. In the present review, we have described the various computer-assisted approaches that use ligands and targets properties to predict endocrine disruptor activities. (***Endocrinology*** **160: 2709–2716, 2019)**

Downloaded from https://academic.oup.com/endo/article-abstract/160/11/2709/5526762 by Institut de Génétique Humaine user on 29 October 2019

During the past decades, a large number of observations have shown that many exogenous substances can interfere with hormone levels or hormone action and, in turn, induce toxic effects. This has led to the identification of endocrine disrupting chemicals (EDCs) as a new class of toxic agents that will not be recognized, at first, by their chemical structure or by a specific type of usage but, rather, by their mechanisms of action (1–3). EDCs are exogenous substances that interfere with the function of hormonal systems and produce a range of developmental, reproductive, neurologic, immune, or metabolic diseases in humans and wildlife (4). Most EDCs are man-made chemicals produced by industry and released into the environment. However, some naturally occurring EDCs can also be found in plants or fungi. Exposure to EDCs occurs through ingesting food, drinking water, breathing contaminated air, or skin contact. The group of molecules acting as EDCs is highly heterogeneous and includes compounds that are often distantly related to endogenous ligands in terms of size or chemical structure. This group contains substances such as plasticizers (*e.g.*, bisphenols, phthalates), preservatives (*e.g.*, parabens), the byproducts of various industrial processes (*e.g.*, dioxins), surfactants (*e.g.*, alkylphenols, perfluoroalkyls), biocides (*e.g.*, organotins), flame retardants (*e.g.*, halogenated bisphenols), and ultraviolet filters (*e.g.*, benzophenones) and natural compounds such as the phytoestrogens genistein and daidzein or the mycoestrogen zearalenone.

ISSN Online 1945-7170
Copyright © 2019 Endocrine Society
This article has been published under the terms of the Creative Commons Attribution License (CC BY; https://creativecommons.org/licenses/by/4.0/)
Received 20 May 2019. Accepted 26 June 2019.
First Published Online 2 July 2019

Abbreviations: 2D, two-dimensional; 3D, three-dimensional; BPA, bisphenol A; EDC, endocrine-disrupting chemical; ER, estrogen receptor; MD, molecular dynamics; PPARγ, peroxisome proliferator–activated receptor-γ; QM, quantum mechanics.

EDCs can affect the endocrine systems of an organism in a wide variety of ways, for example, by mimicking natural hormones, antagonizing their action, or modifying their synthesis, metabolism, and transport through their interference with multiple cellular targets. These include membrane and nuclear receptors, the aryl hydrocarbon receptor, the enzymatic machineries involved in hormone biosynthesis and metabolism, and various carriers. Within the chemical regulations, criteria to identify EDCs have been recently proposed, which require information on a chemical's endocrine mode of action and related adverse effects relevant for human health. This involves the screening and testing of EDCs and mainly incorporates internationally accepted test methods developed under the Organization for Economic Cooperation and Development. In this context, the development of accurate *in silico* testing strategies could help to elucidate or confirm the suspected mode of actions and might suggest associated adverse effects by predicting the repertoire of molecular targets of EDCs. It might also provide guidelines to select or optimize molecule usage or designed to prevent unwanted activities.

Approaches to predict toxicity or activity against a particular target for a putative EDC can be divided according to the nature of data they are using and by their demand in computational resources. One of the simplest tools is ADME (Absorption, Distribution, Metabolism, Excretion)-Tox filters often used by pharmaceutical companies. Those can be based on composition rules (5), for example, specific chemical groups that should be avoided because they have shown adverse effects in the past (6). Another method of investigating the problem is drug-induced metabolic perturbation studies. These are based on metabolic network modeling using large-scale "omics" data, metabolic stability estimations, and mode of action analyses (7–10). Some of them have been shared with usual ADME-Tox issues such metabolization by cytochromes P450. The general methods for *in silico* toxicity prediction have been previously reviewed (7, 11–13). EDCs fall into particular niches of the available chemical space optimized for other properties and only partially mimicking natural hormones. They often differ in chemical structure from most medicinal and endogenous compounds and are encountered at unexpectedly high concentrations in the environment and living organisms [*e.g.*, bisphenol A (BPA), organotins]. Therefore, dedicated approaches are needed to detect the endocrine disruption potential.

The focus of the present review was centered on the field of prediction methods that aim to qualify the interaction between given small molecules, as potential EDCs, and a focused set of macromolecular targets. This is a very large field of research with many different methods that have been developed. Each method has its strengths, limitations, scope of application, and specificity of interpretation. The first questions to be asked upfront include the following: How much data are available? What is the nature of this data? How fast are results required? What is the minimal required accuracy of the prediction? What resources are available? Having those questions in mind, the goal is to find the most effective method. In addition to the classification into high-, medium-, and low-throughput methods, the available approaches can be classified further according to the type of data used. Most often, chemoinformatics methods will be classified as ligand-based and target structure-based approaches (Fig. 1) (14). Depending on the amount of data and the need for screening large data sets, the corresponding method should be chosen. This clearly involves a tradeoff between the amount of molecules, speed, and accuracy. However, combinations of techniques are emerging to improve overall efficiency and applicability. We first surveyed ligand-based virtual screening techniques as quick filters and then the role of structure-based virtual screening and discussed their potential combination. In both cases, one must adequately describe the studied molecules, which will usually start by extracting or writing its chemical formula as a linear string of atoms, such as SMILES (simplified molecular-input line-entry system) (Fig. 2), to be subsequently transformed into various other representations [two-dimensional (2D), three-dimensional (3D)] either for comparison with other molecules (*i.e.*, similarity searches, properties comparisons) in ligand-based virtual screening or by docking into putative targets (*i.e.*, in structure-based virtual screening).

Ligand-Based Methods

The so-called quantitative structure activity relationship (QSAR)/quantitative structure property relationship prediction models have been developed to predict a particular activity or property of the molecule in question. The simplest approaches have been based on the calculation of molecular descriptors that consider the molecule as a whole entity and calculate one value for the whole molecule (*e.g.*, molecular weight). The least expensive in terms of computational cost are models based on binary representations of molecules, called molecular fingerprints, or molecular descriptors (Fig. 2). These fingerprint representations can be binary in nature (property present or absent, yes or no, or 1 or 0), which only reflects the presence (or not) of a given feature or a count representation (sum of the instances for each feature). Millions of compounds can be screened within a

Downloaded from https://academic.oup.com/endo/article-abstract/160/11/2709/5526762 by Institut de Génétique Humaine user on 29 October 2019

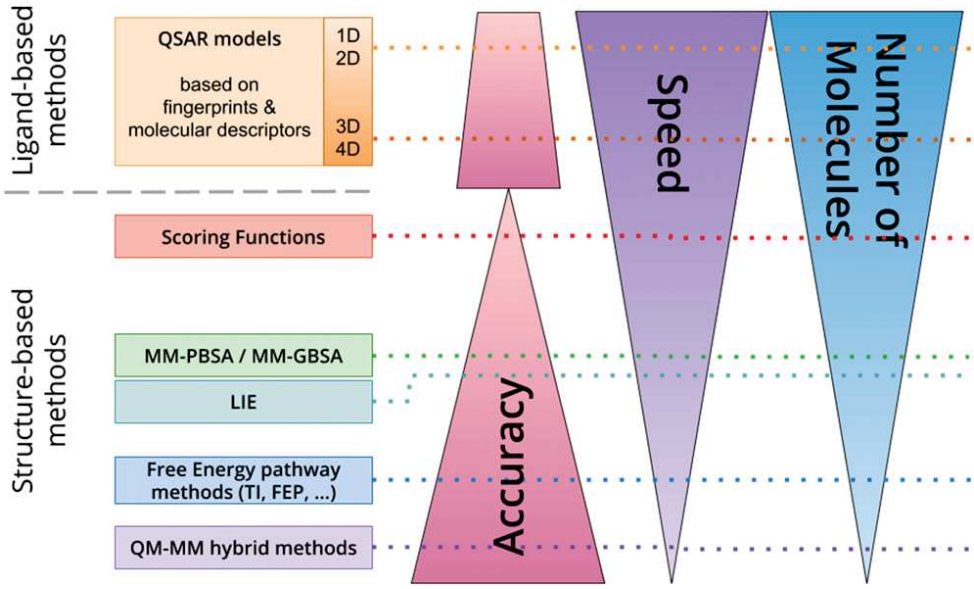


Figure 1. Affinity prediction methods grouped by ligand-based and structure-based methods and ranked by accuracy, computational effort and speed, and number of molecules that can be used. 1D, one-dimensional; 4D, four-dimensional.

reasonable period. Different types of fingerprints represent different properties of molecules, and it is, therefore, crucial to select an adequate type for modeling the desired activity. Many different types of molecular descriptors are available and might already be an output of a property prediction. Molecular descriptors and chemical fingerprints can be classified according to their dimensionality (Fig. 2). One-dimensional descriptors are scalars that describe the molecule according to its chemical formula (*e.g.*, molecular weight, atom counts, or bond counts). Two-dimensional descriptors are based on the structural topology, such as fragment counts or functional group counts (*e.g.*, alcohol function or aromatic ring). Three-dimensional descriptors extract information from 3D coordinate representations and are, therefore, based on the molecule's geometry. Four-dimensional descriptors are an extension of the 3D descriptors, which consider multiple conformations. In the case of 3D and four-dimensional descriptors, the computational effort will have already increased substantially and the borders toward the so-called structure-based methods will tend to vanish. All these descriptors allow for a rather rapid similarity search and classification to deduce or predict functional properties. Various *in silico* QSAR tools and, even, servers, namely the Organization for Economic Cooperation and Development QSAR toolbox (https://qsartoolbox.org/), VEGA HUB (https://www.vegahub.eu/), or CAESAR (http://www.caesar-project.eu/), to cite a few, are available, and open challenges have now been implemented to evaluate them more fairly, such as the Tox21 ("toxicity testing in the twenty-first century" initiative) project. DeepTox, the winner of the "Tox21 Data Challenge 2014" obtained excellent performances with a deep multitask neural network using ECFP4 fingerprint features (15).

In general, ligand-based methods will be very restricted to the chemical space of the molecules used for

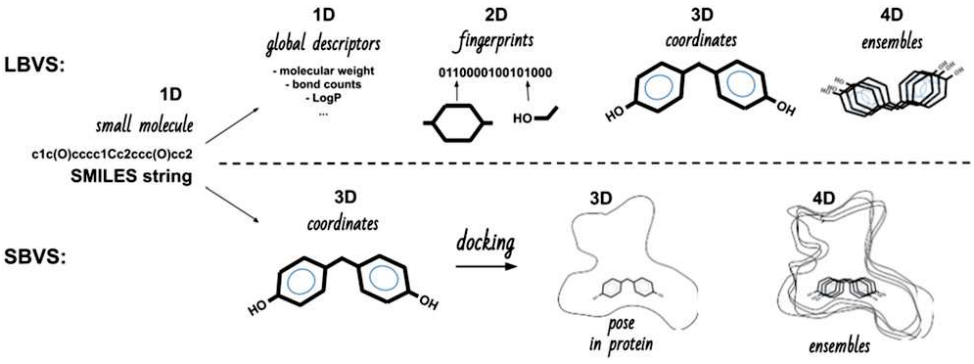


Figure 2. Molecular representations used by different methods, which were classified by the overall methodology (ligand-based virtual screening and structure-based virtual screening) and according to the dimensionality [one-dimensional (1D), 2D, 3D, four-dimensional (4D)] of the variables used.

Downloaded from https://academic.oup.com/endo/article-abstract/160/11/2709/5526762 by Institut de Génétique Humaine user on 29 October 2019

method development, especially if only a limited amount of data are available for model training. This can cause disappointing performances, especially in projections or extrapolation to new and dissimilar compounds (16). Therefore, the definition and declaration of an applicability domain—a region in the chemical space for which a QSAR model should make predictions with a given reliability—is considered as a necessary good practice for those model types (17). The quality of experimental data is also essential for valuable modeling as recently illustrated on the estrogen receptors (ERα, ERβ), which are two of the most extensively studied targets with respect to endocrine disrupting effects (18, 19). Regulation rules have been devised by the US Food and Drug Administration that require the assessment of estrogenic activity, and effort have been made to predict for ER binding (20, 21), including a large collaborative project (22). The latter, which compared numerous models and data sets, showed that poorly evaluated data sets are of little help for improving prediction quality despite providing experimental data for thousands of ligands. Similarly, other steroid hormone receptors, such as the androgen receptor, have been targeted for model development (23–26). To evaluate the risk of being EDCs, the prediction of a specific mechanism such as binding to a particular receptor is preferred for its expected greater accuracy and low cost. General models that aim at predictions on large protein families are less common. EDCs are active against specific targets of diverse nature (enzymes such as cytochromes P450 or DNA-binding proteins such as nuclear receptors). Accordingly, dedicated models might be required in agreement with their experimental characterization.

Structure-Based Methods

The increasing knowledge of functional and structural data has allowed for the evaluation or prediction of the potential interactions of known or putative EDCs to various targets using docking or more demanding approaches [*e.g.*, molecular dynamics (MD); see the next paragraph]. Structure-based methods, also called target-based methods, use information from a protein target 3D structure and are spanning a large scale in terms of computational cost. Docking procedures are the most widely used in virtual screening campaigns and can manage to thousands of ligands. They are based on sampling the conformational space of a given ligand in the binding pocket of a target molecule and a subsequent pose evaluation performed by scoring functions. Although the sampling of many widely used algorithms has seemed to be sufficient to find accurate poses (defined by reproducing crystallographic poses), the scoring functions still seem to suffer from diverse approximations (27–29). Accordingly, docking, followed by various rescoring procedures, is now commonly used to screen large molecular data sets in drug discovery (27, 29). This has been applied for endocrine disruption prediction on the androgen receptor (24, 25, 30) and other nuclear receptors (31–33). Automatic docking to 16 putative targets of EDCs or 14 distinct nuclear receptors has been made user-friendly through two servers, the OpenVirualToxLab (34) and Endocrine Disruptome (35). However, structure analysis has also revealed the importance of protein flexibility. Adequately modeling target flexibility is a major limitation that has been addressed using structure ensembles, instead of single conformations (36). One approach is to use multiple experimental conformations in parallel for docking and gather the results to extract the best or more likely poses. A derivative of our server for comparative modeling "@TOME" (37) now includes a docker (to be described in more detail elsewhere). This allows for the selection of the protein conformation best suitable to accommodate a given ligand. This dedicated server called EDMon (Endocrine Disruptor Monitoring; available at: http://edmon.cbs.cnrs.fr/) is now available to screen for ERα, ERβ, and peroxisome proliferator–activated receptor-γ (PPARγ). It predicts for affinities using a rescoring approach based on machine learning (38). However, the problem is still severe for promiscuous proteins, such as the nuclear receptors CAR (constitutive androstane receptor) and PXR (pregnane X receptor) (39). The dozen of structures described to date for these receptors have shown dramatic structural rearrangements on ligand binding, and more experimental 3D structures are necessary to reach a better description of the conformational landscape they could access.

Alternatively, to unravel or model intrinsic protein flexibility and possible ligand-induced fit, MD simulations can be used but at a significantly greater computational cost (*e.g.*, one to several weeks using a standard workstation). MD-based prediction methods require more effort with respect to system setup and analysis, and they are usually not provided as simple "plug and play" modules, such as is the case for many commercial or noncommercial docking tools. To date, MD simulations have already been used to study the structural flexibility and the dynamics of binding events of several nuclear receptors (40–45), with and without further investigation of small molecule-binding affinities. The server OpenVirualToxLab (34) provides easy access to focused MD, which is used to refine and evaluate theoretical complexes deduced from docking into 16 EDC targets. In general, MD-based affinity estimation protocols can be divided into two major groups: endpoint methods and

Downloaded from https://academic.oup.com/endo/article-abstract/160/11/2709/5526762 by Institut de Génétique Humaine user on 29 October 2019

free energy pathway methods. The endpoint methods, as already indicated by the name, consider the two "end" states of the system: the bound and the unbound molecules. Two commonly used ones include the MM-PBSA (molecular mechanics Poisson-Boltzmann surface area) (46, 47) and MM-GBSA (molecular mechanics generalized born surface area) (47–49). These computations can be adjusted to a particular system through parametrization within the so-called linear interaction energy method (24, 50–52). For example, MD simulations, followed by MM-PBSA calculations, have been used to study the structural effects and interaction mechanism of BPA with three human nuclear receptors, ERα, ERRγ (estrogen-related receptor-γ), and PPARγ (53) or to determine the binding of bisphenols BPA, bisphenol AF, and bisphenol S to ERα (54). These computations require some expertise but can be performed using a personal workstation and are now often applied on several dozens of compounds against a given target. They allow for rescoring of docking poses using physics-based approaches; however, their usefulness has continued to be debated. Furthermore, the standard MD techniques can suffer from an insufficient sampling of the conformational space of the target molecule. This can occur for different reasons, such as large conformational movements during binding, slow transitions between states, rare events, or high-energy barriers that must be overcome. In such cases, a set of different computational methods has been proposed—the free energy pathway methods such as transition path sampling, umbrella sampling, steered-MD, and funnel-metadynamics (55–59). Among the free energy pathway methods is a subgroup of alchemical methods represented by the thermal integration (60, 61) and free energy perturbation (62, 63) methods. Recently, a combination of methods has been applied to toxicity studies for the identification of possible ligand binding modes to PPARγ (64). However, those approaches are even more demanding in central processing unit time and are not commonly performed for toxicity predictions.

Finally, extremely precise energy estimations can be computed using quantum mechanics (QM) but at huge computational cost. Thus, QM is often restricted to modeling of the binding site. Mixed/hybrid approaches will allow for computation locally of a QM procedure, and a standard MD approach is applied to the rest of the molecular system under study. Quantum effects might be required to correctly estimate particular molecular interactions when atomic bonds are broken or reformed during the binding event or for predicting the reaction rates in drug metabolism, which is the case for cytochromes P450 (65–67). Free-energy estimation and QM have been performed on a very limited number of complexes. However, their exquisite characterization of molecular structures and interactions might help to precisely define various chemical properties (*i.e.*, conformation, charge, reactivity) and/or to parametrize quicker methods (*e.g.*, for scoring or docking).

Current Limitations and Future Directions

Because the US ToxCast program and the European Union's Registration, Evaluation, Authorization, and Restriction of Chemicals regulation aim to assess the toxicity of more than 100,000 synthetic chemicals, a strong demand exists for alternative test methods and, in particular, such computational tools that will allow for the reduction of the cost of the evaluation and in animal lives. Despite recent major advances in the field of affinity prediction resulting in numerous tools and diverse approaches, one must remember their limitations. One of the major concerns of ligand-based *in silico* prediction methods is its high dependency on experimental data. The presence of inconsistent and erroneous data during the training process can lead to biased and inaccurate predictions and the applicability domain is a major prerequisite that needs to fit for reliable predictions. Large-scale high-throughput experimental testing to generate coherent databases and curation of the existing ones would help generate more accurate prediction models. The QSAR approaches available usually display applicability domain centered on the training data and struggle to yield reasonable predictions for highly unbalanced data sets. The current development and combination of novel statistical and machine/deep learning approaches are likely to generate novel *in silico* models that could manage highly unbalanced data sets, allowing for the applicability domain to expand beyond the training data.

Concerning target-based methods, its dependency is more reduced. However, the issue of potentially unknown structural changes still exists. An inherent limitation of any modeling tool is its parametrization for all possible chemistries. Currently, knowledge is lacking regarding the proper evaluation of protein–ligand interactions involving halogen atoms, metals (*e.g.*, organotins), or newly used entities such as organoborans. In addition, such compounds have been previously demonstrated to act as EDCs. More crystal structures would be necessary to reflect the conformational landscape of the target receptors in a more comprehensive manner and help in training docking tools with exotic atoms. This suggests the need for tighter interactions between structuralists and predictors to tune experimental works to fill in the gaps in structural and/or functional data. Another difficulty not easily manageable, especially for

Downloaded from https://academic.oup.com/endo/article-abstract/160/11/2709/5526762 by Institut de Génétique Humaine user on 29 October 2019

large chemical data sets, is the possibility for simultaneous binding of two or more cases of the same molecule (especially for small compounds) and/or of distinct molecules (mixtures) in a cooperative and/or allosteric fashion. Developing dedicated tools will be necessary to manage this task correctly to predict potential "cocktail effects" (68). For different protein targets (16 listed to date for EDCs), different techniques are already available and have been applied with varying rates of success. Not only for nuclear receptors or cytochromes P450, but also for ion channels such as the *hERG* (human ether-a-go-go-related gene) potassium channel, different methods from ligand-based and target-based to systems biology have been applied (69).

Finally, cascading prediction tools and filters will be necessary to (i) account for potential metabolization that creates unexpected or new chemical entities with new properties, (ii) detect nonclassic properties [*e.g.*, covalent attachment (70), multiple binding], (iii) combine QSAR and docking, or (iv) start to predefine structural ensembles for quicker estimation of receptor flexibility to derive more accurate predictions. The latter might help in accessing better description of flexible complexes and avoid the burden of long simulations. Next, one could dream of combining those studies with mathematical models at the cellular level and in the endocrine system. Hence, room exists for further improvements in which a fruitful interplay between modeling and experimental characterization should be promoted.

Acknowledgments

Financial Support: The present project received funding from the European Union's Horizon 2020 research and innovation programme (Grant Agreement, GOLIATH No. 825489 to W.B.), the Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Grants ANSES EST-2016/1/162-XENOMIX and EST-2018/1/095-TOXCHEM to W.B.).

Additional Information

Correspondence: Gilles Labesse, PhD, or William Bourguet, PhD, Centre de Biochimie Structurale, 29 rue de Navacelles, 34090 Montpellier, France. E-mail: gilles.labesse@cbs.cnrs.fr or william.bourguet@cbs.cnrs.fr.

Disclosure Summary: The authors have nothing to disclose.

Data Availability: Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

References and Notes

1. Zoeller RT, Bergman Å, Becher G, Bjerregaard P, Bornman R, Brandt I, Iguchi T, Jobling S, Kidd KA, Kortenkamp A, Skakkebaek NE, Toppari J, Vandenberg LN. A path forward in the debate over health impacts of endocrine disrupting chemicals. *Environ Health*. 2014;**13**(1):118.
2. Gore AC, Chappell VA, Fenton SE, Flaws JA, Nadal A, Prins GS, Toppari J, Zoeller RT. EDC-2: The Endocrine Society's second scientific statement on endocrine-disrupting chemicals. *Endocr Rev*. 2015;**36**(6):E1–E150.
3. Delfosse V, Maire AL, Balaguer P, Bourguet W. A structural perspective on nuclear receptors as targets of environmental compounds. *Acta Pharmacol Sin*. 2015;**36**(1):88–101.
4. Schug TT, Janesick A, Blumberg B, Heindel JJ. Endocrine disrupting chemicals and disease susceptibility. *J Steroid Biochem Mol Biol*. 2011;**127**(3-5):204–215.
5. Yang H, Sun L, Li W, Liu G, Tang Y. *In silico* prediction of chemical toxicity for drug design using machine learning methods and structural alerts. *Front Chem*. 2018;**6**:30.
6. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*. 2001;**46**(1):3–26.
7. Reisfeld B, Mayeno AN. What is computational toxicology? In: Reisfeld B, Mayeno AN, eds. *Computational Toxicology: Volume I. Methods in Molecular Biology*. Totowa, NJ: Humana Press; 2012:3–7.
8. Wu Y, Wang G. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. *Int J Mol Sci*. 2018;**19**(8):2358.
9. Raies AB, Bajic VB. In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip Rev Comput Mol Sci*. 2016;**6**(2):147–172.
10. Quignot N, Bois FY. A computational model to predict rat ovarian steroid secretion from in vitro experiments with endocrine disruptors. *PLoS One*. 2013;**8**(1):e53891.
11. Roncaglioni A, Toropov AA, Toropova AP, Benfenati E. In silico methods to predict drug toxicity. *Curr Opin Pharmacol*. 2013;**13**(5):802–806.
12. Zhang L, Zhang H, Ai H, Hu H, Li S, Zhao J, Liu H. Applications of machine learning methods in drug toxicity prediction. *Curr Top Med Chem*. 2018;**18**(12):987–997.
13. Myatt GJ, Ahlberg E, Akahori Y, Allen D, Amberg A, Anger LT, Aptula A, Auerbach S, Beilke L, Bellion P, Benigni R, Bercu J, Booth ED, Bower D, Brigo A, Burden N, Cammerer Z, Cronin MTD, Cross KP, Custer L, Dettwiler M, Dobo K, Ford KA, Fortin MC, Gad-McDonald SE, Gellatly N, Gervais V, Glover KP, Glowienke S, Van Gompel J, Gutsell S, Hardy B, Harvey JS, Hillegass J, Honma M, Hsieh J-H, Hsu C-W, Hughes K, Johnson C, Jolly R, Jones D, Kemper R, Kenyon MO, Kim MT, Kruhlak NL, Kulkarni SA, Kümmerer K, Leavitt P, Majer B, Masten S, Miller S, Moser J, Mumtaz M, Muster W, Neilson L, Oprea TI, Patlewicz G, Paulino A, Lo Piparo E, Powley M, Quigley DP, Reddy MV, Richarz A-N, Ruiz P, Schilter B, Serafimova R, Simpson W, Stavitskaya L, Stidl R, Suarez-Rodriguez D, Szabo DT, Teasdale A, Trejo-Martin A, Valentin J-P, Vuorinen A, Wall BA, Watts P, White AT, Wichard J, Witt KL, Woolley A, Woolley D, Zwickl C, Hasselgren C. In silico toxicology protocols. *Regul Toxicol Pharmacol*. 2018;**96**:1–17.
14. Parenti MD, Rastelli G. Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnol Adv*. 2012;**30**(1):244–250.
15. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. *Front Environ Sci*. 2016;**3**.
16. Maggiora GM. On outliers and activity cliffs—why QSAR often disappoints. *J Chem Inf Model*. 2006;**46**(4):1535–1536.
17. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model*. 2008;**26**(8):1315–1326.
18. Shanle EK, Xu W. Endocrine disrupting chemicals targeting estrogen receptor signaling: identification and mechanisms of action. *Chem Res Toxicol*. 2011;**24**(1):6–19.
19. Delfosse V, Grimaldi M, Cavaillès V, Balaguer P, Bourguet W. Structural and functional profiling of environmental ligands for

Downloaded from https://academic.oup.com/endo/article-abstract/160/11/2709/5526762 by Institut de Génétique Humaine user on 29 October 2019

estrogen receptors. *Environ Health Perspect*. 2014;**122**(12): 1306–1313.
20. Zhang Q, Yan L, Wu Y, Ji L, Chen Y, Zhao M, Dong X. A ternary classification using machine learning methods of distinct estrogen receptor activities within a large collection of environmental chemicals. *Sci Total Environ*. 2017;**580**:1268–1275.
21. Balabin IA, Judson RS. Exploring non-linear distance metrics in the structure-activity space: QSAR models for human estrogen receptor. *J Cheminform*. 2018;**10**(1):47.
22. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, Trisciuzzi D, Fourches D, Horvath D, Benfenati E, Muratov E, Wedebye EB, Grisoni F, Mangiatordi GF, Incisivo GM, Hong H, Ng HW, Tetko IV, Balabin I, Kancherla J, Shen J, Burton J, Nicklaus M, Cassotti M, Nikolov NG, Nicolotti O, Andersson PL, Zang Q, Politi R, Beger RD, Todeschini R, Huang R, Farag S, Rosenberg SA, Slavov S, Hu X, Judson RS. CERAPP: collaborative estrogen receptor activity prediction project. *Environ Health Perspect*. 2016;**124**(7):1023–1033.
23. Chen Q, Tan H, Yu H, Shi W. Activation of steroid hormone receptors: shed light on the in silico evaluation of endocrine disrupting chemicals. *Sci Total Environ*. 2018;**631-632**:27–39.
24. Lill MA, Winiger F, Vedani A, Ernst B. Impact of induced fit on ligand binding to the androgen receptor: a multidimensional QSAR study to predict endocrine-disrupting effects of environmental chemicals. *J Med Chem*. 2005;**48**(18):5666–5674.
25. Yang X, Liu H, Yang Q, Liu J, Chen J, Shi L. Predicting antiandrogenic activity of bisphenols using molecular docking and quantitative structure-activity relationships. *Chemosphere*. 2016; **163**:373–381.
26. Grisoni F, Consonni V, Ballabio D. Machine learning consensus to predict the binding to the androgen receptor within the CoMPARA Project. *J Chem Inf Model*. 2019;**59**(5):1839–1848.
27. Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr. Computational methods in drug discovery. *Pharmacol Rev*. 2013;**66**(1):334–395.
28. Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: a critical review. *Curr Med Chem*. 2013;**20**(23): 2839–2860.
29. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today*. 2006;**11**(13-14):580–594.
30. Wahl J, Smieško M. Endocrine disruption at the androgen receptor: employing molecular dynamics and docking for improved virtual screening and toxicity prediction. *Int J Mol Sci*. 2018; **19**(6):1784.
31. Park S-J, Kufareva I, Abagyan R. Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *J Comput Aided Mol Des*. 2010;**24**(5):459–471.
32. Trisciuzzi D, Alberga D, Leonetti F, Novellino E, Nicolotti O, Mangiatordi GF. Molecular docking for predictive toxicology. In: Nicolotti O, ed. *Computational Toxicology: Methods and Protocols. Methods in Molecular Biology*. New York, NY: Springer New York; 2018:181–197.
33. Delfosse V, Grimaldi M, Pons J-L, Boulahtouf A, le Maire A, Cavailles V, Labesse G, Bourguet W, Balaguer P. Structural and mechanistic insights into bisphenols action provide guidelines for risk assessment and discovery of bisphenol A substitutes. *Proc Natl Acad Sci USA*. 2012;**109**(37):14930–14935.
34. Vedani A, Dobler M, Hu Z, Smieško M. OpenVirtualToxLab—a platform for generating and exchanging in silico toxicity data. *Toxicol Lett*. 2015;**232**(2):519–532.
35. Kolšek K, Mavri J, Sollner Dolenc M, Gobec S, Turk S. Endocrine disruptome—an open source prediction tool for assessing endocrine disruption potential through nuclear receptor binding. *J Chem Inf Model*. 2014;**54**(4):1254–1267.
36. Korb O, Olsson TSG, Bowden SJ, Hall RJ, Verdonk ML, Liebeschuetz JW, Cole JC. Potential and limitations of ensemble docking. *J Chem Inf Model*. 2012;**52**(5):1262–1274.
37. Pons J-L, Labesse G. @TOME-2: a new pipeline for comparative modeling of protein-ligand complexes. *Nucleic Acids Res*. 2009; **37**(Web Server issue, suppl 2):485–W491.
38. Schneider M, Pons JL, Bourguet W, Labesse G. Towards accurate high-throughput ligand affinity prediction by exploiting structural ensembles, docking metrics and ligand similarity [published online ahead of print 26 July 2019]. *Bioinformatics*. doi: 10.1093/bioinformatics/btz538.
39. Ekins S, Kortagere S, Iyer M, Reschly EJ, Lill MA, Redinbo MR, Krasowski MD. Challenges predicting ligand-receptor interactions of promiscuous proteins: the nuclear receptor PXR. *PLoS Comput Biol*. 2009;**5**(12):e1000594.
40. Zheng L, Lin VC, Mu Y. Exploring flexibility of progesterone receptor ligand binding domain using molecular dynamics. *PLoS One*. 2016;**11**(11):e0165824.
41. Blondel A, Renaud J-P, Fischer S, Moras D, Karplus M. Retinoic acid receptor: a simulation analysis of retinoic acid binding and the resulting conformational changes. *J Mol Biol*. 1999;**291**(1):101–115.
42. Martínez L, Sonoda MT, Webb P, Baxter JD, Skaf MS, Polikarpov I. Molecular dynamics simulations reveal multiple pathways of ligand dissociation from thyroid hormone receptors. *Biophys J*. 2005;**89**(3):2011–2023.
43. Celik L, Lund JDD, Schiøtt B. Conformational dynamics of the estrogen receptor α: molecular dynamics simulations of the influence of binding site structure on protein dynamics. *Biochemistry*. 2007;**46**(7):1743–1758.
44. Costantino G, Entrena-Guadix A, Macchiarulo A, Gioiello A, Pellicciari R. Molecular dynamics simulation of the ligand binding domain of farnesoid X receptor. Insights into helix-12 stability and coactivator peptide stabilization in response to agonist binding. *J Med Chem*. 2005;**48**(9):3251–3259.
45. Motta S, Callea L, Giani Tagliabue S, Bonati L. Exploring the PXR ligand binding mechanism with advanced molecular dynamics methods. *Sci Rep*. 2018;**8**(1):16207.
46. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE III. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res*. 2000;**33**(12):889–897.
47. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M. Validation and use of the MM-PBSA approach for drug discovery. *J Med Chem*. 2005; **48**(12):4040–4048.
48. Hou T, Wang J, Li Y, Wang W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J Chem Inf Model*. 2011;**51**(1):69–82.
49. Sun H, Li Y, Shen M, Tian S, Xu L, Pan P, Guan Y, Hou T. Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Phys Chem Chem Phys*. 2014;**16**(40):22035–22045.
50. Åqvist J, Medina C, Samuelsson J-E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng*. 1994; **7**(3):385–391.
51. Gutiérrez-de-Terán H, Åqvist J. Linear interaction energy: method and applications in drug design. In: Baron R, ed. *Computational Drug Discovery and Design. Methods in Molecular Biology*. New York, NY: Springer New York; 2012:305–323.
52. Rifai EA, van Dijk M, Vermeulen NPE, Geerke DP. Binding free energy predictions of farnesoid X receptor (FXR) agonists using a linear interaction energy (LIE) approach with reliability estimation: application to the D3R Grand Challenge 2. *J Comput Aided Mol Des*. 2018;**32**(1):239–249.
53. Li L, Wang Q, Zhang Y, Niu Y, Yao X, Liu H. The molecular mechanism of bisphenol A (BPA) as an endocrine disruptor by interacting with nuclear receptors: insights from molecular dynamics (MD) simulations. *PLoS One*. 2015;**10**(3):e0120330.
54. Li Y, Perera L, Coons LA, Burns KA, Tyler Ramsey J, Pelch KE, Houtman R, van Beuningen R, Teng CT, Korach KS. Differential

Downloaded from https://academic.oup.com/endo/article-abstract/160/11/2709/5526762 by Institut de Génétique Humaine user on 29 October 2019

in vitro biological action, coregulator interactions, and molecular dynamic analysis of bisphenol A (BPA), BPAF, and BPS ligand-ERα complexes. *Environ Health Perspect*. 2018;**126**(1):017012.
55. Woo H-J, Roux B. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc Natl Acad Sci USA*. 2005;**102**(19):6825–6830.
56. Lee MS, Olson MA. Calculation of absolute protein-ligand binding affinity using path and endpoint approaches. *Biophys J*. 2006; **90**(3):864–877.
57. Limongelli V, Bonomi M, Parrinello M. Funnel metadynamics as accurate binding free-energy method. *Proc Natl Acad Sci USA*. 2013;**110**(16):6358–6363.
58. Aldeghi M, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem Sci (Camb)*. 2016;**7**(1):207–218.
59. Brotzakis ZF, Limongelli V, Parrinello M. Accelerating the calculation of protein-ligand binding free energy and residence times using dynamically optimized collective variables. *J Chem Theory Comput*. 2019;**15**(1):743–750.
60. Straatsma TP, Berendsen HJC. Free energy of ionic hydration: analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *J Chem Phys*. 1988;**89**(9):5876–5886.
61. Bhati AP, Wan S, Wright DW, Coveney PV. Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *J Chem Theory Comput*. 2017; **13**(1):210–222.
62. Cournia Z, Allen B, Sherman W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J Chem Inf Model*. 2017;**57**(12):2911–2937.
63. Manzoni F, Ryde U. Assessing the stability of free-energy perturbation calculations by performing variations in the method. *J Comput Aided Mol Des*. 2018;**32**(4):529–536.
64. Fratev F, Steinbrecher T, Jónsdóttir SÓ. Prediction of accurate binding modes using combination of classical and accelerated molecular dynamics and free-energy perturbation calculations: an application to toxicity studies. *ACS Omega*. 2018;**3**(4):4357–4371.
65. Bathelt CM, Mulholland AJ, Harvey JN. QM/MM modeling of benzene hydroxylation in human cytochrome P450 2C9. *J Phys Chem A*. 2008;**112**(50):13149–13156.
66. Shaik S, Cohen S, Wang Y, Chen H, Kumar D, Thiel W. P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations. *Chem Rev*. 2010;**110**(2):949–1017.
67. Lonsdale R, Oláh J, Mulholland AJ, Harvey JN. Does compound I vary significantly between isoforms of cytochrome P450? *J Am Chem Soc*. 2011;**133**(39):15464–15474.
68. Delfosse V, Dendele B, Huet T, Grimaldi M, Boulahtouf A, Gerbal-Chaloin S, Beucher B, Roecklin D, Muller C, Rahmani R, Cavaillès V, Daujat-Chavanieu M, Vivat V, Pascussi J-M, Balaguer P, Bourguet W. Synergistic activation of human pregnane X receptor by binary cocktails of pharmaceutical and environmental compounds. *Nat Commun*. 2015;**6**(1):8089.
69. Taboureau O, Jorgensen ES. In silico predictions of hERG channel blockers in drug discovery: from ligand-based and target-based approaches to systems chemical biology. *Comb Chem High Throughput Screen*. 2011;**14**(5):375–387.
70. le Maire A, Grimaldi M, Roecklin D, Dagnino S, Vivat-Hannah V, Balaguer P, Bourguet W. Activation of RXR-PPAR heterodimers by organotin environmental endocrine disruptors. *EMBO Rep*. 2009;**10**(4):367–373.

Downloaded from https://academic.oup.com/endo/article-abstract/160/11/2709/5526762 by Institut de Génétique Humaine user on 29 October 2019

1.3 Fighting cancer: drug targets, antitargets and resistance

Cancer is an umbrella term for a large family of diseases that involve abnormal cell growth with the potential to spread and invade other tissue. It is characterized by the presence of one or several tumors that are formed upon transformation of initially normal cells. In a healthy organism the immune system can usually cope with and eliminate cancer cells, nevertheless progression can occur giving rise to tumor formation. The tumor transformation is initially caused by mutation and results into a loss of control of the cell cycle, insensitivity to apoptosis, and abnormalities of DNA repair, which finally leads to the abnormal cell growth. Cancers are classified according to the type of cell in which the first transformation occurred, forming the primary tumor. The spreading of tumor cells from the primary site to different sites, which may involve the invasion of other tissue, is called metastasis and often means that chances for a complete cure of the cancer decrease. The mayor strategies to fight cancer are surgery, immunotherapy, chemotherapy and radiotherapy.

Most chemotherapeutic drugs work by impairing cell division in different ways and tumors with high growth rates are more sensitive to chemotherapy. Examples of non-specific drugs are cisplatine that works in part by binding to DNA and inhibiting its replication, fluorouracil that is believed to block DNA production by inhibiting thymidylate synthase, and antifolates that are antimetabolite medications competing with folic acid and thus inhibiting cell division, DNA/RNA synthesis and repair and protein synthesis. At the same time, many chemotherapeutic side effects can be attributed to the damage of normal cells that divide rapidly, such as cells in the bone marrow, digestive tract and hair follicles.[14] Targeting the proteins that are activated upon mutation or proteins involved in the affected pathways is a prominent approach when developing chemotherapeutic drugs. Unfortunately, chemotherapy is not always effective, and it may not completely destroy the cancer. Further problems related to ADME and toxicity issues may occur and differ between patients. Resistance is also a major cause of treatment failure in chemotherapeutic drugs. There are different possible causes of resistance in cancer, e.g gene amplification or alteration of gene expression (such that cell division is not impaired), the drug's metabolism (inactivation), the drug's export from the cells by pumps, and further mutations of involved proteins.

Within the presented work the focus is on three targets that are involved in cancer treatment in different ways: The oncogenic protein kinase BRAF, which is the primary target of the presented drug design project; the Pregnane X Receptor (PXR), a nuclear receptor that is dedicated to cell detoxification (small molecule clearance) and therefore a potential secondary target to be avoided by most drugs; the Estrogen Receptor alpha (ERα), a nuclear receptor as primary target for different cancer types, in particular breast cancer, and a secondary target to be avoided by a broad spectrum of drugs in order to avoid endocrine disruption.

1.3.1 Oncogenic protein kinase BRAF

BRAF is one of the three isoforms (with CRAF and ARAF) of the Rapidly Accelerated Fibrosarcoma (RAF) family of catalytically competent serine/threonine protein kinases (two pseudokinases, KSR1

and KSR2, are also included in the RAF family). BRAF plays a vital role in the RAS/RAF/MEK/ERK signalling cascade, which is also known as mitogen-activated protein kinase (MAPK) pathway (Figure 1.1), and participates in cell proliferation and survival.[131] Upon induction of conformational changes by RAS binding, stimulating the formation of active RAF homodimers or heterodimers, RAF changes its phosphorylation status, which triggers its kinase activity that activates MEK (MEK1 and MEK2), which in turn phosphorylates downstream ERK (ERK1 and ERK2). In contrast to the RAF and MEK kinases, ERK has a broad substrate specificity and is able to phosphorylate houndreds of different proteins.[132] As RAS is mutated in approximately 30% of human cancers, the development of inhibitors has been under investigation for a long while, but without significant success.[133] In addition, the oncogenic activation of BRAF induces constitutively and RAS-independently the MAPK pathway leading to the uncontrolled amplification of downstream signalling, which involves an increase of proliferation and finally tumorigenesis.[134] Many mutations (>30) of the BRAF gene associated with human cancers have been identified.[135] These are involved in approximately 100% of hairy cell leukemia,[136] 50% of melanomas, 45% of thyroid, 10% of colon, and 8% of ovarian carcinomas.[137] The most common mutation, accounting for approximately 90% of the detected BRAF mutated cases, is the replacement of valine with glutamic acid at position 600 (shortly V600E), which is located within the activation segment of the kinase domain and destabilizes the inactive conformation. This mutation leads to a constitutive kinase activity that is about 500-fold increased compared to wild type (WT) BRAF. Moreover, in contrast to the WT BRAF-V600E is signalling as a monomer and insensitive to ERK negative feedback mechanisms.[138] Therefore, inhibiting BRAF-V600E is a promising strategy for cancer treatment.

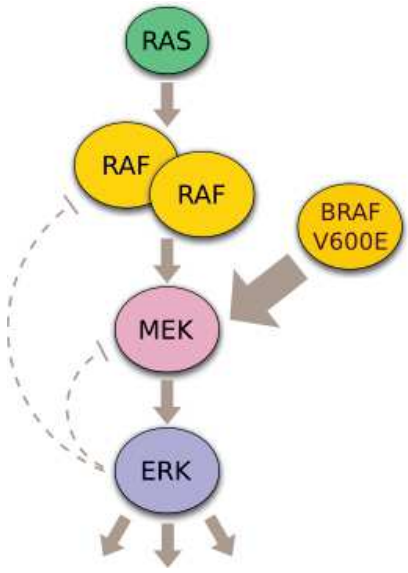


Figure 1.1: MAPK pathway activation. RAS activation promotes the formation of RAF dimers. RAF phosphorylates MEK, which upon activation phosphorylates ERK. ERK phosphorylates many different targets and also exerts a direct negative feedback by phosphorylating and thus regulating CRAF and MEK activity. The BRAF-mutant V600E is constitutively signalling as a monomer and insensitive to ERK negative feedback mechanisms. Paradoxical activation of ERK in BRAF-wild-type cells occurs via transactivation of RAF dimers. The dotted hammers represent pathway-induced feedback inhibition.

1.3.1.1 Structural basis of BRAF activation

BRAF is composed of a total of 766 amino acids and the overall architecture (from N to C terminus) is the following: The most N-terminal can be found a BRAF specific domain (being absent in ARAF and CRAF) that mediates homo- and hetero-dimerization[139] and interacts with the two pseudokinases (KSR1 and KSR2) of the RAF family.[140] All three RAF isoforms (ARAF, BRAF, CRAF) share three conserved regions, namely CR1, CR2, and CR3. CR1 contains a RAS-binding domain and a cysteine-rich domain, interacting with RAS and membrane phospholipids. CR2 is a serine-threonine rich domain, whose binding to the regulator protein 14-3-3 is followed by phosphorylation of serine 365 that inactivates the kinase.[141] CR3 is composed of the protein kinase domain and a C-terminal tail, which contains a regulatory serine residue (S729).

The protein kinase domain consists of 261 amino acids (residues 457–717), it has a typical kinase structure with two domains, the N-terminal and C-terminal lobe, linked via a flexible hinge segment. The small N-terminal lobe contains five β-strands (β1-5) and one α helix (αC). The large C-terminal lobe contains seven helices (αD-I and αEF) and four β-strands (β6-9). The deep cleft between the two lobes forms the active site pocket where ATP and peptide substrate or an inhibitor can bind. The kinase domain has several structurally and functionally important regions: 1) the flexible glycine-rich loop GSGSFG (residues 464-469), permitting ATP binding and ADP release during a catalytic cycle, 2) the regulatory αC helix with a conserved glutamate (E501) forming a salt bridge with the catalytic lysine (K483) in the β3-strand (as for all active kinases), 3) the hinge QWCEG (residues 530-534) that permits the movement of the two lobes with respect to each other, 4) the catalytic loop HRDLKSNN (residues 574-581), and 5) the flexible activation loop (a-loop) (residues 594-623), containing the activation segment, which is usually the phospho-acceptor site, starting with the DFG-motif containing the magnesium-binding D594 and the regulatory F595. BRAF also possesses the signature K/E/D/D (residues 483/501/576/594) required for catalysis, with lysine and glutamate being located in the N-terminal lobe and the two aspartates in the C-terminal lobe. The relative positioning of the αC helix and the DFG-motif orientation are important factors for kinase activation. When the αC helix is positioned close to the ATP site (αC-in position), which is required for an active conformation, the mentioned salt bridge between the β3-lysine and the αC-glutamate is established. If this interaction is not possible, the αC helix is shifted away from the ATP site (αC-out position), adopting an inactive conformation. The αC-in conformation is required, but not sufficient for catalytic activity. In the active conformation the DFG-D is directed towards the base of the αC helix and the active site (DFG-in conformation), where it can bind one of the two magnesium ions required for catalysis, but it can also shift outwards (DFG-out conformation), where the motif itself occupies part of the ATP binding site rendering the conformation inactive. Moreover, the relative orientation and vertical alignment of certain hydrophobic residues, forming the catalytic spine (residues A481, V471, F583, L584, I582, L537, L649, V645) and the regulator spine (residues F516, L505, F595, H574, D638), is equally important for kinase activation. In brief, structural requirements for a catalytically active kinase conformation are a closed configuration of the two lobes, a DFG-in/αC-in conformation, and an extended a-loop with an unfolded activation segment. The a-loop is highly flexible and large parts of it are unresolved in most BRAF crystal structures. The BRAF active site can be subdivided into the adenine (ATP base), the ribose (ATP ribose), hydrophobic, inhibitor type I and type II subpockets.

Structural consequences of the BRAF V600E mutation

Several structural consequences may arise from the BRAF V600E mutation. In the WT structure the activation segment (containing V600) can fold into a short helix and form hydrophobic contacts with the αC helix stabilizing the inactive αC-out conformation. The consequent steric clashes within this conformation upon V600E mutation may induce the active αC-in conformation. Another mode of interaction of V600 that seems to facilitate an inhibited kinase state (with folded a-loop) includes interactions of V600 with G-rich-loop residues. Accordingly, the V600E mutation provokes unfolding of the activation segment and extension of the a-loop, promoting the active state, which may be stabilized by a salt bridge between V600E and αC-lysine 507.[142]

1.3.1.2 BRAF inhibitors

ATP-competitive BRAF inhibitors, such as vermurafenib,[143] sorafenib[135] and dabrafenib[144] have been developed in order to block the MAPK signalling pathway and decrease tumor cell growth in cells expressing the BRAF mutant V600E. Selective targeting of BRAF-V600E is a proven therapeutic strategy for the treatment of metastatic melanoma and the drugs vemurafenib and dabrafenib have been approved by the U.S. Food and Drug Administration (FDA) for treatment of late-stage melanoma in 2011 and 2013, respectively.[145–147] Both drugs show improved response rates and overall survival of BRAF-V600 mutant melanoma patients, but unfortunately, due to rapidly acquired resistance most patients relapse within a year.[148]

The paradoxical effect

The formation of homo- or heterodimers is an important step in the activation of wild type BRAF, as the monomers are generally inactive due to autoinhibition by the N-terminal domain. Vemurafenib and dabrafenib produce the paradoxical activation of the MAPK pathway in wild type BRAF cells. While inhibiting the BRAF-V600E mutant the drugs induce the opposite behaviour in wild type cells, leading to skin lesions and promoting growth and metastasis of tumor cells with RAS mutations. Here, the clinical importance of RAF dimerization is apparent as essential factor, since the inhibitor is most likely unsuccessful in targeting BRAF homodimers or BRAF-CRAF heterodimers. Activation may occur when the inhibitor is not able to effectively inhibit both protomers of the dimer, leaving the unoccupied protomer active. It has also been suggested (e.g. for vemurafenib) that the inhibitor may induce a transactivation of the unoccupied protomer within the dimer as a result of a conformational change.[133,142,148]

To avoid this paradoxical effect and to increase treatment efficacy and finally survival time, combination therapies have been developed that target both BRAF and downstream MEK. The FDA has approved three combinations BRAF and MEK1/2 inhibitors for the treatment of advanced melanoma with a BRAFV600E mutation: vermurafenib and cobimetanib, dabrafenib and trametinib, and encorafenib and binimetinib.[133] Dabrafenib in combination with trametinib is since recently also approved for metastatic non-small cell lung cancer harboring BRAF V600E mutations.[149]

To date, it can be distinguished between three generations of RAF kinase inhibitors. The first generation contains only one inhibitor that is approved by the FDA, sorafenib, a biarylurea derivative.

As it has demonstrated only weak affinity for the mutated BRAF-V600 and shows a broad specificity, the clinical effects are supposed to arise from multikinase targeting. The second generation includes vemurafenib and dabrafenib that show remarkable effectiveness in BRAF-V600E tumors, but no effect on non-V600 mutants and the paradoxical activation in WT cells may lead to secondary cancers. Both drugs vemurafenib and dabrafenib demonstrate a type I1/2 binding mode with BRAF in αC-out/DFG-in conformation. They occupy the type I subpocket with their sulfonamide head group and H-bonding to the DFG motif, the hydrophobic subpocket is occupied by the adjacent di-/fluorophenyl group, and the adenine subpocket by the azaindole/aminopyrimidyl moiety forming H-bonds with the backbone of hinge residues. Their binding differs only with respect to the occupation of the ribose subpocket, which is only occupied by the butylthiazol group of dabrafenib. Encorafenib, another second generation inhibitor, shows a particularly low off-rate from BRAF compared to vemurafenib and dabrafenib,[150] leading to a longer residence time and increased selectivity for BRAF-V600E.[151] Encorafenib, containing also a sulfonamide head group, is expected to show a similar binding mode as vemurafenib and dabrafenib, type I1/2, with the BRAF conformation αC-out/DFG-in.

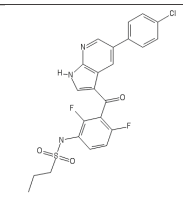
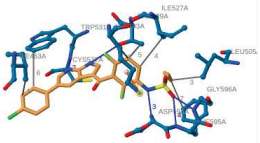
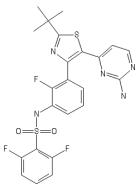
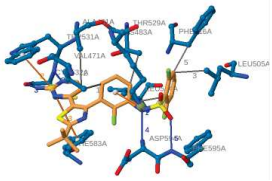
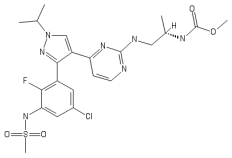
Drug name	**Chemical structure**	**Binding mode**	**PDB codes** (Resolution)	**in vitro** $IC50/K_i$ **[nM]**
Vermurafenib			3OG7 (2.45Å) 4RZV (2.99Å)	31
Dabrafenib			4XV2 (2.5Å) 5CSW (2.66Å) 5HIE (3.0Å)	0.8
Encorafenib		-	-	0.3

Table 1.1: Details of the three FDA approved second generation BRAF inhibitors vermurafenib, dabrafenib and encorafenib, with affinities for BRAF-V600E.[142]

Pharmacokinetics and metabolic drug-drug interactions of dabrafenib

Dabrafenib is a potent and selective inhibitor for BRAF-V600E, but it has been found that its bioavailability decreases rather rapidly (with a half-life of ∼5 hours[152]), which is likely due to induction of its own metabolism through cytochrome P450s (CYPs).[152–158] It is estimated that CYP-mediated oxidation contributes to over 70% to the metabolism of dabrafenib in vivo.[154] The metabolic pathway of dabrafenib (DB) and its three identified major metabolites hydroxy-dabrafenib (HDB), carboxy-dabrafenib (CDB), and desmethyl-dabrafenib (DDB) is depicted in Figure 1.2. Dabrafenib metabolism is mediated by CYP3A4 and CYP2C8. DB is metabolized to HDB (by

CYP3A4 and CYP2C8), further oxidized to CDB (by CYP3A4), and decarboxylated to DDB (pH-dependant). CYP3A4 is also involved in further metabolism of DDB to minor oxidative metabolites.[153,155] The half-lifes of HDB, CDB, and DDB are estimated to 5.7, 17.5, and 20.4 hours, respectively (based on a body mass study).[152] CDB and DDB display elimination rate-limited pharmacokinetics, whereas the pharmacokinetic profile of HDB parallels the one of DB. The relative potency of of DB and its metabolites is found to be ranked as DB > HDB $\sim$ DDB $\gg$ CDB.[156] Furthermore, DB was shown to induce CYP3A4 and 2B6 in hepatocytes, and to inhibit CYP2C8, 2C9, 2C19, and 3A4 in human liver microsomes.[154] Thus, DB is supposed to be subject of drug-drug interactions with strong inhibitors of CYP2C8 and/or CYP3A4.[154–156] CYP3A4 and CYP2B6 mRNA induction is indicating interactions of DB with the nuclear receptors Pregnane X Receptor (PXR) and/or Constitutive Androstane Receptor (CAR).[156]

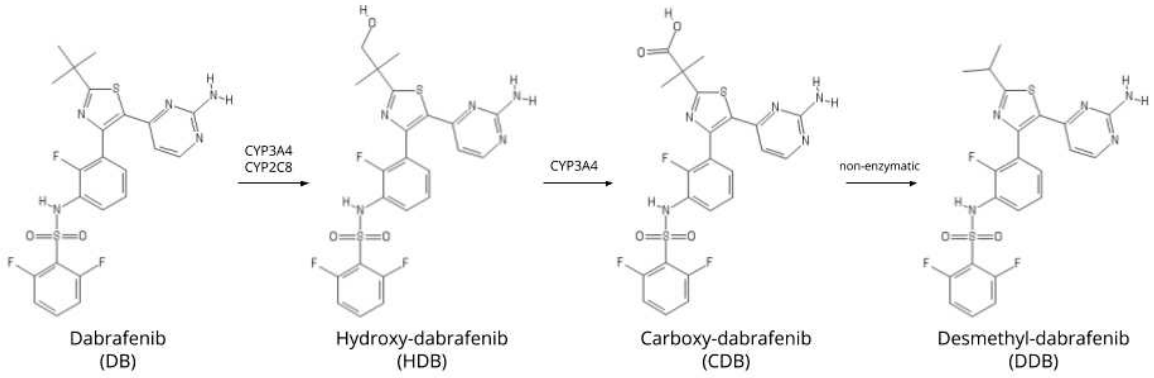


Figure 1.2: Metabolic pathway of dabrafenib (DB) and its three major metabolites hydroxy-dabrafenib (HDB), carboxy-dabrafenib (CDB), and desmethyl-dabrafenib (DDB). DB is metabolized by CYP3A4 and CYP2C8 to HDB, further oxidized to CDB (by CYP3A4), and decarboxylated to DDB (pH-dependant).[153]

1.3.2 Nuclear receptors

Nuclear receptors (NRs) are members of a large superfamily of evolutionary related DNA-binding transcription factors. NRs mediate the effects of hormones and other endogenous ligands to regulate the expression of specific genes and play an essential role in virtually all aspects of mammalian development, metabolism, and physiology. Therefore, their dysfunction and the subsequent aberrant signaling is associated with many diseases concerning reproduction, proliferation and metabolism. Due to their ligand binding ability they are of interest for a broad scientific field as potential pharmaceutical targets and for drug development and in toxicology and environmental science for risk assessment.[159]

Regarding the mechanisms of action, NRs can either be inactive cytoplasmic receptors which, upon ligand binding, are translocated to the nucleus, and activate gene transcription, or they are permanently located in the nucleus, which is the case for most NRs, and are often bound to DNA in the absence of any ligand.

Structural aspects

NRs have a modular structure consistent of the functional domains from the N to C termini: the variable modulator domain (referred to as A/B), the DNA-binding domain (DBD) (referred to as C), the variable hinge region (referred to as D), and the ligand-binding domain (LBD) (referred to as E).[160] Some NRs have additional extensions at the N- or C-terminus. The N-terminal modulator domain is the most variable domain and contains a ligand-independent transcriptional activation function, referred to as AF-1. A second ligand-dependent AF-2 surface is located in the C-terminal domain. Those activation domains are often responsible for mediating the binding of coactivators. The DBD and LBD can function independently and are connected by the very flexible hinge region.[160] The DBD is the region of highest sequence conservation including two zinc finger motifs and being responsible for direct recognition of the target DNA sequence.[161] The LBD, which is crucial for most of the receptor functions because it binds the ligand, performs dimerization and interacts with coregulators. The LBD has a general fold of a three-layered α-helical sandwich composed of 12 helices (H1-H12).

Current mechanistic view

Binding of agonist ligands to the ligand binding pocket of the LBD induces a conformation that preferentially binds coactivator proteins. In contrast, binding of antagonist ligands induces a conformation that prevents the binding of coactivator proteins and prefers the binding of corepressor proteins. Furthermore, NRs can also directly interact with other transcription factors to regulate gene transcription.

NRs are of high interest in the pharmaceutical field as both, primary target for direct treatment of diseases and secondary target for avoiding side effects that can be provoked e.g. by endocrine disruption or by induction of drug metabolism.

1.3.2.1 The Pregnane X Receptor (PXR)

The Pregnane X Receptor (PXR) belonging to NR subfamily I plays an unusual and outstanding role as master regulator for xenobiotic metabolism. It is responsible for the organism's defense against foreign substances and therefore a main regulator for detoxification, acting as sensor to a broad spectrum of ligands (endogenous metabolites, drugs and xenobiotics) with very diverse characteristics (concerning composition, shape and size). Unfortunately, undesired drug binding to PXR is causing many adverse effects. PXR forms heterodimers with the Retinoid X Receptor α (RXRα) and subsequently binds to PXR responsive elements. As main transcriptional inducer of cytochrome P450 enzyme CYP3A4, one of the main metabolizing enzymes for many drugs in clinical use, it acts as key player for inducing drug degradation and can potentially cause undesirable drug-drug interactions.[162] Rapid metabolism decreases efficacy for many drugs, but drugs with active metabolites can display increased drug effect and/or toxicity upon metabolism. Undesirable drug-drug interactions are also a metabolic issue. When two drugs sharing a metabolism pathway via the same enzyme compete for the same binding site, the one with higher potency predominates and the metabolism of the competing drug is decreased. This, in turn, can lead to increased risks

for toxic effects of the unmetabolized compound, as serum levels may be elevated. PXR is also widely expressed in many different tumors (breast, colon, prostate and ovary) where it has been shown to be involved in both the development of multi-drug resistance and enhanced cancer cells aggressiveness.[163–166] An increasing number of drugs are clinically tested in cancers with sometimes rather limited success and it was also shown recently that some of them could be direct ligands of PXR, thereby inducing their own metabolism or the metabolism of co-administered drugs. PXR is classified as unwanted and harmful secondary target whose activation needs to be avoided in order to simultaneously avoid the activation of the degradation pathway via CYP450 enzymes. Accordingly, a limited interaction with PXR is required additionally to a drug's efficient binding to its primary target. Therefore, an improvement of drugs includes a fine tuning with chemical changes that do not perturb other important characteristics, such as stability, bioavailability, etc., but prevent PXR binding. Nonetheless, PXR has not yet been studied extensively and only a limited number of 25 crystallographic structures are available.

1.3.2.2 The Estrogen Receptor alpha (ERα)

The Estrogen Receptors (ERs) belong to the type I nuclear receptors, also called steroid receptors, and occur in two subtypes, α and β. The Estrogen Receptor alpha (ERα) is among the most studied NRs. It has been linked to osteoporosis, breast cancer, prostate cancer, obesity, inflammation, menopausal problems and other diseases and is therefore an important target for medical treatment.[167] ERα and ERβ are similar in structure and sequence. They have 56% amino acid sequence identity in their LBD and the residues that surround the ligand are nearly identical, varying only at two positions. The LBD of the ER has the NR general fold of a three-layered α-helical sandwich including the 12 helices (H1-H12).[168] The ligand binding site is a mostly hydrophobic cavity located in the lower half of the domain. Upon ligand binding a conformational change is induced in the receptor that promotes translocation into the nucleus, homodimerization and subsequent binding to to hormone response elements within the promoter of a target gene in order to regulate transcription. However, the direct binding of an additional interaction partner, a coactivator protein, is needed for ligand-dependent signaling to occur.[168] In the case of agonist-bound structures the ligand-binding cavity is sealed by the C-terminal helix H12 which is then referred to as the active conformation. This conformation favors the recruitment of coactivators to the AF-2 surface. Cell-type and promoter-context dependent activity of both ER AF-1 and AF-2 has been demonstrated.[169] In case of antagonist-bound structures the sealing of the binding cavity by H12 is not possible, because the usually larger antagonists bind in such a mode that they reach further out of the binding cavity and occupy the space where H12 would be located in the agonist conformation.

Key points

⇒ All three therapeutic targets - BRAF, PXR and ERα - dimerize to exert their biological function. Nevertheless, they are studied as protomeric entities.

⇒ The high complexity of mode of actions does not yet permit to deal with all levels of flexibility, in particular dimerization and allosteric effects e.g. via cofactors.

⇒ The amount of available data and the level of flexibility for each target impact the applicability of diverse computational methods to predict/estimate binding affinities, including MD based MM-PBSA calculations or supervised machine learning approaches.

2

ERα - A WELL STUDIED TARGET

This chapter focuses on the largely studied drug target Estrogen Receptor alpha (ERα) that serves as ideal target for the development of an affinity prediction tool. Also the protein's intrinsic flexibility is further studied in detail.

2.1 Affinity prediction method development on ERα

An integrated affinity prediction method was developed by exploiting data from structure ensembles from the target's and the ligands' point of view and combined with a Random Forest (RF) machine learning algorithm. The complete method development, testing and validation is described in the publication "Towards accurate high-throughput ligand affinity prediction by exploiting structural ensembles, docking metrics and ligand similarity", in the journal *Bioinformatics*, included on the following pages.

Key points

⇒ SBVS and LBVS are combined by employing ML for affinity predictions.

⇒ Predictions are based on ensembles: structural (ligand and target ensembles) and computational (averages from multi-docking, consensus scoring, and RF as ensemble learner).

⇒ The presented approach is very general and is extended to other well characterized targets.

Bioinformatics, 2019, 1–9
doi: 10.1093/bioinformatics/btz538
Advance Access Publication Date: 26 July 2019
Original Paper

OXFORD

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

Structural bioinformatics

Towards accurate high-throughput ligand affinity prediction by exploiting structural ensembles, docking metrics and ligand similarity

Melanie Schneider*, Jean-Luc Pons, William Bourguet and Gilles Labesse*

Centre de Biochimie Structurale, CNRS, INSERM, Univ Montpellier, 34090 Montpellier, France

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on March 25, 2019; revised on May 29, 2019; editorial decision on June 18, 2019; accepted on July 19, 2019

Abstract

Motivation: Nowadays, virtual screening (VS) plays a major role in the process of drug development. Nonetheless, an accurate estimation of binding affinities, which is crucial at all stages, is not trivial and may require target-specific fine-tuning. Furthermore, drug design also requires improved predictions for putative secondary targets among which is Estrogen Receptor alpha (ERα).

Results: VS based on combinations of Structure-Based VS (SBVS) and Ligand-Based VS (LBVS) is gaining momentum to improve VS performances. In this study, we propose an integrated approach using ligand docking on multiple structural ensembles to reflect receptor flexibility. Then, we investigate the impact of the two different types of features (structure-based and ligand molecular descriptors) on affinity predictions using a random forest algorithm. We find that ligand-based features have lower predictive power ($r_P = 0.69$, $R^2 = 0.47$) than structure-based features ($r_P = 0.78$, $R^2 = 0.60$). Their combination maintains high accuracy ($r_P = 0.73$, $R^2 = 0.50$) on the internal test set, but it shows superior robustness on external datasets. Further improvement and extending the training dataset to include xenobiotics, leads to a novel high-throughput affinity prediction method for ERα ligands ($r_P = 0.85$, $R^2 = 0.71$). The presented prediction tool is provided to the community as a dedicated satellite of the @TOME server in which one can upload a ligand dataset in mol2 format and get ligand docked and affinity predicted.

Availability and implementation: http://edmon.cbs.cnrs.fr.

Contact: schneider@cbs.cnrs.fr or labesse@cbs.cnrs.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Despite the fact that the efforts invested in drug development have constantly increased during the last decades, the number of drug approvals stays almost constant (Munos, 2009). Indeed, about 81% of all new drug candidates fail (DiMasi *et al.*, 2010), mainly due to a lack of drug efficiency and/or side effects associated with off-target binding. In order to reduce time and cost of drug development process, various computer aided methods have been implemented. Two main techniques, namely structure-based and ligand-based virtual screening, are widely used (Lavecchia, 2015; Lionta *et al.*, 2014). They are now routinely used for hit identification in order to prioritize compounds for experimental assays and they are also gaining interest for lead optimization.

© The Author(s) 2019. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Ligand-based virtual screening (LBVS) methods are based on analyzing features of substructures and chemical properties related to activity of the ligand. They are useful to search chemical libraries using global or substructure similarity (Mestres and Knegtel, 2000), shape-matching (Nicholls *et al.*, 2010) or pharmacophores (Yang, 2010). The algorithms used in those methods are in constant development and recent LBVS methods are based on data mining and machine learning (Lavecchia, 2015). They do not require structural knowledge of the receptor.

Structure-based virtual screening (SBVS) can be used to predict the binding mode of drugs, to define the important specific interactions between ligand and target and finally to discover a way to improve a given drug by guiding further optimization. SBVS includes docking of candidate ligands into a protein target, followed by evaluation of the likelihood of binding in this pose using a scoring function with an important trade-off between speed and accuracy (Cerqueira *et al.*, 2015). Compared to LBVS, which is restricted to similar molecules the training had been performed on, SBVS is applicable to completely new molecules but it requires knowledge of the targeted structure (or reliable theoretical models). Moreover, very small changes or addition of the molecules that can create strong repulsions (e.g. steric clashes) are more likely to be identified by SBVS methods than by LBVS.

Combining LBVS with SBVS is emerging as a way to compensate limitations of each of these complementary approaches. Indeed, there are new attempts to combine both, thanks to the increasing number of both atomic structures and affinity measurements. Usually, the combination of LBVS and SBVS is performed in a sequential or parallel manner (Yu *et al.*, 2018; Zhang *et al.*, 2017). The sequential approach uses both methods as filter steps in a hierarchical procedure with increasing refinement. The parallel approach compares the selected compounds of both methods and retrieves either a consensus (selected by both) or a complementary selection (top molecules from each approach) (Lavecchia and Di Giovanni, 2013). Alternatively, one might apply a weak similarity restraint such as a molecular shape restraint for the ligand (to be classified as a shape-matching LBVS method) during the docking process in SBVS as it is implemented in the docking software PLANTS (Korb *et al.*, 2009).

In the present study, we take advantage of a new interface between PLANTS and the web server @TOME (Pons and Labesse, 2009) to screen multiple conformations in parallel (to be described in more details elsewhere). It also allows us to systematically deduce a shape restraint and binding site boundaries based on the geometry of the original ligand from each crystal structure in a fully automatic manner. Subsequent postprocessing is performed using various chemoinformatics tools including several scoring functions to predict protein–ligand affinity and select an optimal pose.

Ultimately, all the parameters computed to evaluate a ligand pose can be used for machine learning. Indeed, the combination of LBVS and SBVS with machine learning is an emerging approach to improve affinity prediction (Wójcikowski *et al.*, 2017). Therefore, we evaluate applicability of machine learning on the docking outputs of @TOME and PLANTS and ligand similarity measurements. In order to set up and evaluate this development, we focused on a well-known therapeutic target—the estrogen receptor ERα.

The ERα is a steroid binding receptor playing a key role in a variety of diseases due to its important role in development and physiology. The most prominent examples are ER-based cancer therapies that focus on blocking estrogen action in targeted tissues, with ERα being the main target for treatment of ER-positive breast cancer (Ma *et al.*, 2009). The development of new and improved selective ER modulators is therefore still of high interest for pharmaceutical companies to target tissues selectively and to avoid resistance and adverse effects (Baker and Lathe, 2018; Katzenellenbogen *et al.*, 2018; Wang *et al.*, 2018).

Moreover, ERα can also be an unwanted target of drugs or xenobiotics (Baker and Lathe, 2018; Delfosse *et al.*, 2012) and has been identified as an anti-target that should be considered in toxicity tests during drug development. Thus, a better understanding of the mechanism of ligand recognition by ERα is of paramount importance for safer drug design. Previously, dedicated prediction methods have been addressing the question of whether a molecule is binding or not (Mansouri *et al.*, 2016; Niu *et al.*, 2016; Pinto *et al.*, 2016; Ribay *et al.*, 2016), and traditional structure-activity relationship (QSAR) modeling studies have also been performed with varying success on this nuclear receptor (Asikainen *et al.*, 2004; Hou *et al.*, 2018; Waller *et al.*, 1995; Waller, 2004; Zhang *et al.*, 2013, 2017).

Despite the fact that ERα is an already well characterized therapeutic target (Ekena *et al.*, 1997; Nettles *et al.*, 2004), we are still lacking an efficient and robust method for predicting the binding mode and affinity of docked ligands. A large number of ERα crystal structures in complex with ligands are now known and the binding affinity of hundreds of chemical compounds have been experimentally determined. Therefore, ERα represents a perfect example to attempt a full characterization by combining SBVS with LBVS and employing machine learning in order to better predict binding affinity and potential future drug profiles.

2 Approach

Here, we present an integrated approach for high accuracy affinity predictions on the well-known and intensively studied drug target ERα. First, a training set and several testing sets were built by systematic docking of chemical compounds extracted from the BindingDB, the FDA and from in-house experiments, into the available crystal structures of ERα. An interesting feature of the approach is the fact that we take advantage of structural ensembles for the receptor and the ligand to simulate flexibility. Scoring functions and other chemometric information were gathered for the corresponding complexes through the @TOME server and for the ligands through the CDK. All virtual screening results are made available at http://atome4.cbs.cnrs.fr/htbin-post/AT23/MULTI-RUN/FILTER/showform.cgi? WD=AT23/EG/38751543. Seeing that the accuracy of scoring functions is not sufficient for fine ligands ranking, we employ a random forest machine learning algorithm on diverse features, including structure-based docking metrics and ligand-based molecular descriptors. We also tested various subsets of descriptors, such as MACCS fingerprints, and different algorithms. The developed prediction tool is provided to the community as an automatic prediction extension within the @TOME-EDMon server (http://edmon.cbs.cnrs.fr). The developed machine learning method is equally applied on and provided for ERβ and PPARγ.

3 Materials and methods

3.1 Ligand datasets

3.1.1 BindingDB dataset

Two sets of experimentally tested ligands for the human ERα (UniProtID: P03372) were extracted from BindingDB (Gilson *et al.*, 2016; Liu *et al.*, 2007) (2018 dataset, updated 2018-04-01). One set contains ligands with known inhibitory constant (Ki) as affinity

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

measure, and a second set contains ligands with half maximal inhibitory concentration (IC50) as an affinity proxy.

A few peptides and a series of boron cluster containing molecules were removed from both datasets, as it was not possible to generate proper 3D conformations or charges for these molecules. The final sets contained 281 ligands (Ki set) and 1641 ligands (IC50 set), respectively, with an overlap of 48 common compounds. Overall, both datasets show a similar compound diversity (compare Supplementary Fig. S1). For training, we preferred to focus on the Ki dataset since it corresponds to more direct measurements while the IC50 dataset was used as a larger dataset for method testing.

3.1.2 In-house xenobiotic dataset

The xenobiotic chemical data that was used first as an external testing dataset and afterwards to build an extended training set, is an in-house dataset of 66 ligands with measured affinities for ERα (Grimaldi *et al.*, 2015). These extra compounds correspond mostly to bisphenols, halogenated compounds, as well as phytoestrogens (natural fused heterocycles and macrocycles partially micmicking estradiol).

3.1.3 FDA ERα dataset

In order to have a second external validation, we used the Estrogen Receptor targeted dataset from the Endocrine Disruptor Knowledge Base (EDKB) provided by the U.S. Food & Drug Administration (FDA), named here ER-EDKB dataset. The dataset contains 130 ER binders and 101 non-ER binders including natural ligands and xenochemicals that are structurally different from drug-like molecules. For ER binders, the binding affinity measure is reported as a relative binding activity (RBA), which is based on an assay using rat uteri. Those cell-based measurements are influenced by different factors, such as cellular permeability, and are unfortunately not directly comparable with direct Ki measures. Nevertheless, we predicted affinities using all models and transformed the measured RBA values back to pIC50 values ($pIC50 = \log_{10}(RBA) - 8$).

Interestingly, affinity distributions of the datasets cover a wide range of about ten orders of magnitude without major gaps for distinct affinity ranges (see Supplementary Fig. S2).

3.2 Generation of ligand conformations

On the ligand side, there are two factors that can have an impact on docking. One is the initial conformation submitted to a docking program. Indeed, providing the bound conformation is a well-known bias to improve the success of a docking tool as previously recognized (Cross *et al.*, 2009; Plewczynski *et al.*, 2011). The generated ligand conformations can also differ significantly due to ambiguities in molecular descriptions (e.g. multiple conformations of heterocyclic alkyl moieties are possible from usual SMILES strings) and to the optimization procedure after *ab initio* building. Indeed, we notice that some steroid compounds highlighted improperly distorted conformation (data not shown). The second factor is the atomic partial charges that have an impact on ligand pose evaluation and can be calculated using different models (e.g. Gasteiger and MMFF94). In PLANTS, the atomic partial charges affect hydrogen bond donor/acceptor properties (e.g. for aromatic carbons) impacting significantly the docking itself and hence its subsequent scoring.

The initial ligand sets were downloaded from BindingDB (BDB) and have 3D conformations generated by VConf and partial charges generated by VCharge (Chang and Gilson, 2003). We also tested two other charge models (Gasteiger and MMFF94 charges) instead of the default charge for the 3D conformers built by Vconf.

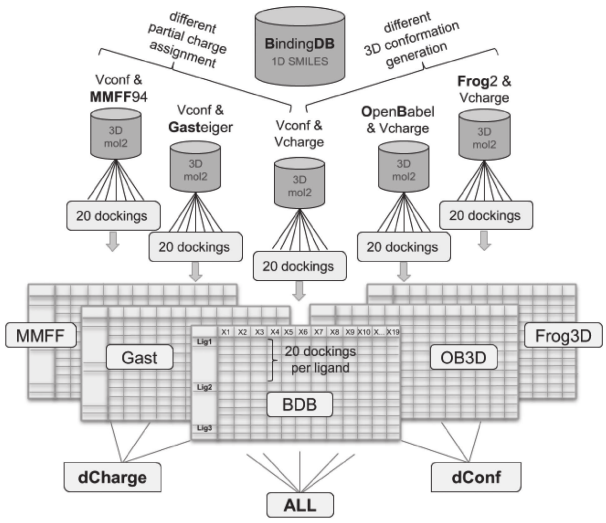


Fig. 1. Structure-based dataset generation approach. The ligand dataset was extracted from the BindingDB (BDB), which uses VConf for 3D conformation generation and VCharge for charge assignment. Two more partial charge models (MMFF and Gasteiger) and two other 3D conformation generators (OpenBabel and Frog2) were employed to generate a total of five ligand sets. Those were submitted to the @TOME server for docking and complex evaluation. The @TOME output datasets 'MMFF', 'Gast', 'BDB', 'OB3D' and 'Frog3D' (containing the results of 20 dockings per ligand in different structures) were grouped in three combined datasets, a different charge dataset 'dCharge', a different conformation dataset 'dConf', and an 'ALL' dataset

Two other 3D generators [OpenBabel (O'Boyle *et al.*, 2011) and Frog2 (Miteva *et al.*, 2010)] using their default charge. This resulted in a total of five ligand sets. The ligand sets were then grouped based on variation on their 3D generation, their charges or all together as depicted in Figure 1.

Noteworthy, by training on distinctly generated datasets in parallel we prevent dependencies and bring more versatility. Consequently, the user would be able to provide compounds without the need for further conversion that could possibly introduce errors.

3.3 Structure-based ligand docking

3.3.1 Ensemble docking

First, all liganded ERα structures available in the PDB (461 monomers) were gathered using the @TOME server (Pons and Labesse, 2009) by submitting the 'canonical' amino acid sequence of ERα (UniProt identifier: P03372-1) with a specified sequence identity threshold of 90%. All gathered 461 monomers had a sequence identity between 95 and 100% with the submitted sequence and correspond to point mutants of the human ERα. Missing or substituted side-chains were modeled using SCWRL 3.0 (Wang *et al.*, 2008) using the strictly conserved side-chains fixed. By default, for each ligand to be docked (e.g. from BDB), a set of 20 different template structures were automatically selected among all available PDB structures. This selection is based on the highest similarity (Tanimoto score) between the uploaded ligand and the co-crystallized ligand present in a template. The automatic virtual screening procedure implemented in the @TOME server uses the docking program PLANTS with its shape restraint functionality (with a weighting of -3, the default value suggested by the software manual), using the original ligand of the screened structure as a pharmacophore. Of note, this ligand is also used to define the

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

boundaries of the binding site to be screened (using a distance cutoff of 8 Å). So, not only the protein conformation is (slightly) distinct but various cavity volume and extent are used in this parallel docking procedure. For each template screened, only one pose was computed by PLANTS. After docking and structure alignment, the 20 computed poses were clustered by conformation similarity, and the most likely pose is selected automatically among the largest cluster using a dedicated heuristics. Accordingly, we perform ligand docking on conformational ensemble as an optimal procedure for SBVS.

3.3.2 Structure-based molecular descriptors

Each docking pose is evaluated by various chemoinformatics tools (see Table 1). Here, in order to predict protein–ligand affinities, we take advantage of several re-scoring functions [namely MedusaScore (Yin *et al.*, 2008), DSX (Neudert and Klebe, 2011) or X-SCORE (Wang *et al.*, 2002)) recently embedded in @TOME to derive a consensus score [including also ChemPLP as used in PLANTS (Korb *et al.*, 2006)]. Here, we used both, raw output from these scoring functions, and linear regression based on a study of PDBbind (to be described elsewhere). In addition, other evaluations are performed on the server, such as the model quality of the receptor [QMean (Benkert *et al.*, 2008)] and the evaluation of the ligand conformation [such as LPC (Sobolev *et al.*, 1999) or AMMP energy computed by AMMOS (Pencheva *et al.*, 2008)]. Other scores measure the similarity between the docked ligand and the pharmacophoric anchor used to guide the docking. For instance, AnchorFit (as computed by PLANTS) evaluates their shape similarity, and the Tanimoto score (computed by OpenBabel) evaluates compositional similarity. Finally, we also implemented two new scoring metrics, one comparing ligand–receptor interactions as a sequence-based profile (named PSim for Profile Similarity), and the other predicting a pose RMSD based on a support vector machine (named LPE for Ligand Position Error). These evaluation metrics will be described in more detail elsewhere.

The above parameters were important for structure-based screening, and they were complemented by other information regarding the chemical nature of ligands using additional molecular descriptors.

Table 1. Structure-based docking metrics

Metric name	Short description
PlantsFull	PLANTS score (with anchor weight) (Korb *et al.*, 2006)
Plants	PLANTS ChemPLP score (without weight)
PlantsLR	PLANTS pKa (calculated by linear regression on PDBbind)
MedusaScore	Medusa original score (Yin *et al.*, 2008)
MedusaLR	MedusaScore pKa (calculated by linear regression on PDBbind)
XScore	XScore affinity score (pKa) (Wang *et al.*, 2002)
DSX	DSX original score (Neudert and Klebe, 2011)
DSXLR	DSX pKa (calculated by linear regression on PDBbind)
AtomeScore	@TOME pKa = mean(PLANTS, XScore, MedusaScore, DSX)
Tanimoto	Similarity between candidate ligand and anchor ligand
AtomSA	S.A. @TOME score
QMean	QMean score of receptor model
AnchKd	Affinity calculated between receptor/anchor (pKa)
AnchorFit	Candidate/ligand superimposition score (PLANTS software)
LigandEnergy	Internal energy of ligand (AMMP force field)
LPC	LPC software score (receptor/ligand complementarity function)
PSim	Similarity to receptor/ligand interaction profile in PDB template
CpxQuality	Complex quality consensus score
LPE	Ligand Position Error (SVM multi-variable linear regression)

3.3.3 Ligand molecular descriptors

In order to include more information about the small molecules being screened, molecular descriptors were calculated using the Chemistry Development Kit (CDK) (Willighagen *et al.*, 2017), a collection of open source Java libraries for chemoinformatics, through its R interface rcdk (Guha, 2007). The descriptors were selected based on their ability to represent the diversity of the ligand dataset, taking into account their orthogonality, and based on their variable importance score during model training. The final set of 11 QSAR molecular descriptors includes topological, geometrical, constitutional and charge based descriptors and is listed in Table 2 with CDK descriptor name, the used abbreviation and a short description.

3.3.4 Combined structure/ligand descriptors

All 5 docking datasets (originating from the 5 different ligand sets) provided 19 structure-based docking metrics for the 20 docking poses computed for each ligand. For each metric, median and standard deviation were computed and used as a unified instance. Ligand-based variables (11 CDK molecular descriptors) were added to the 19 structure-based metrics. A correlation matrix with all descriptors used for the Ki-BDB dataset is provided as heatmap (see Supplementary Fig. S5). Alternatively, the commonly used MACCS fingerprints (166 features) (Durant *et al.*, 2002; Taylor, 2007) were also tested for comparison.

3.4 Machine learning approaches

3.4.1 Algorithm selection and training

For all analyses, calculations and machine learning, the R language (version 3.2.4) was used with RStudio (employed packages are listed in Supplementary Table S1). In an initial test on the BDB Ki dataset we assessed the performance of 7 machine learning algorithms (see Supplementary Fig. S3): Random Forest (RF), Gradient Boosting

Table 2. Ligand-based molecular descriptors

Abbrv.	CDK descriptor name	Short description
MW	Weight	molecular weight
VABC	VABC	volume descriptor
nAtom	AtomCount	number of atoms
nBond	BondCount	number of bonds
nRotBond	RotatableBondsCount	number of rotatable bonds
nAromBond	AromaticBondsCount	number of aromatic bonds
nHBDon	HBondDonorCount	number of hydrogen bond donors
nHBAcc	HBondAcceptorCount	number of hydrogen bond acceptors
TPSA	TPSA	Topological Polar Surface Area
XLogP	XLogP	prediction of logP based on the atom-type method called XLogP
HybRatio	HybridizationRatio	fraction of sp3 carbons to sp2 carbons

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

Machine (GBM), support vector machine (SVM) with a radial kernel (SVM_R), a polynomial kernel (SVM_P) and a linear kernel (SVM_L), linear regression (LinReg), decision tree (CARTree) and Partial Least Squares (PLS). All algorithms were employed with default variable settings with the R package 'caret'. In order to avoid over-fitting of the models, we used stratified 10-fold cross validation repeated 10 times for all models in this study (unless otherwise indicated).

Alternatively, an external test set was built by taking a stratified selection of 20% of the whole dataset. The remaining 80% was used as training set for the models.

3.4.2 Comparison of different tree-based algorithms

The RF algorithm we used, has only one tunable hyperparameter that can be adjusted for the present dataset. Therefore, we wondered whether other tree-based ensemble algorithms with more tunable hyperparameters offer an improved prediction accuracy when tuned more carefully. In total, five different tree-based algorithms were employed on the same Ki BindingDB2018 dataset for affinity prediction and subsequent performance comparison. They are: random forest (RF), regularized random forest (rRF), global regularized random forest (rRFglobal), Extreme Gradient Boosted Trees (xgbTree) and Extreme Gradient Boosted Trees with dropout (xgbDART). Here, Bayesian optimization was employed to select the best hyperparameters (5 to 7 depending on the method), which demands a substantially increase in computational expense compared to the one-variable optimization required for the RF algorithm. The performance of the different models are compared based on the left out data during cross-validation (see Supplementary Fig. S4).

3.5 Random Forest regression modeling

Random forest models were trained on each dataset separately ('MMFF', 'Gast', 'BDB', 'OB3D' and 'Frog3D'), on the combination of the three different 3D conformation datasets ({'BDB', 'OB3D', 'Frog3D'} = 'dConf'), on the combination of the three different partial charge datasets ({'MMFF', 'Gast', 'BDB'} = 'dCharge') and on all five datasets combined ('ALL') (compare Fig. 1).

Besides the Pearson correlation (r_P), two further regression evaluation metrics were used to evaluate the model performance on the external test set. First, the coefficient of determination (R^2) is calculated using the sum of squares method. The second metric, the Root Mean Square Error (RMSE) is the average deviation of the predictions (predicted affinities) from the observations (measured affinities). In some cases, we also indicate the Spearman rank correlation (r_S).

4 Results and discussion

We developed and tested an automated and integrated structure- and ligand-based approach to predict quickly accurate binding affinities for ERα. This approach takes into account structural variability from the ligand side by using different 3D generators and different charge models, and from the receptor side by using 20 structures for each ligand to be docked. Here, we give access to the docking poses while we evaluate thoroughly the affinity predictions performed using various methods.

4.1 Predictions using re-scoring methods

In a first attempt, the predictive power of the four different scoring functions implemented in the @TOME server was assessed.

Table 3. Pearson correlations (r_P) on all five datasets between experimental affinities and scores from four scoring functions Plants, MedusaScore, DSX and XScore, of (1) the best pose selected by @TOME, and of (2) the median scores of the four scoring functions, calculated on 20 dockings per ligand on all five datasets

Dataset name	Plants	MedusaScore	DSX	XScore
(1)	r_P on predictions for the best pose			
Gast	0.042	0.154	0.129	0.060
MMFF	0.063	0.182	0.157	0.082
BDB	0.038	0.111	0.118	0.076
OB3D	0.109	0.180	0.143	0.129
Frog3D	0.022	0.132	0.118	0.040
(2)	r_P on predictions over 20 poses			
Gast	−0.031	0.204	0.019	0.049
MMFF	−0.025	0.192	0.038	0.054
BDB	−0.019	0.087	0.022	0.059
OB3D	−0.017	0.175	0.008	0.050
Frog3D	−0.048	0.199	0.005	0.036

The Pearson correlations between affinity measurement and the median scores (calculated on 20 docking poses per ligand) are very low for all generated datasets (see Table 3). Even the most recent scoring functions (MedusaScore and DSX) performed poorly in this test. Interestingly, the selection of the best pose among the 20 computed ones slightly improves the correlation between predicted and measured affinities for 3 scoring functions but for MedusaScore which appeared as the most robust and the best for the various ligand description schemes.

However, the overall correlation is too low for fine ligand affinity prediction and indicates a limitation of the general-purpose scoring functions, but the better Spearman correlation (see Supplementary Table S2) suggested a better ranking that could be useful to guide machine learning. This prompted us to develop a more sophisticated method that should be able to combine advantages of different docking evaluations (structure-based and ligand-based ones) and potentially take into account specific features.

4.2 Random Forest regression—model training

First, we did some model optimization using parameter tuning, variable selection and engineering (e.g. to better take into account rotatable bounds, see below).

4.2.1 Structure-based and ligand-based partial models

To investigate the actual affinity prediction capabilities of structure-based and ligand-based variables, partial models were trained using the 19 structure-based metrics or the 11 ligands-based metrics on the same dataset named BDB. Affinity predictions made on the held-out 20% test set are shown in Supplementary Figure S6. The docking-metrics only model ($r_P = 0.78$, $r_S = 0.77$ and $R^2 = 0.60$) outperforms the molecular-descriptors only model ($r_P = 0.69$, $r_S = 0.70$ and $R^2 = 0.47$). Interestingly, the combined model (trained on BDB using simultaneously Vconf and Vcharge data) has Pearson correlation coefficient, Spearman's rank and R^2 value in between the reduced-variable models ($r_P = 0.73$, $r_S = 0.74$ and $R^2 = 0.50$).

4.2.2 Random Forest model trained on MACCS fingerprints

In this context, it might be interesting to add more information regarding the chemical nature of the ligands studied. Instead of using a reduced set of ligand-based parameters, we turned to use a more

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

thorough description based on an extended and popular fingerprints: MACCS. A new random forest model was trained on MACCS fingerprints representing the ligands only, without providing any structural docking data. This resulted in a Pearson correlation (r_P) of 0.76, Spearman's coefficient (r_S) of 0.76 and an R^2 of 0.57 on the Ki test set, midway between the two partial models compared above (molecular descriptors only model and docking metrics only model). Combining MACCS with docking-based features improves the overall performance on the training and left-out testing dataset ($r_P = 0.81$, $r_S = 0.82$ and $R^2 = 0.61$) but further evaluation using external datasets suggested some overfitting (see below).

4.2.3 Combined models—trained on single and multiple combined datasets

Then, we compared the various models trained on either single datasets ('MMFF', 'Gast', 'BDB', 'OB3D' and 'Frog3D') or multiple combined datasets ('nf', 'dCharge' and 'ALL'). Whereas the five models trained on single datasets have an R^2 of 0.66 (±0.01), an RMSE of 0.82 (±0.01) and an explained variance of 63.4 (±0.8) during training, the three models trained on multiple datasets have a better R^2 of 0.68 (±0.004), a lower RMSE of 0.78 (±0.008) and an explained variance of 90.6 (±3.7). Also evaluation on the 20% left-out test set demonstrates improved predictions for the models trained on multiple combined datasets ('dConf', 'dCharge' and 'ALL') with a mean Pearson correlation (r_P) of 0.77 (and standard deviation of 0.014), compared to the models trained on single datasets with a mean r_P of 0.73 (and standard deviation of 0.029).

The boosted tree models xgbTree and xgbDART appeared to outperform slightly the RF model on this 'ALL' dataset. But the reverse was true when evaluating the corresponding models onto the FDA dataset (see below). Most of the differences are weak and may not be significant. Accordingly, the more complex implementations did not provide significant increase in performance and they were not studied further.

4.3 Random Forest regression—model testing

Most remarkable is the strong increase in accuracy when using either the 'dConf' model trained on the three different 3D conformation datasets ('BDB', 'OB3D' and 'Frog3D') or the model trained on the fully combined 'ALL' dataset comprising all five datasets ('MMFF', 'Gast', 'BDB', 'OB3D' and 'Frog3D') (compare also Supplementary Fig. S7). Interestingly, using different charge models improves affinity predictions, but slightly less efficiently than using different 3D conformations. This is probably due to the fact that the binding pocket of ERα is mostly hydrophobic and therefore the ligands show the same property and partial charges are predominantly found to differ only marginally.

4.4 Analysis of variable importance

To assess the impact of the various parameters from structure-based and ligand-based scoring functions, the variable importance was tracked during training of the RF models. The 30 most important variables for the models trained on the 'ALL' dataset is shown in Figure 2. Overall, all models have a rather similar variable importance profile (data not shown).

Noteworthy, the most important variable *'Tanimoto_Med'* is the same for all trained models showing its outstanding importance. It represents the median Tanimoto score calculated between the docked ligand and the 20 shape restraints (or 'anchors') present in the targeted structure. This may reflect the importance of using structures bound to similar ligand to ensure proper affinity predictions.

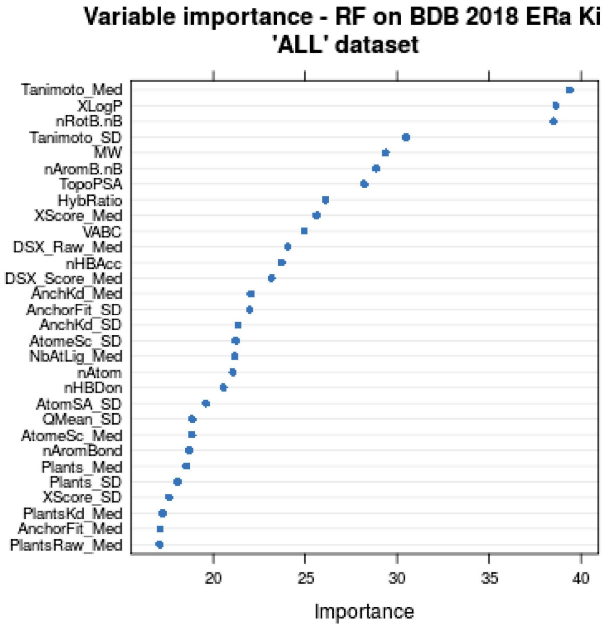


Fig. 2. Variable importance of the top 30 variables, tracked during model training for the model trained on the 'ALL' dataset with the full variable set. Structure-based docking metrics have an extension (*_Med* or *_SD*). The suffix *_Med* stands for the calculated median of the variable for a ligand's 20 dockings and *_SD* is the respective standard deviation of this variable

The second and third most important variables are *'nRotB.nB'* and *'XLogP'*. *'nRotB.nB'* estimates ligand flexibility, deduced from the number of rotatable bonds *'nRotB'* and the total number of bonds *'nB'* by simply dividing them (*'nRotB'/'nB'*). During variable testing, this combined variable showed an increased importance compared to the original variables (data not shown), which were therefore removed for the final model training. The particular importance of *'nRotB.nB'* indicates the important role of entropy cost for binding flexible ligands. Obviously, this parameters is not easily handled in a systematic manner by general scoring functions while it is an important parameter for affinity predictions. In the particular case of ERα, it likely discriminates rather small and rigid agonists from larger and more flexible antagonists to prevent overestimating the affinity of the latter. In agreement, the fifth variable is the molecular weight (*'MW'*) which may also compensate for the additive terms of most scoring functions dedicated to affinity predictions.

Another predominantly important and high-rank variable (second in the 'ALL' model and third in the 'dCharge' and 'dConf' models) is *'XLogP'*. Representing hydrophobicity and solubility of the ligand, it is expected to be an important factor with respect to the mainly hydrophobic binding pocket of ERα. Moreover, *'XLogP'* may reflect solvent-driven entropic effects that are not easily taken into account by usual scoring functions. Indeed, flexibility and solvation-linked metrics can be regarded as useful for a crude estimate of some entropic effects and counterbalance the enthalpy-oriented affinity prediction approach of usual scoring functions.

Finally, the different scoring functions (DSX, Plants, MedusaScore and X-score; through their means and standard deviations) show a smaller importance than the three above parameters, which could be in agreement with the poor correlations described above. It may also arise from the intrinsic redundancy of our selected variables as several affinity predictions are used in parallel.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

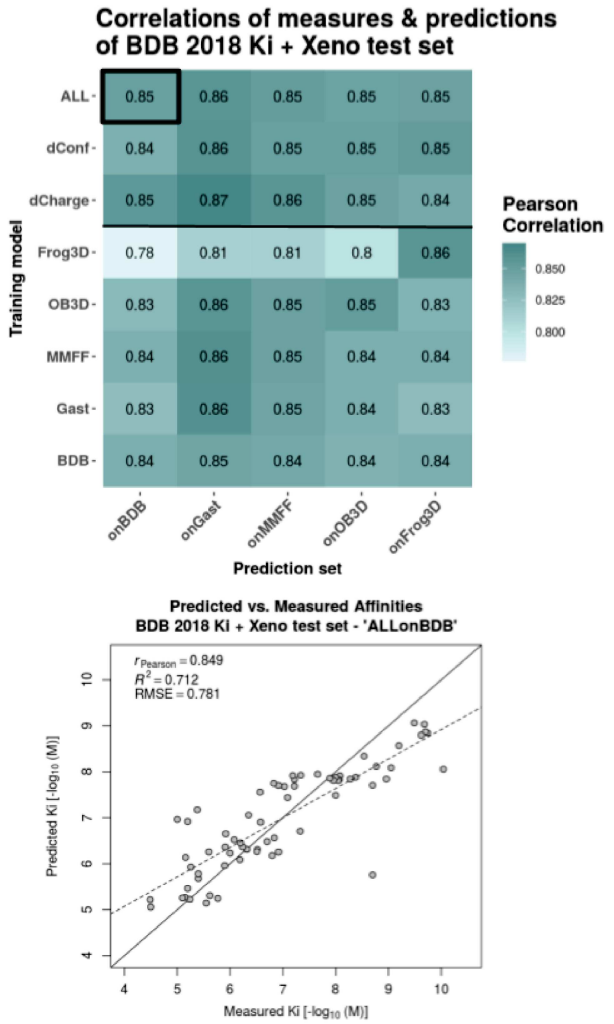


Fig. 3. Performance evaluation of extended models on their respective 20% left-out test sets. The initial dataset of 281 ligands is extended by a set of 66 xenochemicals. The heatmap shows Pearson correlations between predictions and measures for all combinations of training model and prediction set. The different training models are listed as rows and the test sets, on which the predictions were made, are listed as columns. RF models were trained on each dataset separately ('MMFF', 'Gast', 'BDB', 'OB3D', 'Frog3D'), on the combination of the three different 3D conformation datasets ({'BDB', 'OB3D', 'Frog3D'} = 'dConf'), on the combination of the three different partial charge datasets ({'MMFF', 'Gast', 'BDB'} = 'dCharge') and on all five datasets combined (= 'ALL'). The predictions with the Pearson correlation highlighted in the heatmap (black box) is plotted as scatter-plot for details below. The scatter plot shows the actual predicted versus measured affinities together with a regression line (dashed line), the optimal prediction line (solid diagonal) and the evaluation metrics—Pearson correlation coefficient (r_P), coefficient of determination (R^2) and root-mean-square error (RMSE). All evaluation metrics were calculated with respect to the actual values (solid diagonal), not the regression line

Overall, this result underlines the importance of developing dedicated models for each target under investigation, in order to account for some specific features including particular desolvation and flexibility properties.

4.5 Model evaluation on different datasets

4.5.1 Model evaluation on an in-house xenobiotic dataset

We took advantage of a complementary and independent dataset—the xenobiotic chemical data of 66 ERα binders to evaluate the robustness of our models. Our models performed rather poorly on this dataset with a Pearson correlation (r_P) of 0.48 for the best Random Forest model BDB-Ki (and 0.40 with the BDB-Ki+MACCS model). For the partial models, docking-metrics-only and molecular-descriptors-only, as well as the MACCS-only model, correlations are even lower with r_P of 0.45, 0.31 and 0.13, respectively. This underlines the improved robustness of the BDB-Ki model combining SBVS and LBVS features (compare also Supplementary Fig. S8). Importantly, the chemical nature of most xenobiotics differs significantly from most of the drug-like compounds from the BDB dataset used for training. As such, small xenobiotics (including the small bisphenols) occupy only partially the hydrophobic cavity and often also present numerous halogen substitutions (that are notoriously hard to model). Furthermore, for some of the small xenobiotics we cannot rule out the possibility that two molecules may bind simultaneously (with synergetic effects). This result prompted us to combine these xenobiotics and BDB Ki dataset into an extended training set to build a new RF model with improved performance (Fig. 3).

Table 4. Model performances on the FDA ER-EDKB test set

Algorithm	Training set	Variable type	Pearson correlation
RF	ALL+Xeno	@TOME+LD	0.748
RF	ALL+Xeno	@TOME+LD+MACCS	0.740
RF	ALL	@TOME+LD	0.663
RF	ALL	@TOME+LD+MACCS	0.648
RF	BDB+Xeno	@TOME+LD	0.712
RF	BDB+Xeno	@TOME+LD+MACCS	0.688
RF	BDB	@TOME+LD	0.584
RF	BDB	@TOME+LD+MACCS	0.542
RF	BDB+Xeno	MACCS only	0.487

Note: The presented models employ all the RF algorithm and differ in training set composition concerning used molecules and in type of variables used. @TOME+LD = docking evaluation variables from the @TOME server + ligand descriptors calculated with CDK.

4.5.2 Model evaluation on FDA ER-EDKB dataset

We then evaluated our two best models on a reference dataset comprising both 322 drug-like and xenobiotic compounds (see Table 4 and Supplementary Table S3). At the first glance, the predictions made using the original model (trained on only BDB-Ki) showed a lower performance especially on the edges of the affinity ranges with both overestimated affinities for small and weak binders (e.g. alkylphenol) and underestimated predictions for tight binders such as rigid and compact agonists. Indeed, the BDB dataset is mainly composed of large and high-affinity antagonists. Accordingly, some FDA compounds such as high affinity agonists, appear as strong outliers.

Most remarkable is the benefit of adding a complementary dataset of 66 xenobiotic compounds to the initial 281 ligands from BindingDB (see Table 4). Here, the best Pearson correlation (r_P) of 0.75 is attained with the model trained on 'ALL' datasets including the xenobiotics and the model trained on a single dataset (BDB) including the xenobiotics also shows a high r_P of 0.71. Accordingly, the nature and diversity of the ligands matter, so that, proper coverage of the studied chemical space, in the training dataset compared to the testing one is essential. The model trained on a single dataset without the xenobiotics has already a decreased r_P of 0.58, whereas the partial models, docking-metrics-only and molecular-descriptors-

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

only, and the MACCS-only model, show poor performances with r_P of 0.49, 0.47 and 0.41, respectively (compare Supplementary Fig. S9). This underlines again the increased robustness of our feature type combination.

4.5.3 Model evaluation on BindingDB—IC50 dataset

Finally, the most extended and reliable dataset we used for evaluating our RF models is the BindingDB 2018 IC50 dataset which includes 1641 entities. Interestingly, the model trained on the Ki dataset already performed well against IC50 data suggesting a strong robustness.

Training and testing the IC50 dataset (1641 compounds versus 281 for the Ki dataset) also provides some insights into dataset size requirements for the studied target. First, the performance on the IC50 test set ($r_P = 0.87$) is better than on the Ki dataset ($r_P = 0.77$) (compare Table 5). Then, cross-predictions were computed by either using the model constructed on the Ki dataset for predictions on the IC50 dataset, or employing the model constructed on the IC50 dataset for predicting the Ki dataset. In that case, it seems that the small Ki test set (56 compounds) does not allow optimal validation as it shows a significant drop in performance compared to the Ki training set (0.49 versus 0.64). On the contrary, the Ki-ALL model showed similar performance on both the IC50 training and testing sets (1319 versus 322 entities).

We also evaluated our last model trained on the extended dataset including both the Ki dataset and the xenobiotic dataset on the largest available IC50 dataset from BindingDB (compare Table 6). Good predictions were observed for the IC50 dataset although the addition of the xenobiotic dataset did not bring any improvement (nor any deterioration) for that particular dataset. For comparison of all trained model see Supplementary Figure S10. Again, these results suggests that our final model is rather robust.

5 Conclusion

We provide an original *in silico* method for accurate binding affinity predictions that takes advantage of structural ensembles, of a limited number of structure-based metrics (19) and of ligand-based descriptors (11) in a unique combination. This combination led to a prediction tool outperforming our other models based either on SBVS or on LBVS features when we take into account not only the overall performance on the internal testing set but also the robustness on a range of distinct datasets. This is true also with the use of many more features as exemplified here with the MACCS fingerprints (166 bits). Our work also confirmed the performance of Random Forest over other machine learning approaches as previously noticed (Russo *et al.*, 2018). In some cases, higher accuracy was reported but for smaller compound libraries (Hou *et al.*, 2018). Accordingly, our results present one of the largest validation surveys (1641 ligands from the BDB IC50 dataset) and best performing tools for affinity prediction against ERα. As major advantages, RF algorithms handle non-linearities, numerical and categorical variables, and they give estimates of variable importance and generalization error.

By training our model in parallel on various types of partial charges and/or 3D builders, we believe our tool will be more versatile and robust to variations in the way the submitted compound libraries are generated. Noteworthy, the user has simply to upload one single dataset to EDMon, where the submitted chemical compounds will automatically be docked and their theoretical affinity for ERα be computed. With this tool, one can easily and rapidly evaluate new compounds either to find putative binders of ERα or to check the absence of binding to this frequent secondary target, in order to avoid potential side-effects.

Interestingly, the same approach yields very similar performances on two other nuclear receptors (ERβ and PPARγ) (see Supplementary Figs S11 and S12, respectively) and their automatic affinity prediction is also implemented in EDMon. The method also provided excellent results for a protein-kinase (to be described elsewhere) and we see no reason for any limitation as soon as dozens or hundreds of structures and affinity points are known. Areas for further improvements are probably: increasing the accuracy in ligand docking, a possible addition of complementary evaluation metrics for the protein–ligand interactions, as well as using deep learning. Testing challenging compounds is also an important way to guide improvement and we expect our web server to be thoroughly tested with novel compounds.

Table 5. Comparison of cross-predictions between the Ki and IC50 models and datasets

	BDB Ki training set	BDB Ki test set	BDB IC50 training set	BDB IC50 test set
number of compounds	225	56	1319	322
Ki ALL model	0.99	0.77	0.64	0.69
IC50 ALL model	0.64	0.49	1.00	0.87

Note: Pearson correlations between experimental affinities and the random forest predictions are reported.

Table 6. Evaluation of best RF models on various datasets

Prediction set	Xeno	FDA	IC50	Ki
RF ALL model				
Ki+Xeno	0.98	0.75	0.65	0.96
Ki	0.48	0.66	0.65	0.77*
IC50	0.25	0.35	0.87*	0.61

Note: Pearson correlations between experimental affinities and the RF predictions are reported for the whole datasets but for values marked with '*' that indicates values for a 20% test set.

Acknowledgements

We thank Muriel Gelin, Corinne Lionne, Dominique Douguet, Matteo Paloni and Rafaela Salgado for careful reading of the manuscript. We are grateful for the helpful comments from the referees.

Funding

This work was supported by the CNRS, INSERM and the University of Montpellier. This project has also received funding from the EU Horizon 2020 research and innovation program under grant agreement GOLIATH [825489] and from the ANSES (EST-2016/1/162-XENomix).

Conflict of Interest: none declared.

References

Asikainen,A.H. *et al.* (2004) Consensus kNN QSAR: a versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. *Environ. Sci. Technol.*, **38**, 6724–6729.

Baker,M.E. and Lathe,R. (2018) The promiscuous estrogen receptor: evolution of physiological estrogens and response to phytochemicals and endocrine disruptors. *J. Steroid Biochem. Mol. Biol.*, **184**, 29–37.

Benkert,P. *et al.* (2008) QMEAN: a comprehensive scoring function for model quality assessment. *Proteins*, **71**, 261–277.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

Cerqueira,N.M.F.S.A. *et al.* (2015) Receptor-based virtual screening protocol for drug discovery. *Arch. Biochem. Biophys.*, **582**, 56–67.

Chang,C.-E. and Gilson,M.K. (2003) Tork: conformational analysis method for molecules and complexes. *J. Comput. Chem.*, **24**, 1987–1998.

Cross,J.B. *et al.* (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.*, **49**, 1455–1474.

Delfosse,V. *et al.* (2012) Structural and mechanistic insights into bisphenols action provide guidelines for risk assessment and discovery of bisphenol A substitutes. *Proc. Natl. Acad. Sci. USA*, **109**, 14930–14935.

DiMasi,J.A. *et al.* (2010) Trends in risks associated with new drug development: success rates for investigational drugs. *Clin. Pharmacol. Therap.*, **87**, 00362.

Durant,J.L. *et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, **42**, 1273–1280.

Ekena,K. *et al.* (1997) Different residues of the human estrogen receptor are involved in the recognition of structurally diverse estrogens and antiestrogens. *J. Biol. Chem.*, **272**, 5069–5075.

Gilson,M.K. *et al.* (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, **44**, D1045–D1053.

Grimaldi,M. *et al.* (2015) Reporter cell lines for the characterization of the interactions between human nuclear receptors and endocrine disruptors. *Front. Endocrinol.*, **6**, 62.

Guha,R. (2007) Chemical informatics functionality in R. *J. Stat. Softw.*, **18**, 1–16.

Hou,T.-Y. *et al.* (2018) Insight analysis of promiscuous estrogen receptor α-ligand binding by a novel machine learning scheme. *Chem. Res. Toxicol.*, **31**, 799–813.

Katzenellenbogen,J.A. *et al.* (2018) Structural underpinnings of oestrogen receptor mutations in endocrine therapy resistance. *Nat. Rev. Cancer*, **18**, 377–388.

Korb,O. *et al.* (2006) PLANTS: application of ant colony optimization to structure-based drug design. In: *Ant Colony Optimization and Swarm Intelligence*. Springer, pp. 247–258.

Korb,O. *et al.* (2009) Empirical scoring functions for advanced protein–ligand docking with PLANTS. *J. Chem. Inf. Model.*, **49**, 84–96.

Lavecchia,A. (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today*, **20**, 318–331.

Lavecchia,A. and Di Giovanni,C. (2013) Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.*, **20**, 2839–2860.

Lionta,E. *et al.* (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.*, **14**, 1923–1938.

Liu,T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.

Ma,C.X. *et al.* (2009) Predicting endocrine therapy responsiveness in breast cancer. *Oncology (Williston Park, N.Y.)*, **23**, 133–142.

Mansouri,K. *et al.* (2016) CERAPP: collaborative estrogen receptor activity prediction project. *Environ. Health Perspect.*, **124**, 1023–1033.

Mestres,J. and Knegtel,R.M.A. (2000) Similarity versus docking in 3D virtual screening. *Perspect. Drug Discov. Des.*, **20**, 191–207.

Miteva,M.A. *et al.* (2010) Frog2: efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Res.*, **38**, W622–W627.

Munos,B. (2009) Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.*, **8**, 959–968. 00701.

Nettles,K.W. *et al.* (2004) Allosteric control of ligand selectivity between estrogen receptors α and β: implications for other nuclear receptors. *Mol. Cell*, **13**, 317–327.

Neudert,G. and Klebe,G. (2011) DSX: a knowledge-based scoring function for the assessment of protein–ligand complexes. *J. Chem. Inf. Model.*, **51**, 2731–2745.

Nicholls,A. *et al.* (2010) Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.*, **53**, 3862–3886.

Niu,A.-Q. *et al.* (2016) Prediction of selective estrogen receptor beta agonist using open data and machine learning approach. *Drug Des. Dev. Ther.*, **10**, 2323–2331.

O'Boyle,N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminf.*, **3**, 33–00943.

Pencheva,T. *et al.* (2008) AMMOS: automated molecular mechanics optimization tool for in silico screening. *BMC Bioinformatics*, **9**, 438.

Pinto,C.L. *et al.* (2016) Prediction of estrogenic bioactivity of environmental chemical metabolites. *Chem. Res. Toxicol.*, **29**, 1410–1427.

Plewczynski,D. *et al.* (2011) Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.*, **32**, 742–755.

Pons,J.-L. and Labesse,G. (2009) @TOME-2: a new pipeline for comparative modeling of protein–ligand complexes. *Nucleic Acids Res.*, **37**, W485–W491.

Ribay,K. *et al.* (2016) Predictive modeling of estrogen receptor binding agents using advanced cheminformatics tools and massive public data. *Front. Environ. Sci.*, **4**, 12.

Russo,D.P. *et al.* (2018) Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. *Mol. Pharmaceut.*, **15**, 4361–4370.

Sobolev,V. *et al.* (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics (Oxford, England)*, **15**, 327–332.

Taylor,K.T. (2007) chminf-l@list.indiana.edu—*description of public MACCS keys*. https://list.indiana.edu/sympa/arc/chminf-l/2007-11/msg00058.html (6 May 2019, date last accessed).

Waller,C.L. (2004) A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. *J. Chem. Inf. Comput. Sci.*, **44**, 758–765.

Waller,C.L. *et al.* (1995) Using three-dimensional quantitative structure-activity relationships to examine estrogen receptor binding affinities of polychlorinated hydroxybiphenyls. *Environ. Health Perspect.*, **103**, 702–707.

Wang,L. *et al.* (2018) New class of selective estrogen receptor degraders (SERDs): expanding the toolbox of PROTAC degrons. *ACS Med. Chem. Lett.*, **9**, 803–808.

Wang,Q. *et al.* (2008) SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.*, **3**, 1832–1847.

Wang,R. *et al.* (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.*, **16**, 11–26.

Willighagen,E.L. *et al.* (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.*, **9**, 33.

Wójcikowski,M. *et al.* (2017) Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.*, **7**, 46710.

Yang,S.-Y. (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov. Today*, **15**, 444–450.

Yin,S. *et al.* (2008) MedusaScore: an accurate force-field based scoring function for virtual drug screening. *J. Chem. Inf. Model.*, **48**, 1656–1662.

Yu,M. *et al.* (2018) Discovering new PI3kα inhibitors with a strategy of combining ligand-based and structure-based virtual screening. *J. Comput. Aided Mol. Des.*, **32**, 347–361.

Zhang,L. *et al.* (2013) Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol. Appl. Pharmacol.*, **272**, 67–76.

Zhang,X. *et al.* (2017) Computational insight into protein tyrosine phosphatase 1b inhibition: a case study of the combined ligand- and structure-based approach. *Comput. Math. Methods Med.*, **2017**, 1.

Zhao,Q. *et al.* (2017) Rational design of multi-target estrogen receptors ERα and ERβ by QSAR approaches. *Curr. Drug Targets*, **18**, 576–591.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz538/5539694 by guest on 07 August 2019

Supplement

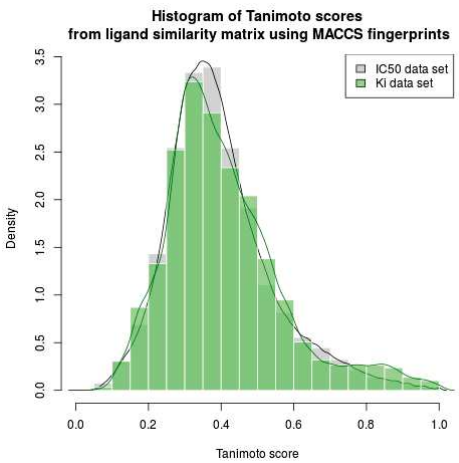


Figure S1. Compound diversity of the BDB-IC50 data set and the BDB-Ki data set, evaluated with Tanimoto score on MACCS fingerprints.

Table S1. R packages used for computation and visualization

Package name	Version	Short description from package authors
plyr	1.8.4	Tools for Splitting, Applying and Combining Data
dplyr	0.7.8	A Grammar of Data Manipulation
tidyr	0.8.2	Easily Tidy Data
ggplot2	3.1.0	Create Elegant Data Visualisations
stringr	1.3.1	Wrappers for Common String Operations
lattice	0.20-35	Trellis Graphics for R
caret	6.0-81	Classification and Regression Training
randomForest	4.6-14	Random Forests for Classification and Regression
rcdk	3.4.7.1	Interface to the 'CDK' Libraries
rpubchem	1.5.10	An Interface to the PubChem Collection
doSNOW	1.0.16	Foreach Parallel Adaptor for the 'snow' Package
magrittr	1.5	A Forward-Pipe Operator for R
tidyverse	1.2.1	A system of packages for data manipulation
corrr	0.3.0	Correlations in R
scales	1.0.0	Scale Functions for Visualization

Table S2. Spearman's rank correlation coefficients (r_S) on all five datasets between experimental affinities and scores from four scoring functions Plants, MedusaScore, DSX and XScore, of (1) the best pose selected by @TOME, and of (2) the median scores of the four scoring functions, calculated on 20 dockings per ligand on all five datasets.

Dataset name	Plants	MedusaScore	DSX	XScore
(1)	r_S on predictions for the best pose			
Gast	0.130	0.381	0.187	0.105
MMFF	0.151	0.391	0.180	0.106
BDB	0.137	0.357	0.204	0.102
OB3D	0.158	0.310	0.108	0.121
Frog3D	0.057	0.232	0.106	0.020
(2)	r_S on predictions over 20 poses			
Gast	0.083	0.154	0.164	0.085
MMFF	0.092	0.150	0.197	0.096
BDB	0.121	0.105	0.200	0.091
OB3D	0.134	0.127	0.127	0.087
Frog3D	0.027	0.139	0.135	0.059

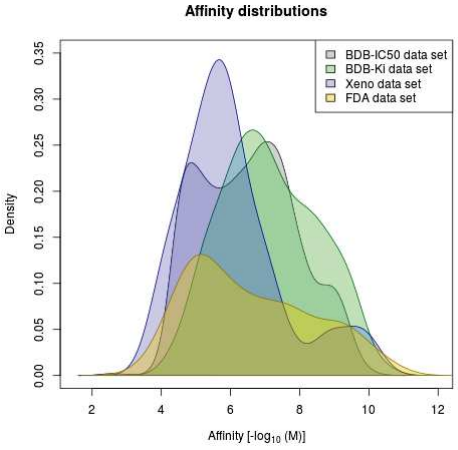


Figure S2. Affinity distribution of all 5 data sets: BDB-IC50 data set, BDB-Ki data set, Xeno data set, and FDA data set (with 1641, 281, 66, and 130 compounds, respectively).

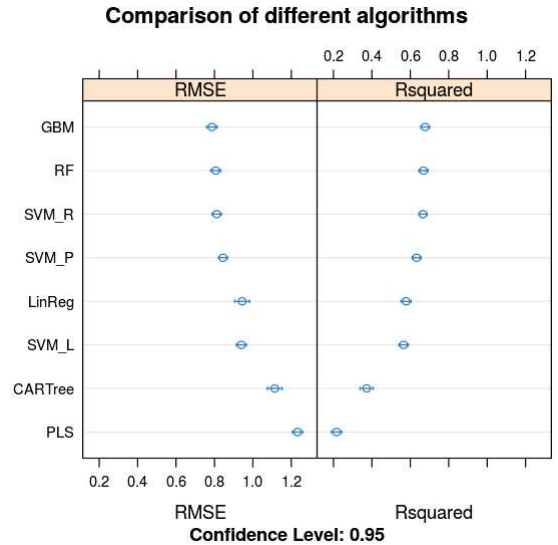


Figure S3. Performance comparison of different algorithms for binding affinity prediction (regression) - as dotplot. Models are ranked according to their cross-validation performance: general boosted machine (GBM), random forest (RF), support vector machine with radial kernel (SVM_R), with polynomial kernel (SVM_P), linear regression (LinReg), SVM with linear kernel (SVM_L), Classification and regression tree (CARTree) and Partial Least Squares (PLS).

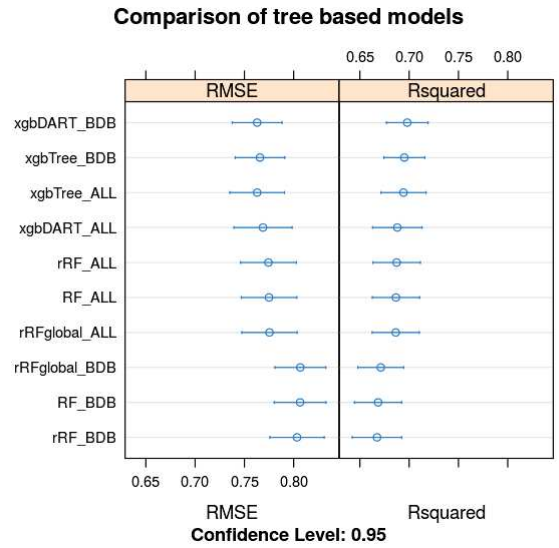


Figure S4. Comparison of five different tree-based algorithms with performance metrics calculated on cross-validation samples. All algorithms are trained on the 'BDB' dataset and on the combined 'ALL' dataset as indicated by the respective suffix after the algorithm name.

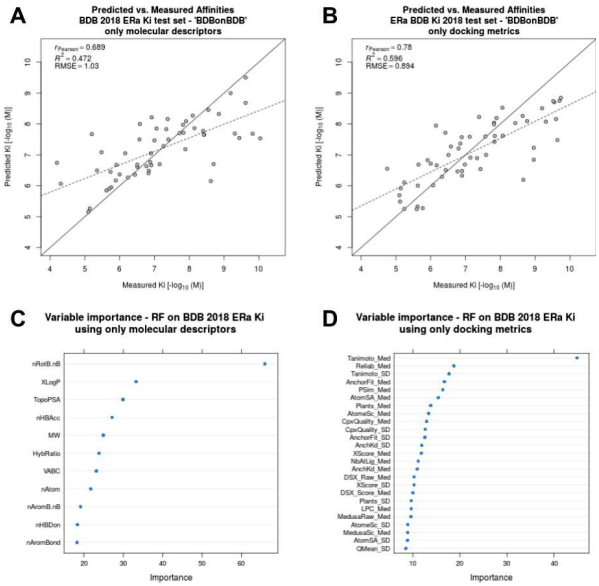


Figure S6. Correlations between measured and predicted affinities for the external Ki test set. The predictions were generated by models that were trained on a subset of descriptors. A) model trained only on the set of ligand-based molecular descriptors, and B) model trained only on the set of structure-based metrics from the @TOME server. C) and D) show the ranked variable importance to the trained models A) and B), respectively.

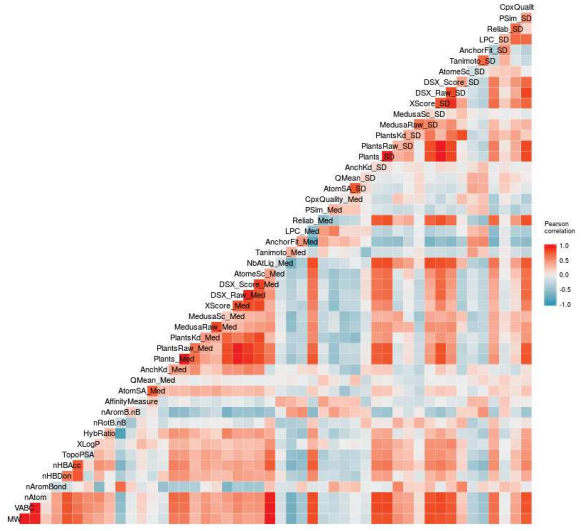


Figure S5. Descriptor correlation matrix for the Ki-BDB dataset as heatmap.

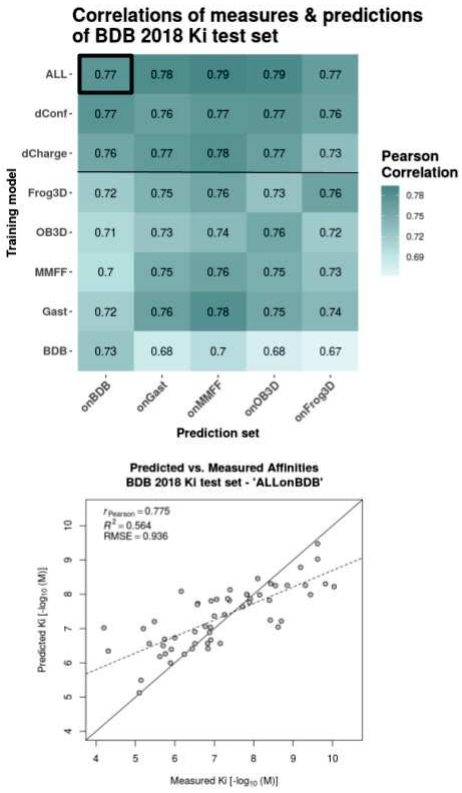


Figure S7. Correlations between measured and predicted affinities for the external Ki test set. The heatmap shows Pearson correlations between predictions and measures for all combinations of training model and prediction set. The different training models are listed as rows and the test sets, on which the predictions were made, are listed as columns. Random forest models were trained on each dataset separately ('MMFF', 'Gast', 'BDB', 'OB3D', 'Frog3D'), on the combination of the 3 different 3D conformation datasets ({'BDB', 'OB3D', 'Frog3D'} = 'dConf'), on the combination of the 3 different partial charge datasets ({'MMFF', 'Gast', 'BDB'} = 'dCharge'), and on all 5 datasets combined (= 'ALL'). For one prediction set the scatter plot below shows the actual predicted versus measured affinities together with a regression line (dashed line), the optimal prediction line (solid diagonal) and the

evaluation metrics - Pearson correlation coefficient (r_P), coefficient of determination (R^2) and root-mean-square error (RMSE). All evaluation metrics were calculated with respect to the actual values (solid diagonal), not the regression line.

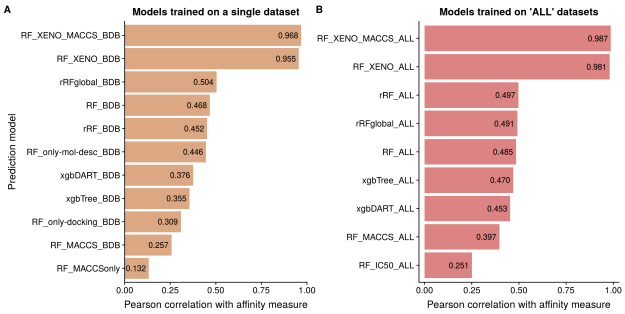


Figure S8. Correlations between measured and predicted affinities for the in-house xenobiotic dataset (66 compounds). The prediction models are named by 'algorithm_trainingset'.

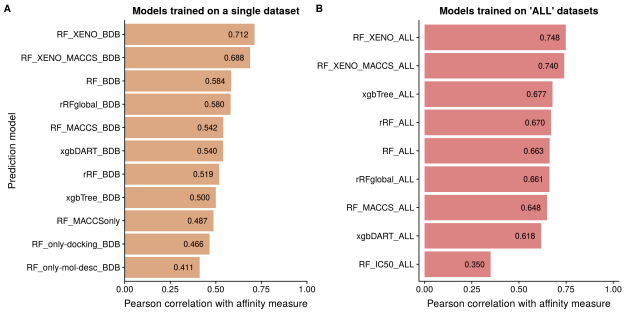


Figure S9. Correlations between measured and predicted affinities for the FDA ER-EDKB dataset (131 compounds). The prediction models are named by 'algorithm_trainingset'.

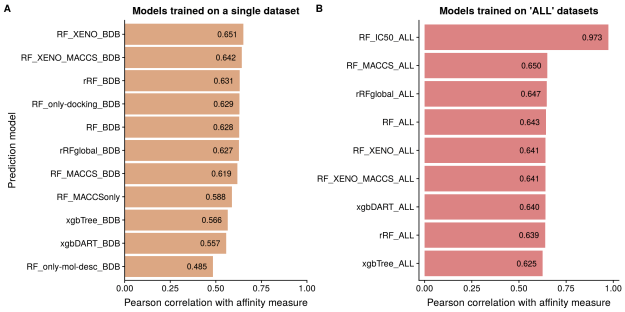


Figure S10. Correlations between measured and predicted affinities for the BindingDB-2018 IC50 dataset (1640 compounds). The prediction models are named by 'algorithm_trainingset'.

Table S3. Model performances on the FDA ER-EDKB test set. The presented models differ in algorithm usage, amount of cross-validation folds (by default 10-fold unless indicated differently as 3-CV), and training set composition concerning used molecules. The type of variables used remains unchanged. @TOME+LD = docking evaluation variables from the @TOME server + ligand descriptors calculated with CDK.

Algorithm	Training set	Variable type	Pearson correlation
xgbTree	ALL	@TOME+LD	0.677
rRF	ALL	@TOME+LD	0.670
RF	ALL	@TOME+LD	0.663
RF(3-CV)	ALL	@TOME+LD	0.661
rRFglobal	ALL	@TOME+LD	0.661
xgbDART	ALL	@TOME+LD	0.618
RF(3-CV)	BDB	@TOME+LD	0.592
RF	BDB	@TOME+LD	0.584
rRFglobal	BDB	@TOME+LD	0.580
xgbDART	BDB	@TOME+LD	0.540
rRF	BDB	@TOME+LD	0.519
xgbTree	BDB	@TOME+LD	0.500

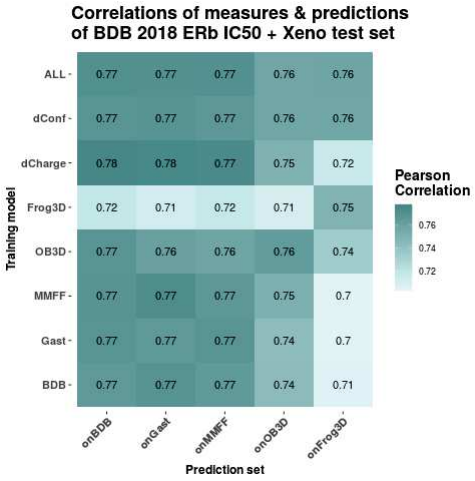


Figure S11. Correlations between measured and predicted affinities for the ERβ IC50 + XENO test set. The heatmap shows Pearson correlations between predictions and measures for all combinations of training model and prediction set. The different training models are listed as rows and the test sets, on which the predictions were made, are listed as columns. Random forest models were trained on each dataset separately ('MMFF', 'Gast', 'BDB', 'OB3D', 'Frog3D'), on the combination of the 3 different 3D conformation datasets ({'BDB', 'OB3D', 'Frog3D'} = 'dConf'), on the combination of the 3 different partial charge datasets ({'MMFF', 'Gast', 'BDB'} = 'dCharge'), and on all 5 datasets combined (= 'ALL').

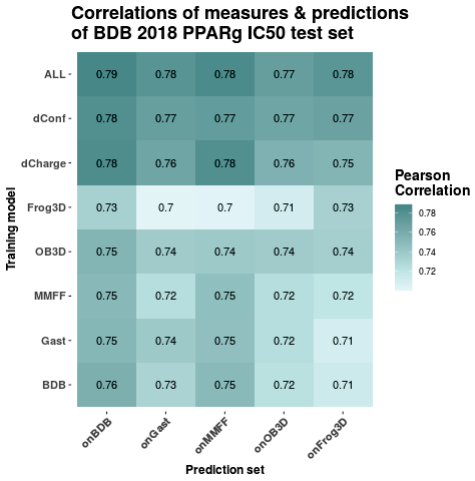


Figure S12. Correlations between measured and predicted affinities for the PPARγ IC50 test set. The heatmap shows Pearson correlations between predictions and measures for all combinations of training model and prediction set. The different training models are listed as rows and the test sets, on which the predictions were made, are listed as columns. Random forest models were trained on each dataset separately ('MMFF', 'Gast', 'BDB', 'OB3D', 'Frog3D'), on the combination of the 3 different 3D conformation datasets ({'BDB', 'OB3D', 'Frog3D'} = 'dConf'), on the combination of the 3 different partial charge datasets ({'MMFF', 'Gast', 'BDB'} = 'dCharge'), and on all 5 datasets combined (= 'ALL').

2.2 The flexibility universe of ERα

Protein structure flexibility was probed for the drug target ERα using different computational techniques that revealed particular aspects of distinct conformational states (agonist vs antagonist) on a global (secondary structure) and local (binding pocket side-chain) level.

Key points

⇒ A detailed comparison of differently generated structure ensembles is provided (set of static X-ray structures, refined ensembles, MD simulations, NMA).

⇒ The presented approach is very general and can be extended to other well characterized targets.

The flexibility universe of ERα probed with ensemble analysis and virtual screening

Melanie Schneider [1,*], Jean-Luc Pons [1] and Gilles Labesse [1,*]

[1]Centre de Biochimie Structurale (CBS), CNRS, INSERM, Univ Montpellier, 34090 Montpellier, France.

ABSTRACT

Motivation: Protein flexibility is challenging for both experimentalists and modellers and represents an emerging issue especially for drug design. Estrogen Receptor alpha (ERα) is a well-known therapeutic target with an important role in development and physiology and an extensively studied Nuclear Receptor (NR). It is also one of the main off-targets in standard toxicity tests trying to detect endocrine disruption. Here, we aim to evaluate the possibility to accurately describe the conformational space and macromolecular flexibility of this well-characterized drug target by exploiting information from hundreds of crystallographic structures, molecular dynamics simulations and exhaustive virtual screening.

Results: The analysis of hundreds of crystal structures enables us not only to distinguish two main conformational states 'agonist' and 'antagonist', but also to highlight the most frequent local flexibility patterns. Indeed, besides the large reorientation of the C-terminal helix H12, the receptor showed small loop rearrangements as well as side-chain displacements in the active site. Interestingly, standard molecular dynamics simulations and crystal structure refinement as ensemble recapitulate most of the features involved in loop mobility detected by crystal structure superpositions. In parallel, we investigate on the kind and extent of flexibility that is required to achieve convincing docking for all high affinity ERα ligands present in BindingDB. Exhaustive docking for these ERα ligands is achieved by incorporating flexibility in two ways: either by using one antagonist conformation but including side-chain flexibility or, by using parallel docking on conformational ensembles. Both approaches result in precise and highly similar pose predictions. Accordingly, we identified the minimal and optimal flexibility requirements to accommodate all known high affinity ligands. Moreover, the identified focused flexibility is in agreement with the overall conformational landscape available for ERα. The molecular details provided here could guide new drug design strategies for ERα, both as a primary and as secondary target.

Contact: schneider@cbs.cnrs.fr, labesse@cbs.cnrs.fr

Keywords: ERα, flexibility, crystallographic structures, ensembles, molecular dynamics, virtual screening

1 INTRODUCTION

In the physiological cellular environment, proteins are dynamic, which is often crucial for their function and activity. For a long while, structure-based virtual screening of small organic compounds has been performed on single rigid conformations of the target. However, various biophysical characterizations of biological macromolecules have illustrated their intrinsic flexibility - at various scales. The protein binding pocket often adapts to accommodate an entering ligand or a certain conformation of the conformational space of the protein is stabilized by the bound ligand. Erickson et al. [1] showed that docking accuracy falls off dramatically if an "average" or apo structure is used instead of an experimental crystal structure with a bound ligand.Those conformational changes can range from minor movements of single side-chains to large shifts of whole secondary structures or even domains. Therefore, the experimentally obtained crystal structures can be regarded as static snapshots of the whole dynamic conformational space of the protein. Theoretical models have been developed to simulate those conformational changes but accuracy and speed are still important limitations for general use in combination with virtual screening. In parallel, the number of atomic structures for many therapeutic targets of interest has been growing, bringing a very detailed view of their structure but also their conformational variability, although crystal packing can represent a severe limitation in that case. Nevertheless, refinement as structure ensemble demonstrate some structural breathing even in crystal structures solved at 100K. While some proteins show large structural rearrangements it seems that most (potentially two-thirds [2]) experience only limited variations with low overall RMSD (less than 3 Å) between known end-points. In this context, it is still unknown whether the conformation space accessible for structurally well-characterized proteins can be already described accurately and efficiently for drug design and virtual screening. Indeed, improving virtual screening will require to better model protein conformational flexibility. This would open up access to better entropy estimation on top of standard enthalpy computation, although the latter is sometimes complemented with imperfect extrapolation of the entropy part. In order to do so, the use of structure-ensembles is gaining popularity nowadays. Nevertheless, extracting the most fruitful target conformations is not yet straightforward and easily feasible in routine.

Docking involves a trade-off between the speed of the docking algorithm and its accuracy. Therefore, by adding flexibility to the protein structure used for docking higher accuracy can be achieved, but this also adds noise to the computation and usually increases the computational cost intensively. Especially in large-scale virtual database screening this comes in to play a role, since due to the high number of compounds to be screened there is a practical limit of available computational time per compound [3]. Unfortunately, the degree of required flexibility is not known in beforehand for new ligand types. All this underlines the fact that target flexibility represents one of the greatest challenges for docking programs.

One way to circumvent the problem of small rearrangements is to perform "soft docking". Implemented in several docking programs, it allows a small overlap of the ligand and the receptor by reducing the actual volume of the atom spheres [4]. Unfortunately,

© .

this could introduce errors, such as the detection of false positives, and it also does not even account for slightly larger conformational changes, such as side-chain rotations. Nowadays, the ability to include side-chain flexibility (for a limited amount of side-chains) by using libraries of preferred conformational states (e.g. sets of torsion angles) is fortunately implemented in several docking programs, such as PLANTS [5], which is used in this study. Another way to take into account the flexibility of a macromolecular structure is to built an ensemble of static models, called "ensemble docking". The ensemble structures can be generated, for example, by molecular dynamics (MD) simulation [6, 7]. This addresses both flexibility problems, the flexibility of the receptor and of the ligand, at once. Additionally, it should avoid a bias towards one protein conformation while implicitly including protein flexibility. One approach is to perform a long MD simulation that could help to sample the conformational space of the receptor prior to docking [8]. The difficulty in ensemble docking lies in the selection of appropriate target structures e.g. from a MD trajectory. One attempt to select relevant target conformations is normal mode analysis, which has been demonstrated to be an effective tool [9]. Unfortunately, additional noise is introduced by each extra conformation added to an ensemble, which may mask/counterbalance the beneficial information it provides. Therefore, the choice of the most appropriate receptor conformations is key for the success of the VS experiments and for the results to be representative. This problem highlights the need for clear guidelines to select the experimental structures that should compose an ensemble.

Thorough analysis of available experimental structures (especially high resolution crystal structures) remain to be performed to evaluate the conformational space eventually observed in those frozen conformations, and compare it with the one actually explored by the protein. Docking successfully all known high-affinity ligands could be a step toward delimiting the type and number of conformations truly accessible by a given target. The approach discussed above was implemented here on a well-characterized therapeutic target with hundreds of crystal structures already solved as well as hundreds of ligands with high affinity among which many lack an experimentally observed binding mode.

The Estrogen Receptor alpha (ERα) is among the most studied NRs. It has been linked to osteoporosis, breast cancer, prostate cancer, obesity, inflammation, menopausal problems and other diseases and is therefore an important target for medical treatment [10]. NRs have a modular structure consistent of the functional domains from the N to C termini: the variable modulator domain (referred to as A/B), the DNA-binding domain (DBD) (referred to as C), the variable hinge region (referred to as D), and the ligand-binding domain (LBD) (referred to as E) [11]. The LBD, which is crucial for most of the receptor functions because it binds the ligand, performs dimerization and interacts with coregulators. The LBD of the ER has the NR general fold of a three-layered α-helical sandwich including the 12 helices (H1-H12) [12]. The ligand binding site is a mostly hydrophobic cavity located in the lower half of the domain. Upon ligand binding a conformational change is induced in the receptor that promotes homodimerization and subsequent binding to hormone response elements within the promoter of a target gene in order to regulate transcription. However, the direct binding of an additional interaction partner, a coactivator or corepressor protein, is needed for ligand-dependent signaling to occur [12]. In the case of agonist-bound structures the ligand-binding cavity is sealed by the C-terminal helix H12 which is then referred to as the active conformation. This conformation favors the recruitment of coactivators (having a short leucine-rich motif) to the AF-2 surface. Cell-type and promoter-context dependent activity of both ER AF-1 and AF-2 has been demonstrated [13]. In case of antagonist-bound structures the sealing of the binding cavity by H12 is not possible, because the usually larger antagonists bind in such a mode that they reach further out of the binding cavity and occupy the space where H12 would be located in the agonist conformation. This antagonist conformation prevents the binding of coactivators and leads to preferential binding of corepressors (having a longer leucine-rich motif) [14]. Nevertheless, besides full agonists or antagonists there is a third group of molecules, called selective estrogen receptor modulators (SERMs) that show mixed agonistic/antagonistic behavior depending on the tissue [15, 16, 14]. Thus, SERMs can show agonist behavior in a tissue rich in coactivators but have antagonist effects in tissues rich in corepressors [15, 17, 14, 10]. SERMs were the first examples of selective modulators identified [18], but due to the development of resistances [16] and severe adverse effects there is still a high need for developing new SERMs [15]. This reveals a complex mechanism of regulation despite minor apparent changes besides H12 relocation that is also investigated by Srinivasan et al. [19]. Nonetheless, a possible full coverage of the conformational space is not interrogated.

2 APPROACH

In the present study we aim to probe the complete conformational space of the promiscuous nuclear receptor ERα based on freely available experimental data. We are addressing the problem from two sides: from the protein side by exploiting all available crystallographic data and performing MD simulations, and from the ligand side by making use of the known ligand space.

2.1 Exhaustive structure ensemble analysis

As a first approach for probing the receptor's flexibility all available crystallographic structures complexed with a ligand are analyzed as ensemble. Light is shed on the whole receptor, ranging from a global perspective to a focused binding site analysis by taking into account differences of the two dominant conformations, agonist and antagonist.

2.2 Molecular dynamics for ensemble generation

As a second approach 5 replicas of 50ns molecular dynamics (MD) simulations are performed on three liganded ERα complexes in agonist and antagonist conformation (PDB-IDs: 2YJA, 2OUZ and 3UUC) to investigate the intrinsic dynamics and the ensemble generation capabilities. Additionally, to investigate the effect of the ligands, the same protocol is repeated for the three structures without their respective ligands (in their apo form).

2.3 X-ray structure refinement for ensemble generation

The refinement of crystal structures as structure-ensembles has been recently described [20] and should provide an additional view of protein flexibility taking advantage of short molecular dynamics

supplemented with X-ray data. Ensemble refinement is equally performed on the three liganded ERα complexes 2YJA, 2OUZ and 3UUC.

2.4 Exhaustive virtual screening

As a third approach for probing the receptor's flexibility information from known ligands is used in virtual screenings (VS), while aiming for exhaustiveness. For the local flexibility VS two representative crystallographic target structures (agonist 2YJA and antagonist 3UUC) are selected and in the global flexibility VS all available PDB structures are used within the @TOME server.

The two main research questions are:

- Can we sufficiently cover the conformational space of our target receptor ERα?
- What is needed to perform successful exhaustive ligand docking in terms of flexibility?

3 MATERIALS AND METHODS

3.1 Ensemble analysis of 440 ERα structures

All liganded ERα structures currently available in the PDB (461 protomers) were gathered using the @TOME server by submitting the canonical amino acid sequence of ERα (UniProt identifier: P03372-1) with a specified sequence identity threshold of 90%. 19 protomers (originating from 12 PDB entries) did not contain the C-terminal H12 and were therefore removed from the analysis dataset. Additionally, 2 outlier protomers (chain A and B, originating from PDB-ID: 1A52) were identified, having an ambiguous electron density for H12 (an incorrectly modelled domain swap of H12), a rather low resolution (2.8 Å) and containing gold atoms, and were therefore also excluded for further analysis. The resulting structural dataset contained 440 ERα protomers.

All analysis and image generation was performed using R, R-Studio, in particular the 'bio3d' package, and PyMOL.

3.2 Selected crystallographic structures

For MD simulation, ensemble refinement and local flexibility virtual screening (with PLANTS) representative structures of ERα are selected from the PDB (listed in Table 1). As representative agonist conformation 2YJA is used, containing the natural ligand estradiol (EST). Two representative antagonist are chosen, 2OUZ in complex with Lasofoxifene (C3D) a selective estrogen receptor modulator (SERM) and approved drug that is representative for antagonists in terms of size and shape, and 3UUC in complex with bisphenol C 2 (0D1), an endocrine disruptor that represents the smallest pharmacophore structure as antagonist. Besides criteria such as the agonist/antagonist functionality, a wild-type sequence and the nature of their co-crystallized ligands, they are selected due to their good resolution (1.82 Å 2.1 Å and 2.1 Å respectively). The structures have a Diffraction Precision Index (DPI) [21] of 0.14 Å for 2YJA, 0.18 Å for 2OUZ and 0.27 Å for 3UUC.

Table 1. Selected ERα crystal structures. The calculated Diffraction Precision Index (DPI) [21] is listed for each structure. The respective ligands present in the structures are listed with their Ligand IDs and their activity described in the respective publications.

PDB-ID	Resolution	DPI	Ligand ID	Activity
2YJA	1.82 (Å)	0.14 (Å)	EST	agonist
2OUZ	2.0 (Å)	0.18 (Å)	C3D	antagonist
3UUC	2.1 (Å)	0.27 (Å)	0D1	antagonist

3.3 Molecular dynamics simulation

The crystal structures 2YJA, 2OUZ and 3UUC are downloaded from the RCSB protein data bank (PDB). Structure 2YJA does contain only one ERα monomer (chain B) and no missing residues and can directly be used for MD. Structure 2OUZ also contains only one ERα monomer (chain A), but several side chains are missing. Therefore, the completed and re-refined structure is downloaded from PDB-REDO databank [22] (created with version 7.15). For 3UUC, the most complete protomeric structure (chain D) (with gaps in 4 loop regions) is prepared for MD with an in-house script using Modeller [23] for modelling missing residues with sequences given in the PDB file. Hydrogen atoms of the respective ligands were modeled with OpenBabel at pH 7. All simulations were carried out with Gromacs 2018 [7]. The ligand topologies were generated using the ACPYPE/ANTECHAMBER [24] program of AmberTools17 [25] with partial charges generated by the empirical charge model AM1-BCC. The ligands parameters are based on the General Amber Force Field (GAFF) and the Amber FF14SB force field was employed for the proteins. Each complex was solvated in a TIP3P water dodecahedral box, with periodic boundary conditions and a minimum distance of 1.0 nm from the surface of the complex to the edge of the box. Each system was neutralized by adding NA+ and Cl- ions to physiological concentration of 150 mM. A completely free steepest descent energy minimization for 2000 steps was followed by a 100-ps NVT equilibration and a 100-ps NpT equilibration with Parrinello-Rahman pressure coupling. NVT and NpT equilibrations were performed at a reference temperature of 300 K with ligand restraints of 1000 kJ/mol nm^2 in x,y,z directions. Finally, 50 ns unrestrained production runs were performed with a 2 fs time-step in the NpT ensemble and snapshots were saved every 10 ps. For the simulations without ligands, the ligands are simply removed from the initial structures before starting the MD protocol. Analysis and plotting is performed with Gromacs tools, R and Python scripts.

3.4 Ensemble refinement of X-ray data

The Phenix tool `phenix.ensemble_refinement` models the experimental X-ray data by an ensemble of structures obtained by maximum-likelihood time-averaged restrained MD simulation. Within the calculations, a large amount of sets of coordinates are sampled and the reported number of structures is reduced by selecting the minimal number of structures, equally distributed over the sampling time, that reproduces the R_{free} value of the whole trajectory within a 0.1% tolerance [20]. In order to run the ensemble refinement a structure .pdb file, a structure .cif file and a ligand .cif file are required. The structure files are downloaded from the PDB and the ligand .cif file is generated using the Grade webserver (http://grade.globalphasing.org). The generation of structure-ensembles is streamlined by the in-house script, which runs the ensemble refinement simulation for different sets of parameters, since they affect heavily the running and the outcome and cannot be determined *a priori*. The two main empirical parameters are p_{TLS}, which defines the fraction of atoms included in the TLS fitting, and w_{xray} (or T_{bath}), which controls the X-ray weight by a temperature bath offset (in K). Tested values for p_{TLS} are (0.6, 0.7, 0.8, 0.9 and 1.0) and for w_{xray} (2.5, 5 and 10). A third variable, the relaxation time (or memory time) t_x (in ps) of the time-averaged restraints, is used in the simulation. t_x changes the amount of structures contributing to the target function and is automatically selected based on the dataset resolution. Additionally, values of 2 × automated t_x and 0.5 × automated t_x are also tested, as suggested by the authors. In total, 45 separate simulations are performed for an exhaustive parameter test. The ensemble with the lowest R_{free} is selected for further analysis.

3.5 The BindingDB dataset

Two affinity containing dataset were chosen from the BindingDB 2018: One containing all available Ki affinity measures for ERα comprising 283 molecules and a second one containing all available IC50 affinity measures for ERα comprising 1641 molecules. Since the two affinity measures are not directly comparable, the datasets are kept separately, but compared in terms of structural diversity and spread of binding affinities. The smaller Ki dataset is used in for the manual screening with PLANTS while investigating local side-chain rearrangements and the larger IC50 dataset is additionally used in the automated structure ensemble approach on the @TOME server in order to extrapolate to larger quantities and better statistics.

3.6 Virtual screening

3.6.1 Local flexibility VS on ERα with PLANTS

In order to obtain a first insight into receptor specific, structure dependent difficulties within the ligand screening, involving characteristics such as flexibility of single side-chains but also the movement of whole protein parts, a ligand screening is performed on a two selected ERα structures. The workflow consists of the following steps:

1. Extraction of experimentally tested ligands for the respective receptor from BindingDB. (Extracted ligands are numbered according to decreasing binding affinity.)
2. Selection of different protein structures from the PDB with focus on good resolutions of the crystal structures and the discrimination between agonist, reverse agonist and antagonist conformations.
3. Structure preparation using Spores and Babel.
4. Performing the ligand docking using PLANTS.
5. Visual validation of results using PyMOL.
6. Re-performing step 4 and 5, the ligand docking with PLANTS and subsequent validation, using improved input settings as for example a different protein conformation, different ligand restraints and different flexible protein side chains, until all ligands can be docked in a plausible binding position.

3.6.2 Global flexibility VS on ERα with @TOME server

In order to obtain a broad overview of possible binding modes in addition with more sophisticated information, such as binding affinities, the previously selected set of ligands is screened on respective structure-ensembles using the meta server @TOME instead of only taking single structures. For a large range of NRs @TOME already provides a pre-calculated set of complex supports which are either crystallographic or modeled structures. The envisioned ensemble of structures, here, all crystallographic ligand binding domain (LBD) structures of ERα containing a ligand, is gathered using the @TOME's Modelling module. The settings contain an activated screening module, the 'selected tools' are *Psi-Blast (PDB)* and *HHSearch (PDB)*, with a 'low limit of identity between query and template' of *90%*, and the 'maximum number of additional complexes' set to '*All*'.

One further feature of the @TOME server is the comparative virtual screening module, which is based on the PLANTS software. It performs docking of candidate ligands that are uploaded by the user, taking into consideration the crystallographic ligands being present in the templates and their profile of contacts with the protein. The selection of the model complexes to guide the docking can be done manually according to different criteria. In the automatic mode, which is used here, @TOME selects the structures containing the closest ligands, defined by the Tanimoto similarity score between the uploaded ligand (termed as candidate) and the crystallographic ligand (termed as anchor). Here, the number of template structures to be selected for each ligand is set to 20. So, each candidate ligand is docked into the 20 structures containing the anchors with the highest similarity score, meaning that the selected structures are not necessarily the same for all uploaded ligands and can differ depending on the chemical and structural nature of the candidate.

@TOME furthermore performs a ligand position clustering. This means that for each candidate ligand the newly calculated complexes are superimposed. Based on the superimposition different orientation clusters are detected and ranked. The cluster containing the highest number of detected orientations is termed as P1, which should represent the most probable orientation. Then follows P2, P3, and so on. During and after this processes a range of different parameters are calculated by @TOME. In order to evaluate the calculated complexes different descriptors are taken into account:

- the calculated binding affinity (pKa, also termed AtomeScore), which is an average of the results of the 4 scoring functions PLANTS, MedusaScore, XScore and DSX,
- the similarity to the crystallographic template (PSim), showing the similarity of the ligand-receptor contact profile towards homologous complexes in the PDB (in %), and
- the Tanimoto score between the candidate and the anchor (calculated with Open Babel using FP2 fingerprints)
- the AnchorFit, which is a candidate/anchor superimposition score (provided by PLANTS)
- the Quality of complex, which is a consensus score including features such as internal energy of the ligand, complementarity function between ligand and binding site, quality of receptor structure and type of contacts. It ranges from 0 to 1, while 1 represents the best quality and 0 the lowest quality.

Furthermore, @TOME calculates a Ligand Position Error (LPE) in Ångstrom, which is a theoretical RMSD, using a support vector machine multi-variable regression method (JL. Pons & G. Labesse, to be described elsewhere). The average LPE for each position cluster indicates the most probable orientation of the ligand. The final output of the @TOME server is analyzed using the interactive web interface and contains a high quantity of information (20 instances - docking structures - for each screened ligand and 47 variables).

4 RESULTS AND DISCUSSION

4.1 Global flexibility analysis

4.1.1 Structural variability of 440 crystal structures

In order to get a complete view on ERα's intrinsic flexibility all available crystallographic structures (which were also available for the automated virtual screening) are clustered and analyzed as structural ensemble.

All liganded ERα structures currently available in the PDB (461 protomers) had a sequence identity between 95% and 100% (of them 99% or 100%) with the canonical amino acid sequence of ERα (UniProt identifier: P03372-1) and correspond to point mutants of ERα. This set of structures is also freely available for automated virtual screening campaigns on the @TOME server. After removal of 21 protomers (originating from 13 PDB entries) from the analysis dataset as they are lacking H12 or have ambiguous solutions (details in Methods), the resulting structural dataset contains 440 ERα protomers (originating from 232 PDB entries) with 210 different co-crystallized ligands.

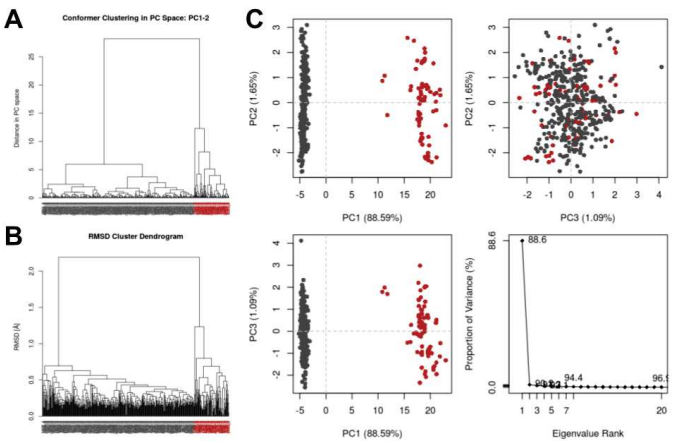


Figure 1. Hierarchical clustering of all 440 protomeric structures based on (A) distance in PC space and (B) RMSD. (C) Principal Component Analysis with plots for PC 1 to 3 and their respective proportion of variance. Coloring is based on RMSD cluster attribution.

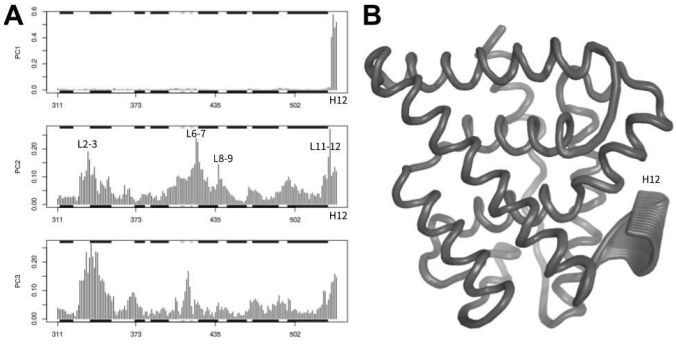


Figure 2. A) Principal Component (PC) residue contribution for the first three PCs and B) PC1 represented as trajectory.

To reveal ERα's global flexibility concerning larger rearrangements such as domain movements, secondary structure repositioning, or loop conformation variability, the following analysis is performed on 440 protomeric structures:

- clustering based on RMSD and Principal Component Analysis (PCA)
- ensemble Normal Mode Analysis (eNMA)
- RMSF analysis
- detailed comparison of agonist and antagonist subsets

Two different clustering methods (hierarchical and k-means) are employed on two different distance measures (RMSD and PC) to identify the main structural conformations of ERα. All four clustering approaches result in the same partitioning of the 440 structures (compare Figure 1 and S3) into two main subsets, a larger one with 358 protomers and smaller one with 82 protomers. The two subsets are identified as ERα active agonist and inactive antagonist conformations, respectively, as shown by the superimposed structures colored by conformer cluster (Figure S4).

Principal Component Analysis (PCA) of the full ensemble of 440 structures demonstrates an outstanding role of the first principal component PC1 (compare Figure 2C). PC1 has a high proportion of variance of 88.59%, compared to the second PC, which reflects only 1.65% of structural variance. Furthermore, when looking at the overall residue contribution to the first three principal components (compare Figure 2A), it is remarkable that PC1 is solely directed by the protein's C-terminus and thus represents the movement of H12, as demonstrated by the PC1 trajectory representation (see Figure 2B).

Ensemble Normal Mode Analysis (eNMA) of the 440 protomeric structures (see Figure 3) grouped by conformer cluster (agonist and antagonist) does not reveal any concerted movement, such as sub-domain or secondary structure rearrangements (except for the C-terminal H12), but instead shows distinct tendencies of loop variability. Agonist conformations show highest fluctuations within

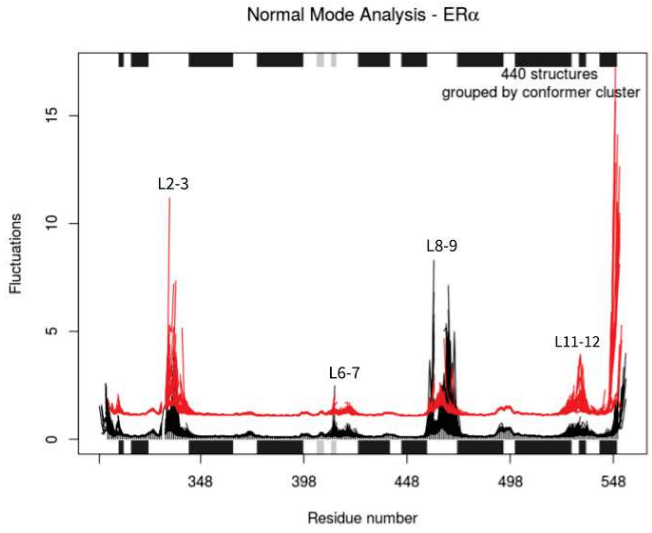


Figure 3. Ensemble Normal Mode Analysis (eNMA) with fluctuations per residue. The 440 protomeric structures are grouped by conformer cluster (agonist - black and antagonist - red) and spread for better comparison.

loop L8-9, followed by a second peak at loop L2-3. Antagonist however, show lower fluctuations for loop L8-9, but even higher fluctuations for loop L2-3. Fluctuation values for loop L11-12 are also increased for antagonists compared to agonists, and most extreme values are attained for the C-terminal H12.

The RMSF analysis of all 440 aligned structures points out the high importance of the C-terminal H12, as depicted in Figure 4. The C-terminus of the protein shows a sharp increase in RMSF with values of up to 20.4 Å for Cα atoms, whereas fluctuations for the rest of the protein's Cα atoms stay below 5.2 Å, with maxima at the N-terminus and loop L11-12 (adjacent to H12) and two further main peaks at loops L2-3 and L8-9.

When comparing RMSFs of the two subsets, agonists (with 358 protomers) and antagonists (with 82 protomers), similar tendencies as revealed by eNMA are apparent (see Figure 5). Agonists and antagonists show RMSF peaks within the same regions (the termini and loop regions L2-3, L6-7, L8-9 and L11-12), whereat the amplitude of fluctuation differs for the two conformer clusters. Agonists

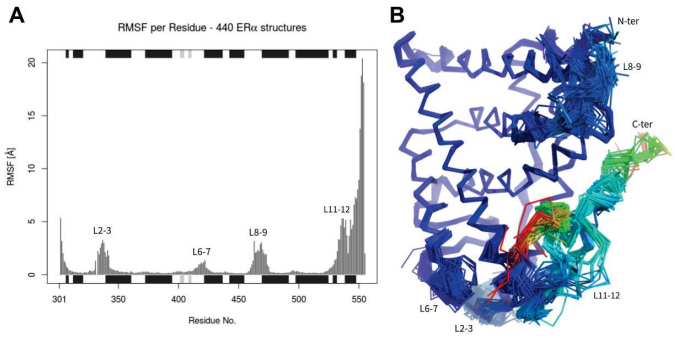


Figure 4. Cα Root Mean Square Fluctuations (RMSF) calculated on all 440 protomeric ERα structures, superimposed on their common core Cα atoms. A) Cα RMSF plotted by residue with annotated secondary structures (α helix in dark grey, β strand in light grey). Gap positions (residues not present in all structures) are excluded from the plot. B) Superimposed structures colored by Cα RMSF including gap positions (coloring scheme = rainbow, with a range of 0 to 20.4 Å).

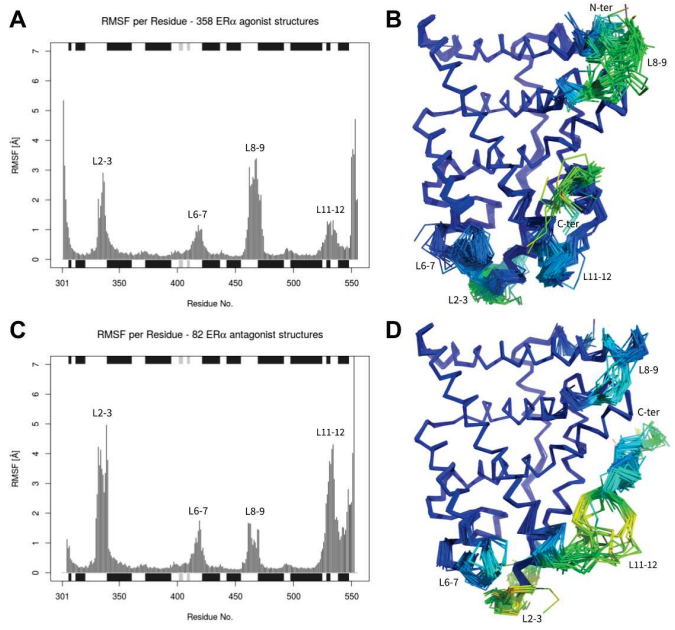


Figure 5. Cα Root Mean Square Fluctuations (RMSF) of common-core superimposed ERα structures calculated on either the agonist subset (A and B) or the antagonist subset (C and D). Gap positions are included. The superimposed structures are colored by Cα RMSF for the agonist subset (B) and the antagonist subset (D) (coloring scheme = rainbow, with a range of 0 to 7.39 Å).

show higher fluctuations for the N-terminus and L8-9. Antagonists have maximal RMSF values for the C-terminus (the extension of H12), helix H12, and the loops L2-3 and L11-12.

4.1.2 Molecular dynamics of ligand-bound complexes

Three systems are investigated, the agonist 2YJA with the natural ligand estradiol (EST) and the two antagonists 2OUZ with Lasofoxifene (C3D) and 3UUC with bisphenol C 2 (OHT), by performing

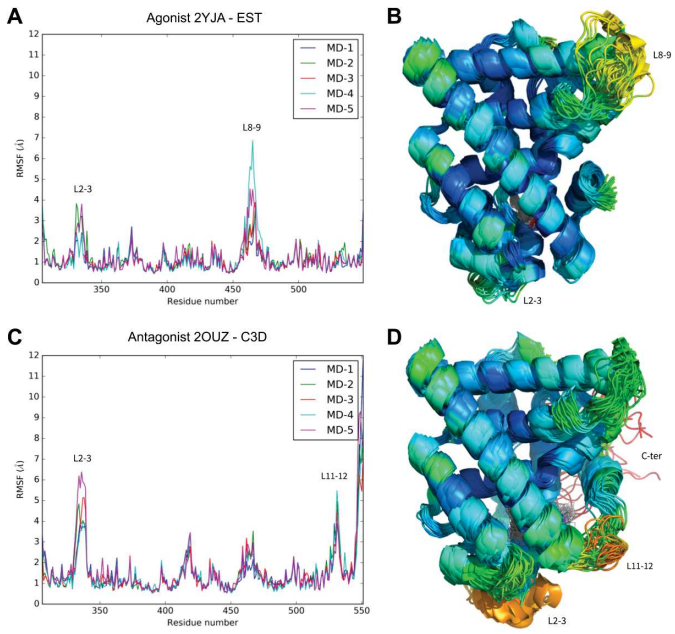


Figure 6. A) & C) RMSF averaged per residue of 5x 50ns MD simulations for agonist conformation 2YJA and antagonist conformation 2OUZ, respectively. B) & D) respective visualization of 55 frames extracted from MD-1 to MD-5 (1 frame every 5ns = 11 frames per MD) colored by RMSF average per residue averaged across all 5 MD simulations (coloring scheme = rainbow, with a range of 0 to 6 Å for B and D).

50ns MD runs, repeated 5 times with different initial velocities, resulting in a total of 250ns simulation time for each of the three systems. To investigate the effect of the ligands, the protocol is repeated for the tree systems without their respective ligands. For the six simulated systems all 5 replicas show a stable and rather rigid behavior over 50ns simulation time, as backbone RMSDs stay low along the trajectories with fluctuations of usually less than 2.5 Å (compare Figure S5). RMSF values averaged per residue, including backbone and side-chain atoms, are showing baseline fluctuations between about 1 - 2 Å (see Figure 6 A and C, and Figure 4). The largest RMSF contributions come from H12 and flexible loop regions (compare Figure 6 B and D). While the mean RMSF per residue averaged over all 5 MD simulations attains a maximum of 3.8 Å for the agonist 2YJA at residue Lys467, located within loop L8-9, the antagonist 2OUZ attains a maximum of 8.6 Å at the C-terminus. If we exclude H12 for 2OUZ, two mean RMSF peaks are visible with maxima of 4.5 Å at Phe337/Ser338, within loop L2-3, and 4.6 Å at Lys531, within loop L11-12. Interestingly, antagonist 3UUC shows rather similar behavior as the agonist 2YJA with a maximum of 5.5 Å at residue Lys467. It has to be pointed out that the ligand in 3UUC is bisphenol C 2, a rather unusual and very small antagonist, an endocrine disruptor and environmental pollutant.

4.1.3 Ensemble refinement of X-ray data

The crystallographic structures 2YJA, 2OUZ and 3UUC were submitted to the ensemble-refinement procedure implemented in the crystallographic refinement package Phenix (for more details see Methods and [20]). For 2YJA the selected ensemble with the lowest

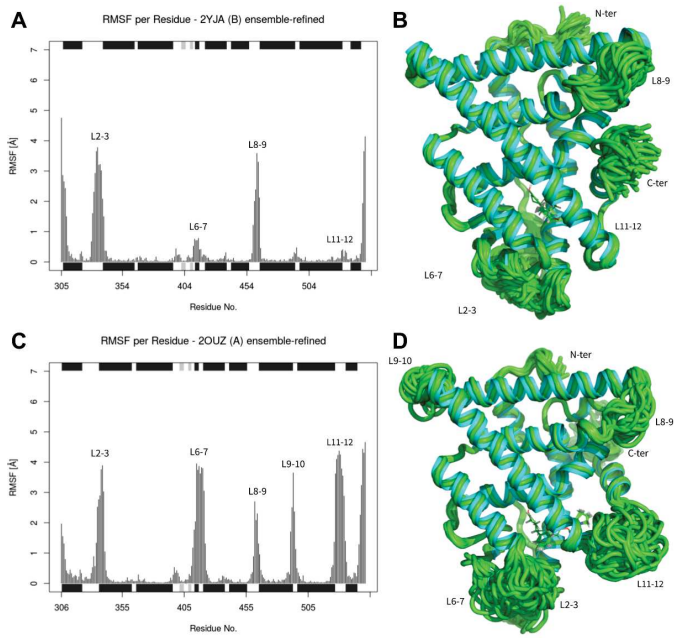


Figure 7. A) & C) RMSF averaged per residue of ensemble-refined agonist conformation 2YJA and antagonist conformation 2OUZ, respectively. B) & D) respective visualization of the single-refined (cyan) and the ensemble-refined (green) structures.

R_{free} (of 0.203) has an R_{work} of 0.142 and the refinement parameters: $p_{\text{TLS}} = 0.9$, $w_{\text{xray}} = 5$ and $t_{\text{x}} = 1.2$. The ensemble refined structure shows an improved agreement between the model and the experimental diffraction data compared to the initial refinement of 2YJA as single model with $R_{\text{work}}/R_{\text{free}} = 0.198/0.234$. For 2OUZ the selected ensemble has an R_{free} of 0.267 and an R_{work} of 0.199 and the refinement parameters are: $p_{\text{TLS}} = 1.0$, $w_{\text{xray}} = 2.5$ and $t_{\text{x}} = 1.0$. Here, the ensemble refined structure does not show any improvement compared to the initial refinement of 2OUZ as single model with $R_{\text{work}}/R_{\text{free}} = 0.199/0.269$. For 3UUC the selected ensemble with the lowest R_{free} (of 0.263) has the following refinement parameters: $p_{\text{TLS}} = 0.9$, $w_{\text{xray}} = 5$ and $t_{\text{x}} = 0.5$. The best refined ensemble with its $R_{\text{work}}/R_{\text{free}} = 0.184/0.263$ lies in the same range as the initial refinement of 3UUC as single model with $R_{\text{work}}/R_{\text{free}} = 0.214/0.255$. This may indicate that the ERα structure 3UUC has a rather restrained flexibility within the crystal and a refinement as single model is sufficient to represent this crystallographic data.

Overall, the ensemble-refinement explores a slightly larger conformational space compared to the standard MD simulations with larger RMSF values and increased variability in loop regions (compare Figure 7, S7, and S8).

4.2 Ligands and conformational preferences

One would expect that a certain ligand favors one over the other conformer and therefore only crystallizes in one protein conformation. Nonetheless, we find three co-crystallized ligands to be present in both, the agonist and the antagonist conformer. The PDB ligand IDs are ”EST”, ”KN1” and ”KN3”. ”EST”, the natural 17-beta-estradiol is a pure agonist, but crystallized in a triple cysteine to serine mutant which adopts an antagonist conformation (PDB-ID: 1QKT) [26]. ”KN1” and ”KN3” are dynamic WAY-derivatives that are partial agonists and cristallyzed in both the canonical active and inactive antagonist conformations (PDB-IDs: 2QZO, 3OS9, 3OSA and 4IW8) [27, 19]. Similarly, we expected the different protomers of a single PDB entry to adopt the same overall conformation and thus being all attributed the same class. Nonetheless, also here we find an exception: The PDB entry 5TLP contains two ERα protomers in complex with two different ligands that induce the distinct conformers agonist and antagonist [28]. Those findings suggest that agonist-antagonist classification of ligands is not a trivial task and that the two ligand binding concepts ’induced-fit’ and ’conformation selection’ are likely inter-knotted in a complex way that also depends on the local environment, such as the presence of specific cofactors.

Finally, the structures that were initially excluded from the analysis dataset (13 PDB entries) are manually attributed to the agonist or antagonist cluster according to the annotation of the bound ligand in the respective publication. This complete classified list of ERα structures enables the analysis of target selection preferences within the automated virtual screening process in @TOME. Here we find that for the xenobiotics and the FDA dataset primarily agonist conformations are selected as support (across 20 dockings per ligand), whereas the BindingDB Ki and IC50 dataset have a rather even selection distribution among agonist and antagonist supports (compare Figure 8). When only taking into account the ’best’ docking per ligand selected by the @TOME server (instead of the mean of all 20 dockings) the mentioned agonist-antagonist support partitioning is even more pronounced. The proportion of agonist vs. antagonist as docking support are for the BDB-IC50 dataset 53.1% vs. 46.9%, for the BDB-Ki dataset 45.2% vs. 54.8%, for the FDA dataset 91.8% vs. 8.2%, and for the Xeno dataset 87.9% vs. 12.1%.

4.2.1 Ligands’ chemical space

Agonists, antagonists, as well as all currently known groups of SERMs, Triphenylethylen SERMs, Benzothiophene SERMs, and Indole and tetrahydronaphthalene SERMs are present in both data sets (Ki and IC50). SERMs are chemically diverse compounds that lack the steroid structure of estrogens and are therefore classified based on the chemical structure of their scaffold.

Appropriate structural diversity of screened ligands

The Tanimoto score is the most common metric in chemoinformatics for comparing molecules and has been shown to be an appropriate choice for fingerprint-based similarity calculations [29]. Thus, it is used here to report structural diversity of the used data set and applied on different molecular fingerprints. For all used fingerprint types (CDK extended, circular, PubChem and MACCS) the Tanimoto score distributions report a diverse dataset as shown in Figure S2. Moreover, when comparing the Ki dataset (containing 281 molecules) to the IC50 dataset (containing 1641 molecules), a very high similarity between the Tanimoto score distributions across all fingerprint types can be stated (compare Figure S1). This indicates a similar composition and similar chemical diversity of the two datasets.

Appropriate spread of binding affinities

Even though the affinity measures Ki and IC50 cannot directly be compared, the histogram of affinities (Figure S2) demonstrate a

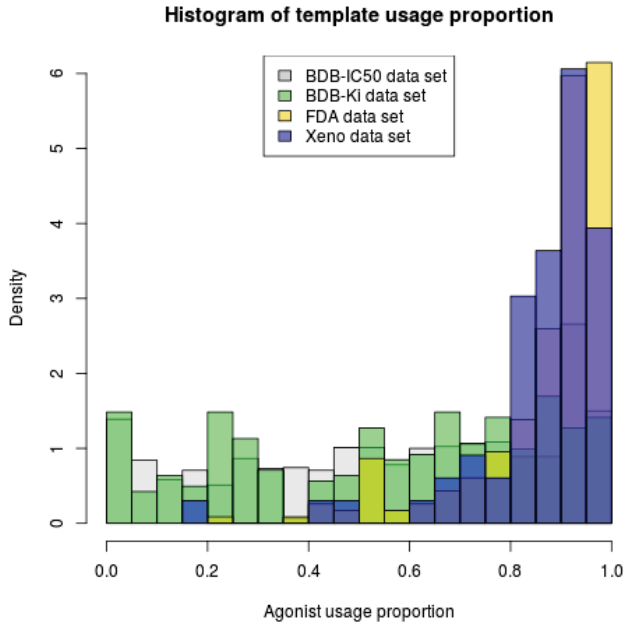


Figure 8. Template class usage profiles for screened ERα ligands across 20 dockings on the @TOME server.

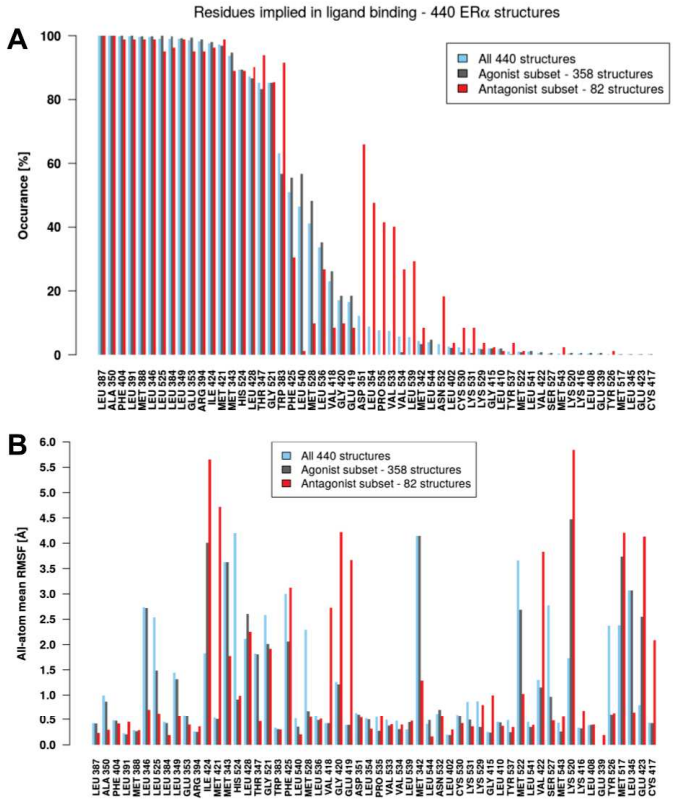


Figure 9. A) Frequency of involvement in ligand binding per residue (ordered from highest to lowest) for all 440 structures and for the conformer cluster subsets agonist and antagonist. B) All-atom mean RMSF per binding site residue, with residue ordering as in A.

wide spread covering more than six orders of magnitude for both datasets, IC50 and Ki.

4.3 Local flexibility within the binding pocket

4.3.1 Structural variability of 440 crystal structures

The ER ligand binding pocket is manly hydrophobic. Nevertheless, two polar residues located at the end of the pocket play a major role in ligand binding. Glutamic acid 353 on helix H3 and arginine 394 on helix H5 contribute to the formation of two hydrogen bonds with the respective ligand (also called tweezers).

A more detailed regard onto the flexibility of the binding pocket reveals further pathways for drug design campaigns. Here we define the binding pocket of ERα by taking into account all residues within a distance of 4 Å of any co-crystallized ligand. This results in a list of 56 residues that potentially contribute to ligand binding.

Particularly, the frequency of involvement in binding for all the identified residues (shown in Figure 9 A) reveals insights into binding requirements and conformer cluster particularities. For example, Leu540 is identified as binding site residue in 56.7% of agonist structures, but only in a single antagonist structure. The antagonist conformations also have their unique fingerprint in the binding pocket. Five residues (Asp351, Leu354, Pro535, Val533 and Leu539) are uniquely identified in antagonists with occurrences of over 30%. Pro535 and Val533 are part of the hinge between H11 and H12 and Leu539 is located within H12, which are for antagonist in a ligand accessible position. On the other hand, Asp351 and Leu354 are part of H3, but with a spacial close location to Pro535 and Leu539 in antagonist conformations. Thus, those residues form a distinct area that is only accessed by ligands crystallized in the antagonist conformation.

Moreover, the flexibility of the 56 identified residues contributing to ligand binding is reflected by their all-atom mean RMSF across all 440 structures and across the conformer subsets (compare Figure 9 B).

The four selected side chains for flexible virtual screening (Met343, Met421, Met528 and His524) are among the frequently identified residues (93.6%, 97.3%, 41.1%, and 89.3% occurrence, respectively) and additionally have increased RMSF values. In particular, Met343 has a rather high RMSF across all structures, which is the same for the agonist subset, but reduced by half for the antagonist subset. This indicates that the variability is dominated by the agonists. In contrast Met421 has a remarkably high RMSF (of 4.7 Å) only across the antagonist subset. Met528 and His524 show a third behavior with high RMSF values for the whole set but low values within both of the subsets. This indicates that the variability is mainly observed when a switch between the two main conformations (agonist and antagonist) occurs.

4.3.2 Molecular dynamics shows rigid ligand-bound state

Concerning the 56 identified binding pocket residues 2OUZ shows only larger RMSF values for some residues (Asn532, Lys531, and Glu339). Otherwise, the binding pocket residues show in general

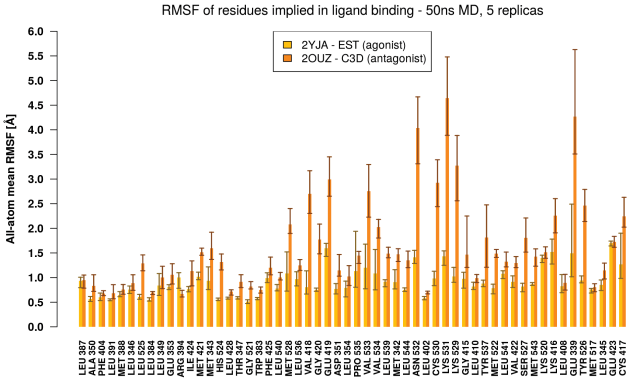


Figure 10. All-atom mean RMSF per binding site residue from 5x 50ns MD simulations, with residue ordering as in Figure 9. The height of the bars is the mean of the 5 replica simulations and the error bars indicate lowest and highest values of the 5 replicas.

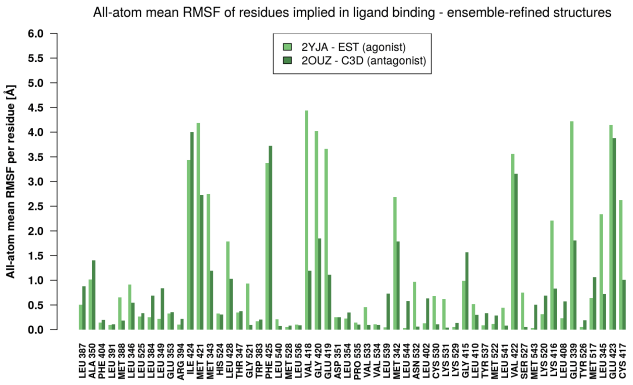


Figure 11. All-atom mean RMSF per binding site residue of the ensemble-refined crystal structures 2YJA and 2OUZ, with residue ordering as in Figure 9.

rather low RMSF values, especially for 2YJA and 3UUC (compare Figure 10 and S10). This indicates a rather rigid ligand-bound conformation that might be restrained by the ligand itself.

4.3.3 Ensemble refinement of X-ray data

Accordingly, the 56 identified binding site residues show increased RMSF values, as depicted for 2YJA and for 2OUZ in Figure 11, and for the four protomers (chain A-D) of 3UUC in Figure S12. RMSF values are very variable among the four protomers of 3UUC (chain A-D), which may be due to chain breaks, as none of the four protomers is complete, and due to lack of crystallographic data for those areas.

4.4 Required side-chain flexibility in the binding pocket

Upon analysis of different crystal structures, it seems to be crucial that the ligand's polar moiety (if present) is placed in such a way that the hydrogen bonds can be established. Also the docking results from PLANTS seem to confirm this observation, since all the binding ligands (extracted from BindingDB) show a very similar positioning of the polar moiety, usually a hydroxyl group, towards the tweezer formed by the two residues involved in hydrogen bonding (Arg and Glu). The hydroxyl group is often bound to a phenyl ring. Therefore, the ligands seem to be oriented by two types of contacts: hydrogen bonding at one (or two) end(s) and hydrophobic van der Waals contacts along the body of the molecule.

In order to probe to which extend local flexibility is required in the binding pocket VS is performed on two selected crystallographic structures in the two main conformations (agonist and antagonist) using the PLANTS software. In this approach an anchor is used in form of a co-crystallized ligand and selected side-chains are additionally set flexible to accommodate more ligands. Flexibility is added in a step-wise manner by choosing different sets of flexible side-chains in an iterative manner for each screening in order to stay at a minimal level of punctually introduced flexibility and to minimize the production of artifacts generated by improbable side-chain orientations. In general, rotamer libraries are often insufficient to sample the conformations of protein side-chains finely enough to yield in natural, probable and collision-free conformations.[30] Alternatively, it is possible that the used side-chain flexibility is required to compensate for minor main-chain motions in the ERα structures upon ligand binding.

The ERα LBDs of the two selected structures 3UUC and 2YJA are shown in Figure 12. The two structures are superimposed underlining the conformational difference of H12 positioning (on the left hand) being the the characteristic difference between agonist and antagonist conformations. The close-up on the binding pocket shows the same positioning for the hydrogen bond forming side-chains Glu353 and Arg394 and the ligand's hydroxyl group for both conformations.

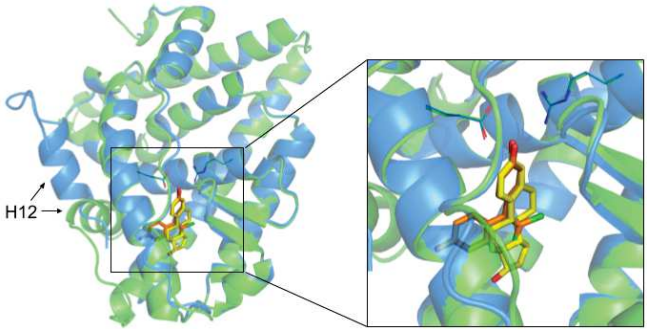


Figure 12. Superimposed ERα LBD domains of 3UUC (blue) in complex with bisphenol C 2 (orange) and 2YJA (green) in complex with estradiol (yellow). The inset shows a close-up on the binding pocket with the two superimposed ligands depicted as sticks and the two hydrogen bond forming side-chains Glu353 and Arg394 in line representation.

Virtual screening results

Small ligands, usually agonists can be docked successfully in a very similar and convincing mode into both selected conformations, the agonist (2YJA) and the antagonist (3UUC) as shown in Figure 13 A. In contrast, larger ligands, having a bulkier extension are usually antagonist ligands and do only fit properly into the binding pocket of the antagonist conformation 3UUC (compare Figure 13 B and

C). Docking results for such ligands into 2YJA either show very unusual and unconvincing binding modes (see Figure 13 B) or they are placed outside the binding pocket (see Figure 13 C).

The docked ligand poses generated using the 3UUC conformation are considered here as more reliable since the ligand poses show common, frequently appearing features. The ligands' polar moieties are positioned similarly, providing hydrogen bonding to Glu353 and Arg394, which is apparently important for good binding affinities, since more than a hundred of the strongest binding ligands (down to 25 nM) all contain a hydrogen bond forming moiety at this position. Moreover, a conjugated ring system, representing a common core structure, is usually oriented in the same plane indicating preferential hydrophobic interactions connected to this orientation. Most extensions from this core structure are characteristic for antagonist ligands, since they are preventing the positioning of helix H12 at the same position to close the binding pocket. Therefore, those extensions are usually pointing out of the binding pocket, in the same direction. A frequently occurring piperidine moiety located at the end of such an outreaching extension (pointing towards the solvent) supports the reliability of the binding pose, since this moiety is usually used to increase the solubility of a ligand.

It has to be mentioned that certain ligands are discarded for the analysis of virtual screening results, because their stereochemistry is not represented properly in three dimensions. This concerns especially carborane containing ligands due to the improper representation of boron atoms and some steroids.

As one might expect, not all the ligands with high affinities can be docked successfully into 3UUC, as exemplified in Figure 13. Therefore, in order to be able to dock more ERα ligands into 3UUC two side-chains in the binding pocket are set as flexible, Met343 and Met421. As result, more convincing binding modes are obtained for most ligands.

Nevertheless, this is not sufficient for one class of large ligands being composed of a steroid core with a large extension connected to the D-ring via a rigid alkyne entity (-C≡C-), which is often followed by a phenyl ring. The strongest binder among them shows an affinity of 25 pM (K_i). The fact that those ligands contain the same steroid moiety, but are often not fitting into the binding pocket suggests that further flexibility is needed to accommodate their extensions. Therefore, in order to dock this class of ligands four side-chains are set flexible, the previously mentioned two methionines (Met343 and Met421) together with two further side-chains, Met528 and His524, which are all located in close proximity to the D-ring. This results in more convincing ligand binding modes, as the binding pocket is extended, enabling the accommodation of the ligands.

The choice for introducing flexibility to certain side-chains follows certain criteria. These are the proximity of the side-chains to the docked ligands, an appearing difference of the positioning of the side-chains between the agonist and the antagonist conformations, and the proximity to the position of ligand extensions when comparing ligands with similar core structures. Moreover, side-chains having an elevated intrinsic flexibility are selected preferentially.

The selection of the flexible side-chains was validated a posteriori by comparing the initial side-chain conformations of 2YJA and 3UUC. The comparison of side-chain positions (see Figure 14) shows larger variations for the conformations of Met343, Met421, Met528 and His524 between structure 2YJA and 3UUC, whereas other binding pocket side-chains are rather equally oriented.

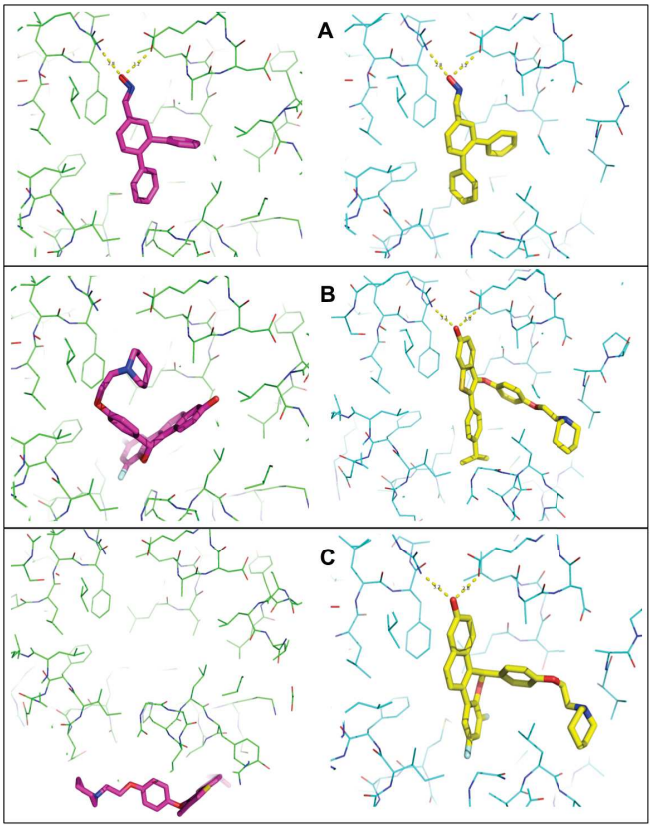


Figure 13. Exemplified docking results for three ERα ligands (A,B,C). The docked ligands are shown in purple (docked into 2YJA) and yellow sticks (docked into 3UUC), the protein structures are shown in green (2YJA) and cyan (3UUC) in line representation. Expected hydrogen bonds to Glu353 and Arg394 are depicted as yellow dotted lines. Ligand A shows convincing poses in both structures; whereat ligands B and C can only be docked successfully into 3UUC.

Virtual screening on multiple crystallographic conformations is performed to account for structural flexibility on a more global level - including possible protein backbone fluctuations. Here, structure-ensembles are used for virtual screening in an automatic manner, a functionality implemented in the @TOME server that is also based on the PLANTS docking software. Assuming that crystal structures show the protein side-chains in favorable conformations, this second approach attempts to sample the available side-chain conformational space by the use of side-chain conformers from multiple X-ray structures. It makes use of all available conformations of ERα (being present in the PDB). Here we find about 70% of poses selected as best pose by the server equivalent to the manual screening with included flexibility for the BindingDB ERα Ki dataset. Most of the remaining ligands have a convincing pose among the remaining 19 @TOME poses that were not selected.

5 CONCLUSION

For ERα large scale movements are dominated by a major rearrangement of helix H12 that distinguishes the two main conformations - agonist and antagonist - that has an important impact on the

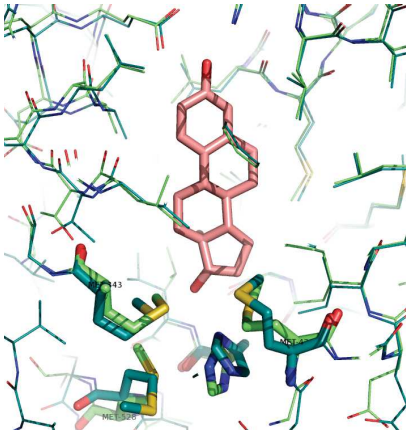


Figure 14. Comparison of side-chain conformations between structure 2YJA (in dark green) and 3UUC (in light green). The natural ligand estradiol is represented (in pink) for binding pocket localization. The selected four flexible side-chains (Met343, Met421, Met528 and His524), are located at the bottom of the binding pocket (highlighted in thicker stick representation).

feasibility of ligand accommodation. The comparison of the various ensembles of conformations (as extracted from the whole crystal structure set, from the refined ensembles, and from the classical MD simulations) highlights a very good qualitative agreement. Indeed, the same regions are found variable and/or flexible (loops L2-3, L6-7, L8-9, and L11-12) with distinct intensity variation patterns for agonist and antagonist conformations. Interestingly, the refined ensembles seem to recapitulate the side-chain flexibility better than classical MD simulations with respect to the variability found in the whole crystal structure set. Accordingly, they may be used for ligand docking and possibly MM-PBSA affinity calculations in the near future. We shall evaluate their use via the @TOME server very soon. Concerning small scale side-chain flexibility within the binding pocket, our exhaustive virtual screening approach revealed the required flexibility of residues Met343, Met421, Met528 and His524, which are equally found to display increased variability within the crystal structure set.

For the first time, to our knowledge, an exhaustive virtual screening campaign is performed to probe a receptor's intrinsic flexibility dictated by its ligands' nature. The results obtained from the virtual screening approach (as view from the ligand side) are confirmed by the results from the protein structure ensemble analysis (as view from the protein side). Those two complementary approaches provide a complete view of the receptor's flexibility. We show that making use of experimental data (ligands with measured affinities from BindingDB and crystallographic structures from the PDB). In this paper, we have described exhaustive docking results on ERα that will serve as reference for detailed VS campaigns and are made available online.

REFERENCES

[1]Jon A. Erickson, Mehran Jalaie, Daniel H. Robertson, Richard A. Lewis, and Michal Vieth. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry*, 47(1):45–55, January 2004. 00293.

[2]Peter Cimermancic, Patrick Weinkam, T. Justin Rettenmaier, Leon Bichmann, Daniel A. Keedy, Rahel A. Woldeyes, Dina Schneidman-Duhovny, Omar N. Demerdash, Julie C. Mitchell, James A. Wells, James S. Fraser, and Andrej Sali. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *Journal of Molecular Biology*, 428(4):709–719, February 2016.

[3]Sam Z. Grinter and Xiaoqin Zou. Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design. *Molecules*, 19(7):10150–10176, July 2014. 00019.

[4]F. Jiang and S. H. Kim. "Soft docking": matching of molecular surface cubes. *Journal of Molecular Biology*, 219(1):79–102, May 1991. 00391.

[5]Oliver Korb, Thomas Sttzle, and Thomas E. Exner. Empirical Scoring Functions for Advanced ProteinLigand Docking with PLANTS. *Journal of Chemical Information and Modeling*, 49(1):84–96, January 2009. 00321.

[6]David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, December 2005. 04085.

[7]Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, March 2008. 07109.

[8]A. Lavecchia and C. Di Giovanni. Virtual screening strategies in drug discovery: a critical review. *Current Medicinal Chemistry*, 20(23):2839–2860, 2013. 00063.

[9]Olivier Sperandio, Liliane Mouawad, Eulalie Pinto, Bruno O. Villoutreix, David Perahia, and Maria A. Miteva. How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *European biophysics journal: EBJ*, 39(9):1365–1372, August 2010. 00055.

[10]B. Lawrence Riggs and Lynn C. Hartmann. Selective Estrogen-Receptor Modulators Mechanisms of Action and Application to Clinical Practice. *New England Journal of Medicine*, 348(7):618–629, February 2003.

[11]D. J. Mangelsdorf, C. Thummel, M. Beato, P. Herrlich, G. Schtz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon, and R. M. Evans. The nuclear receptor superfamily: the second decade. *Cell*, 83(6):835–839, December 1995. 06444.

[12]Vanessa Delfosse, Marina Grimaldi, Jean-Luc Pons, Abdelhay Boulahtouf, Albane le Maire, Vincent Cavailles, Gilles Labesse, William Bourguet, and Patrick Balaguer. Structural and mechanistic insights into bisphenols action provide guidelines for risk assessment and discovery of bisphenol A substitutes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37):14930–14935, September 2012. 00075.

[13]M. Berry, D. Metzger, and P. Chambon. Role of the two activating domains of the oestrogen receptor in the cell-type and promoter-context dependent agonistic activity of the anti-oestrogen 4-hydroxytamoxifen. *The EMBO journal*, 9(9):2811–2818, September 1990. 00776.

[14]Qin Feng and Bert W. OMalley. Nuclear Receptor Modulation - Role of Coregulators in Selective Estrogen Receptor Modulator (SERM) Actions. *Steroids*, 90:39–43, November 2014.

[15]Hitisha K. Patel and Teeru Bihani. Selective estrogen receptor modulators (SERMs) and selective estrogen receptor degraders (SERDs) in cancer treatment. *Pharmacology & Therapeutics*, December 2017.

[16]Joan S. Lewis and V. Craig Jordan. Selective estrogen receptor modulators (SERMs): Mechanisms of anticarcinogenesis and drug resistance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 591(1):247–263, December 2005.

[17]Philipp Y Maximov, Theresa M Lee, and V. Craig Jordan. The Discovery and Development of Selective Estrogen Receptor Modulators (SERMs) for Clinical Practice. *Current Clinical Pharmacology*, 8(2):135–155, May 2013.

[18]Thomas P. Burris, Laura A. Solt, Yongjun Wang, Christine Crumbley, Subhashis Banerjee, Kristine Griffett, Thomas Lundasen, Travis Hughes, and Douglas J. Kojetin. Nuclear Receptors and Their Selective Pharmacologic Modulators. *Pharmacological Reviews*, 65(2):710–778, April 2013.

[19]Sathish Srinivasan, Jerome C. Nwachukwu, Alex A. Parent, Valerie Cavett, Jason Nowak, Travis S. Hughes, Douglas J. Kojetin, John A. Katzenellenbogen, and Kendall W. Nettles. Ligand-binding dynamics rewire cellular signaling via estrogen receptor-. *Nature Chemical Biology*, 9(5):326–332, May 2013.

[20]B. Tom Burnley, Pavel V. Afonine, Paul D. Adams, and Piet Gros. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife*, 1:e00311, December 2012. 00092.

[21]Dinesh Kumar Kala Sekar, Gurusaran Manickam, S.N. Satheesh, P Radha, S Pavithra, K P. S. Thulaa Tharshan, John Helliwell, and K Sekar. Online-DPI:

A web server to calculate the diffraction precision index for a protein structure. *Journal of Applied Crystallography*, 48:939–942, June 2015.

[22]Robbie P. Joosten, Fei Long, Garib N. Murshudov, and Anastassis Perrakis. The PDB_redo server for macromolecular structure model optimization. *IUCrJ*, 1(Pt 4):213–220, May 2014. 00148.

[23]A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, December 1993. 09586.

[24]Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling*, 25(2):247–260, October 2006. 01693.

[25]D. A. Case, D. S. Cerutti, T. E. Cheatham, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, D. M. York, and P. A. Kollman. Amber 2017, University of California, San Francisco, 2017. 00000.

[26]M. Gangloff, M. Ruff, S. Eiler, S. Duclaud, J. M. Wurtz, and D. Moras. Crystal structure of a mutant hERalpha ligand-binding domain reveals key structural features for the mechanism of partial agonism. *The Journal of Biological Chemistry*, 276(18):15059–15065, May 2001.

[27]John B. Bruning, Alexander A. Parent, German Gil, Min Zhao, Jason Nowak, Margaret C. Pace, Carolyn L. Smith, Pavel V. Afonine, Paul D. Adams, John A. Katzenellenbogen, and Kendall W. Nettles. Coupling of receptor conformation and ligand orientation determine graded activity. *Nature Chemical Biology*, 6(11):837–843, November 2010.

[28]Jerome C. Nwachukwu, Sathish Srinivasan, Nelson E. Bruno, Jason Nowak, Nicholas J. Wright, Filippo Minutolo, Erumbi S. Rangarajan, Tina Izard, Xin-Qui Yao, Barry J. Grant, Douglas J. Kojetin, Olivier Elemento, John A. Katzenellenbogen, and Kendall W. Nettles. Systems Structural Biology Analysis of Ligand Effects on ER Predicts Cellular Response to Environmental Estrogens and Anti-hormone Therapies. *Cell Chemical Biology*, 24(1):35–45, January 2017.

[29]Dvid Bajusz, Anita Rcz, and Kroly Hberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7:20, 2015. 00000.

[30]Qiang Wang, Adrian A. Canutescu, and Roland L. Dunbrack. SCWRL and MolIDE: Computer programs for side-chain conformation prediction and homology modeling. *Nature protocols*, 3(12):1832–1847, 2008. 00082.

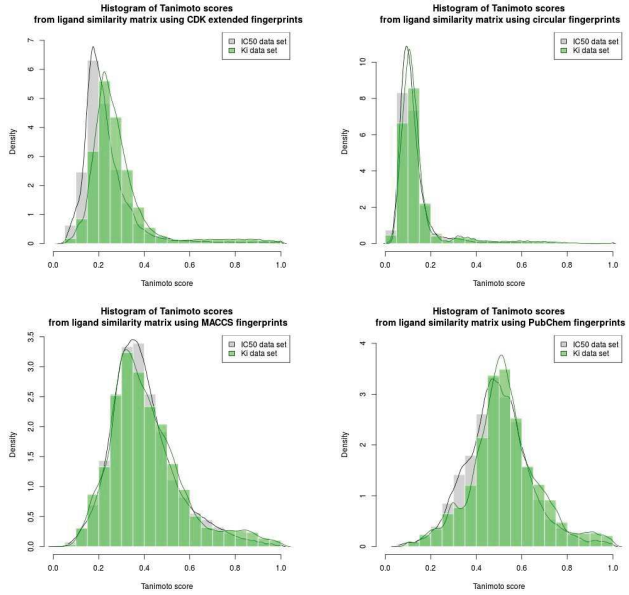


Figure S1. Tanimoto score distributions calculated on four different fingerprints of the Ki dataset (281 molecules) and the IC50 dataset (1641 molecules).

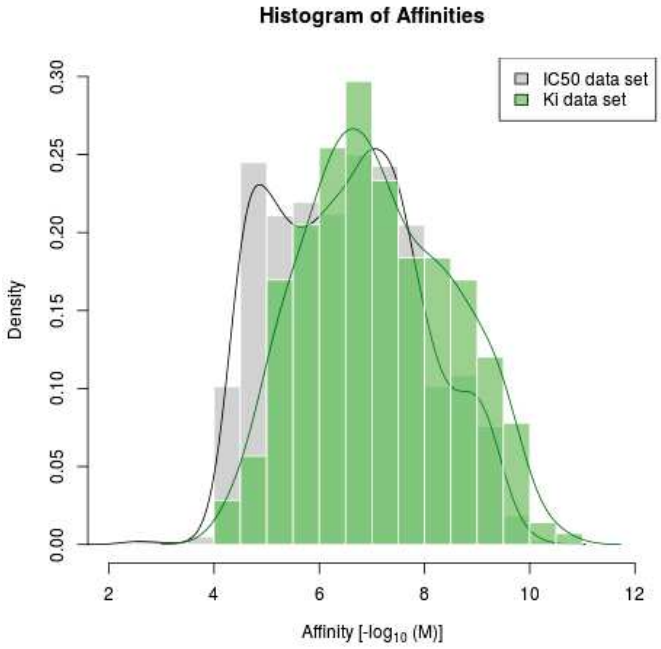


Figure S2. Ligand affinity spread of both datasets, IC50 and Ki

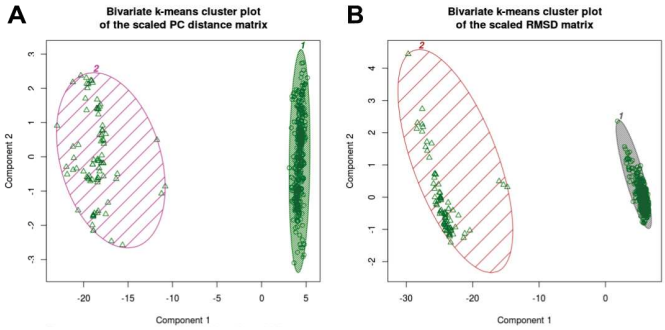


Figure S3. k-means clustering of all 440 protomeric structures based on (A) distance in PC space and (B) RMSD.

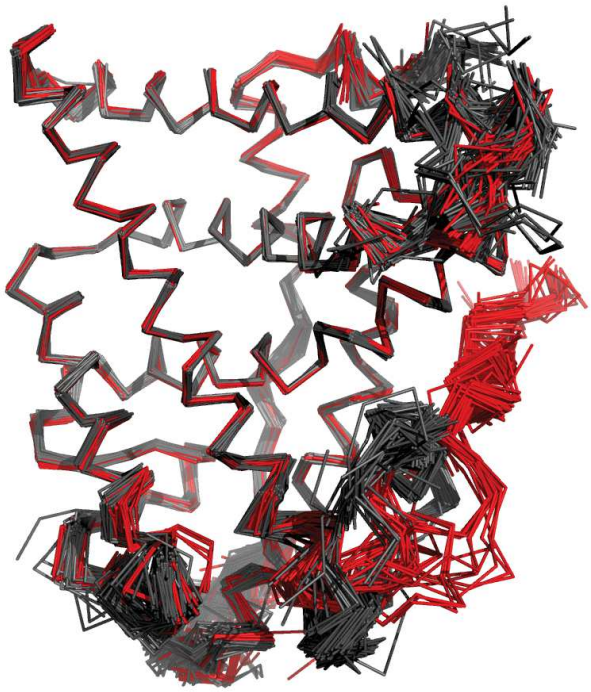


Figure S4. 440 superimposed ERα structures colored by conformer cluster - 358 agonist in grey and 82 antagonist in red.

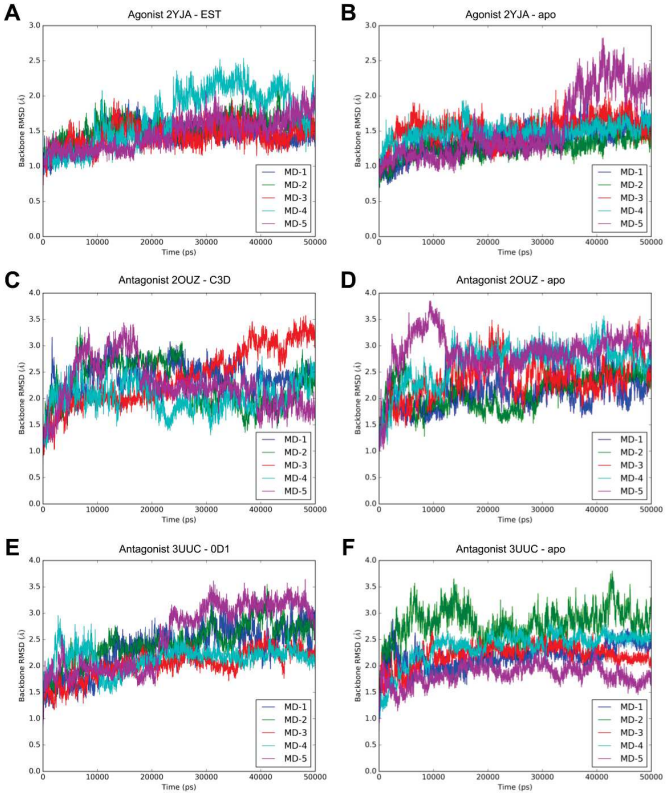


Figure S5. Backbone RMSD over time of 5x 50ns MD simulations for 2YJA (A), 2OUZ (C), and 3UUC (E), and of simulations without the respective ligands (as apo structures), (B), (D), and (F) respectively.

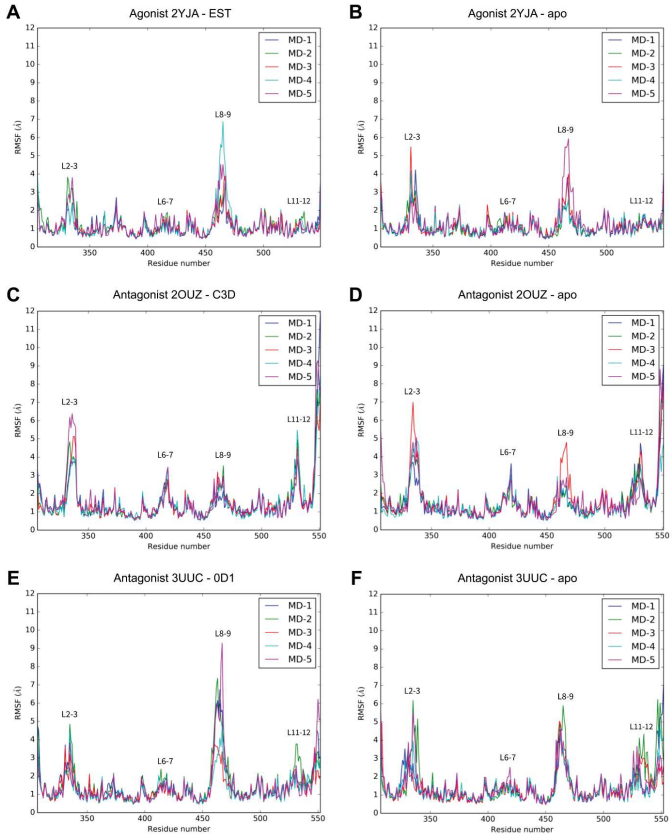


Figure S6. RMSF averaged per residue of 5x 50ns MD simulations for agonist conformation 2YJA (A) and antagonist conformations 2OUZ (C) and 3UUC (E), and of simulations without the respective ligands (as apo structures), (B), (D), and (F) respectively.

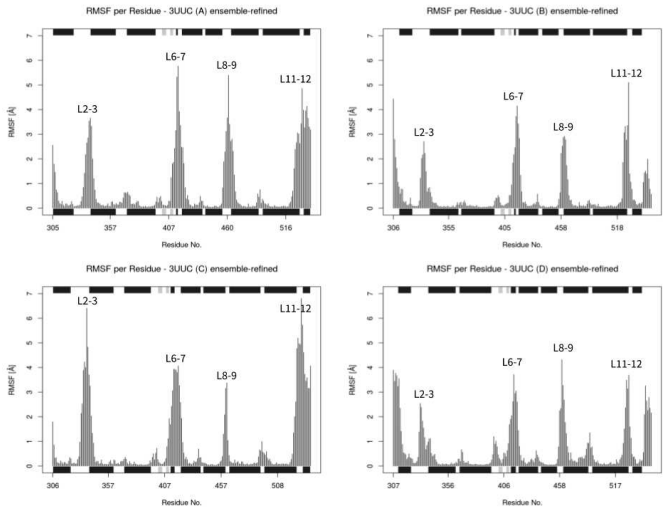


Figure S7. RMSF per residue for the four protomers (chain A-D) of the ensemble-refined crystal structure 3UUC.

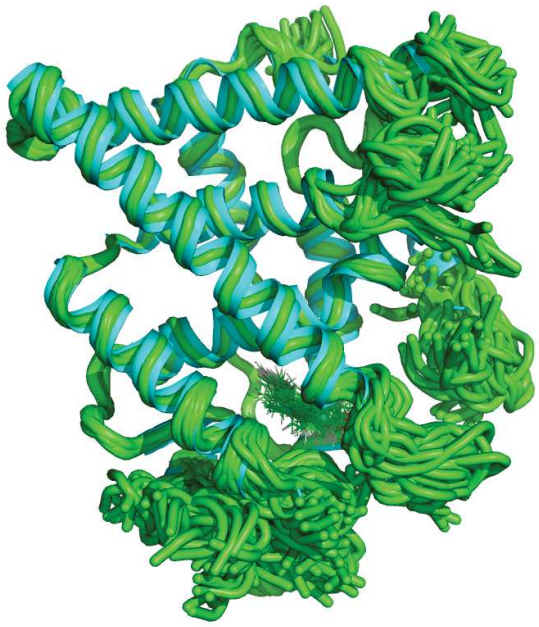


Figure S8. Protomer A of crystal structure 3UUC refined as single structure (cyan) and as ensemble (green).

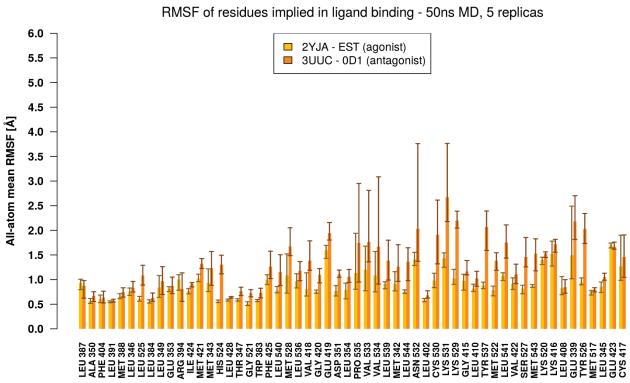


Figure S10. All-atom mean RMSF per binding site residue from 5x 50ns MD simulations, with residue ordering as in Figure 9. The height of the bars is the mean of the 5 replica simulations and the error bars indicate lowest and highest values of the 5 replicas.

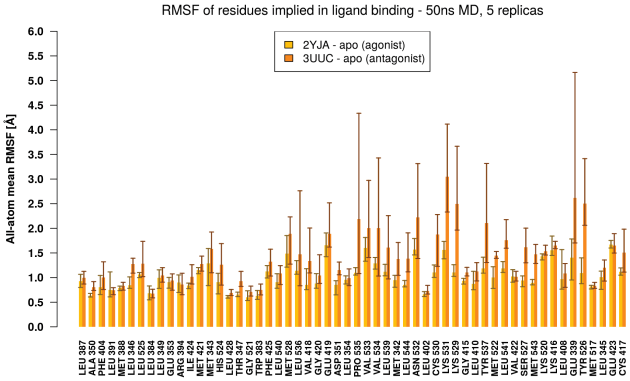


Figure S11. All-atom mean RMSF per binding site residue from 5x 50ns MD simulations, with residue ordering as in Figure 9. The height of the bars is the mean of the 5 replica simulations and the error bars indicate lowest and highest values of the 5 replicas.

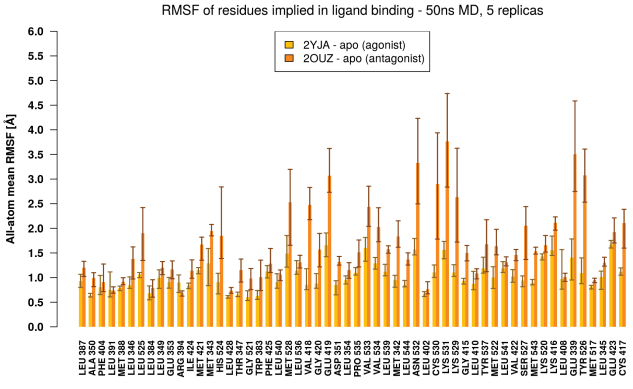


Figure S9. All-atom mean RMSF per binding site residue from 5x 50ns MD simulations, with residue ordering as in Figure 9. The height of the bars is the mean of the 5 replica simulations and the error bars indicate lowest and highest values of the 5 replicas.

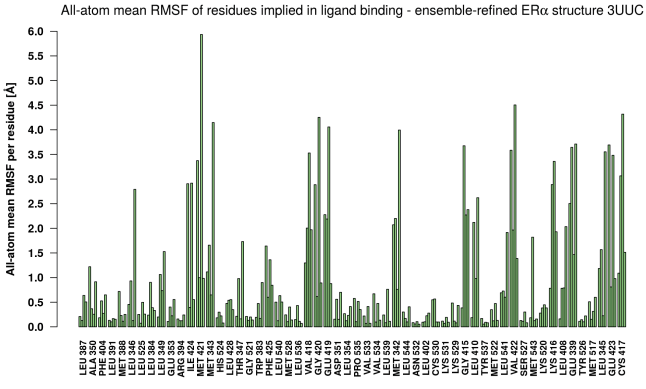


Figure S12. All-atom mean RMSF per binding site residue from the four protomers (chain A-D) of the ensemble-refined crystal structure 3UUC, with residue ordering as in Figure 9. The four protomer RMSF values are grouped per residue (four bars per residue for chain A-D).

3

THE DRUG DESIGN PROJECT

This chapter contains a detailed description of the diverse (sometimes iterated) steps within the drug design project that is aimed at finding a potent inhibitor of the oncogenic protein kinase BRAF-V600E with minimal affinity for the nuclear receptor PXR.

3.1 Motivation

The aim of this project (as depicted in Figure 3.1) is to set up an integrated approach for drug refinement. The biological system under investigation is focused on the protein kinase inhibitor (PKI) dabrafenib and possible derivatives taking into account both the primary target serine/threonine kinase BRAF, but also the unwanted and harmful secondary target PXR, in order to avoid the activation of the degradation pathway via CYP450 enzymes.

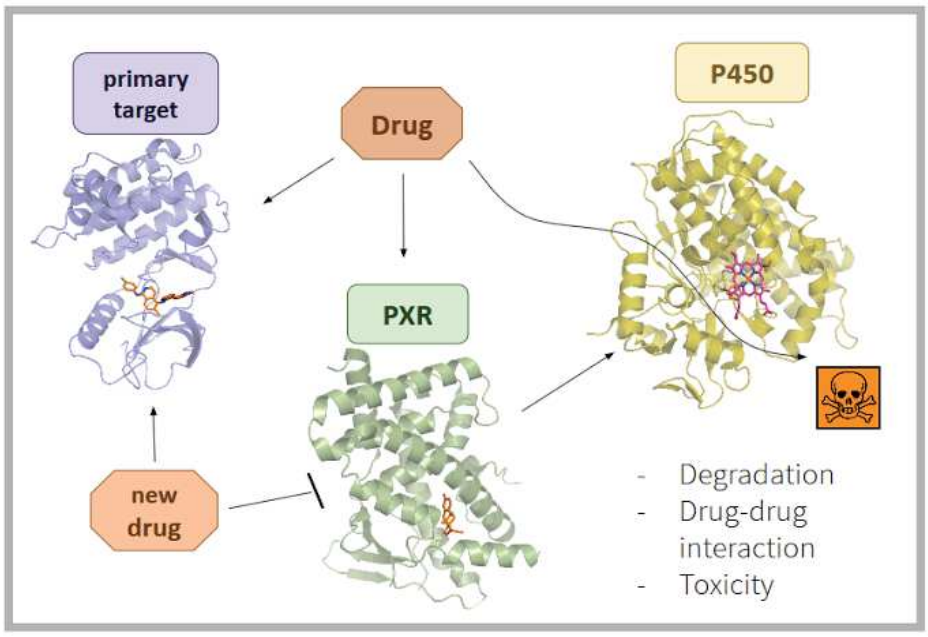


Figure 3.1: Motivation: The anti-cancer drug does not only bind to its primary target BRAF, but also to the nuclear receptor PXR, which subsequently induces the transcription of CYP450 resulting in drug degradation and adverse effects. The new drug should still bind its primary target BRAF, but not any more PXR.

3.2 Protein structure flexibility - a global view on BRAF & PXR

Proteins are not strictly static objects, but populate ensembles of conformations. Which of the states are happening to be the most populated and therefore most occurring depends on their overall energy in equilibrium and on the transition energies between the conformations.

The stability of a certain conformation and the protein's overall flexibility are closely related. Stability of a conformation of a globular protein depends on the interplay of three factors: the unfavorable conformational entropy change, which favors random chains and therefore a rather unfolded state, the favorable enthalpy contribution arising from intramolecular side group interactions, and the favorable entropy change arising from the shielding of hydrophobic groups from the aqueous solvent through burying within the molecule.

Transitions between different conformational states are often functionally relevant and can occur on a variety of time scales (ns to s) and length scales (Å to nm). Furthermore, stability-activity trade-offs can exist, as proteins have not evolved to maximize stability, but rather to preserve adequate stability, and to exert an 'optimal' activity. However, sometimes stability and activity are selected at the expense of the other, such that some mutations in the active site of a protein are more stable but less active (e.g. in the case of a binding pocket charge compensation by mutation, whereas the compensation is usually performed by an oppositely charged ligand), or a mutation leads to a reduced stability of the inactive conformation, provoking in turn to constitutive activity (as for BRAFV600E). This highlights the importance of the intrinsic flexibility and its balanced fine-tuning for the protein's functionality. Within the study of protein dynamics two levels of flexibility are often distinguished:

1. Local flexibility, concerning the movement of atoms and residues (e.g. side-chains switching between separate discrete rotamers and energy minima), usually happening in the picosecond to nanosecond scale;
2. Global flexibility, concerning secondary structure re-locations or domain movements, usually happening in the microsecond to second scale.

However, the transition between the two levels can be continuous. For example, the movement of single residues can be coupled, transferring to a larger scale. The coupling of residue movement may have an effect across the whole protein (e.g. when coupled residues are forming pathways and linking functionally important parts of a protein). By such means, the coupling may also participate or even be the driving force in allosteric signaling.

Different proteins show varying degrees of conformational flexibility on the different levels.

3.2.1 Analysis of X-ray structures

Analyzing the sequence, structure and conformational heterogeneity of proteins is one of the first steps for investigating the protein's potential for rational drug design. The examination of

diverse protein conformers can help to identify important structural and dynamic features that may affect ligand binding and the quantitative comparison of the different conformers may lead to new predictions for structural dynamics. Therefore, the information deduced from interconformer relationships is a valuable factor to take into account for drug design approaches.

3.2.1.1 PXR - global analysis

In the case of PXR we have a completely different starting point compared to the previously studied nuclear receptor ERα. For ERα there are many structures available in the PDB (248) with overall good resolutions (more than 90% $<$ 2.7 Å) and the two dominant conformations, the agonist and the antagonist form are represented and well described with their respective ligands. For PXR there are only a few crystallographic structures available (23 PDB entries in total with 34 protomeric PXR structures) and most of them have a rather low resolution ($\geqslant$ 2.7 Å for 9 PDB entries). Moreover, unlike most NRs that tend to be specialized for a set of ligands with structural homologies, PXR is able to bind a large number of structurally diverse ligands. Its binding pocket is very large, mostly hydrophobic with 8 evenly distributed polar residues and possesses a high deformability and adaptability to accommodate a large variety of chemical scaffolds.

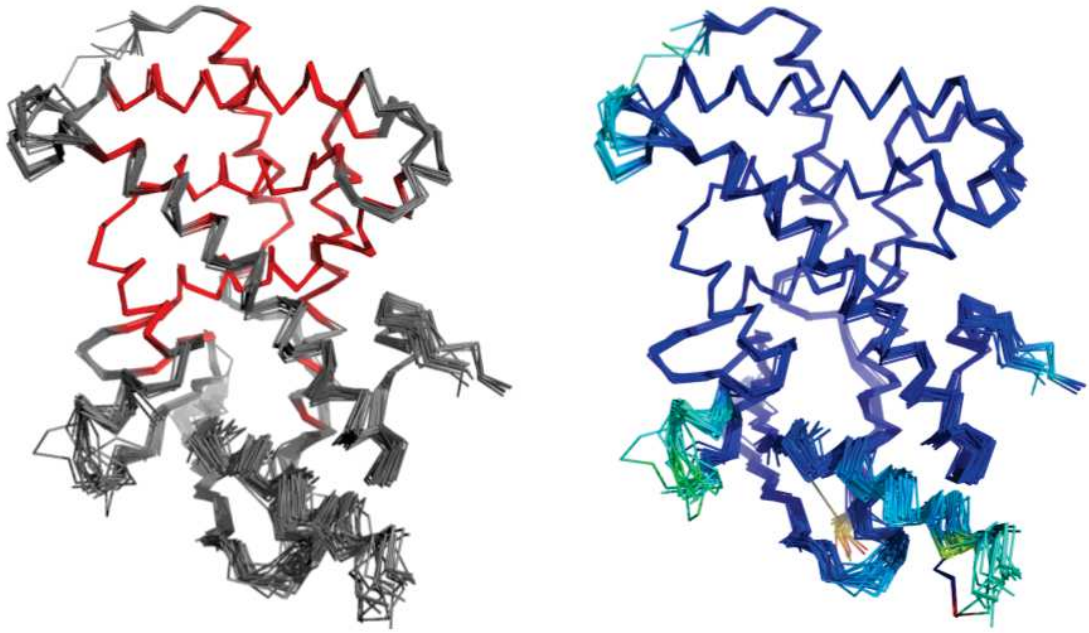


Figure 3.2: Rigid core superpositioning result (left) of 34 PXR protomers with 'rigid core' residues highlighted in red (calculated with a cumulative volume cutoff at 0.5 Å^3); and subsequent RMSF calculation across the protomers with rainbow color coding (from 0Å in blue to a maximum of 9.95Å in red).

In the present structure ensemble analysis all PXR chains/protomeric structures that are present in the PDB entries are taken into account. Therefore, the number of analyzed structures increase from 23 to 34. The protomers are superimposed based on a rigid core calculated using the function *core.find* from the R package *bio3d*. The function *core.find* performs iterated rounds of structural superposition to identify the most invariant region in an aligned set of protein structures. It refines

an initial structural superposition determined from a multiple alignment. At each round of iteration the position(s) displaying the largest differences is(are) excluded from the defined "rigid core", until one of the stopping criteria, either a minimal core size (of 15 residues), or a minimal cumulative volume (e.g. $\leqslant 0.5$ Å^3) is reached. The superimposed structures are subsequently investigated by RMSD/RMSF analysis, Principal Component Analysis (PCA), clustering, and ensemble Normal Mode Analysis (NMA).

The identified rigid core of the PXR ensemble (see Figure 3.2) comprises 107 residues when calculated with a cumulative volume cutoff at 0.5 Å^3 (and 141 residue with a cutoff at 1 Å^3). This represents a large portion of the protein structure with its 231 non-gap residues (excluding all gap residues not resolved in a structure) and a total sequence of 289 analyzed residues.

Based on the structural alignment Root Mean Square Fluctuations (RMSFs) across all 34 PXR protomers are calculated and visualized as color code on the aligned structures in Figure 3.2 and as diagram along the sequence in Figure 3.3. The RMSF analysis of the PXR ensemble pictures a rigid receptor. Only one loop (L2-3), which is not completely resolved in any protomeric structure (residues 179-185 are missing in all the 34 crystallographic protomers), shows a particularly increased variability with large RMSF values (up to 9.95 Å) for adjacent residues (compare Figure 3.2 and 3.3). Along the rest of the sequence RMSF peaks with values of maximal 3.9 Å are attained at the N-terminus, at the beginning of H3 (just after the missing loop), at loop L6-7 (after β5) and at loop L9-10. Large parts of the whole sequence (except the previously mentioned parts) stay even below an RMSF value of 1 Å.

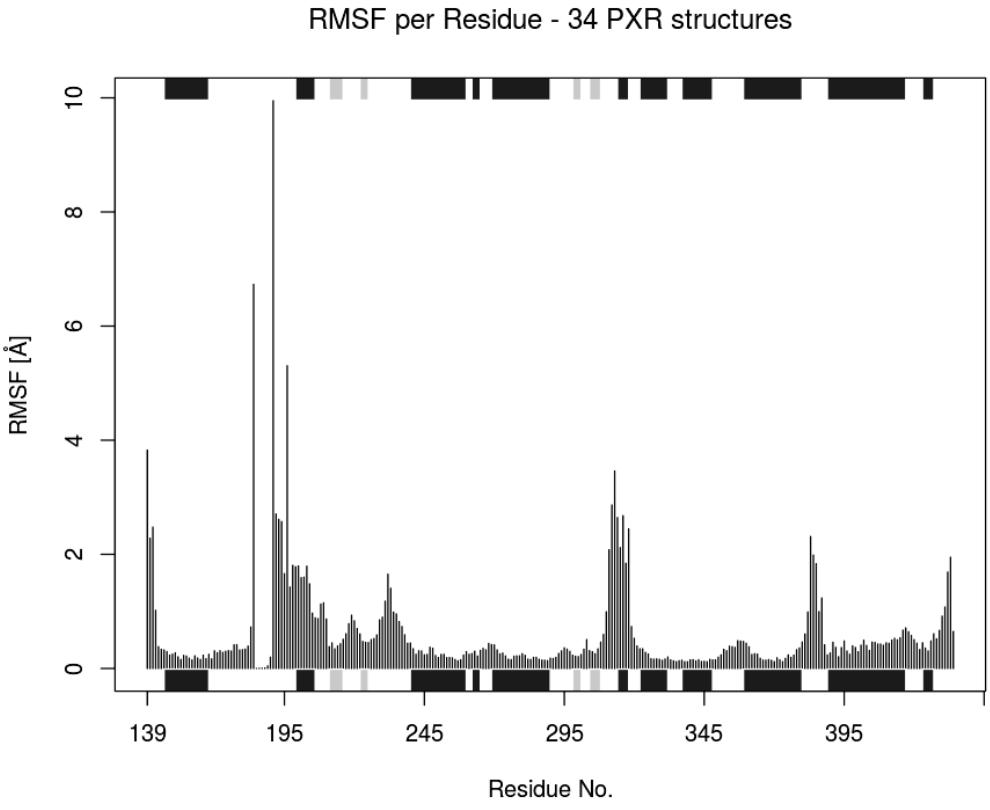


Figure 3.3: RMSF calculation across the 34 PXR protomers plotted along the sequence with secondary structure annotation from a reference structure (PDB-ID: 4S0T chain B).

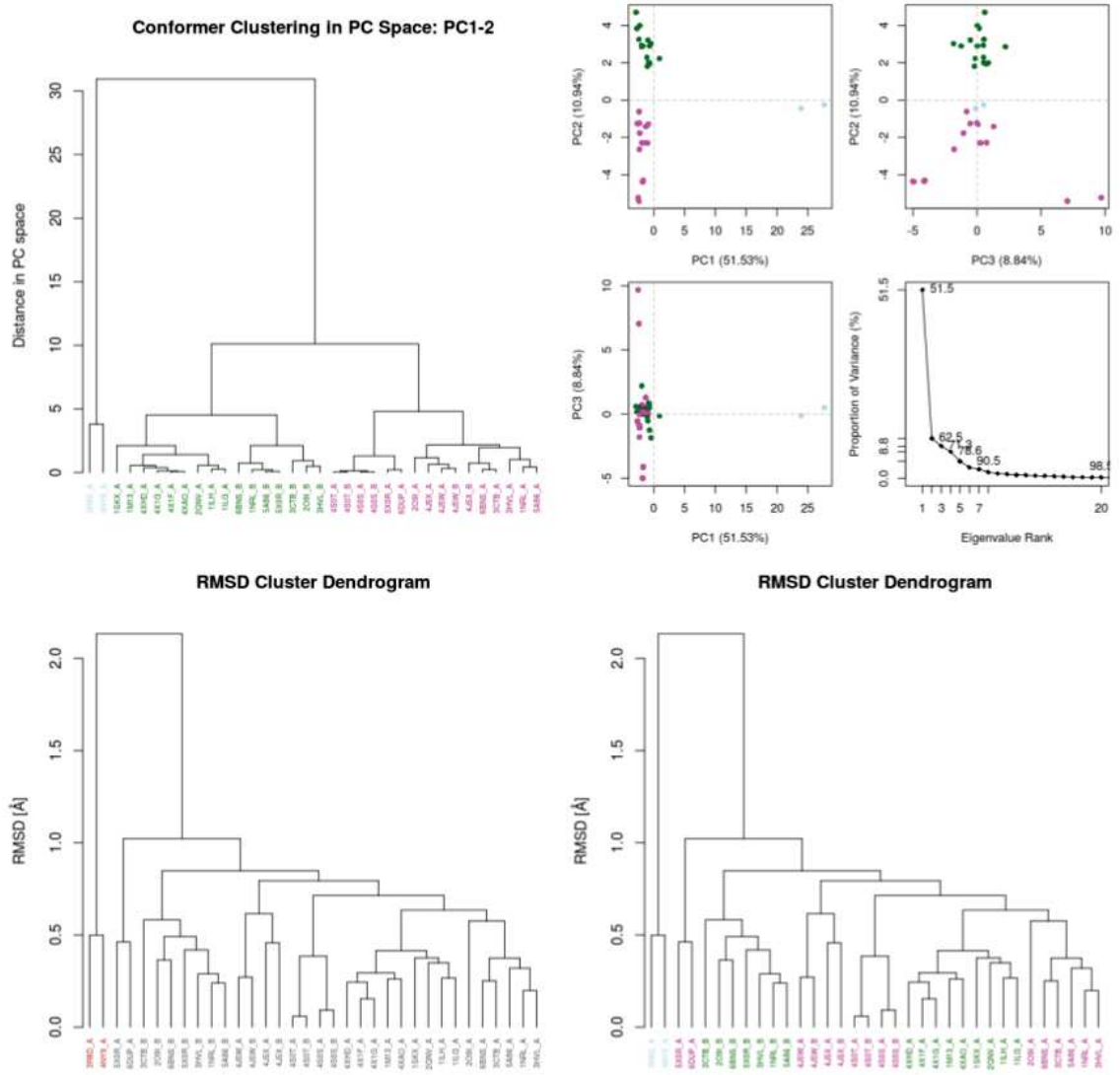


Figure 3.4: Principal Component Analysis (PCA) of the PXR conformational ensemble (top row) with clustering in PC space (PC1-2), 2D plots of the first three PCs and the proportion of the variance covered by the PCs. RMSD conformer clustering (bottom row) with cluster overlap shown by RMSD dendrogram colored by PC cluster.

Visualized on the atomic structure, the PCs do not show significant amplitudes concerning conformational changes, as PC1 is an artifact resulting from modelling issues at a chain break. PC2 can be visualized principally governed by a movement of the external loop L9-10 (see Figure 3.5) and minor contributions of other areas, which due to their small amplitudes can rather be connected to vibrational differences of the structure. Unfortunately, we cannot obtain a complete view of the accessible conformational space for PXR, which is indicated by comparing the PCA plots for PXR, where we see separated sparse dots and for ERα, where whole connected clouds are visible.

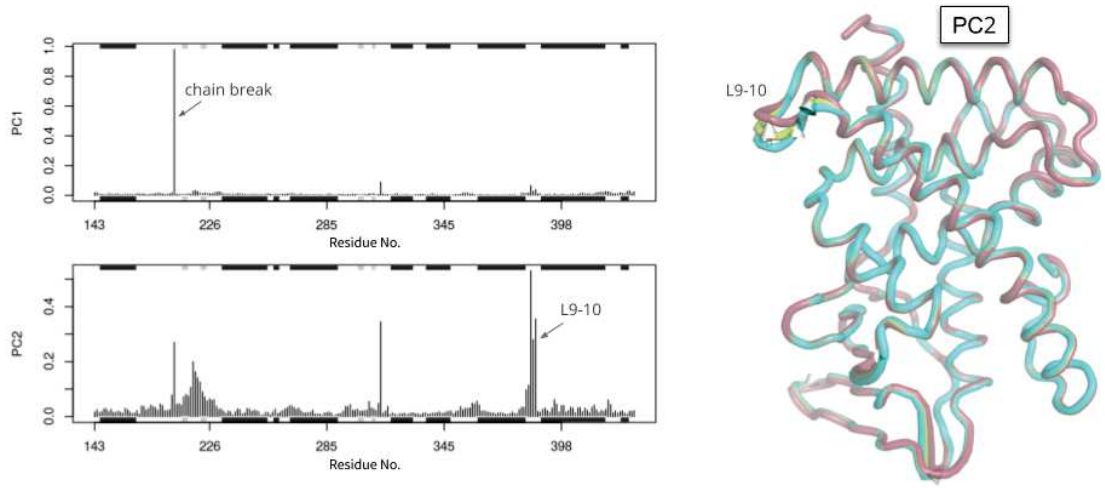


Figure 3.5: Principal Component Analysis of the PXR conformational ensemble: per residue contributions to the first two PCs (left), and structural representations of the second PC with vectors (grey arrows) calculated using the *modevectors* module in PyMol (right).

3.2.1.2 PXR - binding pocket analysis

Also for PXR a binding pocket analysis is performed (with all parameter settings as for the analysis of BRAF and ERα). Among the 23 PXR containing PDB entries six (1ILG, 1M13, 3CTB, 4J5W, 4S0S, 4XAO) are not complexed with a ligand, and can therefore not be used for binding pocket analysis, reducing the set of protomeric complexes to 25.

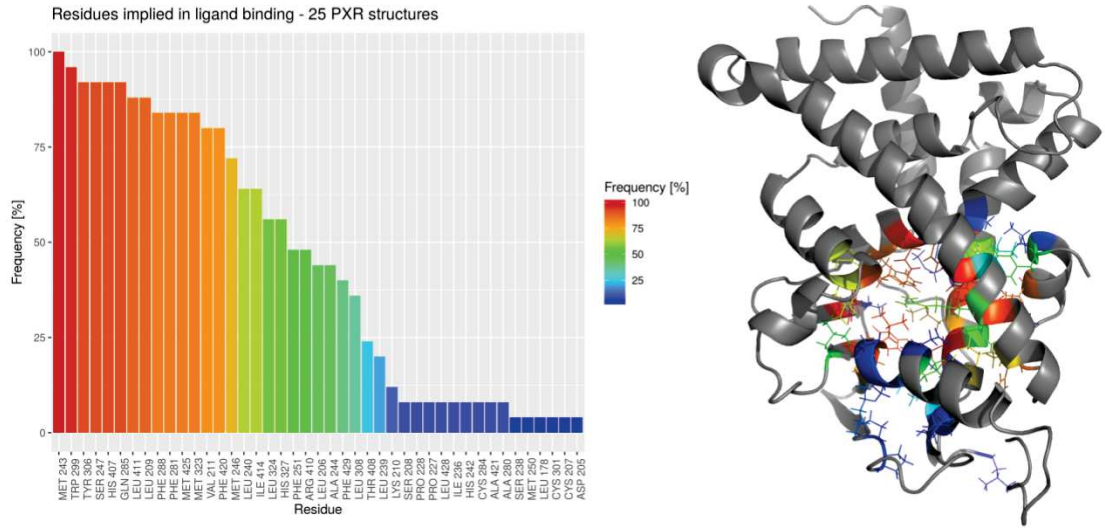


Figure 3.6: Residues implied in ligand binding of 25 liganded PXR protomers (left), colored by the frequency they are identified being within a radius of 4 Å around the ligand. A representative protein structure (PDB-ID: 4S0T, chain B) (right) is used to visualize the location of the identified residues within the structure. The side-chains of the 43 identified residues are shown in line representation on the protein structure and the coloring is also based on identification frequency.

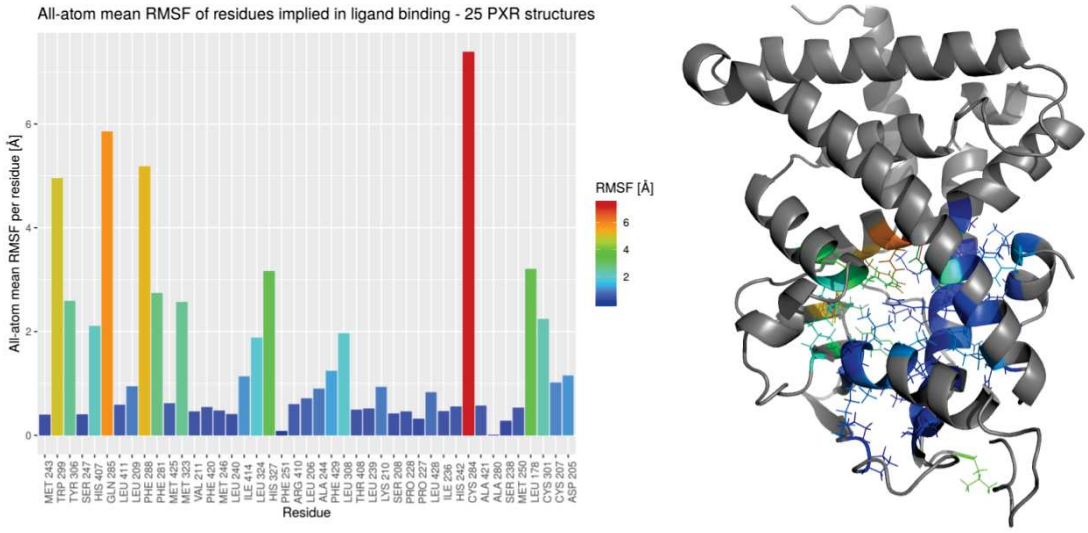


Figure 3.7: All-atom mean RMSF of residues implied in ligand binding of 25 liganded PXR protomers (left). The 43 residues are ordered based on identification frequency (compare Figure 3.6). A representative protein structure (PDB-ID: 4S0T, chain B) (right) is used to visualize the location of the identified residues within the structure. The side-chains of the 43 identified residues are shown in line representation on the protein structure and the coloring is also based on the all-atom mean RMSF in Å.

Residues are identified as involved in binding when located within a radius of 4 Å around the ligand (see Figure 3.6). The identified residues show a rather continuous decrease in their identification frequency, from one residue being identified in all 25 structures (100%) down to nine residues identified in two structures (8%) and six residues identified in only one structure (4%). The maximal RMSF value among the 43 identified residues of 7.40 Å is attained by CYS301, but which is only identified for two structures. Among the frequently identified residues (residues listed towards the left in Figure 3.7) there are three residues with particularly increased RMSF values: TRP299 with 4.96 Å, GLN285 with 5.86 Å, and PHE288 with 5.18 Å (ordered based on identification frequency). Those are followed by a set of 9 residues with slightly increased RMSF values between 1.90 Å and 3.17 Å. All others (30 residues) are showing a very rigid behaviour with RMSF values below 1.24 Å.

3.2.1.3 PXR - ensemble refinement

Since we have to rely on a very limited number of crystallographic structures, it is not possible to extract a rather complete view on the receptor's flexibility like for ERα. In order to obtain a better view on PXR's intrinsic flexibility in the crystalline state ensemble refinement[95] (as implemented in the refinement software PHENIX and automated in an in-house script testing different value ranges for three parameters) is employed. Here, an ensemble generation is considered as successful if both crystallographic R values, R_{work} and R_{free}, decrease during refinement.

The ensemble refinement for most of the PXR structures shows a better agreement with the experimental reflections compared to the refinement as single conformation, as indicated by improved R_{free} values (compare Table 3.1 column 'initial PDB' and column 'ensemble refinement'). Improvements are in the range of 0.012 to 0.058 in R_{free} difference. Only four structures, 2O9I, 4J5W, 4J5X and 4XAO demonstrate worsened R_{free} values upon ensemble refinement with R_{free} differences of 0.016, 0.019, 0.034 and 0.003, respectively.

Furthermore, missing loops are rebuilt and ensemble refinement is performed subsequently. The results show that this rebuilt loops do not contribute to a better agreement with the data as worse R_{free} values are obtained compared to the ensemble refinement without those loops. This can be explained by the fact that the missing loop regions are reflected by equivalently missing reflection data from the experimental side and therefore a reconstruction could lead to a bias during the refinement. In contrary to the missing loops, the building of missing side chains does not worsen R values remarkably compared to the ensemble refinement with the initial PDB structure. Thus, our findings strengthen the phenix.ensemble_refinement authors' recommendation to use completed side chains, but not rebuilding longer sequences of highly disordered regions. In total, 5 of the 20 PXR structures currently available in the PDB could not be used due to missing experimental data.

Additionally, the recent structure of PXR complexed with dabrafenib (PDB-ID: 6HJ2) is submitted to the PDB-REDO webserver[170] (that includes the construction of missing side chains) and subsequently subjected to ensemble refinement. The respective R_{work} and R_{free} values for the initial PDB are 0.190 and 0.232, for the output of PDB-REDO they are 0.193 and 0.221 (note the improved R_{free}), and for the ensemble refined PDB-REDO output they are 0.166 and 0.226 (note the improved R_{work}, but no improvement in R_{free}).

PDB ID & resolution		initial PDB	ensemble refinement	with added side chains	with loops
1ilg [a]	R_{work}	0.215	0.185	0.174	-
2.52 Å	R_{free}	0.279	0.248	0.249	-
1ilh	R_{work}	0.222	0.166	0.166	-
2.76 Å	R_{free}	0.282	0.245	0.249	-
1m13	R_{work}	0.212	0.166	0.167	0.171
2.15 Å	R_{free}	0.246	0.222	0.233	0.248
1nrl [d c]	R_{work}	0.216	0.161	0.160	-
2.0 Å	R_{free}	0.240	0.216	0.221	-
1skx	R_{work}	0.218	0.182	none	-
2.8 Å	R_{free}	0.266	0.241	missing	-
2o9i [d c]	R_{work}	0.228	0.175	none	-
2.8 Å	R_{free}	0.240	0.256	missing	-
3r8d	R_{work}	0.238	0.171	0.170	0.170
2.8 Å	R_{free}	0.289	0.248	0.246	0.263
4j5w [a t]	R_{work}	0.250	0.213	0.224	-
2.8 Å	R_{free}	0.298	0.317	0.323	-
4j5x [t c]	R_{work}	0.245	0.233	0.227	-
2.8 Å	R_{free}	0.298	0.332	0.336	-
4ny9	R_{work}	0.228	0.170	none	0.171
2.8 Å	R_{free}	0.298	0.240	missing	0.275
4x1f	R_{work}	0.182	0.153	0.161	-
2.0 Å	R_{free}	0.210	0.190	0.195	-
4x1g	R_{work}	0.174	0.167	0.166	-
2.25 Å	R_{free}	0.218	0.206	0.211	-
4xao	R_{work}	0.189	0.167	0.163	-
2.58 Å	R_{free}	0.239	0.242	0.246	-
4xhd	R_{work}	0.198	0.163	0.166	0.162
2.4 Å	R_{free}	0.242	0.218	0.217	0.232
5a86 [d]	R_{work}	0.231	0.208	none	-
2.25 Å	R_{free}	0.255	0.233	missing	-

Table 3.1: Summary of ensemble refinement results for PXR. Column 'initial PDB' are the single structure refinement values from the structure deposited in the PDB, column 'ensemble refinement' are the values obtained by refining the initial PDB structure as ensemble, followed by the columns of ensemble refined structures with added side chains and added loops. In all ensemble refinements the dataset with the best R_{free} is chosen. [a] = apo-structure, [d] = homodimer, [t] = heterotetramer with RXRα [c] = with SRC-1 coactivator.

3.2.1.4 BRAF - global analysis

All available liganded crystallographic BRAF structures of the protein kinase domain are gathered using the @TOME server, resulting in a total number of 96 liganded BRAF protomers, originating from 52 unique PDB entries. All protomeric sequences have a sequence identity between 93% and 100% with the canonical BRAF sequence (UniProt-ID: P15056). The liganded protomers are superimposed based on a rigid core calculated using the function *core.find* from the R package *bio3d*[171] (see Figure 3.8). For the BRAF ensemble comprising 96 protomers a rigid core of 82 residues (out of 199 non-gap residues, being present in all protomers and a total sequence of 276 analyzed residues) is identified with a cumulative volume cutoff at 0.5 Å^3 (and 110 positions with a cumulative volume $\leqslant$ 1 Å^3). Subsequently, based on the structural alignment Root Mean Square Fluctuations (RMSFs) across all 96 BRAF protomers are calculated and visualized as color code on the aligned structures in Figure 3.8 and as diagram along the sequence in Figure 3.9.

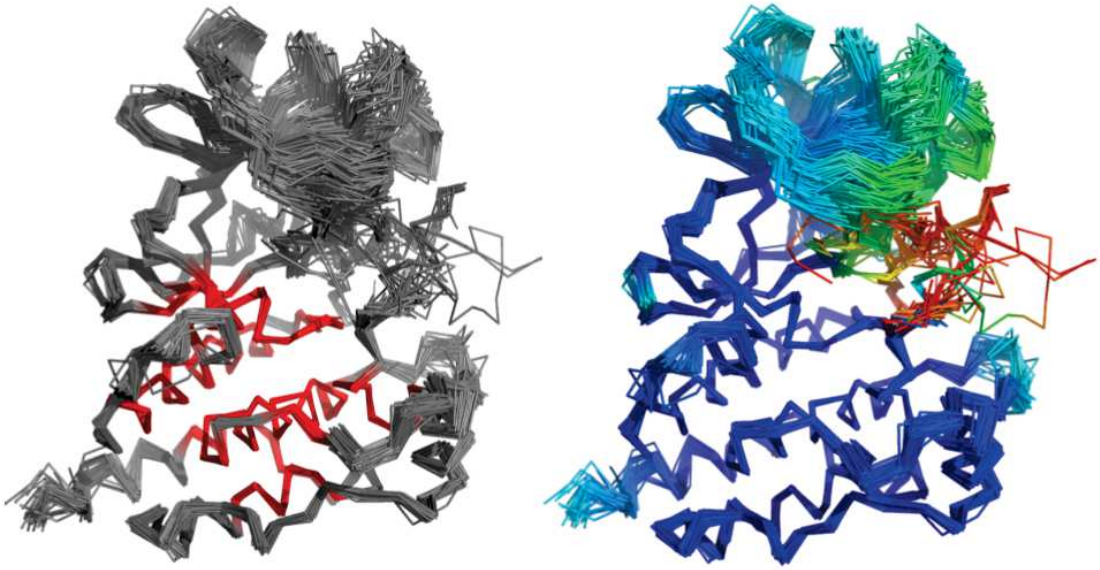


Figure 3.8: Rigid core superpositioning result (left) of 96 liganded BRAF protomers with 'rigid core' residues highlighted in red (calculated with a cumulative volume cutoff at 0.5 Å^3); and subsequent RMSF calculation across the protomers with rainbow color coding (from 0 Å in blue to a maximum of 9.1 Å in red).

For the BRAF ensemble the rigid core is composed of residues primarily located within the C-lobe. Furthermore, the highest RMSF values are detected for the activation loop (residues 594-623). Here, it has to be mentioned that the complete activation loop is only resolved in very few (9) of the 96 protomers, indicating already a high level of flexibility. The N-lobe shows overall an increased flexibility, and when having a closer look, a gradual tendency can be observed, with the highest fluctuations located just above the activation loop, which are decreasing towards the back and the interior of the protein - towards the connection with the C-lobe. This observation suggests a concerted movement of the N-lobe with the conformation of the activation loop.

To investigate this effect in more detail, Normal Mode Analysis (NMA) is performed on the single, complete and representative BRAF structure 5HID (chain B), and Principal Component Analysis (PCA), as well as conformer clustering are performed on the ensemble of 96 protomers. The NMA of the single structure, depicted as vector field representation in Figure 3.10, shows very well the

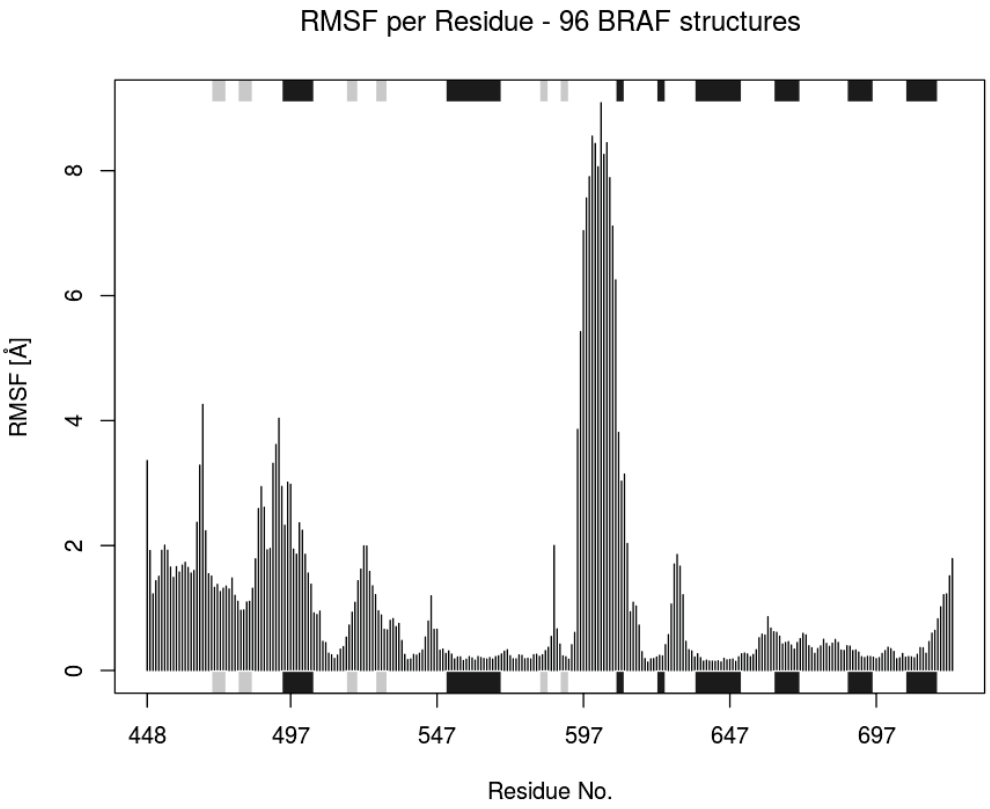


Figure 3.9: RMSF calculation across the 96 BRAF protomers plotted along the sequence with secondary structure annotation from a reference structure (PDB-ID: 4MBJ chain B).

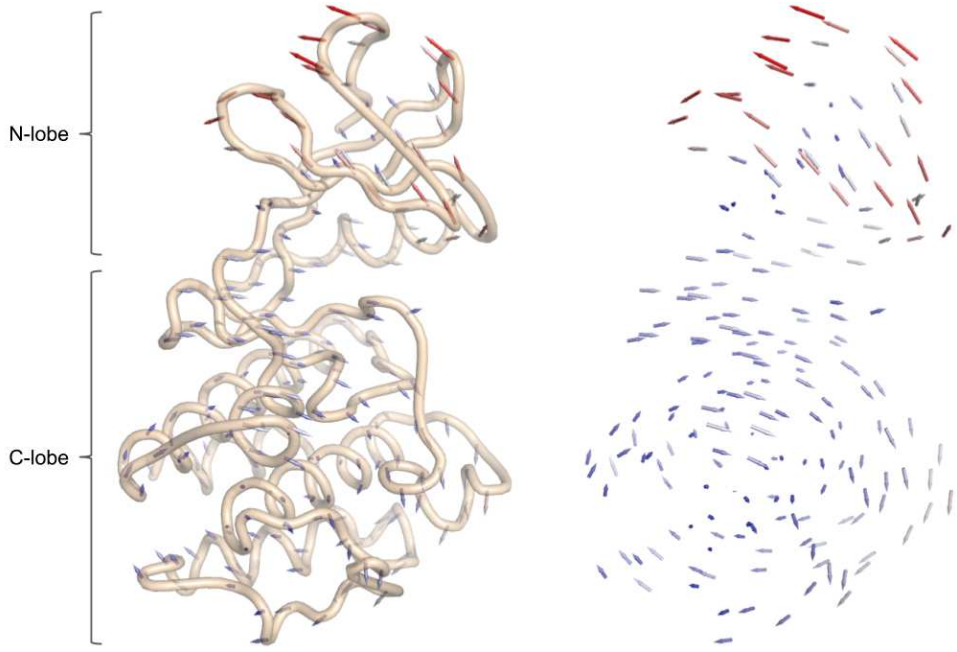


Figure 3.10: Normal mode analysis of BRAF structure 5HID (chain B). The modevectors (color-coded by direction) visualized on the protein structure (in sand colored tube representation) (left) and without the structure (right). The vector field representation is generated with the bio3d function *pymol.modes* and PyMol.

opening of the binding cleft by a concerted, outwards directed rotary movement of the two lobes, with slightly increased amplitudes (longer vectors) within the N-lobe.

Additionally, deformation energies and fluctuations are calculated based on the first three normal modes of 5HID (using the bio3d functions *deformation.nma* and *fluct.nma*). Deformation analysis provides information about the amount of local flexibility in the protein structure, such as atomic motion relative to neighboring atoms. Fluctuations (e.g. RMSF values), in contrast, provide amplitudes of the absolute atomic motion. The deformation analysis of the normal modes of 5HID highlights three high energy hot-spots on the structure (compare Figure 3.11 - left), the activation loop, C532 within the hinge region, and N500 within the αC helix. Concerning the fluctuations (Figure 3.11 - right), the normal mode amplitudes have increased values within the N-lobe compared to rest of the structure, as already reflected by the vector field representation in Figure 3.10.

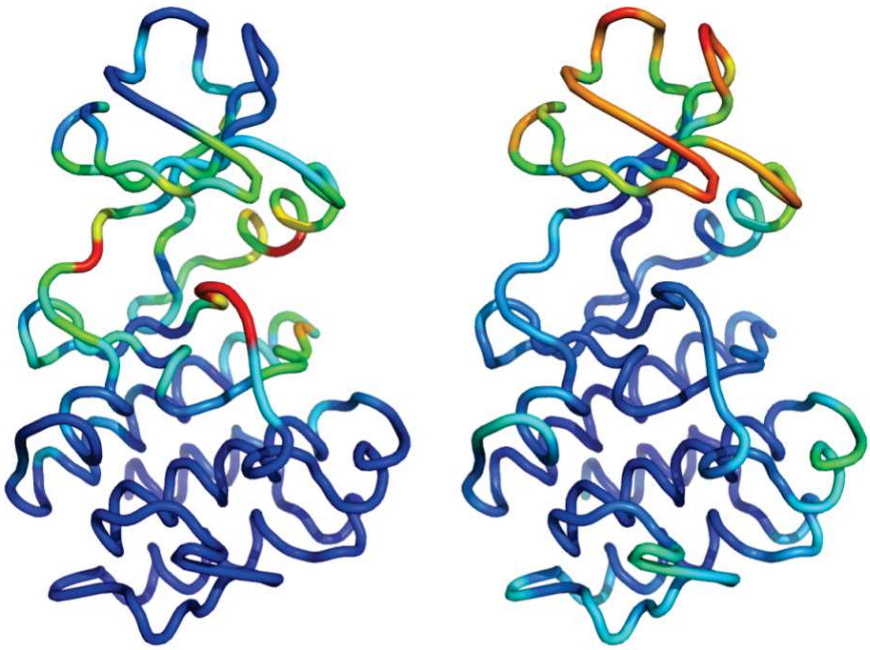


Figure 3.11: Deformation energies (left) and fluctuations (right) based on the first three normal modes of BRAF structure 5HID (chain B) color coded (rainbow: blue to red) onto the structure. Values range from 0.10 to 11.42 for the deformation energies, and from 0.00 to 0.31 for the fluctuations.

To investigate whether the ensemble of 96 BRAF protomers reflects the same tendencies as the NMA on a single structure, or gives different insights into the conformational variability, PCA and conformer clustering are performed on the ensemble. PCA has the advantage that it highlights the regions of the protein which are varying across the ensemble. Therefore, it is a suitable technique to identify outliers. PCA offers even more very useful features: It is a tool for data reduction, as a large amount of variance can often be explained by a small number of PCs; the first PC explains the highest proportion of variance, and subsequent PCs explain decreasing proportions of the variance; and all PCs are uncorrelated with one another, due to their orthogonality. To apply PCA, the protein structures have to be reduced in complexity, which is done here by simply using matrices of the distances between the atomic Cartesian coordinates of the Cα atoms (to avoid that the variation would be dominated by the movements of surface side-chains) and omitting all other information.

PC analysis of the BRAF conformational ensemble (see Figure 3.12) shows a distinct importance of the first PC (PC1), explaining 51.8% of the variance, followed by PC2 with 13.0%. Concerning the residue contribution PC1 is principally governed by the activation loop (a-loop) and by the N-lobe

as a whole, whereas PC2 is primarily directed by movement of the αC helix and slightly also by the a-loop.

As PCA has the potential to classify calculated structures according to the correlated structural variation across the ensemble, it is used for conformer clustering, in parallel to RMSD based hierarchical clustering. The two employed methods, RMSD and PCA clustering (see Figure 3.13) clusters the conformers in a similar way, but not completely identical. Nevertheless, the same trends are visible, as both clustering methods distinguish between differences in N-lobe positioning, especially differences of the glycine-rich loop and of the orientation of helix αC.

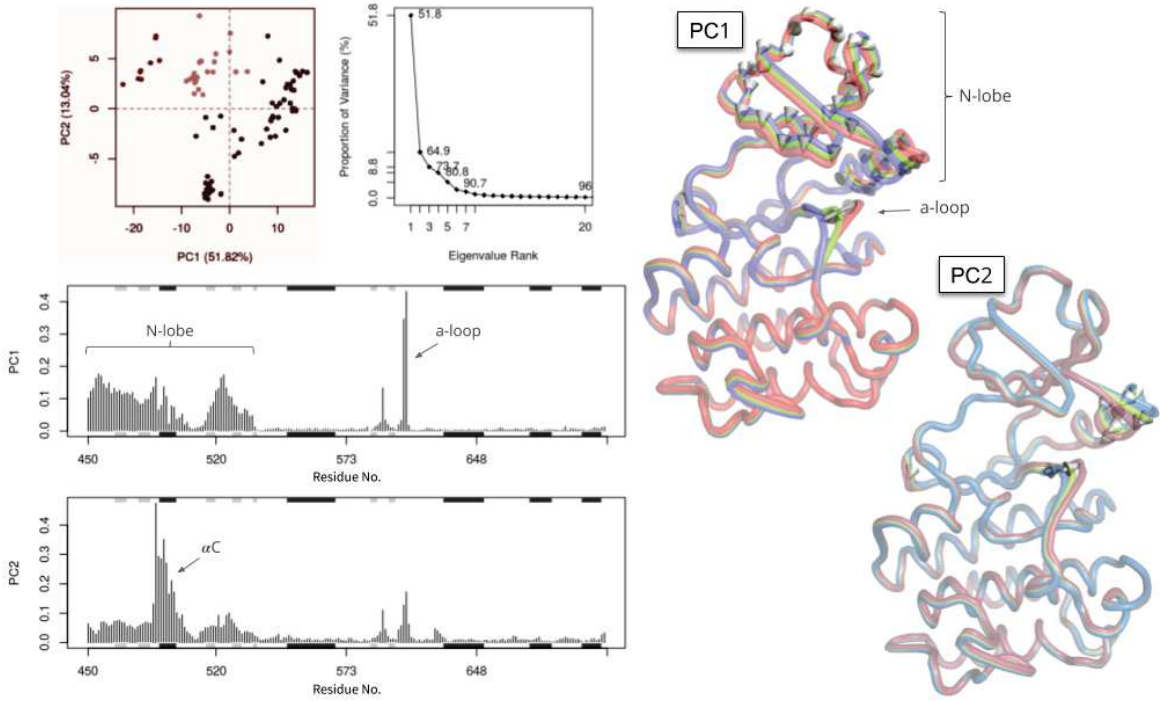


Figure 3.12: Principal Component Analysis of the BRAF conformational ensemble: Proportion of the variance covered by the PCs (top left), per residue contributions to the first two PCs (bottom left), and structural representations of the first two PCs with vectors (grey arrows) calculated using the *modevectors* module in PyMol (right).

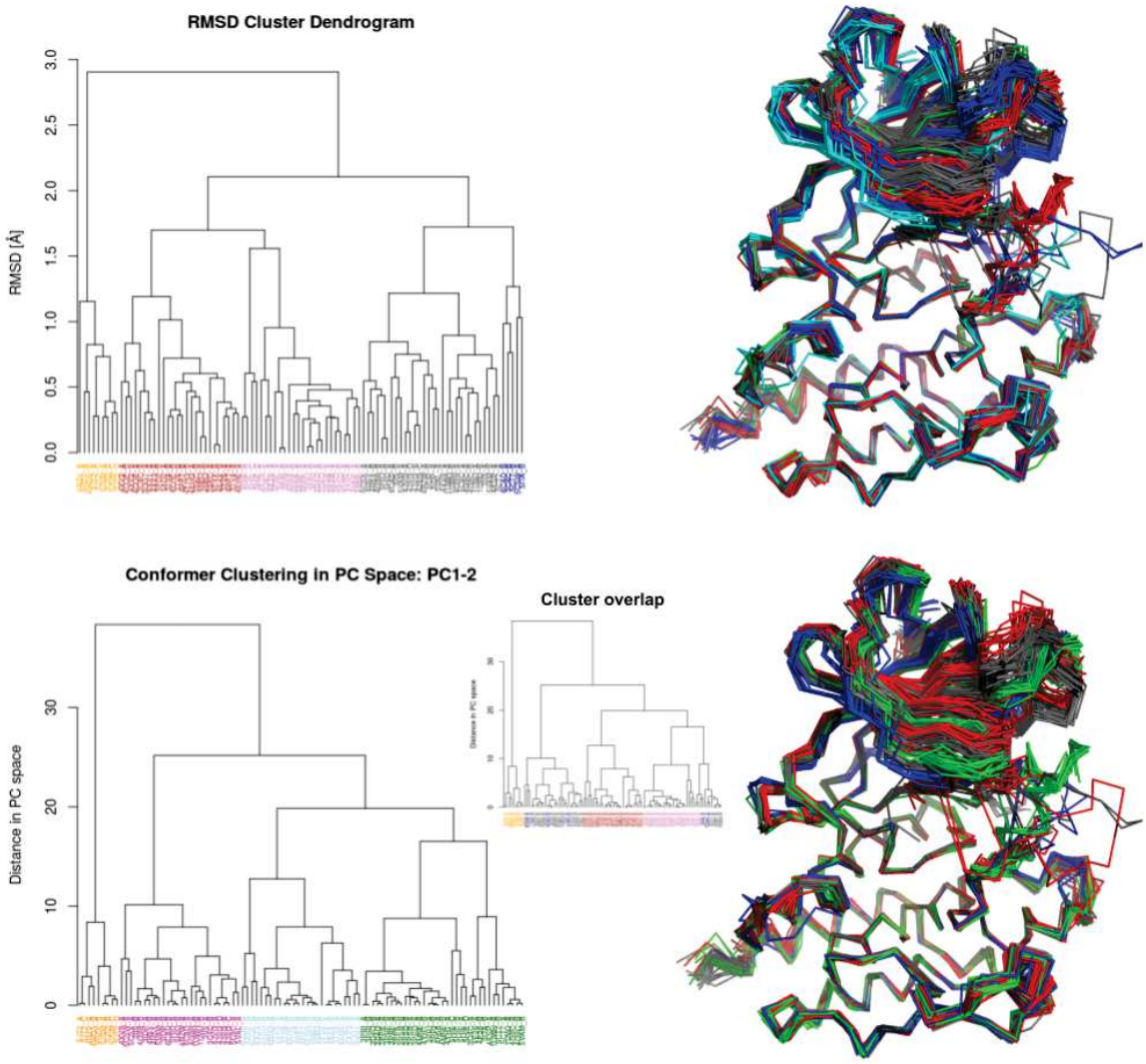


Figure 3.13: Hierarchical clustering based on RMSD (top row) and clustering in PC-space (bottom row) of 96 liganded BRAF protomers. The dendrogram inset in the bottom row shows the cluster overlap of the two methods by coloring the PC-space dendrogram based on RMSD clusters. The superimposed structures are colored based on the identified clusters.

3.2.1.5 BRAF - binding pocket analysis

To investigate the importance of specific residues, the frequency of implication in ligand binding is tracked across 96 liganded BRAF protomers, which are all available liganded PDB structures. A residue is identified as 'implied' if any of its atoms is found within a radius of 4 Å around any atom of the ligand. This results in a list of 52 identified residues with varying implication frequencies (Figure 3.14). Nonetheless, a 'high frequency' stet of 15 residues, identified in more than 79% of the structures, can be distinguished (see Figure 3.14 residues with coloring red to orange), as the next highest frequency is only 50% (green color), marking a significant step. Two further light steps can be observed, one from 30% to 22% (cyan to lightblue), and another from 17% to 10% (lightblue to darkblue).

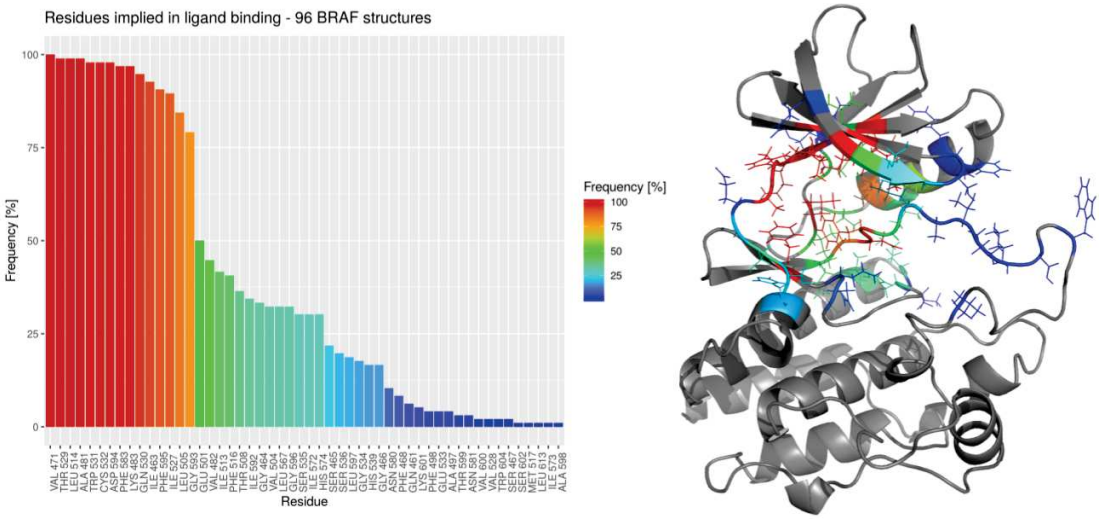


Figure 3.14: Residues implied in ligand binding of 96 liganded BRAF protomers (left), colored by the frequency they are identified being within a radius of 4 Å around the ligand. A representative protein structure (PDB-ID: 5HID, chain B) (right) is used to visualize the location of the identified residues within the structure. The side-chains of the 52 identified residues are shown in line representation on the protein structure and the coloring is also based on identification frequency.

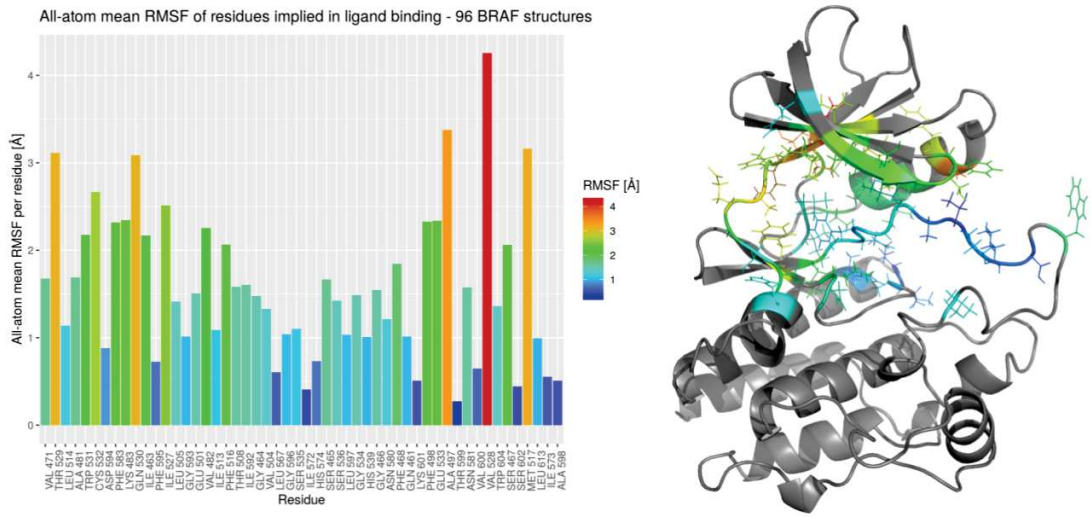


Figure 3.15: All-atom mean RMSF of residues implied in ligand binding of 96 liganded BRAF protomers (left). The 52 residues are ordered based on identification frequency (compare Figure 3.14). A representative protein structure (PDB-ID: 5HID, chain B) (right) is used to visualize the location of the identified residues within the structure. The side-chains of the 52 identified residues are shown in line representation on the protein structure and the coloring is also based on the all-atom mean RMSF in Å.

3.2.1.6 BRAF - ensemble refinement

For protein kinase BRAF six crystallographic structures are subjected to ensemble refinement (see Table 3.2). Missing side chains were not rebuild, as only one or none were missing. Note that large parts of the activation loop are completely unresolved in most structures (no backbone present) and are thus not modelled prior to ensemble refinement (in agreement with author suggestions of ensemble refinement[95]). Only for two of the six crystallographic BRAF structures (4XV3 and 4XV9)

the ensemble refinement shows better agreement with the experimental reflections compared to the refinement as single conformation, as indicated by improved R_{free} values. Thus, BRAF is expected to obtain a rather rigid conformation within the crystal for the resolved parts (excluding e.g. the non-resolved activation loop).

PDB ID & resolution		**initial PDB**	**ensemble refinement**	**with added side chains**
4XV1	R_{work}	0.234	0.201	only LYS 522
2.47 Å	R_{free}	0.273	0.284	missing
4XV2	R_{work}	0.212	0.169	only LYS 522
2.50 Å	R_{free}	0.244	0.233	missing
4XV3	R_{work}	0.258	0.221	only LYS 522
2.80 Å	R_{free}	0.296	0.310	missing
4XV9	R_{work}	0.205	0.142	only LYS 522
2.00 Å	R_{free}	0.238	0.181	missing
5CSW	R_{work}	0.217	0.205	none
2.66 Å	R_{free}	0.282	0.286	missing
NEW*	R_{work}	0.195	0.176	none
2.37 Å	R_{free}	0.250	0.256	missing

Table 3.2: Summary of ensemble refinement results for BRAF. Column 'initial PDB' are the single structure refinement values from the structure deposited in the PDB, column 'ensemble refinement' are the values obtained by refining the initial PDB structure as ensemble. In all ensemble refinements the dataset with the best R_{free} is chosen. * newly solved structure in complex with a drug candidate.

Key points

⇒ For the nuclear receptor PXR there is only a limited number of crystallographic structures available, compared to ERα, and PXR's intrinsic (side-chain) flexibility is expected to be elevated due to the versatility of binding ligands and it's detoxification role within the organism.

⇒ Protein kinase BRAF is a well studied drug target with 52 liganded PDB entries. Its activation loop represents the part with the highest variability/flexibility.

⇒ The efficient, accurate and representative sampling of the conformational space represents a major challenge.

3.3 Where to modify the drug?

3.3.1 Drug binding in target (BRAF) and anti-target (PXR)

The binding modes of dabrafenib (PDB-chemicalID: P06) in both, primary target BRAF and anti-target PXR, were analyzed (see Figure 3.16 and 3.17).

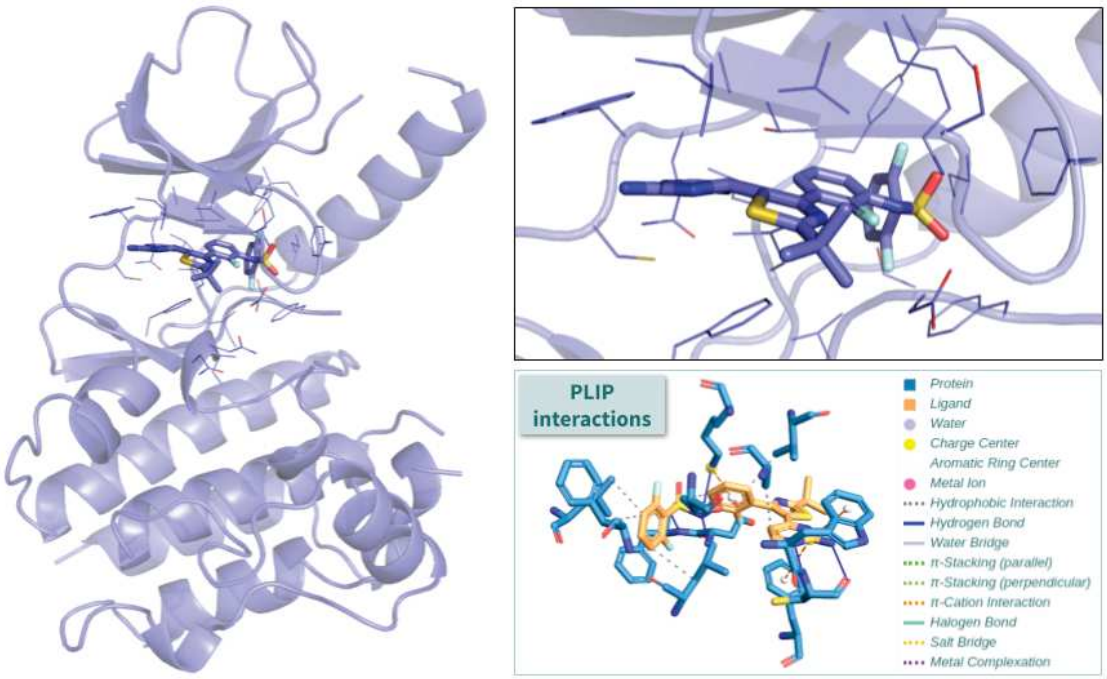


Figure 3.16: The binding mode of dabrafenib in its primary target BRAF (PDB-ID: 4XV2, chain A, downloaded from PDB-REDO).

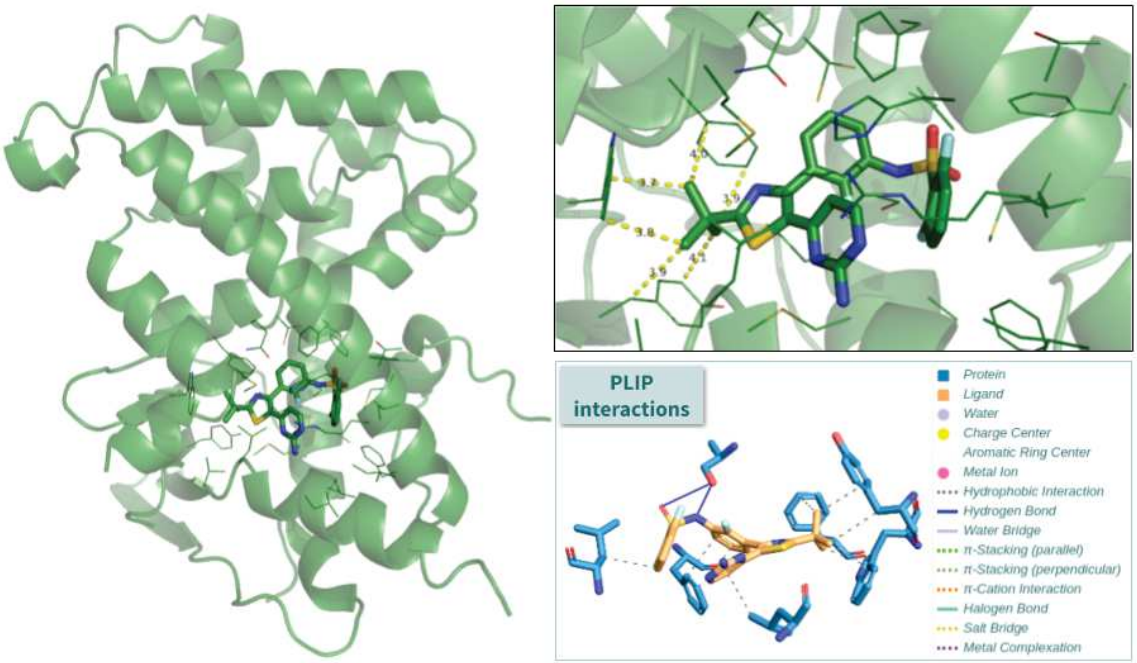


Figure 3.17: The binding mode of dabrafenib in its anti-target, the nuclear receptor PXR (PDB-ID: 6HJ2, chain A, processed with PDB-REDO).

The interactions of P06 with both of its targets that were identified by the PLIP web-server[172] and are depicted in Figure 3.16 and 3.17 are listed with further details in Figure 3.18 and 3.19.

Hydrophobic Interactions ····

Index	Residue	AA	Distance	Ligand Atom	Protein Atom
1	471A	VAL	3.54	3993	170
2	481A	ALA	3.78	3976	252
3	505A	LEU	3.78	3999	422
4	514A	LEU	3.87	4003	499
5	516A	PHE	3.77	3999	515
6	529A	THR	3.80	3992	614

Hydrogen Bonds —

Index	Residue	AA	Distance H-A	Distance D-A	Donor Angle	Protein donor?	Sidechain	Donor Atom	Acceptor Atom
1	483A	LYS	3.25	3.65	105.03	✔	✔	268 [N3]	3996 [Npl]
2	532A	CYS	1.96	2.91	162.24	✔	✖	638 [Nam]	3977 [N2]
3	532A	CYS	2.12	3.06	158.81	✖	✖	3979 [Npl]	641 [O2]
4	594A	ASP	1.90	2.82	153.89	✔	✖	1140 [Nam]	3996 [Npl]
5	595A	PHE	1.93	2.87	159.90	✔	✖	1148 [Nam]	4007 [O2]

π-Cation Interactions ····

Index	Residue	AA	Distance	Offset	Protein charged?	Ligand Group	Ligand Atoms
1	483A	LYS	4.06	2.00	✔	Aromatic	3989, 3990, 3991, 3992, 3993, 3994
2	531A	TRP	3.95	1.44	✖	guanidine	3979, 3975, 3977
3	583A	PHE	5.28	0.74	✖	guanidine	3979, 3975, 3977

Figure 3.18: PLIP interactions of dabrafenib in its primary target BRAF (PDB-ID: 4XV2, chain A, downloaded from PDB-REDO).

Hydrophobic Interactions ····

Index	Residue	AA	Distance	Ligand Atom	Protein Atom
1	209A	LEU	3.83	2275	450
2	281A	PHE	3.57	2291	1012
3	281A	PHE	3.75	2293	1013
4	288A	PHE	3.94	2286	1079
5	288A	PHE	4.00	2285	1076
6	299A	TRP	3.43	2285	1164
7	299A	TRP	3.56	2287	1165
8	306A	TYR	3.92	2286	1225
9	306A	TYR	3.90	2287	1223
10	411A	LEU	3.61	2302	2084

Hydrogen Bonds —

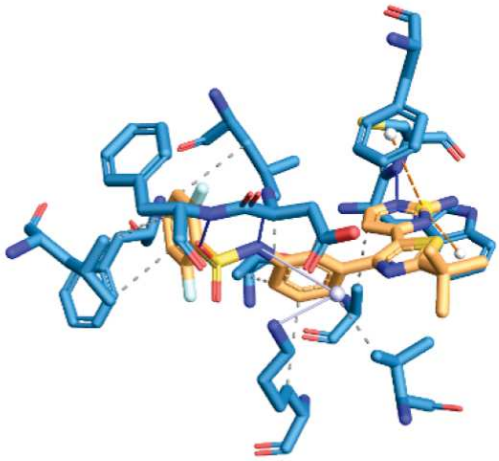
Index	Residue	AA	Distance H-A	Distance D-A	Donor Angle	Protein donor?	Sidechain	Donor Atom	Acceptor Atom
1	247A	SER	3.09	3.80	131.84	✔	✔	742 [O3]	2305 [O2]
2	247A	SER	1.68	2.60	154.11	✖	✔	2295 [Npl]	742 [O3]

Figure 3.19: PLIP interactions of dabrafenib in its anti-target, the nuclear receptor PXR (PDB-ID: 6HJ2, chain A, processed with PDB-REDO).

As BRAF structure 4XV2 is crystallized in a dimeric form, with both chains being complexed with dabrafenib, it is possible to compare the interactions identified by PLIP from chain A (Figure 3.16 and 3.18) with the ones from chain B (see Figure 3.20). In chain B there is a water molecule in the binding site that forms a water bridge from the nitrogen of the aminosulfoxide moiety of the ligand to Lys483, substituting the direct hydrogen bonding to Lys483, which is found in chain A (at a rather large D-A distance of 3.65 Å). The hydrophobic interactions are identical between the two chains (in chain B there are two more interactions listed than in A, which are interactions with the same residues already listed in A - Leu505 and Leu514 - and therefore considered as transient additional interactions).

Hydrophobic Interactions ····

Index	Residue	AA	Distance	Ligand Atom	Protein Atom
1	471B	VAL	3.54	4028	2157
2	481B	ALA	3.78	4011	2240
3	483B	LYS	3.91	4029	2252
4	505B	LEU	3.65	4034	2422
5	505B	LEU	3.77	4036	2423
6	514B	LEU	3.74	4038	2499
7	514B	LEU	3.75	4027	2502
8	516B	PHE	3.70	4036	2515
9	529B	THR	3.68	4027	2614



Hydrogen Bonds —

Index	Residue	AA	Distance H-A	Distance D-A	Donor Angle	Protein donor?	Sidechain	Donor Atom	Acceptor Atom
1	532B	CYS	1.98	2.95	166.32	✔	✖	2638 [Nam]	4012 [N2]
2	594B	ASP	1.95	2.85	152.03	✔	✖	3124 [Nam]	4031 [Npl]
3	595B	PHE	1.89	2.87	171.73	✔	✖	3132 [Nam]	4042 [O2]

Water Bridges —

Index	Residue	AA	Dist. A-W	Dist. D-W	Donor Angle	Water Angle	Protein donor?	Donor Atom	Acceptor Atom	Water Atom
1	483B	LYS	4.09	2.74	149.07	133.13	✔	2256 [N3]	4031 [Npl]	4083

π-Cation Interactions ····

Index	Residue	AA	Distance	Offset	Protein charged?	Ligand Group	Ligand Atoms
1	531B	TRP	3.99	1.47	✖	guanidine	4010, 4012, 4014
2	583B	PHE	5.22	0.47	✖	guanidine	4010, 4012, 4014

Figure 3.20: The binding mode of dabrafenib in its primary target BRAF (PDB-ID: 4XV2, chain B, downloaded from PDB-REDO).

3.3.2 BRAF structures with similar ligands

An additional approach to define the parts of the molecule that can be modified was to search for BRAF structures with similar ligands. These structures give insights on the expected variability of the binding mode and the adaptability of the binding pocket. Upon an inverse screening with P06 on all available liganded BRAF structures on the @TOME server, the four complexes with the

highest ligand Tanimoto scores (ranging from 0.84 to 0.33) were extracted and the protein structures were superimposed (see Figure 3.21). All four ligands show a highly similar binding pose with the ring systems positioned in the same way, the sulfoxide moiety at the same position and with the same orientation, and a highly similar position of the nitrogen atoms within the hinge-binding moiety establishing hydrogen bonds to the hinge backbone residues. The high similarity of the poses (despite larger chemical discrepancies) indicates a good and stable fit within the binding pocket. The highest variability can be found at the two ends of the molecules that are pointing outside the binding pocket, 1) the extension of the hinge binding segment, which is still largely in contact with the protein, and 2) the extension towards the binding pocket entry (a tertiary butyl moiety in P06), which is only present in the two most similar molecules (P02 and CQE), but shows a higher location variability.

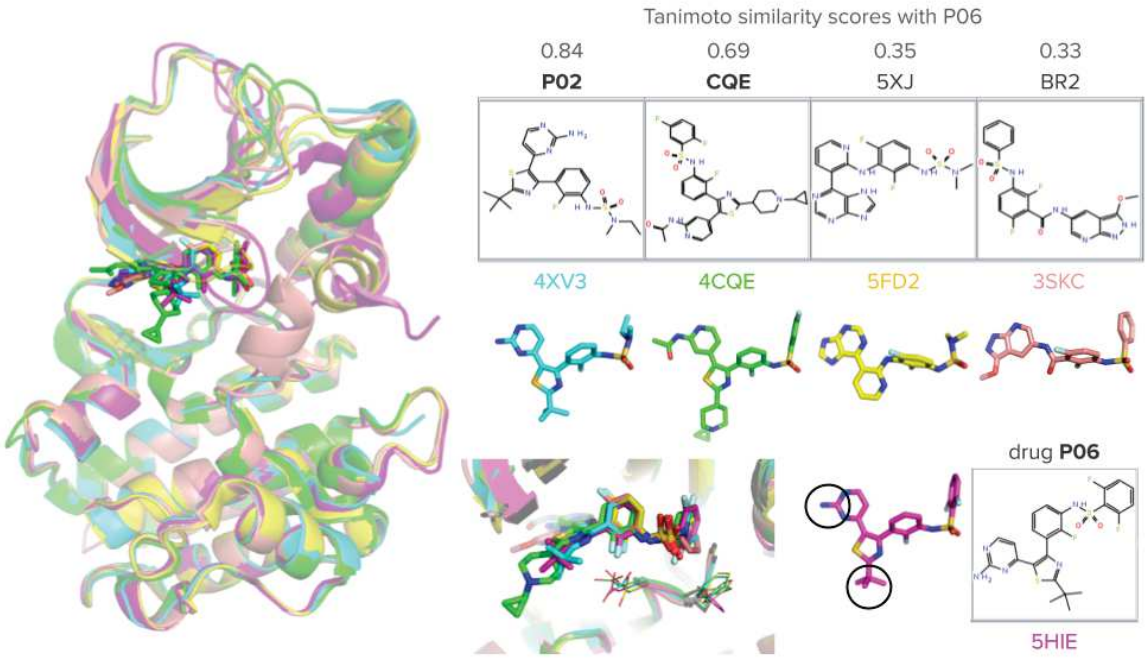


Figure 3.21: PDB structures of BRAF with ligands similar to dabrafenib (P06) with superimposed protein structures and comparison of the crystallographic ligands' chemistry and binding mode. Coloring by crystallographic complex (as for the PDB entry codes) and ligand IDs in grey.

3.3.3 Interfering with drug metabolism

An important aspect of the drug design strategy is the reduction of the fast metabolism of dabrafenib into its three identified major metabolites hydroxy-dabrafenib (HDB), carboxy-dabrafenib (CDB), and desmethyl-dabrafenib (DDB) (see Figure 1.2 in Chapter 1). Since the point of attack for CYP-mediated oxidation is the tertiary butyl moiety of dabrafenib, the modification of this moiety is expected to reduce the metabolism of dabrafenib and should therefore hopefully lead to increased bio-availability and reduced secondary effects.

Key points

⇒ The layout of the drug design approach is based on the attempt to reduce drug metabolism and on the comparison of the binding modes of dabrafenib (P06) in its primary target BRAF and its secondary target PXR, and of similar ligands in BRAF.

3.4 *In silico* synthesis of drug candidates

Potential drug candidates are synthesized *in silico* by performing the following steps:

1. Analyzing dabrafenib's chemical synthesis on SciFinder (`https://www.cas.org/products/scifinder`);
2. Defining the modification step within the synthesis pathway: the tertiary butyl moiety of dabrafenib can be replaced by substituting the pathway reactant containing a thioamide moiety that leads to cyclization of the 2-amino group attached to the pyrimidine. The corpus of the molecule except of the part to be modified (the tertiary butyl moiety) is defined as scaffold;
3. Searching for purchasable fragments containing a thioamide moiety to replace the trimethyl moiety of dabrafenib: 179 molecules were found by substructure search on Enamine Chemical Supplier (`https://www.enaminestore.com/search`) using enamine structure: thioamide "NH2-C(=S)-CH2" and the "Advanced Filters: Stock amount (mg) = 100" (due to availability reasons);
4. Then, the 179 thioamide containing reactants (termed fragments) are used within a KNIME workflow that performs *in silico* chemical reactions with the pre-dabrafenib (termed scaffold) to produce the final three-dimensional drug candidates (see Figure 3.22).
5. An additional step is included within the KNIME workflow that removes three different protecting groups from the drug candidates. A protecting group is usually introduced into a molecule to "protect" a functional/reactive group and therefore obtain chemoselectivity, which ensures the desired reaction. Protecting groups are frequently used in multistep organic synthesis and are removed before the final product is obtained. The three different protecting groups present within the set of 179 fragments are BOC (20x), phthalimide (4x), and benzyl carbamate (1x).

Moreover, the defined scaffold is also modified, resulting in a selection of five different scaffolds (that are named after their PDB-ligand-IDs):

- *P02* - molecule closest related to dabrafenib available in the PDB (PDBID: 4XV3) and supposed to avoid paradoxical activation of WT-BRAF,
- *P02C* - P02 having a carbon atom instead of the nitrogen connecting the methyl-ethyl moiety,
- *P06* - the original dabrafenib scaffold,
- *P06F* - P06 with one fluor atom shifted from cis to trans position at the di-fluorophenyl ring,
- *P06FCl* - P06F with an additional chlorine atom added in para to the central fluorophenyl ring.

This protocol resulted in a final set of 179 drug candidates per scaffold in 3D mol2 or sdf format (among which there are a few duplicates as result from the deprotection step). Each molecule name contains the scaffold ID ('P02', 'P02C', 'P06', 'P06F', or 'P06FCl') and the fragment ID from the supplier to ensure easy identification and purchase later on.

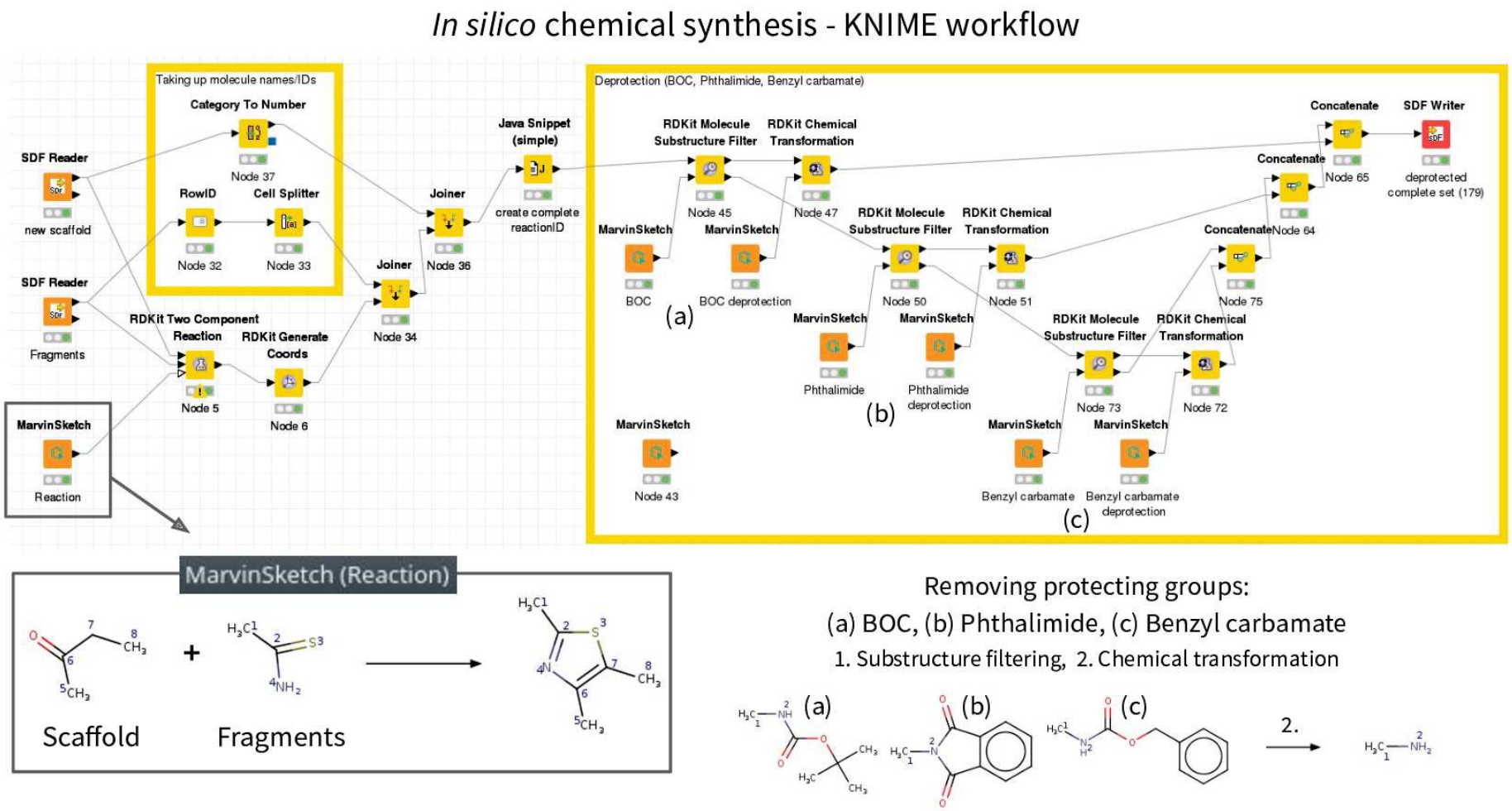


Figure 3.22: *In silico* synthesis of drug candidates - KNIME workflow.

3.5 Computational approaches investigating targets and ligands

3.5.1 Molecular modelling and molecular dynamics

Molecular dynamics (MD) simulations can give valuable insights into the dynamic behavior of the target protein alone, as well as in complex with potential drug candidates. Moreover, the MD simulation snapshots can subsequently be used for affinity estimation calculations (e.g. by MM-PBSA).

Before an MD simulation can be performed, the structure to be simulated needs to be prepared. Important points to be considered are the modelling of missing side-chains, entire residues or even whole loops that are not present in the crystallographic structure, the definition of proper protonation states of protein and ligand, and of the respective parameters that are used during simulation and which need to be compatible with the employed force field.

MD simulations were performed on a variety of complexes, including variations of the BRAF structure and different ligands, which were either already present in the crystal structure, or docked into the protein by using the crystallographic ligand as molecular shape restraint (anchor) with PLANTS.[173]

Initial investigations started with the crystallographic complex of BRAF and dabrafenib (abbreviated as DB or P06) at the best available resolution of 2.5 Å (PDB-ID: 4XV2), and with the use of crystallographic complexes with the highly similar ligands P02 and CQE (PDB-IDs: 4XV3 and 4CQE, with resolutions of 2.8 Å and 2.3 Å, respectively). Subsequently, crystallographic ligands were replaced with DB, its three major metabolites, and several designed drug candidates.

The impact of different protein conformations was evaluated by using different starting models for MD simulation. MD simulation replicas were generated with differing initial velocities that result in different trajectories of the systems. Both approaches help to evaluate the expected error for affinity estimations performed by subsequent MM-PBSA calculations, for which an additional statistical error estimation method, namely bootstrapping, was tested.

3.5.1.1 Molecular modelling - BRAF and its loops

The glycine-rich loop (G-rich loop) and the activation loop (a-loop) are two essential parts for the function of a kinase. The G-rich loop is important for the access of the binding site and often interacting with the ligand. The a-loop is known to be present in an extended, unfolded conformation when the enzyme is active. As very flexible region, it is typically not resolved in the crystal structure, especially not for the constitutively active V600E mutant. Only eleven among 98 available protomeric structures have a completely resolved a-loop (PDB-ID_chain-ID: 3SKC_B, 3TV4_B, 3TV6_B, 4E4X_B, 4EHE_B, 4H58_C, 4MBJ_B, 4MNE_B, 4PP7_B, 4RZV_B, 5HID_B). It is only complete in one protomer

when the structure is resolved at least as dimer. This is indicative for a required stabilizing effect of the dimerization. Furthermore, the a-loop can adopt a folded, helical structure in an inactive kinase conformation. This helical conformation is only observed in wild-type BRAF, as the V600E mutation would lead to steric clashes with the αC helix (see Figure 3.24). Additionally, smaller loops, such as the flexible G-rich loop are also frequently missing in the crystallographic structures. 4XV2 (chain A) has 25 missing internal residues, with 18 of them in the a-loop (597-614), but a completely resolved G-rich loop; 4XV3 (chain A) has 26 missing internal residues, with 17 of them in the a-loop (597-613) and 3 of the G-rich loop (466-468); 4CQE (chain A) has 22 missing internal residues, with 18 of them in the a-loop (597-614) and 2 of the G-rich loop (467,468).

In order to perform MD simulations the structures need to be completed to avoid unnatural artifacts produced by the chain breaks. In the present work, models were generated using the Modeller software[174,175] within an in-house python script. The script employs sequence and PDB utilities (Bio.SeqIO and Bio.PDB) from Biopython,[176] the multiple dynamic programming alignment (MALIGN) of Modeller and customizes Modeller's automodel and loopmodel classes with an automatic loop selection. The selection of loop residues to be modelled is based on an upstream structure analysis and sequence alignment that selects only the internal non-resolved (missing) residues, plus one adjacent residue upstream and downstream the missing part (to avoid "kinking" artifacts from terminal residues modelled wrongly in ambiguous X-ray data, such as the often occurring confusion between main-chain and side-chain density of the terminal residue). This procedure allows for keeping the exact coordinates of atoms present in the template, which may have an important structural and functional impact e.g. in ligand binding.

Five different models of the complete BRAF kinase domain were used for simulations and subsequent MM-PBSA calculations:
First, three BRAF models based on sequence and structure present in the PDB structures 4XV2, 4XV3 and 4CQE were generated, resulting in different loop conformations, in particular for the long and usually unresolved activation loop (see Figure 3.23). The three PDB structures do not only contain the oncogenic V600E mutation but also 16 solubilizing mutations (I543A, I544S, I551K, Q562R, L588N, K630S, F667E, Y673S, A688R, L706S, Q709R, S713E, L716E, S720E, P722S, and K723G - permitting kinase domain overexpression in bacteria), with 13 of them being present in the structure (the last 3 of them, S720E, P722S, and K723G, are located in the unresolved C-terminus).

Furthermore, as crystallographic BRAF structures differ from the canonical BRAF WT sequence in $\sim$16 residues, due to purification/solubility reasons, the modelling approach was extended to the WT sequence and a sequence with the single oncogenic mutation V600E (otherwise as WT).
BRAF-V00E complexed with dabrafenib was modelled based on PDB structure 4XV2 (chain A) with the canonical sequence except for V600 mutated to glutamate (E).
The BRAF-WT model was generated using a further extended structure homology approach. Here, the BRAF-WT sequence, PDB structure 4XV2 (chain A) and PDB structure 3SKC (chain B) were used for building a (homology) model that has the BRAF-WT sequence, atomic coordinates of 4XV2 for all resolved atoms and a-loop coordinates as close as possible to the structure of 3SKC.

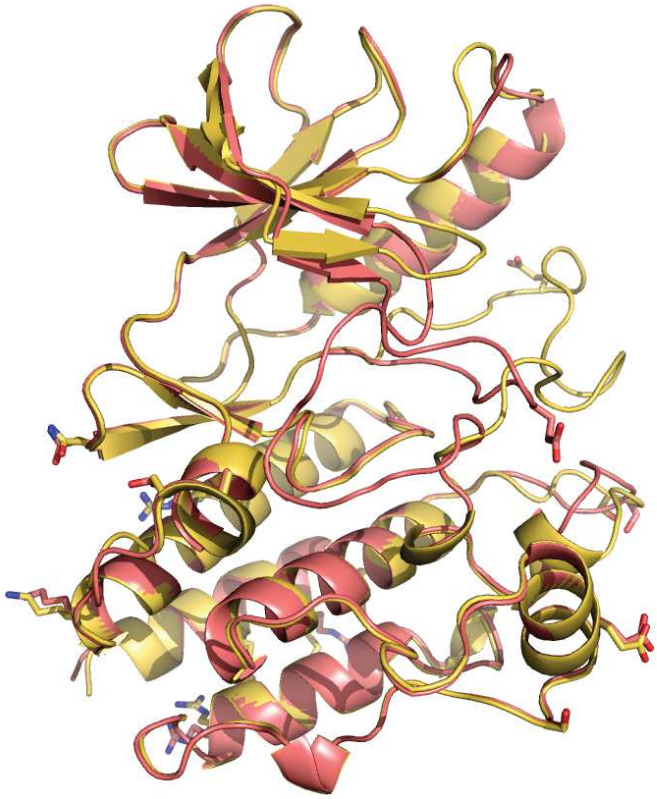


Figure 3.23: Two loop-complete models with mutated BRAF sequences: 1) a loop-model based on PDB structure 4XV2 (chain A) (rose) with an extended a-loop that is folded back onto the structure, and interacting with the G-rich loop and 2) a loop-model based on PDB structure 4CQE (chain A) (yellow) with an extended and free a-loop. All mutated residues are shown in stick representation. These are the V600E mutation and 13 solubilizing mutations (I543A, I544S, I551K, Q562R, L588N, K630S, F667E, Y673S, A688R, L706S, Q709R, S713E, L716E - permitting kinase domain overexpression in bacteria).

3.5.1.2 Molecular dynamics simulations

Molecular dynamics simulation - methods

All simulations were carried out with Gromacs 2018.[97] The ligand topologies were generated using the ACPYPE/ANTECHAMBER[177] program of AmberTools17[112] with partial charges generated by the empirical charge model AM1-BCC. The ligands' parameters are based on the General Amber Force Field (GAFF) and the Amber FF14SB force field was employed for the proteins. Each complex was solvated in a TIP3P water dodecahedral box, with periodic boundary conditions and a minimum distance of 1.0 nm from the surface of the complex to the edge of the box. Each system was neutralized by adding NA+ and Cl- ions to physiological concentration of 150 mM. A completely free steepest descent energy minimization for 2000 steps was followed by a 100-ps NVT equilibration and a 100-ps NpT equilibration with Parrinello-Rahman pressure coupling. NVT and NpT equilibrations were performed at a reference temperature of 300 K with ligand restraints of 1000 kJ/mol nm^2 in x,y,z directions. Finally, 50 ns unrestrained production runs were performed with a 2 fs time-step in the NpT ensemble and snapshots were saved every 10 ps. For each complex, usually five replica simulations were run with different randomly assigned initial velocities, resulting in a total of 250 ns simulation per complex.

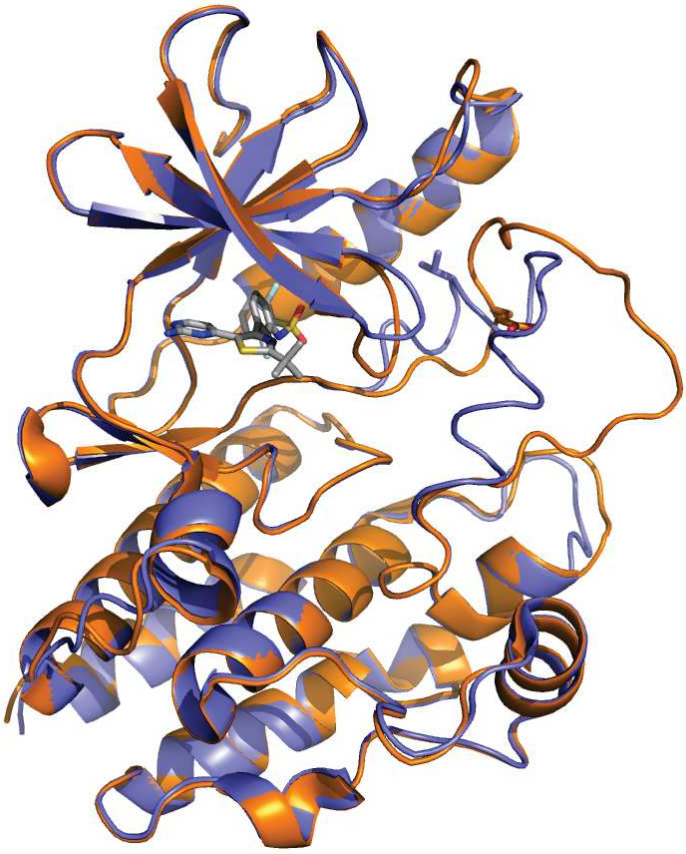


Figure 3.24: Two complete models with naturally occurring sequences: BRAF-V600E with extended a-loop (orange) and BRAF-WT (V600) with structured a-loop conformation (violet). Residue 600 is shown in stick representation and also ligand P06 (grey).

Analysis and visualization was performed with Gromacs tools, PyMol, VMD, Chimera, and Python scripts.

Helix formation of the activation loop

For all simulations of the WT model with the structured helical activation loop, the loop stays stable in this structured form. For the simulation of V600E the unstructured activation loop shows different degrees of mobility with varying fluctuation amplitudes. Interestingly, during one simulations an unstructured activation loop forms a complete helix (residues 611-620) within a simulation time of 100 ns (see Figure 3.25). Visual analysis of the MD trajectory shows that the helix starts forming from residue Met620 and gradually extends upstream the sequence, reaching residue Glu611 at the end of the 100 ns simulation. Thus, a further extension of the helix during longer simulation is probable. Secondary structure analysis was performed using DSSP[178] via the Gromacs suite.

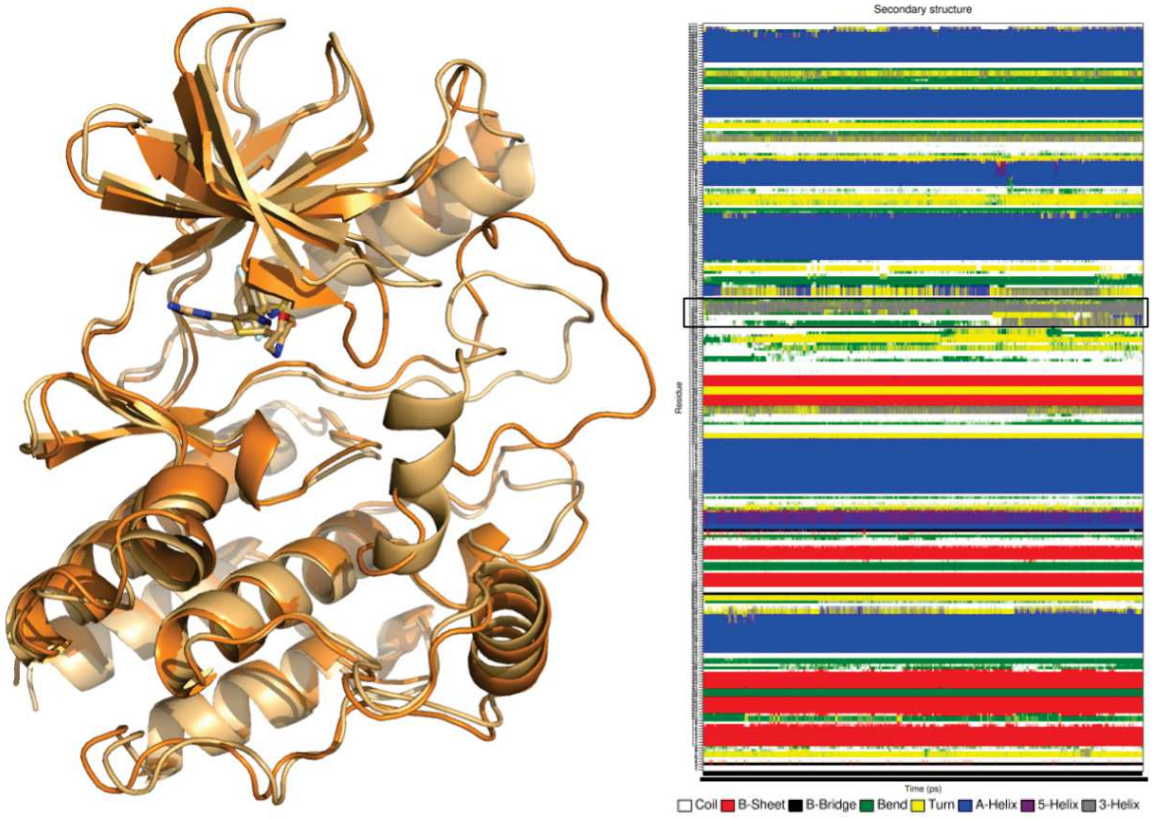


Figure 3.25: Helix formation during MD simulation. Left: The starting conformation of BRAF-V600E (orange) and the last frame after 100 ns of MD simulation (sand). Whereas the starting conformation has a completely unstructured activation loop, the last conformation shows an α-helix for residues Glu611 - Met620. Right: Secondary structure calculation from DSSP along the trajectory for the whole protein sequence, with the residues forming the helix highlighted in a black box.

3.5.2 Machine learning methods for drug design from two perspectives

Machine learning models addressing the affinity prediction and molecule prioritization issue from different angles were trained, evaluated and finally employed to provide ideas for molecular drug design candidates.

Machine learning methods for drug design from two perspectives: broad structure-based and tailored ligand-based

Melanie Schneider [1] and Gilles Labesse [1]

[1]Centre de Biochimie Structurale (CBS), CNRS, INSERM, Univ Montpellier, 34090 Montpellier, France.

ABSTRACT

Predicting the interactions between small molecules and receptors plays a critical role in drug discovery and development. Especially, when having a dedicated set of molecular data with biological activity measurements at hand, machine learning methods became an extensively used tool to exploit the best combinations for guiding new drug designs. In the present study we set up, compare and evaluate commonly used techniques for affinity prediction in drug design by taking the example of a drug design campaign based on the clinical protein kinase inhibitor dabrafenib. We shed light on affinity predictions for drug design from two perspectives: broad structure-based methods that are aimed to give information on any kind of small molecule and are trained on rather large and very diverse molecular datasets, and tailored ligand-based methods that are dedicated to predicting on a defined chemical space with a confined applicability domain. Prediction performances for six different types of fingerprints are compared and evaluated on the basis of newly designed molecule sets. Additionally, we showcase the integrated usage of different feature selection methods, resulting in improved predictions for the tailored models.

Keywords: machine learning, drug design, structure-based, ligand-based, dabrafenib, BRAF kinase

1 INTRODUCTION

Nowadays, computational techniques became crucial for medicinal chemistry and drug design. This is mainly due to development of new software, improvements in hardware, and the constantly increasing amount of available data, which is related to the rise of new experimental techniques that increased experimental throughput, such as the large availability of high-throughput screening platforms. Rational drug design, referring to the development of medications based on the study of the structures and functions of target molecules, should also largely benefit from those developments. Computational methods for filtering drug-like compounds and evaluating their binding to a given target protein can take advantage of the increased amounts of data. Generally they can be classified based on the information content they use for predictions, namely structure-based and ligand-based approaches. The structure-based approaches that are dedicated to treat large amounts of molecules (e.g. *in silico* high-throughput screening of large databases) are usually based on docking of the compounds into the protein binding site, followed by subsequent scoring. As those methods are dedicated to screen large quantities of very diverse molecules, they often go along with a trade-off in accuracy. When the set of molecules is already narrowed down to a more restricted chemical space, but still containing a large number, dedicated ligand-based approaches are often preferred.

In particular Quantitative Structure-Activity Relationship (QSAR) modelling - a ligand-based approach - is one of the major computational tools employed in drug design and medicinal chemistry in general. It involves modelling a continuous activity for quantitative prediction of the activity of previously unseen compounds. Feature selection is an important part of QSAR modelling, as the features are often large in number, with some of them having a rather low information gain for a certain prediction aim, and too many features also reduces the model's interpretability and increase the risk for overfitting [1]. Usually, a feature selection method is chosen and applied to the dataset before employing the actual learning algorithm. There have been a few studies on the combined impact of choices of feature selection method and learning method with different conclusions on the combinations that should work best [2, 3, 4, 5].

The aim of this project is to set up an integrated approach for drug refinement. With this study we want to highlight that the prediction aim and desired accuracy of a model affects the type of model that should be used, and that it becomes a highly important aspect how it is set up, requiring careful investigations on the descriptors importance. We investigate different molecular descriptor types and their ability to distinguish between different newly designed drug scaffolds that are similar to some of the training molecules, which is a standard drug design setting during lead generation and refinement. The approach presented in this study is of interest, as to our knowledge, there are no studies combining different feature selection methods to form a consensus selection, and particularly doing so across different types of fingerprints for targeted QSAR modelling.

Freely available datasets from two widely used databases, BindingDB [6] and PubChem BioAssay [7], are used to showcase two different drug design approaches (that are usually applied at different development stages): 1) the employment of a broad tool that can cope with and is developed on a large chemical diversity and is less restricted by a given applicability domain, as being based on docking results and very global descriptors, and 2) the development of a tailored QSAR approach, which is based on a confined set of molecules and different chemical fingerprint types for model training.

© .

The biochemical system under investigation is the protein kinase inhibitor dabrafenib and possible designed derivatives taking into account the primary target, the oncogenic serine/threonine kinase BRAF, and the available ligand space of this target with experimental affinity measures. Dabrafenib is a drug approved by the U.S. Food and Drug Administration (FDA) for treatment of advanced melanoma and metastatic non-small cell lung cancer with a BRAFV600E mutation [8, 9, 10]. It shows improved response rates and overall survival of BRAF-V600 mutant cancer patients, but unfortunately, resistance is rapidly acquired [11]. Thus, the desire to modify the existing drug in a way that possibly reduces the side effects, e.g. by slowing down its fast metabolism rates (half-life of $\sim$5 hours [12]), which may diminish the acquired resistances. Furthermore, dabrafenib has been shown to produce the paradoxical activation of the downstream pathway in wild type BRAF cells. While inhibiting the BRAF-V600E mutant the drug induces the opposite behaviour in wild type cells, leading to skin lesions and promoting growth and metastasis of tumor cells with RAS mutations [11, 13, 14].

2 METHODS AND RESULTS

2.1 Small molecule datasets

2.1.1 The two ligand datasets used for model generation

In this study two ligand datasets are used for model training that differ in size, molecular diversity and overall data quality.

- BindingDB [6] BRAF-V600E (2018) with annotated IC50 affinity measures - 2193 molecules
- PubChemAssay [7] AID:1257566 with annotated IC50 affinity measures (produced by one laboratory and labelled as confirmatory) - 103 molecules

Note that the BindingDB BRAF-V600E dataset contains also the PubChemAssay molecules.

2.1.2 The prediction datasets - designed candidates

Within the drug design project we aim for molecules that (1) are derivatives of the clinical drug dabrafenib, (2) show improved binding affinities, (3) are supposed to avoid the paradoxical effect through small scaffold variations, and (4) have a modification/extension of dabrafenib's tertiary butyl, as this moiety is the main access point for metabolism, where one methyl is transformed into a hydroxy, a carboxy, and then completely eliminated [15].

The different extensions attached to the scaffold are purchasable fragments that contain the reaction entity (thioamide) for addition to the scaffolds (from Enamine Chemical Supplier: 179 fragments).
As scaffold we define a pre-step in the synthesis pathway of the drug before the part to be modified (the tertiary butyl) is added to the molecule by cyclization of the 2-amino group attached to the pyrimidine. This scaffold is also modified, resulting in a selection of five different scaffolds (see Figure 1): *P02* - molecule closest related to dabrafenib available in the PDB (PDBID: 4XV3) and supposed to avoid paradoxical activation of WT-BRAF, *P02C* - P02 having a carbon atom instead of the nitrogen connecting the methyl-ethyl moiety, *P06* - the original dabrafenib scaffold, *P06F* - P06 with one fluor atom shifted from cis to trans position at the di-fluorophenyl ring, and *P06FCl* - P06F with an additional chlorine atom added in para to the central fluorophenyl ring.
The following molecules are used for visualizations:

- P02 - 179 molecules
- P02C - 179 molecules
- P06 - 179 molecules
- P06FCl - 179 molecules

The complete molecules are synthesized *in silico* with an in-house KNIME [16] workflow, which additionally removes the protecting groups BOC, phthalimide and benzyl carbamate.

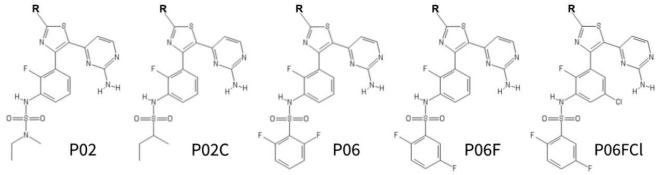


Figure 1. Drug-design scaffolds (P02, P02C, P06, P06F, and P06FCl), where **R** stands for the 179 different fragments that are used for obtaining the 179 drug candidates per scaffold.

2.2 Investigation on applicability domain

To ensure that the training datasets are eligible for prediction we first investigate on the chemical space that is covered by the three different molecular datasets. Here, the chemical space of a molecular datasets is captured by performing principal component analysis on generated PubChem fingerprints of the two training sets BindingDB BRAF-V600E and PubChemAssay AID:1257566. In order to ensure that these sets can be trustfully used to predict the new molecules, the P06FCl molecules (also represented by PubChem fingerprints) are projected onto the PCs of the two training datasets (see Figure 3 and 4). Additionally, for comparison with the P06FCl molecules, the smaller PubChemAssay training set is also projected onto the PCs of the larger BindingDB BRAF-V600E dataset (see Figure 2). Noteworthy, the PubChemAssay molecules are contained in the much larger BindingDB dataset. Indeed, the P06FCl molecules and the PubChemAssay molecules have a very similar location on the PC scatter plots spanned by the BindingDB BRAF-V600E dataset. Furthermore, all P06FCl molecules are covered by the PC scatter plot spanned by the PubChemAssay (Figure 4). Additionally, the affinity ranges covered by the two employed training sets are large and similar for both datasets, as visualized in Figure 5. Nonetheless, the distribution of the PubChemAssay is more skewed to higher affinities than the large BindingDB BRAF-V600E dataset.

2.3 Machine learning for affinity prediction

For all analysis, calculations and machine learning the R language (version 3.4.4) and RStudio are used. First, to obtain an overview of the data exploratory data analysis is performed. For the training of all machine learning algorithms in this study mainly the R package *caret* [17] (version 6.0-81) is used. In order to avoid over-fitting of the models 10-fold cross validation repeated 10 times is used for all models. Training of machine learning algorithms in regression

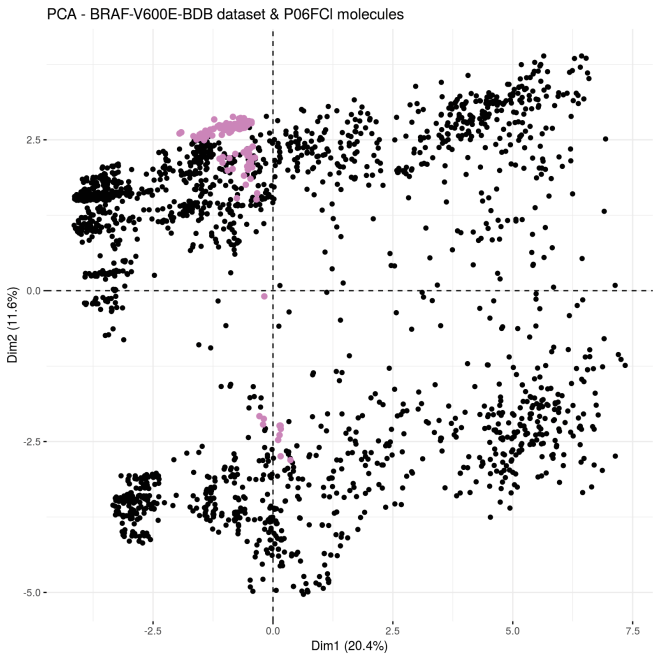


Figure 2. Scatterplot of the first two principal components (PCs) of the BRAF-V600E-BDB molecules (black dots) and the P06FCl-molecules projected onto these PCs (purple dots). The PCs are calculated based on the molecules' PubChem fingerprints.

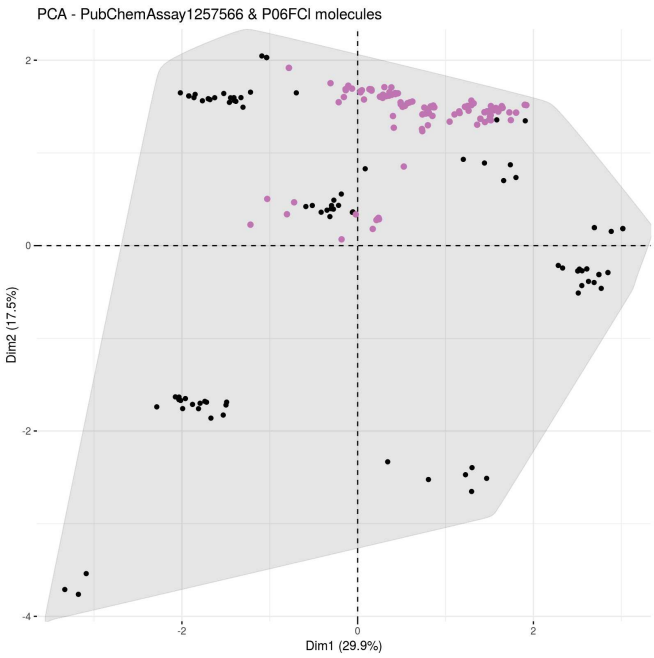


Figure 4. Scatterplot of the first two principal components (PCs) of the PubChemAssay molecules (black dots) and the P06FCl-molecules projected onto these PCs (purple dots). PCs are calculated based on the molecules' PubChem fingerprints. The PubChemAssay points are encircled using R package "ggalt" by drawing a polygon with slightly smoothened corners (s_shape=0.9) and a default expansion factor of 0.05.

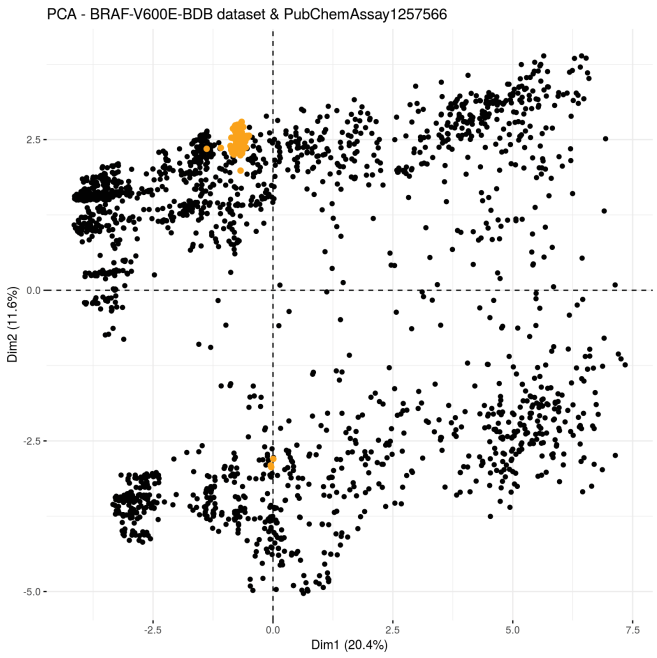


Figure 3. Scatterplot of the first two principal components (PCs) of the BRAF-V600E-BDB molecules (black dots) and the PubChemAssay molecules projected onto these PCs (orange dots). The PCs are calculated based on the molecules' PubChem fingerprints.

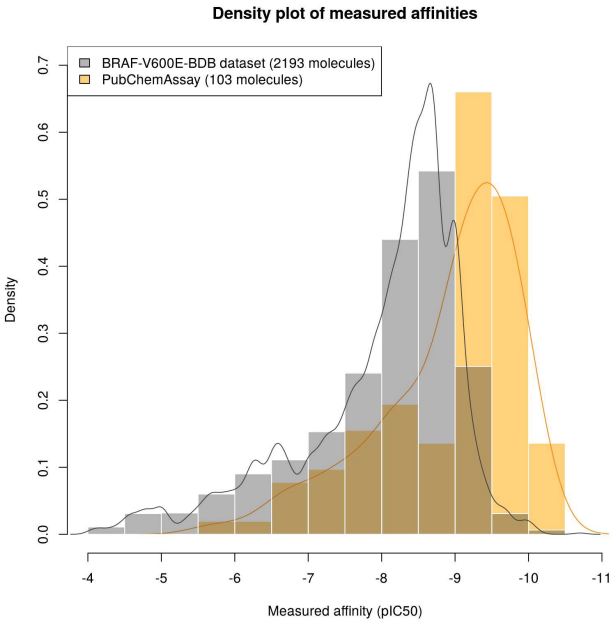


Figure 5. Density distribution of measured affinities for the two training sets BRAF-V600E-BDB and PubChemAssay.

mode was performed based on the two ligand datasets BindingDB BRAF-V600E (2018) and PubChemAssay AID:1257566.

2.4 Broad structure-based affinity prediction

In a first approach structural data from a docking campaign of the BindingDB BRAF-V600E (2018) dataset on our @TOME server [18] is employed. Here, every molecule is docked into 20 different protein structures in parallel using PLANTS [19] and the 20 cocrystallized ligands as molecular shape restraints. Subsequently, the generated complexes are evaluated by different metrics on the server. The whole procedure is repeated for the same set of molecules with different initial 3D conformations and different partial charges. Finally, global ligand-based molecular descriptors (e.g. molecular weight, number of rotatable bonds, XLogP, etc.) are added and engineered. This forms the final dataset that is used for the broad structure-based affinity prediction. The complete procedure is explained in details in a previous publication [20].

2.5 Tailored ligand-based affinity prediction

PubChemAssay AID:1257566 entitled "Raf/Mek amplified luminescence proximity homogeneous assay" is used to set up the tailored ligand-based prediction models. As the assay is labelled as "confirmatory" it represents result from multiple concentration test and we expect the results to be reliable and coherent. The tailored approach follows a rather traditional QSAR approach, as it is based on different types of fingerprints (FPs) from the CDK, computed through the *rcdk* package [21] on the PubChemAssay dataset composed of 103 molecules. The used FP types are:

- MACCS - 166 bit MACCS keys described by MDL [22]
- PubChem - 881 bit FPs defined by PubChem
- extended - hashed FPs, with a default length of 1024 bits and default search depth of 6, considers paths of a given length and takes rings and atomic properties into account
- graph - hashed FPs, with a default length of 1024 bits and default search depth of 6, considers connectivity
- shortestpath - hashed FPs, with a default length of 1024 bits and default search depth of 6, based on the shortest paths between pairs of atoms and takes into account ring systems, charges etc.
- circular - implementation of the ECFP6 fingerprint, with a length of 1024 bits and default search depth of 6 [23]

Choices for initial bit length and search depth are made based on recommendations provided by ChemAxon (`https://docs.chemaxon.com/display/docs/Chemical+Fingerprints`). For all six FP types bits with zero variance across the 103 molecules are removed, which leads to reduced FP bit sizes of 51 for MACCS, 141 for PubChem, 238 for extended, 156 for graph, 691 for shortestpath, and 428 for circular FPs.

In order to exploit the FP data in an exhaustive way, two different, widely used algorithms, Support Vector Machine (SVM) with a radial kernel, and Random Forest (RF) are employed on all six FP sets individually.

2.5.1 Investigation on variable importance

During model training optimal parameters were selected by caret's automatic grid search with 10 values per parameter (*tuneLength=10*). In the case of RF variable importance is tracked during training as the mean decrease in node impurity (see Figure 6).

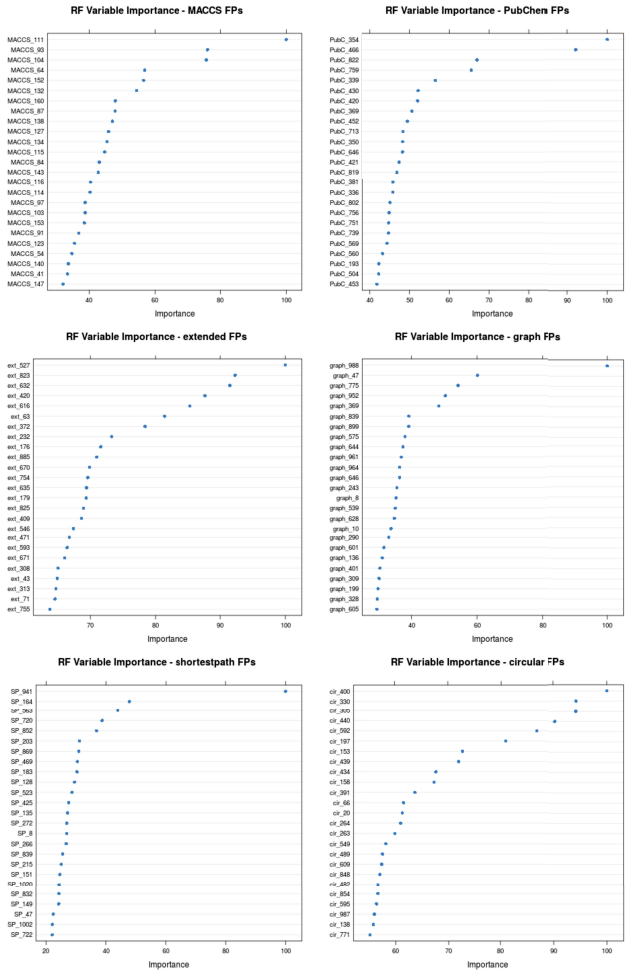


Figure 6. Variable importance tracked by RF for the six FP types MACCS, PubChem, extended, graph, shortestpath, and circular.

Moreover, quantile Random Forest (qFR) [24] is employed on all six FP types with variable importance tracking (see Figure S1). qFR is a generalisation of random forest. It gives a non-parametric way of estimating conditional quantiles for high-dimensional predictor variables. The trained qFR models show lower cross-validation performances than the RF models. Therefore qRF is only used for variable importance confirmation, not for affinity prediction. Aditionally, two further algorithms, Multivariate Adaptive Regression Splines (MARS) [25] and Boruta, an all relevant feature selection wrapper algorithm [26], are employed to identify important variables for all six FP types (see Figure S2 and S3, respectively). MARS is a non-parametric regression technique that automatically models nonlinearities and interactions between variables. It can be seen as an extension of linear models. Boruta iteratively compares importances of attributes with importances of shadow attributes, created by shuffling original ones. It does a sharp classification of features rather than ordering. Being an all relevant method, it aims to find all features connected with the decision and therefore, it also includes redundant features. (By default the ranger package Random Forest implementation is used.)

2.5.2 Tailored FP selection for model training

In analogy to the problem-solving principle of "Occam's Razor" we seek a model with the smallest number of descriptors that yield a reasonable model.

Table 1. Selected FPs for every FP type

FP type	Selected FPs
MACCS	111, 93, 104
PubChem	354, 466, 822, 759, 339
extended	527, 823, 632, 420, 616, 63, 372
graph	988, 47, 775, 952, 369, 839, 899
shortestpath	941, 164, 563, 720, 722, 825, 135
circular	400, 330, 305, 440, 592, 197, 153, 439, 434, 158, 391

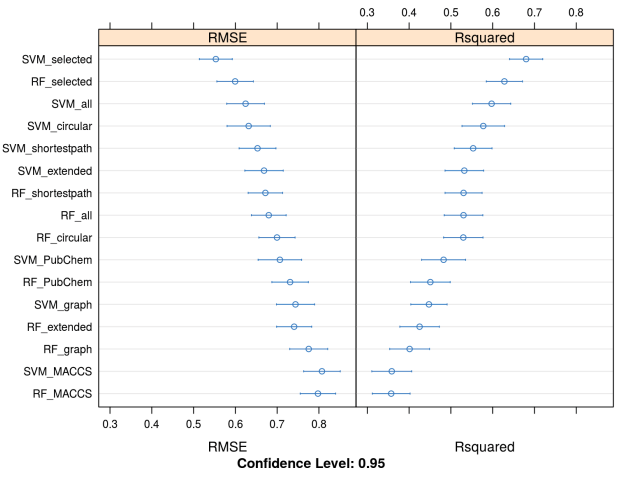


Figure 7. Performance comparison of all trained models on internal cross-validation of training set.

Often, a small number of descriptors affords a model that outperforms more complex ones. Therefore, the results of the four algorithms employed for investigating variable importance are compared and the most distinct features, the most highly ranked by the algorithms are extracted for every FP type (see Table 1) and subsequently combined as custom 'selection' comprising 40 FPs.

On this selection new SVM and RF models are trained. For comparison, SVM and RF models are also trained on the combination of all generated FPs (not including the ones with zero variance). The 'selected' models and 'all' models are compared with the previously trained models on single FP types by performance on cross-validation during training (see Figure 7). Remarkably, the 'selected' SVM and RF models show the best performance, with lowest RMSE and highest R^2 values. They are followed by the SVM_all model, whereas the RF_all model is situated much lower, in the midfield among all trained SVM and RF models. In general, the SVM algorithm shows slightly improved performances over RF on the same FPs.

2.6 Affinity prediction for designed molecules

The broad structure-based prediction model trained on the BindingDB BRAF-V600E (2018) dataset, named 'BDB-IC50', and the two best performing models from the tailored ligand-based approach, 'SVM_selected' and 'RF_selected' are used for affinity prediction of the newly designed molecule sets (named by their molecular scaffold) P02, P02C, P06, and P06FCl (see Figure 8, 9, and 10).

The 'SVM_selected' and 'RF_selected' predictions show clear distribution shifts between the four molecular scaffolds with a ranking of P02 - P02C - P06 - P06FCl, from lowest to highest affinity. This separation is not visible for the predictions of the 'BDB-IC50' model. Only the P02 set is predicted

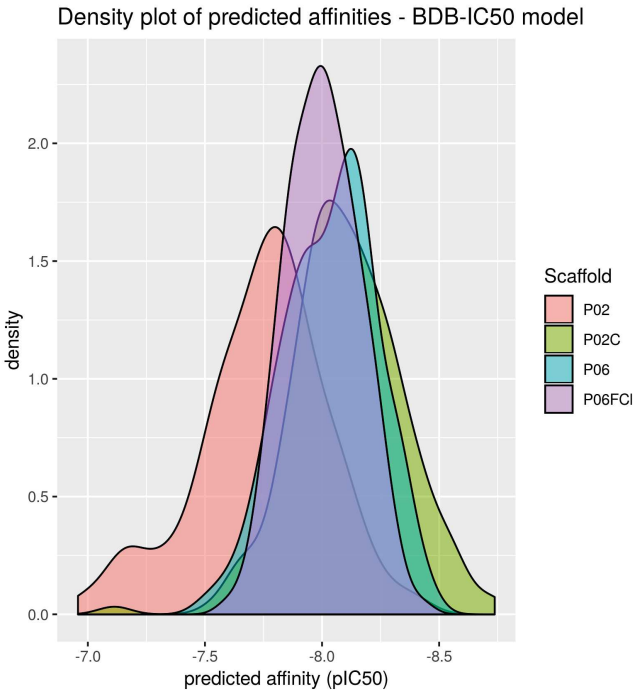


Figure 8. Affinity predictions of 'BDB-IC50' model for four molecule sets differing in their 'scaffold' (P02, P02C, P06, and P06FCl) and containing 179 compounds each.

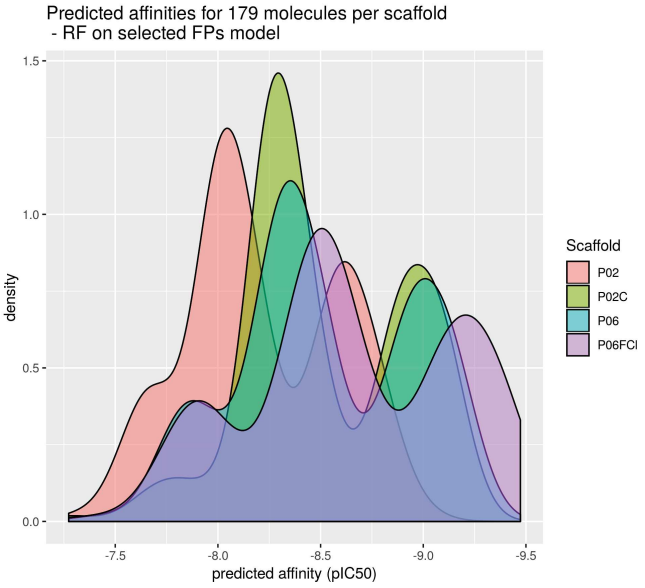


Figure 9. Affinity predictions of 'RF_selected' model for four molecule sets differing in their 'scaffold' (P02, P02C, P06, and P06FCl) and containing 179 compounds each.

with overall lower affinities. The 'RF_selected' predictions have the most pronounced discrimination between the scaffold sets and cover the largest affinity range (7.2 to 9.5), whereas the 'SVM_selected' predictions cover a smaller range (7.8 to 8.6). Molecules with the best predictions from the

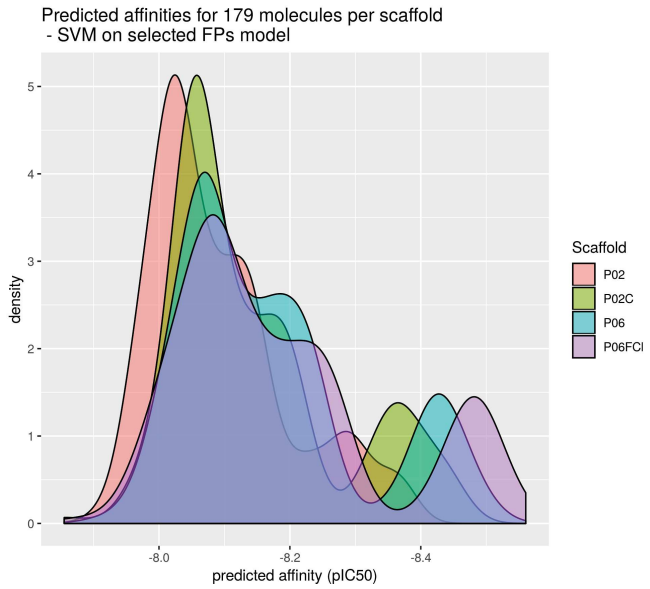


Figure 10. Affinity predictions of 'SVM_selected' model for four molecule sets differing in their 'scaffold' (P02, P02C, P06, and P06FCl) and containing 179 compounds each.

'SVM_selected' and 'RF_selected' model with scaffold P06 and P06FCl are depicted in Figures S4, S5, S6 and S7. These sets of molecules served as inspiration basis for subsequent drug design rounds including the synthesis and experimental testing of selected compounds.

The six RF models trained on different single FP types (MACCS, PubChem, extended, graph, shortestpath, and circular) show different performances with respect to discrimination capacities between the scaffold sets (see Figure 11). Four models (MACCS, PubChem, extended, and graph) are identifying scaffold P02 as the worst binding, whereas the MACCS, PubChem, and extended models also agree on the following scaffold ranking: P02 - P02C - P06 - P06FCl (from worst to best). The graph FP based model sets apart the P02 scaffold as worst binding and identifies P06FCl as best scaffold, but does not distinguish between P02C and P06. The shortestpath FP based model does not discriminate between the four scaffolds at all, in contrast to the circular PF based model that clearly separates all four scaffolds, but in a different order: P02C - P02 - P06FCl - P06 (from worst to best). Interestingly, the models developed on all FPs are showing the scaffold ranking P02 - P02C - P06 - P06FCl (from worst to best), but are not able to discriminate very well between the different scaffolds (see Figure 12), whereas the SVM model shows a slightly better performance, by clearly setting apart the P02 scaffold.

2.6.1 Experimental testing of synthesized molecules

One molecule was synthesized for the P02C scaffold, followed by 12 molecules with the P06F scaffold and finally, two further molecules containing the P06FCL scaffold. Experimental testing on BRAF confirmed the lower activity of P02C (85% inhibition at 1000 nM). The P06 molecules were effective in the low nanomolar range (4-6 nM), while the two compounds with the PO6FCl scaffold were active at 2 nM.

3 DISCUSSION AND CONCLUSION

In this project, we aim to understand the effect of different machine learning approaches with respect to the composition of the molecular training set and variations in the nature of employed features within a drug design pipeline.

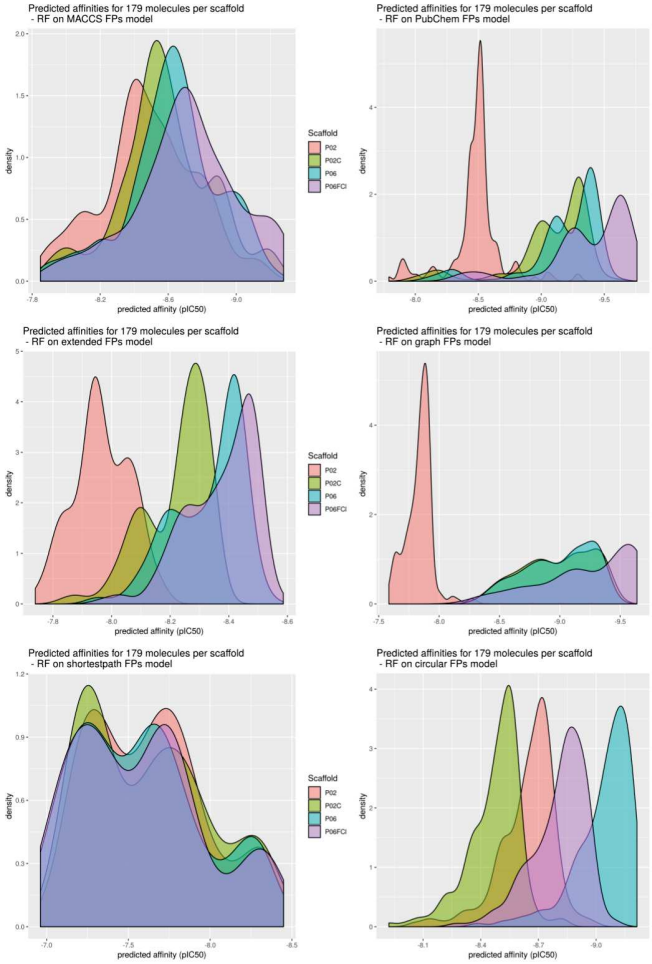


Figure 11. Affinity predictions of six RF models trained on different FP types (MACCS, PubChem, extended, graph, shortestpath, and circular FPs, from upper left to bottom right) for four molecule sets differing in their 'scaffold' (P02, P02C, P06, and P06FCl) and containing 179 compounds each.

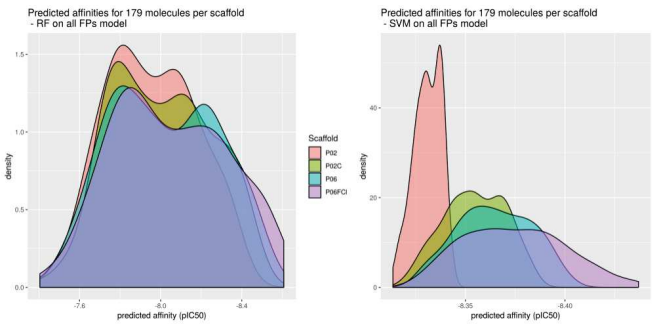


Figure 12. Affinity predictions of 'RF_all' and 'SVM_all' model for four molecule sets differing in their 'scaffold' (P02, P02C, P06, and P06FCl) and containing 179 compounds each.

We make sure that the training datasets are adequate for model development by evaluating the applicability domain based on the molecules' chemical space.

The importance of feature selection for dedicated drug design is highlighted by the improved results from models with particularly selected features. Here we present an approach that employs different feature selection algorithms and methods and combines the results into a consensus selection weighted by the performance of previously trained models. This combination of automated identification of informative features with a balanced evaluation of the models' importance weighting (based on previous performance results) is performed across different types of fingerprints and results into improved prediction performances during cross-validation. Moreover, with the trained models we succeed to distinguish between our designed scaffolds, among which selected molecules were synthesized and their affinities tested *in vitro* on BRAF-V600E.

REFERENCES

[1]Mohammad Goodarzi, Bieke Dejaegher, and Yvan Vander Heyden. Feature selection methods in QSAR studies. *Journal of AOAC International*, 95(3):636–651, June 2012.

[2]Mohammad Goodarzi, Matheus P. Freitas, and Richard Jensen. Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3beta inhibitory activities. *Journal of Chemical Information and Modeling*, 49(4):824–832, April 2009.

[3]Ying Liu. A comparative study on feature selection methods for drug discovery. *Journal of Chemical Information and Computer Sciences*, 44(5):1823–1828, October 2004.

[4]Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. Benchmarking Variable Selection in QSAR. *Molecular Informatics*, 31(2):173–179, February 2012.

[5]Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. Choosing feature selection and learning algorithms in QSAR. *Journal of Chemical Information and Modeling*, 54(3):837–843, March 2014.

[6]Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N. Jorissen, and Michael K. Gilson. BindingDB: a web-accessible database of experimentally determined proteinligand binding affinities. *Nucleic Acids Research*, 35(Database issue):D198–D201, January 2007. 00773.

[7]Yanli Wang, Stephen H. Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A. Shoemaker, Paul A. Thiessen, Siqian He, and Jian Zhang. PubChem BioAssay: 2017 update. *Nucleic Acids Research*, 45(D1):D955–D963, 2017.

[8]Anita D. Ballantyne and Karly P. Garnock-Jones. Dabrafenib: first global approval. *Drugs*, 73(12):1367–1376, August 2013.

[9]Alexander M. Menzies and Georgina V. Long. Dabrafenib and trametinib, alone and in combination for BRAF-mutant metastatic melanoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 20(8):2035–2043, April 2014.

[10]Lauretta Odogwu, Luckson Mathieu, Gideon Blumenthal, Erin Larkins, Kirsten B. Goldberg, Norma Griffin, Karen Bijwaard, Eunice Y. Lee, Reena Philip, Xiaoping Jiang, Lisa Rodriguez, Amy E. McKee, Patricia Keegan, and Richard Pazdur. FDA Approval Summary: Dabrafenib and Trametinib for the Treatment of Metastatic Non-Small Cell Lung Cancers Harboring BRAF V600e Mutations. *The Oncologist*, 23(6):740–745, June 2018.

[11]Weijiang Zhang. BRAF inhibitors: the current and the future. *Current Opinion in Pharmacology*, 23:68–73, August 2015.

[12]Cathrine L. Denton, Elisabeth Minthorn, Stanley W. Carson, Graeme C. Young, Lauren E. Richards-Peterson, Jeffrey Botbyl, Chao Han, Royce A. Morrison, Samuel C. Blackman, and Daniele Ouellet. Concomitant oral and intravenous pharmacokinetics of dabrafenib, a BRAF inhibitor, in patients with BRAF V600 mutation-positive solid tumors. *Journal of Clinical Pharmacology*, 53(9):955–961, September 2013.

[13]Bogos Agianian and Evripidis Gavathiotis. Current Insights of BRAF Inhibitors in Cancer. *Journal of Medicinal Chemistry*, 61(14):5775–5793, 2018.

[14]Robert Roskoski. Targeting oncogenic Raf protein-serine/threonine kinases in human cancers. *Pharmacological Research*, 135:239–258, 2018.

[15]Alicja Puszkiel, Galle No, Audrey Bellesoeur, Nora Kramkimel, Marie-Nolle Paludetto, Audrey Thomas-Schoemann, Michel Vidal, Franois Goldwasser, Etienne Chatelut, and Benoit Blanchet. Clinical Pharmacokinetics and Pharmacodynamics of Dabrafenib. *Clinical Pharmacokinetics*, 58(4):451–467, April 2019.

[16]Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Ktter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 319–326. Springer Berlin Heidelberg, 2008.

[17]Max Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(1):1–26, November 2008.

[18]Jean-Luc Pons and Gilles Labesse. @TOME-2: a new pipeline for comparative modeling of proteinligand complexes. *Nucleic Acids Research*, 37(suppl_2):W485–W491, July 2009.

[19]Oliver Korb, Thomas Sttzle, and Thomas E. Exner. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In Marco Dorigo, Luca Maria Gambardella, Mauro Birattari, Alcherio Martinoli, Riccardo Poli, and Thomas Sttzle, editors, *Ant Colony Optimization and Swarm Intelligence*, Lecture Notes in Computer Science, pages 247–258. Springer Berlin Heidelberg, 2006.

[20]Melanie Schneider, Jean-Luc Pons, William Bourguet, and Gilles Labesse. Towards accurate high-throughput ligand affinity prediction by exploiting structural ensembles, docking metrics and ligand similarity. *Bioinformatics (Oxford, England)*, July 2019.

[21]Rajarshi Guha. Chemical Informatics Functionality in R. *Journal of Statistical Software*, 18(1):1–16, January 2007.

[22]Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, December 2002.

[23]David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010.

[24]Nicolai Meinshausen. Quantile Regression Forests. *J. Mach. Learn. Res.*, 7:983–999, December 2006.

[25]Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67, March 1991.

[26]Miron B. Kursa and Witold R. Rudnicki. Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(1):1–13, September 2010.

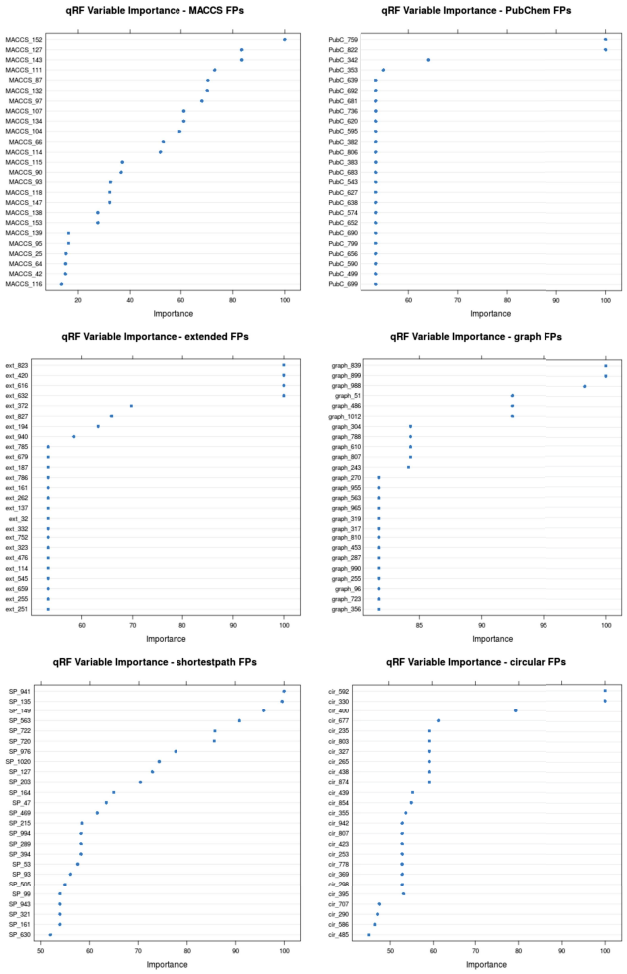


Figure S1. Variable importance tracked by qRF for the six FP types MACCS, PubChem, extended, graph, shortestpath, and circular.

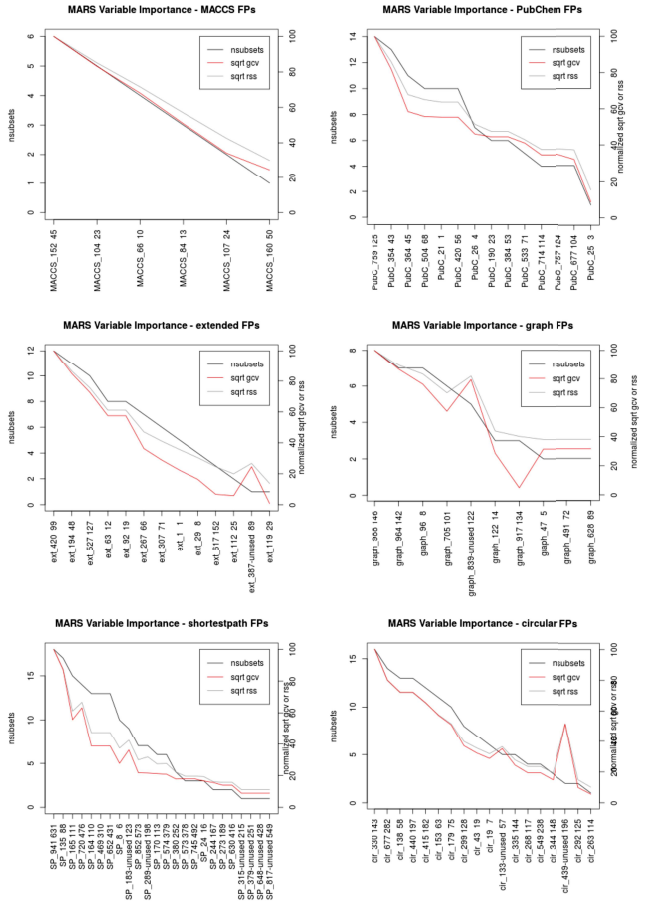


Figure S2. Variable importance tracked by MARS for the six FP types MACCS, PubChem, extended, graph, shortestpath, and circular.

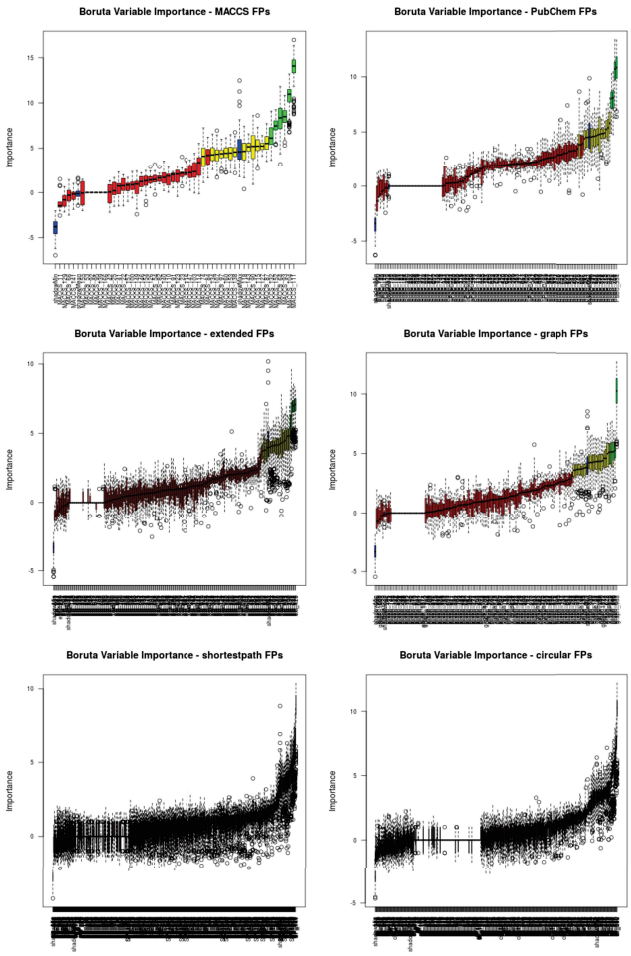


Figure S3. Variable importance tracked by Boruta for the six FP types MACCS, PubChem, extended, graph, shortestpath, and circular.

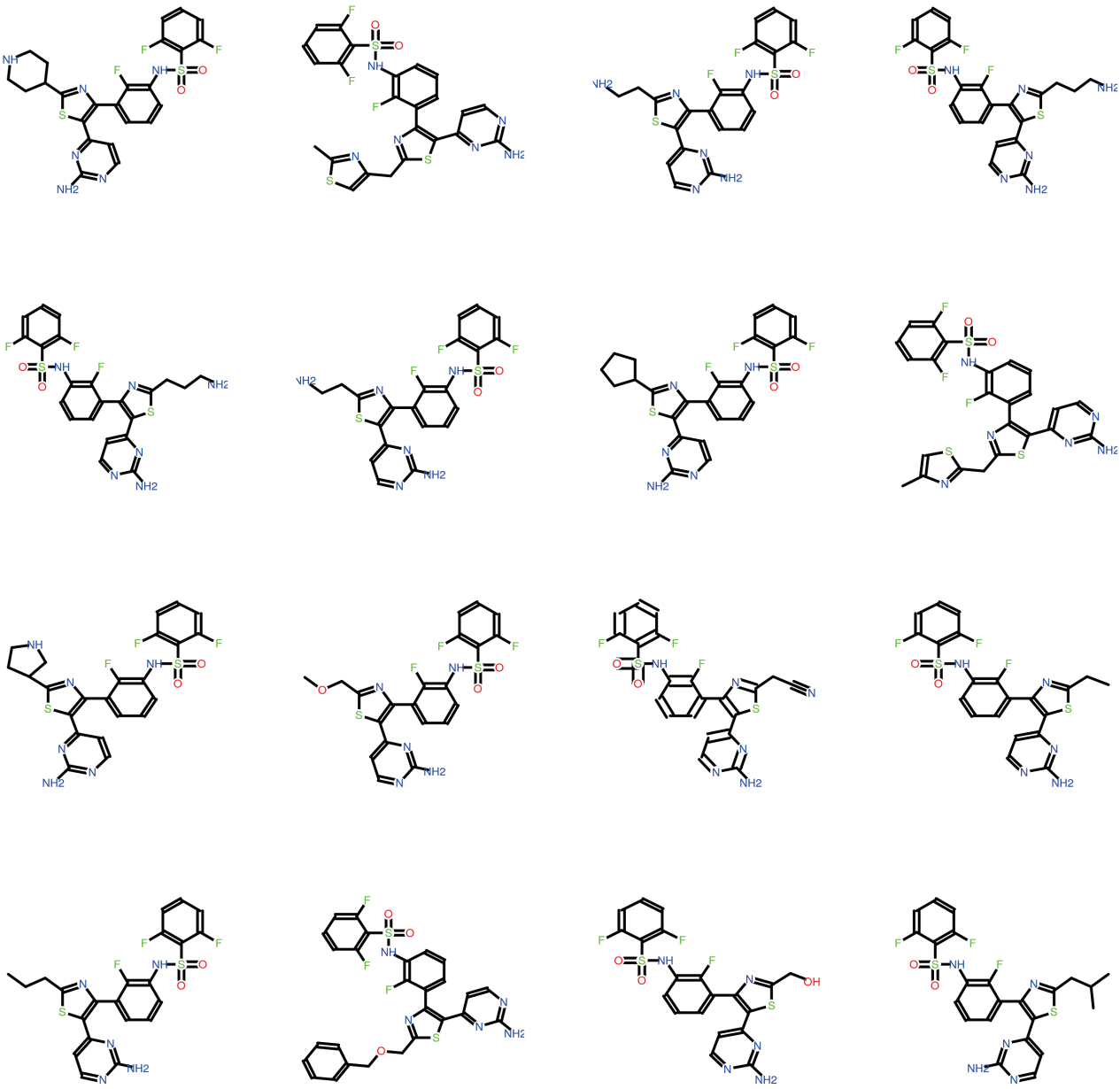


Figure S4. The 16 P06-scaffold molecules with best affinity predictions by the 'RF_selected' model.

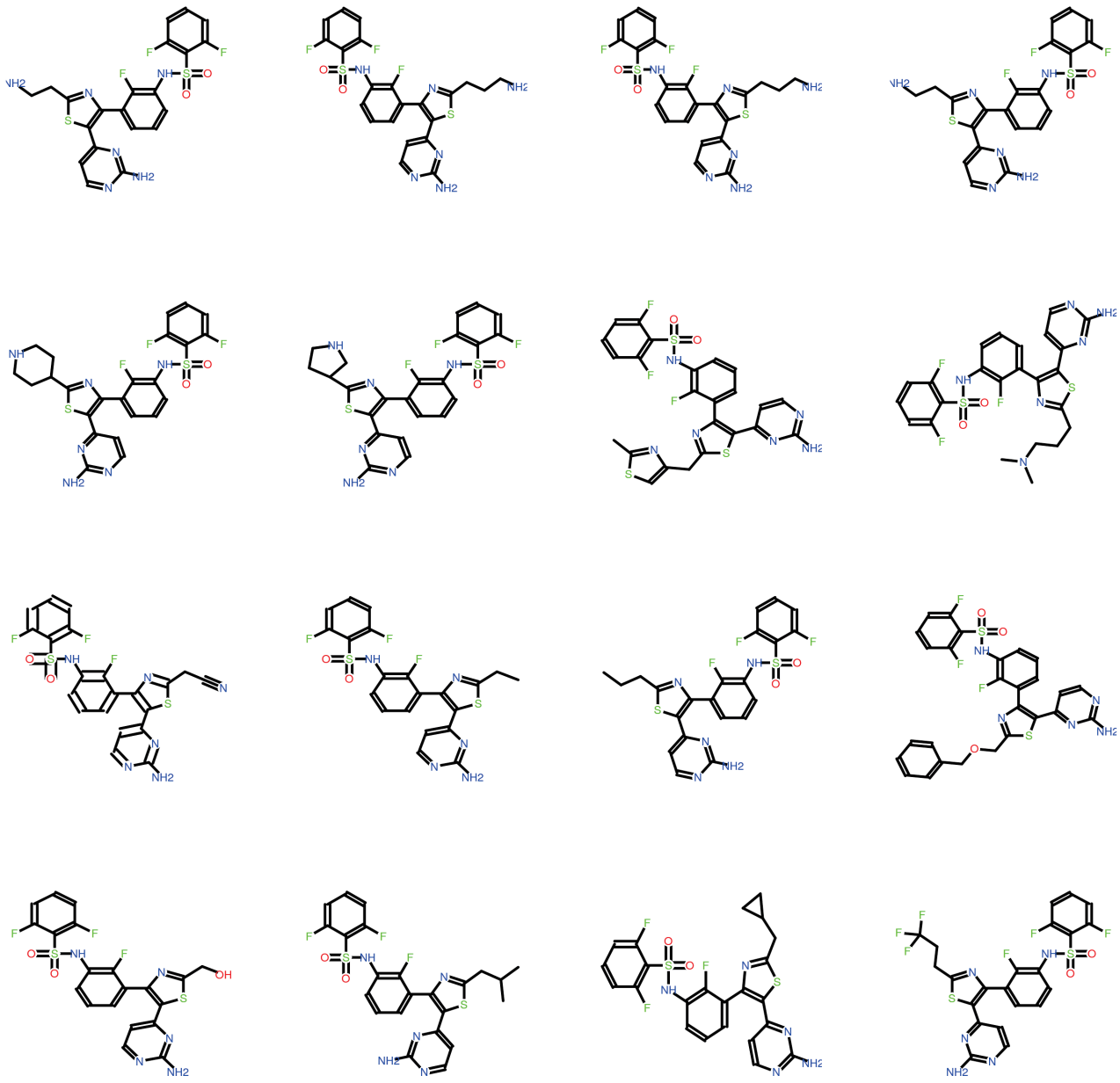


Figure S5. The 16 P06-scaffold molecules with best affinity predictions by the 'SVM_selected' model.

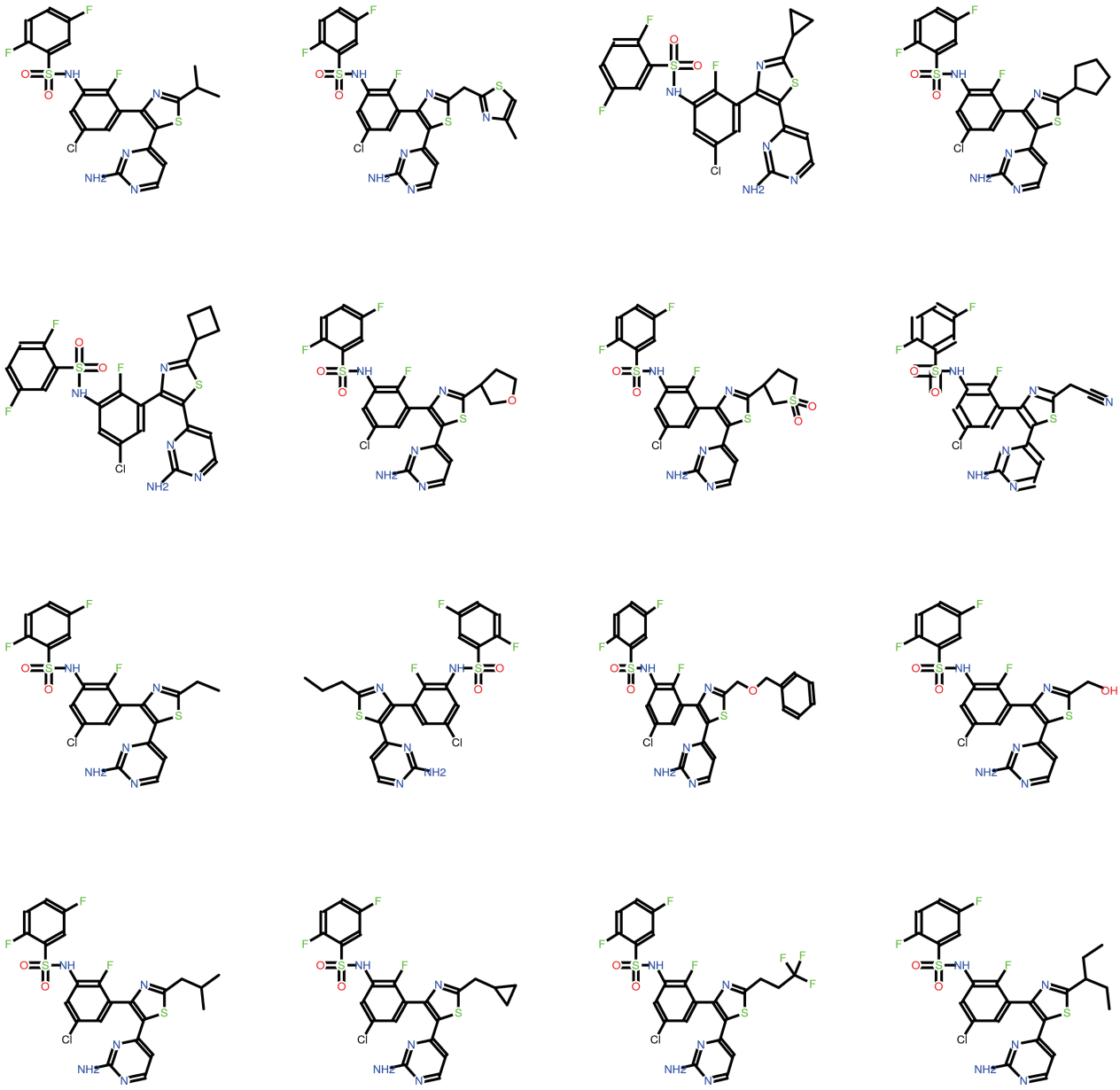


Figure S6. The 16 P06FCl-scaffold molecules with best affinity predictions by the 'RF_selected' model.

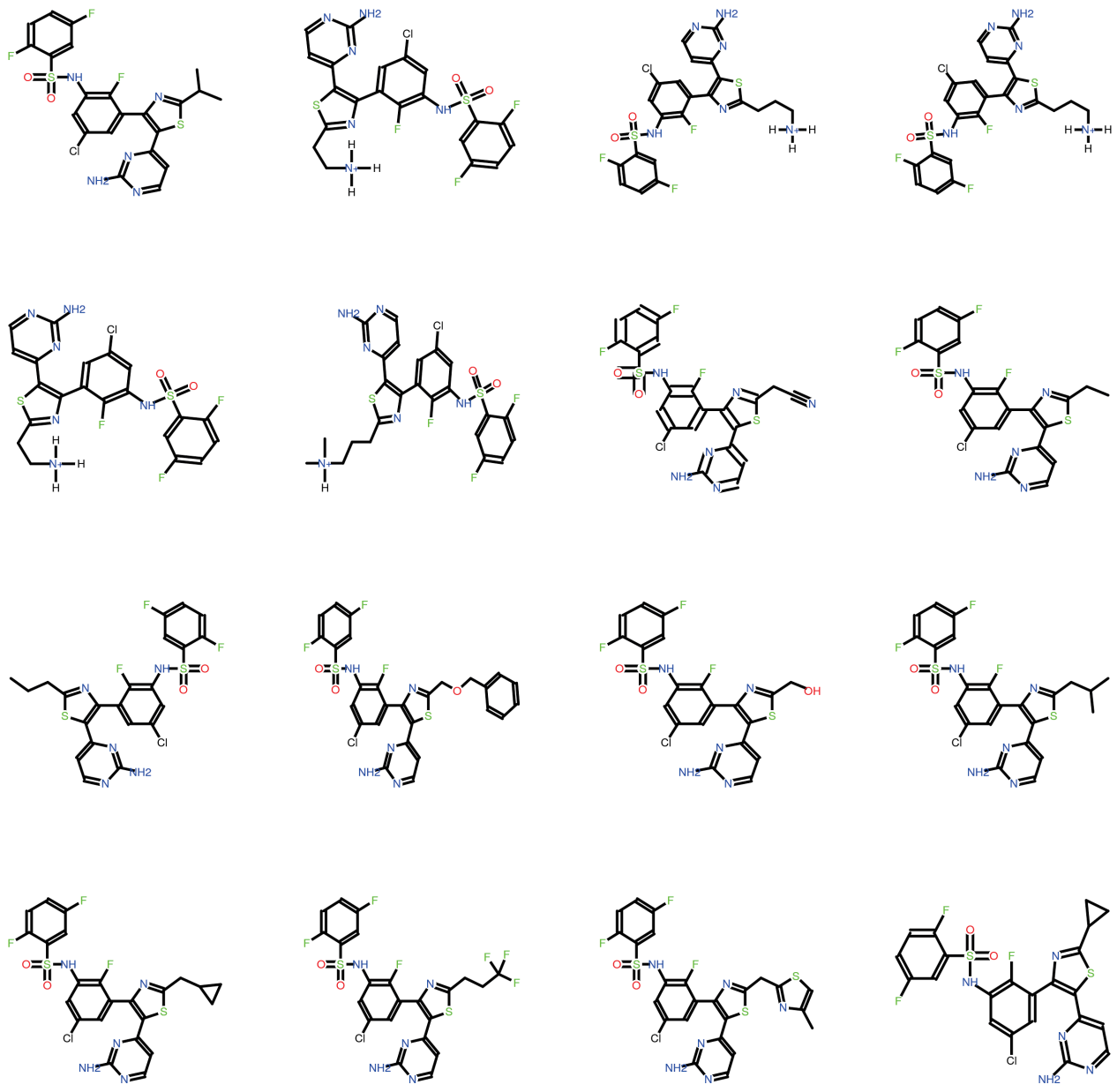


Figure S7. The 16 P06FCl-scaffold molecules with best affinity predictions by the 'SVM_selected' model.

3.6 Drug design synthesis rounds - a chronological overview

Round 1

Drug synthesis round 1 is based on the scaffold P02C (see Figure 3.26), as the publication of P02 suggests its capacity of avoiding the paradoxical effect, and additionally, better affinities are predicted for molecules with a carbon atom instead of P02's original nitrogen. The selection of the attachment fragments forming the molecular extensions is based first on a pre-selection from the RF machine learning approach on BRAF (best predicted ∼20 candidates) and then further refined according to a size criteria (the extension should be larger than the original tertiary butyl moiety) and favorable pharmacokinetics properties of thioamides (regarding cell penetration and solubility), resulting in four candidates.

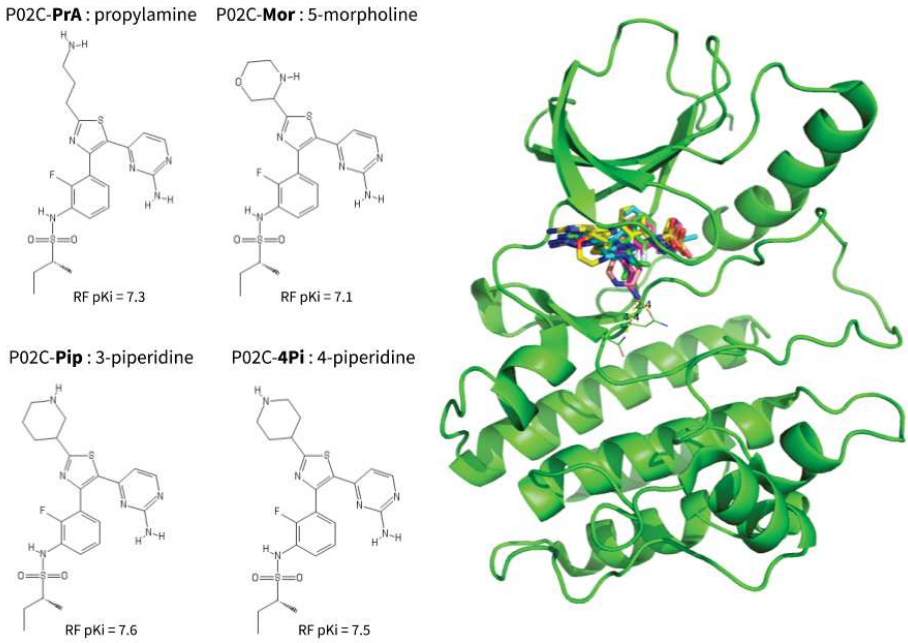


Figure 3.26: The selected molecules for drug synthesis round 1 - scaffold P02C, their predicted affinity by the RF model (trained on the BindingDB dataset), and their docking pose within the BRAF structure.

Round 2

As drug synthesis round 1 showed decreased affinities for BRAF compared to dabrafenib (P06), the P02C scaffold is not further continued. Additionally, machine learning models and MM-PBSA computations predict improved affinities for the P06F scaffold, which is basically P06 with one of the fluor atom shifted from cis to trans position at the di-fluorophenyl ring. Therefore, this scaffold is exemplified with four different extensions as replacement of the tertiary butyl moiety: the extension 1-cyclopropylpiperidine (CPP), which is present in PDB molecule CQE, 3-piperidine (Pip), 5-morpholine (Mor), and propylamine (PrA). On top of that, the pyrimidine moiety is extended by an additional acetyl (-ac) for the four molecules, resulting in a total of eight compounds (see Figure 3.27).

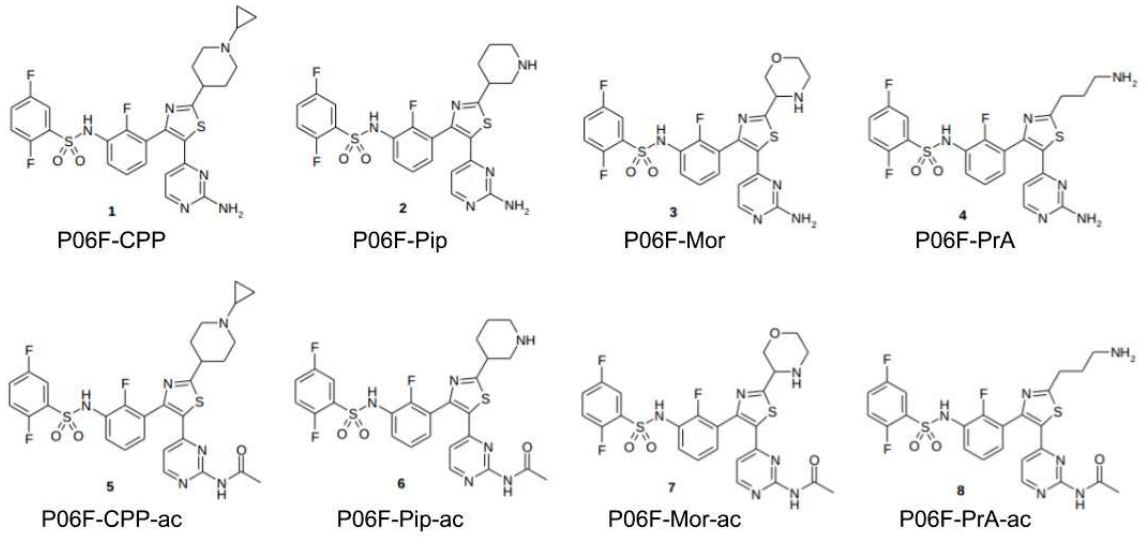


Figure 3.27: Molecules of drug synthesis round 2 - scaffold P06F.

Round 3

To investigate the effect of the different scaffolds drug synthesis round 3 comprises the four modified scaffolds based on P06 (P06F, P06FF, P06FCl, and P06FFCl, where the amount of Fs indicate the number of shifted fluor atoms, and Cl indicates an additional chlorine atom, compared to P06) with the original tertiary butyl extension and the morpholino (Mor) extension, resulting again in eight compounds (see Figure 3.28).

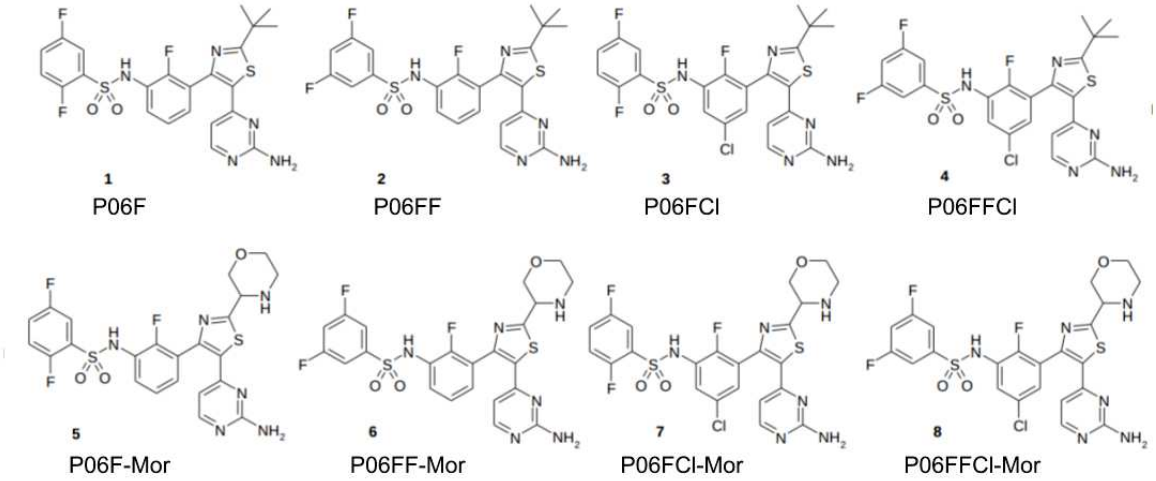


Figure 3.28: Molecules of drug synthesis round 3 - includes the four modified scaffolds based on P06 (P06F, P06FF, P06FCl, and P06FFCl) with the original tertiary butyl extension and the morpholino (Mor) extension.

Round 4

The designed molecules for synthesis round 4 were inspired by the newly solved structure of BRAF with P06F-Mor and the Mor-extension's proximity to polar residues. Four molecules were built by adding new substituents - azetidine (2Az), pyrrolidine (2Py), piperidine (2Pi), piperazine (2PA) - instead of the morpholine group to the P06F scaffold, resulting in a total of four compounds

(see Figure 3.29). Round 4 was aimed to test the interaction between BRAF's G-rich loop and the extensions and attempted to form a hydrogen bond between the extensions and the Asp of the DFG motif.

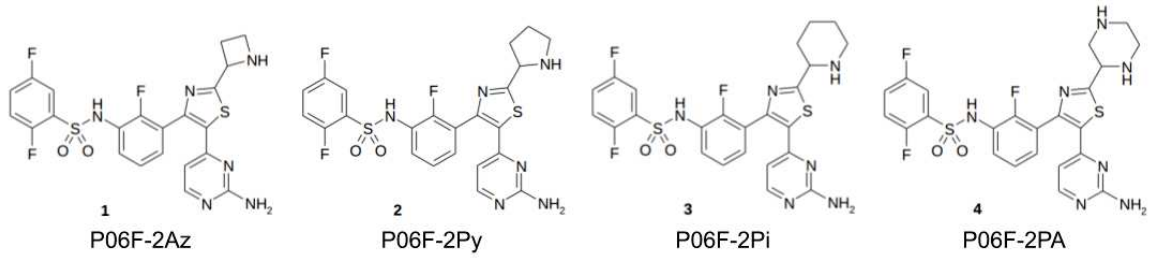


Figure 3.29: Molecules of drug synthesis round 4 - scaffold P06F.

As the synthesis of all proposed molecules would require increased expenses a thorough evaluation by machine learning, molecular dynamics and MM-PBSA affinity calculation was performed.

3.7 MM-PBSA affinity calculations for designed molecules

MM-PBSA is a widely used technique to estimate binding affinities based on structural conformations. Whereas the conformational ensembles used for MM-PBSA calculations are usually generated by molecular dynamics simulations, the technique is applicable to differently generated structures.

MM-PBSA calculation - methods

For MM-PBSA calculation the 50-ns MD production trajectories were reduced to 501 frames each, by extracting a frame every 100 ps. The resulting snapshots of the MD simulations were utilized for post-processing free energies by the single-trajectory MM-PBSA method implemented in g_mmpbsa. Different dielectric constants (ε=(2,4,6,8,12,20)) were tested for the binding pocket, while the solution dielectric constant was kept constant at ε=80. Calculations are performed based on a homogeneous medium with a range of dielectric constants for the solute, an ionic strength of 150 mM, an ionic radius of 0.95 Å for positive charged ions and 1.81 Å for negative charged ions, and a solvent probe radius of 1.4 Å. An example configuration file for g_mmpbsa is provided within the supplements of Section 3.7.1. Other parameters influencing the grid dimensions of the calculation, such as 'cfac', 'gridspace' and 'fadd' were varied from suggested defaults (1.5, 0.5 and 10, respectively) showing only marginal variations in the results and therefore not further changed.

Analysis and visualization is performed with Gromacs tools, PyMol, VMD, Chimera, Python scripts, and provided scripts from the g_mmpbsa package.

3.7.1 MM-PBSA approaches with dabrafenib and its metabolites

The widely used MM-PBSA affinity calculation approach (based on structural snapshots from MD simulations) was evaluated on the drug dabrafenib and its known major metabolites, whereas the importance of the solute dielectric constant became apparent as major concern with respect to ranking of newly designed compounds for protein kinase drug design.

MM-PBSA and the importance of the dielectric constant for kinase drug design

Melanie Schneider [1] and Gilles Labesse [1]

[1]Centre de Biochimie Structurale (CBS), CNRS, INSERM, Univ Montpellier, 34090 Montpellier, France.

ABSTRACT

Predicting the interactions between a set of small molecules and its target plays a critical role in drug discovery and development. Especially in later stages of the drug design process, when a reduced set of molecules is in focus, reliable and accurate binding affinity estimations are important for targeted modifications of given lead molecules. Current limitations in affinity prediction originate from the lack of accurate estimates for solvation energy and entropy. MM-PBSA and the related MM-GBSA aim at providing better estimates. From our studies we infer that the common approach using one dielectric constant for the binding pocket may be misleading (here in the case of a kinase), especially when designed ligands/drugs contain charges. Thus, a range of selected values for the solute dielectric constant is preferred for better and more reliable comparisons.

Keywords: MM-PBSA, drug design, kinase, B-RAF, dielectric constant

1 INTRODUCTION

While a relative and approximate ranking of the stability of different complexes might be sufficient for an initial screening protocol, a finer and more accurate evaluation of the binding free energy may be necessary for a fine tuning in later stages of drug design. Free energy estimates require simulation of complex flexibility and desolvation upon binding of both partners in order to deduce the entropy term instead of a simple extrapolation of the enthalpy term with a very rough and partial prediction of the entropy part as in usual and quick affinity prediction methods. For the most accurate methods, very long simulations are required and limit their use. If accurate energies are needed, the methods of choice are sophisticated MD-based calculations, such as thermodynamic integration (TI) [1, 2] and free energy perturbation (FEP) [3]. Since they are computationally very expensive, extremely time-consuming and exhaustive conformational and statistical sampling is needed to obtain converged results, they are not widely used in structure-based drug design. Among the approximate methods, there are the linear interaction energy (LIE) [4], the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) [5] and the related molecular mechanics generalized Born surface area (MM-GBSA) methods [6]. LIE is a semi-empirical method, based on the assumption that the binding free energy between the ligand and the receptor can be modelled as a linear-response combining weighted electrostatic and Van der Waals interactions with coefficients varying for different systems [4, 7]. Unlike the LIE method, MM-PBSA and MM-GBSA do not employ empirical parameters within their calculations, which makes them promising methods for ranking very different compounds. They both use molecular mechanics force fields with continuum solvent models. The GB equation is simply an approximation of the PB equation [8], resulting in an increased calculation speed (about 5 times faster), but often goes along with an accuracy trade-off [9, 10, 11]. They are both frequently used in structure-based drug design due to their rather high accuracy and relative high computational efficiency. Another advantage is that they have no varying parameters for different protein-ligand systems while using sets of physically well-defined energy terms and they do not require training set calculations.

It has been previously reported that using MM-PBSA long MD simulations seem not to result in better predictions and short MD simulations can be adequate in calculating binding affinities [12, 13]. In order to achieve a higher precision it has been suggested to run many short independent simulations (produced by e.g. replicate sampling) instead of a single long one, which should avoid underestimation of the uncertainty [14]. Additionally, if one is only interested in the relative order of binding affinities, the ranking of compounds with similar structures and binding modes, the entropy contribution to the binding free energy can be omitted, which is often recommended as it reduces the computational cost and avoids adding an additional non-negligible error margin. It has been found that MM-PB/GBSA performances generally vary with the tested system and also depend on the used force fieldand the solute dielectric constant [12, 15, 8].

Here, we wanted to reassess the use of MM-PBSA for fine ranking of a drug, dabrafenib, and its known metabolites in aim at predicting a potential impact of its pharmacokinetics, its metabolization on its efficacy on its primary target the protein-kinase BRAF.

Dabrafenib [16] is a BRAF kinase inhibitor, which inhibits BRAF V600 mutation-positive cancer cell growth. It is an FDA approved drug indicated for the treatment of adult patients with unresectable or metastatic melanoma with a BRAF V600 mutation [17, 18] and as combination therapy since recently also for metastatic non-small cell lung cancer harboring BRAF V600E mutations [19]. Despite improved response rates and overall survival of BRAF-V600 mutant cancer patients, resistance is rapidly acquired, resulting in a relapse of most patients within a year [20]. This effect may be partially due to the fast metabolism of dabrafenib (with a half-life of $\sim$5 hours [21]). There are three major metabolites of dabrafenib that have been identified with potential pharmacological effects: hydroxy-dabrafenib (HDB), carboxy-dabrafenib (CDB), and desmethyl-dabrafenib (DDB), whereas HDB appears to contribute significantly to the pharmacological activity [22].

© .

2 METHODS

2.1 Structure preparation and modelling

The crystal structure of BRAF with dabrafenib as co-crystallized ligand, PDB-ID: 4XV2, was downloaded from the RCSB protein data bank (PDB). The protomeric structure (chain A) was prepared for MD with an in-house Python script using Modeller [23] for modelling missing residues to match the canonical sequence (UniProt identifier: P15056-1), but for the position V600 mutated to E. Here, the coordinates for atoms present in structure 4XV2 are kept fixed and only the missing loops (plus one adjacent residue, to avoid unrealistic geometries caused by ambiguous termini atom positions) are modelled. The complexes with the three metabolites CDB, HDB, and DDB are generated by docking them with PLANTS [24] into the previously generated complete protein structure with dabrafenib as an anchor. Hydrogen atoms of the respective ligands were modeled with OpenBabel at pH 7. This results in zero net charge for DB, HDB, and DDB and a negative net charge for CDB, due to the deprotonated carboxy group.

2.2 Electrostatic potential and dielectric constant maps

The BRAF-DB complex structure was used to calculate electrostatic potential maps and dielectric maps with DelPhi [25, 26]. DelPhi is a free command line tool that calculates the electrostatic potential for biomolecules by solving the Poisson Boltzmann equation.

2.3 Molecular dynamics simulation

All simulations were carried out with Gromacs 2018 [27]. The ligand topologies were generated using the ACPYPE/ANTECHAMBER [28] program of AmberTools17 [29] with partial charges generated by the empirical charge model AM1-BCC. The ligands parameters are based on the General Amber Force Field (GAFF) and the Amber FF14SB force field was employed for the proteins. Each complex was solvated in a TIP3P water dodecahedral box, with periodic boundary conditions and a minimum distance of 1.0 nm from the surface of the complex to the edge of the box. Each system was neutralized by adding NA+ and Cl- ions to physiological concentration of 153.6 mM. A completely free steepest descent energy minimization for 2000 steps was followed by a 100-ps NVT equilibration and a 100-ps NpT equilibration with Parrinello-Rahman pressure coupling. NVT and NpT equilibrations were performed at a reference temperature of 300 K with ligand restraints of 1000 kJ/mol nm^2 in x,y,z directions. Finally, 50 ns unrestrained production runs were performed with a 2 fs time-step in the NpT ensemble and snapshots were saved every 10 ps. For each of the four ligands (DB, CDB, HDB, and DDB) five replica simulations were run with different randomly assigned initial velocities, resulting in a total of 250 ns simulation per ligand.

2.4 MM-PBSA calculation

For MM-PBSA calculation the 50-ns MD production trajectories were reduced to 501 frames each, by extracting a frame every 100 ps. The resulting snapshots of the MD simulations were utilized for post-processing free energies by the single-trajectory MM-PBSA method implemented in g_mmpbsa. Six different dielectric constants (ϵ=(2,4,6,8,12,20)) were used for the binding pocket, while the solution dielectric constant was kept constant at ϵ_s=80. Calculations are performed based on a homogeneous medium with a range of dielectric constants for the solute, an ionic strength of 153.6 mM, an ionic radius of 0.95 Å for positive charged ions and 1.81 Å for negative charged ions, and a solvent probe radius of 1.4 Å. An example configuration file for g_mmpbsa is provided within the supplements (Listing 1). Other parameters influencing the grid dimensions of the calculation, such as 'cfac', 'gridspace' and 'fadd' were varied from suggested defaults (1.5, 0.5 and 10, respectively) showing only marginal variations in the results and therefore not further changed.

Analysis and visualization is performed with provided scripts from the g_mmpbsa package [30], Chimera [31], PyMol [32] and Python scripts.

3 RESULTS

With this study we provide a basis for important considerations when employing MM-PBSA based affinity estimations on kinases. The oncogenic protein kinase BRAF-V600E together with the clinical drug dabrafenib serve as example for pointing out methodical issues that can arise when computing affinities in standard drug design projects.

As the binding mode of dabrafenib in BRAF-V600E is experimentally known (PDB-ID: 4XV2), we take advantage of this complex for further calculations and use it as template for docking the dabrafenib metabolites. Unfortunately the experimental structure is not complete and for chain A the missing loop residues (432-448, 488, 489, 597-614, 627-631, 721-723) had to be modelled.

3.1 Electrostatic potential and dielectric constant distribution of BRAF kinase

Electrostatics plays an important role in regulating interactions between biological macromolecules. The electrostatic potential map of the protein-ligand complex BRAF-dabrafenib at pH 7 shows remarkable variations at different slicing depths within the binding pocket, whereas the values within protein stay rather constant (see Figure 1).

Figure 1. Electrostatic potential (ϕ) map calculated on the protein-ligand complex structure. Six consecutive slices through the protein that sample the depth of the binding pocket, where the coloring shows the electrostatic potential at the slicing surface. The ϕ-map is calculated with DelPhi and visualized with Chimera.

The dielectric constant map equally points out variations inside the binding pocket in contrast to the protein interior (compare Figure 2). In particular, the tri-methyl moiety of dabrafenib is pointing outside the binding pocket and lies outside the cutoff where an epsilon value of 80 is reached (see Figure 2 closeup on binding pocket). This indicates that it is completely solvent accessible and should be considered as solvated.

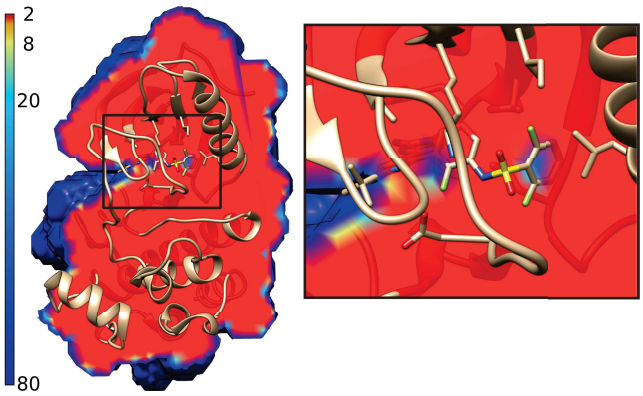


Figure 2. Dielectric constant (ϵ) map calculated on the protein-ligand complex structure with a complete view of the complex (left) and a close-up on the ligand (right). The ϵ-map is calculated with DelPhi and visualized with Chimera.

3.2 MM-PBSA, the dielectric constant and ligand charges

MM-PBSA binding affinity calculations based on 50ns MD simulations for dabrafenib (DB) and its metabolites carboxy-dabrafenib (CDB), desmethyl-dabrafenib (DDB), and hydroxy-dabrafenib (HDB) with BRAF-V600E (see Figure 3) and with BRAF-WT with a structured, helical activation loop (see Figure S2) are dependant on the solute dielectric constant (ϵ). Especially the charged CDB shows an inverted behaviour compared to the other three molecules. Whereas the binding energy for DB, DDB and HDB gradually decreases with increasing ϵ (ranging from 2 to 20), CDB has an extremely low energy at ϵ=2, which rapidly increases when shifting to a slightly higher ϵ=4, but stays constant from ϵ=8 onward.

3.2.1 Energetic contribution of protein residues to ligand binding

To investigate the reason for the extreme discrepancies between calculations at different dielectric constants residue-wise energetic contributions to ligand binding are investigated using the free energy decomposition scheme of g_mmpbsa. For visualization the energies given as kJ per mole are mapped onto the structure (see Figure 4 and S1). The energetic contributions per residue along the protein sequence between a dielectric constant of 2 and 8 (shown in Figure 5) highlights extreme discrepancies for CDB, whereas only the CDB pattern with ϵ=8 agrees with the patterns for DB, DDB and HDB, which are very similar. Therefore, calculations performed with a dielectric constant of 2 are considered as untrustworthy and protein residues with most important energetic contributions are compared between the DB and its metabolites at ϵ=8 (see Figure 6). The contributions appear to be highly similar, except for Lys483, which shows

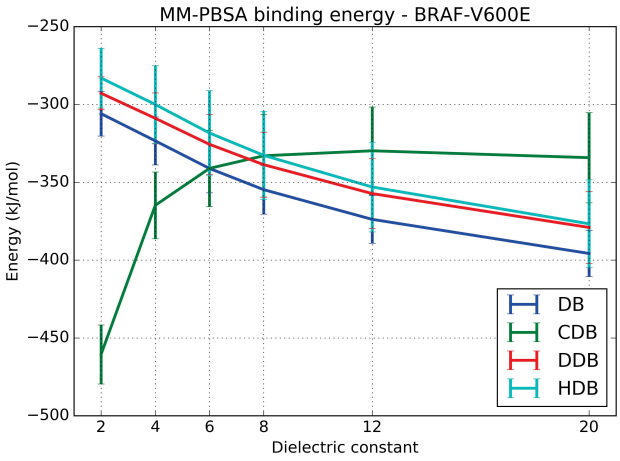


Figure 3. Averaged MM-PBSA binding energies (only enthalpic contribution) for BRAF-V600E with dabrafenib (DB) and its metabolites carboxy-dabrafenib (CDB), desmethyl-dabrafenib (DDB), and hydroxy-dabrafenib (HDB). The averages are calculated for each ligand based on 2505 complex conformations from five replica MD trajectories at six different dielectric constants (2,4,6,8,12, and 20). The error bars are standard deviations across the five replica MD trajectories for each dielectric constant.

increased variations and tends to more favorable energies for DB than its metabolites.

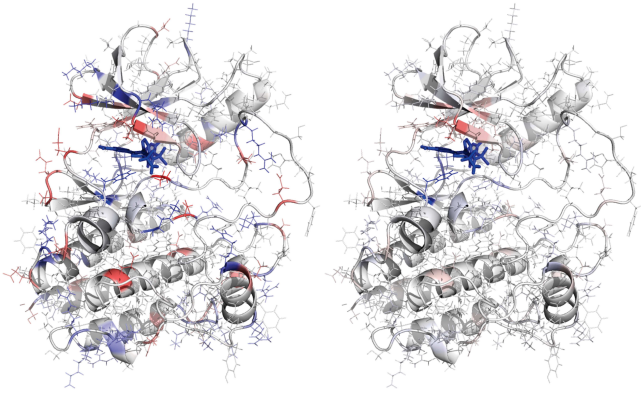


Figure 4. Energetic contribution of protein residues to ligand binding for the charged metabolite carboxy-dabrafenib (CDB) calculated with different dielectric constants of 2 (left) and 8 (right). Color coding = blue-white-red, with a minimum of -33 and a maximum of +33 kJ/mol (visualized with PyMol)

3.2.2 Complex evaluation by scoring function DSX

As external validation of the binding poses of the docked metabolites CDB, DDB and HDB the knowledge-based scoring function DSX (via DSX-online [33]) was used to evaluate the docked complexes that served as starting structures for MD simulations. Providing a score for protein-ligand complexes together with a visualization of the per-atom score contributions DSX-online allows

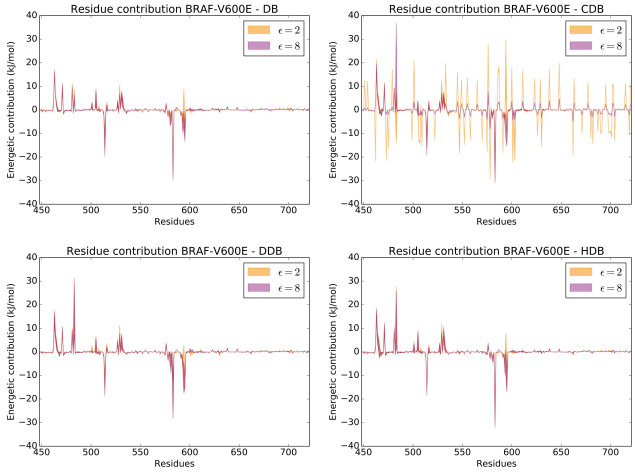


Figure 5. Energetic contribution of protein residues to ligand binding for DB and its metabolites CDB, DDB and HDB, calculated with dielectric constants of 2 and 8.

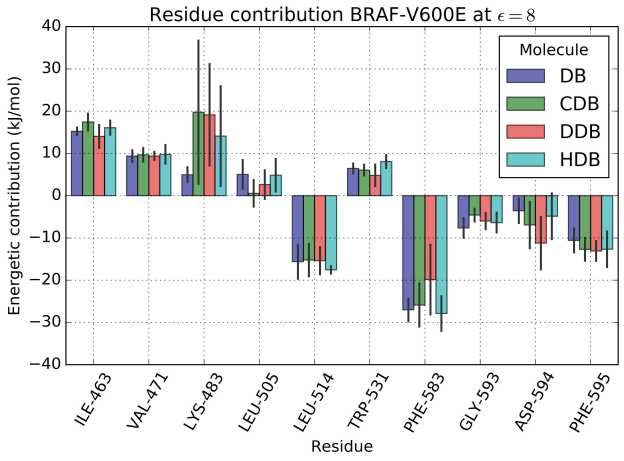


Figure 6. Protein residues with most important energetic contributions (absolute value larger than 10 kJ/mol for at least one simulation) to ligand binding for DB and its metabolites CDB, DDB and HDB, calculated with a dielectric constant of 8. The bar heights are the average values across the 5 replicas and the error

for investigating possible reasons for binding discrepancies. DSX scoring (using CSD potentials) provides the following ranking (from best to worst, with respective scores):

DB (-168.6) - HDB (-137.7) - CDB (-135.7) - DDB (-124.9).

The four molecules are scored highly favorable, very similar and the visualization of the per-atom score contributions showed only marginal variations for the moieties differing between the molecules (see Figure S3). Only minor unfavorable distances were detected between a few atoms of the identical parts of the metabolites and surrounding protein residues (that are supposedly due to tiny pose variations upon docking).
The evaluation by DSX suggests that the metabolites have highly similar binding affinities, which is equally the case for the MM-PBSA calculations with a dielectric constant of 8 (see Figure 3 and S2) and therefore, additionally confirms the parameter choice ϵ=8 for the protein kinase.

3.2.3 Affinity prediction by docking and machine learning

In order to further evaluate the proper ranking of the four compounds (DB, CDB, DDB and HDB), we applied a second completely independent affinity prediction method based on machine learning. The method is described in details in a previous publication [34]. Training of the random forest machine learning algorithm in regression mode was performed based on the ligand dataset BRAF-V600E (with annotated IC50 affinity measures - 2193 molecules) from BindingDB (2018). The method relies on data from multi-structure docking and pose evaluation of the ligand dataset on the @TOME server [35] taking into account all available BRAF structures, and ligand-based molecular descriptors. The machine learning method predicted the following affinity ranking (from highest to lowest) with pIC50 values ($[-log_{10}(M)]$):

DB (8.42) - HDB (8.19) - CDB (8.05) - DDB (8.00).

Remarkably, the machine learning based ranking is the same as for the DSX evaluation, also predicting highly similar binding affinities for the metabolites and an increased affinity for DB. This again, confirms the validity of the choice of the dielectric constant (ϵ=8), at with this tendency is equivalently reproduced.

3.2.4 Comparison with reported affinity measures from literature

GlaxoSmithKline published studies on the activity of their drug Dabrafenib and the three mayor metabolites [36, 37]. Interestingly, measured half-lives for CDB and DDB were longer than for DB and HDB [36]. The study of Ellens et al. [37] reports that, based on *in vitro* antiproliferative IC50 measures, HDB and DDB should be potent inhibitors of BRAF-V600E, slightly less active than DB whereas the activity of CDB is largely reduced. Comparison of the affinity measures with the affinity ranking using MM-PBSA suggests that high values of the dielectric constant ($\epsilon > 8$) are appropriate to obtain the equivalent ranking of DB - HDB/DDB - CDB, from best to worst (compare Figure 3).

4 DISCUSSION

Most biological processes are influenced or even governed by electrostatic effects. Structure-function correlations in general and ligand-receptor interactions in particular are heavily dependant on accurate electrostatic calculations. Just as the electrostatic contribution to the solvation / desolvation process has proved to be of paramount importance. However, the need for discriminatory case studies has not received the attention it deserves. The PB and GB models, provide very performant tools for modelling the effect of the solvent around the protein and they are widely used techniques for binding affinity estimation. The electrostatic contribution is modeled here as a dielectric linear response to the electric field generated by the atomic charges. There have been several previously reported promising results with excellent correlations with experimental data [38, 39, 40].

It has been shown that in implicit solvent simulations that use PB forces employing a dielectric constant of 1.0 (as used in many force fields) the resulting MD trajectories do not always preserve native structures (using different protocols and programs). To reduce solvation forces, a common technique is to raise the dielectric constant [6]. The dielectric constant for a protein has traditionally been estimated below, or at around 4 [41, 42], but can also be significantly larger, as large as 10, in sites of catalytic importance [43]. The optimal value differs between systems and are for example set to 4 [44] or even to 17 [45]. For implicit solvent simulations the proper dielectric constant of the solute is a controversial issue in the literature (see, e.g., [46]).

However, MM-PB/GBSA is a technique mainly used for predicting relative binding energies and not absolute ones, since several effects such as hydration/dehydration, entropy and binding pathway contributions can hardly be taken into account. Thus the dielectric constant becomes only an important factor when the ranking of potential ligands is impacted. This is in particular the case when partial charges differ largely among the ligands to be ranked.

By definition, the concept of a dielectric constant is used to describe the collective behavior of matter and does not describe effects on the atomic level. In practice, using low dielectric constants for the solute (protein), such as ϵ=2, accounts for electronic polarizability and is more sensitive to changes in the molecules, and can therefore be very useful for distinguishing between rather similar ligands. Nonetheless, as we point out with this study, the standard employment of such low dielectric constants may lead to wrong assumptions on the relative ranking of ligands, particularly when they differ in charges. Special care needs to be taken to adjust the dielectric constant for the system under investigation in order to avoid artefacts (e.g. contributions from polar residues located far from the binding site). Based on this study we suggest the use of a rather elevated solute dielectric constant of about 8 for kinases, in particular when investigating charged ligands.

REFERENCES

[1]T. P. Straatsma and H. J. C. Berendsen. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *The Journal of Chemical Physics*, 89(9):5876–5886, November 1988.

[2]Agastya P. Bhati, Shunzhou Wan, David W. Wright, and Peter V. Coveney. Rapid, Accurate, Precise, and Reliable Relative Free Energy Prediction Using Ensemble Based Thermodynamic Integration. *Journal of Chemical Theory and Computation*, 13(1):210–222, January 2017.

[3]Francesco Manzoni and Ulf Ryde. Assessing the stability of free-energy perturbation calculations by performing variations in the method. *Journal of Computer-Aided Molecular Design*, 32(4):529–536, 2018.

[4]Johan qvist, Carmen Medina, and Jan-Erik Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Engineering, Design and Selection*, 7(3):385–391, March 1994.

[5]Peter A. Kollman, Irina Massova, Carolina Reyes, Bernd Kuhn, Shuanghong Huo, Lillian Chong, Matthew Lee, Taisung Lee, Yong Duan, Wei Wang, Oreola Donini, Piotr Cieplak, Jaysharee Srinivasan, David A. Case, and Thomas E. Cheatham. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research*, 33(12):889–897, December 2000.

[6]Samuel Genheden and Ulf Ryde. Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. *Proteins*, 80(5):1326–1342, May 2012.

[7]W. Wang, J. Wang, and P. A. Kollman. What determines the van der Waals coefficient beta in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? *Proteins*, 34(3):395–402, February 1999.

[8]Samuel Genheden and Ulf Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5):449–461, May 2015.

[9]Aaron Weis, Kambiz Katebzadeh, Pr Sderhjelm, Ingemar Nilsson, and Ulf Ryde. Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. *Journal of Medicinal Chemistry*, 49(22):6596–6606, November 2006.

[10]Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *Journal of Computational Chemistry*, 32(5):866–877, April 2011. 00234.

[11]Huiyong Sun, Youyong Li, Mingyun Shen, Sheng Tian, Lei Xu, Peichen Pan, Yan Guan, and Tingjun Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Physical chemistry chemical physics: PCCP*, 16(40):22035–22045, October 2014. 00071.

[12]Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 51(1):69–82, January 2011.

[13]Salla I. Virtanen, Sanna P. Niinivehmas, and Olli T. Pentikinen. Case-specific performance of MM-PBSA, MM-GBSA, and SIE in virtual screening. *Journal of Molecular Graphics & Modelling*, 62:303–318, November 2015. 00005.

[14]Marc Adler and Paul Beroza. Improved ligand binding energies derived from molecular dynamics: replicate sampling enhances the search of conformational space. *Journal of Chemical Information and Modeling*, 53(8):2065–2072, August 2013. 00008.

[15]Krishna Ravindranathan, Julian Tirado-Rives, William L. Jorgensen, and Cristiano R. W. Guimares. Improving MM-GB/SA Scoring through the Application of the Variable Dielectric Model. *Journal of Chemical Theory and Computation*, 7(12):3859–3865, December 2011. 00015.

[16]Geoffrey T. Gibney and Jonathan S. Zager. Clinical development of dabrafenib in BRAF mutant melanoma and other malignancies. *Expert Opinion on Drug Metabolism & Toxicology*, 9(7):893–899, July 2013.

[17]Anita D. Ballantyne and Karly P. Garnock-Jones. Dabrafenib: first global approval. *Drugs*, 73(12):1367–1376, August 2013.

[18]Alexander M. Menzies and Georgina V. Long. Dabrafenib and trametinib, alone and in combination for BRAF-mutant metastatic melanoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 20(8):2035–2043, April 2014.

[19]Lauretta Odogwu, Luckson Mathieu, Gideon Blumenthal, Erin Larkins, Kirsten B. Goldberg, Norma Griffin, Karen Bijwaard, Eunice Y. Lee, Reena Philip, Xiaoping Jiang, Lisa Rodriguez, Amy E. McKee, Patricia Keegan, and Richard Pazdur. FDA Approval Summary: Dabrafenib and Trametinib for the Treatment of Metastatic Non-Small Cell Lung Cancers Harboring BRAF V600e Mutations. *The Oncologist*, 23(6):740–745, June 2018.

[20]Weijiang Zhang. BRAF inhibitors: the current and the future. *Current Opinion in Pharmacology*, 23:68–73, August 2015.

[21]Cathrine L. Denton, Elisabeth Minthorn, Stanley W. Carson, Graeme C. Young, Lauren E. Richards-Peterson, Jeffrey Botbyl, Chao Han, Royce A. Morrison, Samuel C. Blackman, and Daniele Ouellet. Concomitant oral and intravenous pharmacokinetics of dabrafenib, a BRAF inhibitor, in patients with BRAF V600 mutation-positive solid tumors. *Journal of Clinical Pharmacology*, 53(9):955–961, September 2013.

[22]Alicja Puszkiel, Galle No, Audrey Bellesoeur, Nora Kramkimel, Marie-Nolle Paludetto, Audrey Thomas-Schoemann, Michel Vidal, Franois Goldwasser, Etienne Chatelut, and Benoit Blanchet. Clinical Pharmacokinetics and Pharmacodynamics of Dabrafenib. *Clinical Pharmacokinetics*, 58(4):451–467, April 2019.

[23]A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, December 1993. 09586.

[24]Oliver Korb, Thomas Sttzle, and Thomas E. Exner. Empirical Scoring Functions for Advanced ProteinLigand Docking with PLANTS. *Journal of Chemical Information and Modeling*, 49(1):84–96, January 2009. 00321.

[25]Lin Li, Chuan Li, Subhra Sarkar, Jie Zhang, Shawn Witham, Zhe Zhang, Lin Wang, Nicholas Smith, Marharyta Petukh, and Emil Alexov. DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC biophysics*, 5:9, May 2012.

[26]Lin Li, Chuan Li, Zhe Zhang, and Emil Alexov. On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *Journal of Chemical Theory and Computation*, 9(4):2126–2136, April 2013.

[27]Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, March 2008. 07109.

[28]Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling*, 25(2):247–260, October 2006. 01693.

[29]D. A. Case, D. S. Cerutti, T. E. Cheatham, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, D. M. York, and P. A. Kollman. Amber 2017, University of California, San Francisco, 2017. 00000.

[30]Rashmi Kumari, Rajendra Kumar, Open Source Drug Discovery Consortium, and Andrew Lynn. g_mmpbsa–a GROMACS tool for high-throughput MM-PBSA calculations. *Journal of Chemical Information and Modeling*, 54(7):1951–1962, July 2014.

[31]Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, October 2004.

[32]Schrdinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. November 2015.

[33]Gerd Neudert and Gerhard Klebe. DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. *Journal of Chemical Information and Modeling*, 51(10):2731–2745, October 2011.

[34]Melanie Schneider, Jean-Luc Pons, William Bourguet, and Gilles Labesse. Towards accurate high-throughput ligand affinity prediction by exploiting structural ensembles, docking metrics and ligand similarity. *Bioinformatics (Oxford, England)*, July 2019.

[35]Jean-Luc Pons and Gilles Labesse. @TOME-2: a new pipeline for comparative modeling of proteinligand complexes. *Nucleic Acids Research*, 37(suppl_2):W485–W491, July 2009.

[36]David A. Bershas, Daniele Ouellet, Donna B. Mamaril-Fishman, Noelia Nebot, Stanley W. Carson, Samuel C. Blackman, Royce A. Morrison, Jerry L. Adams, Kristen E. Jurusik, Dana M. Knecht, Peter D. Gorycki, and Lauren E. Richards-Peterson. Metabolism and disposition of oral dabrafenib in cancer patients: proposed participation of aryl nitrogen in carbon-carbon bond cleavage via decarboxylation following enzymatic oxidation. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, 41(12):2215–2224, December 2013.

[37]Harma Ellens, Marta Johnson, Sarah K. Lawrence, Cory Watson, Liangfu Chen, and Lauren E. Richards-Peterson. Prediction of the Transporter-Mediated Drug-Drug Interaction Potential of Dabrafenib and Its Major Circulating Metabolites. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, 45(6):646–656, 2017.

[38]Bo Yang, Adel Hamza, Guangju Chen, Yan Wang, and Chang-Guo Zhan. Computational determination of binding structures and free energies of phosphodiesterase-2 with benzo[1,4]diazepin-2-one derivatives. *The Journal of Physical Chemistry. B*, 114(48):16020–16028, December 2010. 00022.

[39]Cristiano R. W. Guimares and Mario Cardozo. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *Journal of Chemical Information and Modeling*, 48(5):958–970, May 2008. 00153.

[40]J. Wang, P. Morin, W. Wang, and P. A. Kollman. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *Journal of the American Chemical Society*, 123(22):5221–5230, June 2001. 00603.

[41]S. C. Harvey and P. Hoekstra. Dielectric relaxation spectra of water adsorbed on lysozyme. *The Journal of Physical Chemistry*, 76(21):2987–2994, October 1972.

[42]M. K. Gilson and B. H. Honig. The dielectric constant of a folded protein. *Biopolymers*, 25(11):2097–2119, November 1986.

[43]Gregory King, Frederick S. Lee, and Arieh Warshel. Microscopic simulations of macroscopic dielectric constants of solvated proteins. *The Journal of Chemical Physics*, 95(6):4366–4377, September 1991.

[44]Federico Fogolari, Alessandro Brigo, and Henriette Molinari. Protocol for MM/PBSA Molecular Dynamics Simulations of Proteins. *Biophysical Journal*, 85(1):159–166, July 2003.

[45]Ben Zhuo Lu, Wei Zu Chen, Cun Xin Wang, and Xiao-jie Xu. Protein molecular dynamics with electrostatic force entirely determined by a single Poisson-Boltzmann calculation. *Proteins*, 48(3):497–504, August 2002.

[46]C. N. Schutz and A. Warshel. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins*, 44(4):400–417, September 2001.

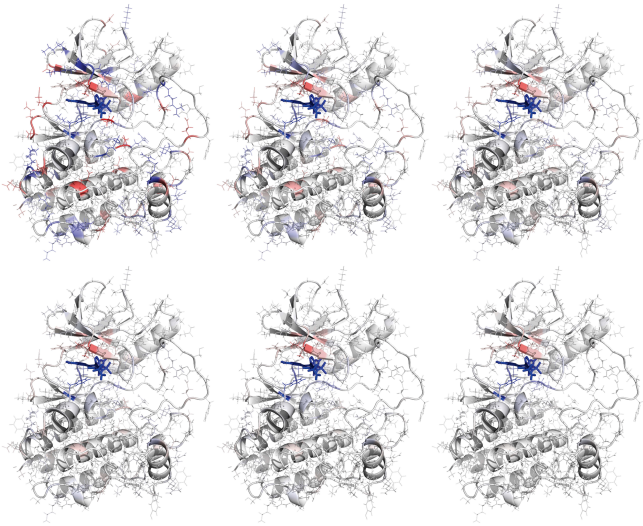


Figure S1. Energetic contribution of protein residues to ligand binding for the charged metabolite carboxy-dabrafenib (CDB) calculated with six different dielectric constants ϵ=(2,4,6) top row, and ϵ=(8,12,20) bottom row, from left to right. Color coding = blue-white-red, with a minimum of -33 and a maximum of +33 kJ/mol (visualized with PyMol).

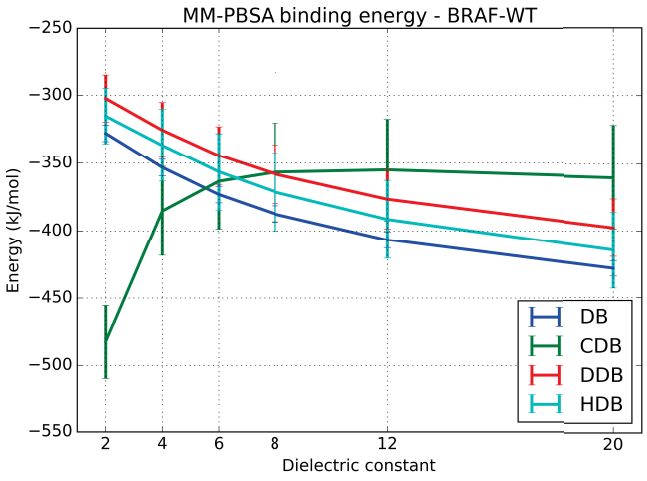


Figure S2. Averaged MM-PBSA binding energies (only enthalpic contribution) for BRAF-WT with dabrafenib (DB) and its metabolites carboxy-dabrafenib (CDB), desmethyl-dabrafenib (DDB), and hydroxy-dabrafenib (HDB). The averages are calculated for each ligand based on 2505 complex conformations from five replica MD trajectories at six different dielectric constants. The error bars are standard deviations across the five replica MD trajectories for each dielectric constant.

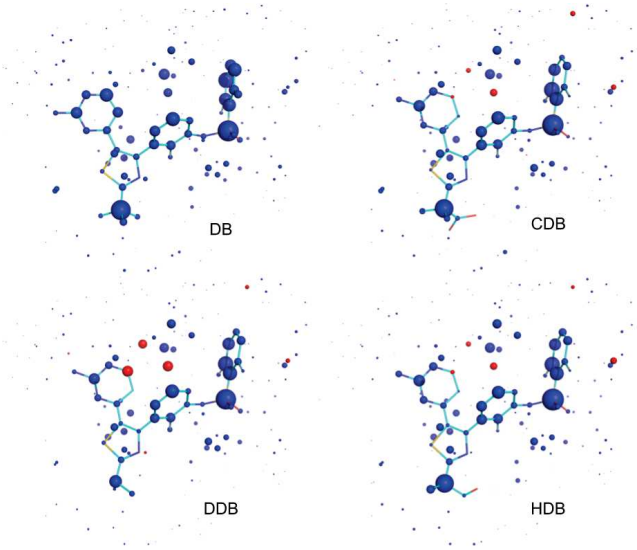


Figure S3. DSX evaluation: visualization of the per-atom score contributions of DB and its metabolites CDB, DDB and HDB in the binding pocket (visualized with PyMol). Favorably interacting atoms are surrounded by blue spheres and disfavorable interactions are shown in red. The sizes of the spheres correspond to the values of the contributing per-atom scores.

Listing 1. MM-PBSA example configuration file for g_mmpbsa with ϵ=6

```
;Polar calculation: "yes" or "no"
polar           = yes

;=============
;PSIZE options
;=============
;Factor by which to expand molecular dimensions to get
     coarsegrid dimensions.
cfac            = 1.5

;The desired fine mesh spacing (in A)
gridspace       = 0.5

:Amount (in A) to add to molecular dimensions to get fine
      grid dimensions.
fadd            = 10

;Maximum memory (in MB) available per-processor for a
     calculation.
gmemceil        = 4000

;============================================
;APBS kwywords for polar solvation calculation
;============================================
;Charge of positive ions
pcharge         = 1

;Radius of positive charged ions
prad            = 0.95

;Concentration of positive charged ions
pconc           = 0.1536

;Charge of negative ions
ncharge         = -1

;Radius of negative charged ions
nrad            = 1.81
```

```
;Concentration of negative charged ions
nconc           = 0.1536

;Solute dielectric constant
pdie            = 6

;Solvent dielectric constant
sdie            = 80

;Reference or vacuum dielectric constant
vdie            = 1

;Solvent probe radius
srad            = 1.4

;Method used to map biomolecular charges on grid. chgm =
     spl0 or spl2 or spl4
chgm            = spl4

;Model used to construct dielectric and ionic boundary.
     srfm = smol or spl2 or spl4
srfm            = smol

;Value for cubic spline window. Only used in case of srfm
      = spl2 or spl4.
swin            = 0.30

;Numebr of grid point per A^2. Not used when (srad = 0.0)
      or (srfm = spl2 or spl4)
sdens           = 10

;Temperature in K
temp            = 300

;Type of boundary condition to solve PB equation. bcfl =
     zero or sdh or mdh or focus or map
bcfl            = mdh

;Non-linear (npbe) or linear (lpbe) PB equation to solve
PBsolver        = lpbe


;============================================================
;APBS kwywords for Apolar/Non-polar solvation calculation
;============================================================
;Non-polar solvation calculation: "yes" or "no"
apolar          = yes

;Repulsive contribution to Non-polar
;===SASA model ====
;Gamma (Surface Tension) kJ/(mol A^2)
gamma           = 0.0226778

;Probe radius for SASA (A)
sasrad          = 1.4

;Offset (c) kJ/mol
sasaconst       = 3.84982

;===SAV model===
;Pressure kJ/(mol A^3)
press           = 0.234304

;Probe radius for SAV (A)
savrad          = 1.29

;Offset (c) kJ/mol
savconst        = 0


;Attractive contribution to Non-polar
;===WCA model ====
;using WCA method: "yes" or "no"
WCA             = no

;Probe radius for WCA
wcarad          = 1.20

;bulk solvent density in A^3
bconc           = 0.033428

;displacment in A for surface area derivative calculation
dpos            = 0.05

;Quadrature grid points per A for molecular surface or
     solvent accessible surface
APsdens         = 20

;Quadrature grid spacing in A for volume integral
     calculations
grid            = 0.45 0.45 0.45

;Parameter to construct solvent related surface or volume
APsrfm          = sacc

;Cubic spline window in A for spline based surface
     definitions
APswin          = 0.3

;Temperature in K
APtemp          = 300
```

3.7.2 MM-PBSA approaches on BRAF and PXR with designed drugs

3.7.2.1 The impact of protein structure loop-model on MM-PBSA results

The impact of the loop-models generated based on the two different crystallographic PDB strctures (4XV2 and 4CQE) is investigated for the molecules from synthesis round 2 (see Figure 3.30). Based on the MM-PBSA calculation on 500 frames extracted from single 50ns MD simulations a general trend of slighly higher affinities in the 4XV2 model compared to the 4CQE model can be stated (which is the case for all P06F molecules except P06F-Pip). However, the error bars, which are based on the variability among the 500 frames per simulation, are overlapping in most cases, indicating an elevated uncertainty for this conclusion.

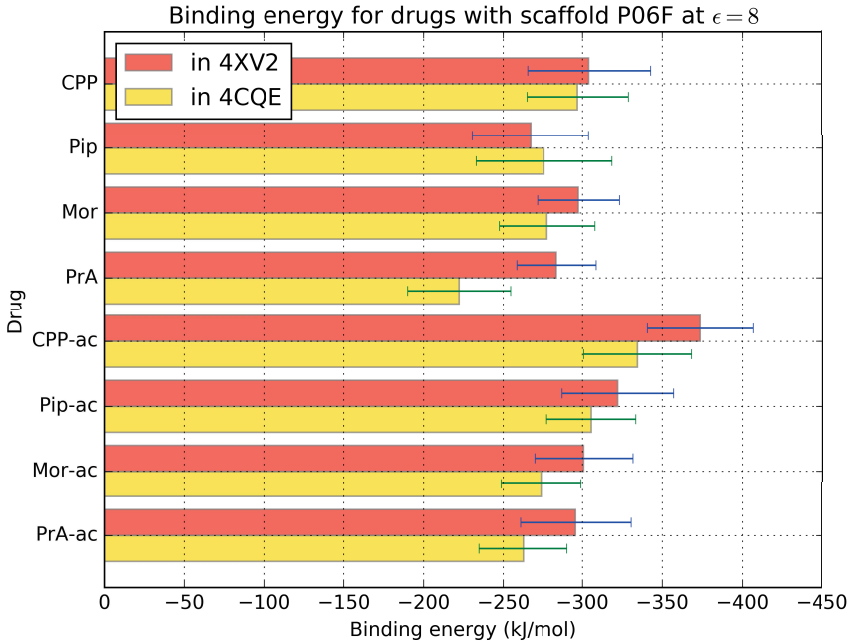


Figure 3.30: MM-PBSA binding energy of the 8 molecules from synthesis round 2 (drug scaffold P06F) at a dielectric constant of 8 in two different loop-model structures: 4XV2 and 4CQE. The errors are based on the variability among the selected 500 frames from a 50ns MD simulation for each complex.

As the loop conformation seems to have an impact on the affinity estimation, P06 is evaluated in the four different loop-model structures: 4XV2, 4CQE, V600E and the structured WT (see Figure 3.31). The tendency of lower affinity estimations for the 4CQE model is confirmed for P06. Furthermore, the V600E model, as well as the structured WT model, seem to result in similar affinity estimations as the 4XV2 model. This indicates that the three models 4XV2, V600E and the structured WT may equally be useful for distinguishing molecules that are derivatives of P06.

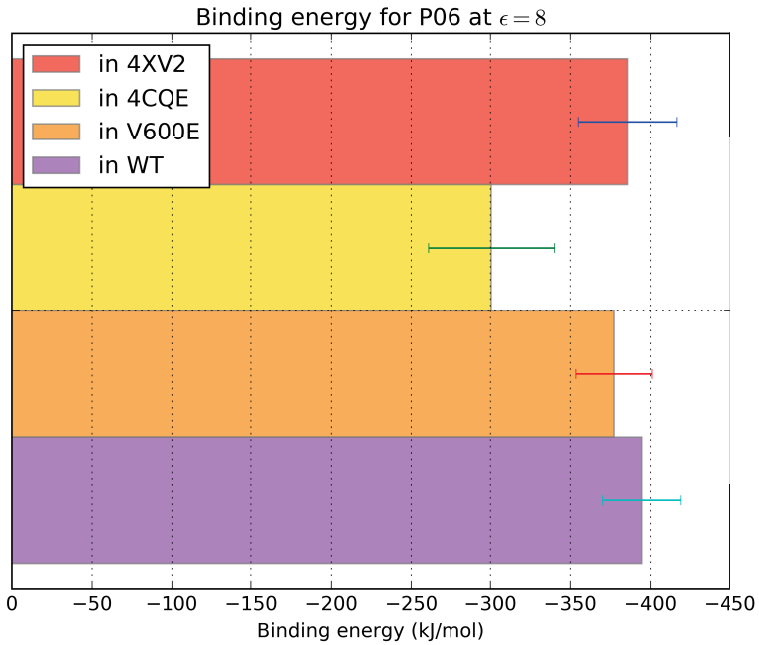
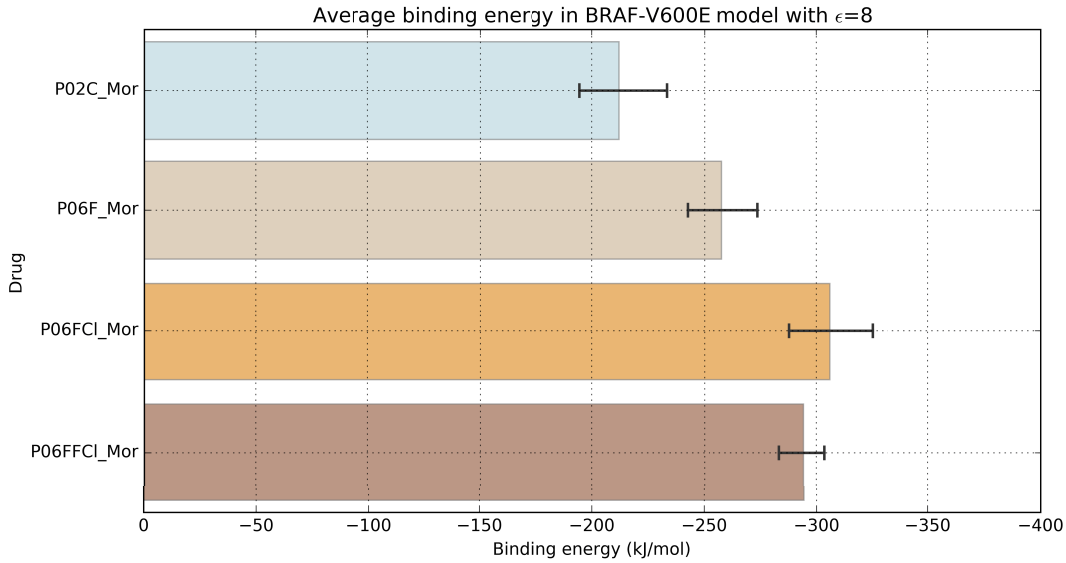


Figure 3.31: MM-PBSA binding energy of P06 at a dielectric constant of 8 in four different loop-model structures: 4XV2, 4CQE, V600E and the structured WT. The errors are based on the variability among the selected 500 frames from a 50ns MD simulation for each complex.

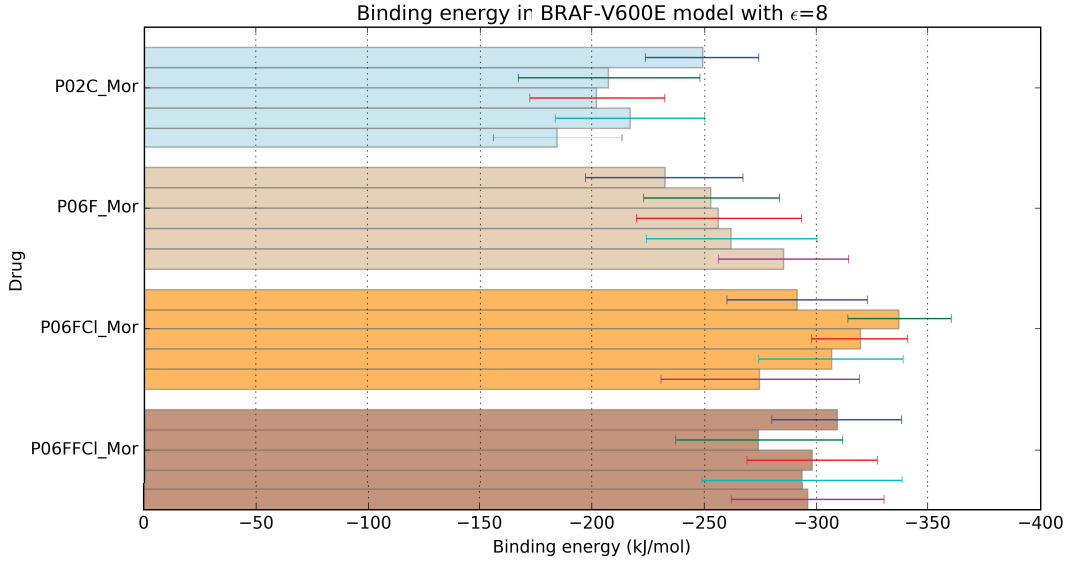
3.7.2.2 The effect of different drug scaffolds

MM-PBSA affinity differences between different drug scaffolds

In order to investigate the impact of the different scaffolds MM-PBSA calculations were performed on 5 replicas of 50ns MD simulations from which 500 frames were extracted for each simulation. The molecules with a morpholine (Mor) extension and the four different scaffolds, P02C, P06F, P06FCl, and P06FFCl, are simulated as complex with the BRAF-V600E model. The error estimates across the 5 replicas (see Figure 3.32a) tend to be smaller, or in a similar range as the error estimates from single MD simulations (see Figure 3.32b) and show lesser overlaps. Thus, when performing several replicas, the molecules with different scaffolds become distinguishable by average binding affinity with a distinct ordering, which is not clearly apparent from calculations based on single simulations.



(a) The bar heights are the MM-PBSA result averages across the 5 MD replicas and the errors are based on the variability across the 5 MD replicas, for each complex.

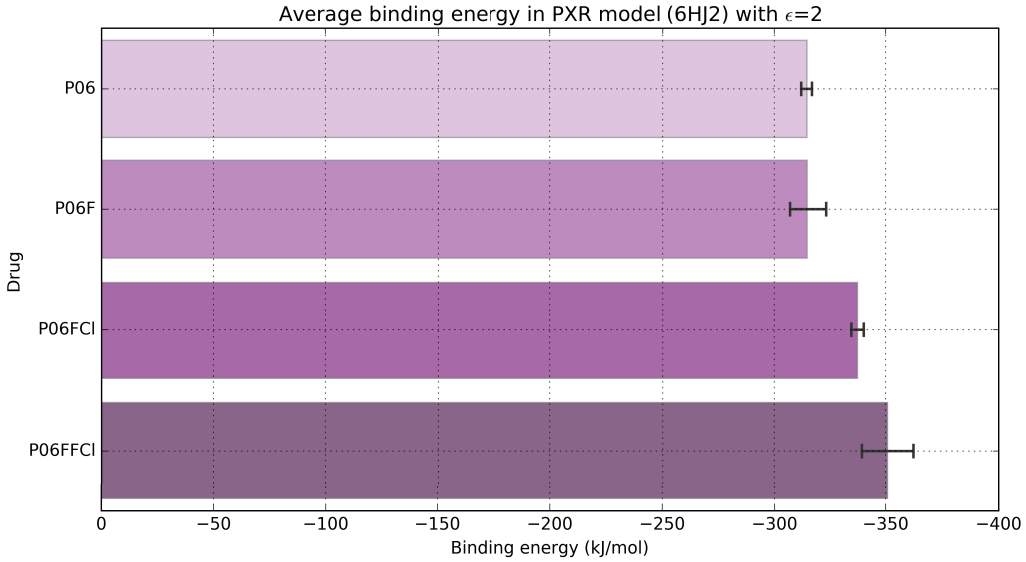


(b) The MM-PBSA results from the 5 MD replicas are plotted separately and the errors are based on the variability among the selected 500 frames from a 50ns MD simulation for each complex.

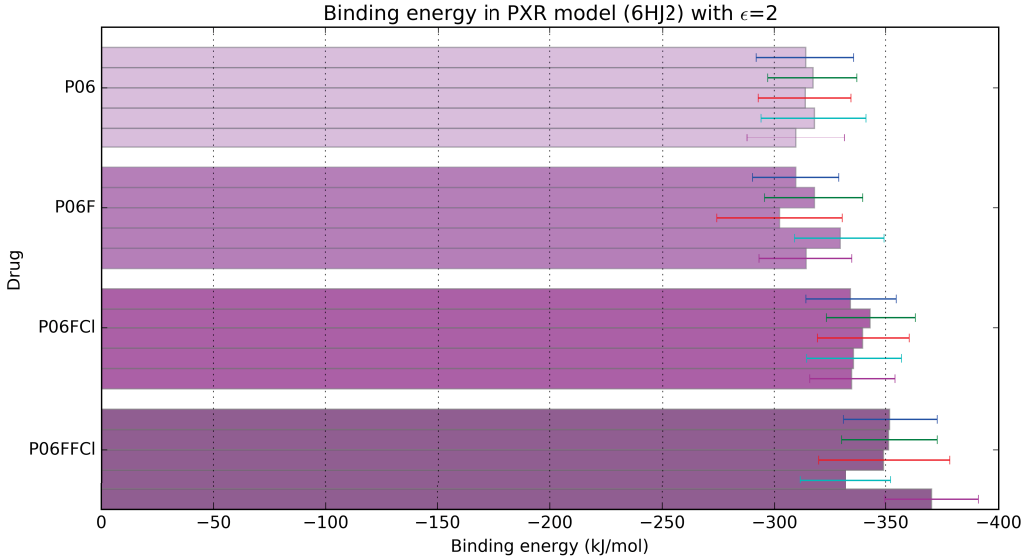
Figure 3.32: MM-PBSA binding energy of molecules with four different scaffolds (P02C, P06F, P06FCl, and P06FFCl) and a morpholine (Mor) as extension in the BRAF-V600E model, at a dielectric constant of 8.

To evaluate the effect of different scaffolds with respect to PXR binding MM-PBSA affinity estimations are performed using the crystallographic PXR structure 6HJ2 that is co-crystallized with P06 (processed by PDB-REDO and completed by the loop-modelling procedure). As the crystallographic pose of P06 is used for the modified drugs and larger replacements of the tertiary butyl moiety of P06 (e.g. by a morpholine) would clash with the PXR structure, the tertiary butyl moiety is kept to investigate scaffold effects in PXR. Without any change of the tertiary butyl moiety we expected the MM-PBSA affinities to stay rather close to the one for P06, which is in particular the case for the very similar molecule P06F (see Figure 3.33), having only one flour atom shifted from cis to trans position at the di-fluorophenyl ring compared to P06. In contrast, adding an additional

chlorine atom in para to the central fluorophenyl ring, as for P06FCl and P06FFCl seems to increase affinity to PXR.



(a) The bar heights are the MM-PBSA result averages across the 5 MD replicas and the errors are based on the variability across the 5 MD replicas, for each complex.



(b) The MM-PBSA results from the 5 MD replicas are plotted separately and the errors are based on the variability among the selected 500 frames from a 50ns MD simulation for each complex.

Figure 3.33: MM-PBSA binding energy of molecules with four different scaffolds (P06, P06F, P06FCl, and P06FFCl) containing the original tertiary butyl moiety of P06 in the PXR model based on 6HJ2, at a dielectric constant of 2.

Hydrogen bonding network between drug and protein

Furthermore, the hydrogen bonding of the ligands with the BRAF protein is investigated along the 50ns MD simulations. To do so, the VMD H-bonds plugin is used with the following criteria for the formation of a hydrogen bond: A hydrogen bond is formed between an atom with a hydrogen bonded to it (the donor, D) and another atom (the acceptor, A) provided that the distance D-A is less than the cut-off distance (default of 3.0 Å, set to 3.5 Å) and the angle D-H-A is less than the cut-off angle (default of 20°, set to 35°). H-bonds are calculated between protein and ligand for all frames, with both molecules as donor and acceptor. The H-bond occupancy is calculated per residue, by summing up the occupancies of all contributions from a protein residue, and averaged across the 5 replicas for each drug. Residues that formed at least in one MD simulation hydrogen bonds with occupancy $\geqslant$10 are listed in Figure 3.34. H-bond occupancy varies largely among the different replicas. Nonetheless, Cys532, Asp594 and Phe595 seem to be of major importance for drug binding, whereas the hydrogen bonding to Phe595 seems to be more transient, as it is not maintained throughout the whole trajectories (maximal 70%). These three residues have also been identified to form hydrogen bonds by the PLIP web server for P06 (compare Figure 3.18).

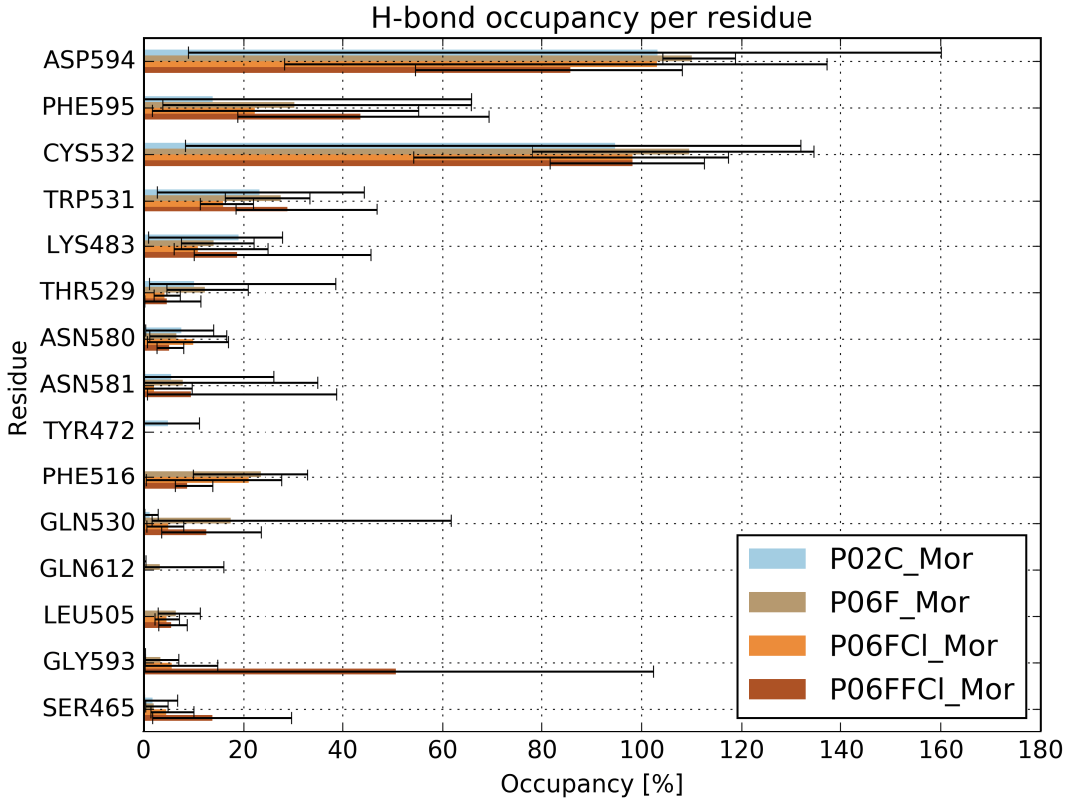


Figure 3.34: H-bond occupancy per residue averaged across the 5 MD replicas per complex. H-bonds are calculated between protein and ligand for all trajectory frames using the VMD H-bonds plugin and occupancies of all contributions from a protein residue are summed up. Residues that formed at least in one MD simulation hydrogen bonds with occupancy $\geqslant$10 are listed. The errors are based on the variability across the 5 MD replicas per complex. Note that the occupancy can be larger than 100% for a residue, as more than one h-bond can be formed per residue.

3.7.3 MM-PBSA approaches on ensemble-refined BRAF structures

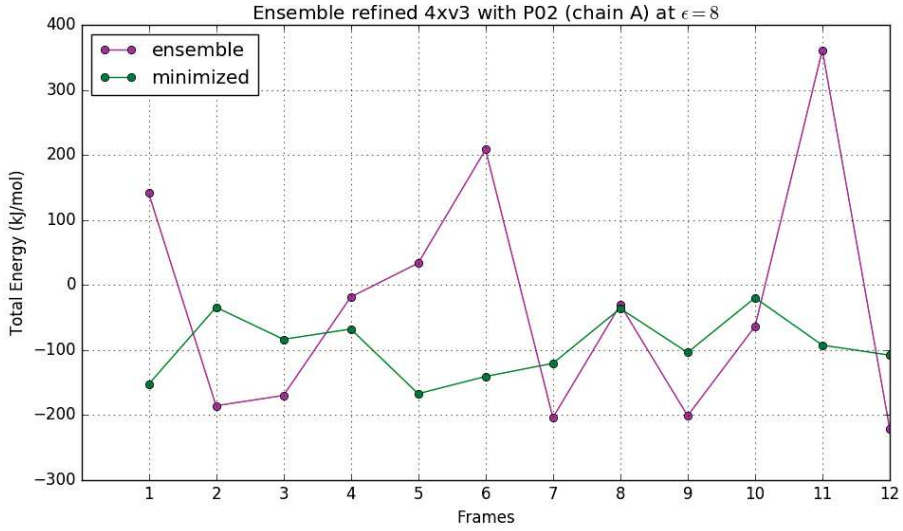
For structure-based VS and affinity estimations, such as MM-PBSA calculations, ensembles of structural conformations are used to be able to provide statistically sound values and to estimate error intervals. Nevertheless, the way of constructing such conformational ensembles is still a highly discussed topic and several approaches for generation of receptor conformational ensembles exist. Usually protein receptor conformations are identified using molecular dynamics (MD) simulations. Within this thesis work, not only conformations extracted from MD simulations, but also ensemble-refined structures were employed for MM-PBSA affinity calculations. Two differently generated ensembles were investigated, one without ligand hydrogens during refinement, and another including ligand hydrogens during refinement. For each refinement the ensemble with the lowest R_{free} value was selected for MM-PBSA calculations. Moreover, the ensemble structures were not only used as is for MM-PBSA calculations, but also in a second approach additionally minimized (for maximal 200 steps, using GROMACS and the Amber14SB force field) before submitting to MM-PBSA calculations. Four crystallographic BRAF complexes were investigated (with their crystallographic ligands): 4XV3 (with P02), 4XV2 (with P06), 5CSW (with P06) and the newly solved structure (with the designed drug candidate P06F-Mor), and the MM-PBSA results are listed in Table 3.3. All MM-PBSA calculations for BRAF complexes were performed with a solute dielectric constant of ε=8, and all protomers (identified here by chain) that are complexed with a ligand were analyzed separately.

PDB-ID - ligand-ID	chain	(-H) ensemble	(-H) minimized	(+H) ensemble	(+H) minimized
4XV3 - P02	A	3253.1 (± 6212.4)	-93.6 (± 70.3)	-29.2 (± 178.9)	-93.8 (± 45.8)
4XV2 - P06	A	-140.8 (± 304.3)	-83.9 (± 85.7)	-232.2 (± 282.3)	-42.8 (± 74.3)
	B	-194.3 (± 192.4)	-372.0 (± 23.1)	-313.3 (± 62.4)	-396.9 (± 19.3)
5CSW - P06	A	-315.4 (± 57.0)	-351.8 (± 128.8)	-344.5 (± 25.4)	-129.2 (± 215.2)
	B	38.9 (± 520.1)	-181.9 (± 178.2)	-289.7 (± 64.6)	-77.1 (± 185.6)
NEW - P06F-Mor	A	-	-	-289.8 (± 122.0)	-360.3 (± 28.1)
	B	-	-	-316.4 (± 38.1)	-361.4 (± 24.0)

Table 3.3: MM-PBSA affinity calculations on ensemble-refined BRAF structures (with solute dielectric constant ε=8). Binding energies are provided in kJ/mol. Ensembles originating from two different ensemble-refinement runs were used, one without ligand hydrogens and one with ligand hydrogens included during refinement (indicated by (-H) and (+H), respectively). The ensembles with lowest R_{free} values were selected for calculations. Each ensemble was employed directly for MM-PBSA calculations (indicated by "ensemble") and the structures were minimized prior to MM-PBSA calculations (indicated by "minimized").

When ligand hydrogens were not included during refinement they were subsequently added to the ensembles, prior to MM-PBSA calculations. This introduced atomic (VdW) clashes for some conformations in most systems (except 5CSW chain A), resulting in high energies (compare Table 3.3). Thus, the more recent protocol, including the ligand hydrogens during ensemble-refinement (indicated by (+H)) is expected to reduce or avoid the additional error caused by misplaced hydrogens (which are generally considered very flexible). Detailed (per conformation) results from those (+H)

refinement ensembles are displayed for structure 4XV3 with P02 in Figure 3.35 (that contains the ligand P02 only in one (chain A) of the two protomers), for structure 4XV2 with P06 in Figure 3.36, for structure 5CSW with P06 in Figure 3.37, and for the newly solved BRAF structure with the designed drug candidate P06F-Mor in Figure 3.38.
For the (+H) refinement ensembles minimization generally reduced the error intervals intensively by solving the atomic clashes, except for structure 5CSW. Investigation of the atomic model deposited in the PDB revealed that the model building was not carefully performed for 5CSW, as even binding site side-chains were not properly placed into the electron densities. Whereas the original ensemble provided rather reasonable values and error intervals, the minimization results in extensive errors. One protomer of structure 4XV2 (chain A) also shows remaining issues, since the error interval stays relatively high upon minimization, although it is largely reduced compared to the error of the original ensemble. Nonetheless, a remarkable difference between the ligands P02 and P06 is in agreement with machine learning predictions and affinity measurements (see Section 3.5.2 and 3.8.1.1, respectively).



(a) chain A.

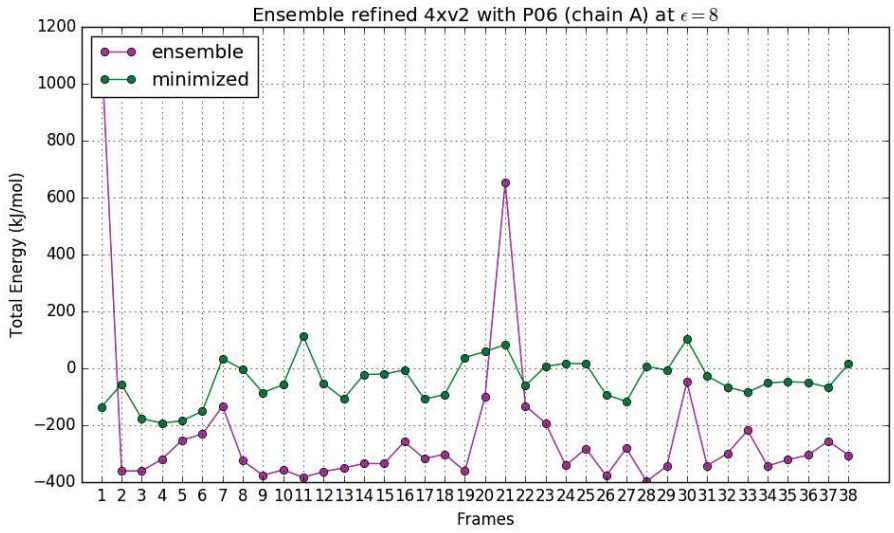
Figure 3.35: MM-PBSA results for the crystallographic ensemble-refined BRAF structure 4XV3 with P02.

Generally, it is important to keep in mind that the resolution of the diffraction data may have a non-negligible impact on the model quality, especially on the accuracy of the positioning of the ligand and protein side-chains. Consequently, the starting model quality may have an important impact on the ensemble-refined structures and on their usability for subsequent MM-PBSA calculations. In order to get a first glance onto general applicability rules the resolution and Diffraction Precision Index (DPI) of the investigated BRAF crystal structures are listed in Table 3.4.

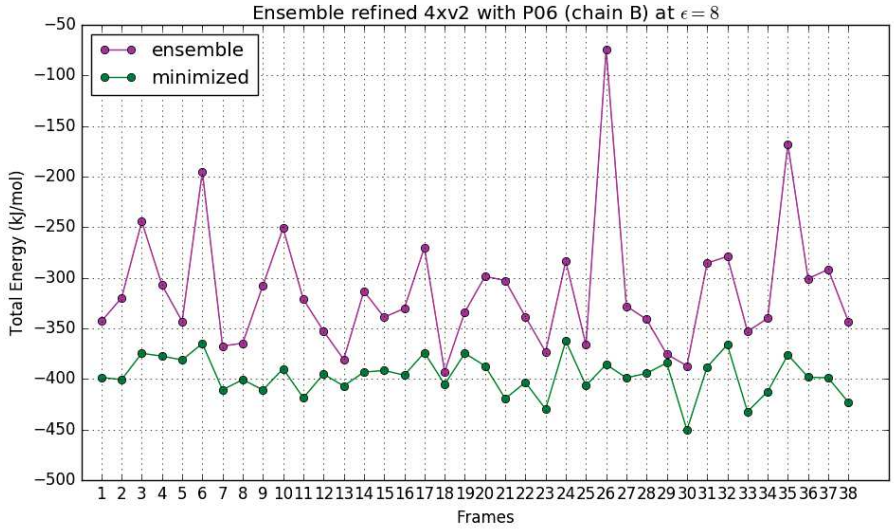
Based on the obtained results a preliminary suggestion for ensemble-refinement applicability would be to use structures with a resolution $\leq$ 2.5 Åand a DPI $\leq$ 0.35, and additionally verify the model quality of the binding site (side-chain positioning within the electron density) in case of binding energy evaluation.

PDB-ID	Resolution	DPI
4XV3	2.80 Å	0.417
4XV2	2.50 Å	0.262
5CSW	2.66 Å	0.385
NEW	2.37 Å	0.290

Table 3.4: Resolution and Diffraction Precision Index (DPI, calculated by Online-DPI[179]) of BRAF crystal structures.

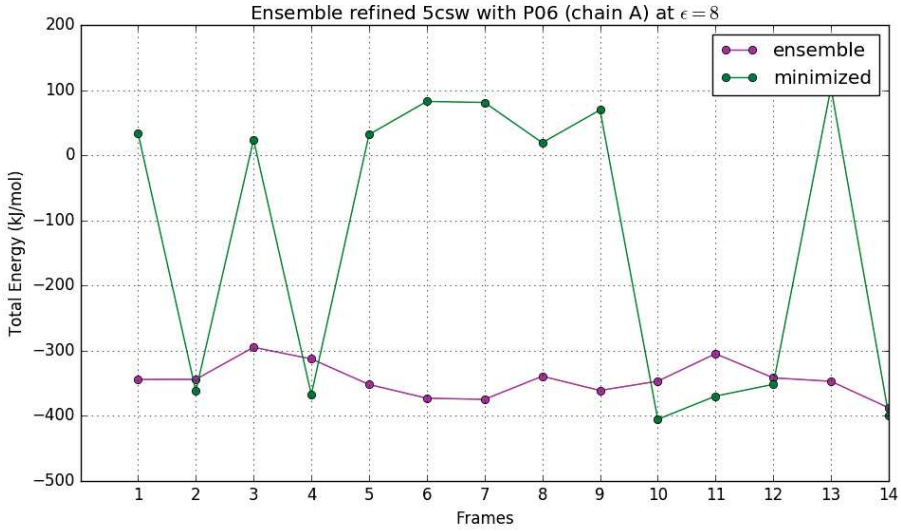


(a) chain A.

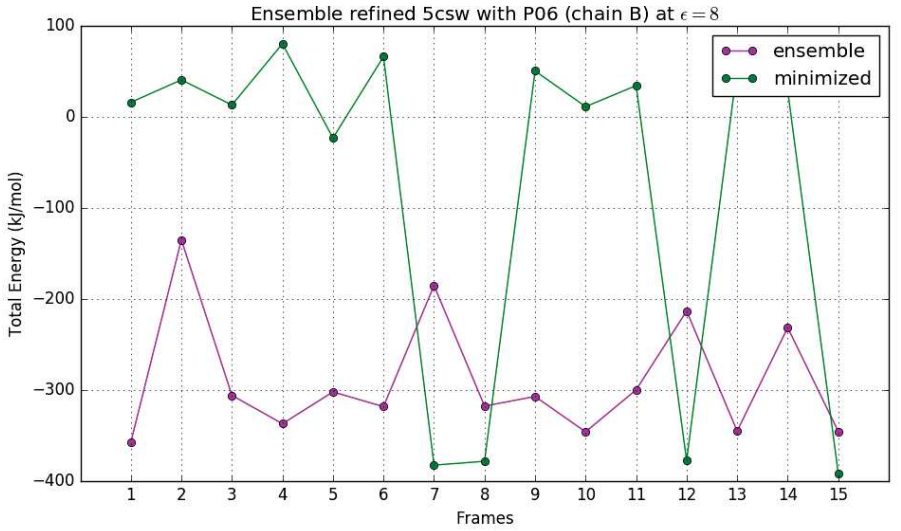


(b) chain B.

Figure 3.36: MM-PBSA results for the crystallographic ensemble-refined BRAF structure 4XV2 with P06.

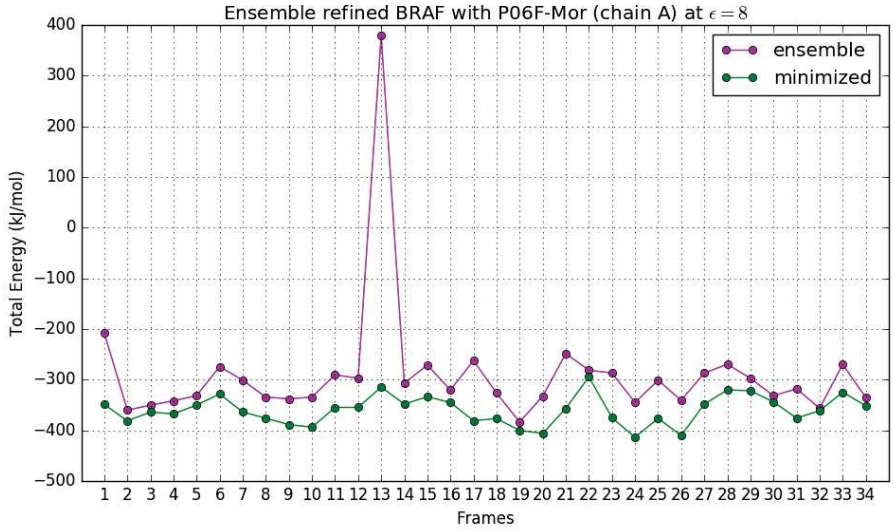


(a) chain A.

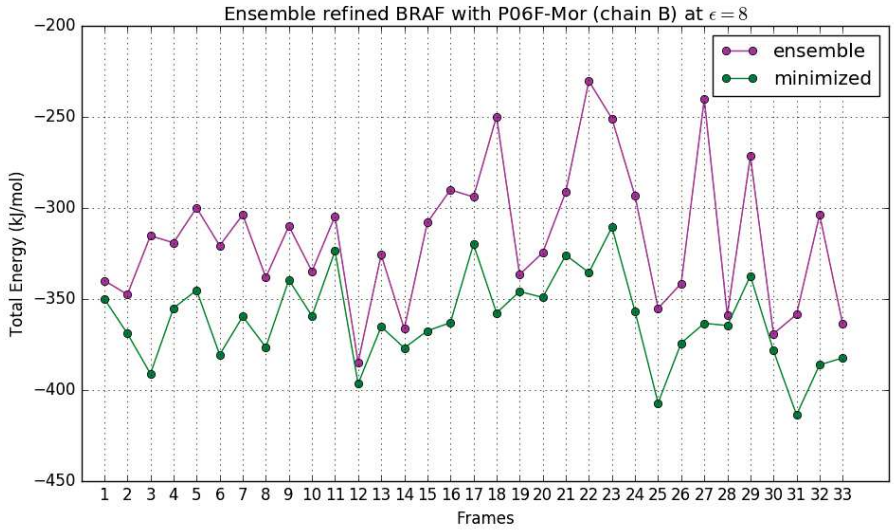


(b) chain B.

Figure 3.37: MM-PBSA results for the crystallographic ensemble-refined BRAF structure 5CSW with P06.



(a) chain A.



(b) chain B.

Figure 3.38: MM-PBSA results for the newly solved crystallographic ensemble-refined BRAF structure with P06F-Mor.

Comparison with MM-PBSA results from MD simulations

For comparison, the MM-PBSA calculations based on 500 snapshots from single 50 ns MD simulations provided the following binding energies (in kJ/mol):

4XV3 (with P02): -384.6 ($\pm$ 22.0)
4XV2 (with P06): -385.7 ($\pm$ 30.8)
5CSW (with P06): -300.5 ($\pm$ 39.7)

The obtained binding energies from minimized ensemble-refined structures and MD simulation snapshots are providing similar value ranges, also concerning the error intervals (compare also with Figure 3.30, 3.31 or 3.32).

Key points

⇒ For the BRAF kinase the conformation of the activation loop and the G-rich loop impact MM-PBSA based binding affinity estimations, shown by differences among four BRAF loop model conformations.

⇒ From the methodological point of view the present study underlines the importance of replica MD simulations for subsequent MM-PBSA calculations and also for direct quantitative investigations, such as hydrogen bond analysis.

⇒ Ensemble-refinement may provide an additional promising tool for generating structural ensembles for structure-based VS and affinity estimations to be further investigated in the near future.

3.8 Experimental work

3.8.1 Activity tests by collaborators

Molecules are synthesized and the affinity on BRAFV600E is tested by AGV discovery. Cellular assays are performed by the team of Patrick Balaguer (IRCM - Institut de Recherche en Cancérologie de Montpellier). For BRAF, proliferation inhibition tests are performed on the cancer cell line A375, and for PXR, induction tests are performed on the cell line HG5LN expressing human PXR.

Kinase activity assay

Upon chemical synthesis, a LanthaScreen kinase activity assay provided by ThermoFisher is performed with BRAFV600E. ThermoFisher explains the assay as following: "In a LanthaScreen kinase activity assay, kinase, fluorescein-labeled substrate, and ATP are allowed to react. Then EDTA (to stop the reaction) and terbium-labeled antibody (to detect phosphorylated product) are added. In a LanthaScreen kinase reaction, the antibody associates with the phosphorylated fluorescein labeled substrate resulting in an increased TR-FRET value. The TR-FRET value is a dimensionless number that is calculated as the ratio of the acceptor (fluorescein) signal to the donor (terbium) signal. The amount of antibody that is bound to the tracer is directly proportional to the amount of phosphorylated substrate present, and in this manner, kinase activity can be detected and measured by an increase in the TR-FRET value."

A375 cells cytotoxicity assays

A375 cell proliferation was assessed using the standard MTT assay as previously described.[168] Briefly, A375 cells were seeded at a density of 500 cells per well in 96-well tissue culture plates and grown in test culture medium. Test compounds were added 24 h after seeding. Cell lines were incubated for 4 days at 37 °C. After the incubation period, the medium containing test compounds was removed and replaced by test culture medium containing 0.4 mg/ml MTT. After incubation (4 h), viable cells cleaved the MTT tetrazolium ring into a dark blue formazan reaction product, whereas dead cells remained colorless. The MTT-containing medium was gently removed and DMSO was added to each well. After shaking, the plates were read in absorbance at 540 nm. Tests were performed in quadriplicate in at least 3 independent experiments. Data were expressed as % of the maximal activity obtained in absence of ligand.

PXR transactivation assays

To characterize the PXR activity, already established HG5LN GAL4-hPXR reporter cell lines[180] were used. In brief, HG5LN cells were obtained by integration of a GAL4-responsive gene (GAL4RE5-bGlob-Luc-SV-Neo) in HeLa cells.[181] The HG5LN GAL4(DBD)-hPXR(LBD) cell line

was obtained by transfecting HG5LN cells with a plasmid [pSG5-GAL4(DBD)-hPXR(LBD)-puro], which enables the expression of the DNA binding domain of the yeast activator GAL4 (Met1–Ser147) fused to the ligand binding domain of hPXR (Met107–Ser434) and confers resistance to puromycin. HG5LN and HG5LN GAL4-hPXR cells were cultured in Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12 (DMEM/F-12) containing phenol red and 1 g/l glucose and supplemented with 5% fetal bovine serum, 100 units/ml of penicillin, 100 μg/ml of streptomycin and 1 mg/ml geneticinin at 5% CO2 humidified atmosphere at 37 °C. HG5LN GAL4-hPXR cells were cultured in the same medium supplemented with 0.5 μg/ml puromycin. For transactivation experiments, HG5LN and HG5LN-PXR were seeded at a density of 25,000 cells per well in 96-well white opaque tissue culture plates (Greiner CellStar) in Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12 (DMEM/F-12) without phenol red and 1 g/l glucose and supplemented with 5% stripped fetal bovine serum, 100 units/ml of penicillin, 100 μg/ml of streptomycin (test medium). Compounds to be tested were added 24 h later, and cells were incubated at 37 °C for 16 h. At the end of the incubation period, culture medium was replaced with test medium containing 0.3 mM luciferin. Luciferase activity was measured for 2 s in intact living cells using a MicroBeta Wallac luminometer (PerkinElmer). Tests were performed in quadriplicate in at least 3 independent experiments. Data were expressed as % of the maximal activity obtained in absence of ligand (HG5LN cells) or with SR12813 3 μM (HG5LN PXR cells).

3.8.1.1 Results of the drug design rounds

Round 1

The only compound tested (P02C-4Pi) was not completely inhibiting (84%) at 1 μM (single measure on BRAF-V600E). Therefore, the P02/P02C scaffold was not further pursued.

Round 2

Molecules 1-4 from synthesis round 2 could be tested (see Table 3.5), but unfortunately, the molecules 5-8 with the pyrimidine moiety extended by an additional acetyl were not stable.

molecule	P06F-CPP	P06F-Pip	P06F-Mor	P06F-PrA
IC50 [nM]	4.40	4.52	5.98	6.27

molecule	P06F-CPP-ac	P06F-Pip-ac	P06F-Mor-ac	P06F-PrA-ac
IC50 [nM]	-	-	-	-

Table 3.5: IC50 affinity measurements for the synthesized drug candidates from synthesis round 2 on BRAF-V600E.

In agreement with the predictions by machine learning (see Section 3.5.2), the various substitutions with the P06F scaffold yielded compounds with affinity in the low nanomolar range. Two compounds (P06F-CPP and P06F-Mor) were also tested successfully against the A375 cell line with

activity ranging from 53% to 79% relative to dabrafenib. This suggests that the cell permeability was not affected, which was expected from the nature of their substituent (cyclopropylpiperidine and morpholine). In parallel, assays on hPXR reporter cells highlighted an intensively decreased activation of PXR compared to dabrafenib (by a factor of 15 and 555 for P06F-CPP and P06F-Mor, respectively).

Round 3

For synthesis round 3, two compounds of the eight designed ones were finally synthesized and tested: P06FCl and P06FCl-Mor.

molecule	P06F	P06FF	P06FCl	P06FFCl
IC50 [nM]	-	-	3.61	-

molecule	P06F-Mor	P06FF-Mor	P06FCl-Mor	P06FFCl-Mor
IC50 [nM]	-	-	2.01	-

Table 3.6: IC50 affinity measurements for the synthesized drug candidates from synthesis round 3 on BRAF-V600E.

The results listed in table 3.6 show a small increase in affinity for molecules with the additional chlorine atom (scaffold P06FCl) compared to the previous synthesis rounds, which is in agreement with predictions by both, machine learning (see Section 3.5.2) and MM-PBSA calculations (see Section 3.7.2). The cell activity was maintained at ∼70% compared to dabrafenib for P06FCl and P06FCl-Mor. Also in agreement with our predictions, the cellular activity against PXR was only marginally lower (by a factor of 2) for P06FCl (containing the original tertiary butyl moiety) and largely decreased (by a factor of 417) for P06FCl-Mor. Indeed, the chlorine atom is rather well accommodated in the PXR pocket.

The two compounds were additionally measured with BRAF-WT giving the following IC50 values: P06FCl 1.71 nM and P06FCl-Mor 1.65 nM.

Round 4

In synthesis round 4, four molecules were built by adding new substituents - azetidine (2Az), pyrrolidine (2Py), piperidine (2Pi), piperazine (2PA) - instead of the morpholine group to the P06F scaffold. The selection was inspired by the newly solved structure of BRAF with P06F-Mor.

molecule	P06F-2Az	P06F-2Py	P06F-2Pi	P06F-2PA
IC50 [nM]	2.10	1.75	1.73	5.92

Table 3.7: IC50 affinity measurements for the synthesized drug candidates from synthesis round 4 on BRAF-V600E.

In agreement with our crystal structure (see Section 3.9 below), the substitution of the morpholine with different heterocycles still harboring a nitrogen atom (in order to maintain a hydrogen bond to

the aspartate D594) improved the affinity toward purified BRAF in three of four cases (as shown in Table 3.7). Only the piperazine containing molecule (P06F-2PA) showed the same affinity as the morpholine derivative (P06F-Mor). Unfortunately, most of these compounds are affected in the cell permeability. They also seemed to still activate PXR.

3.8.2 BRAF crystallogenesis

Expression and purification

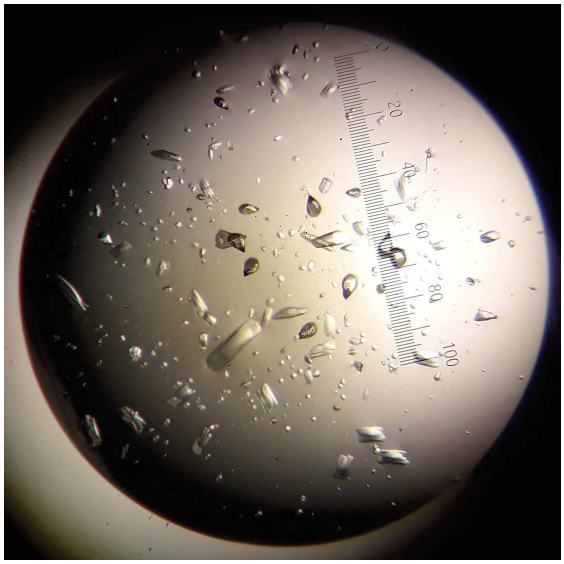
A Pet28a(+) vector with DNA encoding the BRAF kinase domain residues 448-723 containing the V600E mutation, 16 solubilizing mutations (I543A, I544S, I551K, Q562R, L588N, K630S, F667E, Y673S, A688R, L706S, Q709R, S713E, L716E, S720E, P722S, and K723G - permitting kinase domain overexpression in bacteria), encoding as well an N-terminal His tag, and a thrombin cleavage site between the protein and the His tag, was provided by Dr. Michael Grasso (Marmorstein Lab, Department of Chemistry, University of Pennsylvania, Philadelphia, USA). The protein was expressed in *E.coli* (Stellar) cells, with an overnight pre-culture at 37°C in LB, followed by 6h at 37°C and an overnight incubation at 25°C on an auto-inductive kanamycin medium, spun down the next day, lysed in lysis buffer (buffer A supplemented with lysozyme), frozen, thawed, and sonicated. The lysate was then spun down at 18 000 rpm, and the supernatant was incubated on a His-trap nickel column at 4°C for 1h. The supernatant was then eluted, the column washed with buffer A, and the BRAF proteins eluted with buffer B (wich is buffer A supplemented with 300 mM imidazole). Protein was then dialyzed into buffer C (wich does not contain imidazole) and applied to a 16/60 Superdex 75 gel filtration column in a final buffer D. Protein was frozen and stored for future use. Used buffers are:

- Buffer A (50 mM Tris, pH 7.0, 250 mM NaCl, 5% glycerol, 2 mM β-Mercaptoethanol),
- Buffer B (Buffer A and 300 mM imidazole),
- Buffer C (25 mM Tris, pH 7.0, 75 mM NaCl, 5% glycerol, 1 mM EDTA, 10 mM dithiothreitol (DTT)),
- Buffer D (20 mM HEPES at pH 7.0, 150 mM NaCl, 5% glycerol, 10 mM dithiothreitol (DTT)).

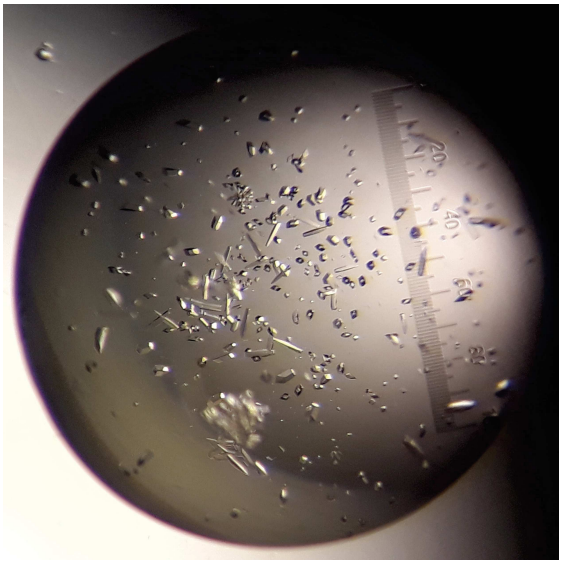
Crystallization and data collection

BRAFV600E-16M at 9 mg/mL was mixed with 10% of a 10 mM DMSO inhibitor solution, and after initial screenings using the commercial kits "PACT", "PEGs-I", and "PEGs-II" from QIAGEN on 96-Well plates, trays were set up screening around a crystallization condition of 100 mM BisTrisPropane at pH 7.0-8.0, 20% PEG monomethyl ether 2000/3350/4000, and 100-350 mM Na-formate using the hanging-drop vapor diffusion method at 18°C. Crystal formation took $\sim$14 days, resulting in maximal crystal sizes of $\sim$50x100 µm (see Figure 3.39). Crystals were flash frozen in liquid nitrogen. X-ray diffraction data was collected at a wavelength of 0.979 Å and a beam size fitted to the crystal dimensions (adjustable between 50-300(H) x 6-100(V) μm^2) at the synchrotron ALBA (Barcelona, Spain), at beamline BL13 - XALOC. Finally, a diffraction dataset was obtained

from a crystal from the drop shown in Figure 3.39b.



(a) Crystallization condition of
100 mM BisTrisPropane at pH 8.0,
20% PEG 3350, and 250 mM Na-formate.



(b) Crystallization condition of
100 mM BisTrisPropane at pH 8.0,
20% PEG 3350, and 300 mM Na-formate.

Figure 3.39: Crystals of BRAFV600E protein with the designed ligand P06F-Mor.

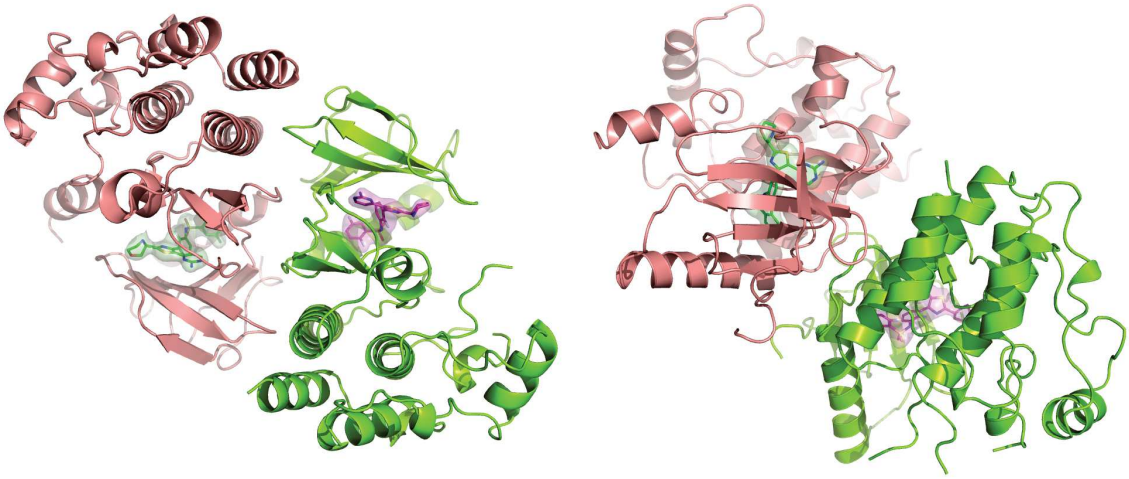
3.8.3 Crystallographic structure determination

The structure was determined by molecular replacement in PHENIX[182] using Phaser using PDB 5ITA as a search model. The molecular replacement search model was used as monomer and had its ligand removed. Model building and refinement were performed using Coot[183] and PHENIX. NCS was used, as two BRAF monomers were present in the asymmetric unit. The CIF file for the inhibitor was generated using the Grade Web Server (at `http://grade.globalphasing.org`). The atomic dimeric structure is refined to a final resolution of 2.37 Å after uncovering and subsequent modelling of a domain swap.

3.9 Crystal structure analysis

The classical BRAF dimer interface ("back-to-back")

As seen in diverse crystallographic BRAF structures, the newly resolved structure also presents the typical "back-to-back" interface with a second monomer in a symmetric unit cell (visualized through Coot by displaying the symmetric molecule, see Figure 3.40).

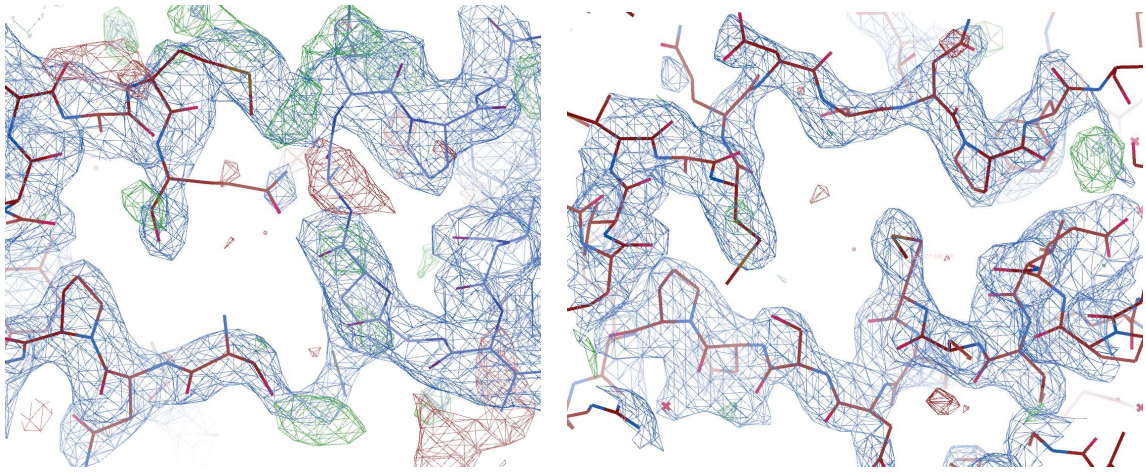


(a) side view **(b)** bottom view

Figure 3.40: The classical BRAF dimer ("back-to-back").

The new BRAF domain swap dimer - a mutual embrace

Interestingly, during refinement it became apparent that the structure forms also another dimeric interface. It can be described as "face-to-face" interface (in contrast to the "back-to-back" interface of the classical BRAF dimer). In this interface the two protomer partners are highly interwoven and even perform a domain swap with one another. The atomic models along with the electron density difference maps ($2F_{obs} - F_{calc}$ and $F_{obs} - F_{calc}$) are shown before and after the domain swap has been modelled in Figure 3.41. After several refinement steps, but before modelling the domain swap there were still several discrepancies apparent between the model and the observed data (as shown by the $F_{obs} - F_{calc}$ difference map in Figure 3.41a) and the $2F_{obs} - F_{calc}$ density map indicated a continuation of the structure from one protomer to the neighboring one. After modelling the domain swap most discrepancies disappeared and the shape of the electron density became more defined, clearly tracing the path of the atomic structure (see Figure 3.41b).



(a) before modelling of the domain swap **(b)** after modelling of the domain swap

Figure 3.41: Modelled atomic structures within the respective electron density difference maps ($2F_{obs} - F_{calc}$ in blue and $F_{obs} - F_{calc}$ in green for positive density and red for negative density, contoured at 1.0σ).

Upon modelling of the domain swap the activation loop protrudes far into the neighboring protomer within the C-lobe (see Figure 3.42). Additionally, the αC helices are having a large contact area with each other (see Figure 3.43) and even the G-rich loop conformation is largely impacted by the dimerization. However, the ligand is not in direct contact with any residue of the partner protomer.

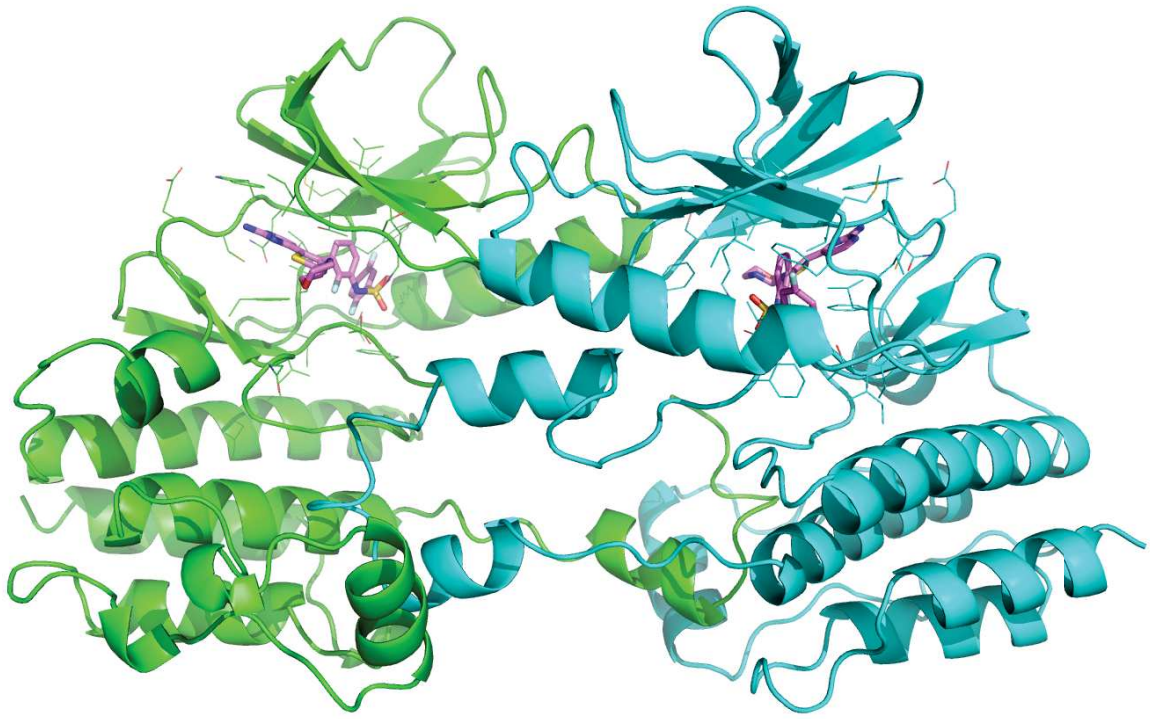


Figure 3.42: The refined crystallographic BRAFV600E structure (chain A in green and chain B in cyan) with the designed ligand P06F-Mor (violet). The dimeric structure shows a domain swap of the activation loop. Residues within a 5 Å distance of any ligand atom are shown in line representation.

In the newly resolved structure the activation loop adopts a partially structured helical conformation (compare with the completely resolved activation loop of chain B). Stating from Trp604 there is a clear

α-helix until Gln612, followed by a helical structure that forms a turn, and again a clear α-helix from Pro622 to Gln628. For chain A residues 598-613 are not resolved, but the residue stretches Ser614 to Ala621 are also forming a helical turn followed by a clear α-helix from Pro622 to Asp629. The latter α-helix is swapped with the partner protomer and located at the exact same position compared to a non-swap conformation (see Figure 3.42).

BRAF domain swap dimer interface ("face-to-face")

The interface of the crystallographic BRAF domain swap dimer is analyzed visually in PyMol (see Figure 3.43) and further explored by using the PDBePISA web server (see Figure 3.44).

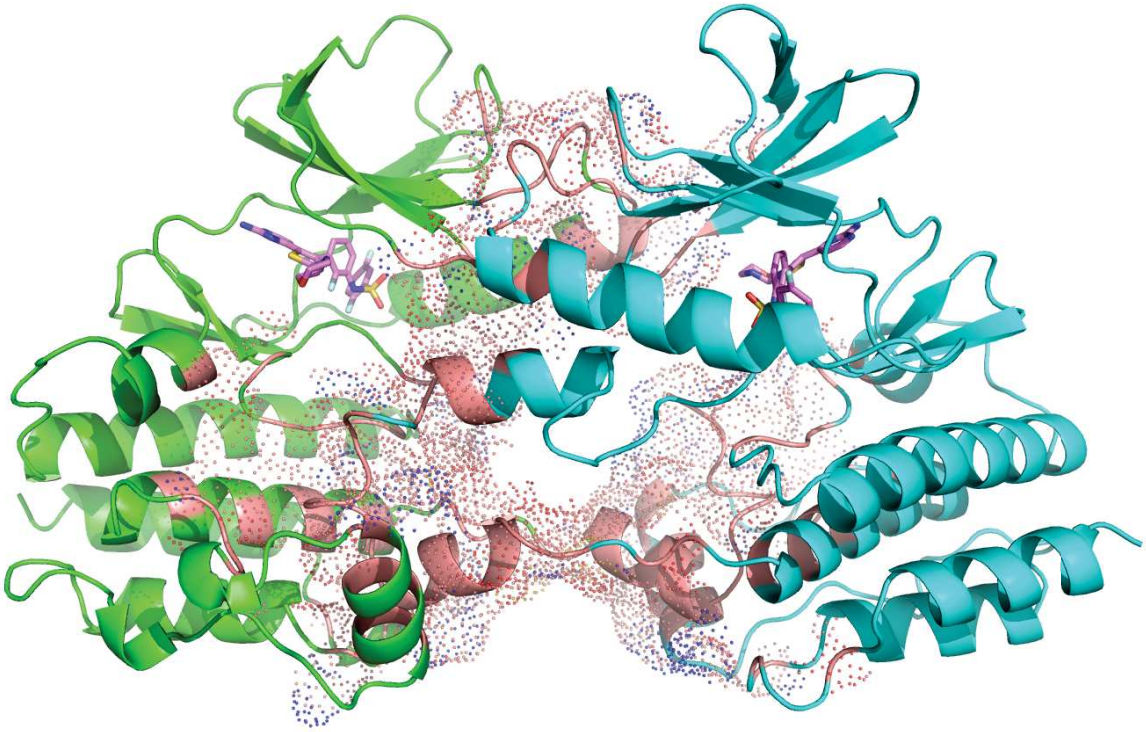


Figure 3.43: The refined crystallographic BRAFV600E structure (chain A in green and chain B in cyan) with the designed ligand P06F-Mor (violet). The interface of the two protomers (as identified by the PyMol tool *interfaceResidue*) is highlighted in rose with its surface in dot representation.

PDBePISA (Proteins, Interfaces, Structures and Assemblies)[184] is an interactive tool for the exploration of macromolecular interfaces. It reports on structural and chemical properties of macromolecular surfaces and interfaces. It is used here, to evaluate the interface between the two protomers A and B of the newly resolved crystallographic BRAFV600E structure (see Figure 3.44). A rather large portion of the total solvent-accessible area of the two chains is identified as interface (15.6% and 15.8%), and the portion of number of residues is even 23.9%. Furthermore, 32 hydrogen bonds and 9 salt bridges are established between the two chains in the crystal structure. Thus, the interface is estimated to be stable in solution.

interface #1/10

Interface Summary XML

	Structure 1		Structure 2	
Selection range	B		A	
class	Protein		Protein	
symmetry operation	x,y,z		x,y,z	
symmetry ID	1_555		0_555	
Number of atoms				
interface	239	11.0%	223	10.8%
surface	1313	60.2%	1265	61.1%
total	2180	100.0%	2072	100.0%
Number of residues				
interface	65	23.9%	62	23.9%
surface	250	91.9%	242	93.4%
total	272	100.0%	259	100.0%
Solvent-accessible area, Å				
interface	2349.9	15.6%	2342.4	15.8%
total	15025.8	100.0%	14867.1	100.0%
Solvation energy, kcal/mol				
isolated structure	-252.4	100.0%	-243.3	100.0%
gain on complex formation	-11.6	4.6%	-14.5	6.0%
average gain	-4.2	1.7%	-4.5	1.8%
P-value	0.068		0.017	

Hydrogen bonds XML

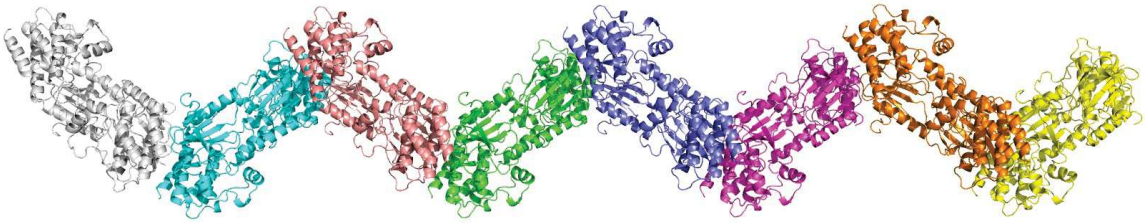
##	Structure 1	Dist. [Å]	Structure 2
1	B:GLN 494[NE2]	2.95	A:SER 467[OG]
2	B:THR 488[N]	3.59	A:ASN 486[O]
3	B:THR 488[N]	2.78	A:ASN 486[OD1]
4	B:ALA 489[N]	2.97	A:ASN 486[OD1]
5	B:ASN 486[ND2]	3.34	A:THR 488[O]
6	B:THR 488[OG1]	3.02	A:THR 488[OG1]
7	B:ASN 486[ND2]	3.46	A:ALA 489[O]
8	B:GLY 466[N]	3.04	A:GLN 493[OE1]
9	B:ASN 486[ND2]	2.96	A:GLN 494[OE1]
10	B:SER 616[OG]	2.61	A:ASP 576[OD1]
11	B:ARG 662[NH1]	2.55	A:SER 614[O]
12	B:LYS 578[NZ]	3.67	A:GLY 615[O]
13	B:LYS 578[NZ]	2.89	A:SER 616[OG]
14	B:ARG 704[NH2]	3.05	A:GLU 623[OE1]
15	B:ARG 704[NH1]	2.48	A:GLU 623[OE2]
16	B:TRP 619[NE1]	2.61	A:GLU 648[OE1]
17	B:ARG 626[NH1]	2.96	A:GLY 670[O]
18	B:GLY 466[O]	2.98	A:THR 491[OG1]
19	B:SER 467[O]	2.92	A:GLN 494[NE2]
20	B:ASN 486[O]	2.86	A:THR 488[N]
21	B:ASN 486[OD1]	3.22	A:GLN 494[NE2]
22	B:THR 488[OG1]	3.36	A:THR 488[N]
23	B:THR 488[OG1]	3.12	A:GLN 524[NE2]
24	B:ALA 489[O]	2.99	A:ASN 486[ND2]
25	B:ASP 576[OD1]	2.61	A:SER 616[OG]
26	B:GLY 615[O]	3.20	A:LYS 578[NZ]
27	B:SER 616[OG]	2.57	A:LYS 578[NZ]
28	B:GLU 623[OE1]	2.79	A:ARG 704[NH2]
29	B:GLU 623[OE2]	3.83	A:ARG 701[NE]
30	B:GLU 623[OE2]	3.03	A:ARG 704[NH1]
31	B:GLU 648[OE2]	2.81	A:TRP 619[NE1]
32	B:GLY 670[O]	2.90	A:ARG 626[NH2]

Salt bridges XML

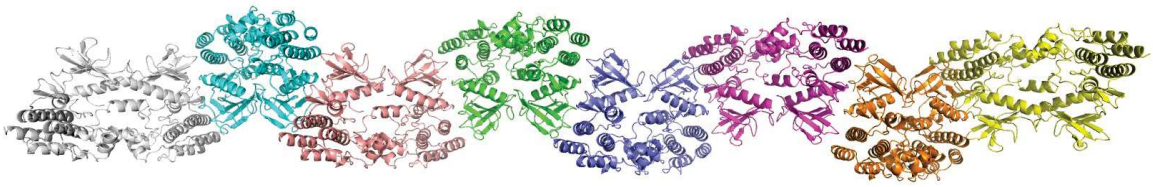
##	Structure 1	Dist. [Å]	Structure 2
1	B:ARG 704[NH1]	3.60	A:GLU 623[OE1]
2	B:ARG 704[NH2]	3.05	A:GLU 623[OE1]
3	B:ARG 704[NH1]	2.48	A:GLU 623[OE2]
4	B:ARG 704[NH2]	3.41	A:GLU 623[OE2]
5	B:GLU 623[OE1]	2.79	A:ARG 704[NH2]
6	B:GLU 623[OE1]	3.74	A:ARG 704[NH1]
7	B:GLU 623[OE2]	3.83	A:ARG 701[NE]
8	B:GLU 623[OE2]	3.62	A:ARG 704[NH2]
9	B:GLU 623[OE2]	3.03	A:ARG 704[NH1]

Figure 3.44: PDBePISA[184] server output of the refined crystallographic BRAFV600E structure with the designed ligand P06F-Mor. The interface of the two protomers A and B is evaluated concerning size and electrostatic bonding.

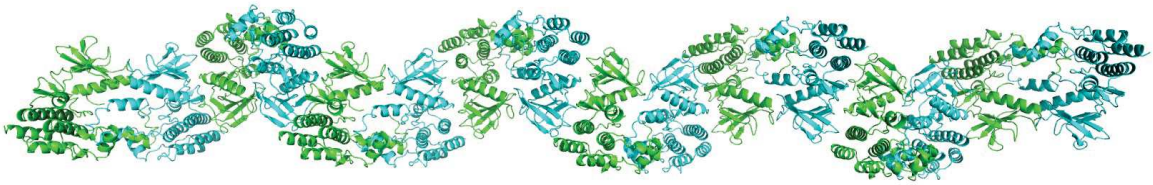
As both dimer interfaces (the classical "back-to-back" and the new "face-to-face") are large in size and are expected to be stable in solution, a fiber-like macromolecular structure is expected (see Figure 3.45).



(a) top view, colored by asymmetric unit present in the crystal ("face-to-face" dimer)



(b) side view, colored by asymmetric unit present in the crystal ("face-to-face" dimer)



(c) side view, colored by protomer (chain A in green and chain B in cyan)

Figure 3.45: The refined crystallographic BRAFV600E structure with the symmetric unit cell replicated in one axis, as present in the protein crystal (generated using Coot by displaying the symmetric molecules and visualized in PyMol).

BRAF dimer flexibility

Visualization of crystallographic b-factors in Figure 3.46 (range: 21.42 - 106.55 Å) shows differences between the two protomers.

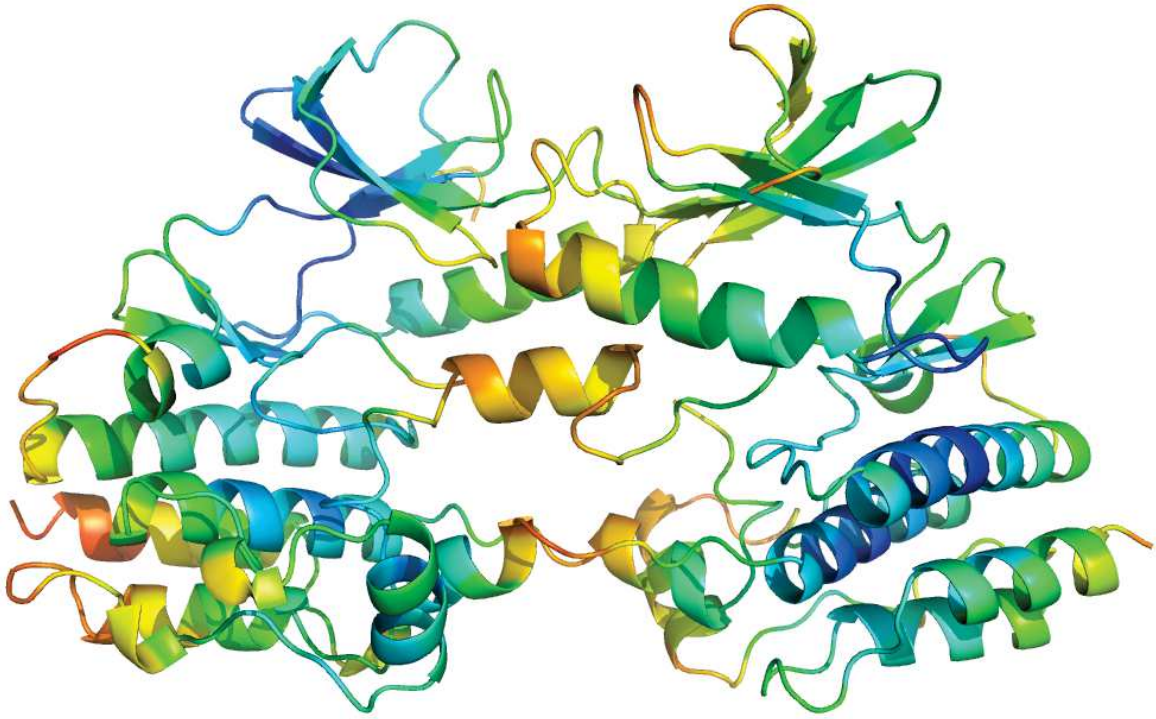
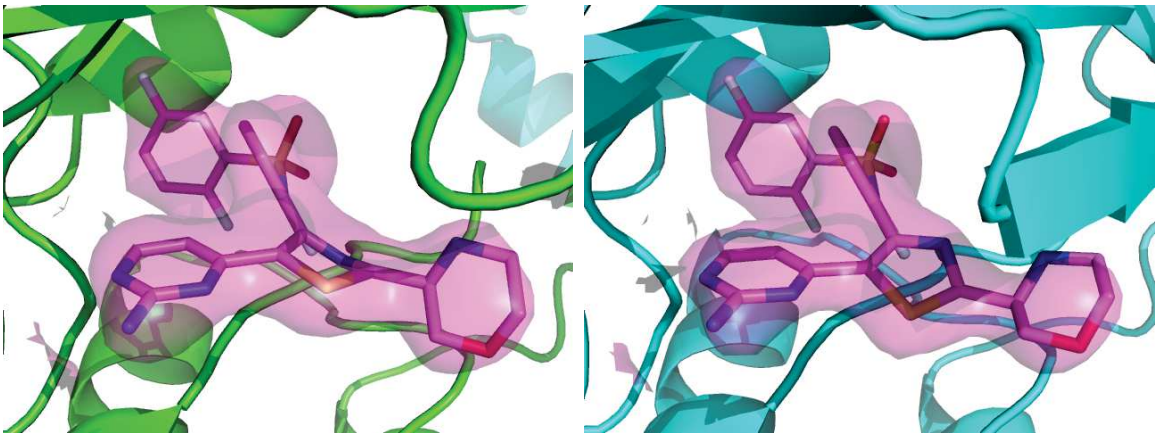


Figure 3.46: The refined crystallographic BRAFV600E structure colored by b-factor with a range of 21.42 - 106.55 Å (rainbow: blue to red).

Ligand binding mode

The electron density map for the ligands are clear in both protomers and permit an accurate positioning of the molecules (see Figure 3.47).



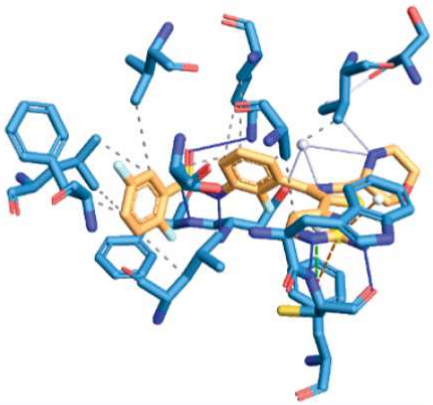
(a) in chain A **(b)** in chain B

Figure 3.47: Electron density and atomic structures of the designed drug P06F-Mor in the two protomers (chain A and B) of the crystal structure.

Molecular interactions detected by the PLIP webserver[172] of the ligand P06F-Mor with chain A (see Figure 3.48) and chain B (see Figure 3.49) are significantly different. Whereas in both binding sites a rather large network of hydrophobic interactions is established, only in protomer A the water molecule present in the binding site, forming three water bridges between protein and ligand, is detected. Nonetheless, the water molecule is present in both protomers (compare Figure 3.50). Remarkably, the nitrogen of the morpholine moiety takes part in the hydrogen bond network formation with the water molecule and the residues Lys483 and Asp594 (D of the DFG motif). Additionally, in protomer A the formation of a π-stacking of the ligands hinge binding pyrimidine moiety and Phe583 is detected by PLIP.

Hydrophobic Interactions ····

Index	Residue	AA	Distance	Ligand Atom	Protein Atom
1	471A	VAL	3.87	4277	188
2	481A	ALA	3.76	4270	271
3	483A	LYS	3.94	4279	285
4	483A	LYS	3.85	4278	283
5	505A	LEU	3.95	4264	454
6	505A	LEU	3.56	4266	453
7	514A	LEU	3.63	4263	530
8	516A	PHE	3.81	4264	546
9	527A	ILE	3.94	4266	634
10	529A	THR	3.61	4278	649



Hydrogen Bonds —

Index	Residue	AA	Distance H-A	Distance D-A	Donor Angle	Protein donor?	Sidechain	Donor Atom	Acceptor Atom
1	483A	LYS	2.89	3.62	129.55	✔	✔	287 [N3]	4289 [O2]
2	532A	CYS	2.02	2.97	160.81	✔	✖	673 [Nam]	4284 [N2]
3	532A	CYS	2.05	3.00	161.68	✖	✖	4285 [Npl]	676 [O2]
4	594A	ASP	2.11	2.86	132.54	✔	✖	1175 [Nam]	4283 [Npl]
5	595A	PHE	2.21	3.03	140.29	✔	✖	1183 [Nam]	4288 [O2]

Water Bridges —

Index	Residue	AA	Dist. A-W	Dist. D-W	Donor Angle	Water Angle	Protein donor?	Donor Atom	Acceptor Atom	Water Atom
1	465A	SER	2.98	4.01	149.10	72.62	✖	4287 [N3]	148 [O2]	4385
2	594A	ASP	3.43	2.94	104.55	109.99	✔	1182 [O3]	4255 [N2]	4365
3	594A	ASP	3.80	2.94	104.55	75.05	✔	1182 [O3]	4287 [N3]	4365

π-Stacking ···· ····

Index	Residue	AA	Distance	Angle	Offset	Type	Ligand Atoms
1	583A	PHE	4.65	25.24	0.94	P	4268, 4270, 4271, 4272, 4284, 4286

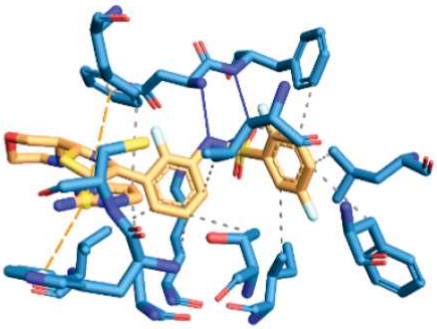
π-Cation Interactions ····

Index	Residue	AA	Distance	Offset	Protein charged?	Ligand Group	Ligand Atoms
1	531A	TRP	3.94	1.82	✖	guanidine	4284, 4285, 4286
2	583A	PHE	5.08	0.76	✖	guanidine	4284, 4285, 4286

Figure 3.48: PLIP interactions of designed drug P06F-Mor with chain A in its crystal structure.

Hydrophobic Interactions ····

Index	Residue	AA	Distance	Ligand Atom	Protein Atom
1	471B	VAL	3.86	4314	2251
2	481B	ALA	3.63	4307	2333
3	483B	LYS	3.60	4315	2345
4	483B	LYS	3.58	4316	2347
5	505B	LEU	3.65	4301	2516
6	514B	LEU	3.80	4316	2595
7	514B	LEU	3.59	4300	2592
8	516B	PHE	3.81	4301	2608
9	527B	ILE	3.75	4303	2696
10	529B	THR	3.90	4315	2711
11	583B	PHE	3.96	4307	3157
12	595B	PHE	3.67	4300	3254



Protein
Ligand
Water
Charge Center
Aromatic Ring Center
Metal Ion
Hydrophobic Interaction
Hydrogen Bond
Water Bridge
π-Stacking (parallel)
π-Stacking (perpendicular)
π-Cation Interaction
Halogen Bond
Salt Bridge
Metal Complexation

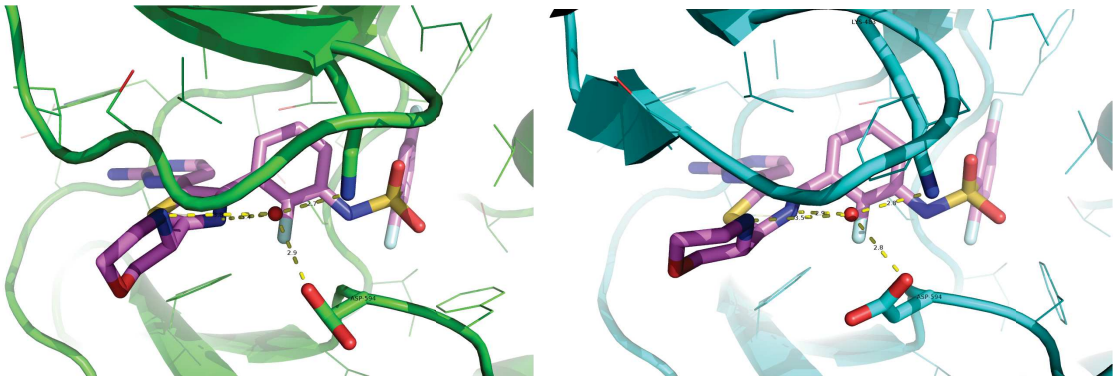
Hydrogen Bonds —

Index	Residue	AA	Distance H-A	Distance D-A	Donor Angle	Protein donor?	Sidechain	Donor Atom	Acceptor Atom
1	483B	LYS	3.28	4.07	135.94	✔	✔	2349 [N3]	4325 [O2]
2	532B	CYS	2.22	3.13	153.39	✔	✖	2735 [Nam]	4321 [N2]
3	594B	ASP	2.02	2.83	138.47	✔	✖	3237 [Nam]	4320 [Npl]
4	595B	PHE	2.28	3.17	149.97	✔	✖	3245 [Nam]	4325 [O2]

π-Cation Interactions ····

Index	Residue	AA	Distance	Offset	Protein charged?	Ligand Group	Ligand Atoms
1	531B	TRP	3.99	1.90	✖	guanidine	4322, 4323, 4321
2	583B	PHE	5.33	1.36	✖	guanidine	4322, 4323, 4321

Figure 3.49: PLIP interactions of designed drug P06F-Mor with chain B in its crystal structure.

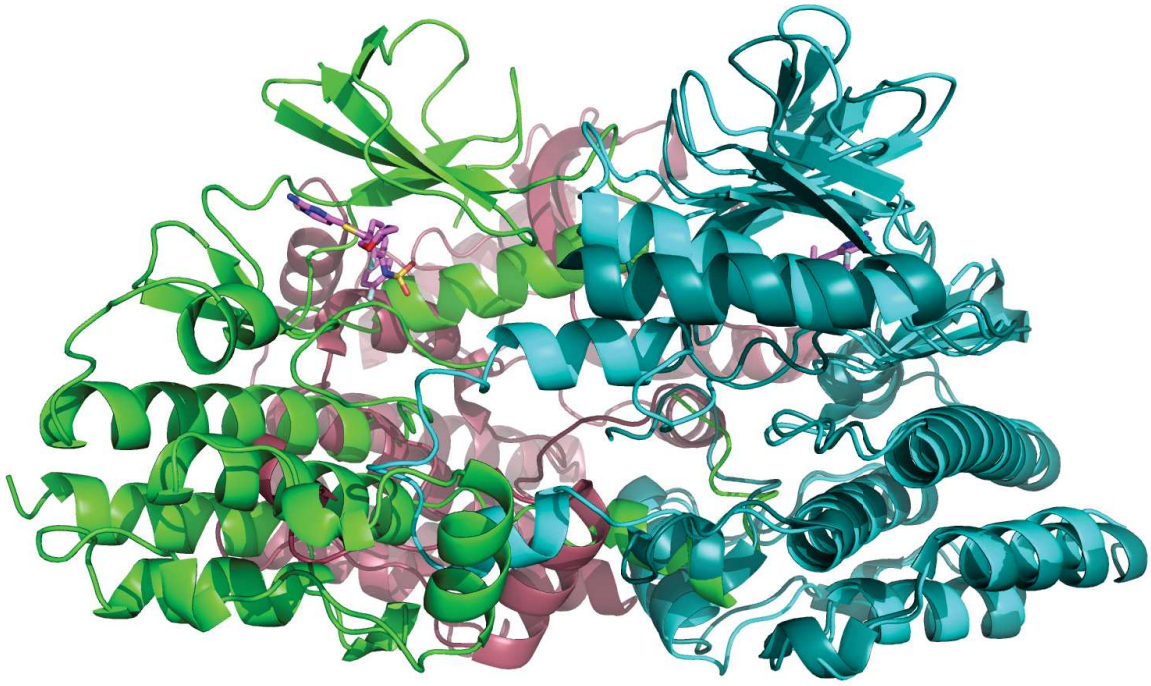


(a) in chain A **(b)** in chain B

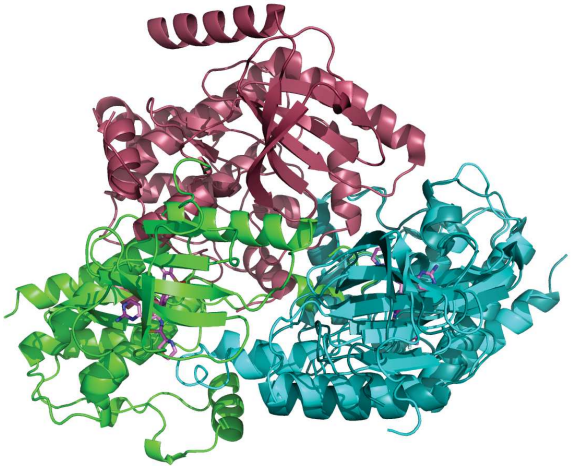
Figure 3.50: Hydrogen bond network (yellow dashed lines) that is presumably established between the designed drug P06F-Mor (in purple), a water molecule (red sphere) present in the binding site and the residues Lys483 and Asp594 (in stick representation) in the two protomers (chain A and B) of the crystal structure.

BRAF domain swap dimer - impact on signalling

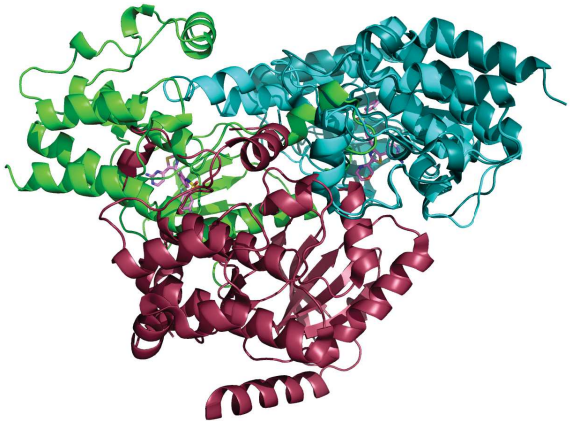
A possible impact of the new "face-to-face" domain-swap dimer on the downstream signalling cascade, especially the activation of the protein kinase MEK (being a direct substrate) was investigated by structural alignment of the new dimer structure with the crystallographic complex of BRAF and MEK (PDB-ID: 6U2G). Apparently, the structure of the crystallographic BRAF domain-swap dimer is not compatible with MEK binding, as the MEK binding site overlaps partially with the second BRAF domain-swap protomer (see Figure 3.51).



(a) side view



(b) top view



(c) bottom view

Figure 3.51: The refined crystallographic BRAFV600E structure with the designed ligand P06F-Mor as domain swap dimer (chain A in green and chain B in cyan, as in Figure 3.42) and crystallographic structure 6U2G containing BRAF (chain B, in dark cyan) and MEK (chain A, in bordeaux), whereas the respective B chains (cyan and dark cyan) are superimposed.

Key points

⇒ The binding mode of the designed drug candidate P06F-Mor in the crystal structure confirms the predicted binding mode (by docking), but an expected hydrogen bond to the Asp of the DFG motif is not formed. (This result inspired synthesis round 4.)

⇒ An original dimeric domain swapping conformation is detected (that was previously unseen and therefore unpredictable), which could potentially lead to fiber formation.

4

Conclusions and Perspectives

This chapter summarizes and concludes on the employed techniques and obtained results within this thesis work and discusses general current issues and future trends within the area of drug design that have been encountered by performing the presented work.

4.1 Summary, discussion and conclusions

In this PhD project, one aim is to understand the effect of differently generated structural ensembles for ligand screening and binding affinity estimation under different conditions in terms of target flexibility. Another aim is the rational design of modified drug candidates by taking into account the primary target and an unwanted secondary target that both represent distinct biological systems.

4.1.1 The biological systems

This thesis presents detailed investigations on two different nuclear receptors and a protein kinase:

- a nuclear receptor - **ER**α - being a primary target and secondary target simultaneously,
- another nuclear receptor - **PXR** - being a secondary target by function/biological role, and
- a protein kinase - **BRAF** - being the primary target for rational drug design (while simultaneously avoiding secondary targets, such as PXR).

For ERα a broad characterization is provided by employing several computational techniques, highlighting the different levels of flexibility, which are important for the proteins functionality. On PXR focused structural studies provide routes for avoiding drug binding, and for BRAF a combination of methods provides both, a broad overview of the protein's conformational flexibility and detailed information for targeted drug design.

The presented PhD project shows how the protein's nature and the available amount of data influence or even dictate the tools that can be used for investigations and the degree of exploitation and instrumentalization that can be attained.
For all three targets there are crystallographic structures available, as well as sets of ligands with associated experimental affinity measures. Nonetheless, the quantity differs largely. For ERα's LBD there are currently 260 PDB entries, for PXR's LBD 23, and for BRAF's kinase domain 65. Concerning tested ligands, for ERα there are 281 ligands with K_i affinity measures and 1641 with IC50 measures available in BindingDB, for PXR 47 with IC50 values, and for BRAFV600E 2193 with IC50 values. The availability of crystal structures permits the employment of MD simulations and the refinement of crystal structures as ensemble (Sections 2.2 and 3.2). Based on the MD simulations on BRAF and PXR together with diverse drug derivatives MM-PBSA affinity calculations are performed (Sections 3.7.1 and 3.7.2). The ensembles of available crystal structures enables the analysis of the structural variability of the proteins on different levels, ranging from side-chain mobility to reorientation of whole loops and secondary structures (Section 2.2 for ERα and 3.2 for PXR and BRAF). The abundant data for ERα and BRAF further allows for the development of machine learning models dedicated to affinity prediction (Sections 2.1 and 3.5.2, respectively).

The designed drugs and perspectives for cancer treatment of BRAFV600 mutants

A result of this thesis work is the proposal of designed drug candidates that bind to the primary target BRAF with high affinities and avoid binding to the unwanted secondary target PXR. The procedure was based on iterative design, computation, synthesis and testing rounds. Concerning the testing (by collaborators), affinities were measured *in vitro* through kinase activity assays and efficacy was tested through A375-cell cytotoxicity assays and PXR transactivation assays (Section 3.8). Moreover, the atomic structure of one designed drug (P06F-Mor) co-crystallized with the BRAF-V600E protein kinase was solved at a resolution of 2.37 Å and a clear electron density for the ligand (in both protomers of the dimeric structure) was obtained, confirming the expected binding mode and allowing for precise analysis.
Based on the presented results, potential future patients should benefit from largely reduced side effects, as avoiding PXR should prevent the induction of cytochrome P450. Additionally, a decrease of the previously fast drug metabolism is expected due to the alteration of the initial drugs site of metabolic modification. This could lead to an increased bioavailability of the drug, requiring potentially a lower administration dose, which in turn should additionally reduce side effects (due to reduced off-target binding) and may also reduce the rate of acquired resistance and relapse.
In the future, antagonists of BRAF mutants (without the paradoxical activation of the wild type) may be the cornerstone in the treatment of many cases of what has up until now been untreatable metastatic.

From single-target to multi-target approaches

Currently, one of the main approaches in drug discovery is the development of target-specific inhibitors with high-fold potency and selectivity towards one isoform or one specific mutant. For the oncogenic mutant BRAF-V600E this is a requirement in order to avoid harming healthy cells through inhibition of the BRAF-WT or other protein kinases and reducing severe side effects. Nevertheless, this is a reductionism approach that could or should be extended, as organisms can affect a drug's effectiveness through compensatory ways. For instance, cancer is a complicated disease, affecting several pathways and moreover, many patients develop resistance to drugs via different ways. This promotes a multi-targeted therapy as a sometimes more promising approach to achieve the desired treatment.[14] The transition from the one-drug-one-target model to a multiple-target approach is gaining momentum within the area of drug development[185] and becomes a highly interdisciplinary task including fields such as systems biology[186] and chemogenomics,[185,187] which already by itself combines chemoinformatics and bioinformatics in an interdisciplinary field. This thesis project aims also for a broader view than the one-drug-one-target model, as not only the primary target BRAF, but also the secondary target PXR are taken into account. Within the presented drug design process approaches from different fields (structural biology, computational biology, chemoinformatics) are combined in an interdisciplinary way. However, predictions on the effects on the whole organism, such as pathway modelling, lies beyond the scope, but would be of interest in future studies.

The reductionist view of the protomeric system

Another reductionist approach, often found in structure-based drug design, is the focus on the protomeric system. Even though all proteins of interest within this study, BRAF, PXR, and ERα, are active as dimers, the focus is on the protomeric structures and their interactions with ligands. This reductionist view of the proteins as protomeric system allows 1) for circumventing technical issues when treating the structures as ensembles, 2) for more extensive MD simulations (as a much smaller box volumes are required, which reduces the simulations computational cost), and 3) for neglecting (to some extent) additional crystallization effects that may play a role for the relative positioning of multimeric structures. Nevertheless, if the proteins occur in a dimeric state within the crystal structure this is taken into account within the refinement of the crystal structures, both as single model, or as ensemble model, and is therefore indirectly also represented in the resulting protomeric structure or structure ensemble. The dimeric conformational state is accounted for, in particular, when calculating MM-PBSA binding affinities based on the ensemble refined structures, since the ensemble is generated in the conformational composition present in the crystal (which is a dimeric state for all the used systems). For the interpretation of results it is important to keep in mind which structural state is investigated to be able to account for particular bias or errors that may occur when either neglecting a di-/multimeric state, or when taking into account the one present in e.g. a crystallographic structure. Thus, within the presented study the systems are investigated under different angles: the protomeric behaviour is analyzed in solution with MD simulations (whereas the initial conformations coming from the crystallographic dimer are free to relax into conformations that are not biased by the crystal packing), and the ensemble refinement is performed and analyzed as pure crystallographic dimer, which comprises conformational dimer restraints (and/or bias).

4.1.2 Molecular modelling

Molecular modelling has become a fundamental tool to medicinal chemists for drug design. Models are central for the understanding of chemistry and biological processes at the molecular level. Molecular modelling provides tools for investigating, explaining and discovering diverse biological processes and new phenomena. Knowledge of the atomic structure of a given target is of mayor importance when designing drugs.

The experimental basis for model generation

In the field of structure determination X-ray crystallography is still ahead of all the different techniques in terms of deposited structures per year and has established computational structure modelling methods and tools. Nevertheless, due to lacking experimental data, not all structures can be solved successfully and many of them are not complete (missing side-chains, residues, or whole segments), which is the case for most BRAF structures. This causes problems in particular when the missing structural parts impact the drug's binding site (directly or indirectly). Detailed investigations are often required for properly taking into account the resulting effects and to obtain accurate models that are able to provide useful insights into the natural process. Worth mentioning is also that an

error in atomic positions specified as Diffraction Precision Index (DPI) of 0.3 - 0.5 Å has to be taken into account for most crystallographic structures.

Extending/completing experimental models

In a nutshell, the difficulty in modelling lies in getting the right model and proper interpretation. Loop modelling is such a case where a multitude of ambiguous solutions are available, among which several may co-exist in nature. In the example of protein kinase BRAF one is confronted by the question weather it is better to take the exact atomic positions of a structure that accommodates a very closely related ligand and model the missing parts (loops) around, or to take a structure that is more complete, but not complexed with a similar ligand. Within this thesis work, the first option is preferred, as one of the aims is the development of an improved drug and the positioning of the binding site residues may have an impact on subsequent affinity predictions. Moreover, as the newly designed drug (as being similar to the already crystallized one) was expected to obtain a highly similar pose within the binding pocket and also to favor a highly similar overall conformation of the protein, starting from the available exact conformation seemed the better choice. The non-resolved loop sections were additionally expected to be highly mobile for the attained conformation within the crystal. Nonetheless, in order to take into account the effect of the modelled parts, several models were generated and MD simulations with subsequent MM-PBSA calculations were compared. One model, named BRAF-WT, is even a "homology fusion" of two crystal structures, taking all present atomic positions from the structure with the drug of interest and for the missing parts adding atomic positions of another structure with a completely resolved and structured loop. Two observations indicated that the final models were appropriate and that additionally the conformational exchange between the unstructured extended loop and the structured one may have a rather low energy barrier and may frequently occur in solution: First, the obtained MM-PBSA affinity calculations were very similar for the different selected models based on snapshots from 50 ns MD simulations. Second, one unstructured loop model folded into the structured helical form within just 100 ns of simulation time without applying any particular restraint.

Conformational ensembles and the combination of experiment and computation

Most molecules exist in multiple conformations, as they experience fluctuations in their natural environment. The preferred conformation(s) of a molecule is/are a structural characteristic feature that is in a balanced equilibrium and arises as a response to the force of attraction and repulsion, and can thus be modified by the environment. As the biological function is tightly connected with the conformational dynamics of a protein, the representation of the protein structure as conformational ensemble may be more adequate and also more informative when investigating the functional activity. Revealing conformationally heterogeneous states experimentally is not a trivial task, since macroscopic properties are usually ensemble averages over a representative statistical ensemble (either equilibrium or non-equilibrium) of molecular systems. Experimental methods that can provide information about conformational fluctuations are NMR, SAXS/WAXS, FRET, and CryoEM.[188] In general, the mayor critical point when trying to distinguish between conformational states is the observation time compared to the conversion between the states, which is usually much longer for the mentioned experimental setups. Nonetheless, single-molecule FRET (smFRET) measuring specific distances at a scale of 1-10 nm of single molecules can yield a distribution of the observable over the complete ensemble with additional time resolution. Unfortunately, those experiments do

not provide a complete picture of the molecule, but can only provide information on small parts of the structures. In contrast, SAXS experiments can provide a probability distribution of the distances between all pairs of atoms and therefore an overall shape of the ensemble can be obtained, but at a much lower resolution. As currently none of the present methods is able to provide a complete picture of all the conformational states at an atomic level, but diverse methods can provide complementary information, a combination of the obtainable information may be highly beneficial. However, one has to take into account that experimental data is not perfect and all experiments are to some extend affected by random and systematic errors and sometimes provide only sparse or even ambiguous data, representing a big challenge for modelling in general and in particular for modelling of conformational ensembles.[189]

Pure computational methods for generating conformational ensembles are already established, such as standard (atomistic) MD simulations, which are employed within this thesis work, coarse-grained MD simulations, and statistical methods, such as Monte-Carlo. Those also have their own limitations, such as force field inaccuracies and a restricted time scale limiting the sampling of the conformational space. Additionally, extensions for several computational tools have been developed that allow for supplying experimental data as additional restraints, such as the PLUMED module PLUMED-ISDB,[190] which enables implementation of several NMR observables, FRET, SAXS and cryoEM data.

Computational methods are already needed for obtaining a single structure based on experimental techniques. For example, in form of refinement tools they help to transform and interpret the obtained data. In the case of NMR, which is often used to characterize conformational fluctuations of proteins, the obtained model is an ensemble of structures. In the case of X-ray crystallography, the data is usually refined as single structure. Nonetheless, computational tools exist that extend the interpretation of the data to obtain ensembles, such as the ensemble refinement tool[95] implemented in the crystallographic refinement software PHENIX, which is used here for all three targets (ERα, PXR and BRAF) and most extensively for PXR, where the method was expected to provide additional information on the dynamics of the protein, as only a limited amount of crystallographic structures was available. The ensemble refinement of PXR structures reflects the protein's increased intrinsic flexibility, as for most structures the agreement between model and experimental data (measured by the crystallographic *R* factors) improved. Such a general improvement tendency is not detected for ERα ensemble-refined structures, neither for BRAF structures. This may indicate that both proteins are rather rigid in their crystals (excluding the unresolved parts). From the performed MM-PBSA approaches on ensemble-refined (BRAF) crystal structures it is apparent that the ensembles may contain artifacts, such as high energy conformations. Nevertheless, those ensembles can be seen as focused and experimentally validated dynamic extract, which could be sufficient to sample the important parts of the conformational space that is needed for successful VS and/or affinity prediction.

Integrated structural approaches can also be a solution to obtain structures or structural ensembles of drug targets that have been impossible to solve and therefore extend opportunities for rational structure-based drug design. Integrated structural biology is an emerging field, in particular for solving the structure of large molecules or multi-molecular assemblies. Recent advances have been made in the development of computational methods that interface different experimental techniques to combine them and bring them in agreement with resulting structural models.

There are no strict borders any more between experiment and computation for studying dynamics in

structural biology, as it becomes hard to distinguish between experimental data that is modelled by the help of computational tools and computations that are supplemented by experimental data, and the transition is rather gradual.
Hybrid methods that combine experiment with computational methods may be highly beneficial to efficiently generate conformational ensembles. Since the proper sampling of the conformational space is one of the most important limitations in free energy calculation, hybrid methods may also have a positive impact on affinity predictions. Such a pioneering usability evaluation for affinity predictions is attempted within the presented study by employing MM-PBSA calculations on ensemble-refined crystallographic structures and comparing them with results obtained based on MD simulations (see Section 3.7.3).

4.1.3 Affinity prediction

Affinity determination is one of the cornerstones of modern drug design. Many experimental and computational techniques are available to evaluate a drug's affinity towards its target, each having advantages and limitations. However, accurately predicting the binding affinity between a drug molecule and target protein remains challenging, also for recent computational methods including machine learning approaches.

The issue of data quality

As machine learning is based on data mining, the performance of the models are directly affected by the amount and quality of the available data. Deep-learning, for example, particularly relies on very large datasets, but the quality is an issue for all machine learning techniques. As experimental data in public databases is often not measured with the same biological assays, methods, or conditions, the data contains very large measurement errors, to the extend that data points are not comparable any more, and on top of that data sets may also contain contradictory entries. This heavily limits the performance of developed models and also makes the pre-cleaning of the data sets a great challenge. In the presented studies the focus is on compound data from BindingDB that for the given target (ERα and BRAFV600E) has either a large enough dataset with direct K_i affinity measurements, or IC50 measures and removed duplicates. Additional targeted prediction methods are based on a particular PubChemAssay with annotated IC50 values, which are produced by a single laboratory and labelled as confirmatory. Therefore, the measurements are expect to be reliable and coherent, and thus suitable for model development.

The required prediction accuracy

In a drug development pipeline requirements for high-throughput and accuracy usually change along the road. First, to screen an initial large amount of putative candidates often faster but less reliable techniques are used. General SBVS methods (involving docking and scoring) are often employed for molecule filtering at the Hit generation stage. Subsequently, to characterize promising Leads it is usually necessary to rely on more precise methods. This can be achieved either by development of dedicated machine learning models (see Section 3.5.2), or by rather precise

case-by-case computations that are based on the molecular structures of ligand and target. Popular examples are quantum mechanics-based methods, MM-PBSA as force field-based method, enhanced sampling MD methods to sample the targets conformational space, and alchemical free energy methods, which are based on rigorous statistical thermodynamics.
MM-PBSA is usually based on MD simulations and intermediate in both accuracy and computational expense between SBVS and alchemical methods. It is limited by several intrinsic approximations, such as the lack of conformational entropy and complete neglect of water molecules in the binding site, but still provides information on binding energies calculated on an atomic level including changes over time and provides error estimates. Therefore, MM-PBSA can be highly valuable for detailed investigations of a drug's behaviour within the binding site and associated flexibility effects and for relating them to a certain binding affinity. Within this thesis work the MM-PBSA method is employed and evaluated on the two drug development targets BRAF and PXR, highlighting limitations and providing valuable insights for directing the drug design strategy (see Section 3.7.1 and 3.7.2).

4.1.4 Selected and tested drug candidates

The aim of the drug design project was to reduce adverse effects provoked by binding of the anti-cancer drug dabrafenib (P06) to the secondary target PXR, while maintaining the drugs activity on the oncogenic primary target BRAFV600E. The design strategy was established based on knowledge about a) the binding mode of P06 in its primary and secondary target, b) the access points for metabolism of P06 (via CYP450s), c) other existing binders of the primary target (that are rather similar to P06) and their binding mode, and d) literature indications about reasons for the "paradoxical effect".

Five series of (theoretically) synthesizeable molecules with differing scaffolds were constructed *in silico* (see also Section 3.4): *P02* - molecule closest related to dabrafenib available in the PDB (PDBID: 4XV3) and supposed to avoid paradoxical activation of WT-BRAF, *P02C* - P02 having a carbon atom instead of the nitrogen connecting the methyl-ethyl moiety, *P06* - the original dabrafenib scaffold, *P06F* - P06 with one fluor atom shifted from cis to trans position at the di-fluorophenyl ring (or *P06FF* with two fluor atoms shifted), and *P06FCl* - P06F with an additional chlorine atom added in para to the central fluorophenyl ring. Subsequent to machine learning based method/tool development (see Section 3.5.2) the molecule series were subjected to affinity prediction by different developed models. The resulting suggested molecules were further filtered and adapted based on criteria concerning a) a sufficiently large and/or polar extension to avoid PXR binding, b) a slightly restricted flexibility with respect to entropy loss, c) expected cell permeability, and d) chemical stability.
The selection of compounds to be synthesized was adapted in iterative rounds based on the feedback from testing on cellular assays (by collaborators). The first synthesis trial (round 1) dedicated to avoid the "paradoxical effect" showed that alteration of P06's scaffold to P02C reduces the affinity to the primary target BRAF, as equally predicted by the machine learning models, and was therefore not further pursued. The second synthesis round indicated that extensions with higher flexibility (such as the propylamine extension of P06F-PrA) may reduce binding affinity, while more restraint, but equally large extensions showed improved IC50 values on the purified protein. Simultaneously, more polar extensions showed extensively decreased cellular activation of the secondary target PXR.

Synthesis round three indicated that the addition of a chlorine atom to the scaffold slightly improves affinity for the primary target, and only marginally decreases cellular activity of the secondary target. The fourth synthesis round revealed that very polar extensions decrease cell permeability, as the cellular assays provided poor results, even though direct affinity measurements on the purified target were promising.

To conclude, the presented synthesis rounds accompanied with detailed computational and experimental investigations revealed that a) the polarity of the molecule's extension heavily impact cell permeability and thus needs to be fine-tuned carefully together with affinity improvement attempts, and b) P06F-CPP, P06F-Mor and P06FCl-Mor represent highly promising drug candidates that are supposedly efficient inhibitors of the primary target BRAF (including cellular activity) and simultaneously avoid binding to the secondary target PXR.

4.2 General current issues and trends

Drug design / development is a multi-disciplinary research field, which highly benefits from a critical point of view (from different angles) that keeps in mind the diverse limitations of employed methods and approaches, and simultaneously tries to improve or overcome the encountered issues. Thus, new trends are also of high interest for both, academic research and the pharmaceutical industry.

4.2.1 Molecular dynamics in drug development

Molecular dynamics simulations with a wide-variety of different approximations, have become increasingly useful in studying biological systems of biomedical interest, and have been particularly successful in studying the impact of protein motions on ligand binding. This fuels the discussion of the interplay of different levels of conformational change, from the local perspective, such as changes in active site geometries, via coupled protein fluctuations, to a rather global perspective, involving secondary structure or domain movements. Moreover, the smooth transition between those levels highlight that within a single protein conformation long-range coupling networks exist that may be sensitive to interactions with different ligands. With increasing computational power MD simulations became an accessible tool for a large amount of different systems and for answering dynamics related questions on larger scales, which involves many use-cases in drug-discovery and development. Nonetheless, challenges remain. Constructing a representative ensemble of structures, covering sufficiently the required conformational space of a system is not always a straightforward task, but important, as subsequent determination of the free energies are impacted by the ensemble quality, particularly when ligands are (or should be) conformationally-selective. Another challenge is the analysis of the MD simulation itself. A major difficulty here is the amount of information that can be obtained from simulations, and the question where to look for dynamic effects or differences with respect to a reference system. When employing MD simulations (especially in rather short time scales) it is of high importance to consider the quantification of uncertainty and sampling quality.
The desired shift from a rather visual and qualitative analysis towards a more quantitative analysis is a task that is increasingly addressed by new statistical methods, such as machine learning methods, applied to MD trajectories. The large amount of data constitutes a valuable source of information, but currently the extracted knowledge is used to conclude only on the particular system at hand. Beyond that, the information from simulations could be used to train machine learning models that enable further predictions, in such a way that the gained knowledge can be used to generalize it to other systems.
Currently, the trade-off between accuracy and sampling limits the possibility to apply MD simulations within drug design campaigns, particularly on a large scale and in a high-throughput mode. Various different methods, with their advantages and limitations, have been developed to address the sampling problem, which still is a large field of research. On top of that, the approximations of the force fields used in MD currently result in the lack of a guarantee that the modelled systems are behaving like in reality and that torn conclusions are correct. Here,

in particular, the parametrization of small molecule ligands, with sometimes complicated and environment dependant polarization effects, represents a mayor challenge. The rather streamlined generation of improved partial charges, chemical topology/geometry, and parameters is addressed, for example, by the recent web tool PrimaDORAC.[191] Still, there are several approximations required to obtain a decent calculation speed and the issue may further be tackled in the future by the use of machine learning force fields, which are trained with quantum mechanics simulations. One type of deep learning method that shows high performances in machine vision, the deep convolutional neural networks (CNNs) have become increasingly popular for learning from structural biology by treating proteins as 3D images. Also other deep neural network types are currently under exploration and combined with different molecular representations (e.g. graphs).

4.2.2 Machine learning in drug development

Big data in medical biology

Over the past decade, following the emergence of new experimental techniques, such as parallel synthesis and high-throughput screening, there has been a remarkable increase in the amount of available biomedical data and compound activity data. When data is growing, at some point a human being is not capable any more of retrieving useful information without the help of computational tools. Therefore, also in the field of biomedicine and drug development the task of efficiently mining large-scale chemistry data becomes a crucial problem.[192] Moreover, the combination of large data volumes and increased automation technology has promoted further the use of machine learning and there are new emerging fields connected to this trend, such as precision medicine.[193] However, data quality stays as major concern.
In the field of medicinal chemistry machine learning has made big leaps for predicting feasible synthesis paths for many new chemicals. Machine learning and particularly performant deep-learning applications connected to drug design are diverse and include already ligand binding site detection, ligand pose prediction, ligand active/inactive classification, ligand binding affinity prediction, protein design, a.o.

The issue of the applicability domain - "conformal prediction" a solution?

Particularly for ligand-based approaches it is crucial to consider the chemical space a model has been developed on. Even if high quality data and meaningful descriptors have been used with careful validation for parameter adjustments, a given model (ligand-based) can not be expected to predict well far outside the modelled domain. This represents one of the major drawbacks of pure ligand-based approaches, as it is not always easy to judge whether a prediction on a given molecule is reliable or not. Outlier detection and similarity approaches can be indicative, but are rather subjective as they can be defined based on various measures. There is a new approach for applicability domain estimation, called "conformal prediction", which transforms classifiers and regressors into confidence predictors by providing error bounds on a per-instance basis.[194–196] It is based on a so called nonconformity score that measures how unusual an example looks relative to previous examples, and the conformal algorithm turns this nonconformity measure into prediction regions. As it can

be used with any machine learning algorithm and does not require prior probabilities (unlike Bayesian learning) - the only requirements are identically distributed training and test data, and exchangeability (order of observations is irrelevant) - it has a wide application field and is gaining momentum for QSAR models in the field of drug discovery.[197]

Information content of descriptors - the reason why there are so many different ones

As each property of a molecule is dependant on some feature in the molecule, the molecular descriptors, representing the molecule, contain some information about the molecule. The multitude of representation possibilities for molecules and modelling purposes are the major reasons for the large selection of very different descriptors that have been developed to find the ones that are most relevant to study a given problem. For instance, to investigate solubility the dipole of the molecule should be represented within the descriptors, and to model affinity of a drug one may need to describe the effect of induced fit and/or conformational selection with respect to ligand and target conformations. However, most descriptors describe a static view of the molecules, while most biological or chemical processes, including drug binding, are dynamic events, where molecules can undergo geometrical changes and can change their ionization state or polarization depending on the local environment. The effects of those dynamic changes on the predicted properties can be large and are not straight forward to incorporate into the models. This challenge is addressed here (Section 2.1), within the development of a machine learning approach taking into account ensembles of the ligand and the receptor.

The future of QSAR - PCM?

QSAR is nowadays a well established method in academia and industry and of high importance particularly in drug discovery and development. Unfortunately, large databases containing valuable high quality information are usually kept private by pharmaceutical companies. This represents a hurdle when aiming for higher productivity for pharmaceutical drug development. For example, toxicity tests with positive results may be not exploitable by the company, but of high value for machine learning attempts that equally need positive and negative data in order to provide reliable predictions. A general limitation for most QSAR models is that they are purely based on information from the ligands' chemical space, whereas information from the target may be required for a more detailed view and the development of models that are more accurate and/or have an increased ability to generalize.

A different way to combine information from the ligand side and the protein side is proteochemometrics (PCM). PCM is an emerging technology that can use genetic information for the targets without the need for solving the 3D structure and provides the possibility to investigate many targets or mutant variants simultaneously. It can be seen as extension of QSAR that uses not only information from series of chemical compounds, but also information from series of biological targets and therefore, requires activity data of the organic compounds with the investigated targets. Like for QSAR models the compounds and targets are represented by chemical descriptors, but as the targets are usually much larger than the compounds, the type of descriptors is different. Target descriptors can be classified into the ones that are based on the primary amino acid sequence, and the ones based on the 3D structure, but the best way to represent them is still an open question and the 3D approach has not been studied extensively yet, as it is again based on the knowledge of the 3D

target structure. PCM adds another dimension to the modelling as it predicts for all targets and provides opportunities to investigate in greater details the chemical interactions of compound and target.

4.2.3 The trend to target biological networks

The identification of the molecular mechanisms of drug action represents the basis for designing therapeutic strategies aimed at modifying disease processes. However, in order to achieve an efficient and safe treatment with the least side effects possible, the therapeutic strategy needs to consider the whole organism. Therefore, the emerging trend of pathology-directed systems pharmacology is based on the combination of the entire research fields of human genetics, molecular biology and systems biology.[198] There are several challenging aspects of diseases, such as the development of resistance, the observation that drug-drug interactions occur and therefore, the combined effect of two drugs might be larger or smaller than the sum of separate effects, that homeostatic feedback mechanisms exist, that different molecular mechanisms and pathways might lead to the same effect, and that all this can differ among patients.[199] Therefore, not only multi-target drugs or combinations of drugs targeting complex biological networks are a big trend, but also personalized precision treatments that are based on the patients particular circumstances and genetic disposition. The modelling of pharmacodynamic interactions becomes an important aspect, as well as safety assessment in early stages of drug development.[200]
Already at the start of a standard drug development pipeline target selection is crucial for estimating the potential and risks in safety and efficacy and is compared with other target alternatives during target validation. Here, in order to perform a sophisticated, save and efficient target selection one important aspect is the extension of the target space and its characterization in terms of tractability and druggability.

4.2.4 Target tractability & druggability

Target identification is the prerequisite for successful drug development. The assessment of target druggability comprises the assessment of the potential of a target to result in a successful delivery of a novel therapeutics, which is not easy to predict, but of high interest for the pharmaceutical sector.[201] Target tractability (a.k.a. ligandability) is defined by Brown et al.[202] as "the likelihood of identifying a modulator that interacts effectively with the target/domain (or pathway)." In contrast to druggability, tractability does not consider whether the modulator molecule would be suitable as potential drug candidate. Thus, tractability is a necessary but not sufficient condition for druggability and is therefore less restrictive. It has the advantage that it focuses on the properties of potential binding sites and is experimentally accessible (e.g. by screening of compound libraries).[203] To predict the level of ligandability most computational methods are based on cavity detection to identify pockets or suitable surface patches to predict their likelihood to act as ligand binding sites.[203] One limitation is that these methods usually require an atomic structure to accurately evaluate the ligandability of a target. Nonetheless, if the target structure is unknown, there may be other available information

that could serve as indicator for target tractability and additionally, there are more parameters that are crucial for successful drug development, such as target location and the potential for off-target effects. This is taken into account in new approaches that combine genome wide assessment of tractability, data mining from different data bases, data integration and structure-based tractability assessment.[202] These approaches are not restricted to targets with available structures and open up possibilities to detect targets previously considered undruggable or simply not known. However, they still benefit from structural information, which may be delivered in larger numbers through the development of new integrative experimental and hybrid computational structural approaches.

Bibliography

[1] Monya Baker. Fragment-based lead discovery grows up. *Nature Reviews. Drug Discovery*, 12(1):5–7, January 2013.

[2] Ansgar Schuffenhauer, Simon Ruedisser, Andreas L. Marzinzik, Wolfgang Jahnke, Marcel Blommers, Paul Selzer, and Edgar Jacoby. Library design for fragment based screening. *Current Topics in Medicinal Chemistry*, 5(8):751–762, 2005.

[3] M. M. Hann, A. R. Leach, and G. Harper. Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of Chemical Information and Computer Sciences*, 41(3):856–864, June 2001.

[4] Harren Jhoti, Glyn Williams, David C. Rees, and Christopher W. Murray. The 'rule of three' for fragment-based drug discovery: where are we now? *Nature Reviews Drug Discovery*, 12(8):644–644, August 2013.

[5] Venkata Velvadapu, Bennett T. Farmer, and Allen B. Reitz. Chapter 7 - Fragment-Based Drug Discovery. In Camille Georges Wermuth, David Aldous, Pierre Raboisson, and Didier Rognan, editors, *The Practice of Medicinal Chemistry (Fourth Edition)*, pages 161–180. Academic Press, San Diego, January 2015.

[6] Bernard Munos. Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery*, 8(12):959–968, 2009. 00701.

[7] Joseph A. DiMasi, Lanna Feldman, A. Seckler, and A. Wilson. Trends in risks associated with new drug development: success rates for investigational drugs. *Clinical Pharmacology & Therapeutics*, 87(3), 2010. 00362.

[8] Gerhard Müller. Medicinal chemistry of target family-directed masterkeys. *Drug Discovery Today*, 8(15):681–691, August 2003. 00213.

[9] A. K. Ghose and G. M. Crippen. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *Journal of Chemical Information and Computer Sciences*, 27(1):21–35, February 1987.

[10] Tiejun Cheng, Yuan Zhao, Xun Li, Fu Lin, Yong Xu, Xinglong Zhang, Yan Li, Renxiao Wang, and Luhua Lai. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *Journal of Chemical Information and Modeling*, 47(6):2140–2148, December 2007.

[11] Maris Lapins and Jarl ES Wikberg. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. *BMC Bioinformatics*, 11:339, June 2010.

[12] Krisna C. Duong-Ly and Jeffrey R. Peterson. The human kinome and kinase inhibition. *Current Protocols in Pharmacology*, Chapter 2:Unit2.9, March 2013.

[13] Frieda A. Sorgenfrei, Simone Fulle, and Benjamin Merget. Kinome-Wide Profiling Prediction of Small Molecules. *ChemMedChem*, 13(6):495–499, 2018.

[14] A. Petrelli and S. Giordano. From single- to multi-target drugs in cancer therapy: when aspecificity becomes an advantage. *Current Medicinal Chemistry*, 15(5):422–432, 2008.

[15] Ernesto Freire. Do enthalpy and entropy distinguish first in class from best in class? *Drug Discovery Today*, 13(19-20):869–874, October 2008.

[16] Ronan O'Brien, Natalia Markova, and Geoffrey A. Holdgate. Thermodynamics in Drug Discovery. In *Applied Biophysics for Drug Discovery*, pages 7–28. John Wiley & Sons, Ltd, 2017.

[17] P. A. Borea, A. Dalpiaz, K. Varani, P. Gilli, and G. Gilli. Can thermodynamic measurements of receptor binding yield information on drug affinity and efficacy? *Biochemical Pharmacology*, 60(11):1549–1556, December 2000.

[18] John D. Chodera and David L. Mobley. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annual Review of Biophysics*, 42:121–142, 2013.

[19] W. H. Ward and G. A. Holdgate. Isothermal titration calorimetry in drug discovery. *Progress in Medicinal Chemistry*, 38:309–376, 2001.

[20] Barbara Wienen-Schmidt, Hendrik R. A. Jonker, Tobias Wulsdorf, Hans-Dieter Gerber, Krishna Saxena, Denis Kudlinzki, Sridhar Sreeramulu, Giacomo Parigi, Claudio Luchinat, Andreas Heine, Harald Schwalbe, and Gerhard Klebe. Paradoxically, Most Flexible Ligand Binds Most Entropy-Favored: Intriguing Impact of Ligand Flexibility and Solvation on Drug-Kinase Binding. *Journal of Medicinal Chemistry*, 61(14):5922–5933, July 2018.

[21] Benjamin Breiten, Matthew R. Lockett, Woody Sherman, Shuji Fujita, Mohammad Al-Sayah, Heiko Lange, Carleen M. Bowers, Annie Heroux, Goran Krilov, and George M. Whitesides. Water Networks Contribute to Enthalpy/Entropy Compensation in Protein–Ligand Binding. *Journal of the American Chemical Society*, 135(41):15579–15584, October 2013.

[22] Hiroyasu Ohtaka and Ernesto Freire. Adaptive inhibitors of the HIV-1 protease. *Progress in Biophysics and Molecular Biology*, 88(2):193–208, June 2005.

[23] Tjelvar S. G. Olsson, Mark A. Williams, William R. Pitt, and John E. Ladbury. The thermodynamics of protein-ligand interaction and solvation: insights for ligand design. *Journal of Molecular Biology*, 384(4):1002–1017, December 2008.

[24] John E. Ladbury, Gerhard Klebe, and Ernesto Freire. Adding calorimetric data to decision making in lead discovery: a hot tip. *Nature Reviews. Drug Discovery*, 9(1):23–27, 2010.

[25] Andrew L. Hopkins, Jonathan S. Mason, and John P. Overington. Can we rationally design promiscuous drugs? *Current Opinion in Structural Biology*, 16(1):127–136, February 2006.

[26] Paul D. Leeson and Brian Springthorpe. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews. Drug Discovery*, 6(11):881–890, 2007.

[27] György G. Ferenczy and György M. Keserű. Thermodynamics guided lead discovery and optimization. *Drug Discovery Today*, 15(21-22):919–932, November 2010.

[28] Christian X. Weichenberger and Bernhard Rupp. Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate. *Acta Crystallographica. Section D, Biological Crystallography*, 70(Pt 6):1579–1588, June 2014.

[29] R. A. Woldeyes, D. A. Sivak, and J. S. Fraser. E pluribus unum, no more: from one crystal, many conformations. *Current opinion in structural biology*, 28:56–62, October 2014.

[30] Daniel A. Keedy, James S. Fraser, and Henry van den Bedem. Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit. *PLoS Computational Biology*, 11(10), October 2015.

[31] P. A. Williams, V. Fülöp, E. F. Garman, N. F. Saunders, S. J. Ferguson, and J. Hajdu. Haem-ligand switching during catalysis in crystals of a nitrogen-cycle enzyme. *Nature*, 389(6649):406–412, September 1997.

[32] Matthew P. Jacobson, Richard A. Friesner, Zhexin Xiang, and Barry Honig. On the Role of the Crystal Environment in Determining Protein Side-chain Conformations. *Journal of Molecular Biology*, 320(3):597–608, July 2002.

[33] Chaya S. Rapp and Rena M. Pollack. Crystal packing effects on protein loops. *Proteins: Structure, Function, and Bioinformatics*, 60(1):103–109, 2005.

[34] Alexander Wlodawer, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski. Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *The FEBS journal*, 280(22):5705–5736, November 2013.

[35] Frank DiMaio, Thomas C. Terwilliger, Randy J. Read, Alexander Wlodawer, Gustav Oberdorfer, Ulrike Wagner, Eugene Valkov, Assaf Alon, Deborah Fass, Herbert L. Axelrod, Debanu Das, Sergey M. Vorobiev, Hideo Iwaï, P. Raj Pokkuluri, and David Baker. Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature*, 473(7348):540–543, May 2011.

[36] Mi Li, Frank Dimaio, Dongwen Zhou, Alla Gustchina, Jacek Lubkowski, Zbigniew Dauter, David Baker, and Alexander Wlodawer. Crystal structure of XMRV protease differs from the structures of other retropepsins. *Nature Structural & Molecular Biology*, 18(2):227–229, February 2011.

[37] Dan Shi, Brent L. Nannenga, Matthew G. Iadanza, and Tamir Gonen. Three-dimensional electron crystallography of protein microcrystals. *eLife*, 2:e01345, November 2013.

[38] M. Jason de la Cruz, Johan Hattne, Dan Shi, Paul Seidler, Jose Rodriguez, Francis E. Reyes, Michael R. Sawaya, Duilio Cascio, Simon C. Weiss, Sun Kyung Kim, Cynthia S. Hinck, Andrew P. Hinck, Guillermo Calero, David Eisenberg, and Tamir Gonen. Atomic-resolution structures from fragmented protein crystals with the cryoEM method MicroED. *Nature Methods*, 14(4):399–402, February 2017.

[39] Mary J. Harner, Luciano Mueller, Kevin J. Robbins, and Michael D. Reily. NMR in drug design. *Archives of Biochemistry and Biophysics*, 628:132–147, 2017.

[40] David J. Craik and Hayden Peacock. Overview of NMR in Drug Design. In Graham A. Webb, editor, *Modern Magnetic Resonance*, pages 1–11. Springer International Publishing, Cham, 2017.

[41] Steven A. Hofstadler and Kristin A. Sannes-Lowery. Applications of ESI-MS in drug discovery: interrogation of noncovalent complexes. *Nature Reviews. Drug Discovery*, 5(7):585–595, July 2006.

[42] Valerie Vivat Hannah, C. Atmanene, D. Zeyer, A. Van Dorsselaer, and Sarah Sanglier-Cianférani. Native MS: an 'ESI' way to support structure- and fragment-based drug discovery. *Future Medicinal Chemistry*, 2(1):35–50, January 2010.

[43] Xin Chen, Shanshan Qin, Shuai Chen, Jinlong Li, Lixin Li, Zhongling Wang, Quan Wang, Jianping Lin, Cheng Yang, and Wenqing Shui. A ligand-observed mass spectrometry approach integrated into the fragment based lead discovery pipeline. *Scientific Reports*, 5:8361, February 2015.

[44] Andrew N. Holding. XL-MS: Protein cross-linking coupled with mass spectrometry. *Methods (San Diego, Calif.)*, 89:54–63, November 2015.

[45] Glenn R. Masson, Meredith L. Jenkins, and John E. Burke. An overview of hydrogen deuterium exchange mass spectrometry (HDX-MS) in drug discovery. *Expert Opinion on Drug Discovery*, 12(10):981–994, 2017.

[46] Tim Gruene, Julian T. C. Wennmacher, Christan Zaubitzer, Julian J. Holstein, Jonas Heidler, Ariane Fecteau-Lefebvre, Sacha De Carlo, Elisabeth Müller, Kenneth N. Goldie, Irene Regeni, Teng Li, Gustavo Santiso-Quinones, Gunther Steinfeld, Stephan Handschin, Eric van Genderen, Jeroen A. van Bokhoven, Guido H. Clever, and Radosav Pantelic. Rapid Structure Determination of Microcrystalline Molecular Compounds Using Electron Diffraction. *Angewandte Chemie (International Ed. in English)*, 57(50):16313–16317, 2018.

[47] Christopher G. Jones, Michael W. Martynowycz, Johan Hattne, Tyler J. Fulton, Brian M. Stoltz, Jose A. Rodriguez, Hosea M. Nelson, and Tamir Gonen. The CryoEM Method MicroED as a Powerful Tool for Small Molecule Structure Determination. *ACS central science*, 4(11):1587–1592, November 2018.

[48] Haixia Su and Yechun Xu. Application of ITC-Based Characterization of Thermodynamic and Kinetic Association of Ligands With Proteins in Drug Design. *Frontiers in Pharmacology*, 9, October 2018.

[49] Mei-Chu Lo, Ann Aulabaugh, Guixian Jin, Rebecca Cowling, Jonathan Bard, Michael Malamas, and George Ellestad. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Analytical Biochemistry*, 332(1):153–159, September 2004.

[50] Chuong Nguyen, Graham M. West, and Kieran F. Geoghegan. Emerging Methods in Chemoproteomics with Relevance to Drug Discovery. *Methods in Molecular Biology (Clifton, N.J.)*, 1513:11–22, 2017.

[51] Simon G. Patching. Surface plasmon resonance spectroscopy for characterisation of membrane protein–ligand interactions and its potential for drug discovery. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1838(1, Part A):43–55, January 2014.

[52] Elisa Michelini, Luca Cevenini, Laura Mezzanotte, Andrea Coppa, and Aldo Roda. Cell-based assays: fuelling drug discovery. *Analytical and Bioanalytical Chemistry*, 398(1):227–238, September 2010.

[53] Gregory Nierode, Paul S. Kwon, Jonathan S. Dordick, and Seok-Joon Kwon. Cell-Based Assay Design for High-Content Screening of Drug Candidates. *Journal of Microbiology and Biotechnology*, 26(2):213–225, February 2016.

[54] Benjamin T. Burlingham and Theodore S. Widlanski. An Intuitive Look at the Relationship of Ki and IC50: A More General Use for the Dixon Plot. *Journal of Chemical Education*, 80(2):214, February 2003.

[55] Cheng Yung-Chi and William H. Prusoff. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical Pharmacology*, 22(23):3099–3108, December 1973.

[56] Evanthia Lionta, George Spyrou, Demetrios K. Vassilatis, and Zoe Cournia. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current Topics in Medicinal Chemistry*, 14(16):1923–1938, 2014. 00018.

[57] Antonio Lavecchia. Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*, 20(3):318–331, March 2015. 00010.

[58] A. Lavecchia and C. Di Giovanni. Virtual screening strategies in drug discovery: a critical review. *Current Medicinal Chemistry*, 20(23):2839–2860, 2013. 00063.

[59] Yvonne Westermaier, Xavier Barril, and Leonardo Scapozza. Virtual screening: An in silico tool for interlacing the chemical universe with the proteome. *Methods*, 71:44–57, January 2015. 00002.

[60] Dagmar Stumpfe and Jürgen Bajorath. Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry*, 55(7):2932–2942, April 2012.

[61] Maykel Cruz-Monteagudo, José L. Medina-Franco, Yunierkis Pérez-Castillo, Orazio Nicolotti, M. Natália D. S. Cordeiro, and Fernanda Borges. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today*, 19(8):1069–1080, August 2014.

[62] Jordi Mestres and Ronald M. A. Knegtel. Similarity versus docking in 3d virtual screening. *Perspectives in Drug Discovery and Design*, 20(1):191–207, December 2000. 00041.

[63] Anthony Nicholls, Georgia B. McGaughey, Robert P. Sheridan, Andrew C. Good, Gregory Warren, Magali Mathieu, Steven W. Muchmore, Scott P. Brown, J. Andrew Grant, James A. Haigh, Neysa Nevins, Ajay N. Jain, and Brian Kelley. Molecular Shape and Medicinal Chemistry: A Perspective. *Journal of Medicinal Chemistry*, 53(10):3862–3886, May 2010. 00165.

[64] C. G. Wermuth, C. R. Ganellin, P. Lindberg, and L. A. Mitscher. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure and Applied Chemistry*, 70(5):1129–1143, 2009. 00288.

[65] Andrew R. Leach, Valerie J. Gillet, Richard A. Lewis, and Robin Taylor. Three-dimensional pharmacophore methods in drug discovery. *Journal of Medicinal Chemistry*, 53(2):539–558, January 2010.

[66] Antonio Carrieri, Violeta I. Pérez-Nueno, Giovanni Lentini, and David W. Ritchie. Recent trends and future prospects in computational GPCR drug discovery: from virtual screening to polypharmacology. *Current Topics in Medicinal Chemistry*, 13(9):1069–1097, 2013. 00016.

[67] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, December 2002.

[68] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010.

[69] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7, May 2015.

[70] Shane Weaver and M. Paul Gleeson. The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling*, 26(8):1315–1326, June 2008.

[71] Angélica Nakagawa Lima, Eric Allison Philot, Gustavo Henrique Goulart Trossini, Luis Paulo Barbour Scott, Vinícius Gonçalves Maltarollo, and Kathia Maria Honorio. Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery*, 11(3):225–239, 2016.

[72] Qurrat Ul Ain, Antoniya Aleksandrova, Florian D. Roessler, and Pedro J. Ballester. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, 5(6):405–424, 2015. 00002.

[73] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. SSS*. Springer, New York, 2009.

[74] Nuno M. F. S. A. Cerqueira, Diana Gesto, Eduardo F. Oliveira, Diogo Santos-Martins, Natércia F. Brás, Sérgio F. Sousa, Pedro A. Fernandes, and Maria J. Ramos. Receptor-based virtual screening protocol for drug discovery. *Archives of Biochemistry and Biophysics*, 582:56–67, September 2015. 00010.

[75] Gerhard Klebe. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today*, 11(13–14):580–594, July 2006. 00525.

[76] Hongjian Li, Kwong-Sak Leung, Man-Hon Wong, and Pedro J Ballester. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. *BMC Bioinformatics*, 15(1), August 2014. 00009.

[77] Martin Cohen-Gonsaud, Vincent Catherinot, Gilles Labesse, and Dominique Douguet. From molecular modeling to drug design. In *Practical bioinformatics*, pages 35–71. Springer, 2008. 00008.

[78] A. Schafferhans and G. Klebe. Docking ligands onto binding site representations derived from proteins built by homology modelling. *Journal of Molecular Biology*, 307(1):407–427, March 2001. 00113.

[79] Connie Oshiro, Erin K. Bradley, John Eksterowicz, Erik Evensen, Michelle L. Lamb, J. Kevin Lanctot, Santosh Putta, Robert Stanton, and Peter D. J. Grootenhuis. Performance of 3d-Database Molecular Docking Studies into Homology Models. *Journal of Medicinal Chemistry*, 47(3):764–767, January 2004. 00097.

[80] Noel M. OLBoyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *J Cheminf*, 3:33, 2011. 00943.

[81] Maria A. Miteva, Frederic Guyon, and Pierre Tufféry. Frog2: Efficient 3d conformation ensemble generator for small compounds. *Nucleic Acids Research*, 38(Web Server issue):W622–W627, July 2010.

[82] Landrum, Greg. RDKit: Open-Source Cheminformatics Software, 2006.

[83] Sereina Riniker and Gregory A. Landrum. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574, December 2015.

[84] Mikko J. Vainio and Mark S. Johnson. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *Journal of Chemical Information and Modeling*, 47(6):2462–2474, November 2007.

[85] Peter Sadowski and Pierre Baldi. Small-molecule 3d Structure Prediction Using Open Crystallography Data. *Journal of chemical information and modeling*, 53(12):3127, December 2013.

[86] Jens. Sadowski and Johann. Gasteiger. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chemical Reviews*, 93(7):2567–2581, November 1993.

[87] Paul C. D. Hawkins, A. Geoffrey Skillman, Gregory L. Warren, Benjamin A. Ellingson, and Matthew T. Stahl. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling*, 50(4):572–584, April 2010.

[88] Chia-En Chang and Michael K. Gilson. Tork: Conformational analysis method for molecules and complexes. *Journal of Computational Chemistry*, 24(16):1987–1998, 2003.

[89] Paul C. D. Hawkins. Conformation Generation: The State of the Art. *Journal of Chemical Information and Modeling*, 57(8):1747–1756, August 2017.

[90] Sam Z. Grinter and Xiaoqin Zou. Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design. *Molecules*, 19(7):10150–10176, July 2014. 00019.

[91] Jon A. Erickson, Mehran Jalaie, Daniel H. Robertson, Richard A. Lewis, and Michal Vieth. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry*, 47(1):45–55, January 2004. 00293.

[92] Gregory L. Warren, C. Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H. Lambert, Mika Lindvall, Neysa Nevins, Simon F. Semus, Stefan Senger, Giovanna Tedesco, Ian D. Wall, James M. Woolven, Catherine E. Peishoff, and Martha S. Head. A Critical Assessment of Docking Programs and Scoring Functions. *Journal of Medicinal Chemistry*, 49(20):5912–5931, October 2006. 01081.

[93] F. Jiang and S. H. Kim. "Soft docking": matching of molecular surface cubes. *Journal of Molecular Biology*, 219(1):79–102, May 1991. 00391.

[94] Oliver Korb, Thomas Stützle, and Thomas E. Exner. Empirical Scoring Functions for Advanced ProteinLigand Docking with PLANTS. *Journal of Chemical Information and Modeling*, 49(1):84–96, January 2009. 00321.

[95] B. Tom Burnley, Pavel V. Afonine, Paul D. Adams, and Piet Gros. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife*, 1:e00311, December 2012. 00092.

[96] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, December 2005.

[97] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, March 2008. 07109.

[98] Olivier Sperandio, Liliane Mouawad, Eulalie Pinto, Bruno O. Villoutreix, David Perahia, and Maria A. Miteva. How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *European biophysics journal: EBJ*, 39(9):1365–1372, August 2010. 00055.

[99] Dariusz Plewczynski, Michał Łaźniewski, Rafał Augustyniak, and Krzysztof Ginalski. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry*, 32(4):742–755, March 2011.

[100] Jason B. Cross, David C. Thompson, Brajesh K. Rai, J. Christian Baber, Kristi Yi Fan, Yongbo Hu, and Christine Humblet. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *Journal of Chemical Information and Modeling*, 49(6):1455–1474, June 2009.

[101] Jacob D. Durrant and J. Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC biology*, 9:71, October 2011.

[102] Matthew J. Harvey and Gianni De Fabritiis. High-throughput molecular dynamics: the powerful new tool for drug discovery. *Drug Discovery Today*, 17(19-20):1059–1062, October 2012.

[103] Hongtao Zhao and Amedeo Caflisch. Molecular dynamics in drug design. *European Journal of Medicinal Chemistry*, 91:4–14, February 2015.

[104] Jennifer L. Radkiewicz and Charles L. Brooks. Protein Dynamics in Enzymatic Catalysis: Exploration of Dihydrofolate Reductase. *Journal of the American Chemical Society*, 122(2):225–231, January 2000.

[105] Matic Pavlin, Angelo Spinello, Marzia Pennati, Nadia Zaffaroni, Silvia Gobbi, Alessandra Bisi, Giorgio Colombo, and Alessandra Magistrato. A Computational Assay of Estrogen Receptor Antagonists Reveals the Key Common Structural Traits of Drugs Effectively Fighting Refractory Breast Cancers. *Scientific Reports*, 8(1):649, January 2018.

[106] Rajesh Kumar Pathak, Ayushi Gupta, Rohit Shukla, and Mamta Baunthiyal. Identification of new drug-like compounds from millets as Xanthine oxidoreductase inhibitors for treatment of Hyperuricemia: A molecular docking and simulation study. *Computational Biology and Chemistry*, 76:32–41, October 2018.

[107] Abdolkarim Farrokhzadeh, Farideh Badichi Akher, and Mahmoud E. S. Soliman. Probing the Dynamic Mechanism of Uncommon Allosteric Inhibitors Optimized to Enhance Drug Selectivity of SHP2 with Therapeutic Potential for Cancer Treatment. *Applied Biochemistry and Biotechnology*, 188(1):260–281, May 2019.

[108] Mala S. Kumar, Amjesh R, Silpa Bhaskaran, Delphin R. D, Achuthsankar S. Nair, and Sudhakaran P. R. Molecular docking and dynamic studies of crepiside E beta glucopyranoside as an inhibitor of snake venom PLA2. *Journal of Molecular Modeling*, 25(4):88, March 2019.

[109] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Elsevier, October 2001. Google-Books-ID: 5qTzldS9ROIC.

[110] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, May 1995.

[111] Jay W. Ponder and David A. Case. Force fields for protein simulations. *Advances in Protein Chemistry*, 66:27–85, 2003.

[112] D. A. Case, D. S. Cerutti, T. E. Cheatham, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, D. M. York, and P. A. Kollman. Amber 2017, University of California, San Francisco, 2017. 00000.

[113] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[114] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690, March 2010.

[115] W. Scott and W. van Gunsteren. The GROMOS software package for biomolecular simulations. *Methods and Techniques in Computational Chemistry: METECC*, 95:397–434, 1995.

[116] Kunihiro Kitamura, Yunoshin Tamura, Tomokazu Ueki, Koji Ogata, Shigeho Noda, Ryutaro Himeno, and Hiroshi Chuman. Binding free-energy calculation is a powerful tool for drug optimization: calculation and measurement of binding free energy for 7-azaindole derivatives to glycogen synthase kinase-3. *Journal of Chemical Information and Modeling*, 54(6):1653–1660, June 2014.

[117] Peter A. Kollman, Irina Massova, Carolina Reyes, Bernd Kuhn, Shuanghong Huo, Lillian Chong, Matthew Lee, Taisung Lee, Yong Duan, Wei Wang, Oreola Donini, Piotr Cieplak, Jaysharee Srinivasan, David A. Case, and Thomas E. Cheatham. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research*, 33(12):889–897, December 2000.

[118] Aaron Weis, Kambiz Katebzadeh, Pär Söderhjelm, Ingemar Nilsson, and Ulf Ryde. Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. *Journal of Medicinal Chemistry*, 49(22):6596–6606, November 2006.

[119] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *Journal of Computational Chemistry*, 32(5):866–877, April 2011. 00234.

[120] Samuel Genheden and Ulf Ryde. Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. *Proteins*, 80(5):1326–1342, May 2012.

[121] Huiyong Sun, Youyong Li, Mingyun Shen, Sheng Tian, Lei Xu, Peichen Pan, Yan Guan, and Tingjun Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Physical chemistry chemical physics: PCCP*, 16(40):22035–22045, October 2014. 00071.

[122] Samuel Genheden and Ulf Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5):449–461, May 2015.

[123] Pin-Chih Su, Cheng-Chieh Tsai, Shahila Mehboob, Kirk E. Hevener, and Michael E. Johnson. Comparison of Radii Sets, Entropy, QM Methods, and Sampling on MM-PBSA, MM-GBSA, and QM/MM-GBSA Ligand Binding Energies of F. tularensis Enoyl-ACP Reductase (FabI). *Journal of computational chemistry*, 36(25):1859–1873, September 2015.

[124] Changhao Wang, Peter H. Nguyen, Kevin Pham, Danielle Huynh, Thanh-Binh Nancy Le, Hongli Wang, Pengyu Ren, and Ray Luo. Calculating protein-ligand binding affinities with MMPBSA: Method and error analysis. *Journal of Computational Chemistry*, 37(27):2436–2446, 2016.

[125] Matteo Aldeghi, Michael J. Bodkin, Stefan Knapp, and Philip C. Biggin. Statistical Analysis on the Performance of Molecular Mechanics Poisson–Boltzmann Surface Area versus Absolute Binding Free Energy Calculations: Bromodomains as a Case Study. *Journal of Chemical Information and Modeling*, 57(9):2203–2221, September 2017.

[126] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 51(1):69–82, January 2011.

[127] Salla I. Virtanen, Sanna P. Niinivehmas, and Olli T. Pentikäinen. Case-specific performance of MM-PBSA, MM-GBSA, and SIE in virtual screening. *Journal of Molecular Graphics & Modelling*, 62:303–318, November 2015. 00005.

[128] Bernd Kuhn, Paul Gerber, Tanja Schulz-Gasch, and Martin Stahl. Validation and use of the MM-PBSA approach for drug discovery. *Journal of Medicinal Chemistry*, 48(12):4040–4048, June 2005. 00315.

[129] Marc Adler and Paul Beroza. Improved ligand binding energies derived from molecular dynamics: replicate sampling enhances the search of conformational space. *Journal of Chemical Information and Modeling*, 53(8):2065–2072, August 2013. 00008.

[130] Yan Li, Zhihai Liu, and Renxiao Wang. Test MM-PB/SA on true conformational ensembles of protein-ligand complexes. *Journal of Chemical Information and Modeling*, 50(9):1682–1692, September 2010. 00022.

[131] M. J. Robinson and M. H. Cobb. Mitogen-activated protein kinase pathways. *Current Opinion in Cell Biology*, 9(2):180–186, April 1997.

[132] Robert Roskoski. A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacological Research*, 100:1–23, October 2015.

[133] Robert Roskoski. Targeting oncogenic Raf protein-serine/threonine kinases in human cancers. *Pharmacological Research*, 135:239–258, 2018.

[134] Helen Davies, Graham R. Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, Hayley Woffendin, Mathew J. Garnett, William Bottomley, Neil Davis, Ed Dicks, Rebecca Ewing, Yvonne Floyd, Kristian Gray, Sarah Hall, Rachel Hawes, Jaime Hughes, Vivian Kosmidou, Andrew Menzies, Catherine Mould, Adrian Parker, Claire Stevens, Stephen Watt, Steven Hooper, Rebecca Wilson, Hiran Jayatilake, Barry A. Gusterson, Colin Cooper, Janet Shipley, Darren Hargrave, Katherine Pritchard-Jones, Norman Maitland, Georgia Chenevix-Trench, Gregory J. Riggins, Darell D. Bigner, Giuseppe Palmieri, Antonio Cossu, Adrienne Flanagan, Andrew Nicholson, Judy W. C. Ho, Suet Y. Leung, Siu T. Yuen, Barbara L. Weber, Hilliard F. Seigler, Timothy L. Darrow, Hugh Paterson, Richard Marais, Christopher J. Marshall, Richard Wooster, Michael R. Stratton, and P. Andrew Futreal. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954, June 2002.

[135] Paul T. C. Wan, Mathew J. Garnett, S. Mark Roe, Sharlene Lee, Dan Niculescu-Duvaz, Valerie M. Good, C. Michael Jones, Christopher J. Marshall, Caroline J. Springer, David Barford, Richard Marais, and Cancer Genome Project. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, 116(6):855–867, March 2004.

[136] Brunangelo Falini, Maria Paola Martelli, and Enrico Tiacci. BRAF V600e mutation in hairy cell leukemia: from bench to bedside. *Blood*, 128(15):1918–1927, 2016.

[137] Maurizio Pulici, Gabriella Traquandi, Chiara Marchionni, Michele Modugno, Rosita Lupi, Nadia Amboldi, Elena Casale, Nicoletta Colombo, Luca Corti, Marina Fasolini, Fabio Gasparri, Wilma Pastori, Alessandra Scolaro, Daniele Donati, Eduard Felder, Arturo Galvani, Antonella Isacchi, Enrico Pesenti, and Marina Ciomei. Optimization of diarylthiazole B-raf inhibitors: identification of a compound endowed with high oral antitumor activity, mitigated hERG inhibition, and low paradoxical effect. *ChemMedChem*, 10(2):276–295, February 2015.

[138] Christine A. Pratilas, Barry S. Taylor, Qing Ye, Agnes Viale, Chris Sander, David B. Solit, and Neal Rosen. (V600e)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11):4519–4524, March 2009.

[139] Kenta Terai and Michiyuki Matsuda. The amino-terminal B-Raf-specific region mediates calcium-dependent homo- and hetero-dimerization of Raf. *The EMBO journal*, 25(15):3556–3564, August 2006.

[140] Hugo Lavoie, Malha Sahmi, Pierre Maisonneuve, Sara A. Marullo, Neroshan Thevakumaran, Ting Jin, Igor Kurinov, Frank Sicheri, and Marc Therrien. MEK drives BRAF activation through allosteric control of KSR proteins. *Nature*, 554(7693):549–553, 2018.

[141] Robert Roskoski. RAF protein-serine/threonine kinases: structure and regulation. *Biochemical and Biophysical Research Communications*, 399(3):313–317, August 2010.

[142] Bogos Agianian and Evripidis Gavathiotis. Current Insights of BRAF Inhibitors in Cancer. *Journal of Medicinal Chemistry*, 61(14):5775–5793, 2018.

[143] Paul B. Chapman, Axel Hauschild, Caroline Robert, John B. Haanen, Paolo Ascierto, James Larkin, Reinhard Dummer, Claus Garbe, Alessandro Testori, Michele Maio, David Hogg, Paul Lorigan, Celeste Lebbe, Thomas Jouary, Dirk Schadendorf, Antoni Ribas, Steven J. O'Day, Jeffrey A. Sosman, John M. Kirkwood, Alexander M. M. Eggermont, Brigitte Dreno, Keith Nolop, Jiang Li, Betty Nelson, Jeannie Hou, Richard J. Lee, Keith T. Flaherty, Grant A. McArthur, and BRIM-3 Study Group. Improved survival with vemurafenib in melanoma with BRAF V600e mutation. *The New England Journal of Medicine*, 364(26):2507–2516, June 2011.

[144] Geoffrey T. Gibney and Jonathan S. Zager. Clinical development of dabrafenib in BRAF mutant melanoma and other malignancies. *Expert Opinion on Drug Metabolism & Toxicology*, 9(7):893–899, July 2013.

[145] Geoffrey Kim, Amy E. McKee, Yang-Min Ning, Maitreyee Hazarika, Marc Theoret, John R. Johnson, Qiang Casey Xu, Shenghui Tang, Rajeshwari Sridhara, Xiaoping Jiang, Kun He, Donna Roscoe, W. David McGuinn, Whitney S. Helms, Anne Marie Russell, Sarah Pope Miksinski, Jeanne Fourie Zirkelbach, Justin Earp, Qi Liu, Amna Ibrahim, Robert Justice, and Richard Pazdur. FDA approval summary: vemurafenib for treatment of unresectable or metastatic melanoma with the BRAFV600e mutation. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 20(19):4994–5000, October 2014.

[146] Anita D. Ballantyne and Karly P. Garnock-Jones. Dabrafenib: first global approval. *Drugs*, 73(12):1367–1376, August 2013.

[147] Alexander M. Menzies and Georgina V. Long. Dabrafenib and trametinib, alone and in combination for BRAF-mutant metastatic melanoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 20(8):2035–2043, April 2014.

[148] Weijiang Zhang. BRAF inhibitors: the current and the future. *Current Opinion in Pharmacology*, 23:68–73, August 2015.

[149] Lauretta Odogwu, Luckson Mathieu, Gideon Blumenthal, Erin Larkins, Kirsten B. Goldberg, Norma Griffin, Karen Bijwaard, Eunice Y. Lee, Reena Philip, Xiaoping Jiang, Lisa Rodriguez, Amy E. McKee, Patricia Keegan, and Richard Pazdur. FDA Approval Summary: Dabrafenib and Trametinib for the Treatment of Metastatic Non-Small Cell Lung Cancers Harboring BRAF V600e Mutations. *The Oncologist*, 23(6):740–745, June 2018.

[150] Tsun-Wen Yao, Jie Zhang, Michael Prados, William A. Weiss, C. David James, and Theodore Nicolaides. Acquired resistance to BRAF inhibition in BRAFV600e mutant gliomas. *Oncotarget*, 8(1):583–595, January 2017.

[151] Jean-Pierre Delord, Caroline Robert, Marta Nyakas, Grant A. McArthur, Ragini Kudchakar, Amit Mahipal, Yasuhide Yamada, Ryan Sullivan, Ana Arance, Richard F. Kefford, Matteo S. Carlino, Manuel Hidalgo, Carlos Gomez-Roca, Daniela Michel, Abdelkader Seroutou, Vassilios Aslanis, Giordano Caponigro, Darrin D. Stuart, Laure Moutouh-de Parseval, Tim Demuth, and Reinhard Dummer. Phase I Dose-Escalation and -Expansion Study of the BRAF Inhibitor Encorafenib (LGX818) in Metastatic BRAF-Mutant Melanoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 23(18):5339–5348, September 2017.

[152] Cathrine L. Denton, Elisabeth Minthorn, Stanley W. Carson, Graeme C. Young, Lauren E. Richards-Peterson, Jeffrey Botbyl, Chao Han, Royce A. Morrison, Samuel C. Blackman, and Daniele Ouellet. Concomitant oral and intravenous pharmacokinetics of dabrafenib, a BRAF inhibitor, in patients with BRAF V600 mutation-positive solid tumors. *Journal of Clinical Pharmacology*, 53(9):955–961, September 2013.

[153] David A. Bershas, Daniele Ouellet, Donna B. Mamaril-Fishman, Noelia Nebot, Stanley W. Carson, Samuel C. Blackman, Royce A. Morrison, Jerry L. Adams, Kristen E. Jurusik, Dana M. Knecht, Peter D. Gorycki, and Lauren E. Richards-Peterson. Metabolism and disposition of oral dabrafenib in cancer patients: proposed participation of aryl nitrogen in carbon-carbon bond cleavage via decarboxylation following enzymatic oxidation. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, 41(12):2215–2224, December 2013.

[154] Sarah K. Lawrence, Dung Nguyen, Chet Bowen, Lauren Richards-Peterson, and Konstantine W. Skordos. The metabolic drug-drug interaction profile of Dabrafenib: in vitro investigations and quantitative extrapolation of the P450-mediated DDI risk. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, 42(7):1180–1190, July 2014.

[155] Daniele Ouellet, Ekaterina Gibiansky, Cathrine Leonowens, Anne O'Hagan, Patricia Haney, Julie Switzky, and Vicki L. Goodman. Population pharmacokinetics of dabrafenib, a BRAF inhibitor: effect of dose, time, covariates, and relationship with its metabolites. *Journal of Clinical Pharmacology*, 54(6):696–706, June 2014.

[156] A. Benjamin Suttle, Kenneth F. Grossmann, Daniele Ouellet, Lauren E. Richards-Peterson, Gursel Aktan, Michael S. Gordon, Patricia M. LoRusso, Jeffrey R. Infante, Sunil Sharma, Kari Kendra, Manish Patel, Shubham Pant, Hendrik-Tobias Arkenau, Mark R. Middleton, Samuel C. Blackman, Jeff Botbyl, and Stanley W. Carson. Assessment of the drug interaction potential and single- and repeat-dose pharmacokinetics of the BRAF inhibitor dabrafenib. *Journal of Clinical Pharmacology*, 55(4):392–400, April 2015.

[157] Jeovanis Gil, Lazaro Hiram Betancourt, Indira Pla, Aniel Sanchez, Roger Appelqvist, Tasso Miliotis, Magdalena Kuras, Henriette Oskolas, Yonghyo Kim, Zsolt Horvath, Jonatan Eriksson, Ethan Berge, Elisabeth Burestedt, Göran Jönsson, Bo Baldetorp, Christian Ingvar, Håkan Olsson, Lotta Lundgren, Peter Horvatovich, Jimmy Rodriguez Murillo, Yutaka Sugihara, Charlotte Welinder, Elisabet Wieslander, Boram Lee, Henrik Lindberg, Krzysztof Pawłowski, Ho Jeong Kwon, Viktoria Doma, Jozsef Timar, Sarolta Karpati, A. Marcell Szasz, István Balázs Németh, Toshihide Nishimura, Garry Corthals, Melinda Rezeli, Beatrice Knudsen, Johan Malm, and György Marko-Varga. Clinical protein science in translational medicine targeting malignant melanoma. *Cell Biology and Toxicology*, March 2019.

[158] Alicja Puszkiel, Gaëlle Noé, Audrey Bellesoeur, Nora Kramkimel, Marie-Noëlle Paludetto, Audrey Thomas-Schoemann, Michel Vidal, François Goldwasser, Etienne Chatelut, and Benoit Blanchet. Clinical Pharmacokinetics and Pharmacodynamics of Dabrafenib. *Clinical Pharmacokinetics*, 58(4):451–467, April 2019.

[159] Pierre Germain, Bart Staels, Catherine Dacquet, Michael Spedding, and Vincent Laudet. Overview of Nomenclature of Nuclear Receptors. *Pharmacological Reviews*, 58(4):685–704, December 2006. 00431.

[160] D. J. Mangelsdorf, C. Thummel, M. Beato, P. Herrlich, G. Schütz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon, and R. M. Evans. The nuclear receptor superfamily: the second decade. *Cell*, 83(6):835–839, December 1995. 06444.

[161] V. Giguère. Orphan nuclear receptors: from gene to function. *Endocrine Reviews*, 20(5):689–725, October 1999. 00874.

[162] Timothy M. Willson and Steven A. Kliewer. PXR, CAR and drug metabolism. *Nature Reviews. Drug Discovery*, 1(4):259–266, April 2002. 00389.

[163] A. Geick, M. Eichelbaum, and O. Burk. Nuclear receptor response elements mediate induction of intestinal MDR1 by rifampin. *The Journal of Biological Chemistry*, 276(18):14581–14587, May 2001. 00846.

[164] Caroline Raynal, Jean-Marc Pascussi, Géraldine Leguelinel, Cyril Breuker, Jovana Kantar, Benjamin Lallemant, Sylvain Poujol, Caroline Bonnans, Dominique Joubert, Frédéric Hollande, Serge Lumbroso, Jean-Paul Brouillet, and Alexandre Evrard. Pregnane X Receptor (PXR) expression in colorectal cancer cells restricts irinotecan chemosensitivity through enhanced SN-38 glucuronidation. *Molecular Cancer*, 9:46, March 2010. 00000.

[165] Hongwei Wang, Madhukumar Venkatesh, Hao Li, Regina Goetz, Subhajit Mukherjee, Arunima Biswas, Liang Zhu, Andreas Kaubisch, Lei Wang, James Pullman, Kathleen Whitney, Makoto Kuro-o, Andres I. Roig, Jerry W. Shay, Moosa Mohammadi, and Sridhar Mani. Pregnane X receptor activation induces FGF19-dependent tumor aggressiveness in humans and mice. *The Journal of Clinical Investigation*, 121(8):3220–3232, August 2011. 00073.

[166] Yakun Chen, Yong Tang, Changxiong Guo, Jiuhui Wang, Debasish Boral, and Daotai Nie. Nuclear receptors in the multidrug resistance through the regulation of drug-metabolizing enzymes and drug transporters. *Biochemical Pharmacology*, 83(8):1112–1126, April 2012. 00125.

[167] B. Lawrence Riggs and Lynn C. Hartmann. Selective Estrogen-Receptor Modulators — Mechanisms of Action and Application to Clinical Practice. *New England Journal of Medicine*, 348(7):618–629, February 2003.

[168] Vanessa Delfosse, Marina Grimaldi, Jean-Luc Pons, Abdelhay Boulahtouf, Albane le Maire, Vincent Cavailles, Gilles Labesse, William Bourguet, and Patrick Balaguer. Structural and mechanistic insights into bisphenols action provide guidelines for risk assessment and discovery of bisphenol A substitutes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37):14930–14935, September 2012. 00075.

[169] M. Berry, D. Metzger, and P. Chambon. Role of the two activating domains of the oestrogen receptor in the cell-type and promoter-context dependent agonistic activity of the anti-oestrogen 4-hydroxytamoxifen. *The EMBO journal*, 9(9):2811–2818, September 1990. 00776.

[170] Robbie P. Joosten, Fei Long, Garib N. Murshudov, and Anastassis Perrakis. The PDB_redo server for macromolecular structure model optimization. *IUCrJ*, 1(Pt 4):213–220, May 2014. 00148.

[171] Barry J. Grant, Ana P. C. Rodrigues, Karim M. ElSawy, J. Andrew McCammon, and Leo S. D. Caves. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England)*, 22(21):2695–2696, November 2006.

[172] Sebastian Salentin, Sven Schreiber, V. Joachim Haupt, Melissa F. Adasme, and Michael Schroeder. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, page gkv315, April 2015. 00013.

[173] Oliver Korb, Thomas Stützle, and Thomas E. Exner. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In Marco Dorigo, Luca Maria Gambardella, Mauro Birattari, Alcherio Martinoli, Riccardo Poli, and Thomas Stützle, editors, *Ant Colony Optimization and Swarm Intelligence*, Lecture Notes in Computer Science, pages 247–258. Springer Berlin Heidelberg, 2006.

[174] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, December 1993. 09586.

[175] Narayanan Eswar, Ben Webb, Marc A. Marti-Renom, M. S. Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, and Andrej Sali. Comparative protein structure modeling using Modeller. *Current Protocols in Bioinformatics*, Chapter 5:Unit–5.6, October 2006.

[176] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–1423, June 2009.

[177] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling*, 25(2):247–260, October 2006. 01693.

[178] Wouter G. Touw, Coos Baakman, Jon Black, Tim A. H. te Beek, E. Krieger, Robbie P. Joosten, and Gert Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43(Database issue):D364–368, January 2015.

[179] Dinesh Kumar Kala Sekar, Gurusaran Manickam, S.N. Satheesh, P Radha, S Pavithra, K P. S. Thulaa Tharshan, John Helliwell, and K Sekar. Online-DPI: A web server to calculate the diffraction precision index for a protein structure. *Journal of Applied Crystallography*, 48:939–942, June 2015.

[180] Géraldine Lemaire, Cindy Benod, Virginie Nahoum, Arnaud Pillon, Anne-Marie Boussioux, Jean-François Guichou, Guy Subra, Jean-Marc Pascussi, William Bourguet, Alain Chavanieu, and Patrick Balaguer. Discovery of a highly active ligand of human pregnane x receptor: a case study from pharmacophore modeling and virtual screening to "in vivo" biological activity. *Molecular Pharmacology*, 72(3):572–581, September 2007.

[181] Mathieu Seimandi, Géraldine Lemaire, Arnaud Pillon, Agnès Perrin, Isabelle Carlavan, Johannes J. Voegel, Françoise Vignon, Jean-Claude Nicolas, and Patrick Balaguer. Differential responses of PPARalpha, PPARdelta, and PPARgamma reporter cell lines to selective PPAR synthetic ligands. *Analytical Biochemistry*, 344(1):8–15, September 2005.

[182] Paul D. Adams, Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, Li-Wei Hung, Gary J. Kapral, Ralf W. Grosse-Kunstleve, Airlie J. McCoy, Nigel W. Moriarty, Robert Oeffner, Randy J. Read, David C. Richardson, Jane S. Richardson, Thomas C. Terwilliger, and Peter H. Zwart. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica. Section D, Biological Crystallography*, 66(Pt 2):213–221, February 2010. 08847.

[183] Paul Emsley and Kevin Cowtan. Coot: model-building tools for molecular graphics. *Acta Crystallographica. Section D, Biological Crystallography*, 60(Pt 12 Pt 1):2126–2132, December 2004. 18475.

[184] Evgeny Krissinel and Kim Henrick. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372(3):774–797, September 2007.

[185] José L. Medina-Franco, Marc A. Giulianotti, Gregory S. Welmaker, and Richard A. Houghten. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discovery Today*, 18(9):495–501, May 2013.

[186] Aislyn DW Boran and Ravi Iyengar. Systems approaches to polypharmacology and drug discovery. *Current opinion in drug discovery & development*, 13(3):297, May 2010.

[187] Edgar Jacoby. Computational chemogenomics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(1):57–67, 2011.

[188] Massimiliano Bonomi and Michele Vendruscolo. Determination of protein structural ensembles using cryo-electron microscopy. *Current Opinion in Structural Biology*, 56:37–45, June 2019.

[189] Massimiliano Bonomi, Gabriella T. Heller, Carlo Camilloni, and Michele Vendruscolo. Principles of protein structural ensemble determination. *Current Opinion in Structural Biology*, 42:106–116, 2017.

[190] Massimiliano Bonomi and Carlo Camilloni. Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics*, 33(24):3999–4000, December 2017.

[191] Piero Procacci. PrimaDORAC: A Free Web Interface for the Assignment of Partial Charges, Chemical Topology, and Bonded Parameters in Organic or Drug Molecules. *Journal of Chemical Information and Modeling*, 57(6):1240–1245, 2017.

[192] Fabricio F. Costa. Big data in biomedicine. *Drug Discovery Today*, 19(4):433–440, April 2014.

[193] Daniel Richard Leff and Guang-Zhong Yang. Big Data for Precision Medicine. *Engineering*, 1(3):277–279, September 2015.

[194] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, softcover reprint of hardcover 1st ed. 2005 edition edition, October 2010.

[195] Ulf Norinder, Lars Carlsson, Scott Boyer, and Martin Eklund. Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *Journal of Chemical Information and Modeling*, 54(6):1596–1603, June 2014.

[196] U. Norinder, A. Rybacka, and P. L. Andersson. Conformal prediction to define applicability domain - A case study on predicting ER and AR binding. *SAR and QSAR in environmental research*, 27(4):303–316, April 2016.

[197] Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence*, 74(1):117–132, June 2015.

[198] Meindert Danhof, Kevin Klein, Pieter Stolk, Murray Aitken, and Hubert Leufkens. The future of drug development: the paradigm shift towards systems therapeutics. *Drug Discovery Today*, 23(12):1990–1995, December 2018.

[199] Meindert Danhof. Systems pharmacology – Towards the modeling of network interactions. *European Journal of Pharmaceutical Sciences*, 94:4–14, October 2016.

[200] Marianne Uteng, Laszlo Urban, Dominique Brees, Patrick Y. Muller, Gerd A. Kullak-Ublick, Page Bouchard, Gervais Tougas, and Salah-Dine Chibout. Safety differentiation: emerging competitive edge in drug development. *Drug Discovery Today*, 24(1):285–292, January 2019.

[201] Robert M. Garbaccio and Emma R. Parmee. The Impact of Chemical Probes in Drug Discovery: A Pharmaceutical Industry Perspective. *Cell Chemical Biology*, 23(1):10–17, January 2016.

[202] Kristin K. Brown, Michael M. Hann, Ami S. Lakdawala, Rita Santos, Pamela J. Thomas, and Kieran Todd. Approaches to target tractability assessment – a practical perspective. *MedChemComm*, 9(4):606–613, April 2018.

[203] Udo Bauer and Alexander L. Breeze. "Ligandability" of Drug Targets: Assessment of Chemical Tractability via Experimental and In Silico Approaches. In *Lead Generation*, pages 35–62. John Wiley & Sons, Ltd, 2016.

List of Figures

1.1 MAPK pathway activation. RAS activation promotes the formation of RAF dimers. RAF phosphorylates MEK, which upon activation phosphorylates ERK. ERK phosphorylates many different targets and also exerts a direct negative feedback by phosphorylating and thus regulating CRAF and MEK activity. The BRAF-mutant V600E is constitutively signalling as a monomer and insensitive to ERK negative feedback mechanisms. Paradoxical activation of ERK in BRAF-wild-type cells occurs via transactivation of RAF dimers. The dotted hammers represent pathway-induced feedback inhibition. 55

1.2 Metabolic pathway of dabrafenib (DB) and its three major metabolites hydroxy-dabrafenib (HDB), carboxy-dabrafenib (CDB), and desmethyl-dabrafenib (DDB). DB is metabolized by CYP3A4 and CYP2C8 to HDB, further oxidized to CDB (by CYP3A4), and decarboxylated to DDB (pH-dependant).[153] 59

3.1 Motivation: The anti-cancer drug does not only bind to its primary target BRAF, but also to the nuclear receptor PXR, which subsequently induces the transcription of CYP450 resulting in drug degradation and adverse effects. The new drug should still bind its primary target BRAF, but not any more PXR. 94

3.2 Rigid core superpositioning result (left) of 34 PXR protomers with 'rigid core' residues highlighted in red (calculated with a cumulative volume cutoff at 0.5 $Å^3$); and subsequent RMSF calculation across the protomers with rainbow color coding (from 0Å in blue to a maximum of 9.95Å in red). 96

3.3 RMSF calculation across the 34 PXR protomers plotted along the sequence with secondary structure annotation from a reference structure (PDB-ID: 4S0T chain B). 97

3.4 Principal Component Analysis (PCA) of the PXR conformational ensemble (top row) with clustering in PC space (PC1-2), 2D plots of the first three PCs and the proportion of the variance covered by the PCs. RMSD conformer clustering (bottom row) with cluster overlap shown by RMSD dendrogram colored by PC cluster. 98

3.5 Principal Component Analysis of the PXR conformational ensemble: per residue contributions to the first two PCs (left), and structural representations of the second PC with vectors (grey arrows) calculated using the *modevectors* module in PyMol (right). 99

3.6 Residues implied in ligand binding of 25 liganded PXR protomers (left), colored by the frequency they are identified being within a radius of 4 Å around the ligand. A representative protein structure (PDB-ID: 4S0T, chain B) (right) is used to visualize the location of the identified residues within the structure. The side-chains of the 43 identified residues are shown in line representation on the protein structure and the coloring is also based on identification frequency. 99

3.7 All-atom mean RMSF of residues implied in ligand binding of 25 liganded PXR protomers (left). The 43 residues are ordered based on identification frequency (compare Figure 3.6). A representative protein structure (PDB-ID: 4S0T, chain B) (right) is used to visualize the location of the identified residues within the structure. The side-chains of the 43 identified residues are shown in line representation on the protein structure and the coloring is also based on the all-atom mean RMSF in Å. 100

3.8 Rigid core superpositioning result (left) of 96 liganded BRAF protomers with 'rigid core' residues highlighted in red (calculated with a cumulative volume cutoff at 0.5 $Å^3$); and subsequent RMSF calculation across the protomers with rainbow color coding (from 0 Å in blue to a maximum of 9.1 Å in red). 103

3.9 RMSF calculation across the 96 BRAF protomers plotted along the sequence with secondary structure annotation from a reference structure (PDB-ID: 4MBJ chain B). 104

3.10 Normal mode analysis of BRAF structure 5HID (chain B). The modevectors (color-coded by direction) visualized on the protein structure (in sand colored tube representation) (left) and without the structure (right). The vector field representation is generated with the bio3d function *pymol.modes* and PyMol. . . . 104
3.11 Deformation energies (left) and fluctuations (right) based on the first three normal modes of BRAF structure 5HID (chain B) color coded (rainbow: blue to red) onto the structure. Values range from 0.10 to 11.42 for the deformation energies, and from 0.00 to 0.31 for the fluctuations. 105
3.12 Principal Component Analysis of the BRAF conformational ensemble: Proportion of the variance covered by the PCs (top left), per residue contributions to the first two PCs (bottom left), and structural representations of the first two PCs with vectors (grey arrows) calculated using the *modevectors* module in PyMol (right). . . . 106
3.13 Hierarchical clustering based on RMSD (top row) and clustering in PC-space (bottom row) of 96 liganded BRAF protomers. The dendrogram inset in the bottom row shows the cluster overlap of the two methods by coloring the PC-space dendrogram based on RMSD clusters. The superimposed structures are colored based on the identified clusters. . . . 107
3.14 Residues implied in ligand binding of 96 liganded BRAF protomers (left), colored by the frequency they are identified being within a radius of 4 Å around the ligand. A representative protein structure (PDB-ID: 5HID, chain B) (right) is used to visualize the location of the identified residues within the structure. The side-chains of the 52 identified residues are shown in line representation on the protein structure and the coloring is also based on identification frequency. . . . 108
3.15 All-atom mean RMSF of residues implied in ligand binding of 96 liganded BRAF protomers (left). The 52 residues are ordered based on identification frequency (compare Figure 3.14). A representative protein structure (PDB-ID: 5HID, chain B) (right) is used to visualize the location of the identified residues within the structure. The side-chains of the 52 identified residues are shown in line representation on the protein structure and the coloring is also based on the all-atom mean RMSF in Å. . . . 108
3.16 The binding mode of dabrafenib in its primary target BRAF (PDB-ID: 4XV2, chain A, downloaded from PDB-REDO). . . . 110
3.17 The binding mode of dabrafenib in its anti-target, the nuclear receptor PXR (PDB-ID: 6HJ2, chain A, processed with PDB-REDO). . . . 110
3.18 PLIP interactions of dabrafenib in its primary target BRAF (PDB-ID: 4XV2, chain A, downloaded from PDB-REDO). . . . 111
3.19 PLIP interactions of dabrafenib in its anti-target, the nuclear receptor PXR (PDB-ID: 6HJ2, chain A, processed with PDB-REDO). . . . 111
3.20 The binding mode of dabrafenib in its primary target BRAF (PDB-ID: 4XV2, chain B, downloaded from PDB-REDO). . . . 112
3.21 PDB structures of BRAF with ligands similar to dabrafenib (P06) with superimposed protein structures and comparison of the crystallographic ligands' chemistry and binding mode. Coloring by crystallographic complex (as for the PDB entry codes) and ligand IDs in grey. . . . 113
3.22 *In silico* synthesis of drug candidates - KNIME workflow. . . . 116
3.23 Two loop-complete models with mutated BRAF sequences: 1) a loop-model based on PDB structure 4XV2 (chain A) (rose) with an extended a-loop that is folded back onto the structure, and interacting with the G-rich loop and 2) a loop-model based on PDB structure 4CQE (chain A) (yellow) with an extended and free a-loop. All mutated residues are shown in stick representation. These are the V600E mutation and 13 solubilizing mutations (I543A, I544S, I551K, Q562R, L588N, K630S, F667E, Y673S, A688R, L706S, Q709R, S713E, L716E - permitting kinase domain overexpression in bacteria). . . . 119
3.24 Two complete models with naturally occurring sequences: BRAF-V600E with extended a-loop (orange) and BRAF-WT (V600) with structured a-loop conformation (violet). Residue 600 is shown in stick representation and also ligand P06 (grey). . . . 120

3.25 Helix formation during MD simulation. Left: The starting conformation of BRAF-V600E (orange) and the last frame after 100 ns of MD simulation (sand). Whereas the starting conformation has a completely unstructured activation loop, the last conformation shows an α-helix for residues Glu611 - Met620. Right: Secondary structure calculation from DSSP along the trajectory for the whole protein sequence, with the residues forming the helix highlighted in a black box. 121
3.26 The selected molecules for drug synthesis round 1 - scaffold P02C, their predicted affinity by the RF model (trained on the BindingDB dataset), and their docking pose within the BRAF structure. 135
3.27 Molecules of drug synthesis round 2 - scaffold P06F. 136
3.28 Molecules of drug synthesis round 3 - includes the four modified scaffolds based on P06 (P06F, P06FF, P06FCl, and P06FFCl) with the original tertiary butyl extension and the morpholino (Mor) extension. 136
3.29 Molecules of drug synthesis round 4 - scaffold P06F. 137
3.30 MM-PBSA binding energy of the 8 molecules from synthesis round 2 (drug scaffold P06F) at a dielectric constant of 8 in two different loop-model structures: 4XV2 and 4CQE. The errors are based on the variability among the selected 500 frames from a 50ns MD simulation for each complex. 147
3.31 MM-PBSA binding energy of P06 at a dielectric constant of 8 in four different loop-model structures: 4XV2, 4CQE, V600E and the structured WT. The errors are based on the variability among the selected 500 frames from a 50ns MD simulation for each complex. 148
3.32 MM-PBSA binding energy of molecules with four different scaffolds (P02C, P06F, P06FCl, and P06FFCl) and a morpholine (Mor) as extension in the BRAF-V600E model, at a dielectric constant of 8. 149
3.33 MM-PBSA binding energy of molecules with four different scaffolds (P06, P06F, P06FCl, and P06FFCl) containing the original tertiary butyl moiety of P06 in the PXR model based on 6HJ2, at a dielectric constant of 2. 150
3.34 H-bond occupancy per residue averaged across the 5 MD replicas per complex. H-bonds are calculated between protein and ligand for all trajectory frames using the VMD H-bonds plugin and occupancies of all contributions from a protein residue are summed up. Residues that formed at least in one MD simulation hydrogen bonds with occupancy $\geqslant$10 are listed. The errors are based on the variability across the 5 MD replicas per complex. Note that the occupancy can be larger than 100% for a residue, as more than one h-bond can be formed per residue. 151
3.35 MM-PBSA results for the crystallographic ensemble-refined BRAF structure 4XV3 with P02. . . 153
3.36 MM-PBSA results for the crystallographic ensemble-refined BRAF structure 4XV2 with P06. . . 154
3.37 MM-PBSA results for the crystallographic ensemble-refined BRAF structure 5CSW with P06. . . 155
3.38 MM-PBSA results for the newly solved crystallographic ensemble-refined BRAF structure with P06F-Mor. 156
3.39 Crystals of BRAFV600E protein with the designed ligand P06F-Mor. 162
3.40 The classical BRAF dimer ("back-to-back"). 163
3.41 Modelled atomic structures within the respective electron density difference maps ($2F_{obs} - F_{calc}$ in blue and $F_{obs} - F_{calc}$ in green for positive density and red for negative density, contoured at 1.0σ). 164
3.42 The refined crystallographic BRAFV600E structure (chain A in green and chain B in cyan) with the designed ligand P06F-Mor (violet). The dimeric structure shows a domain swap of the activation loop. Residues within a 5 Å distance of any ligand atom are shown in line representation. 164
3.43 The refined crystallographic BRAFV600E structure (chain A in green and chain B in cyan) with the designed ligand P06F-Mor (violet). The interface of the two protomers (as identified by the PyMol tool *interfaceResidue*) is highlighted in rose with its surface in dot representation. 165
3.44 PDBePISA[184] server output of the refined crystallographic BRAFV600E structure with the designed ligand P06F-Mor. The interface of the two protomers A and B is evaluated concerning size and electrostatic bonding. 166

3.45 The refined crystallographic BRAFV600E structure with the symmetric unit cell replicated in one axis, as present in the protein crystal (generated using Coot by displaying the symmetric molecules and visualized in PyMol). 166
3.46 The refined crystallographic BRAFV600E structure colored by b-factor with a range of 21.42 - 106.55 Å (rainbow: blue to red). 167
3.47 Electron density and atomic structures of the designed drug P06F-Mor in the two protomers (chain A and B) of the crystal structure. 167
3.48 PLIP interactions of designed drug P06F-Mor with chain A in its crystal structure. 168
3.49 PLIP interactions of designed drug P06F-Mor with chain B in its crystal structure. 169
3.50 Hydrogen bond network (yellow dashed lines) that is presumably established between the designed drug P06F-Mor (in purple), a water molecule (red sphere) present in the binding site and the residues Lys483 and Asp594 (in stick representation) in the two protomers (chain A and B) of the crystal structure. 169
3.51 The refined crystallographic BRAFV600E structure with the designed ligand P06F-Mor as domain swap dimer (chain A in green and chain B in cyan, as in Figure 3.42) and crystallographic structure 6U2G containing BRAF (chain B, in dark cyan) and MEK (chain A, in bordeaux), whereas the respective B chains (cyan and dark cyan) are superimposed. 170

List of Tables

1.1 Details of the three FDA approved second generation BRAF inhibitors vermurafenib, dabrafenib and encorafenib, with affinities for BRAF-V600E.[142] 58

3.1 Summary of ensemble refinement results for PXR. Column 'initial PDB' are the single structure refinement values from the structure deposited in the PDB, column 'ensemble refinement' are the values obtained by refining the initial PDB structure as ensemble, followed by the columns of ensemble refined structures with added side chains and added loops. In all ensemble refinements the dataset with the best R_{free} is chosen. ^{a} = apo-structure, ^{d} = homodimer, ^{t} = heterotetramer with RXRα ^{c} = with SRC-1 coactivator. 102

3.2 Summary of ensemble refinement results for BRAF. Column 'initial PDB' are the single structure refinement values from the structure deposited in the PDB, column 'ensemble refinement' are the values obtained by refining the initial PDB structure as ensemble. In all ensemble refinements the dataset with the best R_{free} is chosen. * newly solved structure in complex with a drug candidate. 109

3.3 MM-PBSA affinity calculations on ensemble-refined BRAF structures (with solute dielectric constant ε=8). Binding energies are provided in kJ/mol. Ensembles originating from two different ensemble-refinement runs were used, one without ligand hydrogens and one with ligand hydrogens included during refinement (indicated by (-H) and (+H), respectively). The ensembles with lowest R_{free} values were selected for calculations. Each ensemble was employed directly for MM-PBSA calculations (indicated by "ensemble") and the structures were minimized prior to MM-PBSA calculations (indicated by "minimized"). 152

3.4 Resolution and Diffraction Precision Index (DPI, calculated by Online-DPI[179]) of BRAF crystal structures. 154

3.5 IC50 affinity measurements for the synthesized drug candidates from synthesis round 2 on BRAF-V600E. 159

3.6 IC50 affinity measurements for the synthesized drug candidates from synthesis round 3 on BRAF-V600E. 160

3.7 IC50 affinity measurements for the synthesized drug candidates from synthesis round 4 on BRAF-V600E. 160