Computational methods for assessing drug-target residence times in bacterial enoyl-ACP reductases and predicting small-molecule permeability for the *Mycobacterium tuberculosis* cell wall

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades der
Julius-Maximilians-Universität Würzburg

vorgelegt von

Benjamin Merget
aus Aschaffenburg

Würzburg, Oktober 2015

Eingereicht bei der Fakultät für Chemie und Pharmazie am

____

Gutachter der schriftlichen Arbeit:
1. Gutachter:

____

2. Gutachter:

____

Prüfer des öffentlichen Promotionskolloquiums:
1. Prüfer:

____

2. Prüfer:

____

3. Prüfer:

____

Datum des öffentlichen Promotionskolloquiums:

____

Doktorurkunde ausgehändigt am

____

“Reality leaves a lot to the imagination.”

—John Lennon

All of the presented work was carried out under the supervision of Prof. Dr. Christoph Sotriffer at the Institute of Pharmacy and Food Chemistry (University of Würzburg, Germany) between October 2011 and October 2015.

Parts of this thesis are published in scientific journals or are in the process of being published. Therefore, chapters contain text excerpts and figures/tables from these publications:

B. Merget, C.A. Sotriffer (2015), Slow-onset inhibition of *Mycobacterium tuberculosis* InhA: Revealing molecular determinants of residence time by MD simulations, *PLoS ONE*, 10(5): e0127009, ref. [1]

Publication	**Use in dissertation**	**Page number**
pp. 1–2	text reproduced, modified and extended	p. 2
p. 2	text reproduced, modified and extended	p. 11
p. 2	text reproduced and extended	p. 14
pp. 3, 9, 10	figures adapted	pp. 15, 31, 38
pp. 4, S1, S2, 7, S3, S4, 12, 13, S5, 14–16, S6, 18, 19, S7, S8	figures reproduced	pp. 12, 33, 34, 36, 37, 40, 41, 42, 43, 44, 44, 45, 46, 48, 49, 51
pp. 2, 4, 5	text reproduced, modified and extended	pp. 29–31
pp. 5, 6, 8	text reproduced	pp. 31–39
p. 8	table reproduced	p. 35
pp. 10–14	text reproduced	pp. 39–47
pp. 16–17	text reproduced	pp. 47–50
p. 16	table reproduced	p. 46
pp. 19–21	text reproduced	pp. 50–53

B. Merget, C.A. Sotriffer, An accurate and quantitative prediction model for drug-target residence time of *Staphylococcus aureus* FabI inhibitors based on Steered Molecular Dynamics, *manuscript in preparation*, ref. [2]

Publication	Use in dissertation	Page number
	text, figures and tables will (partly) be used and modified in the publication currently in preparation	pp. 1ff, 9ff, 17ff, 75ff

B. Merget, D. Zilian, T. Müller, C.A. Sotriffer (2013), MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 29(1): 62–68, ref. [3][1]

Publication	Use in dissertation	Page number
p. 1	text reproduced, modified and extended	p. 119
p. 1	text reproduced, modified and extended	p. 119
pp. 1, 2	text reproduced, modified and extended	p. 127–128
p. S1	table reproduced and modified	p. 129
pp. 2, 3	text reproduced, modified and extended	pp. 128–134
pp. S1, 2, 3, S3	figures reproduced	pp. 130, 131, 132, 133, 144, 145
p. 3	table reproduced	p. 131
pp. 3, 4	text reproduced	p. 132
pp. 4, 5	text reproduced, modified and extended	pp. 134–141
p. S2	table reproduced and modified	pp. 135–138
pp. 4, 5, S3	figures reproduced and modified	pp. 140, 142, 147
pp. 5, 6	text reproduced, modified and extended	pp. 141–143
p. S3	text reproduced, modified and extended	p. 143–146
pp. 6, 7	text reproduced, modified and extended	pp. 146–150

[1] Reproduced, modified and extended from [3] with permission by Oxford University Press.

Statement of individual author contributions

B. Merget, C.A. Sotriffer (2015), Slow-onset inhibition of *Mycobacterium tuberculosis* InhA: Revealing molecular determinants of residence time by MD simulations, *PLoS ONE*, 10(5): e0127009, [1]

Participated in	Author Initials, Responsibility decreasing from left to right			
Study Design	BM, CAS			
Data Collection	BM			
Data Analysis and Interpretation	BM, CAS			
Manuscript Writing	BM	CAS		

B. Merget, C.A. Sotriffer, An accurate and quantitative prediction model for drug-target residence time of *Staphylococcus aureus* FabI inhibitors based on Steered Molecular Dynamics, *manuscript in preparation*, [2]

Participated in	Author Initials, Responsibility decreasing from left to right			
Study Design	BM, CAS			
Data Collection	BM			
Data Analysis and Interpretation	BM, CAS			
Manuscript Writing	BM	CAS		

B. Merget, D. Zilian, T. Müller, C.A. Sotriffer (2013), MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 29(1): 62–68, [3]

Participated in	Author Initials, Responsibility decreasing from left to right			
Study Design	BM, CAS	TM	DZ	
Data Collection	BM	DZ		
Data Analysis and Interpretation	BM, CAS	TM	DZ	
Manuscript Writing	BM	CAS		

The following publications emerged between October 2011 and April 2015, which are not further discussed in this thesis:

- A.C. Braun, D. Ilko, B. Merget, H. Gieseler, O. Germershaus, U. Holzgrabe, L. Meinel (2015), Predicting critical micelle concentration and micelle molecular weight of polysorbate 80 using compendial methods, *European Journal of Pharmaceutics and Biopharmaceutics*, 94: 559–568.

- J. Schiebel, A. Chang, B. Merget, G. Bommineni, W. Yu, L. Spagnuolo, M. Baxter, M. Tareilus, P. Tonge, C. Kisker, C.A. Sotriffer (2015), An Ordered Water Channel in *Staphylococcus aureus* FabI: Unraveling the Mechanism of Substrate Recognition and Reduction, *ACS Biochemistry*, 54(10): 1943–1955.

- A. Ponte-Sucre, H. Bruhn, T. Schirmeister, A. Cecil, C.R. Albert, C. Buechold, M. Tischer, S. Schlesinger, T. Goebel, A. Fuß, D. Mathein, B. Merget, C.A. Sotriffer, A. Stich, G. Krohne, M. Engstler, G. Bringmann, U. Holzgrabe (2015), Anti-trypanosomal activities and structural chemical properties of selected compound classes, *Parasitology Research*, 114(2): 501–512.

- X. Chen, S. Wehle, N. Kuzmanovic, B. Merget, U. Holzgrabe, B. König, C.A. Sotriffer, M. Decker (2014), Acetylcholinesterase Inhibitors with Photoswitchable Inhibition of β-Amyloid Aggregation, *ACS Chemical Neuroscience*, 5(5): 377–389.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor Prof. Dr. Christoph Sotriffer for the constant support and guidance he has provided throughout all our projects. His mentoring allowed me to follow and explore own ideas and interests while simultaneously never feeling lost. Furthermore, I would like to thank him for the immeasurable amount of work he has put into our projects and for constantly pushing me and teaching me not to be satisfied with anything but your best effort.

I am very grateful to Dr. Johannes Schiebel for countless discussions (scientific and otherwise). Our conferences together were always very insightful and, on top of it, great fun. Moreover, I thank Steffen Wagner, Yogesh Narkhede and Sandra Eltschkner (the rest of the "FabI"-bunch) for our many little talks and sharing of ideas about the topic. In addition, I would like to thank Prof. Dr. Caroline Kisker for many insightful meetings and discussions.

I am thankful to Dr. Tobias Müller for his support and his great ideas throughout the MycPermCheck project. Almost everything I know about biostatistics I learned from him. Moreover, I want to thank Dr. David Zilian for his contributions to this project and, particularly, for "having my back" these past years in the office. I want to express my special thanks to Raphael Dives, the most helpful person in the world, for always having an open ear for so many little problems and constantly providing tailored solutions for everyone in our group.

Furthermore, I would like to thank the entire group–Dr. Michael Hein, Dominik Heuler, Maximilian Kuhn, Manuel Krug, Lukas Pason, Christina Plank, Daniel Schuster, Sarah Wehle, Dr. Armin Welker, Thomas Willmes and (again) Raphael Dives, Yogesh Narkhede, Steffen Wagner and Dr. David Zilian–for simply creating a wonderful time. Evening BBQs, jam sessions, camping trips and Laser Tag certainly helped keep the stress at bay.

I am also very grateful to the Collaborative Research Center (SFB) 630, not only for funding, but for creating a very educational interdisciplinary working atmosphere and allowing me to present my research on numerous occasions.

I am extremely grateful to my parents, my family and all my friends for their support throughout the past years. Finally, I thank my wife for her love, her support, her interest and most importantly for not making our stay in Würzburg a mere transit, but as yet the most awesome part of our lives.

Contents

Like everything, for Julia.

Chapter 1

Introduction

1.1 Early antibiotic research and resistances

The targeted design of chemical compounds to cure infectious diseases is a century-old endeavor. After many unsuccessful trials, Paul Ehrlich and Sahachiro Hata described in the year 1910 the compound Ehrlich 606, which was eventually marketed as Salvarsan, the most frequently prescribed drug against syphilis until the emergence of Penicillin [4]. A period in time followed, which is now known as the *Golden Age of antibiotics* from the 1940s to the 1960s, where half of the antibiotics commonly used today were discovered [5, 6]. During this period, various classes of antimicrobial agents were introduced (β-lactams, tetracyclines, chloramphenicol, aminoglycosides, macrolides, glycopeptides, streptogramines and quinolones) [5]. Together, all these classes cover a broad range of modes of action in the target cell.

A common misbelief of this time was that bacterial infections would soon be eradicated. However, since the *Golden Age*, the excessive therapeutical use of antibiotics and also the non-therapeutical use in animals have led to antibiotic resistant strains [6], which are severe health threats and cause a very high mortality rate [7]. Today, increased pharmaceutical marketing, excessive use of antibiotics in agriculture, food preservation or irrational self-medication, contribute to an exacerbation of the problem [5]. The origin of drug-resistances are manifold. *Mycobacterium tuberculosis* is intrinsically resistant to numerous drugs, due to its unique cell wall composition [8–10]. Because of very low fluidity and high fatty acid content, hydrophilic and lipophilic molecules likewise have severe problems passing the cell wall barrier [8]. In contrast to intrinsic resistances, pathogens can also acquire resistances. For example, *M. tuberculosis* strains with mutated catalase-peroxidase (KatG) can show therapy-resistance due to missing catalytic activity in transforming the prodrug isoniazid to its active form [11, 12]. Gram-positive bacteria, such as *Staphylococcus aureus*, generally show rather unrestricted uptake of antimicrobials due to a more permeable cell wall compared to Gram-negative bacteria or mycobacteria [13]. However, mutant strains of this pathogen can show diverse mechanisms for drug-resistance, including production of a thickened cell wall (vancomycin-intermediate resistant *S. aureus*; VISA) or extension of the proteome (methicillin-resistant *S. aureus*) [13, 14].

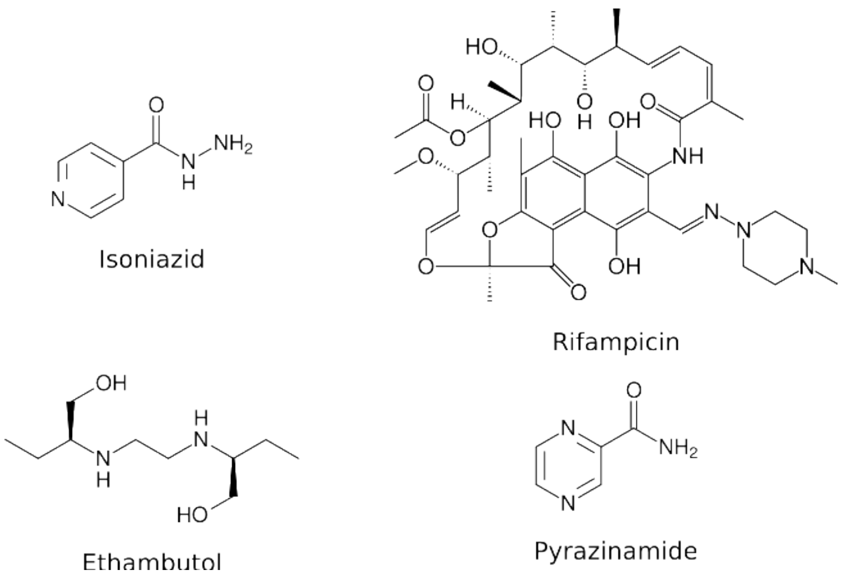


Figure 1.1 First-line drugs against tuberculosis: isoniazid, rifampicin/rifampin, ethambutol and pyrazinamide.

1.2 *Mycobacterium tuberculosis*

Mycobacterium tuberculosis is the primary causative agent of tuberculosis (TB). Although the death rate has dropped by 45% over the past two decades, TB is still a globally present disease. If untreated, the mortality rate can rise up to 66%. In 2013, 9 million new infections were documented and 1.5 million ended lethally. HIV/AIDS patients have a 26 to 31-fold higher probability of developing an active TB [15]. The classical antitubercular therapy–based primarily on cocktails of isoniazid, rifampicin, pyrazinamide, and ethambutol for a period of six months (Figure 1.1)–has cured over 56 million people since 1995, but the emergence of multi- and extensively drug-resistant strains of *Mycobacterium tuberculosis* (MDR-TB and XDR-TB) demands new, high-affinity inhibitor classes, which are unaffected by mycobacterial resistances [15–17].

It is estimated that 480,000 patients developed a multi-drug resistant TB in 2013, underlining the importance of antibiotic agents against MDR-TB and XDR-TB [15]. Resistant TB infections cannot be treated with the effective first-line anti-tuberculosis medication. Rather, the use of second-line antitubercular agents (e.g., fluoroquinolones, amikacin, kanamycin, capreomycin) is necessary, which comes with high cost, limited access and possibly severe adverse effects [15, 16, 18].

1.3 Methicillin-resistant *Staphylococcus aureus*

Staphylococcus aureus infections have severely affected the global health. The great peril of this pathogen are the emerged antibiotic-resistant strains, namely methicillin-resistant *S. aureus* (MRSA or *golden staph*) and vancomycin-resistant *S. aureus* (VRSA). These strains constitute a severe threat in hospital environments, especially for immunocompromised patients [5, 14, 19, 20]. In US hospitals the percentage of *S. aureus* infections caused by MRSA increased from 2.4% to 29% between 1975 and 1991 [14].

Infections with *S. aureus* can generally be treated with β-lactam antibiotics (e.g. methicillin), which is not possible for MRSA infections because of the penicillin-lactam-binding protein PBP2a. Furthermore, over 50% MRSA are also insensitive to macrolides, lincosamides, fluoroquinolones, and aminoglycosides. In many cases, the remaining remedy effective against an MRSA infection is the last-resort-antibiotic vancomycin [14]. This therapy, however, is futile for VRSA, rendering the need for new chemotherapeutics against *S. aureus* very urgent.

In the 2000s, a new lineage of MRSA has emerged, not limiting MRSA related infections to hospital environments and immunocompromised subjects [21]. Community-acquired (CA)-MRSA is globally spread and primarily induces skin infections [21]. Although the typical CA-MRSA is sensitive to most non-β-lactam antibiotics, isolates carrying plasmids with antibiotic-resistance genes have been found in the USA and Europe [22, 23], further highlighting the imperative of new efficacious antibiotic agents against MRSA.

1.4 Scope of this work—rational residence time modulation and permeability prediction to support antibacterial drug design

An early-stage parameter for *in vivo* efficacy profiling of a compound is vital for the effective development of novel therapeutics against these multidrug-resistant pathogens. The drug-target residence time is a valid indicator of *in vivo* activity for many targets (cf. Chapter 2.2 for detailed information) [24, 25]. Rational residence time modulation, however, is still very challenging, mostly due to the lack of structural information about the transition states of ligand dissociation [24, 26]. Molecular dynamics (MD) simulations provide valuable techniques to tackle this issue. With the recent increase in computational power, classical MD simulations are able to provide insight into transition states, as well as metastable intermediate states [26]. Thus, molecular determinants of long residence time can be detected and quantified in terms of receptor flexibility and conformational changes.

Part I of this thesis is focused on revealing the molecular determinants of drug-target residence time in bacterial enoyl-ACP reductases. In case of the mycobacterial enoyl-ACP reductase InhA, an important target for antitubercular drug design, it is still unclear which molecular processes actually govern ligand binding and, thus, the residence time. Since knowing these features is required for the rational optimization of ligand residence time, an extensive MD study was conducted, leading to novel strategies for rational InhA-inhibitor design [1, 26].

With current hardware, MD simulations can nowadays easily reach the microsecond timescale. It was shown in several studies that complete ligand binding events can thus be simulated using classical MD simulations [26–29]. However, drug-target residence times are not confined to microseconds, but can reach seconds, minutes or days [24–26, 30]. Accordingly, enhanced sampling techniques for MD simulations need to be employed to observe ligand association or dissociation (and corresponding transition states) and, thus, gain insight into drug-target kinetics. Although enhanced sampling techniques are numerous, a common difficulty is the feasible definition of one or several reaction coordinates (cf. Chapter 2.3 for further information). Steered molecular dynamics (SMD) simulations [31, 32] are fast and solely need the pulling direction of induced ligand withdrawal as a predefined parameter, while allowing access to free energy profiles along this reaction coordinate [33–36]. In recent studies, several enhanced sampling MD techniques (including SMD) have been employed to computationally assess information about ligand kinetics, although mostly on a qualitative level [37–44]. Here, the SMD methodology was combined with regression techniques to create a linear model for the quantitative prediction of residence time for the enoyl-ACP reductase FabI of *S. aureus*.

Regarding *M. tuberculosis*, slow-onset ligand binding to InhA is assumed to follow a multistep mechanism [17, 24, 45]. For rational residence time optimization it is important to correctly recognize and interpret the intermediate conformations of the protein-ligand complex (EI and EI* state, cf. Section 2.2 for further information). However, binding pocket and substrate binding loop conformations of InhA are highly divergent in several recently published crystal structures [40, 46]. Concluding Part I, the idea of the EI and EI* states of ligand association in the case of InhA is revisited by analysis of recent crystal structures and extensive (Steered) MD simulations.

A further complication for inhibitor design against *M. tuberculosis* is the largely impermeable cell wall (cf. Chapter 8.1 for further information) [8, 16]. Hence, rational efficacy improvement of drug candidates with respect to the drug-receptor residence time may still be ineffective, if the molecules lose their ability to pass the mycobacterial cell wall due to alterations in the physico-chemical properties. Also, the identification of novel inhibitor classes in screening campaigns may be limited in success by lack of compound permeability [16]. Accordingly, a better understanding of the physico-chemical composition of compounds active–and thus very likely permeable–against *M. tuberculosis* is desirable to help define the mycobacterial druggability space.

Part II of this thesis is, thus, focused on a data mining endeavor on physico-chemical descriptor data of active substances, leading to a permeability prediction model, wrapped in the online tool MycPermCheck [3]. To explore the permeability space of *M. tuberculosis*, MycPermCheck was eventually used in a virtual screening endeavor. The quality of initial screening hits as potential InhA inhibitors was investigated via docking, MD and SMD simulations with consideration of results of Part I.

Part I

Revealing the Molecular Determinants of Drug-Target Residence Times of Bacterial Enoyl-ACP Reductases using Molecular Dynamics Simulations

Chapter 2

Background

2.1 FAS II and enoyl-ACP reductases

2.1.1 Fatty acid synthesis type II

In contrast to mammals, in which fatty acid synthesis is based on a multienzyme complex (FAS I), some bacteria, plants and parasites utilize an alternate route for the production of fatty acids: the fatty acid synthesis type II (FAS II) cycle [47]. The elongation cycle of the FAS II consists of four catalytic reactions, while each cycle attaches two carbon atoms to the growing fatty acid chain. The intermediates are transported by the acyl carrier protein (ACP) [47]. First, the growing acyl-ACP is subjected to a condensation with malonyl-ACP to β-ketoacyl-ACP, catalyzed by the β-ketoacyl-ACP synthase I or II (FabB or FabF). Subsequently, the β-ketoacyl-ACP reductase (FabG) catalyzes the reduction of the β-keto moiety, yielding β-hydroxyacyl-ACP, followed by a dehydration by the enzyme β-hydroxyacyl-ACP dehydratase (FabZ or FabA). In the final step of fatty acid elongation, the enoyl-ACP reductase (FabI) catalyzes the hydrogenation of the substrate to acyl-ACP (Figure 2.1) [47].

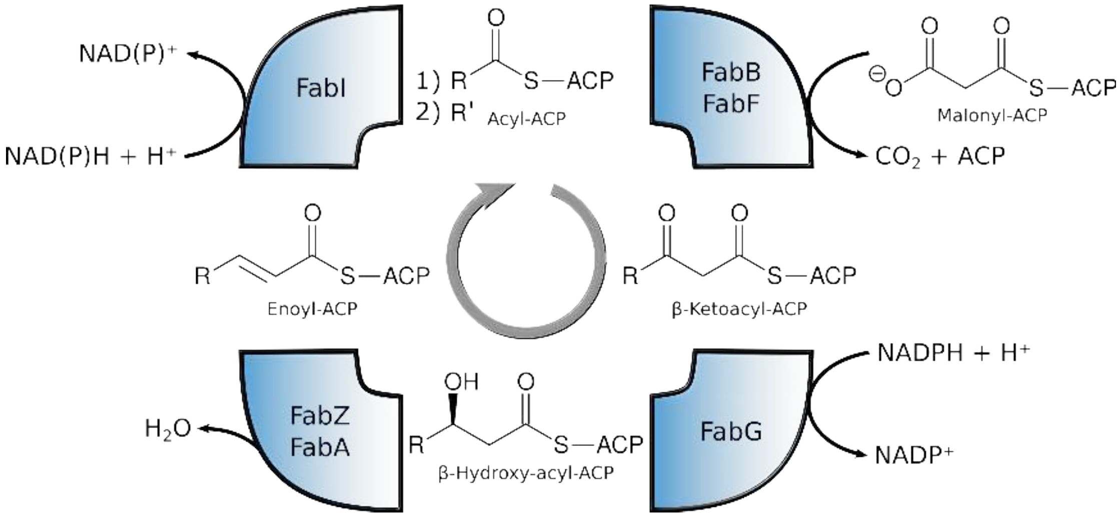


Figure 2.1 Elongation cycle of the fatty acid synthesis type II pathway. Each cycle, consisting of four catalytic steps, elongates the fatty acid substrate by two carbon atoms.

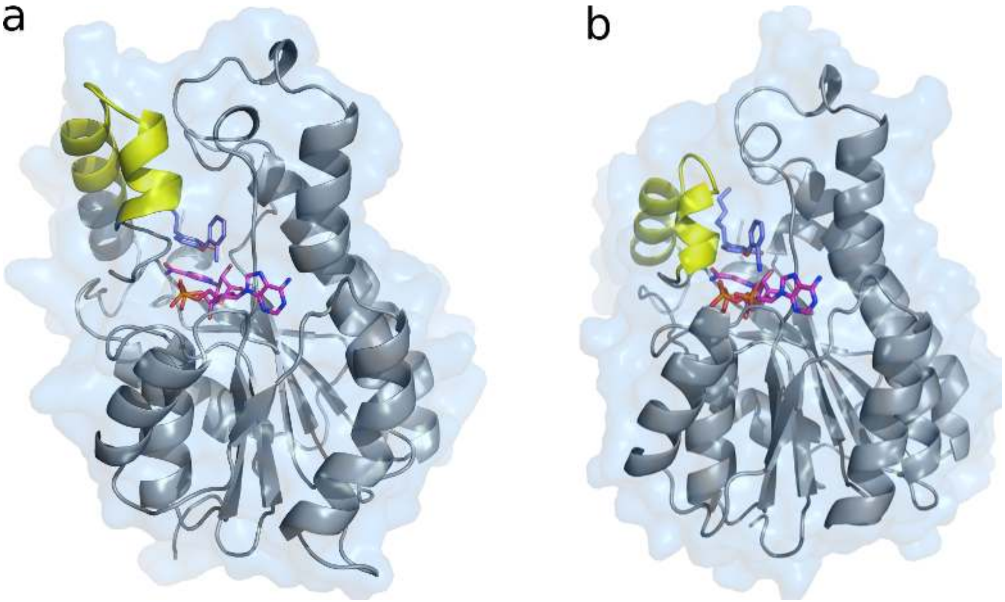


Figure 2.2 Monomeric subunit of (a) *M. tuberculosis* InhA (PDB code 2X23 chain A) and (b) *S. aureus* FabI monomer (PDB code 4BNN, chain A). The protein backbone is illustrated in gray. The ligand (**PT70** and **PT119**, respectively; slate blue) and cofactor (NAD$^+$ and NADP$^+$, respectively; magenta) are illustrated as sticks. The flexible substrate binding loop is colored yellow.

2.1.2 The enoyl-ACP reductases of *M. tuberculosis* and *S. aureus*

Since FAS II-pathogens differ fundamentally in this anabolic pathway from humans, single proteins of the FAS II are excellent drug targets, such as the enoyl-ACP reductase FabI. In *M. tuberculosis* the enzyme FabI is termed InhA (Figure 2.2a). It is inhibited by the first-line antituberculosis prodrug isoniazid (isonicotinic acyl-NADH-adduct after activation by the enzyme catalase-peroxidase KatG) and also weakly inhibited by the broad spectrum biocide triclosan (**TCL**) [48–52]. InhA is a member of the short-chain dehydrogenase/reductase (SDR) superfamily and is bioactive in the homotetrameric form [47]. The monomeric subunits contain an extended Rossmann-fold to bind the nucleotide cofactor [47]. Loop residues of helices $\alpha 6$ and $\alpha 7$ comprise a flexible region, the substrate binding loop (SBL), which closes upon substrate binding [47]. Both ends of the substrate binding pocket of the FabIs are solvent-exposed. These regions were termed major and minor portal, respectively, according to the degree of their exposure [53, 54]. The cofactor is bound at the bottom of the cavity. The nicotinamide moiety is oriented towards the inside of the binding pocket, whereas the adenine is oriented towards the major portal (Figure 2.3).

*Sa*FabI is a homolog to InhA and also a member of the SDR superfamily (Figure 2.2b). A BLAST search [55] of the wild-type sequence (GI: 109157150 [56]) against the UniProt database [57] revealed a sequence identity to *sa*FabI of 32% with a query coverage of

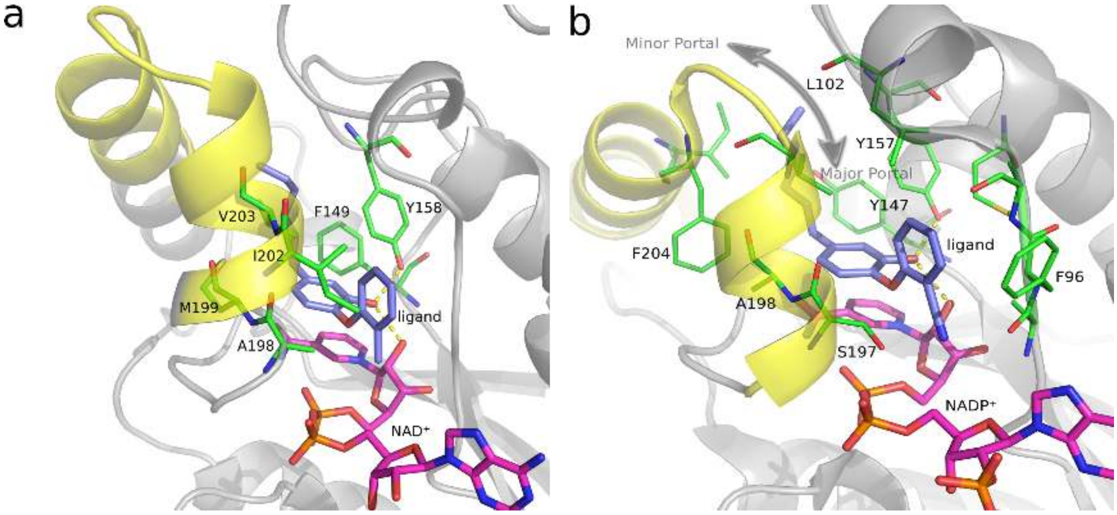
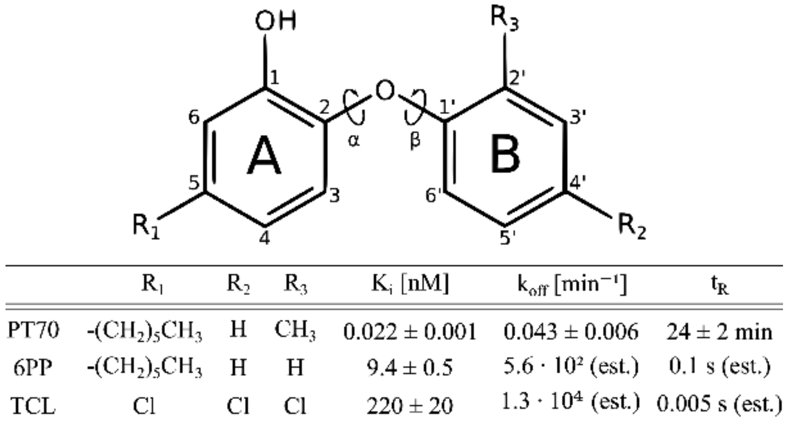


Figure 2.3 FabI binding pockets of (a) InhA crystal structure 2X23 and (b) *sa*FabI crystal structure 4BNN. The protein backbones are illustrated in gray, the substrate binding loops in yellow. The inhibitors **PT70** and **PT119** are represented in slate blue and the cofactors NAD$^+$ and NADP$^+$, respectively, in magenta. Important pocket residues are illustrated in green.

97% and a sequence similarity of 52%. The catalytic reaction of *sa*FabI is NADPH-dependent. The enzyme has a high cofactor specificity towards NADPH instead of NADH, which is a unique behavior among the bacterial FabIs [58]. The catalytic triad of *sa*FabI consists of the residues Tyr147, Tyr157 and Lys164 (Phe149, Tyr158 and Lys165 in InhA) [48, 59]. Substrate binding to *sa*FabI is assumed to take place via the major portal of the binding pocket [59].

2.1.3 Inhibition of *Mycobacterium tuberculosis* InhA by diphenylethers

InhA inhibitors act against mycobacteria by disabling the hydrogenation of the unsaturated precursors of the long and hydrophobic mycolic acids, which are necessary for proper construction of the largely impermeable *M. tuberculosis* cell wall [60]. Diphenylethers are one class of inhibitors currently under investigation. Unlike isoniazid, they bind directly to InhA without the necessity for prior activation by the enzyme catalase-peroxidase (KatG) [17]. Important protein-ligand interactions include a hydrogen bond between the ligand and Tyr158 and between the ligand and NAD$^+$, as well as several hydrophobic contacts to surrounding binding pocket residues, namely Phe149 and the residues Ala198, Met199, Ile202 and Val203, which are located in helix $\alpha 6$ of the SBL (Figure 2.3a) [45].



	R_1	R_2	R_3	K_i [nM]	k_{off} [min^{-1}]	t_R
PT70	-$(CH_2)_5CH_3$	H	CH_3	0.022 ± 0.001	0.043 ± 0.006	24 ± 2 min
6PP	-$(CH_2)_5CH_3$	H	H	9.4 ± 0.5	$5.6 \cdot 10^2$ (est.)	0.1 s (est.)
TCL	Cl	Cl	Cl	220 ± 20	$1.3 \cdot 10^4$ (est.)	0.005 s (est.)

Figure 2.4 Scaffold of diphenylether inhibitors and overview of the experimentally characterized diphenylethers analyzed in Chapter 3. The phenyl rings are referred to as A- and B-ring, respectively. The corresponding ether torsions are symbolized by curly arrows and labeled α and β. Experimental data were taken from [45] and references provided therein. **PT70** is a slow-onset inhibitor, with measured dissociation rate constant k_{off} and residence time t_r. In contrast, **6PP** and **TCL** show rapid-reversible binding kinetics; k_{off} and t_r values were estimated assuming a value of 10^9 $M^{-1}s^{-1}$ for k_{on}, as done by Luckner et al. (2010) [45].

Among the antitubercular diphenylethers, **PT70** displays slow-binding inhibition of InhA with a residence time (t_R; cf. Chapter 2.2 for detailed information) of 24 minutes at a K_i of 0.022 nM [45]. The broad spectrum biocide **TCL**, however, shows a rapid reversible inhibition of InhA, although it is a slow-binder in homologous enoyl-ACP reductases (Figure 2.4) [30, 58, 61–64]. In InhA, slow-binding inhibition is likely associated with the ordering of the SBL, which is the most flexible region of InhA [45, 65]. In fact, the crystal structure of the InhA-NAD$^+$-**PT70** complex (PDB code 2X23) shows an uninterrupted and highly ordered SBL, whereas in the crystal structure of the InhA-NAD$^+$-**TCL** complex (PDB code 2B35) the SBL is unresolved due to disorder [24, 64]. Thus, the highly ordered loop conformation very likely represents the final stage of the two-step binding mechanism (EI*) of the slow-binding inhibitor **PT70**.

Although these observations are experimentally well characterized, it remains unclear how the structural features of a ligand govern the binding mechanism and, hence, the actual residence time. Knowing these features is essential for rationally modulating the residence time as a key parameter in drug design, even more so as small differences in the ligand structure can dramatically affect the dissociation rate constant. Besides **PT70** and **TCL**, the diphenylether **6PP** can serve as an illustrative example: it differs from **PT70** by only a methyl group, but nevertheless shows rapid reversible instead of slow tight binding behavior (cf. Figure 2.4) [17, 64, 66].

Although InhA has been subject of several molecular dynamics studies, drug-target binding kinetics were generally not in the focus of these studies [65, 67–70]. The earliest

MD study of ligand-free InhA systems by Schroeder and colleagues (2005) [65] characterized structural changes over a sampling time in the single-digit nanosecond scale. The work of Pasqualoto *et al.* (2006) investigated thermodynamic properties of InhA systems bound to several isoniazid derivatives [67]. Subba Rao *et al.* (2008) [68] examined the stability and interaction patterns of a tripeptide inhibitor using short MD simulations. In 2010, Punkvang and colleagues [69] employed MD simulations to elucidate the dynamic behavior of arylamide inhibitors in InhA. In a recent study, Kamsri *et al.* (2014) [70] developed a series of diphenylethers and investigated their flexibility and binding free energy using MM/PBSA and normal mode methods [71, 72]. Recently, enhanced sampling methods were used to assess binding kinetics information of InhA computationally [40].

2.1.4 Inhibition of *Staphylococcus aureus* FabI by diphenylethers

As InhA, the enzyme *sa*FabI is part of the fatty acid synthesis (FAS) II cycle, which constitutes a well investigated drug target for *S. aureus* [17, 58, 73]. Currently, there are three *sa*FabI inhibitors in clinical trials: (1) AFN-1252 (Affinium Pharmaceuticals), which stems from the large-scale high-throughput screening (HTS) campaign of GlaxoSmithKline (GSK) between 1995 and 2001 [74, 75]; (2) CG400549 (CrystalGenomics), a diphenylether derivative with a 2-pyridone A-ring instead of a phenol to improve pharmacokinetics [76]; and (3) MUT056399 (Mutabilis), a 4-fluoro-substituted **TCL** derivative [77, 78]. Whereas **TCL** is only a weak inhibitor of InhA, which does not promote ordering of the SBL, it binds to *sa*FabI with a K_i of 0.05 nM and a residence time of 139.5 minutes [30]. In the *S. aureus* binding pocket, Tyr157 forms an important hydrogen bond interaction to the phenolate of the ligand (Figure 2.3b). The ligand also forms several hydrophobic contacts with surrounding residues located in the protein core or the substrate binding loop (SBL), namely with Ala95, Phe96, Leu102, Tyr147, Met160, Ser197, Ala198, Val201, Phe204, Ile207. Furthermore, the backbone of Ala97 is addressed by **TCL** via halogen bonds and the ligand forms a hydrogen bond and a π-π-stacking with the cofactor NADP$^+$ [58]. In the apo-form, the flexible SBL is disordered. Upon slow-onset inhibition, however, the SBL is ordered to helices $\eta 6$ and $\alpha 7$ [48, 58]. In recent literature, a series of diphenylethers was published with high-resolution crystal structures and experimental residence times from 2 to 700 minutes [30, 58]. The binding modes of these diphenylether inhibitors are very similar to that of **TCL**, which leads to two very conserved hydrogen bonds (to Tyr157 and to NADP$^+$) (Figure 2.3) [58]. Diphenylether inhibitors form the ternary complex with *sa*FabI bound to the oxidized cofactor NADP$^+$ [30].

2.2 Binding affinity and drug-target residence time

To obtain highly active inhibitors, projects in early drug discovery generally focus on optimizing the affinity of candidate compounds for a given target, which is the difference in free energy ΔG between the unbound and the final bound state. However, even for high-affinity inhibitors with K_i or K_d values in the low nanomolar range there is a potential activity gap between the *in vitro* assay experiments and a realistic *in vivo* system, where the exposure of target enzymes to drug-like molecules and the subsequent binding event can no longer be correctly described by equilibrium constants like K_d. Rather, the dissociation rate constant (k_{off}) of a protein-ligand complex, the reciprocal value of which describes the residence time (t_R) of a compound at a drug target, should be considered during rational drug-design endeavors [25]. To reduce dosage and increase efficacy, it is, thus, desirable to optimize potential drugs in terms of a long residence time (i.e., low k_{off}). Inhibitors exhibiting such low dissociation (and/or association) rate constants are termed “slow-onset inhibitors”, “slow-binding inhibitors” or briefly “slow-binders”. Although several different kinetic mechanisms are described for slow-binders, most of these inhibitors bind via an induced-fit mechanism [24]. The first initial complex (EI) is formed rapidly, whereupon a slower conformational change of the enzyme allows the ligand to form the final complex (EI*) (Figure 2.5). For such slow-binding ligands, k_{off} is a combination of multiple individual rate constants. In detail, k_{off} can be described by $k_{\text{-}1} \cdot k_{\text{-}2}$ divided by ($k_{\text{-}1} + k_2 + k_{\text{-}2}$); if $k_{\text{-}1}$ is large compared to k_2 and $k_{\text{-}2}$, k_{off} is essentially given by $k_{\text{-}2}$ [24].

The kinetics of slow-binders open a new possibility of efficacy improvement: instead of stabilizing the final state of ligand association EI* (i.e., lowering its free energy level), the transition state between EI and EI* can be destabilized. With a destabilized transition state, the rate constant of the reverse reaction is decreased, thus increasing the residence time of the ligand in the EI* state. However, rational residence time modulation still remains a challenge, since structural information about the transition state is generally not available [24, 26].

2.3 Computational methods for structural, energetic and kinetic characterization of protein-ligand complexes

Today, the field of computer-aided drug design encompasses a plethora of methods to accurately estimate the binding affinity of a compound of interest. Sophisticated scoring functions can evaluate crystal structure and docking poses to quantify the magnitude of enzyme-inhibitor interactions [79–81]. After identifying a valid binding conformation

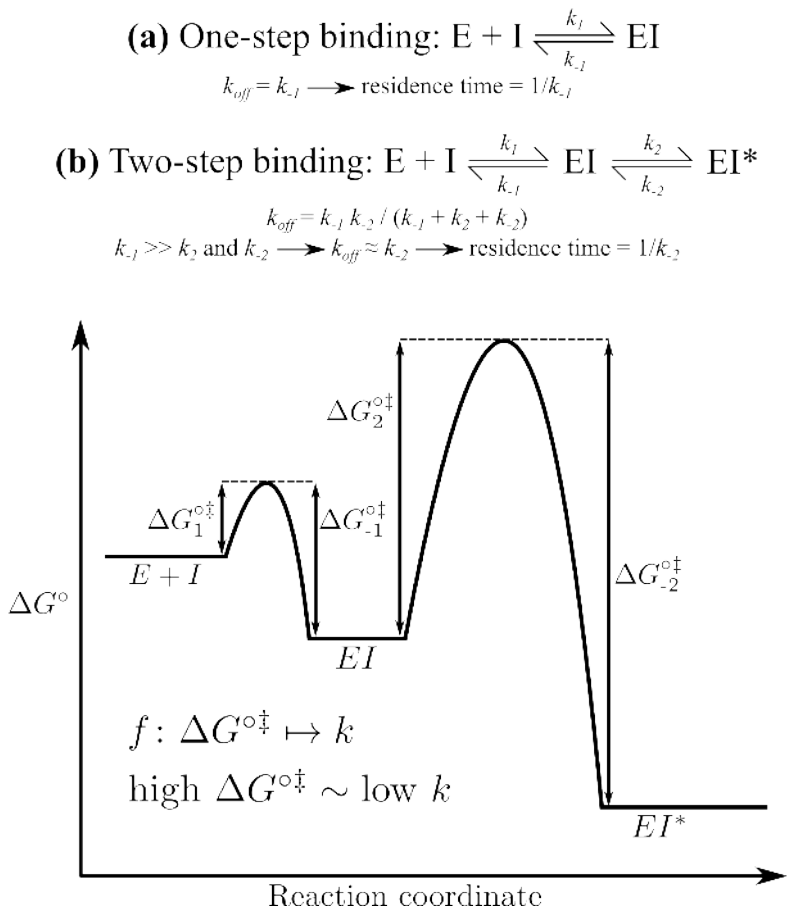

(a) One-step binding: $E + I \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} EI$
$k_{off} = k_{-1}$ → residence time = $1/k_{-1}$
(b) Two-step binding: $E + I \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} EI \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} EI^*$
$k_{off} = k_{-1}\,k_{-2} / (k_{-1} + k_2 + k_{-2})$
$k_{-1} >> k_2$ and k_{-2} → $k_{off} \approx k_{-2}$ → residence time = $1/k_{-2}$
ΔG°
$\Delta G_2^{\circ\ddagger}$
$\Delta G_1^{\circ\ddagger}$
$\Delta G_{-1}^{\circ\ddagger}$
$\Delta G_{-2}^{\circ\ddagger}$
$E + I$
EI
EI^*
$f\colon \Delta G^{\circ\ddagger} \mapsto k$
high $\Delta G^{\circ\ddagger} \sim$ low k
Reaction coordinate

Figure 2.5 Mechanisms of drug-target complex formation. **(a)** Equilibrium of inhibitor binding via a one-step mechanism. **(b)** Equilibrium of inhibitor binding via an induced-fit two-step mechanism represented as equation and schematic free-energy profile for this reaction. E denotes the enzyme, I the inhibitor, EI the initial enzyme-inhibitor complex, and EI* the final enzyme-inhibitor complex. A high energy barrier ($\Delta G^{\circ\ddagger}$) corresponds to a low reaction rate constant k.

with docking studies, a scoring function can estimate the binding affinity very quickly, which makes this method also applicable to large-scale screening endeavors. However, docking and scoring are generally unable to quantify rate constants or residence times of ligands, since they work on static enzyme-inhibitor complex snapshots.

MD simulations extend this static scenario by creating an ensemble of snapshots over time, i.e., a trajectory. Thus, important conformational changes over time can be detected and quantified. Eventually, extensive analysis of such trajectories can be used to qualitatively describe determinants of long residence time for a system. This approach is followed in Chapter 3 for the antimycobacterial drug target InhA. Chapter 4 provides follow-up work on this topic.

MD-based free energy methods, like Linear Interaction Energy (LIE) allow comparison of the free energy of bound and unbound states [80, 82]. The LIE method is based on creating a thermodynamic average of the ligand in its bound and its unbound state.

Thus, it considers only the end-points of protein-ligand-complex formation. The difference of non-bonded energies of the ligand between both states is then evaluated and incorporated into a multiple linear model [82]. The fundamentally different Free Energy Perturbation (FEP) method, on the other hand, divides the transition of the states into several smaller steps (perturbations), which are simulated independently [82]. In general, ligand binding is investigated with said method by use of a thermodynamic cycle to calculate the relative binding free energy between a ligand L and a structurally related ligand L' [83].

Although these methods have proven very valuable to assess binding affinities of ligands to proteins [82], residence times cannot be obtained, since the transition states of inhibitor association or dissociation are not considered.

Steered Molecular Dynamics (SMDs) [31, 32] provide indirect access to residence times by induced extraction of the ligand from the enzyme binding pocket. Thus, a continuous trajectory is generated, simulating a complete unbinding event along a putative reaction coordinate, while the associated conformational changes of protein and ligand can be directly monitored by the dissociating ligand. More importantly, the necessary force for ligand withdrawal is measured during the simulation. After numerical integration over the spatial reaction coordinate of multiple replica simulations, the resulting work profiles of the non-equilibrium SMD simulations can be converted to the Potential of Mean Force (PMF) profiles and hence to a free energy difference, an equilibrium property, using Jarzynski's equality (cf. Chapter 2.3.1 for details) [33–36].

According to Eyring's transition-state theory [84, 85], the necessary Gibbs energy $\Delta G^{\ddagger\circ}$ for a transition between two states of a system correlates to the rate constant k of the reaction:

$$k = \frac{k_b \cdot T}{h} \cdot e^{-\frac{\Delta G^{\ddagger\circ}}{RT}} \tag{2.1}$$

where k_b is the Boltzmann constant, T the temperature, h the Planck constant, $\Delta G^{\ddagger\circ}$ the Gibbs free energy and R the gas constant.

An alternative approach to computationally assess residence times and rate constants are Markov State Models (MSM). These models have been proven in several studies to serve as sophisticated kinetic models for the analysis of large-scale MD trajectory data of multiple replica simulations [86–89]. The ultimate goal of an MSM approach is the generation of a parsimonious model of transition states in MD trajectories to predict experimental data (e.g., residence times) quantitatively. In the last few years, MSMs have successfully been utilized in the field of rational optimization of ligand residence time [29, 90]. A disadvantage of this method is the large amount of trajectory data necessary for MSM generation.

In contrast to MSMs, numerous free energy methods using enhanced sampling techniques are available to address the issue of limited sampling in MD simulations. In recent literature, investigation of ligand unbinding and/or computational estimation of k_{off} values and residence times were published using non-equilibrium SMD simulations [31, 32] or equilibrium methods, such as Metadynamics [91], Partial Nudged Elastic Band (PNEB) [92], Umbrella Sampling [93, 94] and scaled MD simulations [95] (cf. next section for methodological details) [37–44]. Generally, these methods are able to reconstruct free energy surfaces along one or multiple predefined reaction coordinates. Here, SMD simulations are employed to create an accurate and quantitative residence time prediction model for the antistaphylococcal drug target FabI (cf. Chapter 5 for further information). Chapter 6 revisits the concept of the EI and EI* state of InhA inhibition using both classical and steered MD simulations.

2.3.1 Principles of molecular dynamics simulations

Biomolecules are not rigid systems, as static textbook images and 3D structures might suggest. Representations often merely illustrate snapshots among a vast ensemble of possible states in the energetic landscape of conformations [96, 97]. In drug design, consideration of the dynamic nature of proteins is particularly important in studying ligands that bind to their target via an induced-fit mechanism [24]. By leaving a simple two-state model for enzyme-inhibitor association (cf. Figure 2.5a) and extending it to a more complex multi-state model (cf. Figure 2.5b), a more accurate description of the drug-target interaction can be achieved, taking not only the thermodynamic states, but also the transitions into account [96]. The energy barrier between two conformational states of a protein, and thus the likelihood of a transition, strongly depends on the kind of conformational change. Whereas atomic bond vibration happens in femtoseconds, the time scale increases gradually from side chain rotation over loop motion to movement of a whole protein domain (Figure 2.6).

2.3.1.1 Molecular mechanical force fields

In order to capture the motion of a molecular system over time, it is necessary to quantify atomic interactions. A system containing a biological macromolecule (e.g., a protein) surrounded by explicit water molecules and ions, can easily consist of 100,000 atoms and more. Obviously, evaluation of forces between these atoms using a quantum mechanical approach is computationally far too expensive. Molecular mechanical force fields constitute a simplified way to describe atomic interactions [98]. The system is reduced to “balls and springs”. Each atom is represented as a sphere with a fixed

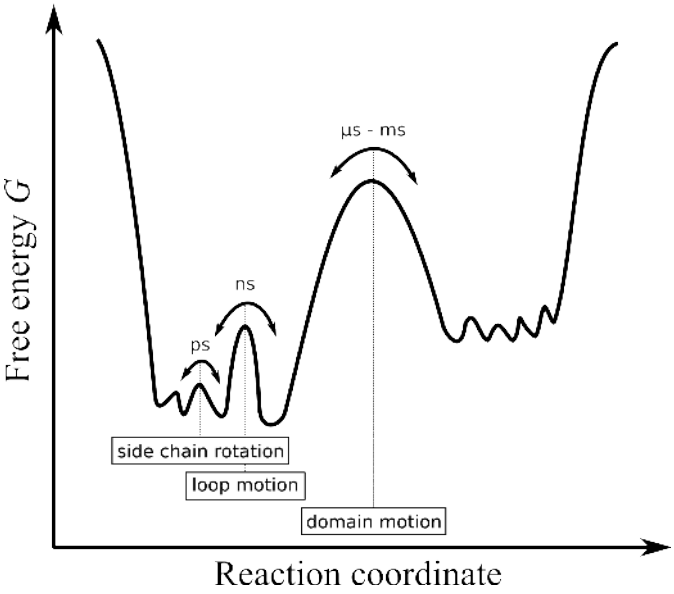


Figure 2.6 Free energy diagram of conformational transitions. A higher energy barrier between equilibrium states reduces the rate constant of the corresponding conformational change. Energy barriers are generally higher for large conformational changes, such as protein domain motion, compared to side chain rotation or loop motion.

radius and charge. Bonded forces that act on these atoms are treated as springs in a classical Newtonian way. The force field potential is therefore a sum of several terms, which generally include the bond lengths, bond angles and torsion angles, as well as non-bonded (electrostatic and van-der-Waals) interactions. The bond length and angle energies are described by a harmonic potential, the torsion potentials follow a periodic cosine function. The electrostatic interactions are evaluated according to the Coulomb-potential and the van-der-Waals interactions are represented by a Lennard-Jones-(6-12)-potential (Figure 2.7) [99, 100]. Hence the interaction potential of each atom is a function of the atomic coordinates. If $E(\mathbf{r})$ is the scalar potential energy function, the forces that act on the atoms are given by the negative gradient $\vec{F} = -\nabla E(\mathbf{r})$ [101, 102].

The total potential energy of the system E_{total} is the sum of all intra- and intermolecular potentials in the system (Figure 2.7). In the microcanonical (NVE) ensemble, the particle number N, the volume V and the potential energy E are constant. The system has no heat exchange, thus the simulation corresponds to an adiabatic process. Since without temperature or pressure control E_{total} is approximately constant, the ensemble is not suited for energetic equilibration of the system. By introducing temperature control to the system via a coupled heat bath, energetic equilibration is possible in the canonical (NVT) ensemble. Temperature control can be achieved, for instance, by use of weak coupling to a thermal bath [103] or Langevin dynamics [104, 105]. After energetic equilibration of E_{total} and heating of the system to the desired temperature

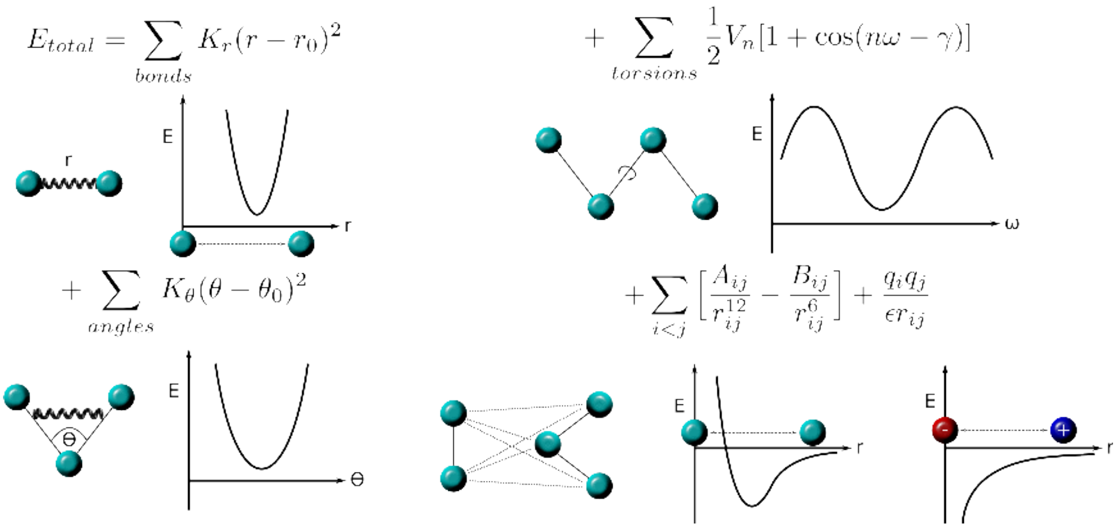

$E_{total} = \sum_{bonds} K_r(r - r_0)^2$
$+ \sum_{torsions} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)]$
$+ \sum_{angles} K_\theta(\theta - \theta_0)^2$
$+ \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right] + \frac{q_i q_j}{\epsilon r_{ij}}$
r
E
r
E
ω
Θ
E
Θ
E
r
E
-
+
r

Figure 2.7 Terms of a force field according to [100]. The total potential is comprised of functions for bond lengths, bond angles, torsion angles and functions evaluating the non-bonded interactions (Lennard-Jones-potential and Coulomb potential). E is the energy, K_r, K_θ and V_n are force constants, r is the distance between two atoms, θ is the bond angle, r_0 and θ_0 are reference values at equilibrium, n is the dihedral multiplicity, ω is the dihedral angle, γ is the dihedral phase, A and B are parameters incorporating the Lennard-Jones potential well and the distance at which the potential is zero, q is the atomic charge and ϵ is the dielectric constant.

T, the isothermal-isobaric (NPT) ensemble can be achieved by additionally introducing pressure control to the system, for instance via the Nosé-Hoover Langevin piston barostat [106, 107].

All particles of the system are contained in a three-dimensional body, e.g., a virtual rectangular box. Periodicity of a system in all dimensions is used to avoid artifacts of the system surfaces [101]. Hence, particles which leave the periodic box on one side are replaced by images on the opposite side [101]. Naturally, the periodicity increases the number of atom pair interactions drastically. Whereas van der Waals interactions are rapidly decaying and can thus be truncated, the evaluation of long-range electrostatic interactions can be computationally expensive, since they decay very slowly (with r^{-1}; cf. Figure 2.7) and their range often spreads over half the box length [98, 101]. To accurately account for these long-range interactions, state-of-the-art MD simulations employ the particle-mesh Ewald (PME) methodology to describe long-range electrostatic interactions for a system with periodic boundary conditions (PBC) [101, 108].

2.3.1.2 Parameterization of protein-ligand systems

Since molecular mechanical force fields are not *ab initio*, but empirical models, accurate determination of the protein, solvent as well as ligand parameters is crucial for the quality of the simulation outcome. The AMBER force field [99] is well validated and widely used for simulation of biomolecular systems in aqueous solution [109]. Atomic charges are derived from electrostatic potentials obtained from quantum mechanical calculations at the Hartree-Fock (HF) 6-31G* level using the restrained electrostatic potential (RESP) methodology [99, 110, 111]. The AMBER force field *ff99SB* is not polarizable and, thus, uses fixed point charges. As a result, the force field model has limited responsiveness to a changing molecular environment [109]. However, it is known that charge fitting to potentials of the HF/6-31G* basis set is apt to overestimation of bond-dipoles with respect to gas phase values. This effect results in bond-dipoles rather comparable to values in empirical water models, like the TIP3P model [109, 112]. TIP3P is a three-site water model and, thus, a simple model compared to four- or five-site water models (e.g., TIP4P or ST2 [113]). A principal disadvantage of water models of higher complexity is, however, the drastically increased computational demands [98].

To assess force field parameters for ligands, Wang *et al.* (2004) developed the General AMBER Force Field (GAFF) [114], which is able to provide AMBER force field parameters for a wide range of organic molecules composed of H, C, N, O, S, P, and halogens. For atomic charge derivation, GAFF uses the RESP method [110, 111] from electrostatic potentials obtained with the quantum mechanical software suite Gaussian [115] at the HF/6-31G* level.

2.3.1.3 Equations of motion and their integration

Based on Newton's second law of motion, $F_i = m_i \cdot a_i$, the acceleration a_i of a particle i with the mass m_i is proportional to the force F_i acting on it [102]. By numerical integration of the acceleration a_i with respect to time t, a new atomic position $\mathbf{r}_i(t_1)$ can be calculated for the next time step $t_1 = t_0 + \Delta t$, depending on the interaction forces at time step t_0 (Figure 2.8):

$$F_i = m_i \frac{d^2\mathbf{r}_i}{dt^2}. \tag{2.2}$$

At the beginning of an MD simulation, a random initial velocity is assigned to each atom based on a Maxwell-Boltzmann-distribution, which provides the probability of an atom i with the mass m_i having a velocity v_i at the temperature T [98]. By choosing a reasonably small time step between iterations and then alternating force evaluation and integration of accelerations, the investigated system can evolve in time and space subject to the interactions defined by the molecular mechanical force field [100, 102]. The time step is an important parameter, which deserves careful consideration. Whereas a too small time step will result in limited conformational space, a too big time step may cause

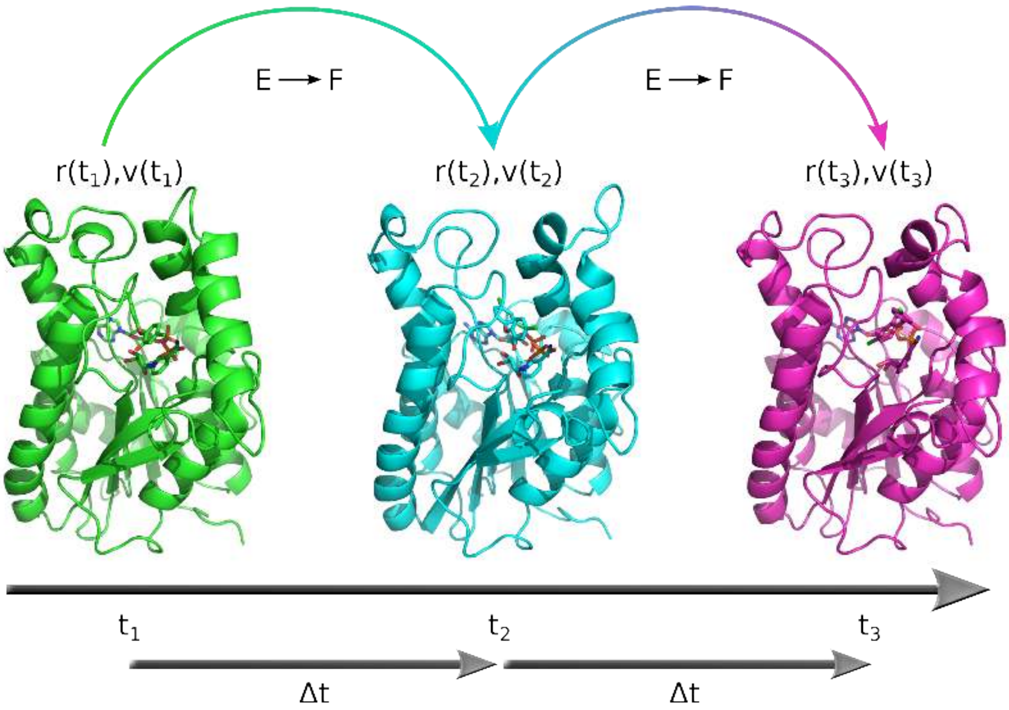


Figure 2.8 Schematic illustration of an MD simulation. For each time step Δt, the atomic position $\mathbf{r}(t_i)$ and velocity $v(t_i)$ is evaluated for every atom i. The underlying forces F_i are derived from the energy function of the molecular mechanical force field. Figure adapted from [102].

instabilities during numerical integration [98]. Depending on the desired resolution of the different types of motion, different time steps are recommended (Table 2.1).

In order to use a time step of ≥2 fs, it is necessary to apply a constraint algorithm on bonds to avoid impairing the accuracy of the simulation. As a result, the spatial motion of these atoms is no longer independent, but coupled [98]. A common implementation of a constraint algorithm in MD is the SHAKE methodology [116]. SHAKE uses holonomic constraints, i.e., the constraints can be expressed as

$$f(q_1, q_2, q_3, \ldots, t) = 0 \tag{2.3}$$

with q_1, q_2, etc., as the coordinates of the particles [98]. Since hydrogen vibrations constitute the oscillation of the highest frequency in the system, it is particularly important to apply constraints to all bonds involving hydrogen.

Table 2.1 Different types of motion with suggested time steps according to [98].

System	Types of motion present	Suggested time step
Atoms	Translation	10 fs
Rigid molecules	Translation, rotation	5 fs
Flexible molecules, rigid bonds	Translation, rotation, torsion	2 fs
Flexible molecules, flexible bonds	Translation, rotation, torsion, vibration	1 or 0.5 fs

2.3.1.4 Enhanced sampling techniques

Although MD simulations are very fast compared to quantum mechanical computations and can to date easily reach the nano- to microsecond time scale, important molecular events, such as protein folding or the dissociation of ligands, may still not be observed due to high energy barriers between the equilibrium states [101]. Accordingly, a plethora of techniques has emerged to enhance the sampling of the phase space in an MD simulation and to allow the evaluation of the free energy change along a reaction coordinate, the Potential of Mean Force (PMF). A traditional method for this purpose is Umbrella Sampling (US) [93, 94]. In US, at least one additional reaction coordinate ξ is introduced to the MD simulation, where ξ can have various forms (e.g., distance, torsion, RMSD) as long as it provides distinction between two thermodynamic states [94]. The probability distribution of the system along ξ is, thus, given by:

$$Q(\xi) = \frac{\int \delta[\xi(\mathbf{r}) - \xi] \exp[(-\beta E) d^N\mathbf{r}]}{\int \exp[(-\beta E) d^N\mathbf{r}]} \tag{2.4}$$

with $\beta = 1/(k_bT)$, k_b being the Boltzmann constant, T the absolute temperature, N the number of degrees of freedom, $\mathbf{r}$ the positional configuration of the system and E the total potential energy, assuming E is independent of the momentum [94]. Hence, the free energy along the reaction coordinate ξ, the PMF, can be assessed by [94]

$$G(\xi) = -k_bT\ ln\ Q(\xi). \tag{2.5}$$

In an ergodic system, the ensemble average of the phase space is equal to its time average $P(\xi)$, assuming infinite sampling, and is thus accessible via MD simulation [94]. In reality, however, regions in configuration space in the vicinity of an energy minimum are generally sampled well, whereas the probability declines for sampling regions with higher energy [94]. To overcome this issue, US is commonly run in parallel in several small windows with varying values of ξ. In each window i, a biasing potential ω_i is added as an additional energy term:

$$E^{biased}(r) = E^{unbiased}(r) + \omega_i(\xi). \tag{2.6}$$

Generally, a harmonic bias of the form $\omega_i(\xi) = \frac{1}{2}k(\xi - \xi_i^{ref})^2$ with a force constant k is used to ensure that the system stays close to the reference ξ_i^{ref} of the respective window i [94]. Using US analysis methods, such as the Weighted Histogram Analysis Method (WHAM) [117, 118] or umbrella integration [119], the multiple MD simulations (one for each window) can be combined and the PMF eventually derived from $P(\xi)$ [94]. A major advantage of US is its wide applicability, due to the variability of the reaction coordinate [94]. However, the application on protein-ligand systems with a dissociating ligand is rather inconvenient, since the *"ligand needs to be equilibrated at each window before moving it to the next window"*[1] [120]. Furthermore, the selection of appropriate reaction coordinates can be challenging for US, as well as related enhanced sampling techniques, such as Metadynamics [91], in which a sum of Gaussians comprises a history-dependent potential added to the MD simulation to eventually escape energy minima [121].

Another approach for enhanced sampling is accelerated MD (aMD) [122, 123]. In contrast to US and related methods, it is not necessary to predefine a reaction coordinate [123]. In aMD, sampling of the phase space is extended by boosting the potential energy function $V(r)$ if $V(r)$ is below a certain threshold. Energy minima are, thus, elevated and transitions occur more frequently [122]. In this context, free energy profiles can be reproduced from the enhanced sampling by post-processing via Boltzmann re-weighting of structural parameters, such as the distribution of atomic coordinates on the PC1-PC2-plane of a Principal Component Analysis (PCA) or torsion angles [123]. Problematic in the aMD approach are, however, the large fluctuations in the boost potential $\Delta V(r)$ [95]. Scaled MD simulations [95] try to address this issue by flattening and smoothing the potential energy surface via scaling $V(r)$ by a factor λ between 0 and 1, giving the population distribution function $p^*(r) = e^{-\beta\lambda V(r)}$ [95]. The canonical population distribution can accordingly be derived with the re-weighting equation $p(r) = p^*(r)^{1/\lambda}$.

Although enhanced sampling methods without prior definition of a reaction coordinate are comfortable and reduce the risk of possible pitfalls, a ligand unbinding event can not be forced. US and Metadynamics are sound methods for this task, require, however, carefully defined reaction coordinates. Hence, the non-equilibrium MD method Steered Molecular Dynamics (SMD) [31, 32] has gained increasing popularity for the simulation of ligand dissociation [120]. In contrast to US and Metadynamics, the reaction coordinate (i.e., the pulling direction of the ligand) is relatively simple to determine. Moreover, only the initial system needs to be equilibrated [120]. The SMD approach will be explained in detail in the following section.

2.3.1.5 Steered Molecular Dynamics

In contrast to unbiased classical MD simulations, SMD introduces a guiding potential to the simulation. In constant velocity SMD (cvSMD), a dummy atom moves along a spatial reaction coordinate with constant momentum, while it is attached to the center of mass of an atom selection ("SMD atoms") via a stiff, virtual spring. As the dummy atom

[1] Baştuğ *et al.*, 2008, J. Chem. Phys. 128(15), 155104.

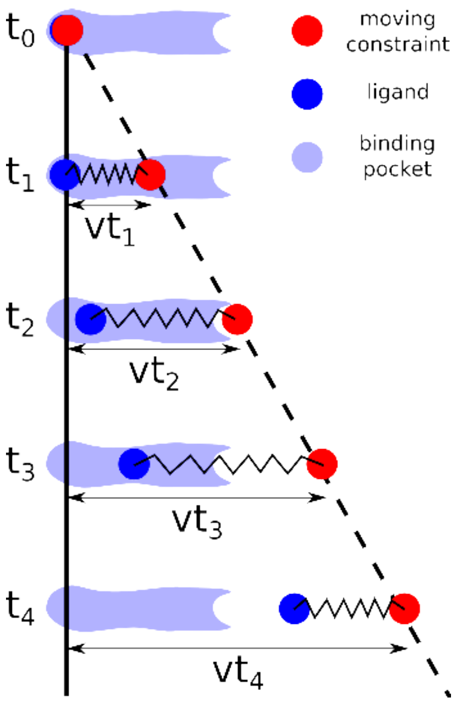


Figure 2.9 Schematic illustration of constant-velocity Steered Molecular Dynamics. The dummy atom (red) is attached to the ligand (blue) via a stiff, virtual spring as it moves along a predefined reaction coordinate, extending the spring and eventually dragging the ligand with it out of the binding pocket. Figure adapted and extended from [101].

or "moving constraint" moves along the pulling direction the spring is being extended and eventually the SMD atoms follow the dummy atom (Figure 2.9).

Necessary conformational changes are, thus, automatically initiated in the process. Simultaneously, the force on the spring is measured in defined intervals, yielding an individual force profile over distance traveled by the moving constraint for each simulation:

$$\vec{F} = -\nabla U \tag{2.7}$$

$$U = \frac{1}{2}k[\lambda - (\vec{r} - \vec{r}_0) \cdot \vec{n}]^2 \tag{2.8}$$

$$\lambda = v \cdot t \tag{2.9}$$

where $\vec{F}$ is the measured force, U the potential energy, λ the distance of the moving constraint along the reaction pathway, k the spring constant, $\vec{r}$ the current position and $\vec{r}_0$ the initial position of the SMD atoms' center of mass, $\vec{n}$ the direction of pulling, v the pulling velocity and t the time.

A challenge for all MD simulations with additional guiding potential is the predefinition of a suitable reaction coordinate. In case of SMD, this reaction coordinate is the direction of pulling, i.e., the vector along which the moving constraint travels with constant velocity. The approach followed in this study is based on Random Accelerated Molecular Dynamics (RAMD) [124, 125]. Here, a constant acceleration is applied on a chosen group of atoms. The method is similar to SMD, however with an important difference: the reaction coordinate (direction of travel) is not defined in advance, but randomly assigned to the chosen group of atoms. If the selected RAMD-atoms do not move a specified minimum distance in time, the direction is mutated; otherwise the reaction coordinate remains unchanged. Applied on a ligand in a binding pocket, RAMD provides a way to find a suitable egress pathway. Hence, the important choice of a proper pulling direction in SMD is well supported by the directionally non-parametric RAMD approach.

2.3.2 Evaluation of MD data

The possibilities for analyzing MD simulations are numerous. In contrast to free energy calculations based on MD simulations, classical MD simulations often have the resulting trajectory of the system in the spotlight of investigation. To quantify the relative atomic displacement over time with reference to another atom, the atomic distance can be utilized. On the other hand, to assess the absolute atomic displacement of an atom selection with reference to the same atoms in the starting structure, the root-mean-square deviation (RMSD) is very useful:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \tag{2.10}$$

where d_i is the distance between each of the N atom pairs of equivalent atoms.

2D-RMSD and clustering. Atomic displacement of a selection can not only be evaluated with respect to the starting structure, but also with respect to all other frames along the trajectory, resulting in a matrix of RMSD values. This matrix can then be illustrated as heatmap with the RMSD values as color code. The major advantage of this analysis is the symmetrical nature of the 2D-RMSD matrix, which can directly be used as a distance metric for a hierarchical clustering to capture recurring molecular conformations.

In general, two important choices have to be made regarding algorithms in a clustering: the distance metric and the linkage method [126]. The distance is the 2D-RMSD, a ready-to-use symmetrical matrix of the RMS deviation of each conformational snapshot against every other. The selection of a proper linkage method needs careful consideration, since different linkage methods may lead to drastically different results. The single-linkage clustering, in which the distance between two clusters is the minimum distance between the respective elements, ensures that all elements of two neighboring clusters have a minimum cutoff distance. A major drawback, on the other hand, is the "chain building" phenomenon, which may lead to many singleton clusters, hampering the interpretability [126]. Thus, clusters produced by a single-linkage algorithm tend to have a large diameter. Conversely, the complete-linkage method, which considers the maximum distance between the respective elements of clusters for fusion, is apt to producing few clusters with similar diameter. A shortcoming of the method is that two elements of different clusters may actually be closer than two elements of the same cluster. The principal advantage, however, is that the method ensures a maximum cutoff diameter for a cluster. Although numerous advanced linkage methods were designed to compensate for the mentioned shortcomings (e.g., UPGMA/Average, Ward, McQuitty), the complete-linkage method is a very good choice for the interpretation of structural data, since all conformational snapshots within a cluster have a maximum RMSD according to a chosen cutoff value, which is an important and intuitive information [127, 128].

2.3.3 Evaluation of SMD data

2.3.3.1 Reconstruction of work profiles

In SMD simulations, the force on the virtual spring between moving constraint and SMD atoms is measured over time. To yield the performed work, this force profile can be numerically integrated over the traveled distance of the moving constraint, using cumulative sums [35]:

$$W_{0\to\lambda} = \sum_{0}^{t} \frac{\vec{F}(\lambda_t) \cdot v \cdot dt'}{69.479\ mol} \tag{2.11}$$

where W is the performed work, t is the time, $\vec{F}$ is the measured force, λ is the distance along the pulling direction, v is the constant pulling velocity, dt' is the time step size and $(69.479\ mol)^{-1}$ is the NAMD conversion factor from $pN \cdot \mathring{A}$ to $kcal/mol$.

2.3.3.2 Reconstruction of PMF profiles using Jarzynski's equality

Since SMD simulations are non-equilibrium simulations, the PMF $\Phi(\lambda_t)$ cannot be extracted directly from the simulation data. Provided that all requirements for the stiff-spring approximation are fulfilled, the PMF relates to the Helmholtz free energy: $\Phi(\lambda) \approx F_\lambda$ [36]. Jarzynski discovered an equality to convert the work performed in an ensemble of SMD experiments to the free energy change, thus offering access to this quantity [33–36]:

$$e^{-\beta\cdot\Delta F_e} = \langle e^{-\beta\cdot W}\rangle \tag{2.12}$$

$$\text{with } \beta = 1/(k_b T) \tag{2.13}$$

where ΔF_e is the free energy change through exponential averaging, W is the performed work in an SMD experiment, k_b is the Boltzmann constant, T is the temperature and brackets $\langle\cdot\rangle$ illustrate the ensemble average. The accuracy of the exponential average is generally limited by high pulling velocities and a small number of replica simulations. However, cumulant expansions of the Jarzynski equality can be used as approximations for finite sampling [35]. Thus, the cumulants up to the second order are typically used for free energy calculation:

$$\Delta F_1 = \langle W\rangle \tag{2.14}$$

$$\Delta F_2 = \langle W\rangle - (\beta/2)(\langle W^2\rangle - \langle W\rangle^2) \tag{2.15}$$

It was proven that the Jarzynski equality also holds true for the change in Gibbs free energy ΔG in an isobaric-isothermal ensemble, besides the Helmholtz free energy change ΔF in a canonical system [36, 129]. Hence, the following modification of Jarzynski's equality are valid in the NPT ensemble:

$$e^{-\beta\cdot\Delta G_e} = \langle e^{-\beta\cdot W}\rangle \tag{2.16}$$

$$\text{with } \beta = 1/(k_b T) \tag{2.17}$$

as well as the corresponding cumulant expansions.

2.4 Analysis tools

All calculations and statistical analyses in Part I of this thesis were conducted using the statistical framework R and the associated plug-ins lattice, cluster, pheatmap, games, gap, aicc, AICcmodavg, vioplot and bio3D [130–139]. Trajectory analyses were carried out with VMD 1.9.1 and the incorporated extensions RMSD Trajectory Tool and Timeline [140, 141]. Visualizations were created with PyMOL [142]. 2D-RMSD plots were drawn with a tailored python script by Raphael Dives (University of Würzburg).

Chapter 3

Slow-onset inhibition of *Mycobacterium tuberculosis* InhA: Revealing molecular determinants of residence time by MD simulations

The contents of this chapter have been published in the open-access journal PLoS ONE in 2015 [1]. The publication has been modified in layout to fit the style of this thesis. Moreover, the supporting information of the publication and previously not shown data have been incorporated into the chapter. The theoretical background of this work is explained in Chapters 1 and 2.

3.1 Introduction

As loop ordering and related conformational changes upon ligand binding are the most likely key factors in the context of slow-onset inhibition, we have conducted an extensive computational survey to elucidate the effects of different ligand structures on InhA conformational dynamics by means of molecular dynamics (MD) simulations. To this aim, five systems were prepared for simulation: (1) the unmodified InhA crystal structure with bound **PT70** and NAD$^+$ (PDB code 2X23) [45], (2) the same crystal structure without inhibitor (i.e., after removing it; hereinafter called *perturbed*), and (3) the same crystal structure without ligand and cofactor (hereinafter called *No NAD$^+$*). Furthermore, based again on PDB structure 2X23, complexes of InhA with NAD$^+$ and the rapid reversible inhibitors (4) triclosan (**TCL**) and (5) **6PP** were setup (cf. Figure 2.4). By starting all simulations from the highly ordered 2X23 crystal structure, it is possible to analyze perturbation effects and to reverse-engineer the potential EI*-complex formation. Placing **TCL** or **6PP** in the closed-SBL conformation of 2X23 enables the simulation of the virtual EI*-state of an InhA-NAD$^+$-**TCL** or -**6PP** complex and examination of the dynamic properties that might eventually lead to loop disordering.

With simulations based on these systems we aim at revealing features of the conformational dynamics of the binding pocket and the SBL of InhA while linking them to structural differences in the respective ligands. Understanding the benefits and disadvantages of ligand properties in this context has implications for inhibitor design and optimization toward a longer residence time.

Accordingly, this study is based on three major hypotheses: (1) The ternary complex of **PT70** with InhA and NAD$^+$ represents the EI* state of the system. According to the current literature, there is no doubt on the validity of this assumption [40, 45, 46]. (2) As recently suggested by Li et al. [40], the EI state most likely corresponds to the open conformation of helix $\alpha 6$ (SBL) with respect to the binding pocket. This open conformation is observed, for example, in a substrate-analogue complex of InhA (PDB code 1BVR [53]). In contrast, the EI* state seen in the **PT70**-complex is characterized by a closed conformation of helix $\alpha 6$. (3) In the presence of inhibitors with rapid reversible binding kinetics, the EI* state is destabilized relative to the EI state. Therefore, after association of such an inhibitor, the EI* state is not reached, at least not to an observable extent. Conversely, placing a rapid reversible inhibitor in an EI* structure should cause its destabilization and eventually lead to the EI state. While experimentally hardly accessible, such a process can be investigated computationally. As illustrated in the schematic free energy profile of Figure 3.1, destabilization of the EI* state in the presence of a rapid reversible inhibitor (or in the absence of an inhibitor) may lower the barrier to such an extent that a transition from EI* to EI could become observable within the time scale of standard unbiased MD simulations. This is the rationale for setting up the simulations with the inhibitors **6PP** and **TCL** placed in the binding pocket of the **PT70**-InhA crystal structure 2X23. The question then is whether and to which extent the EI* state is left under such conditions, whether an EI state is indeed reached and how all of this depends on the nature of the ligand.

In light of these hypotheses and questions, the outline for the analysis of the trajectories and the presentation of the results is as follows: We first focus on the binding pocket dynamics and examine the conformations observed in the **6PP**- and **TCL**-bound systems in comparison to the **PT70**-complex. To this aim, we perform a hierarchical cluster analysis on the basis of 2D-RMSD data of the three trajectories to reveal the conformational families visited by the simulations. This is followed by a closer analysis of the dynamics of the SBL, as well as of the ligand binding modes and the hydrogen-bond interactions. We attempt to link the observations in the different complexes to differences in the ligands, examining in particular the effect of the *ortho*-methyl substitution of **PT70**. We finally discuss the conformational families in the context of available experimental information, especially with respect to the presumed EI and EI* states. We conclude with a discussion of the implications for drug design and rational residence time modulation.

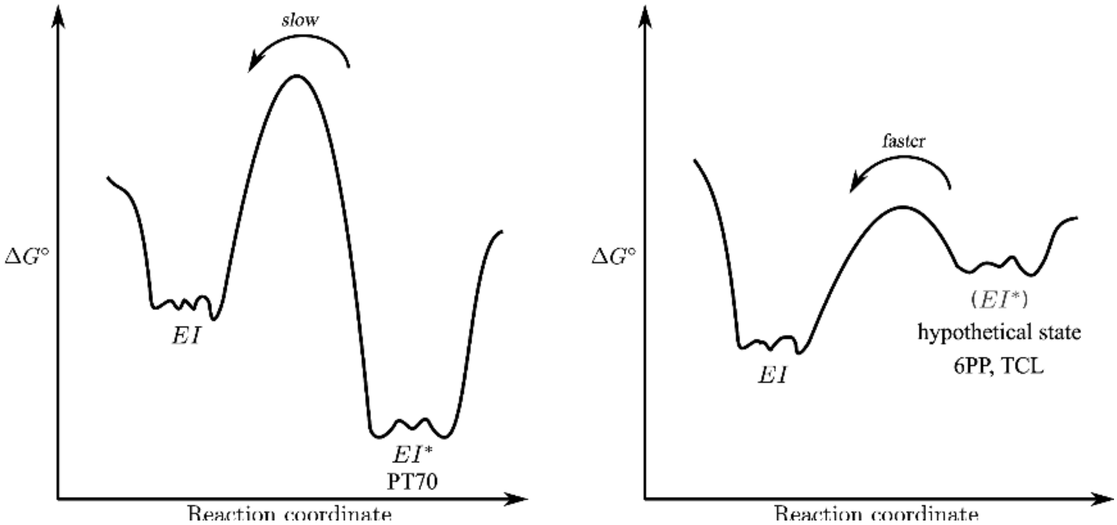
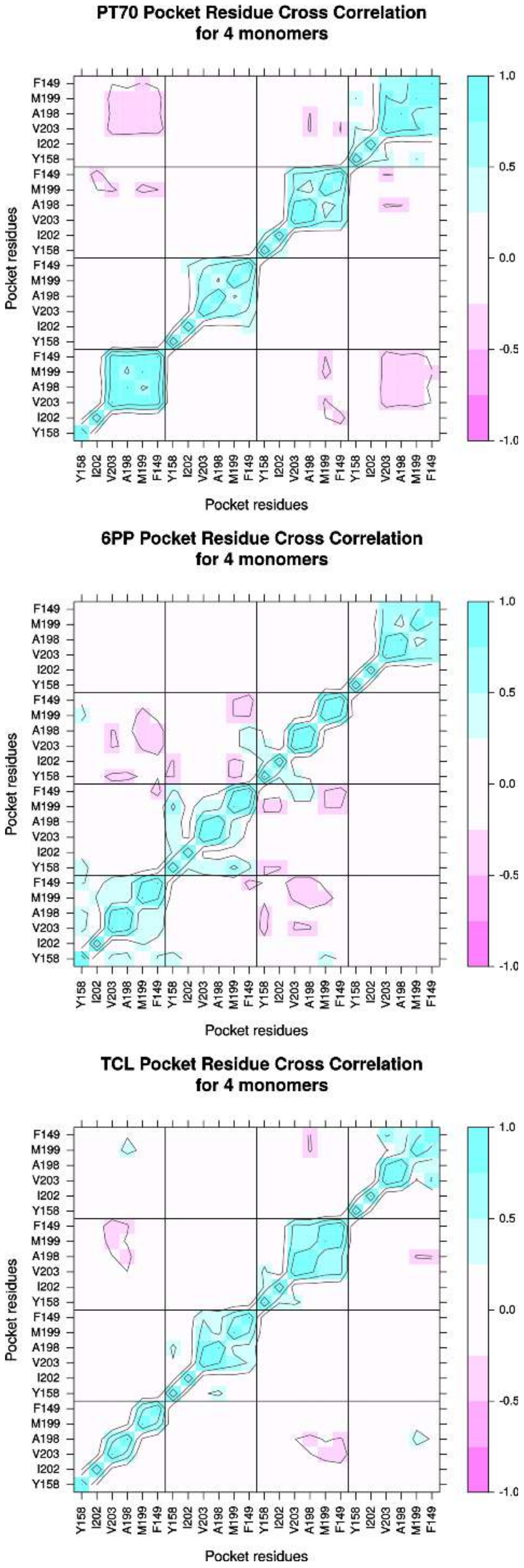


Figure 3.1 Schematic free-energy profiles for a slow-binding inhibitor (left) and a destabilized EI* state as a consequence of the presence of a rapid-reversible inhibitor or ligand removal (right). Each macrostate (EI, EI*) is obviously associated with many microstates.

Because InhA crystallizes as a homotetramer and is known to be active as a homotetramer in solution [53], all simulations were run for the tetramer to best represent the bioactive form of InhA. This has the additional advantage of simultaneously sampling four analogous subunits at the same time. As the active sites of the four monomers are about 40 Å apart from each other, facing opposite sides in the quaternary structure and working independently [143], the 150 ns trajectories of the four binding pockets may be seen as a combined 600 ns sampling for the monomer. In a dynamic cross-correlation analysis of the four binding pockets over the entire trajectory we could not observe any correlated motions among the four pockets, supporting the assumption that their motions can be treated as independent (Figure 3.2) [138, 139]. Therefore, in some of the analyses presented below the combined ensembles of the four monomers were used. In other cases, however, it was more appropriate to follow the monomers individually along their 150 ns trajectory.

3.2 Binding pocket dynamics and conformational families

We first focus on the binding pocket and compare the conformations observed in the different simulations to identify distinct conformational states, viz. recurring conformational families. The InhA binding pocket as defined by Luckner *et al.* (2010) [45] comprises the amino acids Phe149, Ala198, Met199, Ile202, and Val203 of the hydrophobic pocket, as well as the more hydrophilic residue Tyr158, which is an important hydrogen-bonding interaction partner for inhibitors. To detect conformational families



PT70 Pocket Residue Cross Correlation for 4 monomers

6PP Pocket Residue Cross Correlation for 4 monomers

TCL Pocket Residue Cross Correlation for 4 monomers

Figure 3.2 Dynamic cross-correlation analysis of InhA binding pocket in separate homotetrameric systems. Each small box represents a monomer. The color-scale indicates the correlation coefficient. No correlated motion between the four binding pockets of each system can observed.

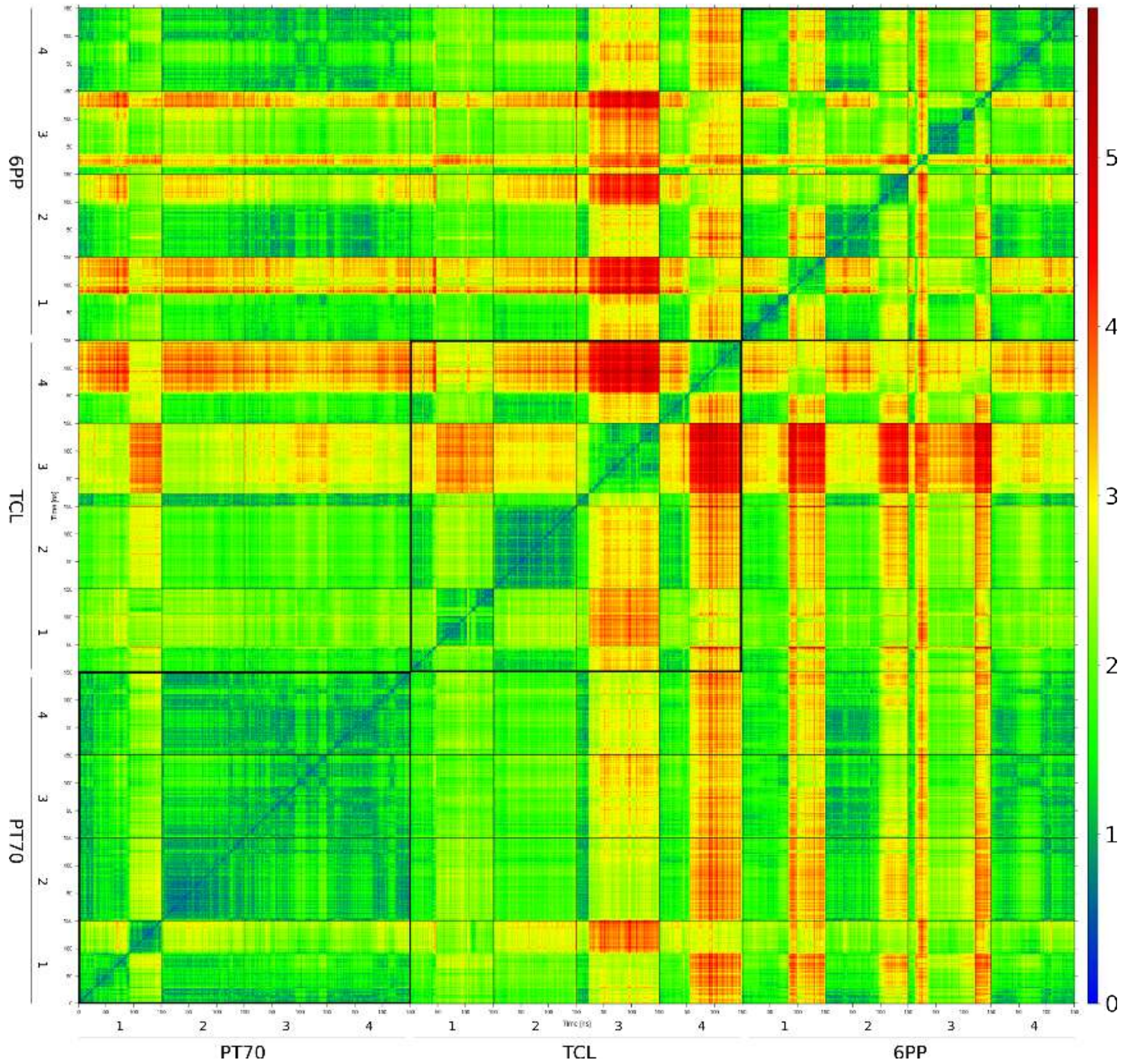


Figure 3.3 12x12 2D RMSD plot of the binding site (defined by the heavy atoms of Phe149, Tyr158, Ala198, Met199, Ile202, and Val203) of all PT70, TCL and 6PP monomers. RMSD values between two frames are illustrated according to the color scale on the right. The axes correspond to the simulation time (0 to 150 ns for each monomer). A single small box (square delimited by thin black lines) represents the comparison of the trajectory snapshots either within a given monomer (boxes along the diagonal) or between two different monomers (off-diagonal boxes). The bold black lines enclose the monomers of a particular homotetramer (i.e., **PT70**, **TCL** or **6PP**).

of the ligand-bound state of the binding pocket, a 12x12 2D-RMSD plot of all against all monomers of the **PT70**-, **TCL**-, and **6PP**-complexes was calculated (Figure 3.3). This allows to compare all conformations occurring in the different simulations and to identify similarities or differences across the systems, which is done most straightforwardly by a hierarchical cluster analysis on the basis of this 2D-RMSD matrix to group the recurring conformations to conformational families.

The hierarchical cluster analysis was carried out with R [130] using the complete linkage method. This method was preferred over others not only because it tends to produce

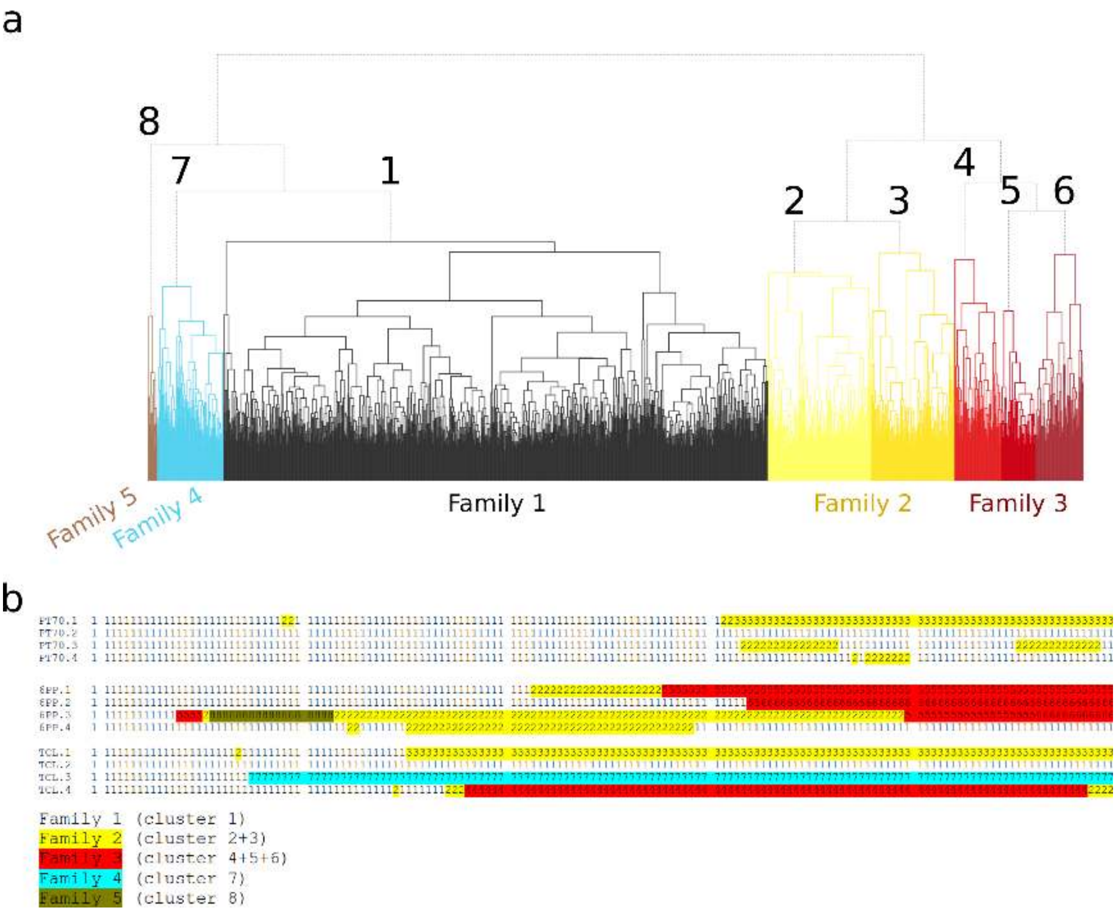


Figure 3.4 Hierarchical clustering analysis of binding-pocket conformers of the PT70, TCL and 6PP simulations based on the mutual RMSD comparison of the individual snapshots as shown in the 2D RMSD plot (Figure 3.3). The calculated RMSD is used as distance measure with complete linkage. The clusters detected at an RMSD cutoff of 3.5 Å are shown in different colors and are numbered as explained in the text. **(a)** Cluster dendrogram. **(b)** Time line of cluster membership. For each monomer of the simulated systems all snapshots included in the analysis from 0 to 150 ns (at intervals of 1 ns) are consecutively written in a line as blocks of 30 ns. The numbers represent the cluster to which a particular snapshot belongs to. Family membership is highlighted by colors according to the legend at the bottom.

clusters with similar diameter, but primarily because it provides readily interpretable results in terms of a maximum RMSD value between members of a cluster. Here, eight clusters of recurring conformations of the InhA binding pocket were identified at an RMSD cutoff of 3.5 Å (cf. Figure 3.4 for further details).

On the basis of the cluster dendrogram and the corresponding structural similarities, the clusters were further summarized to five "monophyletic" conformational families. Subsuming the clusters to monophyletic families was achieved by visual inspection instead of raising the RMSD cutoff, since mere RMSD values might overestimate the importance of minor backbone movements while concealing important side chain flips. These families are hereinafter referred to as Families 1 to 5 (cf. Figure 3.5):

(a) Family 1 (based on cluster 1) corresponds to the crystal structure conformation of

Table 3.1 Occurrence frequencies (in %) of the conformational families of the InhA binding pocket in the three analyzed simulations of the PT70-, 6PP- and TCL-complexes, based on the hierarchical clustering analysis.

	Family 1	Family 2	Family 3	Family 4	Family 5
$\mathbf{PT70}_1$	4.97	3.37	0.00	0.00	0.00
$\mathbf{PT70}_2$	8.33	0.00	0.00	0.00	0.00
$\mathbf{PT70}_3$	6.79	1.55	0.00	0.00	0.00
$\mathbf{PT70}_4$	7.89	0.44	0.00	0.00	0.00
$\mathbf{6PP}_1$	3.53	1.10	3.70	0.00	0.00
$\mathbf{6PP}_2$	5.30	0.00	3.04	0.00	0.00
$\mathbf{6PP}_3$	0.66	4.75	1.93	0.00	0.99
$\mathbf{6PP}_4$	5.85	2.48	0.00	0.00	0.00
$\mathbf{TCL}_1$	2.48	5.85	0.00	0.00	0.00
$\mathbf{TCL}_2$	8.33	0.00	0.00	0.00	0.00
$\mathbf{TCL}_3$	1.27	0.00	0.00	7.06	0.00
$\mathbf{TCL}_4$	2.81	0.44	5.08	0.00	0.00
Sum	58.21	19.98	13.75	7.06	0.99

the **PT70**-complex;

(b) Family 2 (based on clusters 2 and 3) shows a conformation with a slight twist of helix $\alpha 6$ (residues 202-209 in the ascending branch of the SBL), resulting in a shift of Ile202 toward the ligand and a minor displacement of Val203 toward the hydrophobic pocket;

(c) Family 3 (based on clusters 4 to 6) is characterized by a more open conformation of helix $\alpha 6$ and new positions of Ile202 and Val203: Ile202 now adopts the position of Val203 in the **PT70**-crystal structure, and Val203 is shifted to the back, farther away from the binding pocket;

(d) Family 4 (based on cluster 7) represents the conformations with a flip of Tyr158 toward the hydrophobic pocket and an associated conformational change of Phe149 toward the former position of Tyr158;

(e) Family 5 (based on cluster 8) is characterized by another open conformation of helix $\alpha 6$ resulting in a shift of Ile202 and Val203 toward the outside.

The quantitative analysis of the conformational families shows that Family 1 is by far the most frequent conformation, accounting for 58.21% of the frames across all monomers of the three simulations (cf. Table 3.1 and Figure 3.6). Family 2 occurs with a frequency of 19.98%, whereas Family 3 accounts for 13.75%. Families 4 and 5 constitute the minority of conformations with 7.06% and 0.99%, respectively.

Breaking this down to the individual simulations shows a clear difference between **PT70** and the other two complexes: Whereas Family 2 and 3 conformations show an occurrence of only 16.1% and 0.0%, respectively, in the simulation of the **PT70**-complex, values of 25.0% (Family 2) and 26.0% (Family 3) are obtained for the **6PP** simulation and

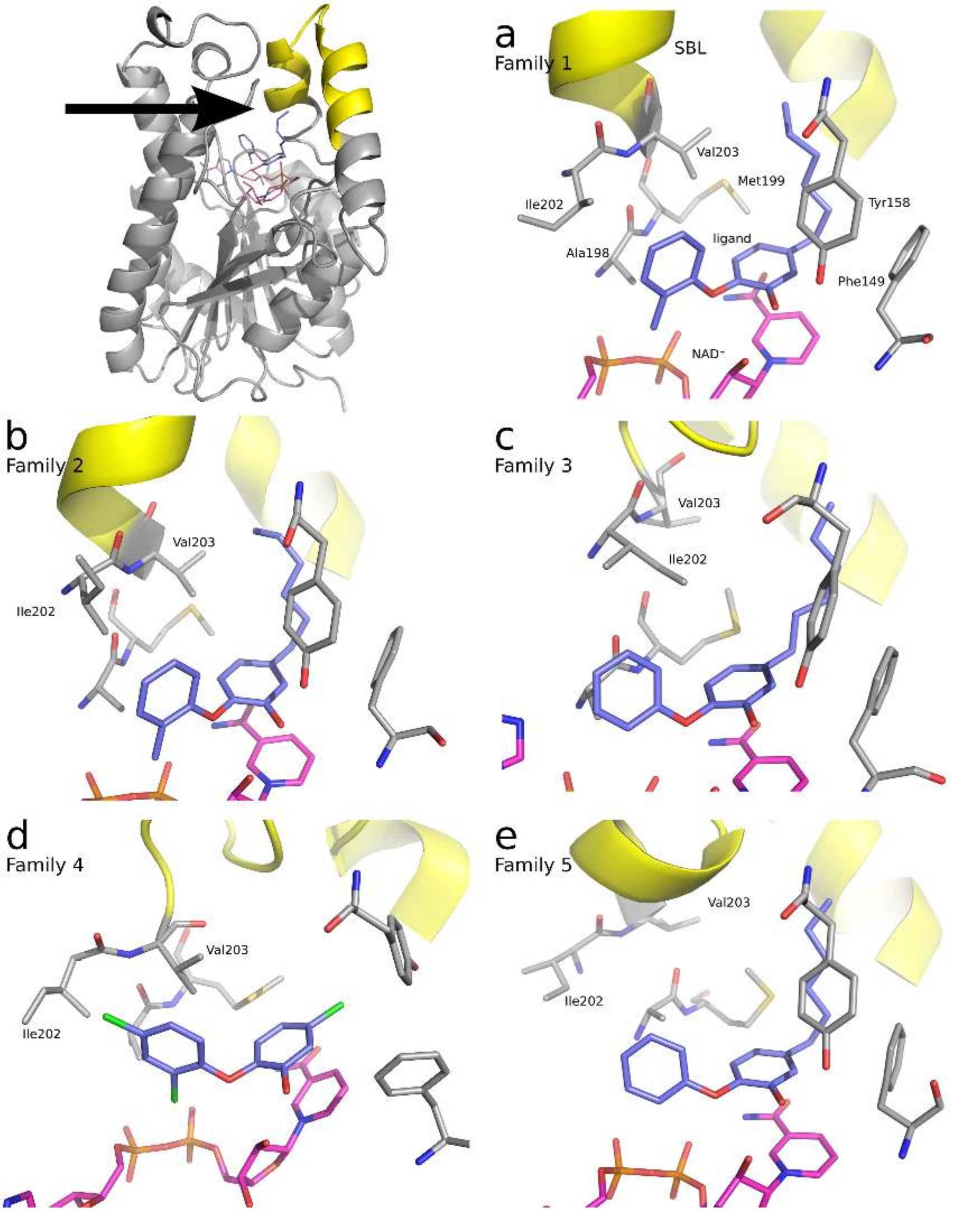


Figure 3.5 Illustration of conformational families of InhA. After summarizing the eight clusters of the hierarchical cluster analysis to five conformational families, a Partitioning Around Medoids (PAM) clustering was performed with R for each conformational family. The resulting medoids are illustrated as cluster representatives. The top left figure shows an entire monomer (A) of InhA from the crystal structure of the complex with **PT70** (PDB 2X23). The substrate binding loop (SBL) is highlighted in yellow. The arrow represents the direction of the view for the subsequent images. **(a)** Family 1: crystal structure conformation; **PT70** monomer 4 after 34 ns of MD simulation. SBL and pocket residues are labeled. The ligand carbon atoms are depicted in slate blue, the cofactor carbon atoms in magenta. **(b)** Family 2: Helical twist of ascending SBL branch with Ile202 shifted toward the ligand; **PT70** monomer 3 after 141 ns of MD simulation. **(c)** Family 3: Enhanced movement of Ile202 far into the hydrophobic cavity; **6PP** monomer 1 after 102 ns of MD simulation. **(d)** Family 4: Flip of Tyr158 toward the hydrophobic pocket; **TCL** monomer 3 after 119 ns of MD simulation. **(e)** Family 5: Ile202 movement toward the outside of the protein into the solvent; **6PP** monomer 3 after 27 ns of MD simulation.

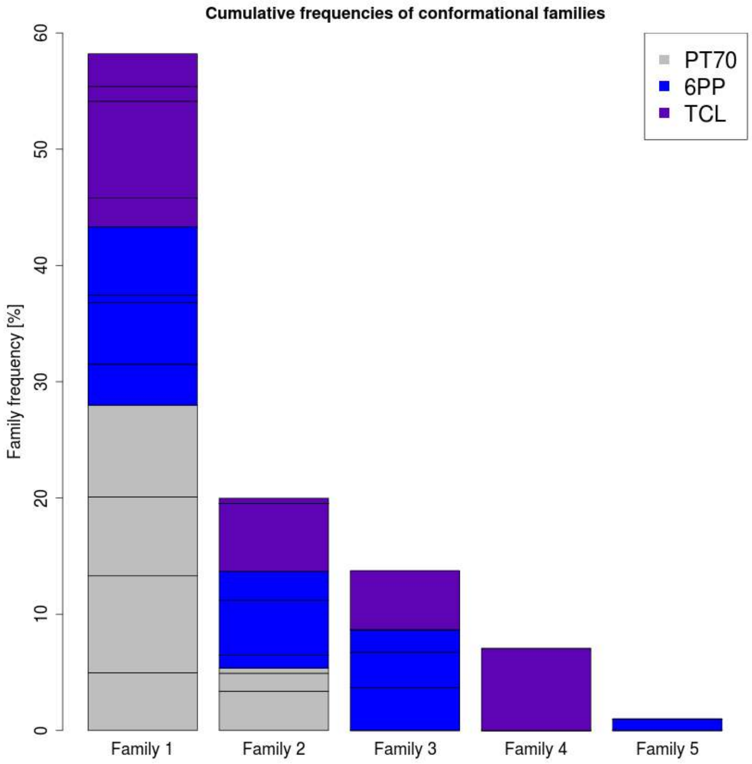


Figure 3.6 Cumulative frequencies of conformational families of the InhA binding pocket in 150 ns of the PT70, 6PP, and TCL MD simulations. Horizontal lines separate the single monomers of each of the three considered homotetrameric complexes.

values of 18.9% (Family 2) and 15.2% (Family 3) for the **TCL** simulation. Besides that, the **TCL** simulation shows 21.2% Family 4 conformations and the **6PP** simulation 3.0% Family 5 conformations. Thus, Family 1 conformations are found to 83.9% in the **PT70** simulation, but only to 46.0% in the **6PP** and to 44.7% in the **TCL** simulation (cf. Figure 3.6).

Apparently, while the state corresponding to conformational Family 1 is stably maintained by the **PT70**-complex, the **6PP**- and **TCL**-complexes have a clear tendency to depart from this state (cf. Figure 3.4b). Interestingly, this is not simply due to a reduced occupation of the hydrophobic pocket, because both **PT70** and **6PP** occupy this site with a hexyl chain, while **TCL** projects only a chlorine substituent into the pocket. However, the space left unoccupied by **TCL** is the reason why the Tyr158 side chain can switch its orientation and lead to a Family 4 conformation, which occurs only in the **TCL**-simulation and is not possible with the other complexes.

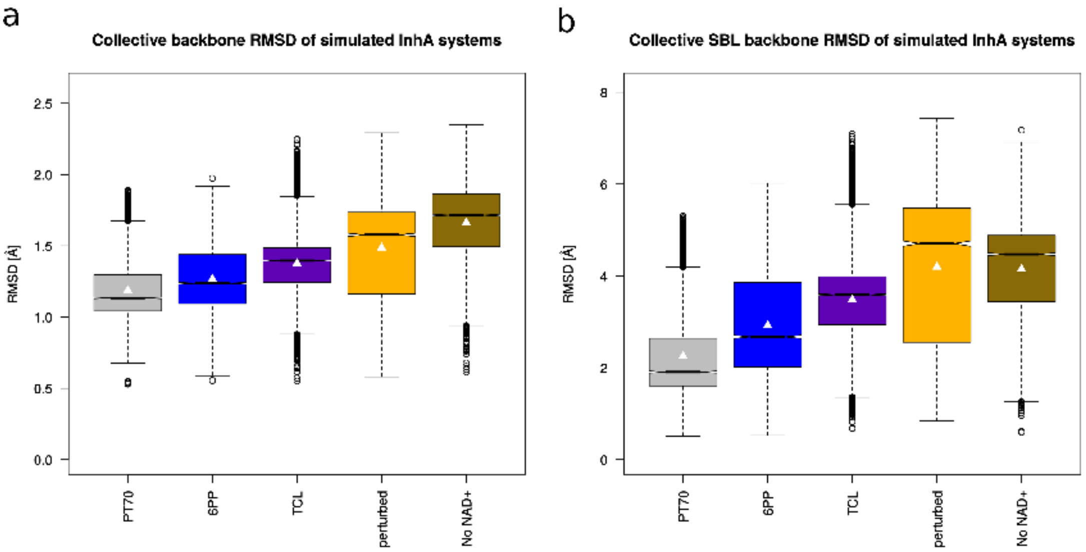


Figure 3.7 Collective RMSD values of (a) backbone (C, N, and C$_\alpha$ atoms) and (b) SBL backbone of InhA monomers. Each monomer of the simulated homotetrameric systems (150 ns) was fitted individually onto chain A of the 2X23 crystal structure as reference for the RMSD measurements and the data of the four monomers were combined to one box plot per system. Boxes indicate the interquartile range (first to third quartile), black lines in the boxes show the median of each distribution. The whiskers extend to values 1.5 times the interquartile range from the box. Significant differences in the medians are indicated by non-overlapping notches. Average values are marked by white triangles.

3.3 SBL dynamics and secondary structure analysis

As the ordering of the SBL is supposed to play an important role in slow-binding inhibition of InhA [24, 45, 64], the dynamic behavior of this structural segment deserves special attention. To look first at the overall backbone dynamics of the entire systems, the RMS deviation of the backbone atoms of each monomer was calculated with respect to chain A of the 2X23 crystal structure (Figure 3.7a). All ligand-bound systems show high stability of the overall structure throughout the entire simulation. With averages of 1.19 Å and 1.27 Å, the **PT70** and **6PP** complexes display slightly lower RMS deviations than the complex with **TCL** (1.38 Å). Not unexpectedly, the perturbed systems without ligand show a clear shift toward higher values and larger fluctuations. Nevertheless, the medians and averages remain well below 2 Å in all cases, indicating reasonable stability of the entire trajectories.

With these values as reference, the large degree of flexibility of the SBL becomes immediately evident. The RMSD of the backbone atoms between residues 202 and 218 (corresponding to the entire SBL) shows similar overall trends as seen in the analysis of the complete backbone, but (much) larger absolute values and fluctuations (Figure 3.7b). In fact, the major mobility of the backbone is observed in the SBL. The highest

RMSD values (as for example in the *perturbed* system) correspond to completely opened loop conformations. Thus, the time scale of the simulation is sufficient to encounter major loop disordering and opening. Furthermore, partial or complete loop closing and rearrangement can be seen after some opening events (e.g., **6PP** monomer 4; cf. Figure 3.8, which shows the RMSD of the SBL as a function of time for each monomer of the simulated systems), emphasizing that the produced trajectories do not simply evolve toward a growing disorder.

Since the ordering of the two helical SBL branches is important for inhibitor binding and happens primarily at the secondary-structure level, a secondary-structure analysis was performed using the VMD plug-in *Timeline* to assign one of six secondary-structure motifs to each atom of the SBL backbone atoms (residues 202 to 218) for each frame of the trajectory: (1) isolated bridge, (2) Coil, (3) 3_{10}-helix, (4) α-helix, (5) π-helix, and (6) turn [141]. The 2X23 crystal structure SBL consists completely (100%) of α-helix and 3_{10}-helix atoms. For the simulations, the average percentage of these two motifs was calculated over the entire sampling time (Figure 3.9). With an average of 69.76% the **PT70**-bound monomers show the highest percentage of α-helix and 3_{10}-helix motifs during the simulation, followed by **6PP** (62.67%). With 49.73% the **TCL** monomers are comparable to the *perturbed* monomers (46.52%). *No NAD*$^+$ shows by far the lowest percentage of these helical motifs (31.55%). This reinforces the notion that the proper occupation of the hydrophobic pocket is an important contributor to the conservation of the helical SBL structure of the final conformational state EI*. The lower helical-motif frequency of **6PP** and **TCL** compared to **PT70** is in line with their differences in binding affinity and residence time, stressing the importance of long-term SBL conservation.

3.4 Hydrogen bond interactions and binding mode analysis

We now focus on the ligand and analyze first the hydrogen bond between Tyr158 and the A-ring phenolic oxygen, which is a highly conserved interaction between diphenyl ethers and InhA. For analysis, the distance between the Tyr158 oxygen (OH) and the phenolic oxygen of the ligands was followed over the entire trajectory (cf. Figure 3.10). **PT70**-bound monomers show by far the lowest distance with medians ranging between 2.82 Å and 2.85 Å, followed by **6PP** (2.84 Å to 3.21 Å) and **TCL** (3.00 Å to 7.34 Å). The bimodal distributions observed for the **TCL** monomers 1 and 4 are caused by the transition to an alternative binding mode of **TCL** further described below. The shorter distances for **6PP** and especially **PT70** evince that the differences in the chemical structures of the ligands directly influence the formation and maintenance of the important

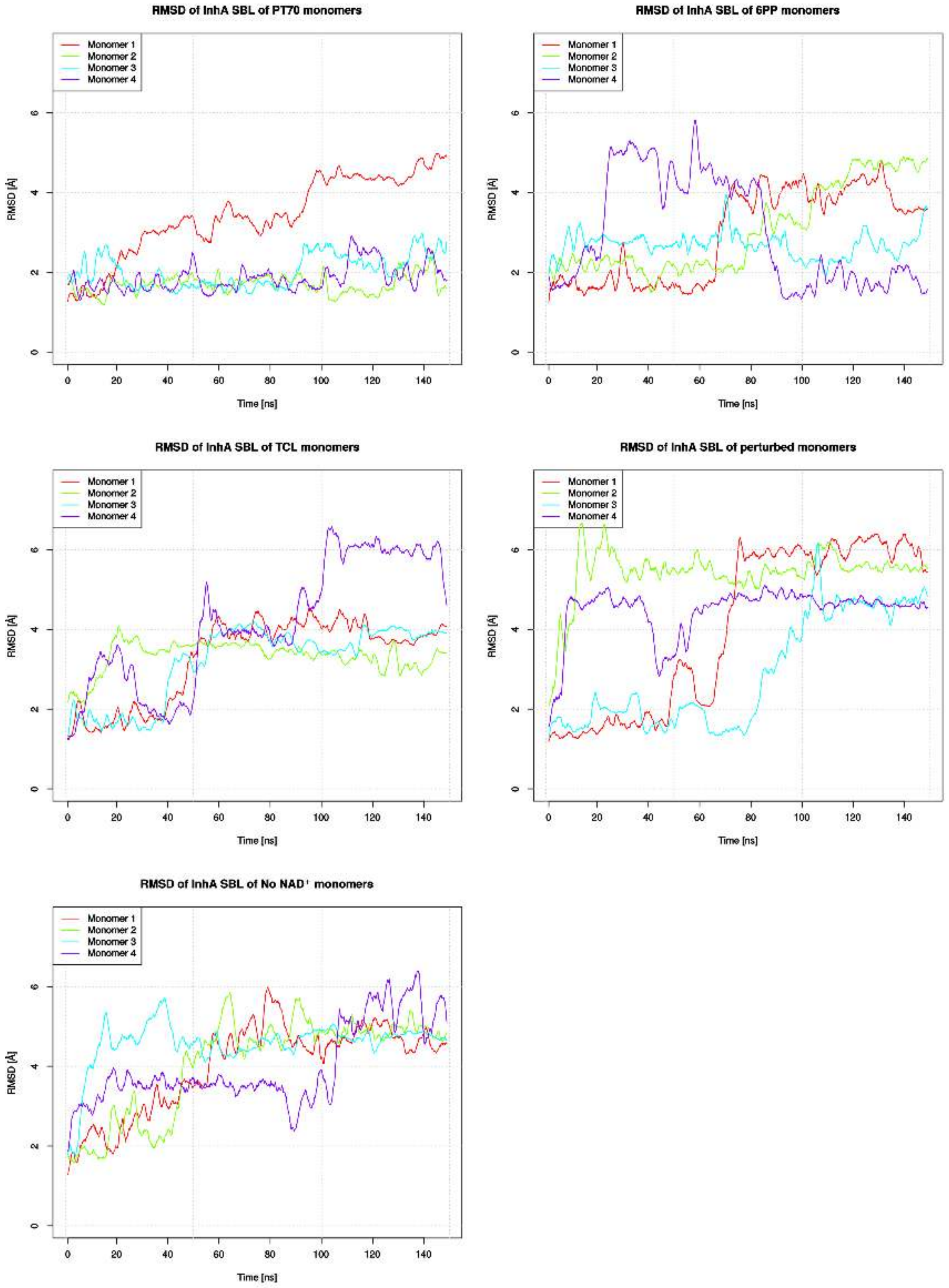


Figure 3.8 Backbone RMSD plots of InhA SBL (residues 202 to 218) of single monomers. A moving average with a window size of 20 frames was used. The RMSD was measured with reference to chain A of the 2X23 crystal structure.

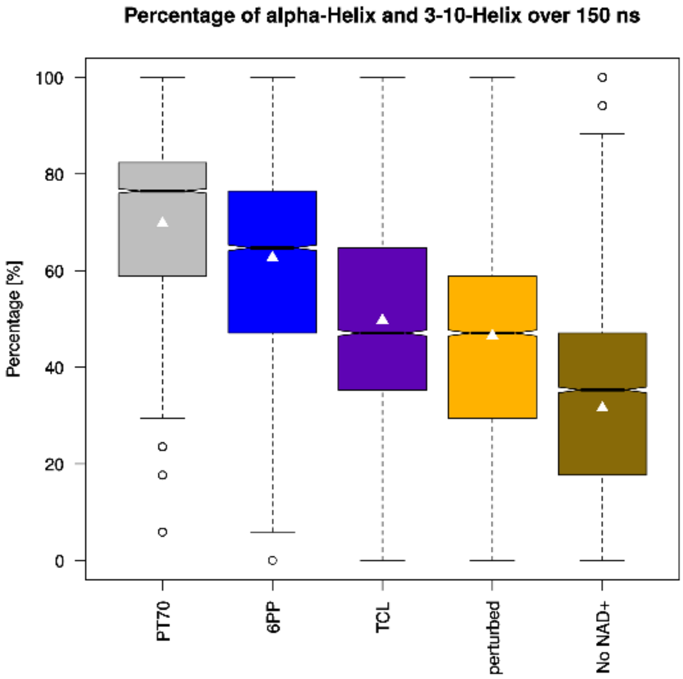


Figure 3.9 Occurrence frequency (in % of the trajectory snapshots) of α-helix and 3_{10}-helix motifs in the substrate binding loop. Each monomer of the simulated homotetrameric systems (150 ns) was analyzed, and data of the four monomers were combined to one box plot per system.

hydrogen bond between Tyr158 and the ligands. The measured distances correlate with the relative affinity of the ligands (Figure 2.4), showing a stably maintained hydrogen bond for **PT70**, a partially maintained hydrogen bond for **6PP**, and a hardly stable interaction for **TCL**.

The second most important aspect of the diphenyl ether binding mode is the occupation of the hydrophobic pocket. While **PT70** and **6PP** both fill the pocket almost completely (a calculation of the free pocket volume with POVME [144] shows virtually no free volume for both complexes, Figure 3.11), the bound **TCL** leaves free space to be occupied. In fact, this space is flooded by water molecules after a few hundred picoseconds (cf. for example **TCL** monomer 2, Figure 3.12). Although this may appear counterintuitive based on the lipophilic character of this area, it is well known that given sufficient space and accessibility, water molecules also occupy lipophilic binding sites [145].

The most drastic effect of the missing hydrophobic moiety of **TCL** can be observed in **TCL** monomers 1 and 4, where the ligand changes its binding mode entirely after around 100 ns and 70 ns, respectively (Figure 3.13). The new binding mode displays a breakage of the hydrogen bond from Tyr158 to the phenolic oxygen of the diphenyl ether with subsequent shift into the hydrophobic pocket. After the scaffold transition, the A-ring, formerly stacked above the nicotinamide ring system, now occupies the hydrophobic

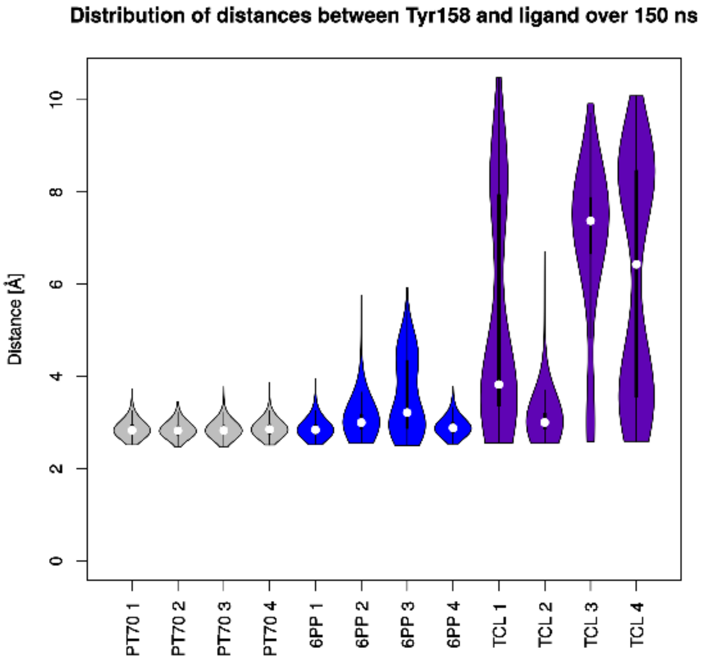


Figure 3.10 Violin plots of distances between the phenolic oxygen of Tyr158 and the respective ligands. White dots depict the medians. Thick vertical lines indicate the interquartile ranges (IQR), thin lines extend to $1.5 \cdot \text{IQR}$ from the third and first quartile, respectively. The shape of the violins illustrates the kernel density estimation of the respective distribution.

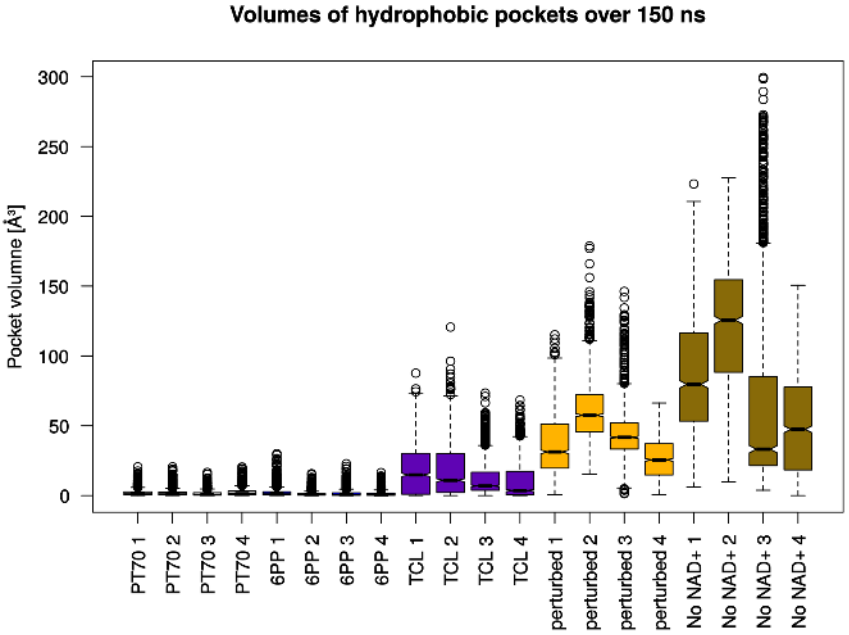


Figure 3.11 Pocket volume analysis of hydrophobic pocket in MD systems. Pocket volume calculations were carried out with the tool POVME [144].

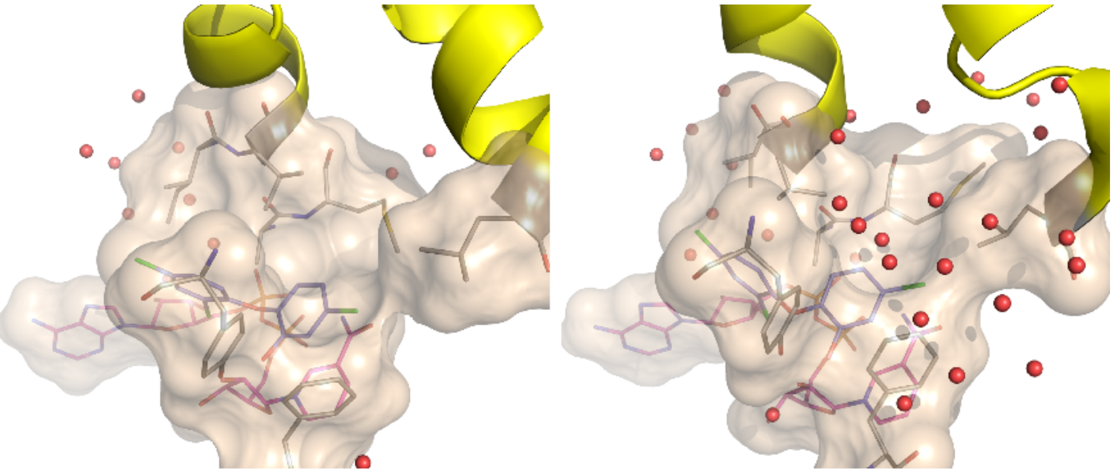


Figure 3.12 Snapshots of TCL monomer 2 after heating (0 ns, left) and after 700 ps of MD simulation (right). The ligand **TCL** is depicted in slate blue, the cofactor in magenta and the pocket residues including Leu218 in gray. The SBL is shown in yellow. Ligand, cofactor, and pocket residues are also shown as surface (wheat), oxygens of water molecules are shown in red. Flooding of the hydrophobic pocket is noticeable after 700 ps (right).

pocket and forms a polar interaction with the nicotinamide oxygen. The B-ring is now placed at the former location of the A-ring. In both cases a stable interaction with the nicotinamide oxygen is observed once the binding mode has changed. This is also represented by heavy-atom distances below 3 Å (Figure 3.14). This new interaction could also be observed in MD simulations of the *Plasmodium falciparum* enoyl-ACP reductase (*Pf*ENR) in complex with NAD$^+$ and the ligands **FT0** and **FT1**, respectively [146]. The novel binding mode of **TCL** co-occurs with the conformational Families 2 and 3, suggesting that a shifted Ile202 is detrimental to ligand stabilization in the pocket. There is indeed a steric hindrance between Ile202 and the B-ring chlorine of the ligand after Ile202 has moved. As a result, the ligand is pushed from "above" and eventually forced to rotate its B-ring, whereupon it yields and moves toward the hydrophobic pocket. Please note that this new binding mode is not postulated as an actual alternative binding mode of **TCL**. Rather, it is a consequence of the instability of the artificial starting structure and only shows that an alternative interaction with the cofactor might be possible in the binding pocket. Because this interaction requires a Family 2 or Family 3 conformation, it does, however, not provide a strategy to increase the residence time of slow-binding inhibitors.

3.5 Influence of *ortho*-substituted B-ring

While **6PP** is a rapid reversible inhibitor, **PT70** binds with a residence time of 24 minutes (Figure 2.4), although the *ortho*-methyl group at the B-ring of **PT70** is the only structural difference [45, 64]. Interestingly, for the **6PP** complex the simulations

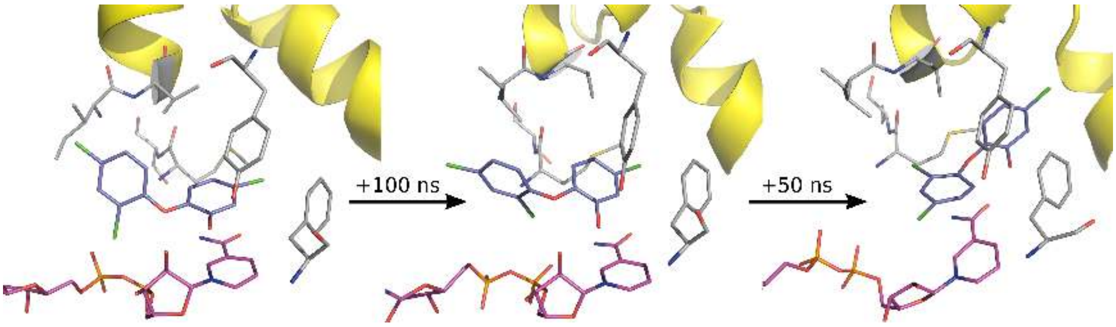


Figure 3.13 TCL of monomer 1 at 0 ns, 100 ns, and 150 ns of MD simulation, respectively. The initial change of binding mode can be observed at 100 ns, resulting in the final binding mode shortly after, which stays stable until the end of the simulation (150 ns). Very similar observations were made for **TCL** monomer 4, but starting already at 70 ns (cf. also Figure 3.14).

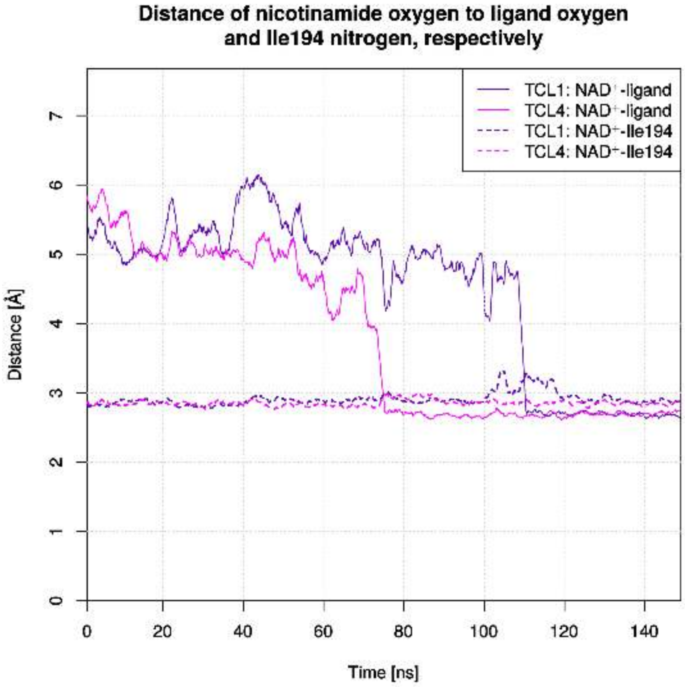


Figure 3.14 Distances between the NAD$^+$ nicotinamide oxygen and the phenolic oxygen of the ligand or the Ile194 backbone nitrogen. Distances are shown as a function of time in a moving-average plot with a window of 20 frames. Monomers 1 and 4 are illustrated for the **TCL** complex. Continuous lines indicate distances to the ligand, whereas dotted lines are used for distances to Ile194. For each illustrated ligand a stable interaction with a distance below 3 Å can be observed after the binding-mode change, while the interaction of NAD$^+$ with Ile194 (present in the starting structure) is only slightly affected.

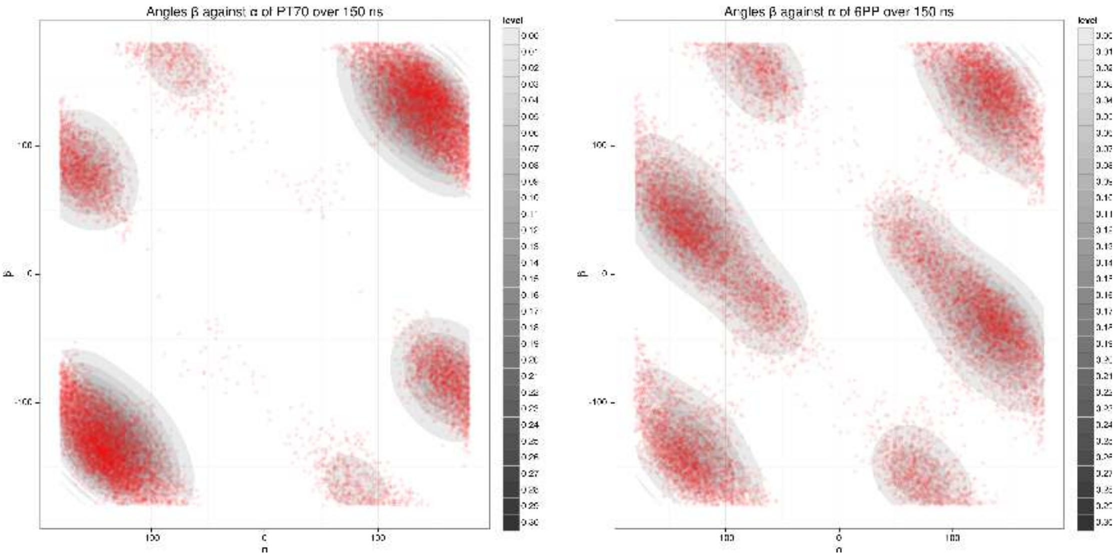


Figure 3.15 2D density plot for the ether dihedral angles α and β of the unbound ligands PT70 (left) and 6PP (right) based on a 150 ns MD simulation in aqueous solution. The dihedral angles α (C_{OH}-C-O-C) and β (C-O-C-$C_{Me/H}$) are illustrated in Figure 2.4.

indicate a reduced stability of the Family 1 state in comparison to the **PT70** complex, and the conformational Families 2 and 3 are significantly more frequent in the **6PP** simulation. With the *ortho*-methyl moiety as the only substitution, this difference appears as the logical origin for these observations. To investigate the effect of *ortho*-methyl substitution on the ligand conformations (which are mainly determined by the torsions around the two ether bonds), two additional 150 ns MD simulations were conducted for each ligand solvated in a water box. By measuring the dihedral angles of the ether moiety along the trajectory, a 2D density map of the (C-O-C-$C_{Me/H}$)-dihedral β *versus* the (C_{OH}-C-O-C)-dihedral α was generated for each ligand (Figure 3.15). The strong peaks in the distribution of the **PT70** angle pairs suggest that fewer conformations are populated compared to **6PP**. Hence, as expected, the *ortho*-substituted **PT70** is more constrained in its intramolecular mobility, hindering the Ile202 movement toward the hydrophobic pocket to a greater extent than the unsubstituted **6PP**. This very likely accounts for the enhanced occurrence of Families 2 and 3 in the case of **6PP** and for the (on average) larger RMS deviations and fluctuations of the ligand in the binding pocket (Table 3.2). Interestingly, also the hexyl chain of **6PP** shows higher mobility than the **PT70** hexyl chain in the binding pocket (Table 3.2, Figure 3.16). In summary, the conformational stabilization of **PT70** by the *ortho*-methyl group appears to translate directly to increased SBL stabilization and retention of a Family 1 conformation.

Table 3.2 Trajectory averages and standard deviations of heavy-atom RMSDs of the PT70 and 6PP ligands and their hexyl chains, respectively. RMSDs were measured individually for each ligand in the four monomers with respect to the corresponding starting structure (after the heating cycles).

	all heavy atoms		**hexyl chain**	
	Avg. RMSD [Å]	**SD [Å]**	**Avg. RMSD [Å]**	**SD [Å]**
PT70$_1$	1.15	0.19	1.50	0.50
PT70$_2$	1.15	0.21	1.15	0.22
PT70$_3$	1.14	0.18	1.20	0.27
PT70$_4$	1.09	0.22	1.17	0.26
6PP$_1$	1.24	0.23	1.77	0.48
6PP$_2$	1.18	0.29	1.31	0.32
6PP$_3$	1.32	0.28	1.91	0.49
6PP$_4$	1.72	0.31	1.41	0.39

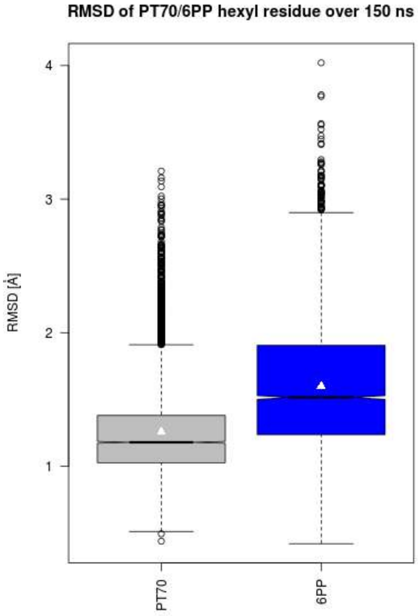


Figure 3.16 Heavy-atom RMSD distributions of hexyl chains of PT70 and 6PP. As references the respective coordinates of the starting structure (after the heating cycles) were used (cf. Figure 3.7 for further explanations).

3.6 Comparison with experimental structures

To further judge the relevance of the simulation results and to discuss the conformational families in the context of the EI and EI* states of the two-step binding process of slow-binding InhA inhibitors (Figure 2.5), a comparison with the experimentally available structural information is important. Most relevant in this context are the very recently released crystal structures of the ternary diarylether complexes with InhA and NAD$^+$

from the studies of Li et al. [40] and Pan et al. [46]. These complexes with the slow-binding inhibitors **PT10** (PDB 4OXY), **PT91** (PDB 4OYR), **PT92** (PDB 4OHU) and **PT119** (PDB 4OIM) and the rapid-reversible inhibitor **PT155** (PDB 4OXN and 4OXK) show differences in the conformations of Ile202/Val203 and the orientation of helix $\alpha6$ of the SBL. The complexes with **PT10**, **PT91** and **PT92** predominantly show the same binding-site conformation and helix orientation as the **PT70**-complex structure, strongly supporting the assumption that this corresponds to the EI* state (the ligands **PT10**, **PT91** and **PT92** differ from **PT70** only by a 2'-nitro, 2'-chloro and 2'-bromo substituent, respectively, instead of the 2'-methyl group) [40]. In contrast, the complex with **PT119** (carrying a 2'-cyano group) displays an alternative arrangement of Ile202 (which adopts the typical position of Val203) and Val203 (which is displaced to the back), but a relatively closed orientation of the helix [46]. Finally, the structures with the rapid-reversible 4-pyridone inhibitor **PT155** (carrying a 4-pyridone as A-ring and an additional 4'-amino substituent on the B-ring in comparison to **PT70**) not only show an unresolved SBL in the monomers of the asymmetric unit, but–for the first time–for one of the monomers also a fully resolved SBL with a widely open orientation of helix $\alpha6$, which has been interpreted as a representation of the EI state by Li et al. [40].

A comparison of this **PT155**-structure with the conformational families suggests that Family 3 indeed captures the characteristics of the EI state: Ile202 is positioned above the ligand, Val203 is moved to the back, and helix $\alpha6$ adopts a very open conformation. Figure 3.17b highlights this open state for a Family 3 representative: it shows a distance between helix $\alpha6$ and strand-4 (used by Li et al. [40] to measure the degree of opening) of 11 Å, whereas only 5 Å are measured for Family 1 (Figure 3.17).

The complex with the slow-binding inhibitor **PT119** constitutes a special case, as it does not show the typical EI* conformation. Whereas in the EI* state Val203 in helix $\alpha6$ is positioned closer to the ligand than Ile215 (located in helix $\alpha7$), in the **PT119** structure Val203 is located far behind and Ile215 is close to the ligand [46]. Together with the altered position of Ile202, this conformation rather reminds of an EI-like state, albeit with a not fully open helix $\alpha6$. Although the authors speculate about the relevance of this structure, they also note that "owing to the crystallization conditions and the potential impact from crystal packing, the observed structure for the **PT119** complex could represent a snapshot along the binding coordinate from EI to EI*" [46]. Indeed, very different crystallization conditions were used for this structure in comparison to the others, limiting the comparability. In particular, the very high acetate concentration leads to the observation of two acetate ions at potentially critical positions of the structure, namely between helix $\alpha6$ and strand-4, as well as between helix $\alpha6$ and helix $\alpha7$ (cf. Chapter 6). Accordingly, we assume that the EI*-state could not be reached

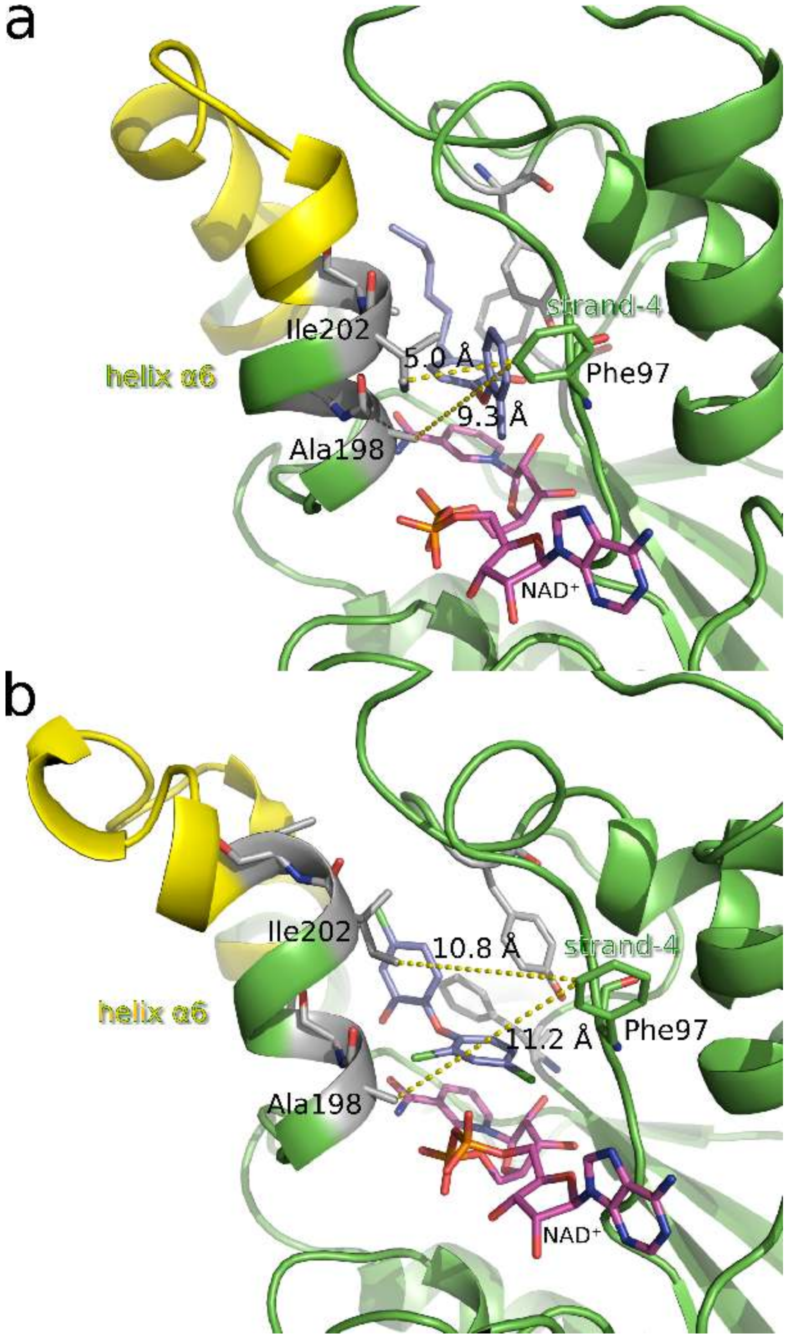


Figure 3.17 Open and closed conformations of InhA observed in the MD simulations. Figure (a) shows the closed state represented by the medoid of conformational Family 1, figure (b) illustrates the open state represented by the medoid of cluster 4 (belonging to conformational Family 3). The same view of the binding pocket as in Figure 3 of Li et al. [40] is used for better comparison. In this view, the portal-forming elements are located left (helix $\alpha6$) and right (strand-4) of the binding site. The distances highlighted as yellow dashed lines were measured between Ala198/Ile202 on helix $\alpha6$ and Phe97 on strand-4. For comparison, in the crystal structure of the **PT70** complex (PDB 2X23) representing the closed state, a distance of 4 Å is found between Ile202 and Phe97, whereas the open state is characterized by a distance of about 10 Å between Ala198 and Phe97 in chain B of the **PT155**-complex crystal structure (PDB 4OXN) [40].

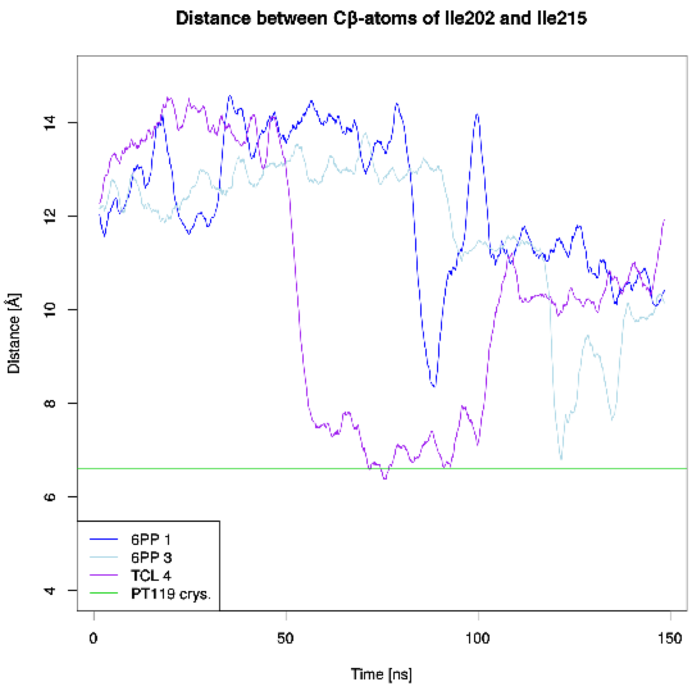


Figure 3.18 Distance between the C_β-atoms of Ile202 and Ile215 over time. **6PP** monomers 1 and 3 are shown in shades of blue, **TCL** monomer 4 is depicted in purple. The green baseline illustrates the **PT119** crystal structure (PDB 4OIM).

under these conditions and that the conformation was frozen in an intermediate, but rather EI-like state. In our simulations, the particular conformational feature of **PT119** occurs only occasionally and only in the context of Family 3 conformations, supporting the EI-likeness (cf. Figure 3.18, which illustrates the distance between Ile202 and Ile215 as a measure for the adoption of a **PT119**-like conformation).

In summary, the comparison with these newly released structures supports the notion that Family 1 corresponds to the EI*-state, whereas Family 3 may be considered as EI state. This has important implications for the interpretation of the simulations and the effects exerted by the different ligands.

3.7 Determinants of residence time and implications for drug design

To optimize potential inhibitors regarding their residence time, it is desirable to understand the reasons which drive the conservation of an EI* state over time [24, 25]. Associating Family 1 conformations with the EI* macrostate and Family 3 conformations with the EI macrostate provides the possibility to interpret the simulation results in this context. Li et al. have carried out partial nudged elastic band MD simulations to investigate the free energy profile for the transition between EI and EI*, illustrating that the energy required for the arrangement of Ile202 and Val203 around the B-ring contributes directly to the height of the energy barrier for the transition from the EI to the EI* state [40]. In contrast, our classical MD simulations were all setup from the EI* state without a biasing potential, but introducing rapid-reversible inhibitors as perturbation to investigate their effects on the stability of the EI* state. As analyzed above, the simulations indeed show a clear tendency for major conformational changes (involving in particular Ile202 and Val203) toward an EI state in the case of the rapid-reversible inhibitors **6PP** and **TCL** and a much higher stability in the case of the slow-binding inhibitor **PT70**. Thus, the dynamic features revealed from the trajectories for the different systems may be linked to the substitution patterns of the examined diphenyl ethers to provide insights for rational ligand optimization toward longer residence times for InhA.

First of all, the *ortho*-methyl group of the B-ring has shown itself to be advantageous as an anchor. A substituent in this position occupies further space between helix $\alpha6$ and cofactor. Moreover, it restrains the phenyl-oxygen-phenyl torsions (as shown in the simulations of the solvated ligands, cf. Figure 3.15), which stabilizes the ligand scaffold and thereby also the *para*-hexyl chain of the A-ring. This appears to improve the stable occupation of the hydrophobic pocket. Proper filling of the hydrophobic pocket is, in fact, a second major determinant, as evidenced by the **TCL**-simulation. In order to lock the binding pocket and the SBL in the EI* state (Family 1), it is desirable to prevent Ile202 and Val203 from moving toward helix $\alpha7$ (residues 210 to 218). Thus, as a third factor and as a suggestion for ligand design, it could be beneficial to introduce a barricade group in 5'-position of the B-ring which might embed itself between Ile202 and Val203 and, thus, further stabilize them, possibly blocking Ile202 from traveling toward the hydrophobic pocket (cf. Figure 2.4). Thereby, the energy barrier between the EI* and the EI state might be significantly increased and the EI* complex could be maintained for a longer time. Notably, a substituent in this particular position is confined in its size by the adjacent Met103. Four *meta*-substituted ligands (fluoro, chloro, methyl and methoxy) were docked exemplarily into the InhA chain A binding pocket using Glide (version 5.8, Schrödinger, LLC, New York, NY, 2012) [147, 148] in extra precision mode (Figure 3.19). All ligands show essentially the same binding mode as **PT70** in the crystal structure, underlining the availability of sufficient space for a small 5'-substituent of diphenyl ethers. Thus, new 2'-substituted diphenyl ethers with an additional small substituent in 5'-position are suggested as inhibitors with potentially further increased residence times.

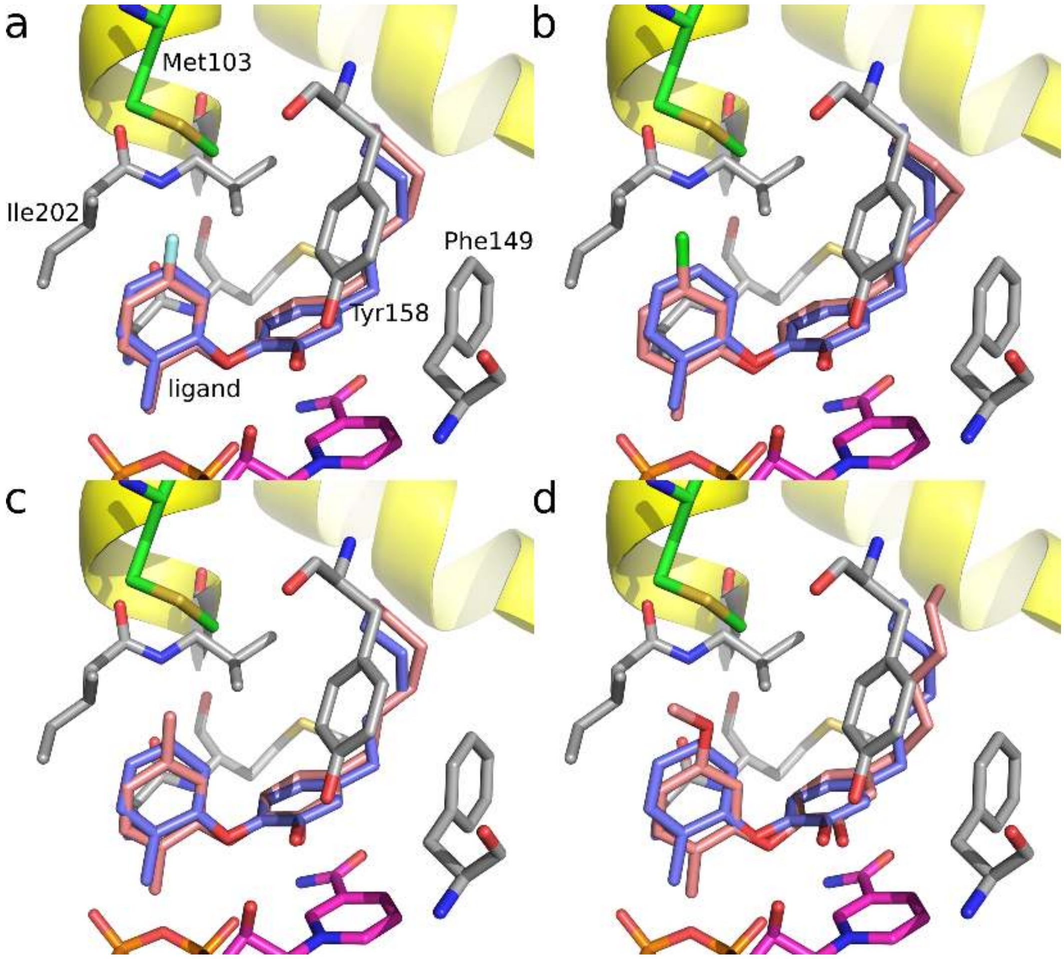


Figure 3.19 Top-ranked docking poses of four 5'-substituted PT70-like diphenyl ethers. Met103 is illustrated in green, the docked ligands are shown in salmon. **PT70** (shown in slate blue as reference) was substituted in 5'-position with a **(a)** fluoro-, **(b)** chloro-, **(c)** methyl-, and **(d)** methoxy-substituent. Docking was carried out with Glide in XP-mode using default settings and a maximum output of 10 poses per ligand.

3.8 Conclusion

By using molecular dynamics simulations with accumulated sampling in the low microsecond time scale, it was possible to unveil previously undetected conformational features of the *Mycobacterium tuberculosis* enoyl-ACP reductase InhA. Starting from an EI* state, the presence of rapid-reversible inhibitors caused an increased tendency for transitions to an EI-like state. The associated conformational changes and dynamic fluctuations of the protein binding pocket and the SBL were illustrated by the MD simulations. Analyses of conformations, pocket volume and secondary structure show different strategies for achieving structural conservation of the EI*-state over time and, thus, increased residence times of inhibitors: firstly, the occupation of the hydrophobic pocket and stabilization of Ile202 and Val203 to prevent these pocket residues from turning over the hydrophobic pocket; secondly, the introduction of a barricade substituent in

5'-position of the B-ring to increase the energy required to arrange helix $\alpha6$ around the B-ring, thus fashioning the final EI* state of **PT70**-like binding modes; and thirdly, the introduction of an anchor in *ortho*-position of the B-ring (methyl in **PT70**) to reduce the degrees of freedom with respect to the central diphenyl ether torsions. This limits the mobility of the bound ligand and, concomitantly, of the hydrophobic pocket, leading to lower fluctuations and an increased stability. These structural features not only keep the InhA binding pocket in the EI* state, but also directly influence the quality of the important hydrogen bond between Tyr158 and the ligand. Taken together, these findings provide valuable insights for future studies of inhibitor design directed against InhA.

3.9 Methods

3.9.1 Protein and ligand preparation

The highly ordered tetrameric InhA crystal structure with bound **PT70** and NAD$^+$ (PDB code 2X23) [45] was used as starting point for the setup of all five simulation systems. Due to the high flexibility of the substrate binding loop the **TCL**-complexed crystal structure of InhA (PDB code 2B35) is incomplete in this crucial area. Therefore, a structural alignment of 2X23 and 2B35 was performed in PyMOL [142]. **TCL** was extracted from the 2B35 structure and placed into the ligand-free 2X23 protein, generating an uninterrupted InhA-NAD$^+$-**TCL** complex. The ligand **6PP** was sketched and docked with Glide (version 5.8, Schrödinger, LLC, New York, NY, 2012) [147, 148] into the 2X23 crystal structure using standard precision. For each monomer the pose with the least RMSD from the crystallized **PT70** was chosen (0.58 Å, 0.47 Å, 1.03 Å, and 0.53 Å, respectively; calculated with fconv [149]). No crystal structure is available for the **6PP** complex, but comparison with the complex structures of the closely related ligands **5PP** and **8PP** [64] show low RMS deviations (of 0.6 Å to 1.0 Å) between the **6PP** binding modes generated by docking and these ligands. Hydrogen atoms were added to **PT70**, **TCL**, and **NAD$^+$** with SYBYL-X. The Amber10 [150] module tleap was used for assigning the parameters of the ff99SB force field [99]. RESP charges [110] were calculated for all three ligands and the cofactor based on HF/6-31G* electrostatic potentials obtained with Gaussian 03 [151]. With the Amber10 module parmchk [152] unavailable force field parameters were calculated according to the General Amber Force Field (GAFF) [114]. Atom and bond types of the ligand were assigned by antechamber [152].

3.9.2 Molecular Dynamics simulations

A short energy minimization of 200 cycles was performed using a generalized Born implicit solvent model [153, 154] as implemented in the Amber11 module sander [155]. Subsequently, the molecules were solvated with tleap using a TIP3P water box [112], retaining all crystallographic water molecules and adding sodium ions to ensure electroneutrality. The resulting systems had dimensions of approximately 110 Å · 112 Å · 89 Å and contained about 101,000 atoms each. For heating-up, water molecules were allowed to move freely in the constant-volume box, while the proteins and ligands were kept rigid for 25 ps. During this step the systems were heated from 100 to 300 K for 20 ps and then cooled to 100 K over 5 ps by means of the Berendsen weak coupling algorithm [103] with a time constant of 0.5 ps. Then the complete systems were treated without constraints and gradually heated to 300 K over a time period of 25 ps. For each system a simulation of 150 ns at 300 K was then carried out, whereby covalent bonds to hydrogen atoms were constrained by the SHAKE algorithm and a time step of 2 fs was used. These simulations were run with NAMD 2.9 [101, 156] using the assigned force field parameters. Energetical equilibration of the simulation box was observed within 1.5 to 3 ns in all cases. The systems were treated with periodic boundary conditions. A van-der-Waals interaction cutoff of 12 Å was used, as well as the particle mesh Ewald methodology (PME) for electrostatic interactions [108]. Constant pressure was assured by the Nosé-Hoover Langevin piston pressure control [106, 107], while constant temperature was achieved by the use of Langevin dynamics. Additionally, two simulations of the uncomplexed ligands **PT70** and **6PP**, respectively, were conducted for 150 ns. Trajectory snapshots were saved every picosecond. For visual and statistical analyses, trajectory snapshots at intervals of 100 ps were considered, resulting in 1500 frames per system. The diphenyl torsion analyses (ligand-only simulations) were carried out with snapshots at 10 ps steps (i.e., 15000 data points). All analyses were performed with VMD and associated plug-ins [140]. All trajectories of the individual monomers were fitted to the chain A backbone atoms (C, N, and C_α) of the 2X23 crystal structure with the *RMSD Trajectory Tool* of VMD for visual inspection and all quantitative analyses. Statistical analysis and plotting was done with the statistical framework R [130, 131, 157]. The pocket volume analysis was performed with POVME [144]. Structural visualizations were created with PyMOL [142].

Chapter 4

MD simulations of 2',5'-disubstituted diphenylethers in InhA

The following chapter is a follow-up study building on a major result of Chapter 3. Extensive structural analyses of the MD simulations carried out for InhA have led to the suggestion of a small 5'-substitution on **PT70** to optimize Family 1 stability of the binding pocket and, thus, the drug-target residence time (cf. Chapter 3.7). In the following chapter the stability of new systems with hypothetical **PT70**-modifications with 5'-substitution is assessed by means of MD simulations.

4.1 System preparation

Two InhA-complexes with different 5'-substituted **PT70**-derivatives were set up: (1) 5'-methyl-**PT70** (2-(2',5'-dimethylphenoxy)-5-hexylphenol) and (2) 5'-chloro-**PT70** (2-(2'-methyl-5'-chlorophenoxy)-5-hexylphenol). The selection of these substituents enables a comparison of the effects of different, but similarly sized moieties. Both systems were prepared in the tetrameric form of InhA. First, the ligands were sketched with Schrödinger Maestro (version 9.7, Schrödinger, LLC, New York, NY, 2014) and docked using Glide in extra precision (XP) mode with default settings (Glide, version 6.2, Schrödinger, LLC, New York, NY, 2014). In the tetrameric systems, a separate docking was carried out for each chain. A maximum number of ten docking poses was generated per docking run. The docking poses with the lowest RMSD with respect to **PT70** of the respective chain were chosen for MD simulation (Table 4.1, calculated with fconv [149]).

The docking poses were energetically minimized for 100 steps in the ff99SB Amber force field and the GAFF [99, 114] using a General Born implicit solvent (GBIS) model [153,

Table 4.1 Minimum RMSD of docking poses of 5'-substituted PT70 ligands to PT70 in Å.

	5'-methyl	**5'-chloro**
Chain A	0.54	0.67
Chain B	0.75	0.67
Chain E	0.82	0.72
Chain G	0.62	0.58

154]. Parameterizations, equilibrations and 50 ns MD simulations were carried out as follows (cf. Chapter 5 for details): both ligands were parameterized in GAFF with RESP atom charges obtained from HF/6-31G* potentials and the protein was parameterized in the ff99SB force field. After 200 steps of energy minimization in the GBIS model, the complexes were solvated with TIP3P water molecules and neutralized with sodium ions, followed by 10,000 steps of energy minimization and a 1 ns equilibration. First, the system was gradually heated from 100 K to 300 K over 500 ps in the NVT ensemble with harmonic constraints on the protein and ligand atoms, which were gradually released. Afterwards, the system was allowed to evolve freely for another 500 ps. MD production runs were performed in the NPT ensemble using NAMD 2.9 [101]. The only difference to the simulation setup followed in Chapter 3 is, thus, the longer and gentler equilibration phase to comply with more recent protocol standards and decrease the likelihood for artifacts in the initial phase of the simulation. Trajectory snapshots were saved every picosecond. For structural and statistical analyses, snapshots at intervals of 100 ps were considered, resulting in 500 frames per system. For the 2D-RMSD analysis of Met103 (Figure 4.4) trajectory snapshots were extracted every 200 ps. Chains A, B, E and G of the simulated systems of InhA bound to 5'-methyl-**PT70** and 5'-chloro-**PT70** are hereinafter referred to as monomers $M1$ to $M4$ and $C1$ to $C4$, respectively.

4.2 Results

4.2.1 The Ile202-Val203-Met103 subpocket

As illustrated in Figures 4.1a and b, the EI* conformation of slow-onset inhibition of InhA exhibits a small subpocket between Ile202 and Val203, limited in size by Met103. The docking of two **PT70**-derivatives shows that small substituents in 5'-position of the B-ring can occupy this space (Figure 4.1c and d). The 5'-substituent, thus, fills the subpocket between Ile202, Val203 and Met103. As a result, the energy barrier from the EI* state towards dissociation of the ligand, which includes a twist of helix $\alpha 6$ with Ile202 and Val203 shifting over the B-ring towards the hydrophobic pocket (cf. Chapters 3 and 6), is assumed to be elevated.

To quantify the occupation of this subpocket, distances were measured between the 5'-substituents of the ligands and the C_β atoms of Ile202 and Val203, respectively (Table 4.2, Figures 4.2 and 4.3). The M-monomers mostly exhibit one distinct peak with medoid distances of 5.53 Å and 5.41 Å, respectively. Compared to the initial distances of the docking poses (Table 4.3), the distance to Ile202 generally increases during the MD simulation, whereas the distance to Val203 decreases or stays similar (monomer

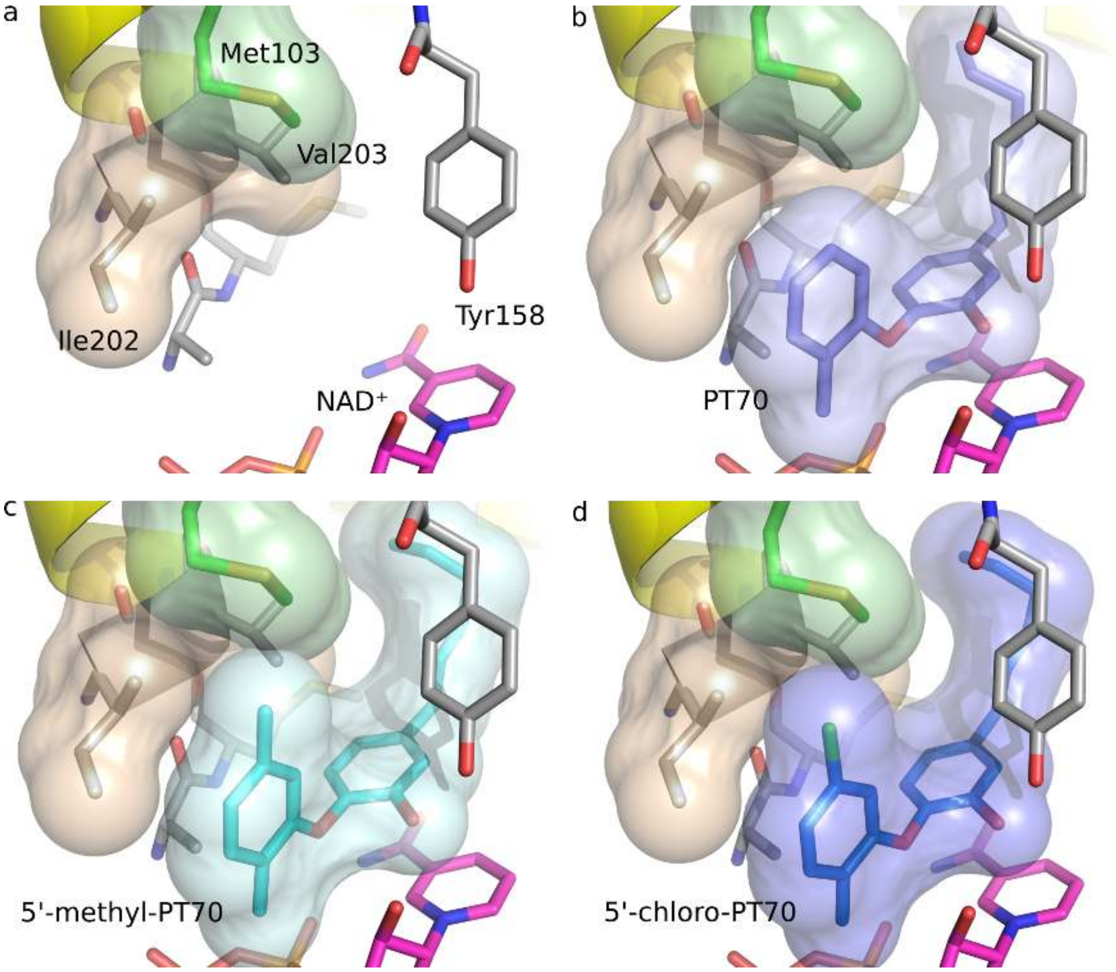


Figure 4.1 Binding pocket of InhA with PT70 and 5'-substituted derivatives. **(a)** Chain A of crystal structure 2X23 displayed without ligand. **(b)** Chain A of crystal structure 2X23 with bound ligand. **(c)** Top docking pose of 5'-methyl-**PT70** in chain A of 2X23. **(d)** Top docking pose of 5'-chloro-**PT70** in chain A of 2X23. Pocket residues are depicted in gray. Met103, limiting the size of potential 5'-substituents, is illustrated in green. Ligands are shown in shades of blue and the cofactor is shown in magenta. Transparent atom surfaces highlight the available space between Ile202, Val203 and Met103 in **(a)** and **(b)**, which is occupied by the 5'-methyl and 5'-chloro-substituent in **(c)** and **(d)**, respectively.

$M3$). Thus, the distances are equalizing over 50 ns sampling time, suggesting that the substituent is embedding evenly in between Ile202 and Val203. The medoid distances measured in the C-monomers behave similarly, although the fluctuations of the distance to Ile202 are higher (cf. Figure 4.3). Monomer 2 of the 5'-chloro-**PT70**-bound system shows two distinct peaks, which stem from a conformational change of the binding pocket (discussed below).

Whereas Ile202 and Val203 exhibit stable and equal distances to the 5'-substituent (except in monomer $C2$), Met103 shows more conformational flexibility. Thus, a 2D-RMSD analysis for the heavy atoms of Met103 was performed (Figure 4.4). After different simulation time, the residue adopts a new conformation in each monomer (indicated by

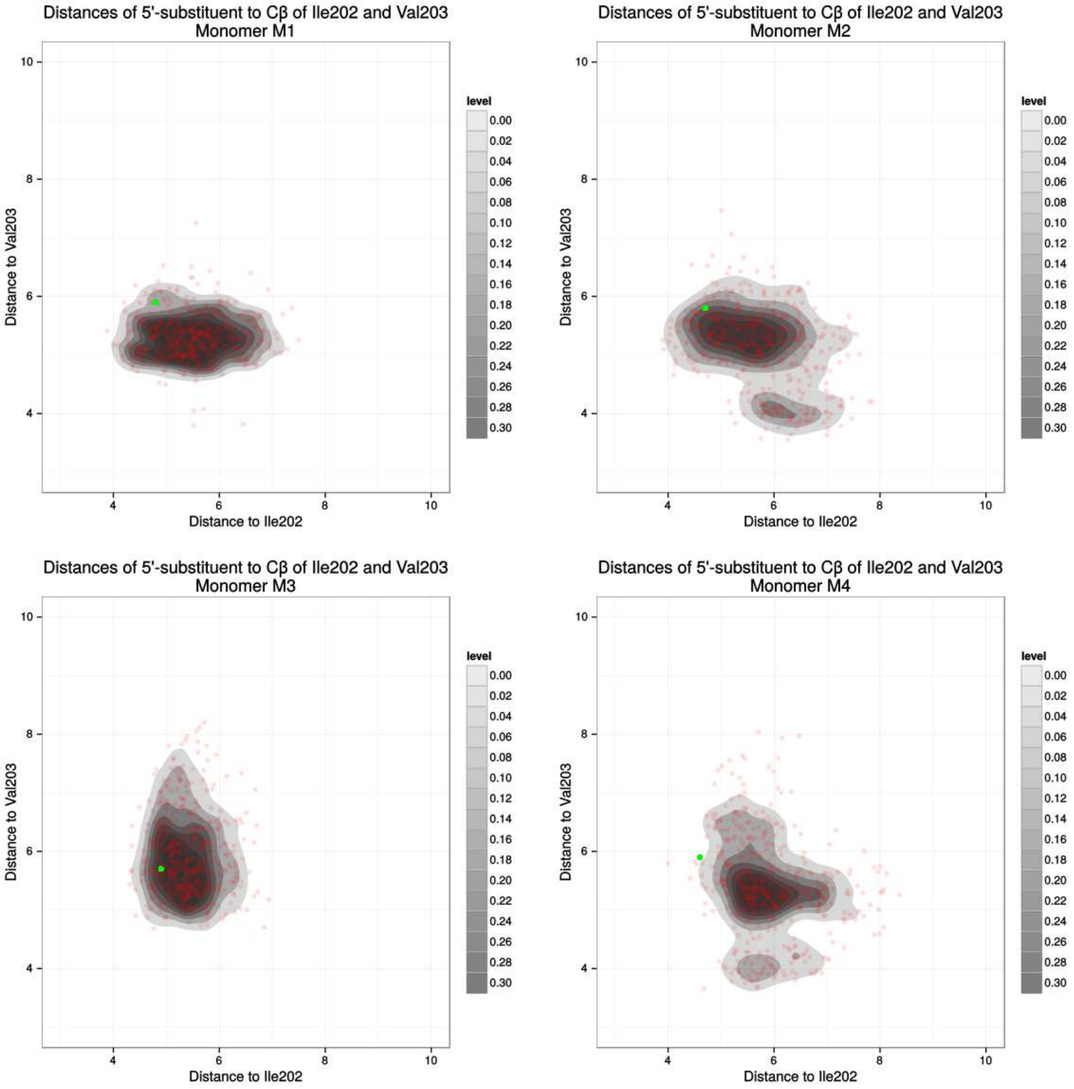


Figure 4.2 Density plots of distances in Å between the 5'-substituents of 5'-methyl-PT70 and the C$_\beta$ atoms of Ile202 and Val203, respectively, for four monomers based on 500 snapshots each. Starting structure values are highlighted as green dots.

Table 4.2 Medoid distances of 5'-substituent to C$_\beta$ atoms of Val202 and Ile203. Medoids are determined using a PAM clustering (cf. Chapter 3.2). Two-dimensional medoids were chosen over one-dimensional medians to only consider actually occurring combinations of the two distances.

	Distance of 5'-methyl		**Distance of 5'-chloro**	
	to Ile202 [Å]	to Val203 [Å]	to Ile202 [Å]	to Val203 [Å]
Monomer 1	5.51	5.29	4.89	5.25
Monomer 2	5.62	5.20	5.58	6.83
Monomer 3	5.34	5.78	5.13	5.42
Monomer 4	5.82	5.31	5.83	5.03
Combined	5.53	5.41	5.40	5.40

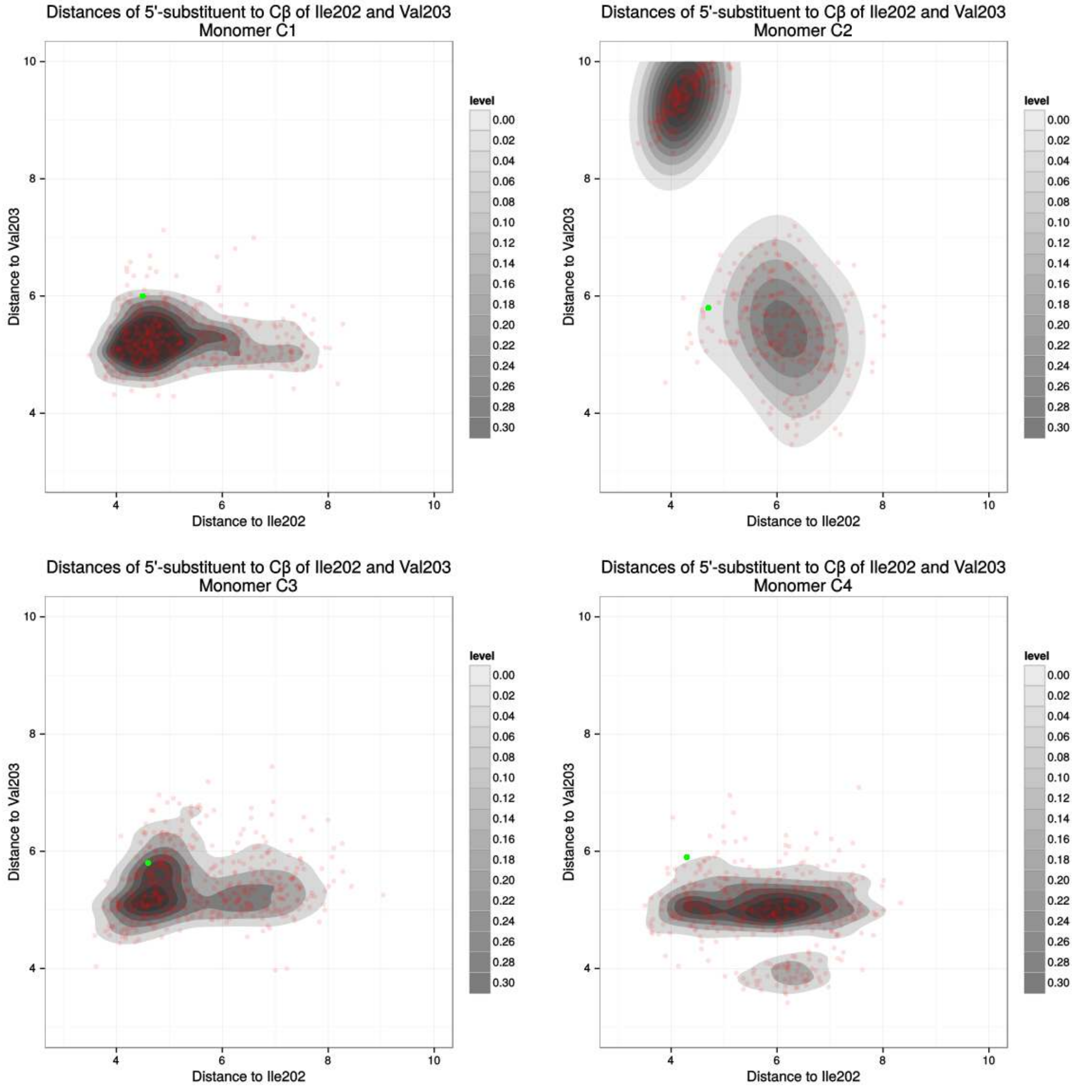


Figure 4.3 Density plots of distances in Å between the 5'-substituents of 5'-chloro-PT70 and the C$_\beta$ atoms of Ile202 and Val203, respectively, for four monomers based on 500 snapshots each. Starting structure values are highlighted as green dots.

Table 4.3 Distances of 5'-substituent to C$_\beta$ atoms of Val202 and Ile203 in docking poses.

	Distance of 5'-methyl		**Distance of 5'-chloro**	
	to Ile202 [Å]	to Val203 [Å]	to Ile202 [Å]	to Val203 [Å]
Monomer 1	4.8	5.9	4.5	6.0
Monomer 2	4.7	5.8	4.7	5.8
Monomer 3	4.9	5.7	4.6	5.8
Monomer 4	4.6	5.9	4.3	5.9

bordeaux rectangles in Figure 4.4). A hierarchical clustering with complete linkage was performed on the 2D-RMSD data, revealing three major conformational clusters of Met103 at a cutoff of 4 Å (Figure 4.5). The first cluster with a frequency of 35.16% represents the crystal structure conformation (Figure 4.5, pale green structure), whereas the smallest cluster (5.18% frequency, white structure in Figure 4.5) shows a side chain flip of Met103 with slightly altered position. The largest cluster (59.66% frequency, beige structure in Figure 4.5) shows a motion of Met103 towards the InhA major portal. After the conformational change, Met103 is no longer facing the edge of the ligand B-ring, but rather occupies the space between Ile202 and the ligand, thus increasing the distance of Ile202/Val203 and the 5'-substituent (Figure 4.6). The conformation of Met103 is seemingly influenced by the presence of a 5'-substituent. In the simulations of the InhA-**PT70** complexes (cf. Chapter 3), Met103 also shows a considerable flexibility (average RMSD of 1.76 Å ± 0.38 Å), which is, however, compensated by the available space due to the lack of a 5'-substituent of the ligand. In the case of the 5'-substituted diphenylethers, the flexible Met103 side chain is seemingly not well stabilized by the close contact to the 5'-substituent, but rather forced to adopt the aforementioned conformations due to potential clashes with the ligand.

4.2.2 Backbone and SBL stability

With respect to chain A of the 2X23 crystal structure, the spatial atomic displacements over time were quantified by RMSD measurements to assess the overall stability of the systems. In general, the *C*-monomers show very stable protein backbones with median RMSD values varying from 1.00 Å to 1.09 Å. The methyl-substituted system is somewhat less stable with median backbone RMSDs between 1.14 Å and 1.39 Å (Figure 4.7). Furthermore, the methyl-substituted monomers 2 to 4 display much higher fluctuations, represented by the high inter-quartile ranges (IQRs) of the respective RMSD distributions (cf. box width in Figure 4.7).

The same trend is visible in the RMSD distributions of the substrate binding loop (Figure 4.8). Obviously, the majority of the protein mobility is defined by the SBL motion, as already described in Chapter 3. Whereas the methyl-substitution is apparently not suited for SBL stabilization (median RMSD values from 2.12 Å to 3.71 Å), the 5'-chloro-substituted systems exhibit a stable SBL with medians below 1.86 Å. These results are comparable to the values of the **PT70** simulation (cf. Chapter 3). Up to 50 ns simulation time, the SBL of the **PT70**-bound monomers exhibit median RMSD values of 2.50 Å, 1.67 Å, 1.78 Å and 1.65 Å, respectively.

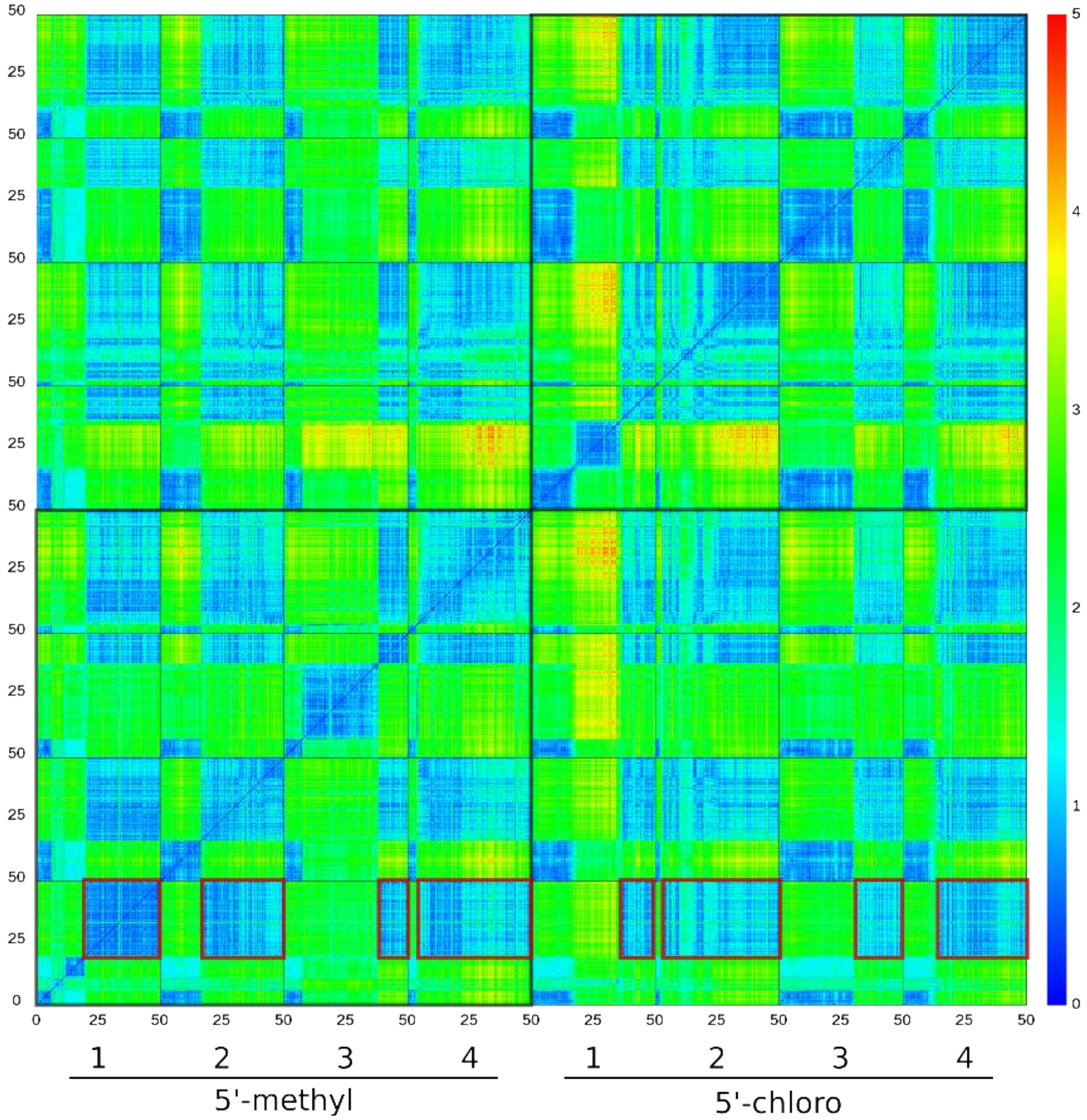


Figure 4.4 Heavy atom 2D-RMSD plot of Met103 of ternary InhA-NAD$^+$-5'-methyl-PT70 and InhA-NAD$^+$-5'-chloro-PT70 in their tetrameric forms in Å over 50 ns. Single monomers are framed by thin black lines. Thus, each small box represents the trajectory of a single chain over 50 ns sampling time. Large black rectangles delimit the tetrameric systems. Small bordeaux rectangles indicate a common conformation of Met103 at the end of each simulation.

4.2.3 Binding pocket stability and conformational family assignment

With one exception, the binding pocket conformation behaves very stably throughout the 50 ns simulation with median heavy atom RMSD values between 1.01 Å and 1.40 Å (Figure 4.9). Excluding monomer $C2$, conformational Family 1 (2X23 crystal structure conformation) is, hence, never left. Monomer $C2$ changes its binding pocket conformation after about 26 ns. By visual inspection of the trajectory and RMSD calculations with respect to the Family medoid structures (cf. Chapter 3), the alternative

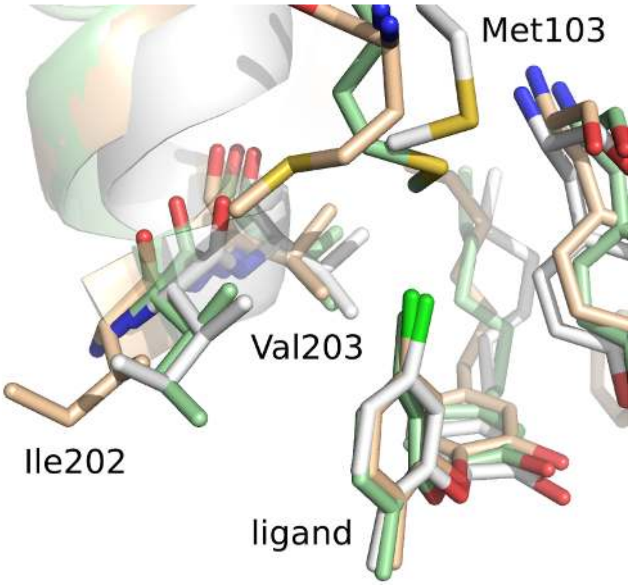


Figure 4.5 Medoid conformations of hierarchical clustering of Met103 2D-RMSD data. The medoid snapshot of cluster 1 is illustrated in pale green, representing the crystal structure conformation. Conformational cluster 2 is depicted in beige. The minor conformational cluster 3 is shown in white.

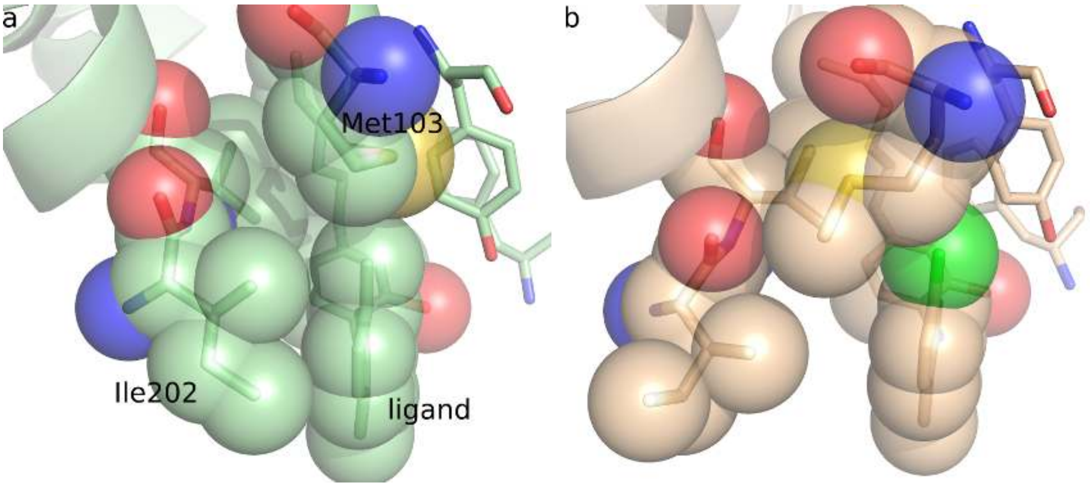


Figure 4.6 Medoid conformations of Met103 clusters 1 and 2. The medoid snapshot of cluster 1 is illustrated in pale green **(a)**. Conformational cluster 2 is depicted in beige **(b)**. Van-der-Waals radii are depicted as spheres.

binding pocket conformation in this monomer could be identified as a Family 3 conformation (Figure 4.10a and Table 4.4).

The atomic distances between Met103 and Ile202 were investigated over time (Figure 4.11). Interestingly, monomer $C2$ shows the lowest median distances between these residues. Moreover, the system exhibits very close contacts of Met103 and Ile202 immediately before the conformational transition to Family 3 occurs (emphasized by black circles in Figure 4.11). After the transition, Met103 occupies space close to the previous area of Ile202 (Figure 4.10b). This suggests that close contact between Met103 and

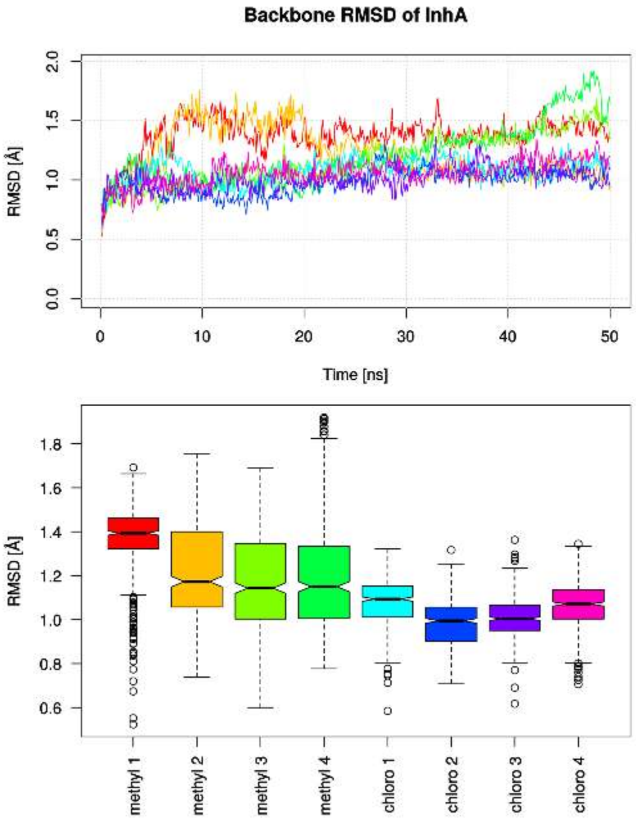


Figure 4.7 Backbone RMSD of single InhA monomers bound to 5'-substituted PT70-derivatives in Å. The RMSD values are illustrated over time (top) and as collective distributions, depicted as boxplots (bottom).

Table 4.4 RMSD of InhA-NAD$^+$-5'-chloro-PT70 binding pocket of monomer 2 at 0 ns and 50 ns compared to selected conformational family medoids (cf. Chapter 3).

	pocket RMSD at 0 ns	**pocket RMSD at 50 ns**
Family 1 medoid	**0.88**	2.82
cluster 4 medoid (Family 3)	3.28	2.09
cluster 5 medoid (Family 3)	3.29	1.80
cluster 6 medoid (Family 3)	2.52	**1.48**
Family 3 medoid	2.93	**1.50**

Ile202, initiated by contacts of Met103 with the 5'-substitution of the ligand, may lead to the conformational transition of the binding pocket.

Interestingly, the SBL of monomer $C2$ does not exhibit higher fluctuations after the appearance of Family 3 conformations of the binding pocket (cf. Figure 4.8). This possibly stems from the hydrophobic interaction between Met103 and Ile202 after the conformational transition, stabilizing helix $\alpha 6$ of the SBL (cf. beige structure in Figure 4.10b). This is underlined by the observation that the adopted conformation of Ile202 and Met103 is very stable throughout the remaining sampling time of the MD

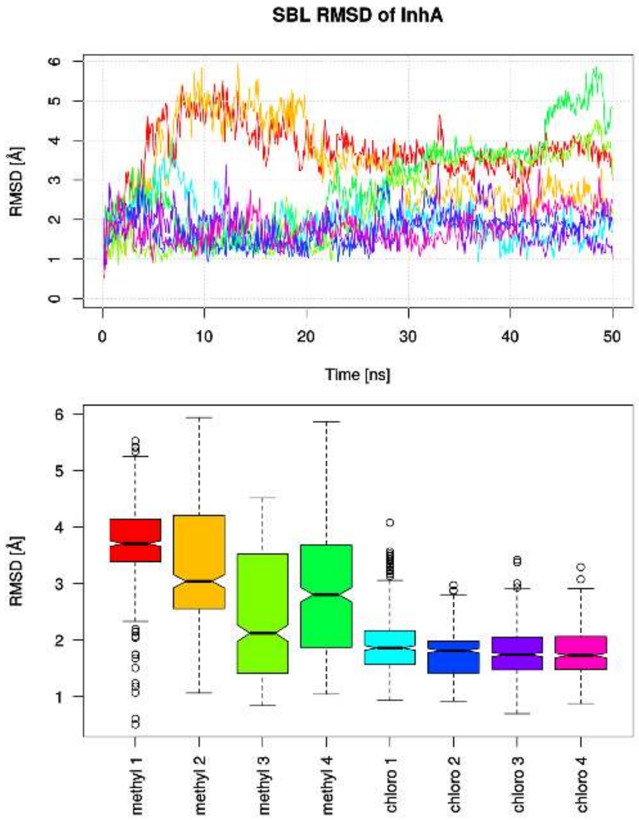


Figure 4.8 SBL backbone RMSD of single InhA monomers bound to 5'-substituted PT70-derivatives in Å. The RMSD values are illustrated over time (top) and as collective distributions, depicted as boxplots (bottom).

simulation (Figure 4.12), suggesting that the new conformation might include a new mechanism of SBL stabilization via an interaction of helix $\alpha6$ with Met103. Thus, a conformational family assignment solely based on the previously considered six pocket residues might not be sufficient in the case of the 2',5'-disubstituted diphenylethers and neighboring residues should be included as well. Accordingly, the observed conformation is hereinafter termed Family 3*.

4.2.4 Ligand stability

In both systems, the ligand itself shows a very high stability, underlining the validity of the used docking poses (Figure 4.13). With median heavy atom RMSD values in the range of 0.98 Å and 1.15 Å the ligands do not exhibit major conformational changes throughout the observed sampling time.

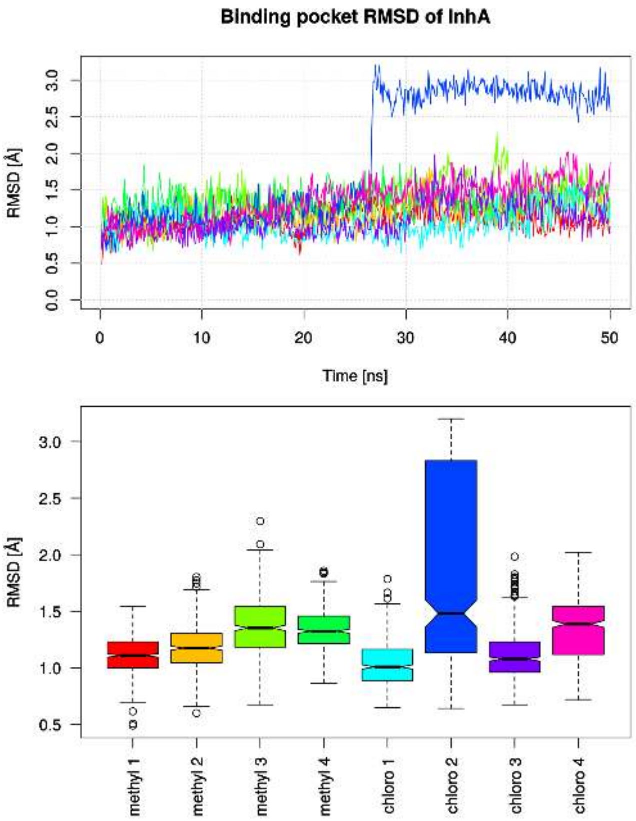


Figure 4.9 Heavy atom RMSD of single InhA binding pockets bound to 5'-substituted PT70-derivatives in Å. The RMSD values are illustrated over time (top) and as collective distributions, depicted as boxplots (bottom).

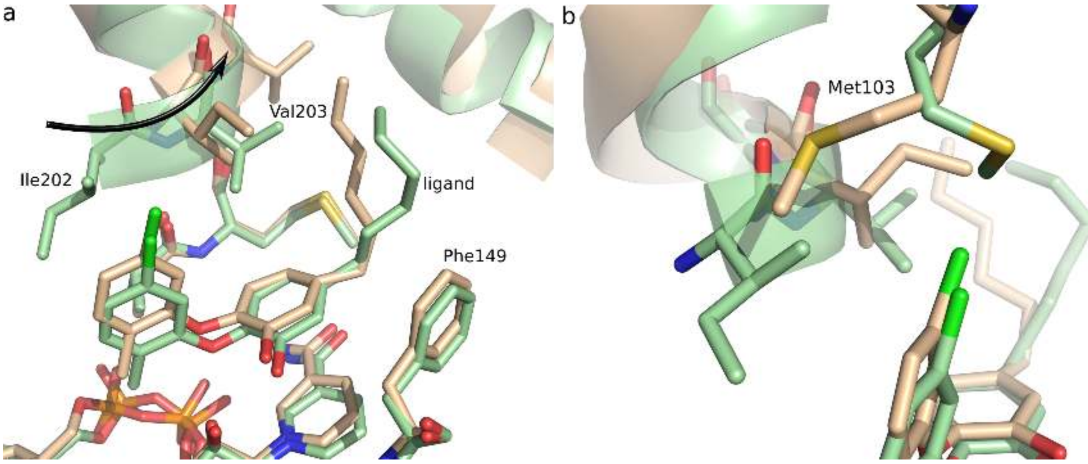


Figure 4.10 Snapshots of InhA-NAD$^+$-5'-chloro-PT70 monomer 2 at the beginning of the simulation (0 ns, green) and after 50 ns of simulation (beige). **(a)** The arrow indicates a Family 3 transition with a shift of Ile202 and Val203 towards the inside of the binding pocket. **(b)** Met103 is illustrated. After the transition, Met103 is located close to the previous area of Ile202.

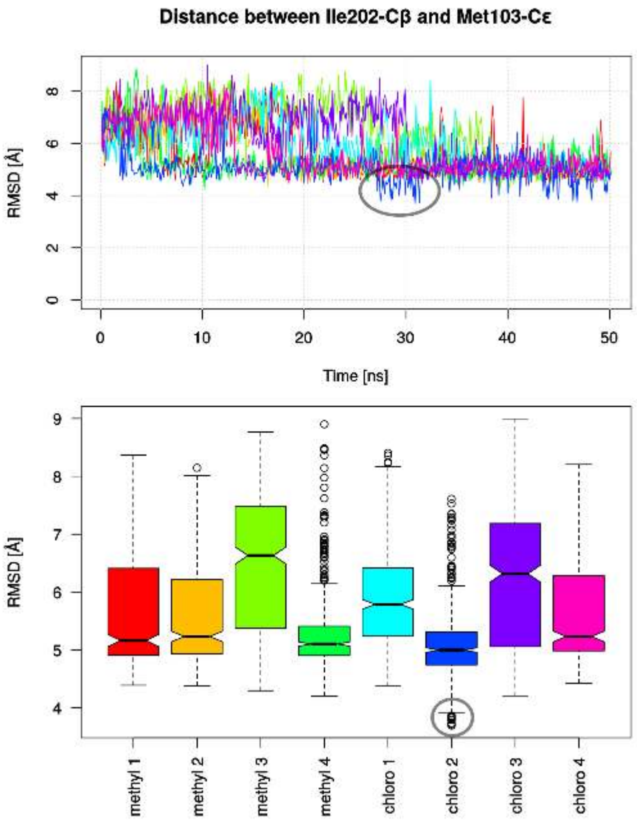


Figure 4.11 Distances between Met103-C$_\epsilon$ and Ile202-C$_\beta$ in InhA systems bound to 5'-substituted PT70-derivatives in Å. The distances are illustrated over time (top) and as collective distributions, depicted as boxplots (bottom). The circled area emphasizes close contacts of Met103 and Ile202 in monomer $C2$, leading to a conformational transition.

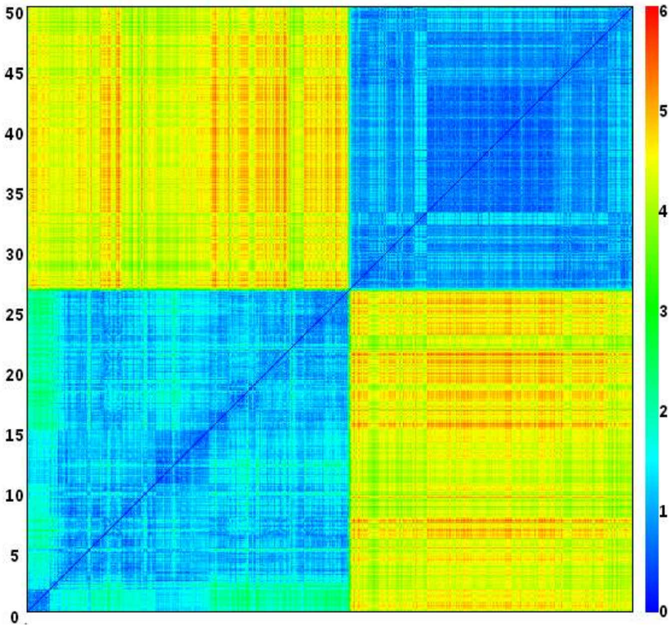


Figure 4.12 2D-RMSD plot of heavy atoms of residues Ile202 and Met103 of monomer *chloro* 2 in Å. The conformational change of these residues after approximately 26 ns is stable throughout the remaining simulation time.

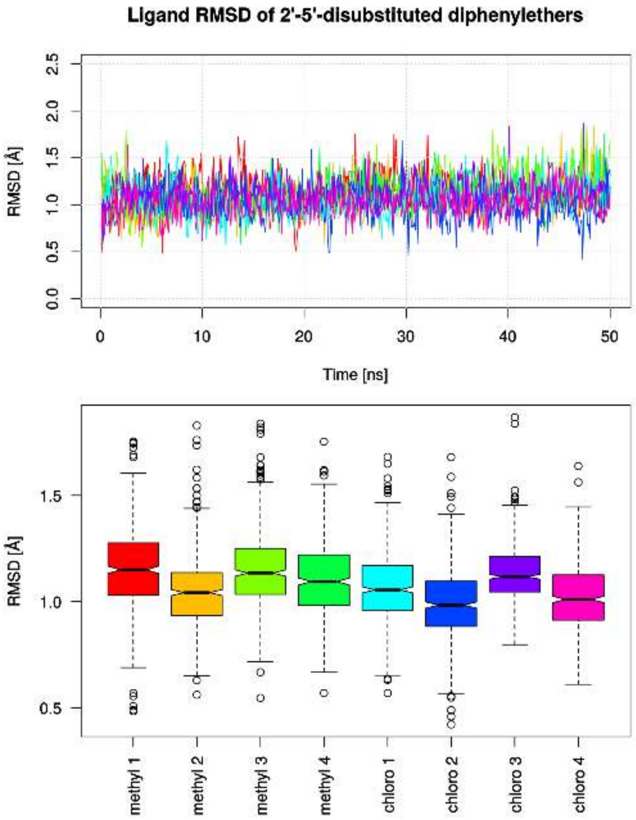


Figure 4.13 Heavy atom ligand RMSD of 5’-substituted PT70-derivatives in Å. The RMSD values are illustrated over time (top) and as collective distributions, depicted as boxplots (bottom).

4.2.5 Analysis of Met161

In the course of the analysis, the side chain of residue Met161 was observed to interact weakly with the inhibitor **PT70** in the crystal structure 2X23, an interaction which has not yet been discussed in the literature for diphenylethers in InhA. Methionine-aryl interactions are important and frequently occurring motifs in crystal structures [158–160]. According to Bissantz *et al.* [158], the C–S–C thioether preferentially binds to the aromatic system in the aryl ring plane, however, there are also complexes described, in which the C–S–C fragment binds to adenine (e.g., as part of cofactors) in a coplanar fashion at a distance of about 4 Å. In a recent survey by Beno and colleagues [160], the authors state that preference regarding the interaction of the sulfur with the π-face or π-edge cannot be clearly derived by means of a crystallographic database analysis. High-level quantum mechanical (QM) calculations, on the other hand, suggest preferential binding of dimethyl sulfide (DMS, mimicking the methionine side chain) to the π-face, rather than the edge [160].

To quantify the stability of the interaction between the sulfur and the 2’,5’-disubstituted

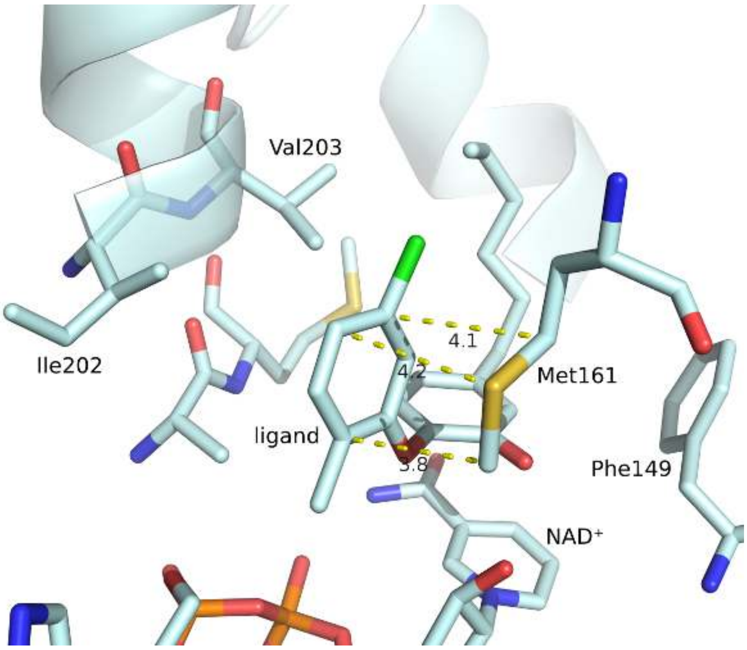


Figure 4.14 InhA-NAD$^+$-5'-chloro-PT70 monomer 1 after 50 ns of simulation. Selected contacts between Met161 side chain atoms and B-ring carbon atoms are depicted as yellow dotted lines, with distances labeled in Å underneath.

diphenylether B-ring (Figure 4.14) over the MD simulation time, the distance between its 4'-carbon atom and the Met161 sulfur atom was measured (Figure 4.15). Furthermore, the absolute deviation and flexibility of Met161 over the entire trajectory was assessed by means of heavy atom RMSD of the residue (Figure 4.16). Median distances in the range of 4.29 Å and 4.74 Å and very low median RMSD values between 0.66 Å and 0.92 Å suggest that an aryl-methionine interaction is maintained stably in all systems.

4.3 Discussion

In Chapter 3, the importance of locking the system in the assumed EI* state of ligand association (2X23 crystal structure conformation) was emphasized. The most frequent binding pocket conformations of InhA bound to diphenylether inhibitors–besides the EI* state–were the conformational Families 2 and 3. These are primarily defined by a shift of Ile202 and Val203 towards the inside of the binding pocket, leading to Ile202 occupying the previous position of Val203. This conformation with an open helix $\alpha6$ is considered the EI state of drug-target complex formation (cf. Chapters 3 and 6). Besides occupation of the hydrophobic pocket of InhA and an anchor-substituent in 2'-position of the diphenylether B-ring, an additional substitution of the 5'-position was suggested as a possible ligand modification to increase the energy barrier between EI and EI* and, thus, the drug-target residence time (cf. Chapter 3).

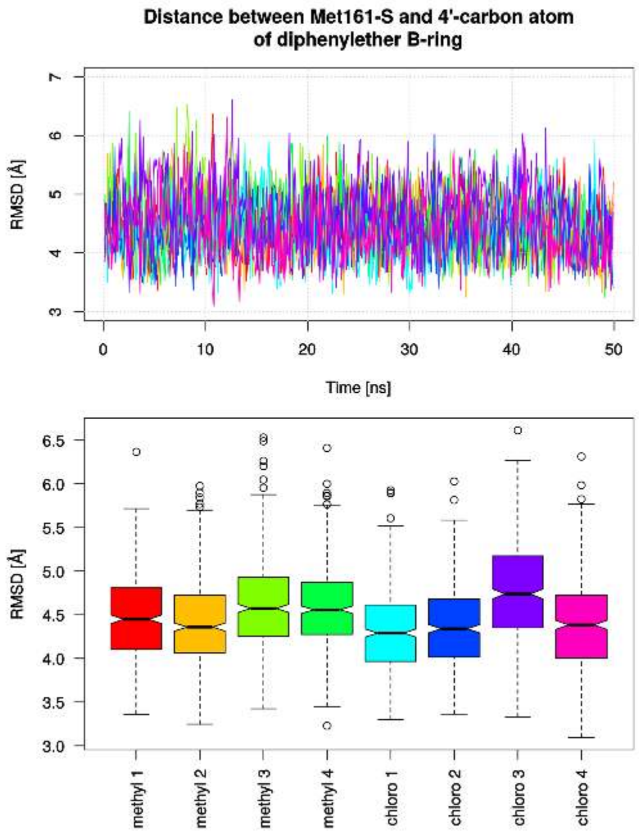


Figure 4.15 Distances between Met161-sulfur and ligand-4'-carbon atom in InhA systems bound to 5'-substituted PT70-derivatives in Å. The distances are illustrated over time (top) and as collective distributions, depicted as boxplots (bottom).

Diphenylethers with a single *meta*-substitution at the B-ring have been evaluated in the literature [17, 161]. In some cases, the *meta*-substitution (**PT11**, nitro; **PT20**, oxamic acid; and **PT29**, isoxazole-5-carboxamide) resulted in improved affinity, compared to the respective *ortho*- or *para*-substituted counterparts (**PT10/PT12**, **PT19/PT21** and **PT28**). In other cases, *meta*-substitutions (**PT14**, amino; **PT17**, acetamide and **PT29**) resulted in a significant decrease in affinity, compared to their respective counterparts (**PT13/PT15**, **PT16/PT18** and **PT30**). However, these inhibitors are all mono-substituted at the B-ring. The effect of a 2',5'-disubstituted B-ring has not yet been investigated experimentally.

Docking of the 5'-substituted **PT70** derivatives yielded valid poses with very low RMS deviations from the **PT70** crystal poses. Although the size of a substituent in 5'-position of the B-ring is confined by the proximity of the Met103 side chain, docking confirmed that sufficient space is available for small substituents (cf. Chapter 3). These substituents are assumed to embed between Ile202 and Val203 to lock this part of the SBL in the crystal structure conformation, i.e. in the EI* state. Whereas the methyl-group forms van-der-Waals contacts with these residues, the chlorine-substituent additionally has a

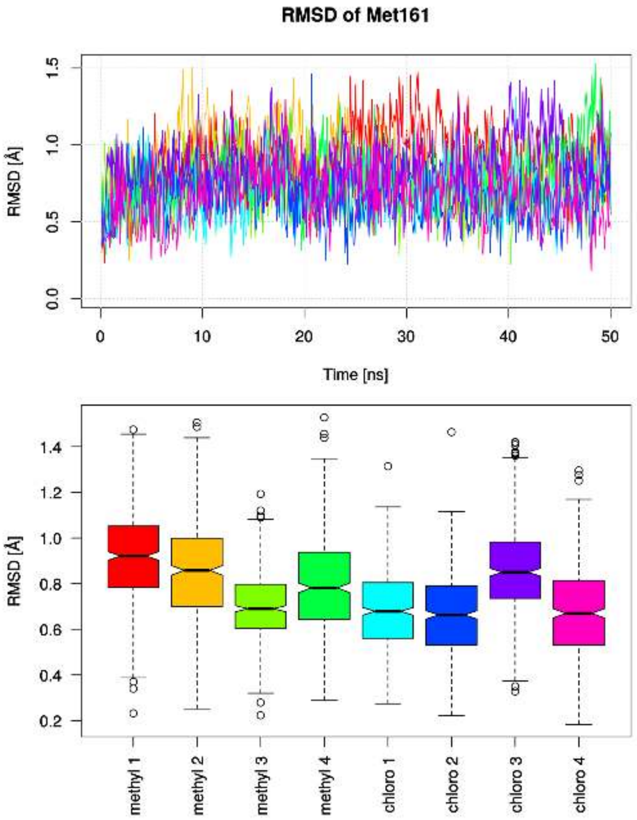


Figure 4.16 Heavy atom RMSD of Met161 in InhA systems bound to 5'-substituted PT70-derivatives in Å. The RMSD values are illustrated over time (top) and as collective distributions, depicted as boxplots (bottom).

$-I$-effect on the aromatic B-ring (exceeding its $+M$-effect), which might further stabilize the not yet described weak electrostatic interaction between the aromatic ring and the sulfur of Met161, detected in the course of the analysis (cf. Figure 4.14).

In general, the 50 ns MD simulations of 5'-methyl-**PT70** and 5'-chloro-**PT70** showed reasonable stability of the InhA binding pocket, substrate binding loop and ligands. In particular, the chlorine-substitution resulted in highly stable SBLs (cf. Figure 4.8), although it should be noted that the sampling time of 50 ns reached in this study is only a third of the previously simulated InhA systems (150 ns per monomer of Chapter 3). Interestingly, the conformational transition of the binding pocket to Family 3* could not be prevented in one of the four 5'-chloro-**PT70**-bound monomers, although the SBL is stable over the entire simulation time. The most likely reason for this transition is a close contact of Met103 to the 5'-substituent of the ligand and Ile202, eventually resulting in displacement of Ile202 towards the inside of the binding pocket and Met103 occupying space close to the former position of the Ile202 side chain. This behavior might suggest that the evaluated 5'-substituted ligands are not sufficiently well able to lock Met103 in the crystal structure conformation, as also shown in Figure 4.4. On the other hand, the new conformation with hydrophobic interactions between Met103 and Ile202 also led to a very stable SBL, which might suggest a new and stable alternative pocket and SBL conformation based on contributions of Met103.

The methyl-substitution resulted in interesting dynamic behavior. Whereas the binding pocket does not leave the Family 1 conformation, i.e., is stable over 50 ns of simulation, the SBLs of the M-monomers show much higher variances and higher absolute RMSD medians (cf. Figure 4.8). Based on the simulations of experimentally characterized diphenylethers (cf. Chapter 3), a Family 1 conformation of the InhA binding pocket (defined by six pocket residues) is a prerequisite for SBL stabilization. Regarding the hypothetical 2',5'-disubstituted diphenylethers, previously undetected additional effects on SBL stability can be observed. By extending the considered pocket residues by Met103, conformational Family 3* was described, which shows a very stable SBL, as suggested by the SBL and pocket dynamics of monomer $C2$. Furthermore, it is notable that 5'-methyl-**PT70** leads to large fluctuations of the SBL, despite a stable Family 1 conformation of the pocket. This suggests that a Family 1 conformation alone is not a sufficient structural characteristic for ordering of the SBL in the case of the 2',5'-disubstituted diphenylethers, as shown by the dynamic behavior of the SBL and pocket of the M-monomers. An RMSD analysis of separate parts of the ligands shows that the M-monomers exhibit significantly more flexible hexyl chains than the C-monomers ($p \ll 0.001$, Mann-Whitney-U test), whereas the remaining diphenylether systems do not behave differently (Figure 4.17; $p = 0.3207$, Mann-Whitney-U test). As described in Chapter 3, a significant difference in the mobility of the hexyl residue could also be observed between the slow-onset inhibitor **PT70** and the rapid reversible ligand **6PP** (cf. Figure 3.16). This observation suggests that the methyl substitution might, thus, translate into an increased flexibility of the hexyl residue and hamper the SBL stability.

In the case of monomers $M3$ and $M4$, the SBLs show increasing deviations from the starting conformation, as represented by the RMSD evolution over time (cf. Figure 4.8). Interestingly, both loops show a steep rise in RMSD after approximately 20 ns of simulation, which is preceded by a close contact between Gln214 of the SBL and Pro156 of the protein core (Figure 4.18). The motion of the Gln214 side chain towards Pro156 seemingly disrupts the order of helix $\alpha7$ of the SBL, which is not restored in the observed simulation time (Figure 4.19).

The analysis of Met161 suggests a stable aryl-methionine interaction between the ligand B-ring and Met161, with a median distance range of 4.29 Å to 4.74 Å (cf. Figures 4.15, 4.16 and 4.14). So far, this interaction has not yet been discussed in the literature for diphenylethers in InhA.

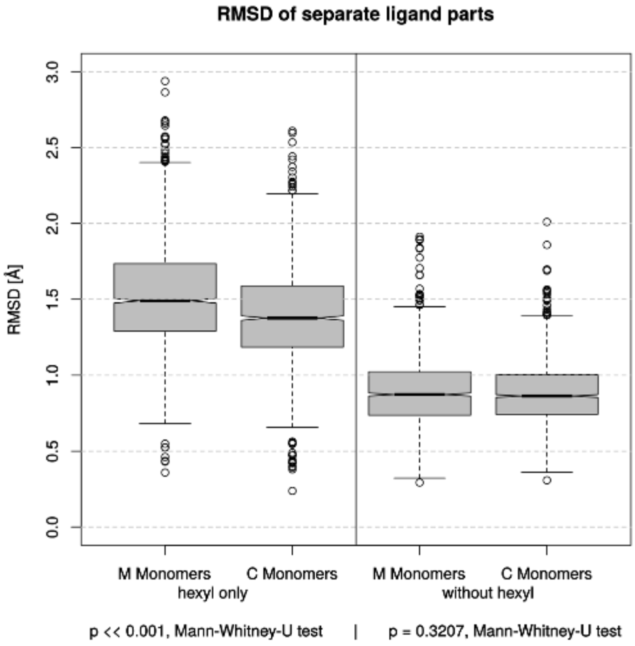


Figure 4.17 Collective heavy-atom RMSD of separate ligand parts of M and C monomers over 50 ns in Å. RMSD values of each ligand were measured separately with respect to each starting structure. A significant difference between the RMSD of the hexyl chains can be observed between the M and the C monomers ($p \ll 0.001$, Mann-Whitney-U test).

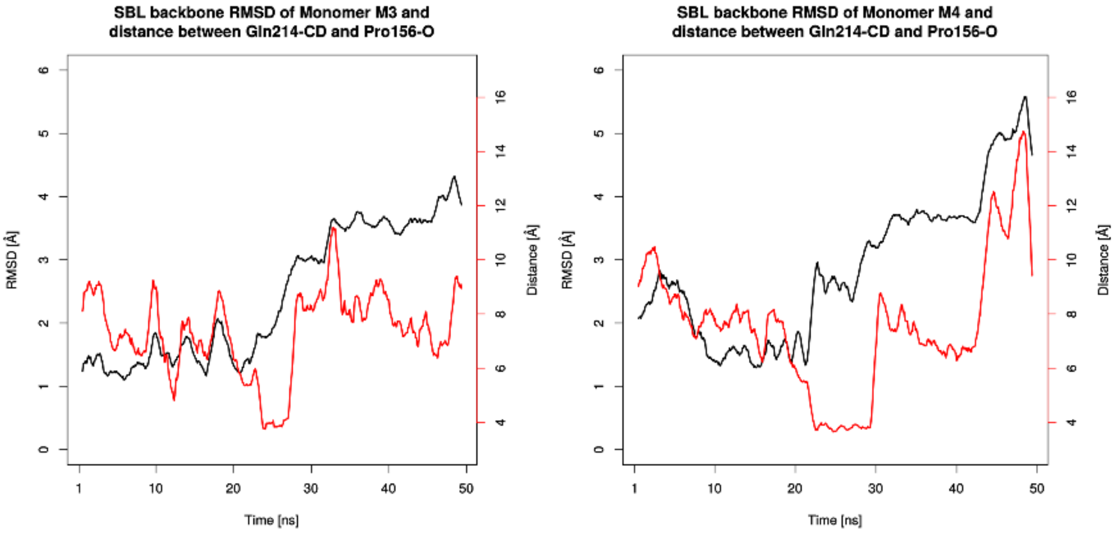


Figure 4.18 SBL backbone RMSD (black) and distance (red) between the Gln214-CD atom and the Pro156-O atom of monomers $M3$ and $M4$ in Å. A moving average with a window size of 10 frames was used.

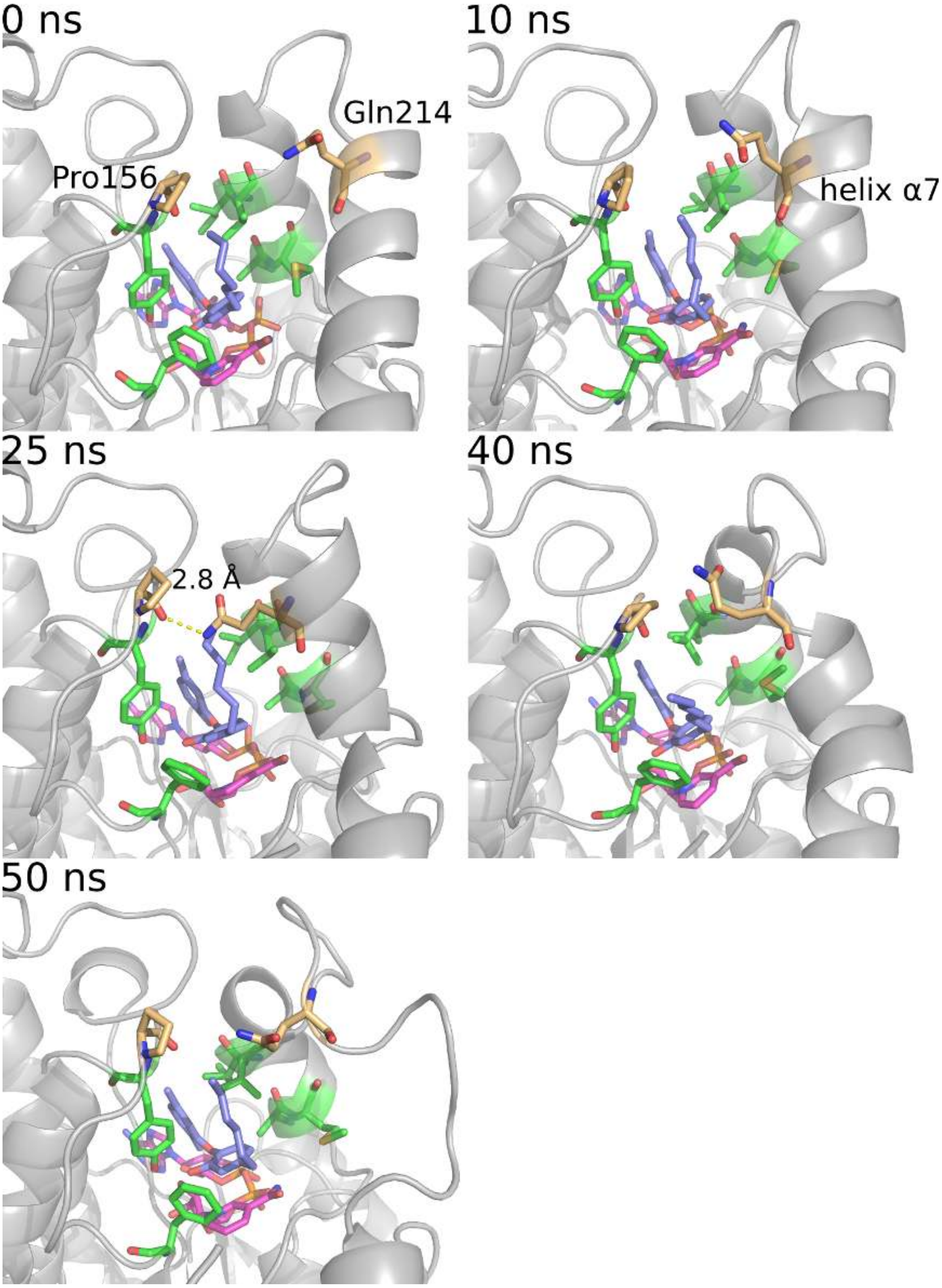


Figure 4.19 Trajectory snapshots of monomer $M4$**.** Residues Pro156 and Gln214 are represented in orange. Pocket residues are depicted in green, the ligand and the cofactor are illustrated in slate blue and magenta, respectively. A dashed yellow line indicates a close contact of 2.8 Å between Pro156 and Gln214 at 25 ns of simulation time, followed by a disordering of helix $\alpha 7$.

4.4 Conclusion

Docking and MD simulations suggest that a small substituent in 5'-position of the diphenylether B-ring generally leads to stable conformational dynamics of the protein-ligand complexes. Of the two exemplary systems, the chloro-substituted **PT70** showed a higher stabilization of the backbone, the SBL and the Met161 side chain. Also the binding pocket was well stabilized in this system, with the exception of monomer 2, where a Family 3* conformation was adopted after half of the sampling time. This conformational family includes an alternative conformation of Met103 interacting with Ile202 and stabilizing the SBL.

Although the occurrence of conformational Family 3* had no negative impact on the stability of the SBL over the entire duration of the simulation after the conformational change (∼24 ns), it also highlights a possible problem of 2',5'-disubstituted diphenylethers. While docking of the ligands resulted in valid poses and most regions of the protein behave stably in MD simulations, Met103 is highly flexible in each monomer and not well stabilized by the 5'-substituent of the ligand. Thus, a possible approach to further improve the dynamic effects of these ligands on InhA might also have to include better stabilization of Met103 in the crystal structure conformation.

Whereas the M-monomers show a stable binding pocket, the SBL shows large fluctuations. Although a Family 1 conformation is necessary for SBL stabilization (cf. Chapter 3), analysis of the M-monomers suggests that a Family 1 pocket conformation alone is apparently not a sufficient structural feature for proper stabilization of the SBL in case of the hypothetical 2',5'-disubstituted diphenylethers. The origin of the difference in SBL stabilization between the M and C-monomers might stem from a different stabilization of the flexible hexyl residue of the ligands in the hydrophobic pocket and an electrostatic interaction between the Pro156 backbone and Gln214.

Chapter 5

An accurate and quantitative prediction model for drug-target residence time of *Staphylococcus aureus* FabI inhibitors based on Steered Molecular Dynamics

To rationally modify the drug-target residence time of a protein-ligand complex, consideration of the energy barriers and transition states of ligand dissociation is imperative [26]. However, common computational methods for the structural and energetic characterization of protein-ligand complexes, such as scoring functions, LIE or FEP, are not able to assess residence times, since they only consider end-points of ligand binding. Hence, the non-equilibrium MD variant Steered Molecular Dynamics (SMD) has gained increasing attention for the investigation of ligand kinetics (cf. Chapter 2.3) [38, 39, 162]. Free energy profiles can be calculated from SMD simulations with the ligand pulling direction as the reaction coordinate. As opposed to Umbrella Sampling (US) or Metadynamics, SMD simulations are relatively straightforward to set up.

In the following chapter, SMD simulations are used to reconstruct multiple ligand dissociation events for protein-ligand complexes, starting from crystal structure conformations, i.e., presumed EI* states, to obtain the corresponding free energy profiles along the direction of extraction (Figure 5.1). This study is carried out for an experimentally well characterized series of diphenylether inhibitors (cf. Figure 2.4) bound to the enoyl-ACP reductase FabI of *Staphylococcus aureus* (cf. Figure 2.2).

5.1 Protein and ligand preparation

The available crystal structures of *sa*FabI bound to diphenylethers with experimentally measured residence times [30, 58] are summarized in Table 5.1. Chain A was extracted from each of the 11 crystal structures, including all crystallized water molecules, as well as the ligand and the cofactor. Hydrogen atoms were added to ligands and cofactors with the Protonate3D tool implemented in MOE [163]. The Amber12 [164] module

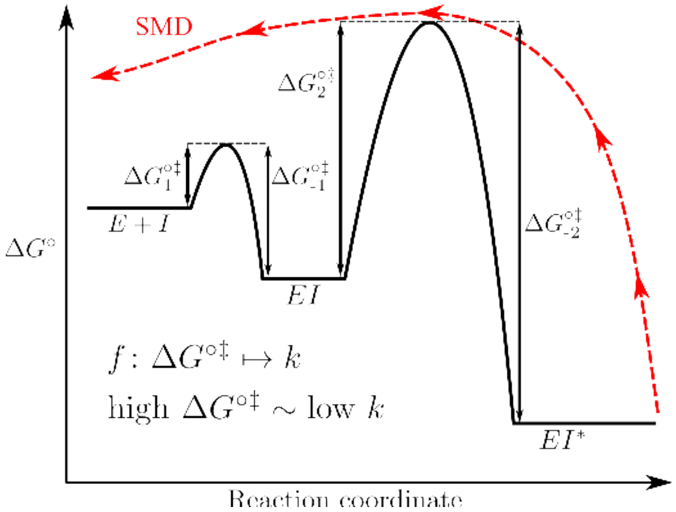

SMD
$\Delta G_2^{\circ\ddagger}$
$\Delta G_1^{\circ\ddagger}$
$\Delta G_{-1}^{\circ\ddagger}$
$\Delta G_{-2}^{\circ\ddagger}$
ΔG°
$E + I$
EI
EI^*
$f: \Delta G^{\circ\ddagger} \mapsto k$
high $\Delta G^{\circ\ddagger} \sim$ low k
Reaction coordinate

Figure 5.1 Two-step mechanism of drug-target complex formation. Equilibrium of inhibitor binding via an induced-fit two-step mechanism represented as schematic free-energy profile for this reaction. E denotes the enzyme, I the inhibitor, EI the initial enzyme-inhibitor complex, and EI* the final enzyme-inhibitor complex. A high energy barrier ($\Delta G^{\circ\ddagger}$) corresponds to a low reaction rate constant k. The red dashed line schematically represents the induced extraction of the ligand along a putative reaction coordinate using SMD simulations.

tleap was used for assigning the parameters of the ff99SB force field. His253 was protonated in δ-position and His247 was doubly protonated to maintain a hydrogen bond network to surrounding residues. RESP charges [110] were calculated for all ligands and cofactors based on HF/6-31G* electrostatic potentials obtained with Gaussian 09 [115]. With parmchk [152] unavailable force field parameters were calculated according to the General Amber Force Field (GAFF) [114]. Atom and bond types of the ligands were assigned by antechamber [152]. All diphenylethers were parameterized in their deprotonated form at the A-ring, resulting in a negative charge [30]. A short energy minimization of 200 cycles was performed on the entire complex using a generalized Born implicit solvent model [153, 154] with the Amber module sander [164]. Afterwards, the system was solvated with tleap in a TIP3P water box [112] with a margin of 10 Å and Na^+ counterions.

For all subsequent equilibration and production runs, the MD package NAMD 2.9 [101, 156] was employed with the previously assigned Amber force field parameters [99]. After 10,000 steps of energy minimization, each system was heated from 100 K to 300 K in 500 ps in a constant-volume box, while harmonic constraints were applied to all non-solvent atoms with a force constant of 0.5 kcal/(mol Å^2). Simultaneously, the constraints were gradually released during the heating period. Afterwards, the systems were allowed to evolve freely for another 500 ps, resulting in 1 ns of equilibration. The systems were treated with periodic boundary conditions. A van der Waals interaction cutoff of 12 Å was used, as well as the particle mesh Ewald methodology (PME) for electrostatic

Table 5.1 Used crystal structures and corresponding diphenylether ligands. The diphenylether scaffold is shown in Figure 2.4. Kinetic data are taken from [30]. For linear regression model generation (cf. Chapter 5.6) the average of the available t_R values of a given complex was used. Complexes are sorted by increasing residence time t_R.

PDB code	**ligand**	$\mathbf{R_1}$	$\mathbf{R_2}$	$\mathbf{R_3}$	K_i **[nM]**	t_R **[min]**
4BNK	PT55	F	H	H	1.42	2.4 - 6.2
4BNJ	PT53	CH_3	H	H	0.38	13.2 - 17.3
4ALJ	PT52	Cl	H	H	0.12	25.8 - 28.2 - 35.3
4BNI	PT13	$(CH_2)_5CH_3$	H	NH_2	0.12	68.5
4BNL	PT68	$CH_2CH{=}CH_2$	H	H	0.11	68.7
4ALK	PT01	CH_2CH_3	H	H	0.09	63.8 - 83.3
4BNF	PT02	$(CH_2)_2CH_3$	H	H	0.07	60.9 - 105.3
4ALI	TCL	Cl	Cl	Cl	0.05	139.5
4BNG	PT03	$(CH_2)_4CH_3$	H	H	0.04	187.5 - 300.0
4BNH	PT04	$(CH_2)_5CH_3$	H	H	0.01	461.5
4BNN	PT119	$(CH_2)_5CH_3$	H	CN	0.01	750

interactions [108]. Constant temperature was achieved by the use of Langevin dynamics with a damping coefficient of 5 ps^{-1}. A 2 fs time step was used for all simulations. SHAKE constraints were applied on covalent bonds to hydrogen atoms. For all subsequent simulations the MD conditions were changed from a constant-volume (NVT) to a constant-pressure ensemble (NPT). Constant pressure was assured by the Nosé-Hoover Langevin piston pressure control with a barostat oscillation time of 100 fs and a barostat damping time of 50 fs [106, 107]. Trajectory snapshots were written at an interval of 1 ps.

5.2 Determining the ligand egress route

The complex 4BNF with the medium sized ligand **PT02** was chosen for a Random Accelerated Molecular Dynamics (RAMD) experiment to determine potential exit pathways using the RAMD tcl script of NAMD [124, 125]. In this setup, a constant acceleration of 0.2 kcal/(mol Å amu) was applied to the ligand **PT02**, starting after 5 ns of unbiased MD simulation. The direction of the acceleration was mutated every 100 steps, unless the ligand showed an RMSD larger than 1 Å from the previous position. The simulation was terminated if the ligand had traveled more than 30 Å from its starting coordinates. Trajectory snapshots were written every 0.2 ps.

Two ligand egress routes could be detected. The exit event in RAMD simulation 1 occurred via the major portal of the FabI binding pocket, whereas exit pathway 2 emerged in a second RAMD simulation through the minor portal (cf. Figure 2.3). The exit pathway via the major portal was further used, since the natural substrates of *sa*FabI are assumed to bind via this route [59]. Accordingly, the vector of the detected exit pathway was applied on all complexes for SMD simulations. Due to the accuracy of the resulting regression model (cf. Chapter 5.6) the chosen exit pathway is assumed to be valid. An extensive RAMD approach with multiple replica simulations [162], however, might allow to further optimize the reaction coordinate for SMD simulations.

5.3 Steered Molecular Dynamics simulations

The last snapshot of the used RAMD simulation (i.e., protein with fully dissociated ligand) was aligned to each of the respective end-points of the 1 ns equilibrations to derive the pulling direction of the ligands for constant-velocity SMD simulations (cvSMD). Force profiles for the induced withdrawal of each ligand were obtained while applying the same pulling direction to all prepared complexes. All ligands were connected to the SMD spring with a large force constant of 10,000 pN/Å to fulfill a requirement of the stiff-spring approximation [35]. The pulling speed was set to a constant 10 Å/ns. Thus, by using 2 ns simulations, a pulling distance of 20 Å could be achieved. To avoid translation of the protein itself, mild harmonic constraints with a force constant of 1 kcal/(mol Å^2) were assigned to all C$_\alpha$-atoms. Each of the 11 complexes was subjected to SMD simulations in 30 independent replicas, resulting in 660 ns of SMD simulation data. For structural analyses, snapshots at intervals of 10 ps were considered, resulting in 200 frames per replica simulation. The force on the SMD spring was measured every 10 time steps (20 fs). See Chapter 5.9 for a discussion of the selected SMD parameters.

5.4 Structural dynamics of the FabI binding pocket

According to Schiebel *et al.* [58], the following residues interact with the diphenylether **TCL** and can thus be considered the core binding pocket: Ala95, Phe96, Leu102, Tyr147, Met160, Ser197, Ala198, Val201, Phe204, Ile207 (hydrophobic contacts), as well as Ala97, Tyr157 and the cofactor NADP$^+$ (hydrogen bonds) (Figure 2.3). Particularly important is here a conserved hydrogen bond between the hydroxyl moiety of Tyr157 and the phenolic oxygen of the diphenylether scaffold, which is assumed to bind in its deprotonated form to mimic the enolate intermediate of substrate turnover [30]. Hence, all ligands were parameterized accordingly with a negative charge.

During the second phase of equilibration (500 ps of unconstrained simulation after 500 ps of constrained simulation), the FabI backbone as well as the pocket residues are very

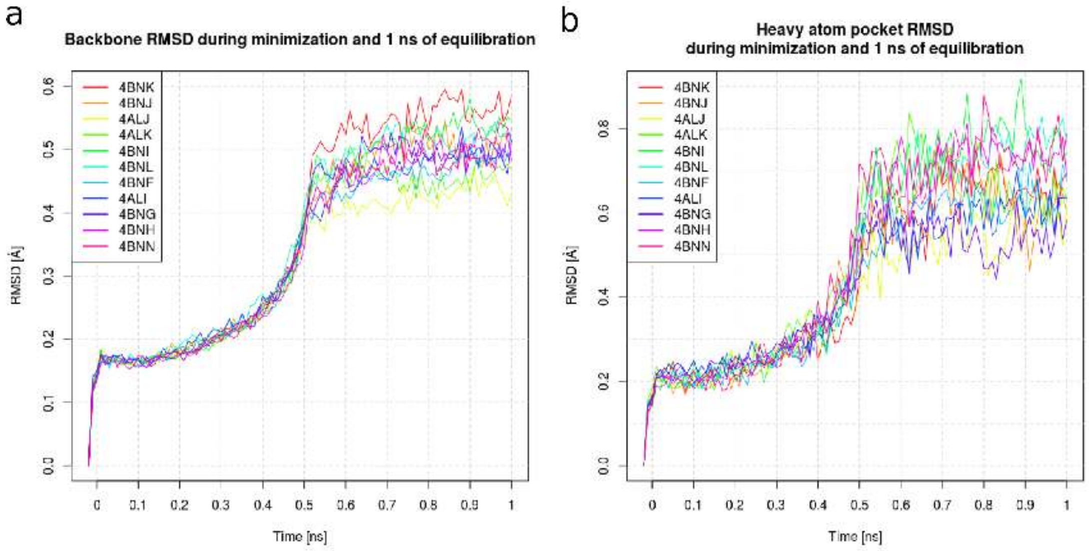


Figure 5.2 RMSD of complexes during minimization and 1 ns of equilibration. (a) Protein backbone atoms. **(b)** Binding pocket heavy atoms.

stable with average RMSD values ranging from 0.42 Å to 0.54 Å and 0.54 Å to 0.74 Å, respectively (Figure 5.2).

During the SMD experiments, the protein backbone and pocket heavy atoms generally behave very stably with RMSD values in the range of 0.40 Å ± 0.03 Å and 0.42 Å ± 0.04 Å, as well as 0.65 Å ± 0.15 Å and 0.93 Å ± 0.22 Å, respectively (average over time and 30 replicas, each). Some of the pocket residues show high deviations from the crystal structure conformation: (i) regarding side chain movement, residue Phe96 displays an elevated average RMSD (Figure 5.3a). Throughout all replicas of all complexes, Phe96 seems to have a gate keeper function. Upon dissociation, the passing ligand displaces the residue, thus opening the major portal. Also Phe204 and Ile207 in helix $\alpha7$ of the SBL exhibit large movements, which can be explained by the increase in available space after the ligand has left the binding pocket. Particularly ligands with large substituents in R_1-position occupying the hydrophobic pocket of FabI are affected. Furthermore, Ser197 shows an elevated average RMSD, which, however, is also due to backbone movements. As opposed to the remaining complexes, residue Leu102 only shows high deviations in the complex 4ALI. This behavior most probably stems from the different crystal structure conformation in this region, compared to the otherwise conformationally very similar remaining complexes (Figure 5.4). (ii) Regarding pure backbone movements of the pocket residues, it is notable that the residues of the substrate binding loop (Ser197, Ala198, Val201, Phe204, Ile207) generally show the highest average deviations from the crystal structure conformations (Figure 5.3b), particularly the pocket residues of helix $\eta6$ of the SBL (Ser197 and Ala198). Hence, the structural

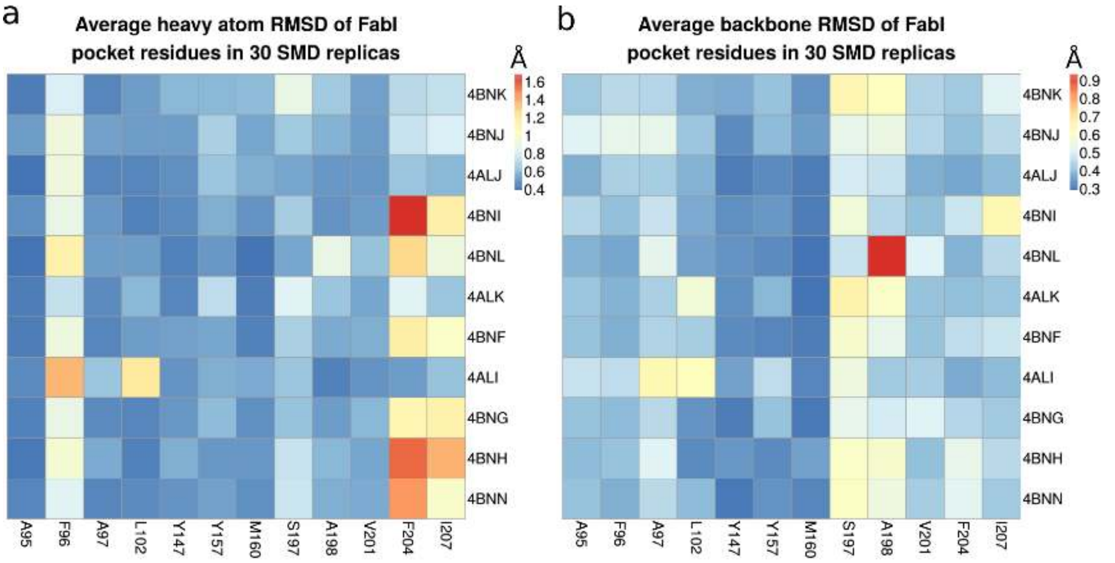


Figure 5.3 Heatmap of (a) average heavy atom RMSD and (b) of average backbone RMSD of binding pocket residues over 30 SMD replicas of all complexes. The color-scale indicates the RMSD value.

analysis of the SMD simulations reveals that the major gate keepers of ligand unbinding through the chosen exit pathway are Phe96, accompanied by the residues of the ascending branch of the SBL, i.e., helix $\eta6$ (cf. Figure 2.3).

It should be noted that the structural dynamics of the SMD simulations are influenced by the harmonic constraints applied on the C_α atoms of the protein backbone, avoiding translation of the protein-ligand complex, but also keeping the protein backbone conformation close to the starting structure. Nonetheless, the described residues open the major portal by displacing the Phe96 side chain and helix $\eta6$ of the SBL during ligand dissociation. Conversely, the SBL and a second SBL (residues 94–108, i.e., including Phe96) are assumed to show concerted closure upon inhibitor binding [58]. Also, in the homologous enzyme *M. tuberculosis* InhA, increased distance between helix $\alpha6$ and Phe97 (Phe96 in *sa*FabI) is assumed to be a conformational characteristic of the EI state of slow-onset inhibition [40], which could also be observed in unguided classical MD simulations (cf. Chapter 3, Figure 3.17). These observations provide evidence for the validity of the chosen dissociation pathway.

Since it contains the ligand of the examined series with the longest t_R (**PT119**), the complex 4BNN was examined exemplarily in detail. Figure 5.5 represents the RMSD evolution of the aforementioned key residues during the egress of **PT119** (Phe96, Ser197 and Ala198). Phe96 exhibits high deviation from the starting structure during the SMDs. In several trajectories the RMSD rises above 2.0 Å at around 0.5 ns of simulation. The backbone atoms of Ser197 and Ala198 behave in a similar way and reach their RMSD

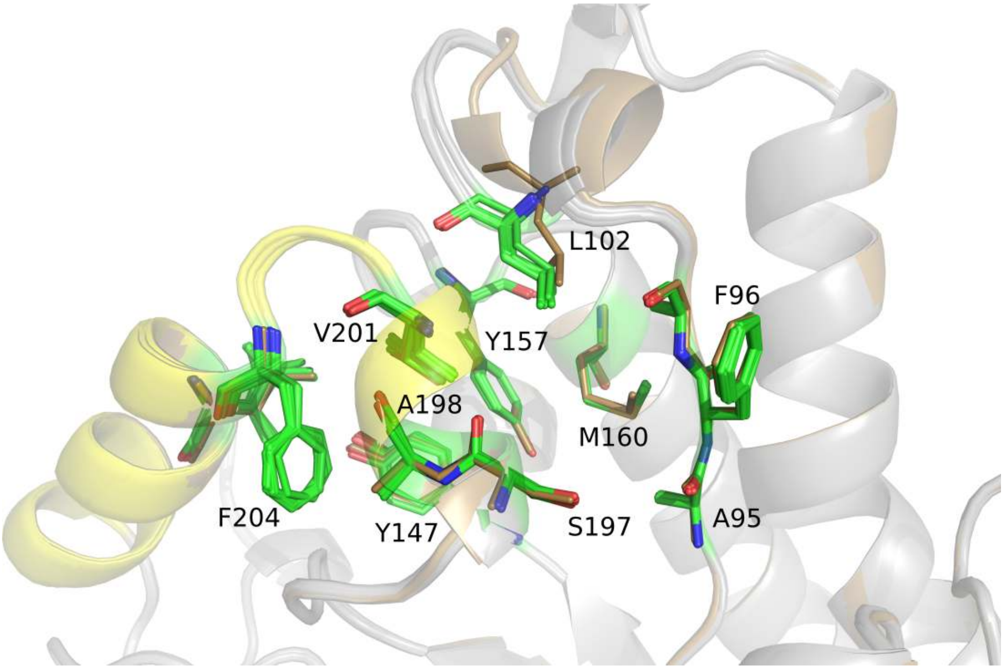


Figure 5.4 Binding pocket of aligned monomers in crystal structure conformation. Pocket residues are represented in green, the substrate binding loop is shown in yellow. The complex 4ALI, deviating in the region around pocket residue L102, is illustrated in light brown.

peak at around 0.4 ns. These peaks and the associated conformational changes of the residues are accompanied by a specific behavior of the ligand during extraction: after the strongest interactions (i.e., the hydrogen bonds of the phenolate to Tyr157 and the ribose-hydroxyl of NADP^{+}; cf. Figure 2.3) are broken at around 0.4 ns of simulation, the ligand travels further outside the binding pocket, displacing Phe96, Ser197 and Ala198 along the way. As soon as the scaffold leaves the core binding pocket at around 0.5 ns, the residues can relax to their original conformation (cf. Figure 5.5).

5.5 Free energy profiles

Using SMD simulations (protocol summarized in Appendix A, for details see Chapters 2.3.1.5 and 5.1) it was possible to determine free energy profiles for each complex. Since there are contradictory results in the literature on which free energy estimator is the most robust for limited sampling [36, 165], the exponential average as well as cumulant expansions up to the second order of the Jarzynski equality were evaluated (Figure 5.6). These quantities are hereinafter referred to as ΔG_e, ΔG_1 and ΔG_2, respectively. For free energy reconstruction the ensemble average of the performed work was calculated using the inner 50% of all available data points of $\langle W_{0\rightarrow\lambda}\rangle$ (i.e., the performed work depending on the position of the moving constraint λ_t). Hence, the

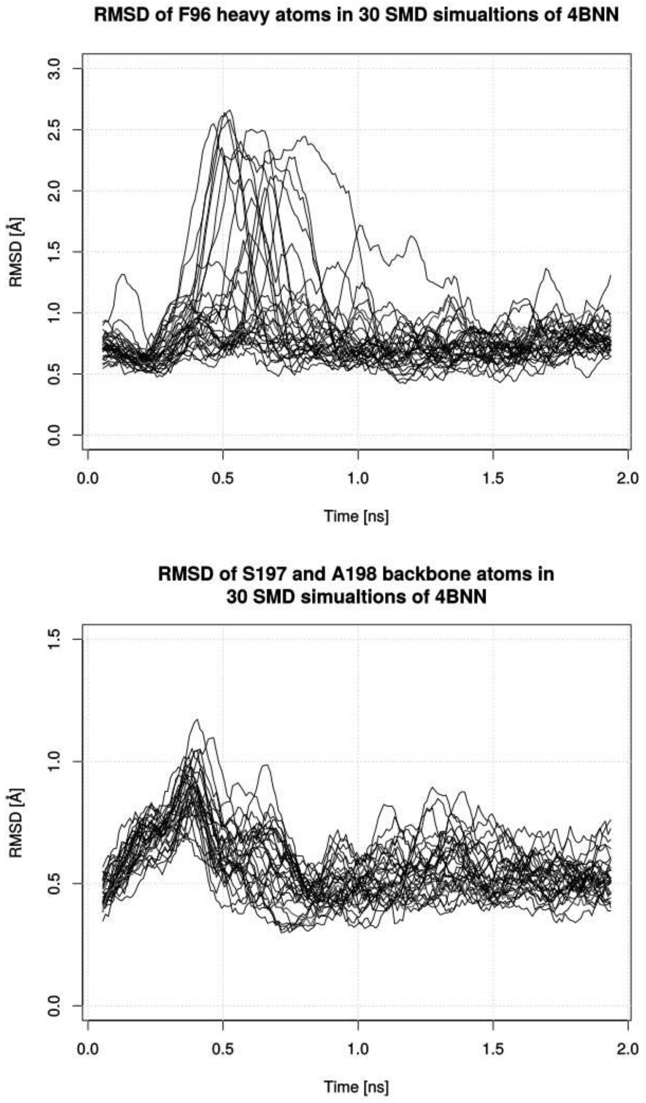


Figure 5.5 (a) RMSD of Phe96 heavy atoms in 30 SMD simulations of 4BNN. (b) RMSD of Ser197 and Ala198 backbone atoms in 30 SMD simulations of 4BNN. A moving average with a window of 20 ps was used.

ensemble average $\langle \cdot \rangle$ was trimmed by 25% on both sides and outlying values of $\langle W_{0\rightarrow\lambda} \rangle$ are dismissed, unless stated otherwise. The dismissal of outlying work profiles is necessary due to a large variance in the work profile ensembles, which would otherwise limit the applicability of Jarzynski's equality (discussed in detail in Chapter 5.7) [35].

All complexes show similar trends, although the curves exhibit different slopes. Generally, the steepest ascent occurs at the beginning until around 0.5 ns, i.e., after 5 Å traveled distance of the moving SMD constraint. At this point the virtual spring linking the ligand to the moving constraint is stretched at the maximum and the core scaffold of the ligand is forced to leave its crystal pose binding mode. From this point on the

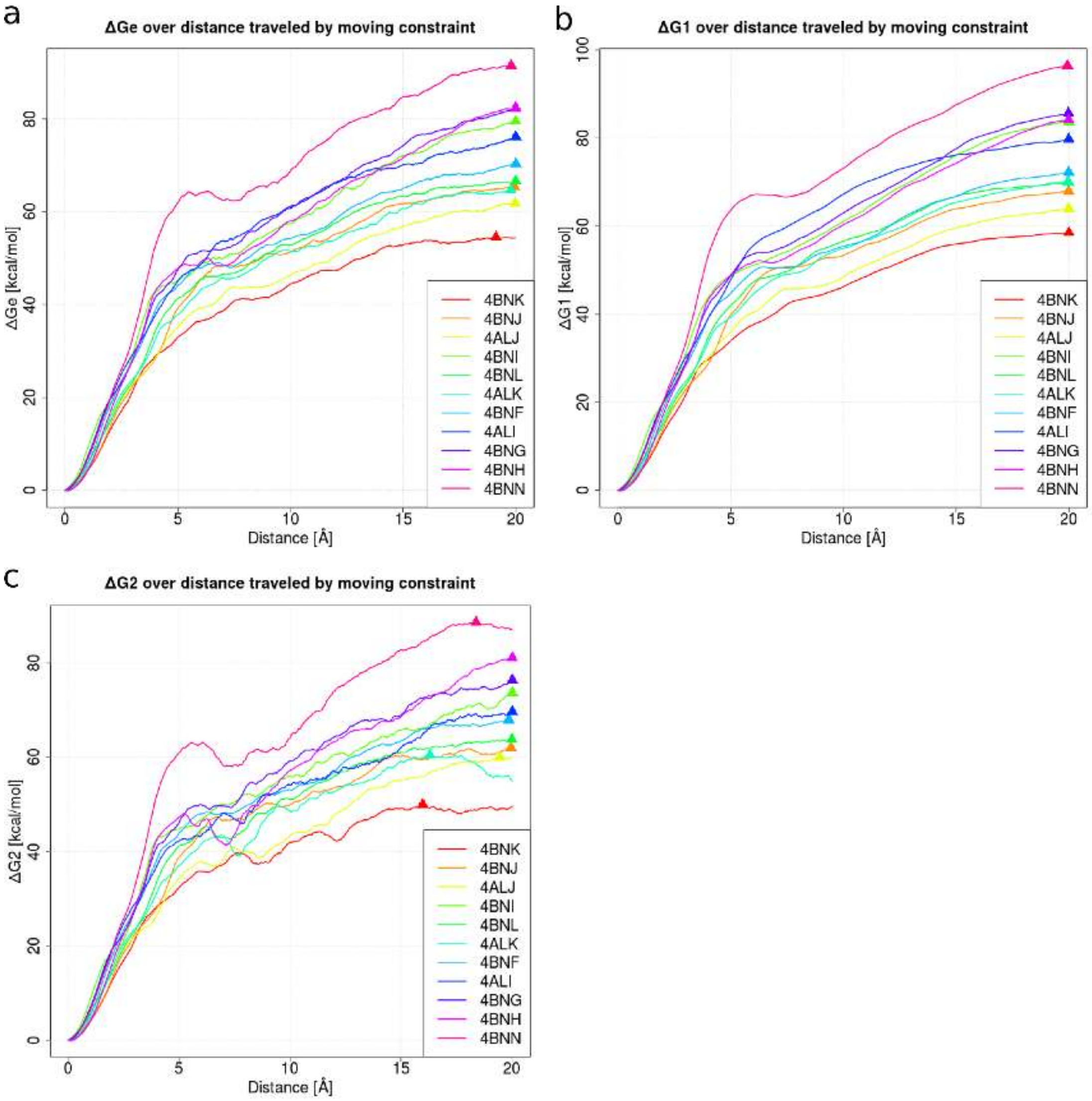


Figure 5.6 Calculated free energy profiles of all complexes. **(a)** Exponential average ΔG_e of Jarzynski equality, **(b)** first cumulant expansion ΔG_1 and **(c)** second cumulant expansion ΔG_2, respectively, over distance traveled by the moving SMD constraint (dummy atom). Triangles illustrate the maximum free energy change for each complex.

moving constraint drags the ligand further outside the binding pocket through the FabI major portal, resulting in different values of free energy change along the reaction coordinate for each ligand, which depends on the respective protein-ligand interactions and the conformational changes that are induced. Of most interest is the maximum free energy change upon induced extraction of each ligand from its pocket, since k_{off} mostly depends on the height of the highest free energy barrier [26]. Maximum values and standard deviations at the respective times of measurement are summarized in Table 5.2.

For all complexes and using each Jarzynski estimator, the maximum free energy changes

Table 5.2 Maximum free energy change of complexes and standard deviations in kcal/mol. ΔG_e is the maximum free energy change reconstructed using the exponential Jarzynski estimator, ΔG_1 and ΔG_2 are calculated using the first and second order cumulant expansion, respectively. SD are the corresponding standard deviations. Values are reconstructed from 25% trimmed averages of work profiles. Rows are sorted by experimental t_R (cf. Table 5.1).

PDB code	**ligand**	ΔG_e	ΔG_1	ΔG_2	SD_e	SD_1	SD_2
4BNK	PT55	54.20	61.00	47.20	4.16	4.64	3.45
4BNJ	PT53	64.31	68.80	58.76	3.57	3.68	3.42
4ALJ	PT52	61.77	64.11	59.76	2.28	2.28	2.27
4BNI	PT13	66.82	71.43	59.21	3.89	4.24	3.73
4BNL	PT68	77.50	84.21	70.01	4.44	4.47	3.77
4ALK	PT01	65.74	70.61	61.10	3.44	3.63	3.34
4BNF	PT02	68.46	71.48	66.71	2.41	2.47	2.38
4ALI	TCL	78.15	80.59	73.56	2.89	3.02	2.89
4BNG	PT03	80.99	84.42	75.85	3.30	3.46	3.14
4BNH	PT04	81.74	84.26	79.03	2.50	2.51	2.50
4BNN	PT119	89.13	97.02	76.73	5.51	5.51	4.37

used for model building were all extracted from the last quarter of the simulations along the reaction coordinate (i.e., >15 Å traveled distance of the moving constraint, cf. triangles in Figure 5.6). Interestingly, earlier local free energy maxima (i.e., maxima obtained from a reduced pulling distance with $\lambda < 15$ Å) do not correlate well with the experimental t_R (cf. Chapter 5.6), although the core scaffold of the ligand has already left the binding pocket after around 0.5 ns (5 Å traveled distance). The different slopes of the free energy curves of the ligands in the last quarter of the simulation time naturally originate from different residual interactions with the protein. Seemingly, the simulation period with 15 Å $\leq \lambda \leq$ 20 Å is as important for accurate model building as the initial induced removal from the core binding pocket. This might be evidence that some ligands have long residence times not only due to a large energy barrier for initial ligand unbinding, but also due to stronger interactions with the protein outside the core binding pocket, diminishing the probability of complete ligand dissociation and, thus, increasing the probability of ligand rebinding [166].

In most cases, the maximum free energy change of ligand extraction is the very last value of the free energy profile ($\lambda = 20$ Å). The PMF profiles are, hence, continuously rising and may not reach the maximum for $\lambda < 20$ Å. Given the chosen reaction coordinate, this is, however, not unexpected, since a further extraction of the ligand from the protein is generally associated with additional work (cf. continuously rising average work estimator, Figure 5.6b). The average work curves reach plateaus towards the end of the simulation, indicating that no further contributions from protein-ligand interactions can be expected (cf. Figure 5.6b). Since the second order cumulant expansion subtracts the variance term from the contribution of work (cf. Chapter 2.3.3.2), the free energy curves of ΔG_2 can rapidly decline when the fluctuations of the work profiles become too high, which may lead to rather artificial values of ΔG_2 after the maximum value. Furthermore, the exponential average ΔG_e is largely dominated by trajectories with small work values due to its exponential nature [35]. Thus, a slight drop in the work profiles of low-work trajectories may lead to a decrease of ΔG_e. For these reasons, maxima of ΔG_2 and ΔG_e can occur for values of $\lambda < 20$ Å.

It should be noted here that the SMD reaction coordinate (i.e., the pulling direction) does not necessarily represent the natural dissociation pathway of EI* to E+I on the free energy hypersurface (including the associated conformational changes of the enzyme), in which the end-point E+I is energetically lower than the transition state EI‡. The one-dimensional and potentially unnatural SMD exit pathway might, thus, lead to artificially elevated absolute values of ΔG compared to the natural energy barriers of ligand dissociation. However, since the same reaction coordinate is applied on all complexes, relative evaluation of ΔG is possible. The harmonic constraints on the C$_\alpha$-atoms of the protein backbone are expected to further contribute to the high absolute values of ΔG.

5.6 Linear models of residence time

In the next step, linear regression models of the maximum free energy change of a ligand and its experimental residence time were generated. For this purpose the average of all available t_R values per complex (summarized in Table 5.1) was used to avoid dismissal of experimental data for model generation. First, the free energy change was calculated according to the exponential average of the Jarzynski equality ΔG_e and used to generate a linear model to the residence time. In agreement with the equations of Eyring's transition state theory, the free energy change was related to $ln(t_R)$ [84, 85]. Regression analysis indicates a linearity of these two measures with an R^2 of 0.8283 (Figure 5.7a), yielding a very good correlation between the drug-target residence time and the simulated free energy change:

$$ln(t_R[min] \cdot min^{-1}) = 0.1228[mol\ kcal^{-1}] \cdot \Delta G_e[kcal\ mol^{-1}] - 4.5170 \tag{5.1}$$

The model shows a Pearson correlation coefficient of $r = 0.9101$ and significant values for slope ($p = 0.000101$) and intercept ($p = 0.008956$). The maximum Cook's distance of the model is 0.5163, which can be measured for the first observation (4BNK) [167]. Hence, no observation has a Cook's distance above 1, i.e., the model is unlikely to contain outliers [168].

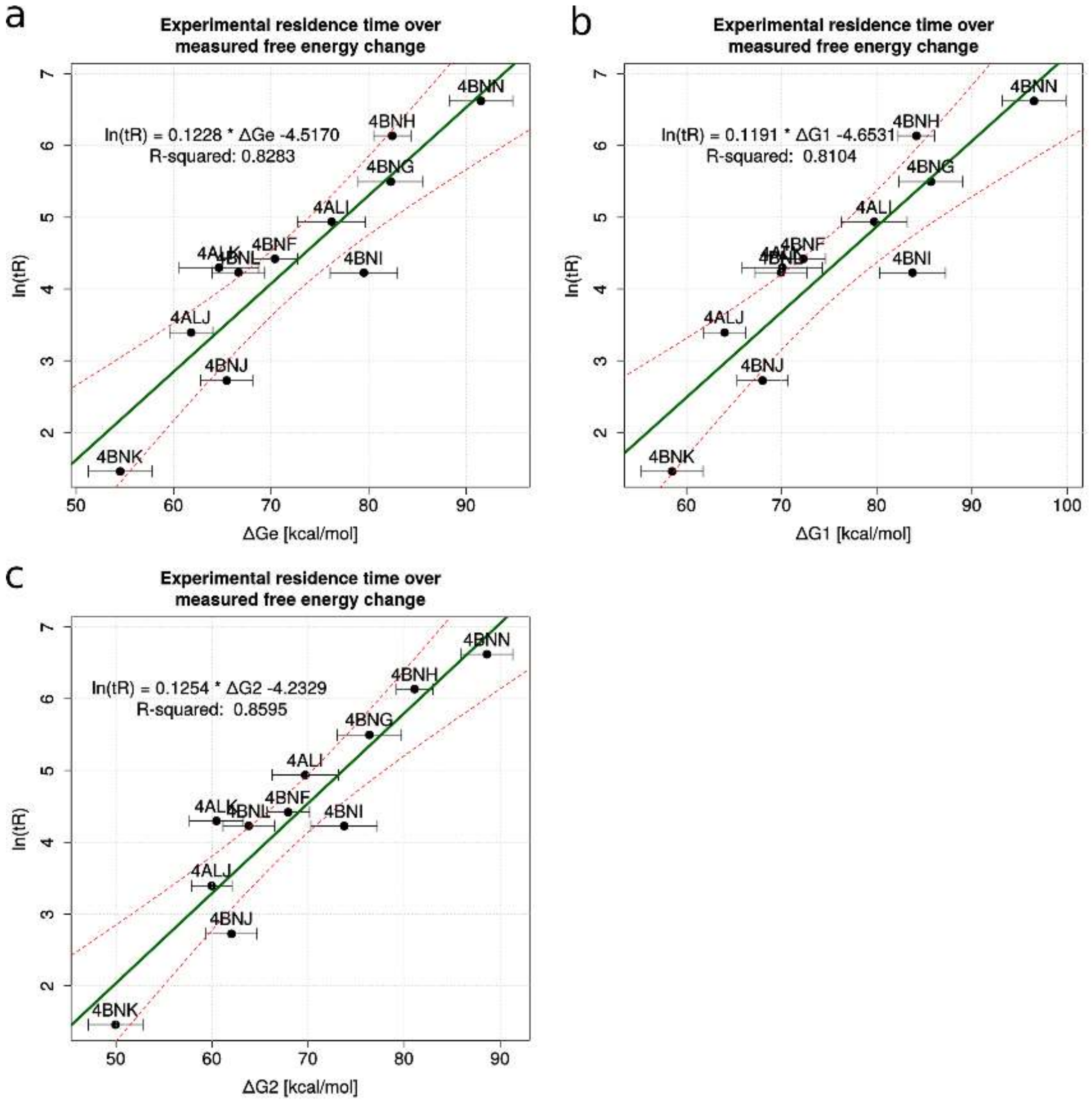


Figure 5.7 Linear models of the maximum free energy change of a ligand upon induced extraction and $ln(t_R)$. (A) $ln(t_R)$ against ΔG_e. The standard deviation of energy calculations is depicted as horizontal error bars. The linear model is illustrated as a green line, accompanied by the 95% confidence interval of the model as dashed red lines. **(B)** $ln(t_R)$ against ΔG_1. **(C)** $ln(t_R)$ against ΔG_2.

Moreover, the maximum free energy change based on the first and second cumulant expansion of the Jarzynski equality, respectively, was used to generate a linear model (Figure 5.7b and c). The ΔG_1 model yields an R^2 of 0.8104, which is below the model using ΔG_e (Figure 5.7b). The combination of high residual and leverage increases the Cook's distance of the first observation (4BNK) to 0.5650, indicating that this model is more influenced by a single observation than the ΔG_e model.

The use of the second order cumulant ΔG_2 outperforms both preceding models (Figure 5.7c):

$$ln(t_R[min] \cdot min^{-1}) = 0.1254[mol\ kcal^{-1}] \cdot \Delta G_2[kcal\ mol^{-1}] - 4.2329 \qquad (5.2)$$

with a Pearson correlation coefficient of $r = 0.9271$ and significant values for slope ($p = 4.02 \cdot 10^{-5}$) and intercept ($p = 0.00562$). The calculated R^2 of this model (0.8595) is the highest observed. In addition, the Cook's distance of 4BNK drops below the threshold of 0.5 with a value of 0.4669, underlining that this model is not overly influenced by one single observation.

Regarding single observations within the models, a few features are worth noting: the complex 4ALI with its small ligand triclosan (**TCL**) bearing a chlorine in R_1-position exhibits similar maximum free energy changes ΔG_e, ΔG_1 and ΔG_2 as the complex 4BNI bound to the much larger ligand **PT13**, which possesses a R_1-hexyl-residue occupying the hydrophobic pocket of FabI. 4ALI also shows a similar maximum free energy change as 4BNF bound to **PT02** with a propyl-residue in R_1-position in all Jarzynski estimators; the most significant difference between these ligands can be observed within the ΔG_e and the ΔG_1 model. Both examples are evidence that the presented methodology does not simply correlate to ligand size. Although, in general, longer ligands need longer pulling distances, and thus more work, to leave the pocket entirely, the SMD approach in the applied force field evidently captures more sophisticated contributions to ligand binding than mere ligand size. Furthermore, the ligands of complexes 4BNK and 4ALJ, which differ solely in a fluorine/chlorine substitution at R_1-position are recognized in the correct order in all three calculated Jarzynski estimators, highlighting that even small differences in the ligands can be evaluated properly with the used method.

The exponential Jarzynski expression can in general be difficult to estimate due to insufficient sampling and too high pulling velocities [35]. Indeed, the presented models suggest that the second order cumulant expansion yields a more accurate linear model to $ln(t_R)$ compared to the exponential average and first cumulant expansion. However, based on Clarke's test for non-nested model selection, none of the three models is significantly closer to the true model and, thus, preferred over the other models ($p > 0.05$) [134, 169]. Furthermore, the slopes of the three models are not significantly different, based on Chow's test for heterogeneity in two regressions ($p > 0.05$ for ΔG_1 or ΔG_2 vs. ΔG_e; $p > 0.01$ for ΔG_1 vs. ΔG_2) [135, 170].

The choice of the maximum ΔG of each profile for model generation over the free energy change at the simulation end-point ($\Delta G_{\lambda=20\text{Å}}$) has no significant influence on the model quality in the case of ΔG_e and ΔG_1 (R^2 of 0.8251 and 0.8101, respectively). Due to large fluctuations in the work profiles towards the end of the simulations, the regression model based on the second order cumulant expansion $\Delta G_{2;\lambda=20\text{Å}}$, on the other hand, shows a decline in R^2 (0.7907) compared to the model based on the maximum ΔG_2, validating the choice of the maximum ΔG values for model generation.

5.7 Fluctuations of work values

By using all available data of $\langle W_{0\rightarrow\lambda}\rangle$ for PMF reconstruction, i.e., an untrimmed ensemble average (Table 5.3), the model quality drops significantly (R^2 of ΔG_e model: 0.7589; R^2 of ΔG_1 model: 0.8097; Figure 5.8). Particularly the model using ΔG_2, which showed the highest predictive power with a 25% trimmed ensemble average, suffers a drastic decline of R^2 (0.3940). The reason for this loss of predictive power is the much higher variance of the measured work profiles of the complexes, which is directly incorporated into PMF reconstruction by the second order cumulant expansion of Jarzynski's equality and, thus, the corresponding maximum free energy change. The fluctuation of work values is an indicator for the applicability of Jarzynski's equality and should be comparable to the temperature k_bT [35]. Here, the standard deviations using untrimmed ensemble averages lie between 2.79 and 9.53 kcal/mol (cf. Table 5.3), which corresponds to 4.68 to 15.99 k_bT. By trimming the distributions of work values, the standard deviations can be reduced to values between 3.20 and 7.08 k_bT, which is very comparable to the values obtained by Park *et al.* (2003), who addressed the issue of work value fluctuations for different pulling speeds [35]. Since the standard deviation of the untrimmed work values is high compared to the temperature, the Jarzynski equality might not be fully applicable for this system without robust measures of average and variation (i.e., trimmed average and standard deviation) using the current simulation setup. Particularly the second order cumulant expansion is affected by the high fluctuations and, thus, rendered inapplicable. It can be expected that a higher number of replica simulations would decrease the fluctuations of the work profiles within one complex, reducing the sampling error [36]. However, enhanced sampling is primarily hindered by limited computational power.

It is known from Jarzynski's equality and Jensen's inequality[1] that $\Delta G \leq \langle W\rangle$ (with $\langle W\rangle = \Delta G_1$) [36], i.e., ΔG_1 must be the upper limit of the true maximum free energy change. Moreover, the mean work estimator ΔG_1 should usually not be used due to unrobust results [165]. Nevertheless, ΔG_1 yields the most accurate correlation to $ln(t_R)$ with untrimmed work profiles in terms of R^2, although the ΔG_1 model is again not significantly preferred over the ΔG_e model ($p > 0.05$, Clarke's test) [134, 169]. Hence, the untrimmed ΔG_e and ΔG_1 models show a basically unchanged performance compared

[1] $\langle e^x\rangle \geq e^{\langle x\rangle}$

Table 5.3 Maximum free energy change of complexes and standard deviations in kcal/mol without trimming.

PDB code	**ligand**	ΔG_e	ΔG_1	ΔG_2	SD_e	SD_1	SD_2
4BNK	PT55	48.85	58.62	28.07	6.44	6.87	5.69
4BNJ	PT53	58.49	68.07	40.01	5.80	5.82	5.76
4ALJ	PT52	52.74	64.60	25.82	7.71	7.76	3.85
4BNI	PT13	60.43	70.16	33.43	7.05	7.13	3.55
4BNL	PT68	69.59	84.09	27.36	9.49	9.53	4.40
4ALK	PT01	57.75	69.91	39.93	6.47	6.94	4.73
4BNF	PT02	63.84	72.56	47.52	6.15	6.20	5.01
4ALI	TCL	59.40	79.39	31.93	8.16	8.20	2.79
4BNG	PT03	69.64	85.26	36.19	7.98	8.15	3.39
4BNH	PT04	71.25	84.49	50.82	6.44	6.45	6.24
4BNN	PT119	80.67	97.16	50.84	9.02	9.29	3.98

to the corresponding models from trimmed values, underlining the robustness of these two free energy estimators, as opposed to ΔG_2. Due to the very large fluctuations in the untrimmed work profiles, many maxima of ΔG_2 are reached early (λ values between 4 Å and 6 Å), leading to highly artificial values for 15 Å $\leq \lambda \leq$ 20 Å.

5.8 Correlation to experimental K_i

Chang *et al.* (2013) reported a strong double logarithmic correlation for *S. aureus* FabI diphenylether inhibitors between k_{off} and the inhibition constant K_i ($r = 0.95$) [30]. Furthermore, in recent studies, the SMD method has been employed to describe a very good linear correlation of the binding affinity of influenza virus neuraminidase inhibitors to the maximum rupture force of ligand pulling experiments [171, 172], although in these studies the plain pulling force, i.e., the unintegrated measure of the pulling experiment, was evaluated, and not the PMF, i.e., the free energy as a function of the pathway. Thus, it had to be clarified whether the used SMD methodology simply reconstructs K_i values or binding affinities, since obviously much faster computational methods are available for this task. The exponential average and both cumulant expansions of Jarzynski's equality were, thus, correlated to the $ln(K_i)$ values of the examined ligands (Figure 5.9).

In each case, the K_i correlation model is outperformed by the respective t_R model. The exponential average reaches an R^2 of 0.7410, whereas the first and second order cumulants exhibit R^2 values of 0.7054 and 0.8096, respectively. Statistical quantities of linear model quality were calculated for the best pair of models (ΔG_2) and summarized in Table 5.4. In this context, the significance of slope and intercept in terms of the p-value, the Pearson and adjusted R^2, the residual standard error (RMSE), F-test

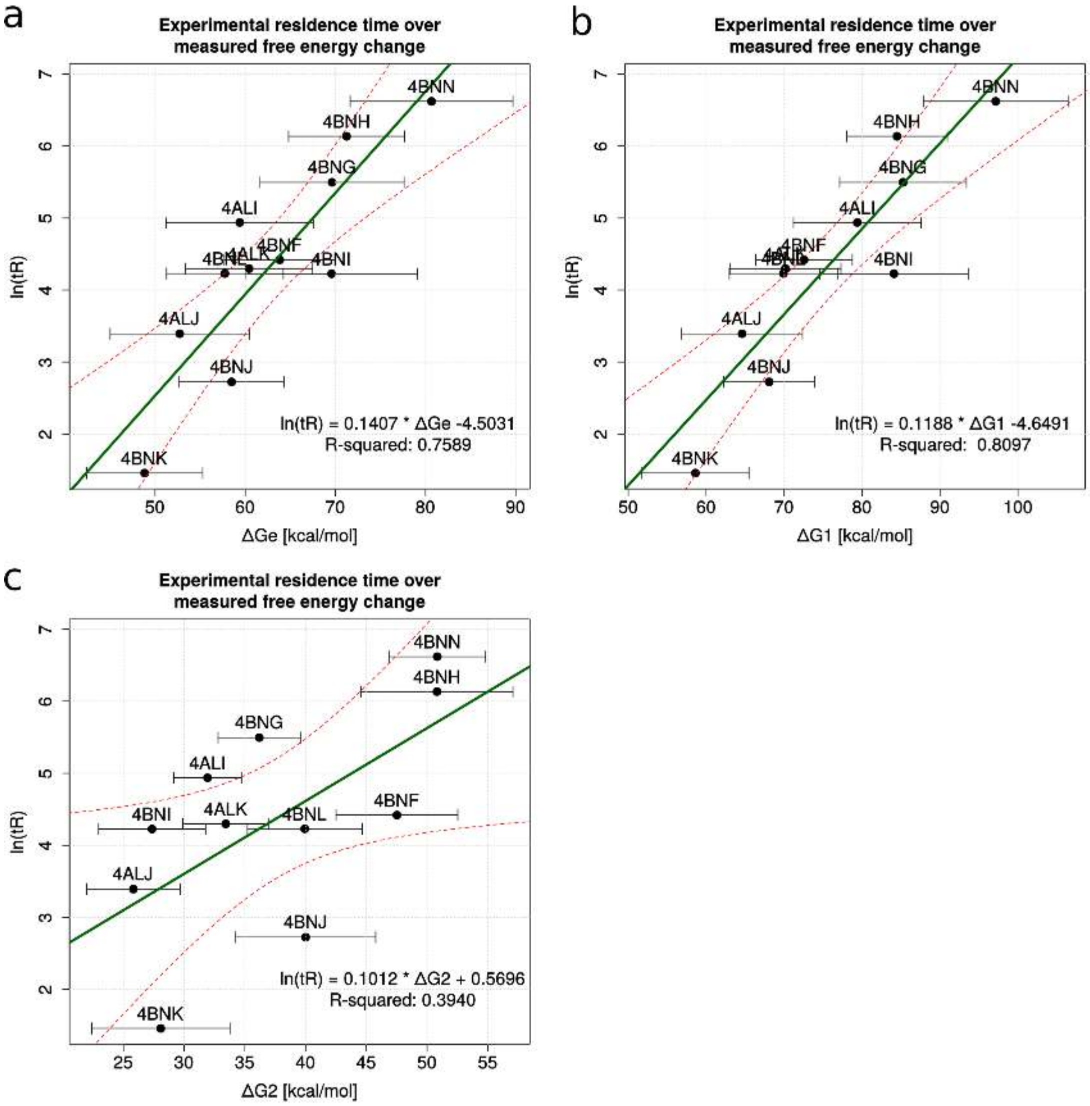


Figure 5.8 Linear models of the maximum free energy change without trimmed ensemble average $\langle W_{0\rightarrow\lambda}\rangle$ and $ln(t_R)$. Cf. Figure 5.7 for full explanation of the plots.

statistics and the first and second order Akaike information criterion (AIC and AICc) were evaluated [173, 174]. In each of the considered items, the t_R model is superior to the K_i model. The only exception is the intercept p-value, which is slightly higher in the t_R model.

For further investigation, two SMD experiments were set up for the homologous enoyl-ACP reductase *Mycobacterium tuberculosis* InhA. Two crystal structures (2X23 and 4OHU) with bound diphenylether inhibitors of comparable size and binding mode (**PT70** and **PT92**) were selected and prepared in the same manner [40, 45]. In contrast to *sa*FabI, the residence time of InhA inhibitors is generally not correlated to their affinity,

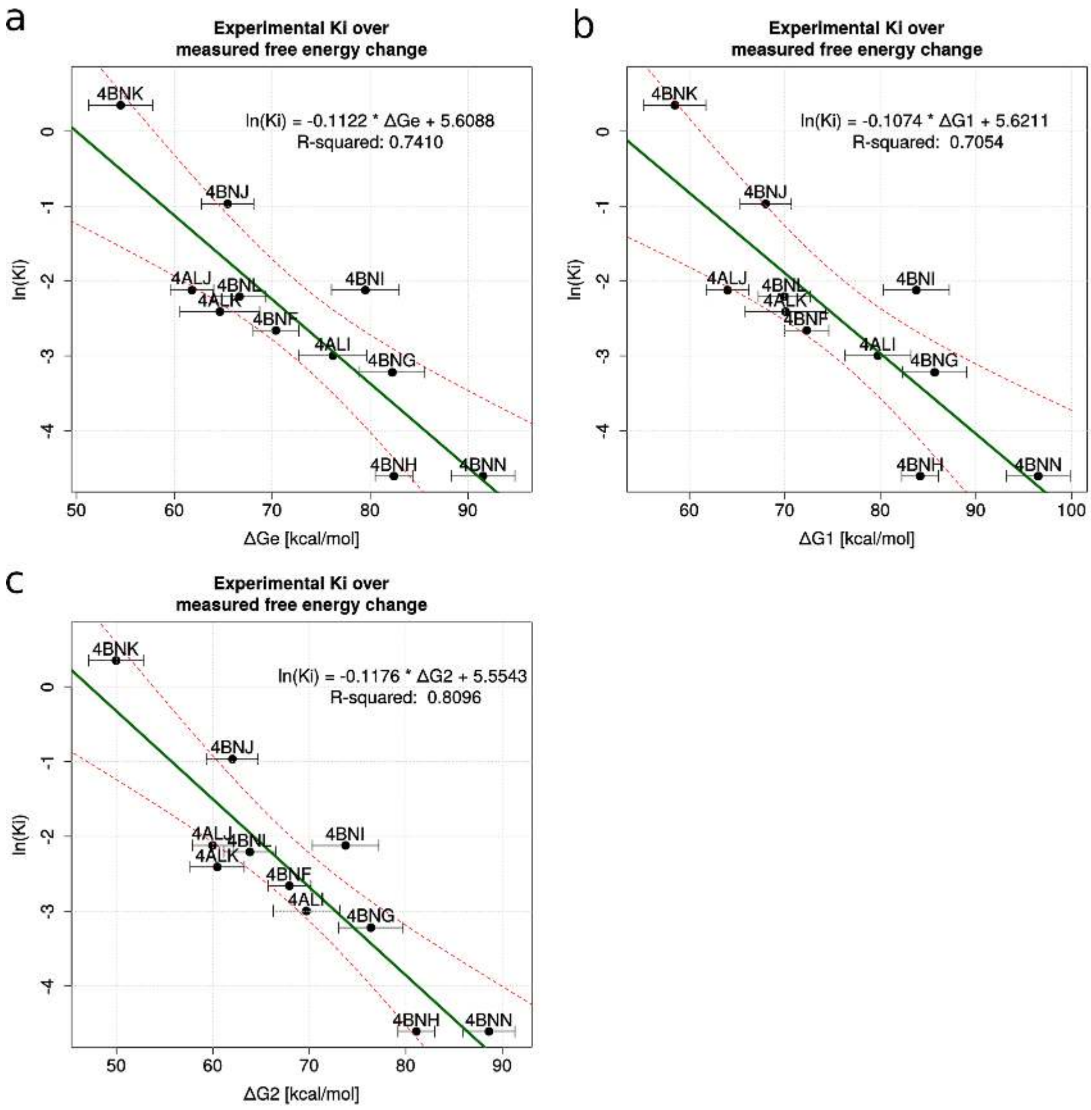


Figure 5.9 Linear models of maximum free energy changes of a ligand upon forceful extraction and its experimental $ln(K_i)$ value. Cf. Figure 5.7 for full explanation of the plots.

Table 5.4 Statistics summary of linear models of ΔG_2 to $ln(t_R)$ and $ln(K_i)$, respectively. Except for the intercept p-value, the $ln(t_R)$ model outperforms the corresponding $ln(K_i)$ model.

	$ln(t_R)$ **model**	$ln(K_i)$ **model**
Intercept p-value	0.005620	0.002256
Slope p-value	$4.02 \cdot 10^{-5}$	$1.60 \cdot 10^{-4}$
Pearson R^2	0.8595	0.8096
Adjusted R^2	0.8439	0.7885
RMSE [ln units]	0.5854	0.6586
F value	55.06	38.28
F-test p-value	$4.02 \cdot 10^{-5}$	$1.61 \cdot 10^{-4}$
AIC	23.23	25.82
AICc	26.66	29.25

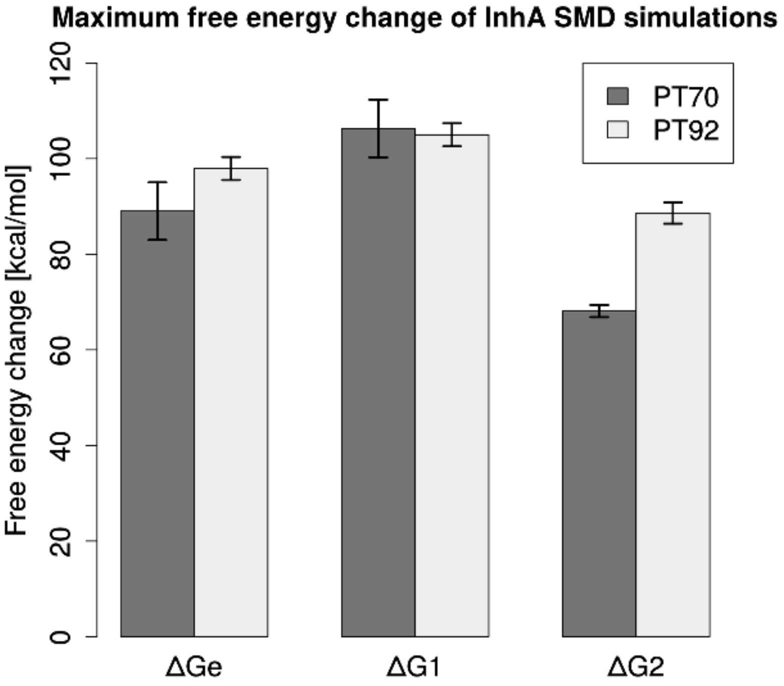


Figure 5.10 Maximum free energy changes of InhA SMD simulations with PT70 and PT92, respectively. 95% confidence intervals are indicated by error bars. The most significant difference can be observed for ΔG_2.

which is also represented by the chosen ligands. Whereas **PT70** (2-(*o*-Tolyloxy)-5-hexylphenol) is a high-affinity compound (K_i = 0.044 nM) with a residence time of 24 minutes, **PT92** (2-(2-Bromophenoxy)-5-hexylphenol) shows a comparable residence time of 30 minutes at a 4.5-fold increased K_i of 0.20 nM [46]. The outlined protocol was applied to both InhA complexes and the maximum free energy changes were determined according to all Jarzynski estimators (Figure 5.10).

A suitable egress pathway was determined by RAMD simulation of the complex 2X23. Again the exit route via the major portal was chosen, according to the substrate delivery route of InhA [53]. Although **PT70** exhibits a higher affinity than **PT92**, the calculated maximum free energy change is equivalent (ΔG_1 and ΔG_e), which is well in line with the experimental t_R values of these complexes. This underlines the hypothesis that the presented SMD approach is rather suited to reconstruct and predict residence times than to calculate binding affinities.

5.9 Parameter selection and optimization

Although SMD simulations are relatively straightforward in defining a suitable reaction coordinate, several other parameters need to be specified as well, namely the number of replica simulations, pulling speed, pulling distance, pulling direction, simulation time, harmonic constraints on the protein, force constant of SMD spring and choice of starting structures. However, modified parameters must be validated by computationally very expensive simulations, limiting the possibilities for a systematic study of parameter combinations. Therefore, only selected parameters were optimized in preliminary simulations. The overall quality criterion for parameter selection during study design was the coefficient of determination R^2 of the maximum free energy change to $ln(t_R)$ of the resulting linear models (data not shown) (cf. Chapter 5.6).

The number of replica simulations was increased from initially 10 to 30 with the tradeoff of increasing the pulling speed from 1 Å/ns to 10 Å/ns, which is a commonly used velocity for irreversible pulling [35, 36]. The higher velocity reduces the simulation time from 20 ns to 2 ns. Although Park and Schulten (2004) stated that fewer trajectories with slower speed yield more accurate results, they based this conclusion on increasing the velocity from 10 to 100 Å/ns, which is 10-fold above the velocity used here and 1000-fold above reversible pulling [36].

The SMD starting structures were initially selected from a snapshot after 5 ns of classical MD simulation for each complex. To reduce the large diversities of the binding pocket and ligand conformations, new simulations were set up with starting structures taken directly after the 1 ns equilibration protocol.

An extensive RAMD approach using multiple simulations and the use of an "average exit pathway" might further improve the validity of the applied pulling direction. The accuracy of the free energy declines with the pulling distance, because the sampling error increases [35, 36]. Hence, a shorter exit pathway might reduce fluctuations of the work profiles and, thus, increase the accuracy of the free energy profiles. In SMD simulations, harmonic constraints must be applied on the protein to avoid translation of the entire complex during ligand extraction. In this study, mild harmonic constraints were assigned to all C_α atoms of the protein backbone, limiting the mobility of otherwise very flexible regions of the enzyme, particularly the SBL and the binding pocket. Hence, future steps of parameter optimization might include the more careful selection of harmonically constraint protein atoms, i.e., keeping the aforementioned key regions of the protein flexible. The choice of a proper SMD force constant is a minor issue, provided the stiff-spring approximation is fulfilled [35]. Park *et al.* (2003) have shown that PMF profiles obtained from SMD simulations with a force constant of 500 pN/Å and 35000 pN/Å, respectively, at a speed of 10 Å/ns essentially show the same results. However, the PMF obtained at a high force constant of 35000 pN/Å shows larger fluctuations [35]. Thus, the large used force constant of 10000 pN/Å might also contribute to the high fluctuations of the work values of the *sa*FabI systems, which make the use of a trimmed ensemble average necessary. A carefully lowered SMD force constant might reduce fluctuations and, thus, further improve the resulting correlation between the experimental residence time and the simulated maximum free energy change. Finally, the application of a finite-sampling correction [35] or extension of Jarzynski's equality by weighted histograms ("Hummer and Szabo-method") [175, 176] might help to further improve the data evaluation.

5.10 Conclusion

Using SMD simulations, it was possible to extract the maximum free energies of ligand dissociation from free energy profiles as a function of a unified reaction coordinate for a series of diphenylethers from *Staphylococcus aureus* FabI. Accurate linear regression models could be generated of the calculated maximum free energy change to the experimentally determined residence time of the respective ligands in the target, which can be used to predict residence times of novel inhibitors from crystal structures or docking poses, representing a to the author's knowledge unprecedented approach. RMSD analyses of all SMD replica simulations revealed Phe96, Ser197 and Ala198 as gate keepers of ligand dissociation, which are forced to adopt new conformations upon unbinding. Conversely, the displacement of these residues is in agreement with the proposed binding mechanism of induced-fit ligand binding to *sa*FabI, in which the SBL and a second SBL (containing Phe96) are assumed to exhibit concerted closure [58], validating the major portal as a suitable exit pathway.

Although the presented FabI residence time models are very accurate, an extensive parameter study regarding the effects of, e.g., the force constants of harmonic constraints, equilibration time, number of replica simulations, different exit pathways, dissociation through the minor portal of FabI, etc., might further improve the outcome. However, each attempted parameter optimization goes along with very high computational costs.

The outlined protocol is thought to serve as a how-to for general application on any desired target and inhibitor series. Hence, the presented SMD simulations on *Mycobacterium tuberculosis* InhA bound to **PT70** and **PT92** could be extended by additional InhA-ligand-complexes, when new kinetic data is being published. In doing so, the work-flow could be used on this target to derive an InhA residence time model.

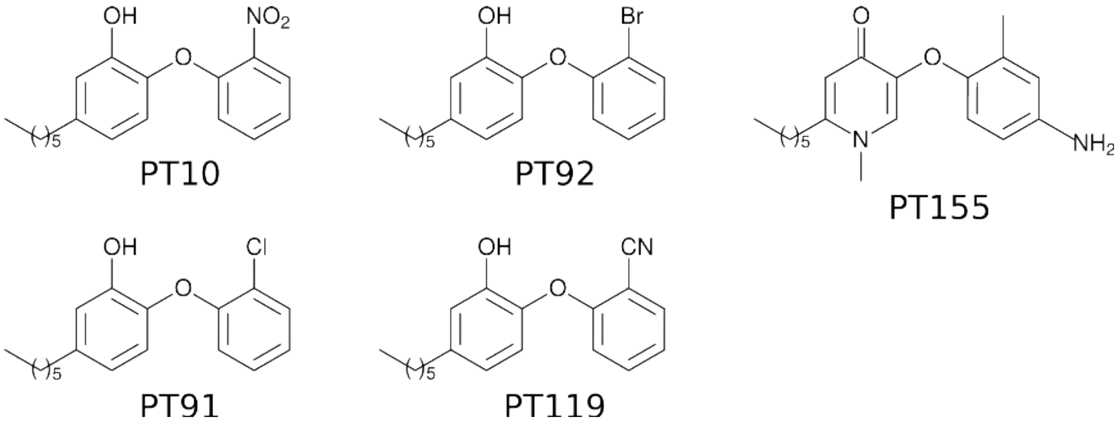
Chapter 6

The EI and EI* states of InhA inhibition revisited

According to current literature, ligand binding of slow-onset diphenylether inhibitors to InhA follows a multistep mechanism (cf. Figure 2.5) [17, 24, 45]. In the process of inhibitor association, the otherwise very flexible SBL is ordered and, hence, generally resolved in crystal structures containing a slow-onset inhibitor (e.g. **PT70** in structure 2X23 [45]). In contrast, this part of the enzyme is missing in crystal structures with rapid reversible diarylether inhibitors, most likely due to high mobility (cf. Chapter 3) [1]. In this context, it is assumed that the 2X23 (**PT70**) crystal structure represents the EI* state [40, 45, 46].

In recent studies several crystal structures of ternary InhA complexes bound to new slow-onset inhibitors and one rapid reversible inhibitor of the diarylether class have been published (Figure 6.1 and Table 6.1) [40, 46]. Interestingly, the binding pocket and SBL exhibit highly divergent conformations among different systems and also in different chains of the same system.

The following chapter revisits the concept of the EI and EI* states of ligand association in the case of InhA by analysis of different recent crystal structures and extensive MD and SMD simulations.



PT10

PT92

PT155

PT91

PT119

Figure 6.1 Structures of new published InhA diarylether inhibitors. PT155 is a rapid reversible ligand, whereas the remaining structures show slow-onset kinetics in InhA.

Table 6.1 Summary of crystal structures used in this analysis. Kinetic data is taken from the respective publication. Symmetry mates were created with PyMOL [142]. Sym. = Symmetry; Res. = Resolved; rev. = reversible.

PDB	Chain	Ligand	Space group	Resolution	Sym. contact	Res. SBL	t_R	Ref.
2X23	A	PT70	$P12_11$	1.81 Å	yes	yes	24 min	[45]
2X23	B	PT70	$P12_11$	1.81 Å	no	yes	24 min	[45]
2X23	E	PT70	$P12_11$	1.81 Å	yes	yes	24 min	[45]
2X23	G	PT70	$P12_11$	1.81 Å	no	yes	24 min	[45]
4OIM	—	PT119	$I4_122$	1.85 Å	yes	yes	80 min	[46]
4OXK	A	PT155	$P2_12_12_1$	1.84 Å	yes	yes	rapid rev.	[40]
4OXK	B	PT155	$P2_12_12_1$	1.84 Å	yes	no	rapid rev.	[40]
4OXK	C	PT155	$P2_12_12_1$	1.84 Å	yes	yes	rapid rev.	[40]
4OXK	D	PT155	$P2_12_12_1$	1.84 Å	yes	yes	rapid rev.	[40]
4OXN	A	PT155	$I2_12_12_1$	2.29 Å	yes	no	rapid rev.	[40]
4OXN	B	PT155	$I2_12_12_1$	2.29 Å	yes	yes	rapid rev.	[40]
4OHU	A	PT92	$P2_12_12_1$	1.60 Å	yes	yes	30 min	[40]
4OHU	B	PT92	$P2_12_12_1$	1.60 Å	yes	yes	30 min	[40]
4OHU	C	PT92	$P2_12_12_1$	1.60 Å	no	nearly	30 min	[40]
4OHU	D	PT92	$P2_12_12_1$	1.60 Å	no	no	30 min	[40]
4OXY	A	PT10	$P2_12_12_1$	2.35 Å	yes	yes	27 min	[40]
4OXY	B	PT10	$P2_12_12_1$	2.35 Å	yes	yes	27 min	[40]
4OXY	C	PT10	$P2_12_12_1$	2.35 Å	no	no	27 min	[40]
4OXY	D	PT10	$P2_12_12_1$	2.35 Å	no	no	27 min	[40]
4OYR	A	PT91	$P2_12_12_1$	2.30 Å	yes	yes	21 min	[40]
4OYR	B	PT91	$P2_12_12_1$	2.30 Å	yes	yes	21 min	[40]
4OYR	C	PT91	$P2_12_12_1$	2.30 Å	no	yes	21 min	[40]
4OYR	D	PT91	$P2_12_12_1$	2.30 Å	no	no	21 min	[40]
2AQ8	—	—	$P6_222$	1.92 Å	yes	yes	—	[56]

6.1 Comparison of experimental InhA crystal structures

All chains of the crystal structures 2X23, 4OIM, 4OXK, 4OXN, 4OHU, 4OXY, 4OYR, 2AQ8 [40, 45, 46, 56] as well as the conformational Family medoids described in Chapter 3 were aligned to chain A of the 2X23 structure with respect to their C_α atoms using PyMOL for binding pocket and SBL RMSD calculations [142].

As shown in Figure 6.2a, a clustering of heavy atom RMSD values of the binding pocket clearly distinguishes two clusters of binding pocket conformations. The first cluster with very small internal differences solely consists of chains binding to slow-onset inhibitors and the Family 1 medoid. The medoid snapshots of Families 2, 4 and 5 medoids are the closest outgroups to this cluster. Given the low inner-cluster RMSD and a visual inspection of the binding pockets, the crystal structures in this cluster exclusively exhibit a **PT70**-like binding of the inhibitor and, thus, an EI* conformation.

The second cluster consists of the Family 3 medoids, accompanied by all six chains of InhA bound to the rapid reversible **PT155** and the binary InhA-**NAD**$^+$ structure 2AQ8. Furthermore, two chains of InhA with the slow-onset inhibitors **PT10** and

PT92, respectively, are present in this cluster, as well as the system 4OIM, bound to the inhibitor **PT119**, which has a drug-target residence time of 80 minutes [46]. The occurrence of slow-binders in this cluster leads to the hypothesis that these chains might not represent a binding pocket conformation in the EI* state, despite the presence of a slow-onset inhibitor.

An analogous clustering of the flexible SBL does not result in equally sharply separated clusters. However, the same chains containing slow-binders, which clustered before by means of pocket RMSD, can now be found in two very similar clusters as well (Figure 6.2b, indicated by green rectangles). Again, 4OIM (**PT119**) and one chain bound to **PT10** and **PT92**, respectively, are clustered with chains containing the rapid reversible **PT155** or no ligand at all (2AQ8) (Figure 6.2b, indicated by cyan rectangle).

The conformational space within the cluster containing the rapid reversible **PT155**, 4OXY-A (**PT10**), 4OHU-B (**PT92**) and 4OIM (**PT119**) is much more diverse, compared to the cluster of slow-onset inhibitors. In particular, the pocket residues located in the SBL adopt various conformations throughout the different chains (Figure 6.3). Whereas the EI* state represented by 2X23-A (**PT70**) shows van der Waals contacts between Ile202/Val203 and the diphenylether B-ring [45], helix $\alpha6$ is twisted and relocated in these chains, leading to different conformations in this region. In three of four **PT155**-bound monomers of crystal structure 4OXK, Met199 occupies the space of Val203 in the 2X23 structure and forms contacts with the ligand B-ring (Figure 6.3a). Chain B of complex 4OXK is captured in a different conformation, in which Ile202 occupies the space of Val203 in the 2X23 structure and interacts with the B-ring (Figure 6.3b). This rather EI-like conformation is also very similar to those of 4OXY-A (**PT10**), 4OHU-B (**PT92**) and 4OIM (**PT119**), i.e. the InhA monomers bound to slow-onset inhibitors, which might not exhibit the EI* state. The subunit 4OXN-B (**PT155**) and the ligand-free structure 2AQ8 show a conformation similar to the aforementioned chains, however with helix $\alpha6$ further opened, i.e. Ile202 is shifted farther away from the ligand (Figure 6.3c). Indeed, Li *et al.* (2014) [40] interpreted this conformation as the actual EI state of InhA inhibition by slow-onset diphenylethers and used it in Partial Nudged Elastic Band (PNEB) MD simulations and Umbrella Sampling (US).

Furthermore, Pan *et al.* (2014) [46] state that the crystal structure 4OIM might not represent the final EI* state, but a snapshot along the reaction coordinate of ligand binding, due to the crystallization conditions and crystal packing. In fact, the crystallization conditions of the ternary InhA-**NAD$^+$-PT119** complex (PDB 4OIM [46]) differ significantly from the remaining complexes (cf. Chapter 3). A very high acetate concentration of 2.4 *M* at pH 5.0 was used for **PT119** crystallization. As a result, two acetate ions occupy positions in the major portal of the binding pocket (Figure 6.4). One

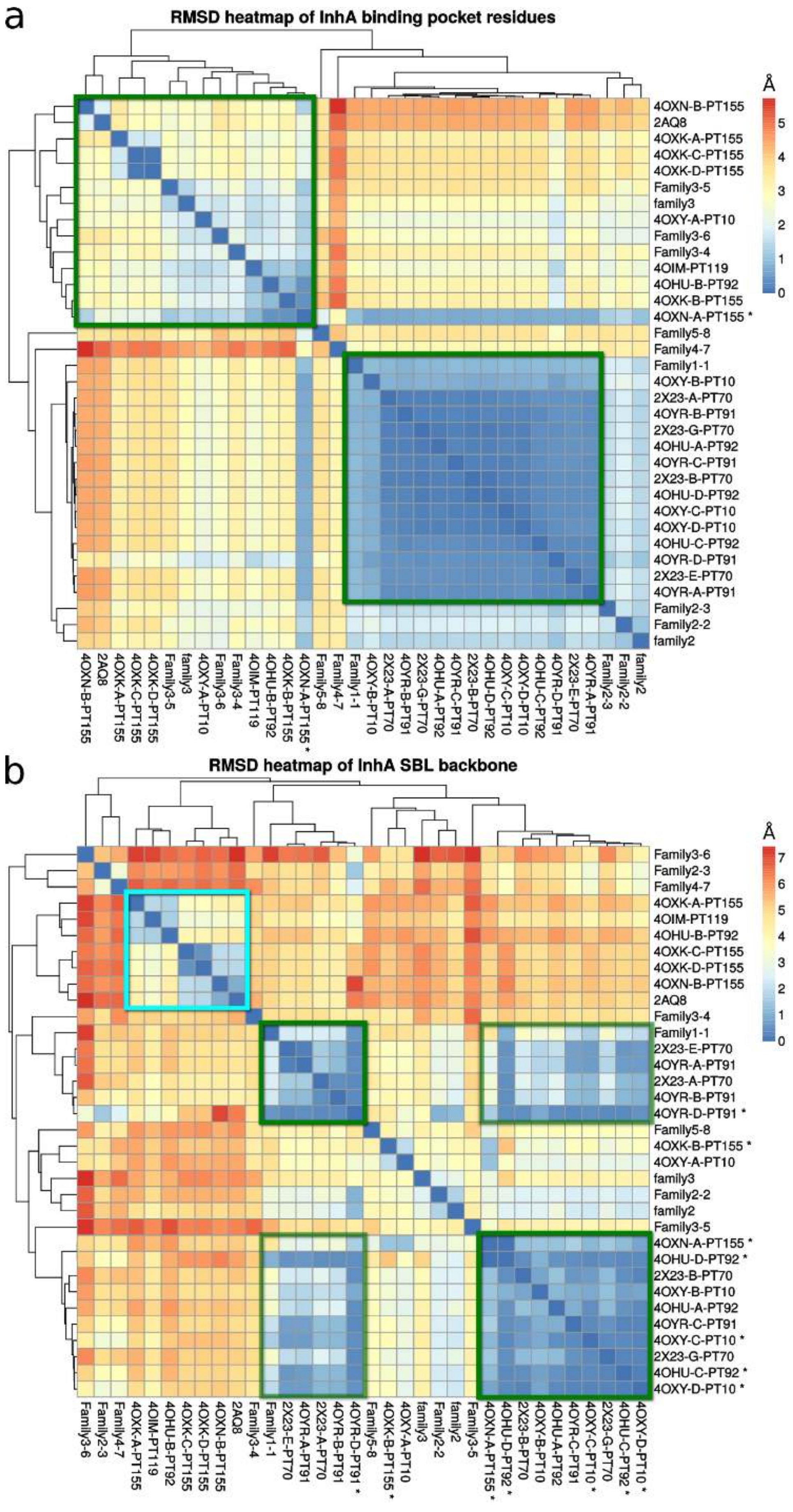


Figure 6.2 RMSD heatmaps of InhA crystal structures. **(a)** Heavy atom RMSD of InhA binding pocket residues (cf. Chapter 3). **(b)** Backbone RMSD of InhA SBL. Green and cyan rectangles indicate separate clusters. Chains marked with an asterisk are incomplete in the region of interest. RMSD values are calculated for the maximum number of common atoms.

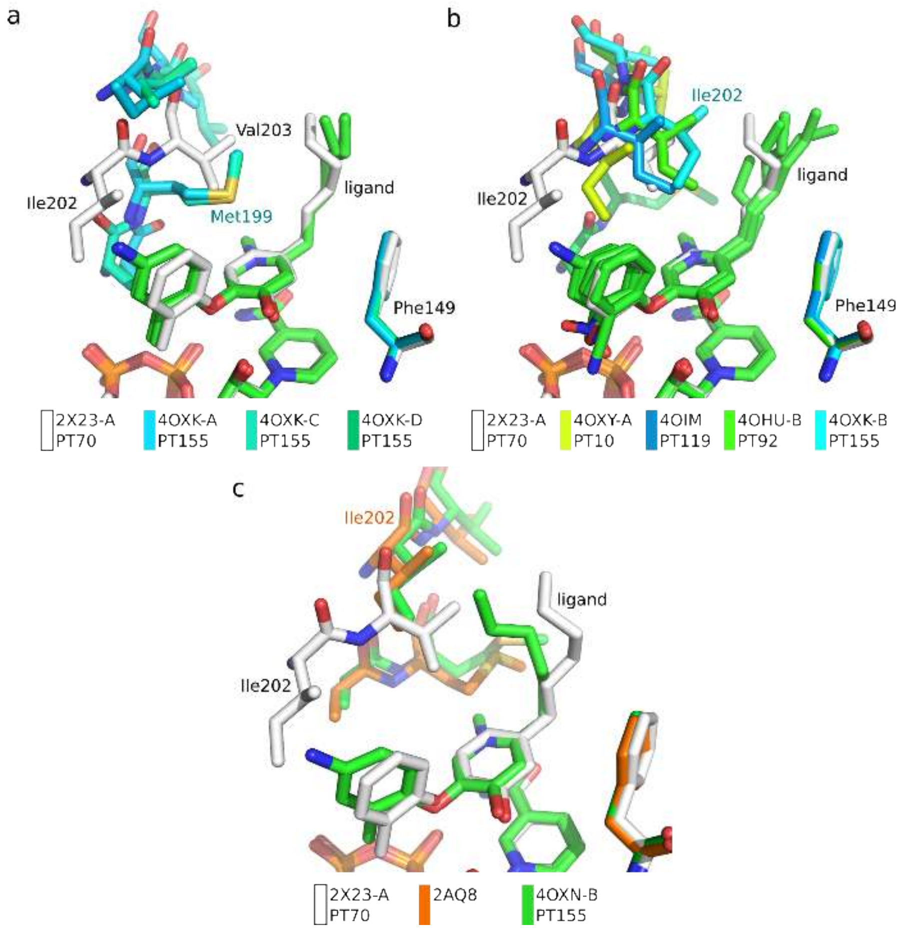


Figure 6.3 Different pocket conformations of InhA crystal structures that do not cluster to the 2X23 monomers according to the 2D-RMSD analysis (cf. Figure 6.2). Chain A of the 2X23 crystal structure is represented in white as the EI* state of InhA inhibition.

acetate is embedded between Ala201 and Phe97/Met98 and forms a polar interaction to the Met98 backbone nitrogen with a distance of 2.9 Å, whereas the other acetate interacts with the cofactor. Moreover, a third acetate ion is located near the turn between helices $\alpha 6$ and $\alpha 7$, interacting with the backbone-NH of Gly208. As already mentioned in Chapter 3, the EI* state can probably not be reached under the used crystallization conditions, leading to a rather EI-like crystal structure conformation.

As described in the literature, loop ordering is a result of slow-onset inhibition of InhA [17]. However, whereas four chains of the examined crystal structures with slow-onset binders do not promote loop ordering, the binary complex (2AQ8) and four of six **PT155**-bound structures show a fully ordered loop (cf. Table 6.1). With only three non-consecutive missing residues, the SBL of Chain C of the crystal structure 4OHU (**PT92**) may be considered nearly resolved. Given the high flexibility of the SBL, loop ordering with a rapid reversible diphenylether or no inhibitor at all is a peculiar phenomenon. Indeed, a symmetry mate analysis with PyMOL revealed that every chain with a resolved

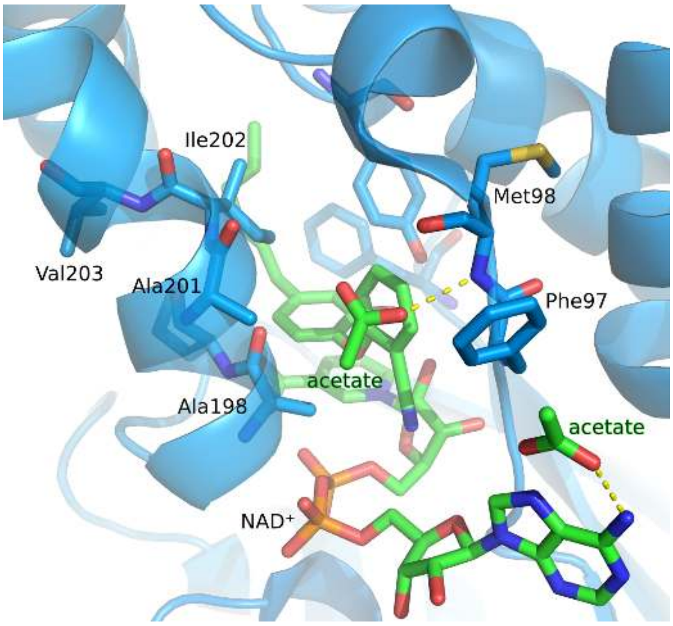


Figure 6.4 Binding pocket of ternary InhA-NAD$^+$-PT119 crystal structure (4OIM). Acetate buffer molecules are illustrated as sticks.

SBL is in close contact to a symmetry mate in this key region (cf. Table 6.1). The only exceptions to this rule are one **PT91** and two **PT70**-bound monomers, which show an ordered SBL without close contacts to a symmetry mate, underlining that **PT70** does indeed promote loop ordering [17, 45]. Conversely, most chains without fully resolved loop do not exhibit close contacts to symmetry mates (cf. Table 6.1), with two exceptions: in 4OXN-A, the chain as well as the symmetry mate show incomplete SBLs. In the case of 4OXK-B, the SBL is not fully resolved, the SBL of the adjacent symmetry mate, however, is complete. Taken together, this supports the assumption that crystal packing in this region may indeed have a stabilizing effect on the SBL. Figure 6.5 exemplarily illustrates the crystal packing on the SBL for monomer A of the complex 4OXK (**PT155**). The effect of crystal packing is further analyzed below by means of MD simulations.

6.2 MD simulations

6.2.1 System preparation

Three of the examined crystal structures were prepared for 150 ns of MD simulations in the homotetrameric assembly: 4OXK (**PT155**), 4OIM (**PT119**) and 4OHU (**PT92**) to cover a system with a rapid reversible inhibitor, the ligand with the longest residence time of the series and a ligand with a residence time comparable to **PT70**. The crystal structures 4OXK and 4OHU were preferred over 4OXN and 4OXY/4OYR, respectively,

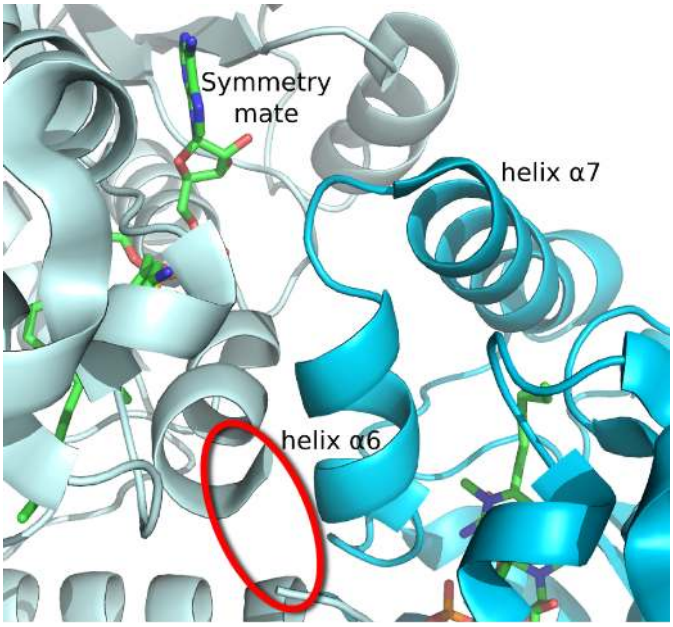


Figure 6.5 Chain A of complex 4OXK with an adjacent symmetry mate chain B' (light cyan) near the SBL. The chain B' displays missing residues in the SBL (circled in red).

due to higher resolution (1.84 Å and 1.60 Å, respectively). Since the 4OIM crystal structure has a monomeric asymmetric unit, PyMOL was used to generate the tetrameric form of the system. Missing residues in 4OXK and 4OHU were freely modeled with Modeller 9.14 [177], while keeping the protein rigid. Thus, only freely modeled residues were subjected to subsequent refinement by Modeller 9.14 using default settings. By analyzing the monomers separately, a total sampling time of 1.8 μs is reached. All systems were set up according to the simulation protocol introduced in Chapter 5. In brief, ligands were parameterized according to the GAFF with RESP atom charges obtained from potentials at the HF/6-31G* level and the protein was parameterized according to the Amber ff99SB force field. After 200 steps of energy minimization, the complexes were solvated with TIP3P water molecules and neutralized with sodium ions, followed by 10,000 steps of energy minimization and two phases of 500 ps equilibration. During the first phase, the system was heated from 100 K to 300 K in the canonical ensemble with harmonic constraints on the protein and ligand atoms, which were gradually released. During the second phase, the system was allowed to evolve freely. MD production runs were performed in the NPT ensemble using NAMD 2.9 [101]. It is still unclear, which diphenylether protomer binds to the InhA binding pocket. Since the phenol moiety is protonated under physiological pH conditions, ligands were parameterized accordingly, also to ensure comparability to the previous InhA MD setups containing **PT70**, **6PP** and **TCL** (Chapter 3) and to other publications investigating InhA-diphenylether complexes [40, 70]. As described in Chapter 4, this MD setup differs solely in the longer and gentler equilibration phase from the protocol followed in Chapter 3 to comply with more

Table 6.2 Protein backbone, binding pocket heavy atom and SBL backbone RMSD values of InhA systems in Å. Monomers were fitted to the backbone of chain A of crystal structure 2X23.

	Protein backbone		Pocket heavy atoms		SBL backbone	
	Avg. RMSD	SD	Avg. RMSD	SD	Avg. RMSD	SD
PT155$_A$	1.80	0.14	3.43	0.24	5.85	0.68
PT155$_B$	1.42	0.13	3.00	0.42	4.21	0.54
PT155$_C$	1.73	0.23	5.52	0.80	4.74	0.48
PT155$_D$	1.77	0.18	3.90	0.61	5.61	0.74
PT92$_A$	1.34	0.17	1.36	0.14	3.39	1.03
PT92$_B$	1.80	0.26	3.02	0.28	5.97	1.07
PT92$_C$	1.32	0.18	1.43	0.21	2.64	0.64
PT92$_D$	1.75	0.11	1.53	0.20	5.48	0.47
PT119$_A$	1.60	0.12	2.88	0.16	4.87	0.58
PT119$_B$	1.65	0.12	3.10	0.15	5.34	0.49
PT119$_C$	1.63	0.16	2.96	0.27	4.93	0.65
PT119$_D$	1.62	0.11	3.17	0.22	5.01	0.32

recent protocol standards and decrease the likelihood for artifacts in the initial phase of the simulation. Trajectory snapshots were saved every picosecond. For analyses, snapshots at intervals of 100 ps were considered, resulting in 1500 frames per system. For 2D-RMSD analyses trajectory snapshots were extracted every nanosecond.

6.2.2 Backbone stability

The trajectories were fitted to the backbone (N, C_α, C) of chain A of the InhA-NAD$^+$-**PT70** complex 2X23, representing the assumed EI* state (cf. Chapter 3). The simulated systems were analyzed regarding their protein backbone RMS deviation from the reference structure (Table 6.2). With average RMSD values below 2 Å, each monomer shows high stability over 150 ns of sampling time. The two most stable monomers are chains A and C of the slow-onset **PT92**-bound system, albeit the pure average backbone RMSD can obviously not be used for delimitation of slow-onset binders from rapid reversible inhibitors (cf. Table 6.2). The **PT119** monomers show very similar RMS deviation and fluctuation, most likely due to the symmetrical constitution of the tetrameric system.

6.2.3 Binding pocket conformations

The simulated systems were analyzed regarding the average RMSD of the binding pocket heavy atoms. The binding pocket was defined according to Chapter 3 as Phe149, Tyr158, Ala198, Met199, Ile202 and Val203. With respect to chain A of the crystal structure 2X23, three of four **PT92**-monomers show very stable binding pockets with average

RMSD values between 1.36 Å and 1.53 Å and low standard deviations (Table 6.2). Chain B of the **PT92**-system exhibits a higher average binding pocket RMSD due to the different pocket conformation in the crystal structure. As shown in the clustering of the binding pockets, this chain is rather comparable to the **PT155**- and **PT119**-conformations (Figures 6.2a and 6.3). Thus, a direct RMSD comparison between single monomers is hampered by the diverse crystal structure conformations of the evaluated chains. However, the fluctuations in the system with the rapid reversible ligand **PT155** (with exception of chain A) are much higher compared to the values of **PT92** chain B or the **PT119** monomers, indicating a much lower binding pocket stability for the **PT155** complex.

A 2D-RMSD analysis was performed using VMD [140] to assess the stability of the single monomeric subunits with respect to the binding pocket residues in more detail. The binding pocket residues show a very high stability in the case of the slow-binder **PT92** (Figure 6.6) over 150 ns of MD simulation. Although **PT92** chain B exhibits a different pocket conformation in the crystal structure, the binding pocket behaves stably over the sampled trajectory time. The four **PT119**-bound monomers show very stable dynamics as well and low deviation among each other, which is not surprising, given the identical starting structures. The close conformational similarity between the **PT119** monomers and **PT92** chain B is indicated by a dark green rectangle in Figure 6.6. Chain A and particularly chain B of the **PT155** monomers also show reasonable stabilization of the binding pocket in their respective starting structure. The other two monomers display very large fluctuations.

6.2.4 Hydrogen bond analysis

The heavy atom distances between the Tyr158-OH and the ligands were measured for each monomer over 150 ns (Figure 6.7). Whereas each monomer with a slow-onset inhibitor exhibits a stable hydrogen bond to Tyr158, the distributions of three **PT155** monomers (chains B, C and D) show clear tendencies towards higher distances. Chain C, in particular, exhibits a highly unstable hydrogen bond between the ligand and Tyr158, as emphasized by the bimodal distribution and the large IQR in Figure 6.7. These results are in agreement with the average RMSD and fluctuations of the binding pocket residues, which are the highest observed in the evaluated systems (cf. Table 6.2).

6.2.5 SBL stability

As expected, the SBL is prone to much higher RMS deviations than the binding pocket residues (Table 6.2 and Figure 6.8). As a result, only **PT92** monomer C shows a

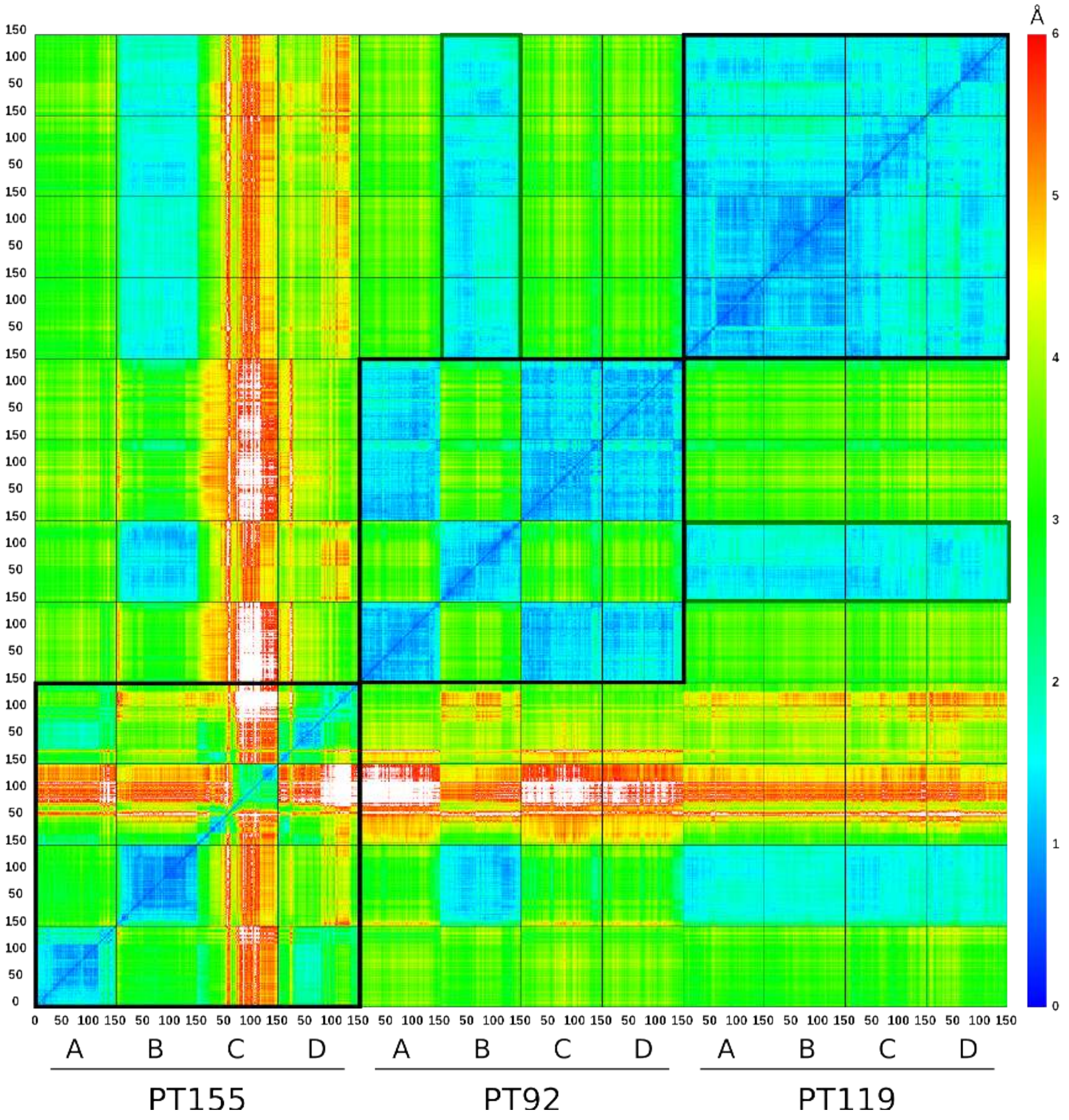


Figure 6.6 Heavy atom 2D-RMSD plot of the binding pocket residues of complexes 4OXK (PT155), 4OHU (PT92) and 4OIM (PT119) in their tetrameric forms over 150 ns. Single monomers are framed by thin black lines. Thus, each small box represents the trajectory of a single chain over 150 ns sampling time. Large black rectangles delimit the tetrameric systems. The smaller dark green rectangle illustrates conformational similarity between **PT92** chain B and the **PT119** monomers. RMSD values above 6 Å are colored white.

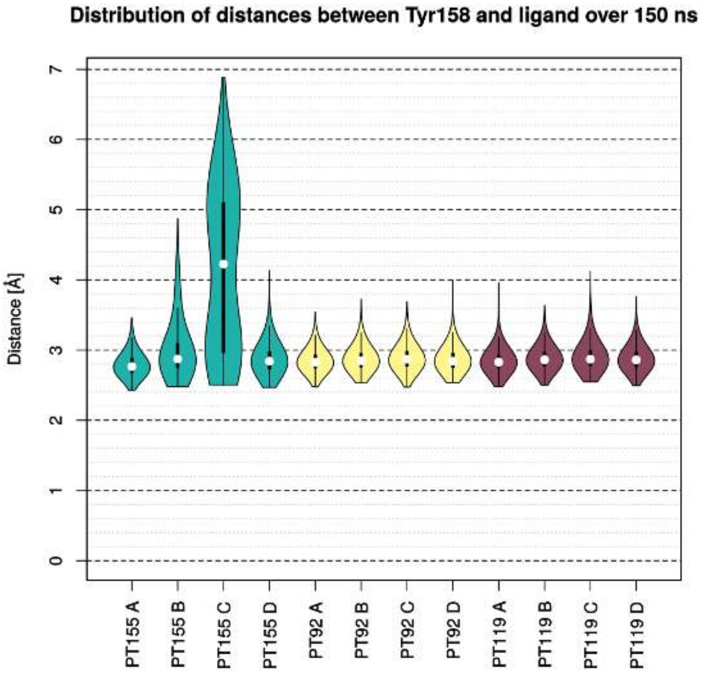


Figure 6.7 Violin plots of distances between the phenolic oxygen of Tyr158 and the respective ligands. White dots depict the medians. Thick vertical lines indicate the interquartile ranges (IQR), thin lines extend to $1.5 \cdot \mathrm{IQR}$ from the third and first quartile, respectively. The shape of the violins illustrates the kernel density estimation of the respective distribution.

reasonably stable SBL with respect to the starting structure, whereas the remaining chains rather display multiple different loop conformations. Interestingly, the SBL of **PT92** chain C is the only nearly complete one without close contacts to symmetry mates in the crystal structure 4OHU (cf. Table 6.1). On the other hand, chain D (which contains an SBL with 13 freely modeled residues) behaves unstably with respect to the modeled starting structure, despite the low standard deviation (cf. Table 6.2 and Figure 6.8).

6.2.6 Analysis of crystal packing effect

Although loop ordering is assumed to be a result of slow-onset inhibition of InhA, several chains of recent crystal structures of InhA bound to the rapid reversible inhibitor **PT155** show a fully resolved SBL as well [17, 40, 45]. A crystal packing analysis of the structures revealed close contacts of the monomers to symmetry mates of the asymmetric unit in the SBL region. To further assess the influence of crystal packing on the crystal structures with rapid reversible inhibitors, a 150 ns MD simulation was set up for chain A of the crystal structure 4OXK and chain B of the adjacent symmetry mate (Figure 6.5).

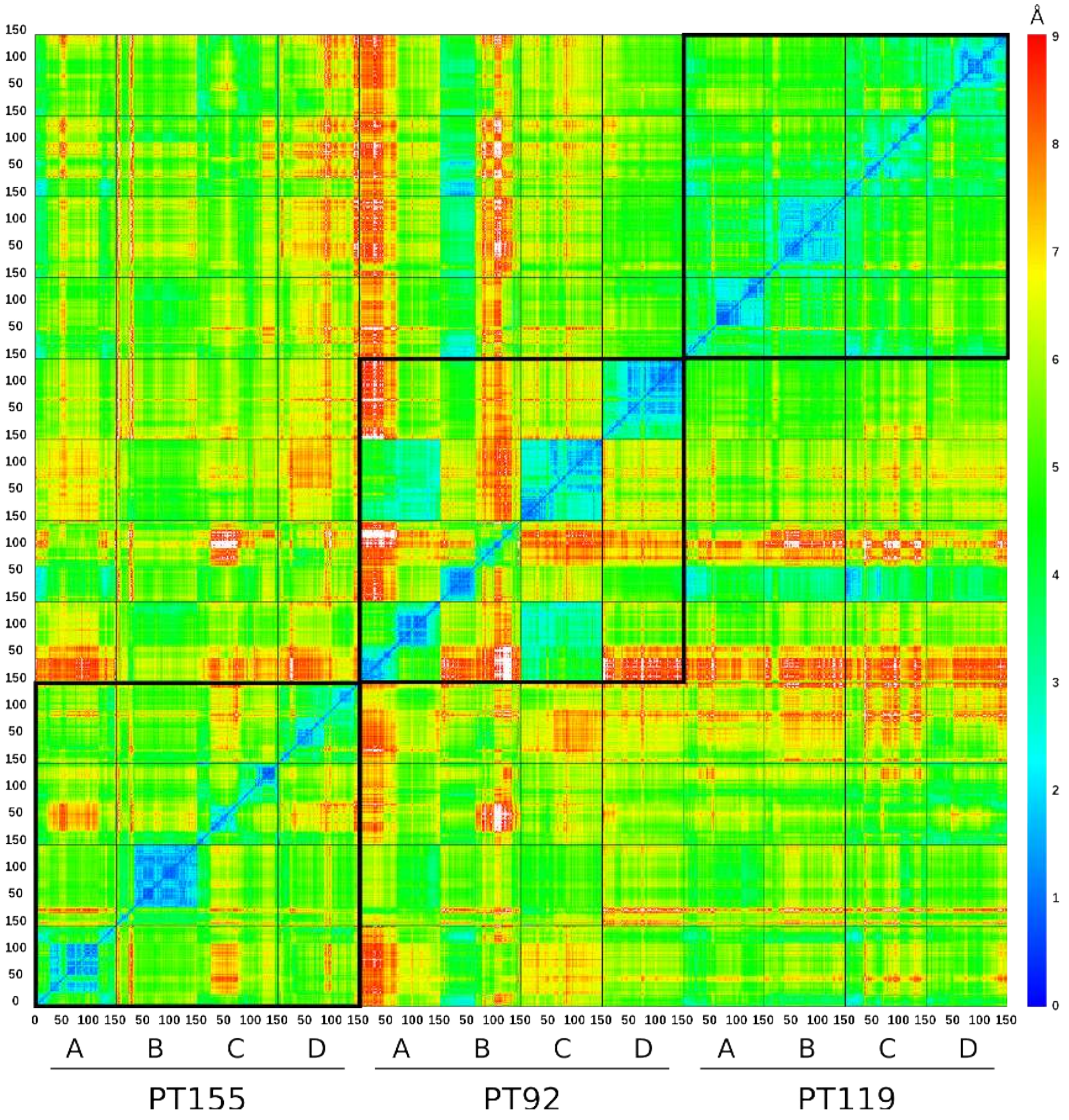


Figure 6.8 Backbone 2D-RMSD plot of the SBL of complexes 4OXK (PT155), 4OHU (PT92) and 4OIM (PT119) in their tetrameric forms over 150 ns. Large black rectangles delimit the tetrameric systems. RMSD values above 9 Å are colored white.

Two missing residues of symmetry chain B (Gly205 and Ala206) were modeled into the structure using Modeller 9.14 with rigid protein and default settings for refinement.

The trajectories of chains A and B were analyzed with respect to their counterparts from the tetrameric setup. The monomers with close contact in the SBL region to a symmetry mate (*symm*-system) exhibit distinctly lower average RMSD values than chains A and B of the previously simulated tetrameric (*tet*) system (Table 6.3), i.e. the divergence from the respective starting structures is much lower for chains with close SBL contacts. Moreover, the lower RMSD standard deviation in the *symm* monomers underlines the higher stability over the sampling time. This is further emphasized by the much lower

Table 6.3 SBL backbone RMSD/RMSF and pocket RMSD values of InhA systems in Å. Monomers were fitted to the backbone of chain A of crystal structure 2X23. Reference for RMSD calculation is the minimized crystal structure conformation of each respective monomer. Reference for RMSF calculation is the trajectory average structure of each respective SBL determined using the RMSD Trajectory Tool plugin of VMD.

	SBL backbone		**SBL backbone**		**pocket heavy atoms**	
	Avg. RMSD	**SD**	**Avg. RMSF**	**SD**	**Avg. RMSD**	**SD**
PT155$_A$ **tet.**	4.36	1.04	2.49	0.74	1.68	0.33
PT155$_B$ **tet.**	4.73	0.54	1.92	1.07	1.71	0.30
PT155$_A$ **symm.**	2.64	0.52	1.60	0.45	1.51	0.26
PT155$_B$ **symm.**	2.15	0.32	1.09	0.31	1.51	0.16

average RMS fluctuation (RMSF), i.e. the average root-mean-squared deviation with respect to the trajectory average structure of the SBL (Table 6.3). Figure 6.9 depicts the distributions of the SBL backbone RMSD and RMSF as box plots. Non-overlapping notches in the boxes illustrate a significant difference between the distributions. Thus, the close contacts in the SBL region, as observed in the crystal structure 4OXK and simulated in the *symm*-system, have a highly significant effect on the stability. This is further affirmed by pairwise Mann-Whitney-U tests of the RMSD and RMSF distributions of *tet* chain A vs. *symm* chain A and *tet* chain B vs. *symm* chain B, respectively ($p \ll 0.001$). Regarding the binding pocket heavy atoms, the *symm*-monomers show lower average RMSD values and standard deviations than the *tet*-monomers, albeit to a lesser extent (Table 6.3). These results provide strong evidence that the SBL of InhA bound to the rapid reversible diarylether **PT155** is ordered in the crystal structure as a result of crystal packing. This is in agreement with the assumption that rapid reversible inhibitors generally do not promote loop ordering in InhA [17]. Conversely, the SBL of **PT92** chain C, which is nearly fully resolved and has no contacts to symmetry mates in the crystal structure, behaves stably with respect to the starting structure over 150 ns of MD simulation, as do the fully resolved SBLs of the **PT70**-bound monomers without contacts to symmetry mates in the crystal structure, as shown in Chapter 3. These results indicate that loop ordering without close crystal contacts in the SBL region might indeed only be achieved by slow-onset inhibitors.

6.2.7 Conclusion

In conclusion, the slow-onset inhibitors show stable behavior in their respective binding pocket conformations in contrast to inhibitor **PT155**, where severe pocket instabilities occur in two of four chains. With respect to the SBL starting structures, only one chain (4OHU-C, **PT92**) displayed a reasonably stable conformation over 150 ns of simulation time. Interestingly, this chain is the only monomer with a nearly complete SBL without

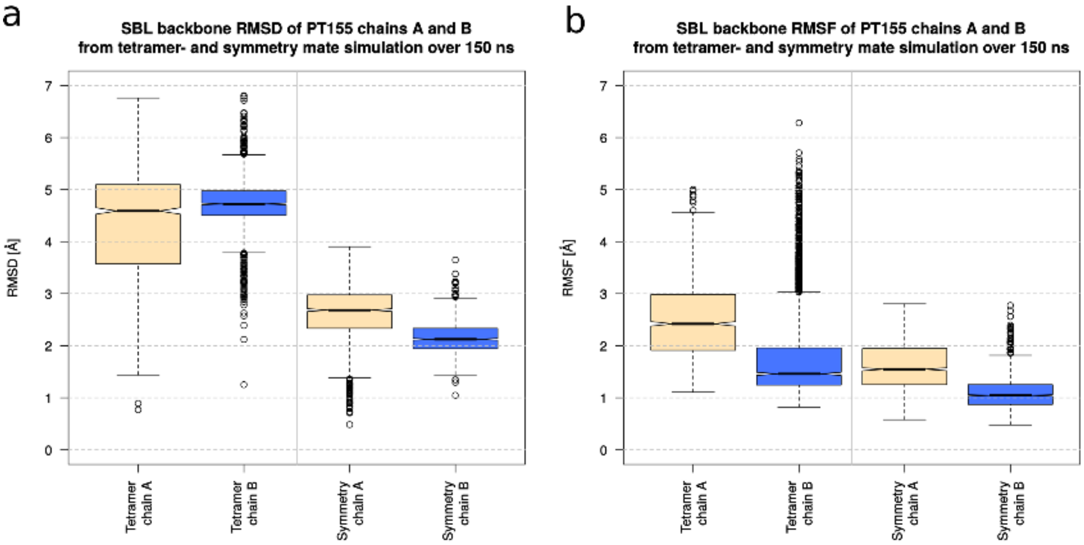


Figure 6.9 Box plots of SBL backbone (a) RMSD and (b) RMSF values of PT155 chains A and B of the tetramer- and symmetry mate simulation, respectively, in Å. Monomers were fitted to the backbone of chain A of crystal structure 2X23. Reference for RMSD calculation is the minimized crystal structure conformation of each respective monomer. Reference for RMSF calculation is the trajectory average structure of each respective SBL determined using the RMSD Trajectory Tool plugin of VMD. Non-overlapping notches indicate significant differences between the simulations of the tetrameric setup and the symmetry mate setup.

close contacts to symmetry mates in the crystal structure, i.e. resolved without crystal packing effects. Whereas the **PT155** monomers showed largely flexible SBLs, a simulation of chain A with a symmetry mate adjacent to this region showed a stabilizing effect of symmetry mate contacts on the SBL. These results underline the hypothesis that loop ordering in crystal structures of InhA bound to a rapid reversible diarylether might be strongly supported by crystal packing in the region of the SBL. Conversely, only chains with resolved SBLs and without close contacts in crystal structures exhibit a stable SBL in MD simulations of the tetrameric assembly, as indicated by simulations of **PT92** chain C as well as **PT70**, in which case chains B and D with no close crystal contacts in the SBL region show the most stable loops (cf. Figure 3.8). Hence, crystal structures containing a resolved SBL without the effect of crystal packing might indeed represent the stable EI* state of slow-onset inhibition of InhA.

6.3 Steered MD simulations

Induced ligand extraction has been proven in Chapter 5 as a useful approach to quantify the inhibitory efficacy of a compound in a protein in terms of residence time. Whereas several crystal structures bound to ligands with experimentally determined residence times are available for *sa*FabI, a similarly extensive dataset is not yet available for the mycobacterial homolog InhA, limiting the possibilities to derive a statistical model. Nonetheless, the maximum free energy changes of induced ligand extraction can be analyzed comparatively for selected InhA crystal structures with kinetics data to qualitatively delimit rapid reversible binders from slow-onset inhibitors.

6.3.1 System preparation

The SMD protocol introduced in Chapter 5 was applied on the ternary InhA complexes examined in Chapter 6.2.1 and two additional InhA complexes bound to the small ligand triclosan (**TCL**). Whereas **TCL** is a slow-binding *sa*FabI inhibitor with a residence time of 139.5 minutes [30], it binds rapid-reversibly without a measurable residence time to InhA [17, 64]. Indeed, the very flexible SBL is not resolved in the ternary InhA-NAD$^+$-**TCL** complex (PDB 2B35), most likely due to very high mobility [24, 64]. For this reason, two different InhA-**TCL** systems were prepared: in one system, the missing loop was freely modeled into 2B35 using Modeller 9.14 [177] with rigid protein, resulting in a very open conformation (Figure 6.10). Subsequently, the modeled residues were refined by Modeller 9.14 using default settings. The other system corresponds to the crystal structure 2X23 with **TCL** placed into the binding pocket after structural alignment of 2X23 with the crystal structure 2B35 (cf. Chapter 3), thus creating a hypothetical EI* state for **TCL**.

In the case of *sa*FabI, diphenylethers are assumed to bind in their deprotonated form [30]. Since the SMD work-flow presented in Chapter 5 was evaluated on deprotonated ligands, all diphenylethers were re-parameterized with a negative charge to use consistent protonation states for SMD simulations. For the induced extraction of ligand **PT155** two separate crystal structures were used. Li *et al.* (2014) [40] defined chain B of the crystal structure 4OXN as the EI state of InhA inhibition. Hence, this chain was prepared for SMD simulations besides chain A of the previously described complex 4OXK. To further evaluate the contribution of the protein conformation to the maximum ΔG, additional systems of **PT119** placed into the crystal structure 2X23 and **TCL** placed into the crystal structure 4OXN-B were prepared via structural alignment of the crystal structures. The previously defined pulling direction of **PT70** (2X23) was used for all complexes as reaction coordinate for induced ligand extraction (cf. Chapter 5).

6.3.2 SMD results

The maximum free energy changes were reconstructed using an untrimmed average of work values, as well as a 25% trimmed average according to the protocol outlined in Chapter 5 to reduce high fluctuations in the work profiles (Table 6.4 and Figure 6.11) [35].

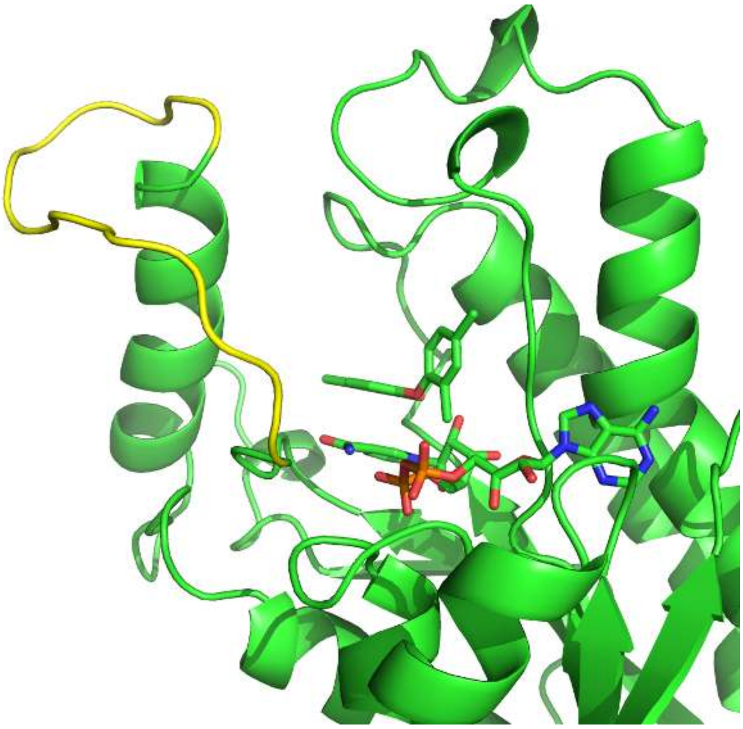


Figure 6.10 Crystal structure of 2B35 with modeled residues of the SBL (yellow). The ligand **TCL** and cofactor **NAD**$^+$are represented as sticks.

It is known that the Jarzynski equality is not applicable, if the fluctuations of the work values are much higher than the temperature k_bT (cf. Chapter 5.7) [35]. Although the exponential Jarzynski expression can generally be difficult to estimate, Gore *et al.* (2003) found that the exponential average is a valid estimator even for limited sampling, in contrast to the first cumulant expansion ΔG_1 [35, 165]. Also the second order cumulant expansion ΔG_2 can show a higher bias, if the variance is not estimated accurately due to small trajectory numbers [165]. Hence, the exponential estimator ΔG_e reconstructed from trimmed averages will be considered primarily in this analysis due to large fluctuations in the work values, as indicated by high standard deviations in the untrimmed average work estimator ΔG_1 (Table 6.4).

Although the ligand **PT119** is characterized by a very high residence time of 80 minutes in InhA [46], its simulated maximum free energy change derived from SMD simulations of the crystal structure 4OIM is similar to or lower than the values obtained for **PT70** and **PT92**, respectively (cf. confidence intervals in Figure 6.11a). Conversely, placing **PT119** into the 2X23 crystal structure leads to the highest observed maximum ΔG_e (Table 6.4). As discussed in detail above and in Chapter 3, the InhA-**PT119** crystal structure 4OIM differs in its binding pocket conformation significantly from the **PT70**/**PT92**-bound crystal structures, and does, hence, not show a typical EI* conformation, but rather characteristics of the assumed EI state. The comparably low

Table 6.4 Maximum free energy change of induced ligand extraction of InhA inhibitors with standard deviation in kcal/mol.

	trimmed			untrimmed		
	ΔG_e	ΔG_1	ΔG_2	ΔG_e	ΔG_1	ΔG_2
PT70 (2X23-A)	**89.02 ± 11.31**	106.29 ± 11.32	68.11 ± 2.38	77.53 ± 20.31	107.15 ± 20.32	39.95 ± 6.09
PT92 (4OHU-A)	**97.94 ± 4.50**	104.98 ± 4.51	88.61 ± 4.15	87.63 ± 12.04	105.39 ± 12.06	52.96 ± 4.40
PT119 (4OIM)	**86.27 ± 5.33**	92.80 ± 5.33	69.44 ± 5.24	72.72 ± 11.92	92.40 ± 11.96	32.48 ± 4.55
PT119 (2X23-A)	**102.19 ± 5.08**	108.80 ± 5.10	88.50 ± 4.74	89.29 ± 12.27	109.95 ± 12.28	40.55 ± 6.77
TCL (2X23-A)	**77.53 ± 5.06**	83.35 ± 5.25	63.21 ± 4.83	66.07 ± 10.97	84.70 ± 11.02	49.49 ± 3.86
TCL (2B35-A)	**46.13 ± 3.07**	49.30 ± 3.07	41.65 ± 2.94	36.95 ± 6.35	49.20 ± 6.34	24.38 ± 3.00
TCL (4OXN-B)	**59.92 ± 2.25**	62.98 ± 2.53	58.76 ± 2.04	50.98 ± 6.67	62.67 ± 6.74	31.19 ± 3.31
PT155 (4OXK-A)	**78.12 ± 3.29**	82.28 ± 3.29	73.22 ± 3.29	56.25 ± 9.01	81.63 ± 9.02	33.85 ± 3.51
PT155 (4OXN-B)	**69.09 ± 2.81**	72.73 ± 2.94	65.65 ± 2.79	65.95 ± 6.30	73.37 ± 6.29	40.24 ± 6.29

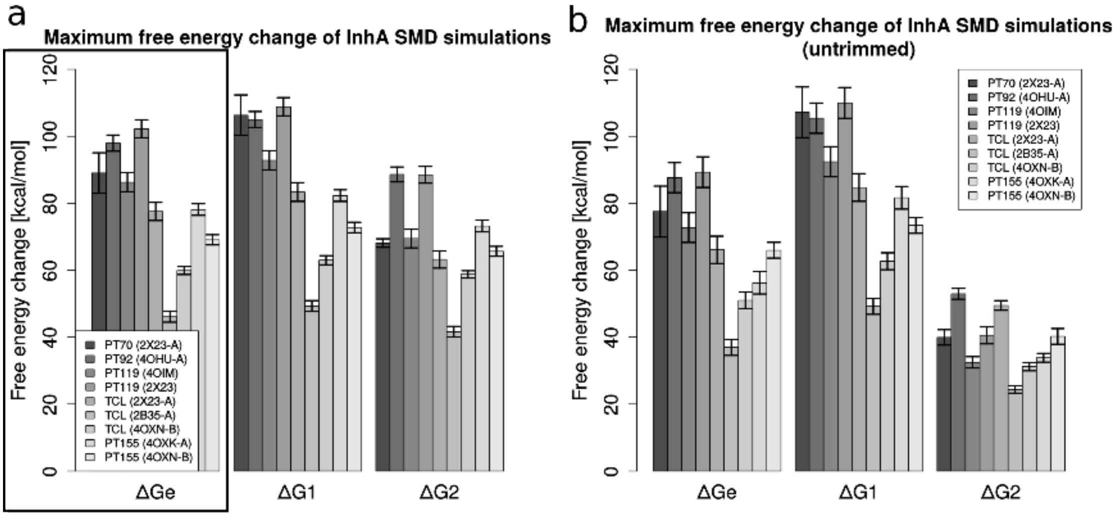


Figure 6.11 Maximum free energy changes of induced ligand extraction of InhA complexes. Error bars indicate the 95% confidence interval. The maximum free energy change according to ΔG_e reconstructed from trimmed averages (black rectangle) is assumed to yield the most accurate results and allows delimitation of slow-onset inhibitors from rapid reversible ligands, indicated by non-overlapping confidence intervals.

maximum free energy change in 4OIM and the high ΔG_e of **PT119** in 2X23 provide strong additional evidence for this hypothesis.

The contribution of the protein conformation to the height of the maximum energy barrier is also apparent from the various systems with rapid reversible inhibitors. In general, systems containing **TCL** and **PT155** show lower maximum free energy changes ΔG_e compared to the slow-onset inhibitors, which is consistent with the kinetic properties of the rapid reversible ligands in InhA. Compared to the 2X23-**TCL** system with a hypothetical EI* state, **TCL** in the more helix-open conformation 4OXN-B and in 2B35 with the freely modeled SBL shows a significant drop of ΔG_e. This emphasizes that dissociation from an EI state happens more easily with respect to dissociation from an (artificial) EI* state. Moreover, the sensitivity of the presented SMD protocol regarding the starting conformation of the protein-ligand complex is underlined by the divergent maxima for these ligands in different crystal structure conformations. With respect to the ligand **PT155**, a drop in maximum free energy change is observed for the monomer 4OXN-B with a more open conformation of helix $\alpha 6$, compared to 4OXK-A.

Based on the exponential estimator ΔG_e reconstructed from trimmed averages, a significant delimitation of slow-onset inhibitors from rapid reversible ligands is possible, as indicated by non-overlapping confidence intervals in Figure 6.11. Since the SMD results are very sensitive to the respective protein starting conformations, separation of ligands with different kinetic profiles is best achieved from valid protein conformations representing the assumed state of the protein bound to the respective ligands.

It is notable that the maximum free energy changes of the rapid reversible ligand **PT155** derived from the second order cumulant expansion ΔG_2 of Jarzynski's equality are relatively high, compared to the results of the slow-onset inhibitors **PT70**, **PT92** and **PT119**. This trend, however, is confined to ΔG_2 and not observable in the results of ΔG_e and ΔG_1, which are qualitatively well in line with the kinetic profiles of the examined ligands. Whereas the second order cumulant (with 25% trimmed average) yields the most accurate model for drug-target residence time prediction for the staphylococcal enoyl-ACP reductase *sa*FabI (cf. Chapter 5), the respective maximum free energy changes in *M. tuberculosis* InhA do not reflect the experimental data, as free energy reconstruction of ΔG_2 is heavily influenced by very high fluctuations in the measured work profiles.

In summary, the systems show maximum free energy changes ΔG_e qualitatively in line with ligand kinetics, provided the simulations are started from valid protein conformations. Thus, comparison of ΔG_e values allows delimitation of slow-onset binders from rapid reversible inhibitors. The outcome of the SMD simulations is sensitive to the protein starting conformation, as underlined by the maximum ΔG_e of the ligands **TCL**, **PT155** and **PT119** in various crystal structures. The second order cumulant expansion ΔG_2 is largely affected by high fluctuations in the replica simulations, rendering its applicability for this target very limited with the current simulation protocol. These results emphasize the necessity for an additional extensive parameter study regarding SMD simulations for residence time prediction as outlined in Chapter 5, e.g. with respect to the number of replica simulations or pulling speed. However, validation of parameter modifications needs computationally very expensive simulations, hampering systematic parameter studies.

6.4 Conclusion

Analysis of various crystal structures combined with MD and SMD simulations were employed to revisit the enzyme-inhibitor complex states EI and EI* of diphenylether ligand binding to InhA. First, the crystal structures were clustered by their RMSD with respect to binding pocket residues and SBL backbone. The slow-onset inhibitors and the rapid reversible inhibitors consistently populated separate clusters with few exceptions, namely 4OIM (**PT119**), 4OXY-A (**PT10**) and 4OHU-B (**PT92**). Subsequently, a symmetry mate analysis was conducted to unveil possible effects of crystal packing on the 3D-structures. Only two 2X23-, one 4OHU- and one 4OYR-chain showed resolved loops without nearby crystal contacts, indicating that these inhibitors indeed promote ordering of the flexible SBL. In large-scale MD simulations the stability of selected systems could be assessed with respect to binding pocket and SBL. Pocket dynamics resulted in very high stability in slow-onset inhibitor-bound systems, whereas the binding pocket of the InhA-NAD$^+$-**PT155** complex displayed large fluctuations. Thus, a stable binding pocket can help delimit slow-onset inhibitors from rapid reversible ones. The SBL, on the other hand, was only found to be stabilized over 150 ns in one **PT92**-bound monomer.

Interestingly, the stable SBL of **PT92** monomer C is the only one of the simulated systems without close symmetry contacts in the crystal structure, alongside chains B and D of the **PT70** simulations described in Chapter 3. Conversely, an MD simulation of **PT155** chain A with a symmetry mate in close proximity of the SBL showed significantly more stable loops, further emphasizing the effect of crystal packing on the SBL in InhA crystal structures. The resolved SBLs in InhA crystal structures with rapid reversible inhibitors might, thus, be ascribed to close contacts to symmetry mates in this region. These results provide evidence that chains containing resolved SBLs without close contacts to symmetry mates might represent a genuine EI* conformation. Moreover, the facts that no **PT119** SBL is entirely stable in its starting conformation (as opposed to **PT92** chain C) and that the crystal structure 4OIM (**PT119**) exhibits close contacts of symmetry mates in the SBL region underline the hypothesis that a 4OIM-like binding pocket/SBL conformation does not represent the final EI* state, as also mentioned by the authors of the corresponding publication [46], whereas a **PT70**-like binding conformation might in fact be considered the final EI* state.

This notion is further affirmed by SMD simulations. The maximum free energy changes ΔG_e of induced inhibitor extraction of these systems are qualitatively well in agreement with the experimental residence times of the ligands in InhA: **PT70** and **PT92** yield comparable maximum free energy changes, which are both higher than the simulated maxima of the structures bound to the rapid reversible inhibitors **TCL** and **PT155**.

PT119 in the crystal structure 4OIM, however, also achieves a maximum free energy change below the other slow-onset inhibitors, although its residence time is 3-fold higher. After placing **PT119** in the crystal structure 2X23, i.e., the assumed EI* state of InhA inhibition by diphenylethers, ΔG_e increases drastically, providing strong evidence that the recorded crystal structure 4OIM does not represent the EI* state, but rather an EI-like state. Moreover, this emphasizes the sensitivity of the presented method on the protein starting conformation.

As further shown in SMD simulations, not only the wide opening of helix $\alpha 6$ leads to much lower maximum free energies (**TCL** in 4OXN-B vs. **TCL** in 2X23 or **PT155** in 4OXN-B vs. **PT155** in 4OXK-A), but also the helical twist with Ile202 and Val203 moving over the ligand toward the inside of the binding pocket (**PT119** in 4OIM vs. 2X23). This provides further evidence for the assumption that wrapping of Ile202 and Val203 around the B-ring of the diphenylether inhibitor is a crucial step in slow-onset inhibition of InhA, underlining the validity of a suggested 5'-substitution to increase the energy barrier for this conformational transition, as proposed in Chapter 3 and evaluated in Chapter 4. In general, SMD simulations were proven as a useful method to delimit slow-onset inhibitors from rapid reversible ligands in InhA based on the maximum ΔG_e of induced ligand egress, although an extensive parameter study is assumed to further improve the accuracy of the simulation results.

Altogether, the assumptions regarding the EI and EI* state presented in this work (cf. Chapter 3) and the work of Li *et al.* (2014) [40] agree with crystal structure analyses, MD and SMD simulations: a **PT70**-like binding conformation of pocket and SBL corresponds to the EI* state, whereas a helical shift of Ile202 and Val203 towards the binding pocket with a more open conformation of helix $\alpha 6$ might be considered the EI state of InhA drug-target association. However, large intra-crystal variations of SBL and binding pocket conformations of crystal structures with slow-onset and rapid reversible inhibitors likewise (cf. Figures 6.2 and 6.3), as well as the contributions of crystal packing to loop ordering indicate that a purely binary classification into the EI and EI* state, respectively, might lead to an oversimplification of the conformational space of InhA inhibition.

Chapter 7

Summary – Part I

The drug-target residence time t_R has gained increasing attention in drug development due to its good correlation to *in vivo* efficacy [24, 25]. However, the lack of structural information about the transition states of inhibitor binding hampers rational optimization of t_R [24, 26]. With rising computational power and enhanced sampling methods, MD simulations provide access to transition and metastable intermediate states [26]. Thus, MD simulations were employed to elucidate the molecular features that govern long drug-target residence time in bacterial enoyl-ACP reductases, a promising drug target for antibacterial drug design [17].

The mycobacterial enoyl-ACP reductase InhA is known to bind to inhibitors of the diphenylether class with different kinetic profiles. Thus, various systems of InhA were set up, bound to either long-binding or rapid reversible inhibitors. An extensive structural analysis of MD trajectories with a total sampling time of 3.0 μs revealed five recurring conformational families. Two of these conformational families correspond to what are assumed to be the EI and EI* states of ligand association of InhA. The dynamic features of the protein-ligand complexes could be linked to the unique substitution patterns of the bound ligands, revealing important insights into the determinants of long drug-target residence time in InhA: (1) occupation of the hydrophobic pocket, (2) introduction of an anchor-moiety in 2'-position and (3) introduction of a small 5'-substituent to embed in between Ile202 and Val203 and prevent these residues from shifting into the hydrophobic pocket.

The latter suggestion was evaluated in additional MD simulations. Two homotetrameric InhA systems were set up with 5'-methyl-**PT70** and 5'-chloro-**PT70**, respectively. Trajectories of 50 ns each resulted in a combined sampling time of 400 ns with respect to the independent monomers. Stabilization of the flexible substrate binding loop (SBL) was achieved in particular in the 5'-chloro-**PT70**-bound monomers, whereas the binding pocket was observed to be very stable in every examined monomer of both systems, except in monomer $C2$, which drifted into a Family 3* conformation. In general, the 5'-substituted ligands show close contacts to Met103, but are not able to prevent instabilities in this residue which translate to the binding pocket, leading to the alternate pocket conformation in the case of monomer $C2$. Thus, the ligands might need further improvement regarding the substitution pattern in order to stabilize Met103 and, hence, the entire binding pocket.

Since drug-target residence times of slow-onset inhibitors are much longer than the time scale achieved with classical MD simulations, enhanced sampling techniques are required to simulate complete ligand-unbinding events. Steered MD (SMD) simulations are a straightforward method for induced extraction of ligands from the binding pocket along a given reaction pathway. Thus, SMD simulations were employed to simulate ligand-unbinding of eleven different *Staphylococcus aureus* enoyl-ACP reductase FabI complexes and reconstruct the maximum free energy change of ligand dissociation. The resulting free energy changes could be associated with $ln(t_R)$ to obtain a very accurate and quantitative regression model. New crystal structures or docking poses of hypothetical ligands in FabI can now be subjected to the outlined protocol to predict their drug-target residence time according to the linear regression model.

Proper characterization of the EI and EI* states of slow-onset inhibition is imperative for rational optimization of t_R. Recently published crystal structures of InhA, however, exhibit largely diverse conformations regarding binding pocket and SBL [40, 46]. In the last chapter of Part I, the concept of the EI and EI* states of InhA inhibition was revisited by means of crystal structure analysis, MD and SMD simulations. Overall, the assumptions regarding the EI and EI* states in the previous chapters could be affirmed: the 2X23 crystal structure conformation corresponds to the EI* state, whereas a twist of Ile202 and Val203 toward the inside of the binding pocket with a more open helix $\alpha6$ represents the EI state. However, SMD simulations showed that not only wide opening of helix $\alpha6$, but also the shift of Ile202 and Val203 itself has a large influence on the maximum ΔG_e. In general, SMD simulations were proven as a useful approach to delimit slow-onset InhA inhibitors from rapid reversible ligands. The effect of crystal packing on the SBL was investigated by means of a symmetry mate MD simulation, showing that the SBL is indeed stabilized by close contacts to adjacent symmetry mates. Conversely, structures containing a resolved SBL without the proximity of a symmetry mate in the SBL region showed stable dynamics and might represent, thus, a genuine EI* state.

In conclusion, MD techniques were applied to investigate the determinants of long-binding kinetics in bacterial enoyl-ACP reductases and, moreover, derive a quantitative prediction model for drug-target residence time. These findings can contribute to future rational drug design endeavors against InhA and *sa*FabI towards inhibitors with longer residence times.

Part II

Prediction of *Mycobacterium tuberculosis* Cell Wall Permeability: Development of MycPermCheck and its Application in Virtual Screening for Antimycobacterial Substances

Chapter 8

Background

8.1 The *Mycobacterium tuberculosis* cell wall hampers antitubercular drug design

An important natural defense mechanism of *M. tuberculosis* is its thick and waxy cell wall which provides a first powerful barrier against antibiotic drugs (Figure 8.1). It consists of a peptidoglycan-arabinogalactan-mycolic acid core as well as the outmost layer, the so-called capsule [60, 178]. It has been shown that not only hydrophilic agents, but also lipophilic agents may have severe problems passing the permeability barrier of the cell wall, owing to the unusually low fluidity of the lipid bilayer [179]. Thus, *M. tuberculosis* is intrinsically resistant to many drugs, due to the unique composition of its cell wall [8–10].

Without the ability to penetrate the *M. tuberculosis* cell wall, even very potent inhibitors of validated mycobacterial drug targets like InhA [181] will not have any efficacy. In 2004, Hong and Hopfinger constructed a complex computational model of the *M. tuberculosis* cell wall and conducted MD simulations to determine diffusion coefficients for 13 different first- and second-line antituberculars [182, 183]. While such *ab initio* approaches yield important results with respect to molecule transport through the cell wall, the computational cost narrows the applicability on large datasets. Thus, a fast, knowledge-based method for permeability prediction is desirable as an additional filter criterion for virtual screening campaigns against *M. tuberculosis*.

The development of knowledge-based methods requires a large datasets. Unfortunately, data about mycobacterial permeability properties of chemical compounds are hardly available. However, as in most of the cases a compound must permeate the mycobacterial cell wall to show antimycobacterial activity, it is reasonable to infer an ability to pass this barrier for compounds active against mycobacteria. In 2010, Ekins and colleagues developed a collaborative database (CDD TB) of >200,000 compounds which had been tested for antibiotic activity against *M. tuberculosis* [184]. Over 3,800 structures showed growth inhibition of $\geq$90% at a concentration of 10 μM. Most likely, these compounds have sufficient permeability to be active against *M. tuberculosis* and may,

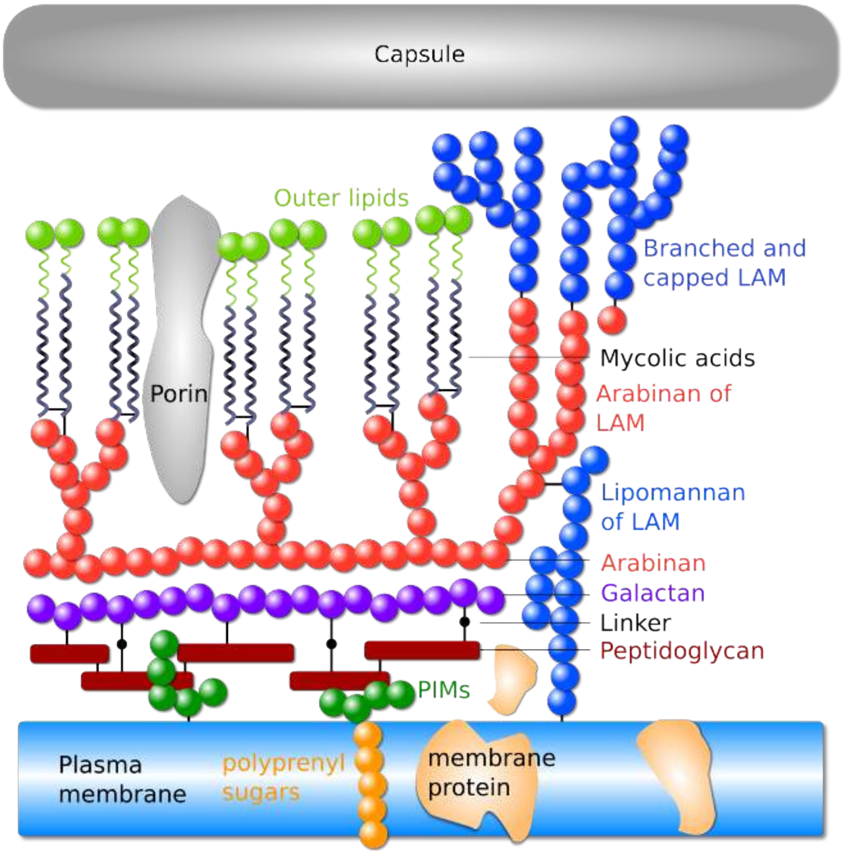


Figure 8.1 Schematic drawing of the *M. tuberculosis* cell wall. LAM: Lipoarabinomannan; PIMs: phosphatidylinositol mannosides. Figure adapted and redrawn from [180].

thus, be used as a knowledge base for analyzing permeability-determining features. Accordingly, an extensive data mining approach based on the physico-chemical properties of this dataset can be performed with the subsequent development of a regression model. This approach was followed herein, leading to the knowledge-based classification system MycPermCheck [3], as described in Chapter 9. An application of this tool to explore the permeability space of *M. tuberculosis* is the subject of Chapter 10.

8.2 Molecular descriptors

8.2.1 Molecular descriptors in drug design

"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." With these words, Todeschini and Consonni introduce a definition of the molecular descriptor [185]. In other words, a descriptor is any kind of reproducible information extracted

from a chemical structure. Generally, the information can be of experimental origin (e.g. logP, pK_a, dipole moment) or of theoretical origin (e.g. calculated logP, number of heavy atoms, polar surface area). The information encoded in theoretical descriptors is manifold. Depending on the underlying algorithm, calculated parameters can base on structural data of different dimensions. One-dimensional parameters include the number of hydrogen bond donors/acceptors or calculated logP; two-dimensional parameters comprise, for example, graph-theoretical diameters or topological surface areas; typical three-dimensional descriptors are the volume or solvent-accessible surface area of a geometrical representation of the chemical structure. Whereas some descriptors are straightforward to calculate and interpret (e.g., number of carbon atoms), others are based on more sophisticated underlying models (e.g., topological polar surface area). In general, physico-chemical descriptors can be used to extract numerical data from chemical structures and find similarities and differences in large datasets of chemical compounds.

The importance of physico-chemical properties and how they govern the pharmacokinetic and pharmacodynamic behavior of drugs is well described in the literature [186–191]. For instance, it is known that the polar surface area (PSA) and the number of rotatable bonds influence the bioavailability of a compound [188]. Furthermore, oral absorption is hindered by high molecular weight (MW) and a logP larger than 5 [186]. Nonetheless, molecular obesity is often a result of drug discovery projects, in which potency is driven by increasing the MW and lipophilicity of compounds, but ultimately leads to high attrition rates in clinical trials [189, 192]. Although the relations of the physico-chemical composition of compounds to solubility and membrane permeability are generally well understood [186–188], it still remains unclear which molecular properties specifically contribute to mycobacterial cell wall permeability.

8.2.2 Descriptor calculation software

There are numerous software packages that calculate countless molecular descriptors based on structural chemical data. A list of selected descriptor calculation packages can be accessed at the free online resource of molecular descriptors by Roberto Todeschini.[1] The proprietary software QikProp of the Maestro Suite (Version 3.4, Schrödinger, LLC, New York, NY, 2011) is designed to process large databases of chemical structures and create a collection of molecular parameters. These encompass physico-chemical properties (such as number of hydrogen bond acceptors and various surface areas), as well as ADMET parameters (such as Lipinski Rule of Five violations [186] or predicted

[1] http://www.moleculardescriptors.eu/softwares/softwares.htm, accessed July, 2015.

IC_{50} value for blockage of hERG K^+ channels). The software, thus, supports rational drug design by identifying molecules with unwanted properties, which can then be dismissed in early stages of drug development to avoid high costs associated with wet lab experiments. All 51 descriptors calculated by QikProp are summarized in the QikProp Manual [193]. In 2007, an independent group of researchers assessed the accuracy of QikProp predictions with respect to selected physico-chemical and ADMET descriptors [194]. Their findings showed very good correlations of experimental data to calculated logP (octanol/water partition coefficient), logS (solubility), dipole moment and ionization potential (IP). Also with respect to the ADME parameters Caco-2 and MDCK cell permeability modeling the gut-blood and blood-brain-barrier, respectively, good results could be achieved [194–197]. The modules predicting hERG K^+ channel blockage and CNS activity, however, did not yield convincing results [194].

Whereas QikProp is part of the proprietary software suite Maestro (Schrödinger, LLC, New York, NY, 2011), the PaDEL-Descriptor package [198] is entirely open-source. The current version of PaDEL (2.21) is able to calculate 1875 different descriptors and 12 types of fingerprints. Descriptors include straightforward atom, bond or ring counts as well as validated prediction methods, e.g., XlogP [199] or the topological PSA [200]. The software is available free of charge from the website of the author [198].[2]

8.3 Statistical methods

8.3.1 Principal component analysis

First described by Karl Pearson [201], Principal Component Analysis (PCA) forms the basis of many analyses of multivariate data [202]. Its primary goal is the simplification of complicated data and the revelation of underlying patterns, which are often concealed [203]. Hence, the data is modified to provide more intuitive access for interpretation.

The method uses an orthogonal transformation of multidimensional data to find the principal components (PC) with the highest variance possible (Figure 8.2). Thus, the first principal component, which is a linear combination of the dataset variables, exhibits the highest variance and hence the highest information content. The remaining principal components are defined according to the same criterion, except that they are restrained to being orthogonal to all previously defined principal components. The resulting principal components are the eigenvectors of the covariance matrix of the examined dataset [204].

[2] http://www.yapcwsoft.com/dd/padeldescriptor/

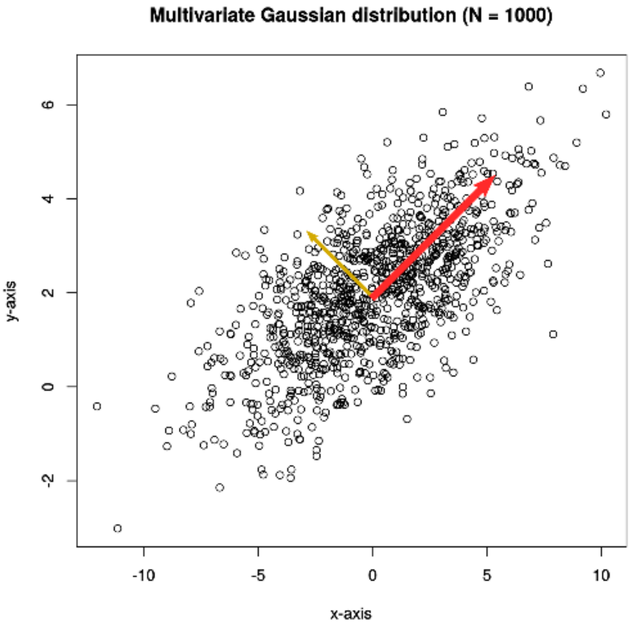


Figure 8.2 Multivariate Gaussian distribution. The vectors schematically illustrate the eigenvectors of the covariance matrix.

The principal component space is always less or equal in dimensions to the original data. However, since a principal component is always defined by maximum variance, the information content may drop quickly for PCs of a higher number, depending on the data. Although the question of how many principal components to include in an analysis is an ongoing debate with no definite answer [205], the most important PC is always the first.

8.3.2 Logistic regression

A logistic regression is used to predict the probability of a positive binary outcome based on one or multiple predictor variables. The logistic function $\sigma(z)$ is as follows:

$$\sigma(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}} \tag{8.1}$$

Here, z is a linear combination of the explanatory variables $(x_1, x_2, ...)$:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \tag{8.2}$$

Thus, instead of hard cutoffs in binary classification (negative/positive), a gradual function is considered. The larger the shift in the distribution of the two datasets in their

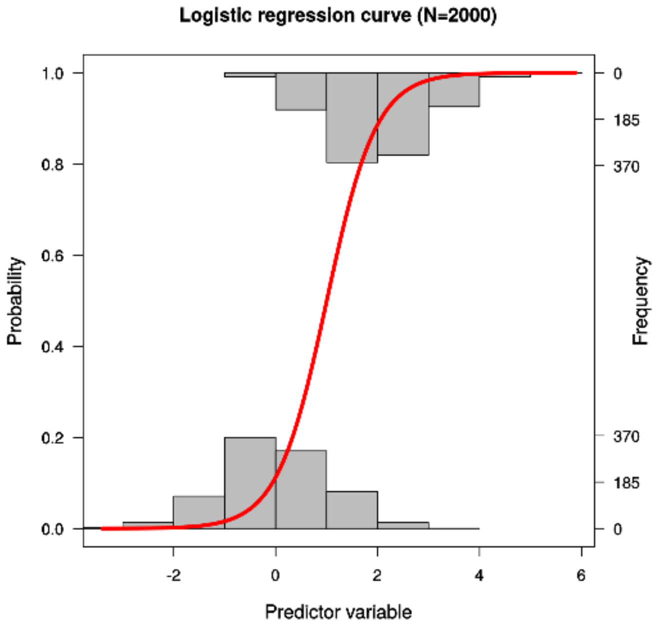


Figure 8.3 Logistic regression analysis on simulated data. Each group (N = 1000) was sampled randomly from a normal distribution centered at 2 and 0, respectively. Histograms represent the distributions of the groups, the red curve illustrates the logistic regression function $1/(1 + exp[-(-2.066 + 2.060 \cdot x)])$.

explanatory variables, the steeper is the logistic regression function. Figure 8.3 provides a logistic regression analysis on simulated data. The groups (N = 1000) were drawn randomly from a Gaussian distribution centered at 2 (positive group) and 0 (negative group), respectively (as shown in Figure 8.2). The red curve represents the logistic regression curve $1/(1 + exp[-(-2.066 + 2.060 \cdot x)])$ modeling the shift in distribution between the positive and the negative group with highly significant slope and intercept ($p \ll 0.001$).

Logistic regression can be combined with a previous PCA. In this case, the regression function is trained on coordinates from the PC space (cf. Chapter 9).

8.3.3 Receiver operating characteristic

A receiver operating characteristic (ROC) is a statistical method to assess the quality of a model with binomial classification [206]. Observations are sorted according to the predictive variable and examined top to bottom, while a varying threshold distinguishes the two classes. The observations are then compared to their actual affiliation. While walking through the sorted observations, the curve grows along the y-axis for each observation of the positive group and along the x-axis for each representative of the negative group that is passed (Figure 8.4).

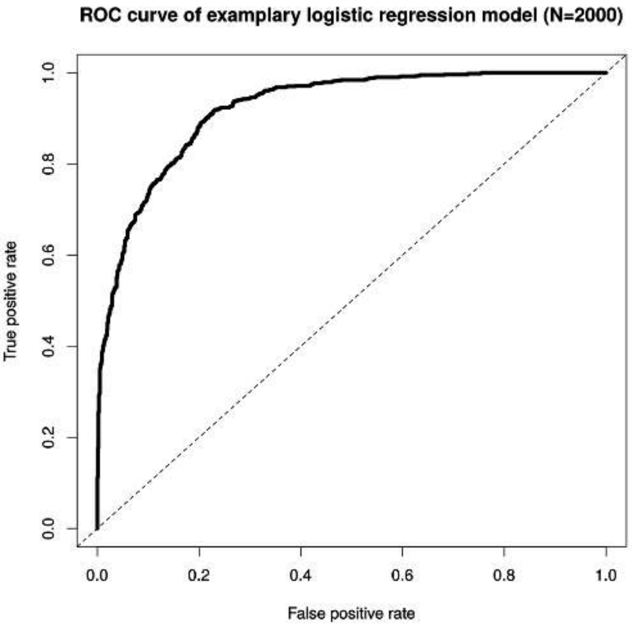


Figure 8.4 ROC curve of exemplary logistic regression of simulated data. Groups were drawn randomly from normal distributions centered at 2 and 0, respectively (cf. Figure 8.3). The dashed line illustrates a random model.

A perfect separation of the two groups would manifest in a ROC curve touching the top left corner of the plot, i.e. all positives are on one side and all negatives are on the other side of a certain threshold.

A helpful measure for the interpretation of a ROC curve is the early enrichment of true positives (graph rises to the top left corner), i.e. a very high true positive rate at a low false positive rate. In the ROC analysis of the simulated data, an enrichment of 73.3% of all positives can be observed at a false positive rate of 10.0% (cf. Figure 8.4). Another approach is the calculation of the area under the curve (AUC) [207], which–as a scalar value between 0 and 1–is easy to compare between multiple models. The simulated model (cf. Figures 8.3 and 8.4) achieves a very high AUC of 0.922. An AUC of 0.5 would correspond to a random model along the diagonal of the ROC plot, i.e. an equal increase of true and false positives.

8.4 Analysis tools

All calculations and statistical analyses in Part II of this thesis were conducted using the statistical framework R and the associated plug-ins vegan, popbio, pheatmap and ChemMineR [130, 133, 208–210]. Trajectory analyses were carried out with VMD 1.9.1

and the incorporated extension RMSD Trajectory Tool [140]. Visualizations were created with PyMOL [142]. 2D-RMSD plots were drawn with a tailored python script by Raphael Dives (University of Würzburg).

Chapter 9

MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules

The contents of this chapter have been published in the Oxford University Press journal *Bioinformatics* in 2013 [3]. The publication has been modified in layout to fit the style of this thesis. Moreover, the supporting information of the publication and previously not shown data have been incorporated into the chapter. Since October 2014, version 1.1 of MycPermCheck with modified PCA scaling of input variables is online. Accordingly, evaluation of the model was repeated using MycPermCheck 1.1. Therefore, numerical results regarding the evaluation dataset may vary with respect to the *Bioinformatics* publication. Additionally, a section describing the stand-alone command-line version of MycPermCheck was included. The theoretical background of this work is explained in Chapter 8.

9.1 Datasets

The MLSMR dataset [211] of the CDD TB database [184, 212] was filtered for compounds that showed a mycobacterial growth inhibition of $\geq$90% at 10 μM and a molecular weight <500 Dalton. This step reduced the total number of considered molecules to 3815 chemical structures. All compounds were converted to 3D structures with the program Corina (available from Molecular Networks GmbH, Erlangen, Germany) [213]. These structures were processed with the tool LigPrep (Version 2.3, Schrödinger, LLC, New York, NY, 2009) for protonation (at pH 7.0 ± 2.0), stereoisomerization, tautomerization and subsequent energy minimization. Physico-chemical descriptors were then calculated for each molecule with Schrödinger QikProp (Version 3.4, Schrödinger, LLC, New York, NY, 2011). Compounds with incomplete descriptor data were removed, leaving 3727 structures. This dataset is hereinafter referred to as *Actives*.

The foundation of this work is the assumption that a compound must sufficiently well permeate the mycobacterial cell envelope (consisting of cell wall, periplasm and inner membrane) to unleash its effect within the target cell. Therefore, the dataset *Actives* can be classified as 'permeable' (i.e. the corresponding compounds have sufficient permeability to be active). Far more difficult is the generation of a sufficiently large 'impermeable' (negative) dataset, as only few studies regarding the permeability of mycobacteria are available (e.g. Refs [183, 214–216]). Simply taking the inactive compounds from *M. tuberculosis* activity tests is obviously not possible, as a lack of permeability may not be the only reason for inactivity. This issue can be addressed by collecting compounds that are active against *M. tuberculosis* targets in target-based (e.g. enzymatic) assays, but inactive in a whole-cell *M. tuberculosis* assay. This approach was indeed followed herein to generate a validation dataset (cf. Section 9.4). The number of compounds obtainable by this way is, however, by far not sufficient for data mining and training-set generation. Accordingly, randomly drawn datasets of drug-like small molecules were used as 'negative' data. These should allow to determine whether the 'permeable' substances show any significant differences with respect to random drug-like compounds. For this purpose, the drug-like subset of the ZINC database ([217], version ZINC12) was processed in the same manner as the *Actives*. Thereby, an extensive table of physico-chemical properties of a randomly distributed dataset of drug-like molecules was obtained. This dataset is hereinafter referred to as *ZINC*. An overview of the used datasets is given in Table 9.1.

To obtain information about the diversity of the *Actives* dataset an all vs. all similarity matrix was generated based on the atom pair similarity of these compounds using ChemMineR [210, 218] (Figure 9.1).

The distribution of all measured atom pair similarities within the *Actives* dataset lies at $\sim$0.2, suggesting that the majority of these compounds have a very low similarity among each other. Furthermore, the heatmap only very rarely shows similarity values of $\geq$0.5 (yellow to red). Altogether, this underlines the diversity of the used positive dataset.

9.2 Descriptor selection and visualization

Pairwise Mann-Whitney-U-tests of *Actives* against *ZINC* (several sets of 100 randomly chosen structures each) were performed for each of the 51 QikProp descriptors. The tests showed consistent results regarding their P-values. Figure 9.2 depicts the distribution of the calculated P-values using the R package BioNet [219, 220] for one representative test set including a fitted beta and uniform distribution. Under the null hypothesis, the P-values are uniformly distributed representing only noise. The remaining part of the P-values describes the signal distribution. The fitted beta-uniform-mixture model [221] shows a strong signal of significant differences in the physico-chemical properties of *Actives* and *ZINC*. Based on a descriptive representation of the 51 distributions (data

Table 9.1 Summary of used datasets. Reproduced and updated from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, pp. 61–68, with permission by Oxford University Press.

	Name	Size	Description
Training sets	*Actives*	3 727	CDD TB compounds with mycobacterial growth inhibition of >90% at a concentration of 10 μM
	ZINC	18 988 507	Prepared structures of ZINC drug-like database
Test sets	*Permeables*	771	Compounds gathered from ChEMBL with antimycobacterial activity (absent in the dataset Actives)
	Impermeables	21	Compounds with in-vitro activity against *M. tuberculosis* targets, but without activity in mycobacterial whole-cell assays
	InhA inhibitors	19	Antimycobacterial InhA inhibitors from selected publications (cf. text for references) absent in the dataset Actives

not shown) and a common understanding of physico-chemical descriptors for drug development, five QikProp descriptors ($P < 0.001$) were further considered:

- FOSA: The hydrophobic part of the solvent accessible surface area (saturated carbon and attached hydrogen atoms);
- QPlogPo.w: The logarithm of the calculated octanol/water partition coefficient (hereinafter called logP);
- PISA: The π-interacting part of the solvent accessible surface area;
- accptHB: The number of H-bond acceptors;
- glob: The generic spherical surface to molecule surface ratio.

Other common molecular descriptors (e.g. molecular weight or the number of H-bond donors) were not considered for model derivation, mostly due to insufficient differences

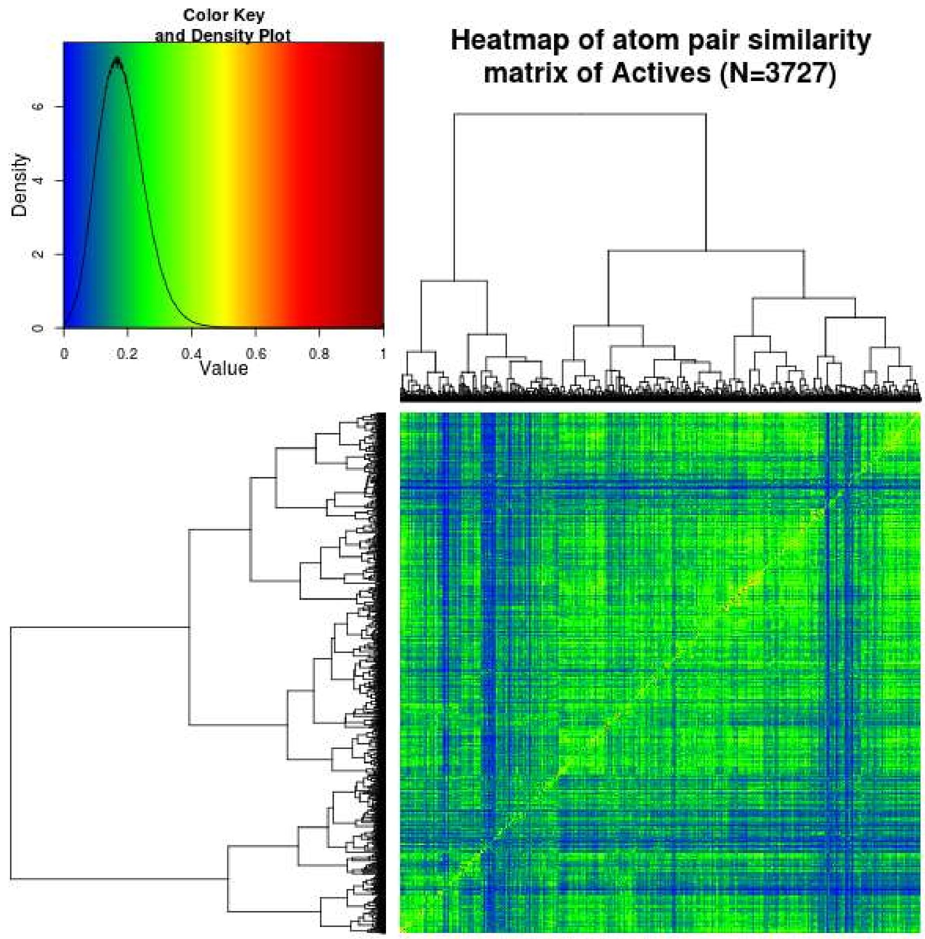


Figure 9.1 Heatmap of atom pair similarity matrix of dataset *Actives* and distribution of similarity values. Reproduced from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, Supplement S1, with permission by Oxford University Press.

between the datasets with respect to these descriptors. To increase statistical significance, in all of the following randomly chosen datasets, the size was increased to 1000 per group. Figure 9.3 illustrates the distribution of the five selected descriptors for a representative randomly chosen test set of *Actives* as well as a randomly chosen *ZINC*-test set of equal size. The non-overlapping box notches show significant differences in the medians of the distributions of these five descriptors for the two datasets.

A first impression whether a potential new inhibitor might show descriptor values typical for permeable compounds can be gained from a comparison with the distribution of the descriptors in the *Actives* dataset. For this purpose, four borders have been defined to better delimit the physico-chemical space of the permeable substances: upper, up, low and lower (Table 9.2). The borders up and low are defined by the 75 and 25% quantile of the training dataset, respectively. Upper and lower represent 75 and 25% quantile ± half the interquartile range, respectively.

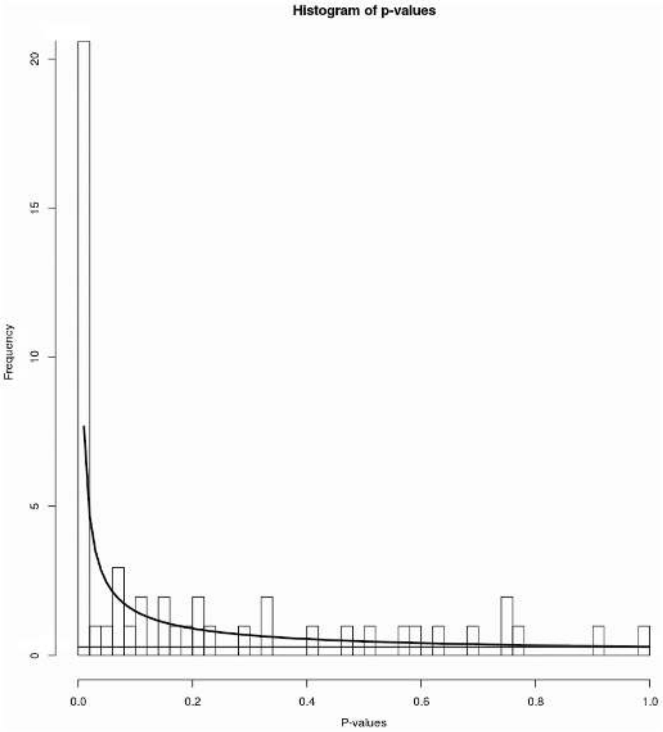


Figure 9.2 Histogram of the P-values of 51 pairwise Mann-Whitney-U-tests of each descriptor of *Actives* against *ZINC*. The black curve indicates the fitted beta distribution (signal + noise), and the gray line indicates the fitted uniformly distributed baseline of noise. A clear deviation of the empirical P-values from the fitted noise distribution is observed, suggesting a strong information content in the differences of *Actives* and *ZINC*. Reproduced from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, pp. 61–68, with permission by Oxford University Press.

Table 9.2 Borders of the five chosen descriptors based on the distributions of the descriptors in the complete *Actives* dataset, as further described in the text. FOSA and PISA are measured in Å^2. Reproduced from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, pp. 61–68, with permission by Oxford University Press.

	FOSA	logP	PISA	accptHB	glob
upper	362.95	5.329	430.66	7.125	0.861
up	272.23	4.479	355.49	6.000	0.839
low	90.80	2.779	205.16	3.750	0.794
lower	0.09	1.929	129.99	2.625	0.772

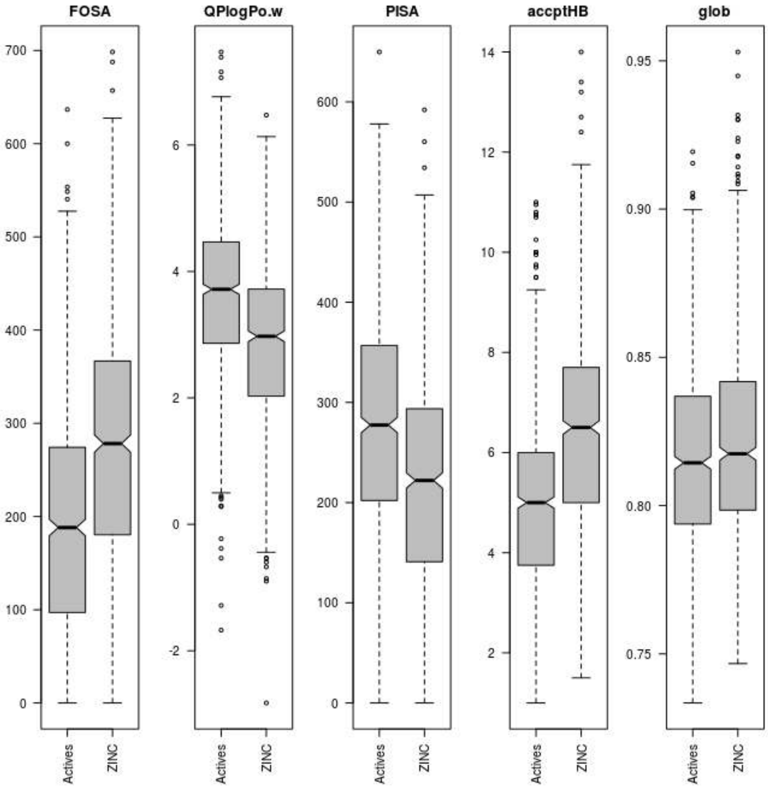


Figure 9.3 Boxplots of the five chosen chemical descriptors. Boxes indicate the interquartile range (25–75% quantile). Black lines indicate the median of each distribution. The whiskers extend to values 1.5 times the interquartile range from the box. A highly significant difference in the medians of *Actives* versus *ZINC* is observed for each descriptor, indicated by non-overlapping notches ($P < 0.001$, Mann-Whitney-U-tests). Reproduced from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, pp. 61–68, with permission by Oxford University Press.

9.3 PCA and logistic regression method

Although the value mapping of each descriptor for a compound of interest is useful for later interpretation of results, a reliable prediction of the permeability cannot be achieved this way. Thus, the permeability prediction approach is based on multivariate statistics. First, 25 principal component analyses (PCAs) were performed based on the five chosen descriptors using random test sets of 1000 permeable substances of the *Actives* dataset and 1000 substances of the *ZINC* dataset. Then, the resulting coordinates were projected to the first principal component. All PCAs showed coherent results: each time a one-dimensional representation of principal component 1 (PC1) showed the best splitting of the two groups *Actives* and *ZINC*. Thus, by reducing the multi-dimensional information space to only the first principal component (42.4% information content, histograms in Figure 9.4), it is possible to achieve a maximum separation of these two groups. All PCA analyses were performed with the vegan R package [208].

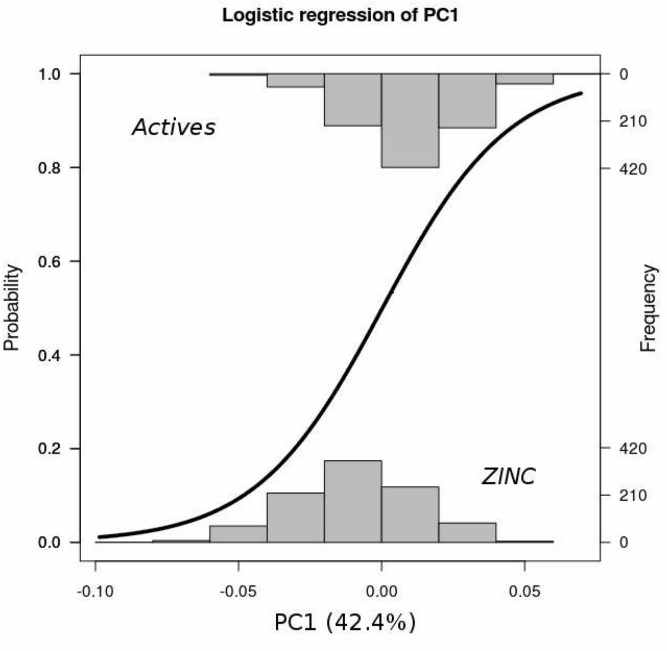


Figure 9.4 Logistic regression model of PCA coordinate 1 (42.4% information content) of 1000 compounds each of the *Actives* and the *ZINC* training sets. The histogram at the top of the plot shows the distribution of the *Actives* dataset, whereas the histogram at the bottom represents the samples from the *ZINC* dataset. A clear separation of the two distributions can be observed. The black curve indicates the calculated logistic regression model based on PC1 of the priorly performed PCA. It is quantified according to the 'Probability' axis, indicating the final result of MycPermCheck. Reproduced from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, pp. 61–68, with permission by Oxford University Press.

The PC1 coordinates of one representative PCA were then used to generate a logistic regression model (Figure 9.4; figure created with the R package popbio [209]) using R [220]. The obtained logistic regression function follows:

$$P(z) = \frac{1}{1 + e^{-z}} \tag{9.1}$$

$$\text{with } z = f(x) = \beta \cdot x \tag{9.2}$$

with a highly significant regression coefficient $\beta = 45.187$ ($P < 2 \cdot 10^{-16}$). The variable x represents the input PC1 coordinate of a given compound. This logistic regression model is the core of the MycPermCheck [3] tool for estimating the likelihood of permeability. During the permeability prediction procedure, any potential inhibitor of interest is processed in MycPermCheck by these steps: (i) first, the principal component coordinates are calculated according to the existing PCA of the training data, (ii) then, the coordinate of PC1 is used as input (x) for the logistic regression model. As a result, the user receives a calculated probability $[0 < P(z) < 1]$ of a compound to be classified as permeable.

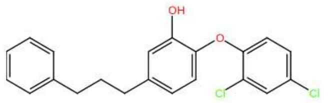
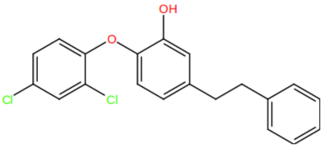
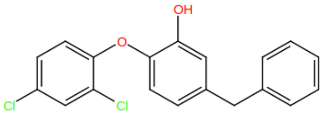
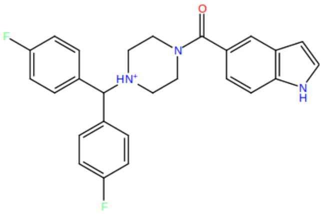
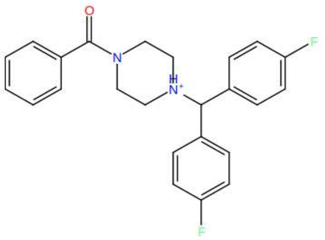
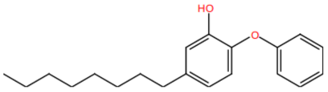
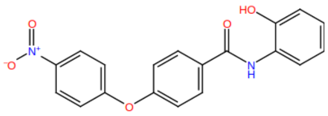
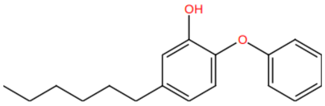
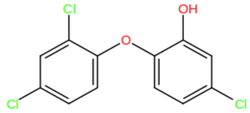
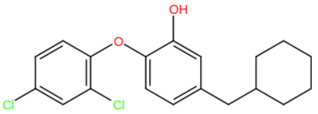
9.4 Evaluation

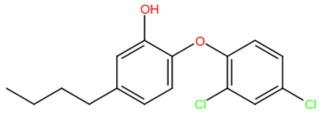
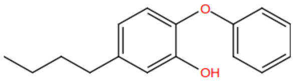
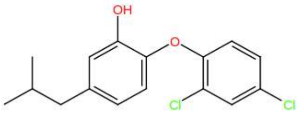
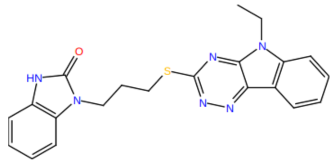
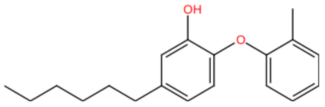
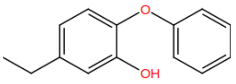
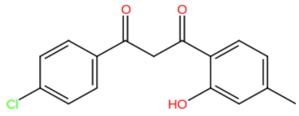
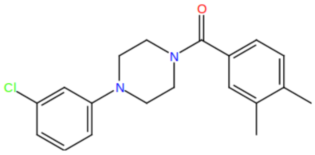
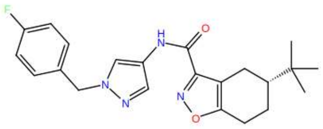
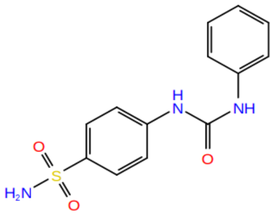
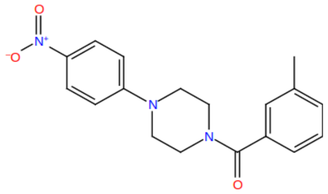
For evaluation of the logistic regression model, the ChEMBL database [222] was browsed for antimycobacterially active compounds with a minimal inhibitory concentration (MIC) $\leq 10\ \mu M$, yielding a total of 771 permeable structures (*Permeables*) absent in the training dataset *Actives*. The compounds were prepared the same way as the compounds of the training set (3D conversion, protonation, stereoisomerization, tautomerization and energy minimization). After descriptor calculation with QikProp, MycPermCheck was used with the option Calculate Mean of all Isomeric Forms (as described in the next section). The calculated permeability probabilities show a median of 0.664 (±0.139 median absolute deviation). Hence, MycPermCheck yields valid predictions for these antimycobacterial and, thus, permeable substances.

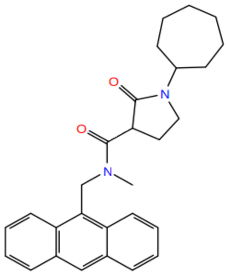
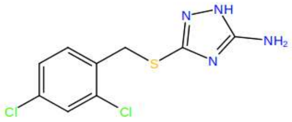
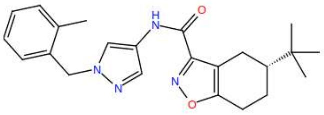
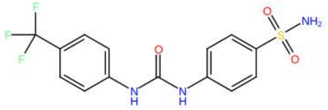
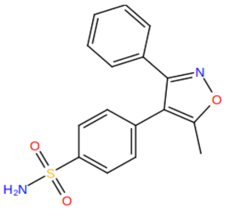
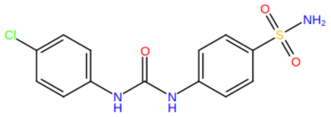
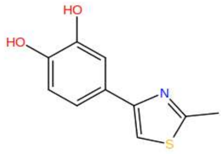
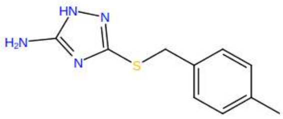
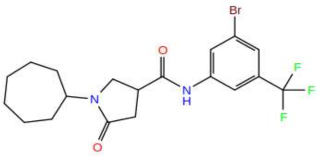
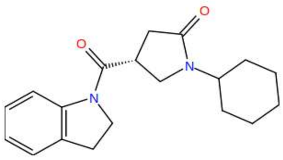
To further evaluate MycPermCheck with biological real-life data, the intersection of two different datasets was generated: first, the CDD TB [184] was filtered for substances which show <10% antimycobacterial activity at $10\ \mu M$, yielding >190 000 compounds. Simultaneously, the ChEMBL database [222] was browsed for assays against *M. tuberculosis* targets and filtered for structures marked as active within the database according to their half-maximal inhibitory concentration (IC_{50} value) or enzymatic inhibition constant (K_i value) of $\leq 10\ \mu M$. On the basis of their International Chemical Identifiers (InChI strings), the intersection of the two datasets was established, yielding 22 compounds. As an additional filter criterion, an all versus all similarity matrix was generated based on the atom pair similarity of these compounds using ChemMineR [210, 218]. A compound showing >80% similarity to another compound was removed. One structure (CHEMBL592712) was affected, yielding a final number of 21 compounds with low IC_{50} or K_i values (i.e. activity against an *M. tuberculosis* target in an *in vitro* enzyme assay), but without antimycobacterial activity. Based on the assumption that the most likely reason for the inactivity of these compounds against *M. tuberculosis* is their inability to penetrate the mycobacterial cell wall, this dataset should be a collection of impermeable compounds. The 21 compounds (*Impermeables*) (see structures IM1-IM21 Table 9.3) were prepared the same way as the *Permeables* and the compounds of the training set. The calculated QikProp descriptors were then processed by MycPermCheck, again with the option Calculate Mean of all Isomeric Forms. The obtained permeability probabilities show a median of 0.444 (±0.172 median absolute deviation).

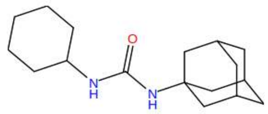
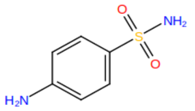
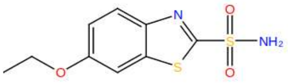
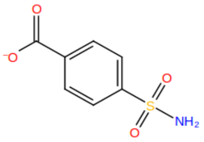
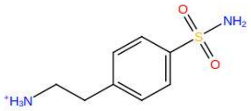
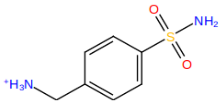
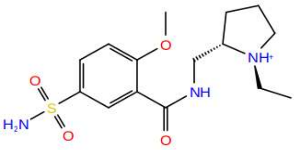
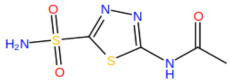
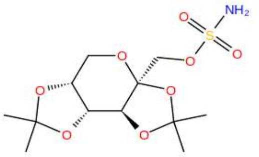
Fifty combined datasets of the 21 *Impermeables* and 21 randomly chosen *Permeables* were then created to perform a multiple Receiver Operating Characteristic (ROC) analysis with the R package ROCR [227] (Figure 9.5a). The single ROC curves were averaged by true-positive rate (black curve) as well as by threshold (colored curve). The color scale illustrated in Figure 9.5 represents the actual permeability probability that is used

Table 9.3 MycPermCheck output for evaluation dataset of 19 InhA inhibitors and 21 impermeable compounds (sorted by probability), including an illustration of the compound. Compounds taken from [a]-[223], [b]-[224, 225], [c]-[64], [d]-[226], [e]-[45]. Modified from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, Supplement S2, with permission by Oxford University Press. Probability values are calculated with MycPermCheck 1.1.

ID	Name	Structure	Prob.	FOSA	logP	PISA	accpt HB	glob
P1	**compound 26 [a]**		**0.923**	90.91	6.676	407.06	1.250	0.812
P2	**compound 25 [a]**		**0.916**	60.03	6.280	403.63	1.250	0.826
P3	**compound 24 [a]**		**0.900**	39.57	5.839	386.66	1.250	0.843
P4	**compound p4 [b]**		**0.855**	102.78	5.415	460.50	5.000	0.807
P5	**compound p6 [b]**		**0.835**	110.25	5.088	450.07	5.000	0.811
P6	**8PP [c]**		**0.804**	325.95	5.951	297.64	1.250	0.794
P7	**VH07 [d]**		**0.801**	0.00	3.180	451.79	4.750	0.812
P8	**6PP [c]**		**0.782**	260.55	5.175	297.64	1.250	0.817
P9	**Triclosan [c]**		**0.774**	0.00	4.738	225.62	1.250	0.898
P10	**compound 7 [a]**		**0.761**	243.11	6.046	200.54	1.250	0.833

P11	**compound 10 [a]**		**0.756**	195.67	5.383	206.91	1.250	0.827
P12	**4PP [c]**		**0.755**	195.62	4.402	297.64	1.250	0.843
P13	**compound 11 [a]**		**0.744**	190.29	5.296	200.89	1.250	0.836
IM1	**CHEMBL 589101**		**0.740**	202.63	4.755	349.54	5.000	0.785
P14	**PT70 [e]**		**0.736**	320.71	5.424	254.59	1.250	0.824
P15	**2PP [c]**		**0.719**	130.20	3.522	297.90	1.250	0.872
P16	**VH04 [d]**		**0.631**	110.86	3.313	255.57	3.750	0.838
IM2	**CHEMBL 239673**		**0.598**	280.71	4.636	222.90	4.000	0.825
IM3	**CHEMBL 259507**		**0.551**	317.17	4.607	231.05	5.500	0.798
IM4	**CHEMBL 25600**		**0.543**	0.00	1.401	319.56	6.000	0.837
P17	**compound a6 [b]**		**0.535**	220.69	3.188	259.43	5.000	0.820

P18	**compound p67 [b]**		**0.531**	376.17	4.184	313.73	6.000	0.812
IM5	**CHEMBL 569750**		**0.516**	25.86	2.221	140.55	3.000	0.858
IM6	**CHEMBL 412059**		**0.501**	373.89	4.579	223.28	5.500	0.808
IM7	**CHEMBL 1337519**		**0.487**	0.30	1.152	271.78	6.500	0.812
IM8	**CHEMBL 865**		**0.479**	80.49	1.669	287.40	6.000	0.847
IM9	**CHEMBL 1446150**		**0.462**	0.00	0.667	289.68	6.500	0.822
IM10	**CHEMBL 495123**		**0.445**	93.80	1.690	167.58	3.000	0.882
IM11	**CHEMBL 568651**		**0.444**	118.96	1.625	165.38	3.000	0.858
P19	**compound d12 [b]**		**0.388**	321.51	4.208	89.26	5.500	0.810
IM12	**CHEMBL 217499**		**0.273**	371.34	2.420	163.93	6.000	0.843

IM13	**CHEMBL 242255**		**0.263**	533.94	3.469	0.00	2.000	0.830
IM14	**CHEMBL 1875592**		**0.229**	0.00	-0.799	149.87	5.500	0.914
IM15	**CHEMBL 1410342**		**0.226**	139.42	0.462	131.75	6.750	0.852
IM16	**CHEMBL 1894686**		**0.210**	0.00	-0.579	132.67	6.500	0.904
IM17	**CHEMBL 7087**		**0.199**	83.52	-0.795	127.20	5.500	0.886
IM18	**CHEMBL 419**		**0.193**	48.96	-1.123	133.03	5.500	0.900
IM19	**CHEMBL 196677**		**0.124**	349.17	0.670	76.84	8.250	0.817
IM20	**CHEMBL 1884503**		**0.091**	81.53	-1.285	10.99	8.000	0.887
IM21	**CHEMBL 220492**		**0.078**	373.44	0.755	0.00	8.250	0.882

as a sliding threshold for establishing the true- versus false-positive rate and, hence, the ROC curve. The average ROC curve shows a fast increase of the true-positive rate without producing an equivalent amount of false positives, indicating that reliable and practically useful results can be obtained with MycPermCheck for randomly selected permeable molecules. At a false-positive rate of $\beta = 10\%$ (a specificity of $1 - \beta = 90\%$), a true-positive rate (sensitivity) of 63.9 ± 10.4% (SD) is achieved (i.e. about two-thirds of all true positives already appear at this cut-off). A less strict false-positive rate of 25% yields a higher sensitivity of 70.2 ± 10.0% (SD). At a permeability probability cut-off of 0.596, a specificity of 90% is obtained, whereas a cut-off of 0.524 matches a specificity of 75%. These two cut-offs (rounded to 0.60 and 0.52, respectively) form the basis of the traffic-lights color code of the web program output, as described below. Altogether, the single ROC-analyses of randomly drawn datasets display a high average area under the curve (AUC) of 0.786 ± 0.072.

For evaluation of the logistic regression model for three well-studied classes of inhibitors of the mycobacterial enzyme enoyl acyl carrier protein reductase (InhA), the chemical structures of 20 mycobacterial inhibitors (not present in the dataset *Actives*) were extracted from the literature [45, 224–226, 228, 229]. Again, the structures were filtered for atom pair similarity <80%. After removing one compound (5PP), 19 mycobacterial inhibitors remained in this test set (see structures P1-P19 in Table 9.3). These inhibitors cover a broad chemical range from triclosan and its derivatives (diphenyl ethers) to arylamides and pyrrolidine carboxamides. Again, the compounds were prepared as before (3D conversion, protonation, stereoisomerization, tautomerization, energy minimization and QikProp descriptor calculation). The calculated permeability probabilities show a median value of 0.761 (±0.043 median absolute deviation). Therefore, MycPermCheck yields valid predictions for these permeable substances.

A second ROC analysis was performed for a combined dataset of these 19 active substances and the previously detected 21 impermeable compounds (Figure 9.5b). Regarding the highly active InhA inhibitors, MycPermCheck shows an even faster increase of true-positive results than for the randomized evaluation test sets with an AUC of 0.945. At a false-positive rate of $\beta = 10\%$, a true-positive rate (sensitivity) of 84.2% is achieved, whereas a false-positive rate of 25% corresponds to a sensitivity of 94.7%. The actual permeability cut-offs at these false-positive rates are very similar to those of the multiple ROC analysis of the randomized evaluation test sets (0.598 and 0.517, respectively). These results illustrate that MycPermCheck is applicable on inhibitors of the *M. tuberculosis* target InhA.

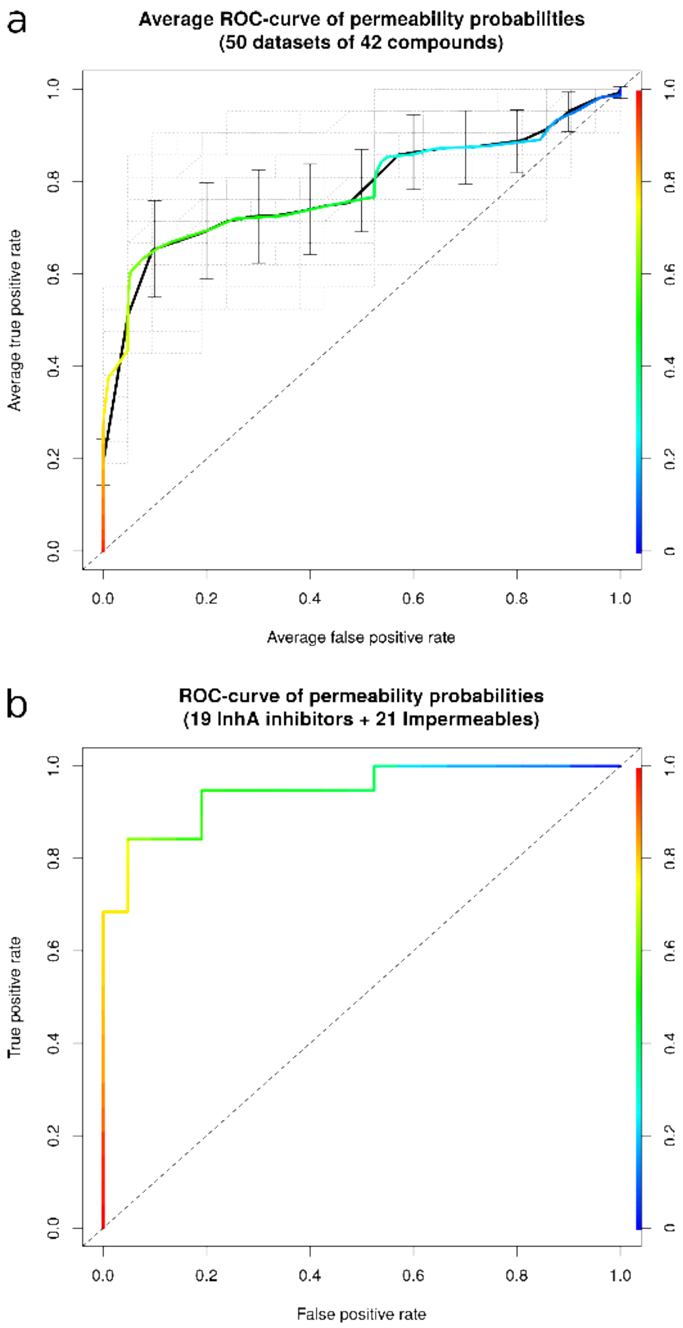


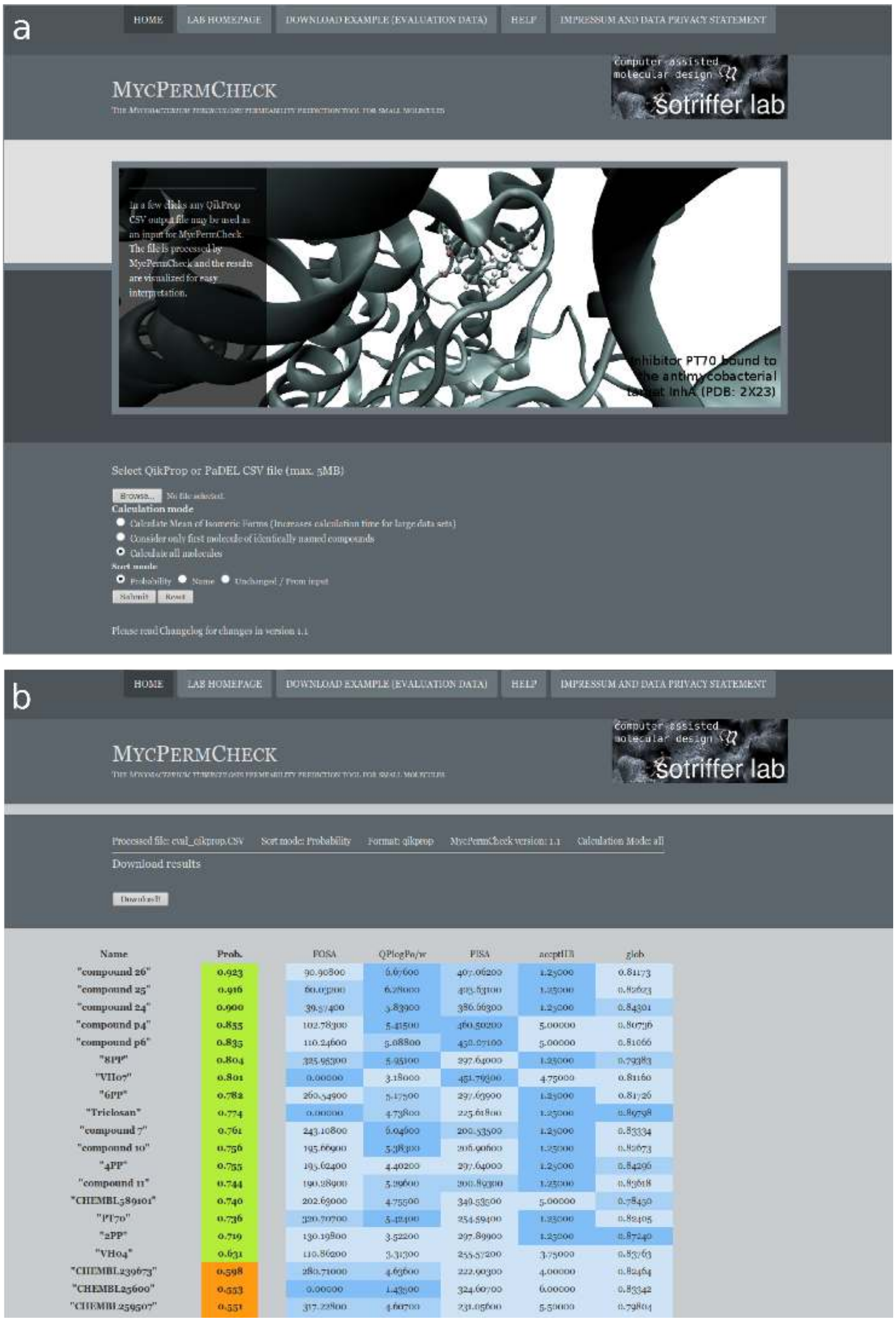
Figure 9.5 **(a)** Multiple ROC analysis of calculated permeability probabilities for 50 datasets of 21 randomly selected *Permeables* and 21 *Impermeables*. The true-positive rate is plotted against the false-positive rate for a rising threshold of the calculated permeability probability (indicated by the color scale). The gray dashed curves illustrate the single ROC analyses. The thick black curve shows the ROC curve averaged by true-positive rate, whereas the thick colored curve represents the calculated average by threshold. Error bars indicate the standard deviation of the true-positives–averaged curve. The dashed angle bisector illustrates a uniform rise of the true-positive and false-positive rate, equivalent to a random model. **(b)** ROC analysis of calculated permeability probabilities for the evaluation dataset of 19 InhA inhibitors and 21 *Impermeables*. The true-positive rate is plotted against the false-positive rate for a rising threshold of the calculated permeability probability (indicated by the color scale). The dashed angle bisector illustrates the random model. Both ROC curves show a clear enrichment of permeable compounds at the top of the permeability-ranked list. Reproduced with updated values (MycPermCheck 1.1) from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, pp. 61–68, with permission by Oxford University Press.

9.5 Implementation

MycPermCheck is a freely accessible online tool. It is programmed entirely in perl, making use of the perl CGI-package for displaying browser contents. Usage of the program begins by accessing the start-up screen (Figure 9.6a). Here, the input data [a QikProp comma-separated values (CSV) file] must be chosen using the browse function of the website. The selection "Calculate Mean of Isomeric Forms" defines whether all 'isomeric' forms of a compound (i.e. tautomers, protomers, stereoisomers, conformers, etc.; indicated and recognized by the same molecule name in the QikProp CSV file) should be considered and averaged. Alternative options are: (i) only the first representative is used for the calculation or (ii) all molecules are processed separately. Clicking the Submit button submits the job to the instant calculation of the permeability probabilities.

Within few seconds, a list of the submitted compounds appears as a result, sorted either by the calculated permeability probability (default), by the compound name or in an unchanged order (optional selection on submission). The list shows the calculated permeability probability in the first column after the compound name, followed by the single descriptor values (Figure 9.6b; detailed list of evaluation data including structures see Table 9.3). For the single descriptor values, blue-scale colors are assigned based on the borders defined in Table 1: (i) if the value lies between the borders up and low, this state is colored light blue; (ii) a value between the borders up and upper or low and lower, respectively, is colored blue; (iii) a value below lower or above upper is illustrated by a dark blue coloring. This graphical illustration represents the chemical similarity of a given compound to the training dataset *Actives* in terms of the five most relevant descriptors (see colored descriptor values in Figure 9.6b). In contrast, the quality of each result is rated according to a simple and intuitive traffic-lights system: for highlighting the permeability probability, two borders have been defined based on the ROC analyses of the evaluation dataset (Figure 9.5). The first cut-off of 0.60 corresponds to a false-positive rate of ∼10%. Results with probabilities above this value (>0.60) are marked green. The second cut-off of 0.52 corresponds to a false-positive rate of ∼25%. Results above this threshold are marked orange. Probabilities below 0.52 are colored red. A download function can be used to save all results in a CSV file for further processing by the user.

Besides the use of Maestro QikProp descriptors for estimating the permeability probability, MycPermCheck is also able to process CSV output files of the open-source java descriptor calculation package PaDEL-Descriptor [198]. A complete evaluation of the PCA and regression model for PaDEL descriptor input is presented below.



MYCPERMCHECK

The Mycobacterium tuberculosis permeability prediction tool for small molecules

In a few clicks any QikProp CSV output file may be used as an input for MycPermCheck. The file is processed by MycPermCheck and the results are visualized for easy interpretation.

Inhibitor PT70 bound to the antimycobacterial target InhA (PDB: 2X23)

Select QikProp or PaDEL CSV file (max. 5MB)

No file selected.

Calculation mode

○ Calculate Mean of Isomeric Forms (Increases calculation time for large data sets)
○ Consider only first molecule of identically named compounds
● Calculate all molecules

Sort mode

● Probability ○ Name ○ Unchanged / From input

Please read Changelog for changes in version 1.1

MYCPERMCHECK

The Mycobacterium tuberculosis permeability prediction tool for small molecules

Processed file: eval_qikprop.CSV Sort mode: Probability Format: qikprop MycPermCheck version: 1.1 Calculation Mode: all

Download results

Name	Prob.	FOSA	QPlogPo/w	PISA	acceptHB	glob
"compound 26"	0.923	90.90800	6.67600	407.06200	1.25000	0.81173
"compound 25"	0.916	60.03200	6.28000	403.53100	1.25000	0.82623
"compound 24"	0.900	39.57400	3.83900	386.66300	1.25000	0.84301
"compound p4"	0.855	102.78300	5.41500	460.50200	5.00000	0.80736
"compound p6"	0.835	110.24600	5.08800	450.07100	5.00000	0.81066
"8PP"	0.804	325.95300	5.95100	297.64000	1.25000	0.79383
"VII07"	0.801	0.00000	3.18000	451.79300	4.75000	0.81160
"6PP"	0.782	260.54900	5.17500	297.63900	1.25000	0.81726
"Triclosan"	0.774	0.00000	4.73800	225.61800	1.25000	0.89798
"compound 7"	0.761	243.10800	6.04600	200.53500	1.25000	0.83334
"compound 10"	0.756	195.66900	5.38300	206.90600	1.25000	0.82673
"4PP"	0.755	193.62400	4.40200	297.64000	1.25000	0.84296
"compound 11"	0.744	190.28900	5.29600	200.89300	1.25000	0.83618
"CHEMBL589101"	0.740	202.63000	4.75500	340.53500	5.00000	0.78450
"PT70"	0.736	320.70700	5.42400	254.59400	1.25000	0.82405
"2PP"	0.719	130.19800	3.52200	297.89900	1.25000	0.87240
"VH04"	0.631	110.86200	3.31300	255.57200	3.75000	0.83763
"CHEMBL239673"	0.598	280.71000	4.63600	222.90300	4.00000	0.82464
"CHEMBL25600"	0.553	0.00000	1.43500	324.60700	6.00000	0.83342
"CHEMBL259507"	0.551	317.22800	4.60700	231.05600	5.50000	0.79804

Figure 9.6 **(a)** Details of the start-up page of the MycPermCheck website. The mask at the bottom of the page is used to upload the input QikProp or PaDEL CSV file. The user can choose between three different calculation and sort modes. With a click on 'Submit', the user can upload the input file to the web server and start the calculation process. **(b)** Details of the results page of the MycPermCheck website. The lower half of the screen depicts the top of the calculated results table. The compounds with the highest permeability probabilities (green) are shown (sorted by probability). In the table, besides the permeability probability, the raw descriptor data of each compound are shown. The blue-scale color code illustrates the deviation of these data from the distribution of the *Actives* training set according to the borders defined in Table 9.2. The provided comma-separated text-file version of the results is accessible through the 'Download' button above the results table. Reproduced with updated values (MycPermCheck 1.1) from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, pp. 61–68, with permission by Oxford University Press.

The program is accessible under the following website:
http://www.mycpermcheck.aksotriffer.pharmazie.uni-wuerzburg.de

9.5.1 Stand-alone version

Besides the graphical online service, a stand-alone command-line interface (CLI) version of MycPermCheck was created. The tool is programmed entirely in perl and is ready-to-use without requiring additional perl packages. It encompasses the same functions regarding calculation and sorting as the web tool. Whereas input files for remote MycPermCheck calculations are limited to 5 MB due to server capacity issues, the local stand-alone installation of MycPermCheck is able to process millions of molecules at once. The open source code of the CLI version is attached in Appendix B.

9.6 PaDEL-Descriptor

9.6.1 Descriptor selection

As an alternative to the QikProp model, which is based on the commercial Schrödinger software, an additional regression model was derived. This model is based solely on descriptors of the open-source descriptor calculation package PaDEL-Descriptor [198]. Since PaDEL 2.7 provides descriptors in an overwhelming quantity (863 descriptors), a first approach of selecting a feasible combination of descriptors was the search for QikProp-equivalents within the PaDEL inventory. In this attempt, the following descriptors were chosen ($p < 0.001$, Mann-Whitney-U tests of *Actives* against *ZINC*):

- HybRatio: The ratio of sp^3 to sp^2 hybridized carbon atoms;
- XlogP: The logarithm of the calculated octanol/water partition coefficient;
- LOBMAX: The maximum length-over-breadth coefficient of the molecule.

The number of H-bond acceptors (nHBAcc) as well as the hydrophobic solvent accessible surface area (THSA) calculated by the PaDEL algorithms did not show significant differences in the two datasets. Hence, these two descriptors were ignored for deriving a PaDEL-Descriptor based model. To further improve the regression model pairwise Mann-Whitney-U-tests of *Actives* against *ZINC* (several sets of 100 randomly chosen structures each) were performed for each descriptor (Figure 9.7). Due to the removal of incomplete descriptor data, the *Actives* dataset comprises 3475 compounds

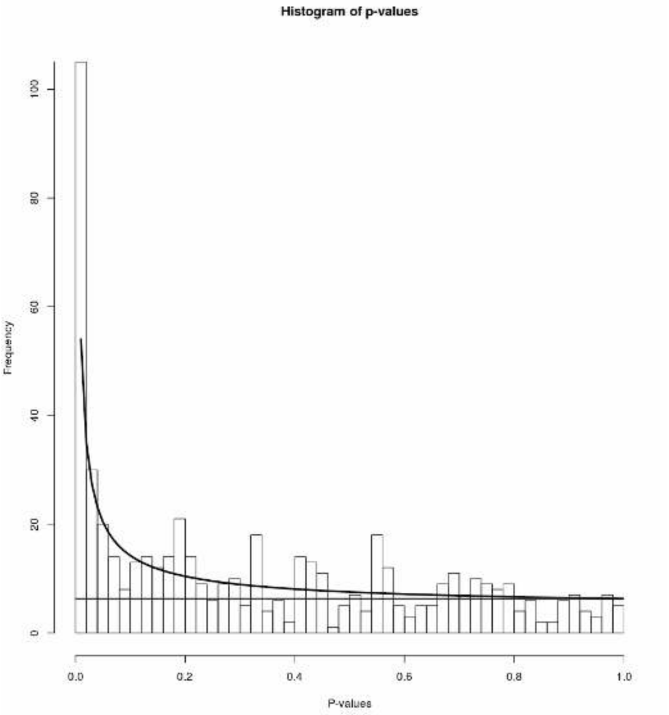


Figure 9.7 Histogram of the p-values of 786 pairwise Mann-Whitney-U-tests of each descriptor of *Actives* against *ZINC* for one test set of 100 compounds per group. The black curve indicates the fitted beta distribution, the grey line indicates the fitted baseline of noise. A clear deviation of the empirical p-values from the fitted noise distribution is observed, suggesting a strong signal in the differences of *Actives* and *ZINC*. Reproduced from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, Supplement S3, with permission by Oxford University Press.

for the PaDEL-based model derivation, and tests were performed for 786 of 863 descriptors. Again, the fitted beta-uniform-mixture model shows a strong signal of significant differences in the physico-chemical properties of *Actives* and *ZINC*. Several new PaDEL-specific descriptors were tested for significant shifts in the distributions ($p < 0.001$, Mann-Whitney-U tests), leading to two additional descriptors used in the PaDEL-Descriptor based regression model:

- C2SP2: Number of doubly bound carbon atoms bound to two other carbon atoms;
- TPSA: Sum of hydrophilic solvent accessible surface areas.

9.6.2 PCA and logistic regression method

Again, Principal Component Analyses were performed on the five chosen PaDEL descriptors using random test sets of 1000 permeable substances of the *Actives* dataset and 1000 substances of the *ZINC* dataset. The first principal component showed the

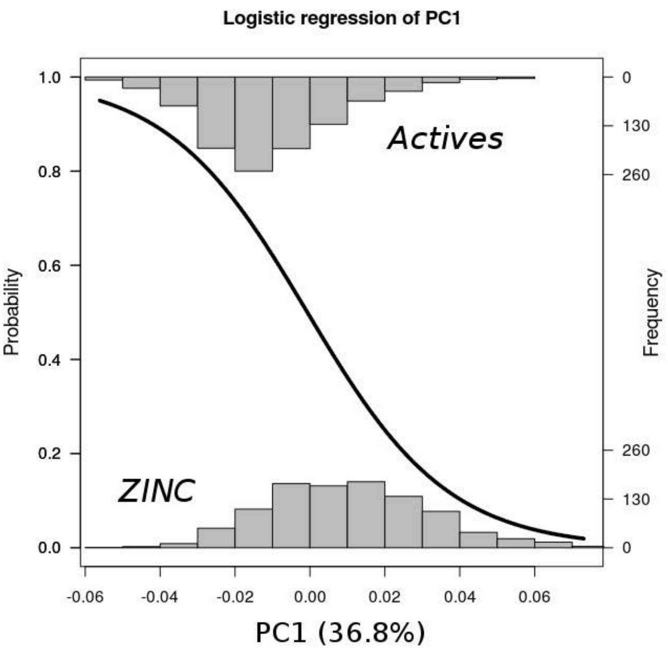


Figure 9.8 Logistic regression model of PCA coordinate 1 (36.8% information content) of 1000 compounds each of the *Actives* and the *ZINC* training sets. The histogram at the top of the plot shows the distribution of the *Actives* dataset, while the histogram at the bottom represents the samples from the *ZINC* dataset. A clear separation of the two distributions can be observed. The black curve indicates the calculated logistic regression model based on principal component 1 of the priorly performed PCA. It is quantified according to the "Probability" axis, indicating the final result of MycPermCheck. Reproduced from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, Supplement S3, with permission by Oxford University Press.

best separation of the two datasets. The PC1 coordinates of one representative PCA were then used to generate a logistic regression model (Figure 9.8; figure created with the R package *popbio* [209]). The logistic regression function follows:

$$P(z) = \frac{1}{1+e^{-z}} \tag{9.3}$$

$$\text{with } z = f(x) = \beta \cdot x \tag{9.4}$$

with a highly significant regression coefficient $\beta = -52.943$ ($p < 2 \cdot 10^{-16}$) and x being the input PC1 coordinate of a compound.

9.6.3 Evaluation

For evaluation purposes, the same three datasets were used as for the QikProp-model (cf. Chapter 9.4). These include (i) 656 compounds with complete descriptor data of the *Permeables* dataset (permeable substances absent in the training dataset extracted from the ChEMBL database [222]), (ii) 21 presumably impermeable compounds, and (iii) 19 highly active antimycobacterial InhA inhibitors [45, 224–226, 228, 229].

For the PaDEL-Descriptor based model, the active compounds gathered from ChEMBL achieved a median permeability probability of 0.715 (±0.116 median absolute deviation). The 21 impermeable compounds showed a median permeability probability of 0.453 (±0.200 median absolute deviation). The 19 antimycobacterial InhA inhibitors obtained high values with a median probability of 0.792 (±0.069 median absolute deviation).

ROC-analyses were performed for 50 combined datasets of the 21 *Impermeables* and 21 again randomly chosen *Permeables* with the R package ROCR [227] (Figure 9.9). The clear enrichment of true positive results indicates that MycPermCheck is able to provide valid results on the basis of PaDEL descriptors. At a specificity of 90%, a true positive rate of 58.4% ± 8.6% (SD) can be achieved. For a decreased specificity of 75%, the true positive rate rises to over 67.0% ± 8.5% (SD). At a permeability probability cutoff of 0.691, a specificity of 90% is achieved, while a cutoff of 0.620 corresponds to a specificity of 75%. These cutoffs (rounded to 0.69 and 0.62, respectively) form the basis of the traffic-lights color code for the PaDEL-Descriptor based model. Altogether, the ROC-analysis shows a high average AUC of 0.819 ± 0.049.

In a second ROC-analysis, the combined dataset of 19 InhA inhibitors and 21 *impermeables* shows an even faster increase of the true positive rate with an AUC of 0.945. A specificity of 90% corresponds to a true positive rate (sensitivity) of 84.2%, whereas a 75% threshold comprises 94.7% of all true positives (Figure 9.9).

9.6.4 Implementation

The PaDEL-Descriptor based regression model is implemented into both the MycPermCheck online tool and the CLI version. The format of the input file is detected automatically and the corresponding model is loaded for all following calculations.

9.7 Discussion

MycPermCheck is an intuitively accessible online tool for knowledge-based estimation of the permeability of potential antimycobacterial compounds with respect to the *M. tuberculosis* cell wall. The program is based on a chemoinformatic data-mining approach without any assumptions regarding the uptake mechanism. It is, hence, generally applicable to drug-like compounds with a molecular weight <500 Dalton. With statistical significance, a training set of permeable compounds (*Actives*) could be delimited from

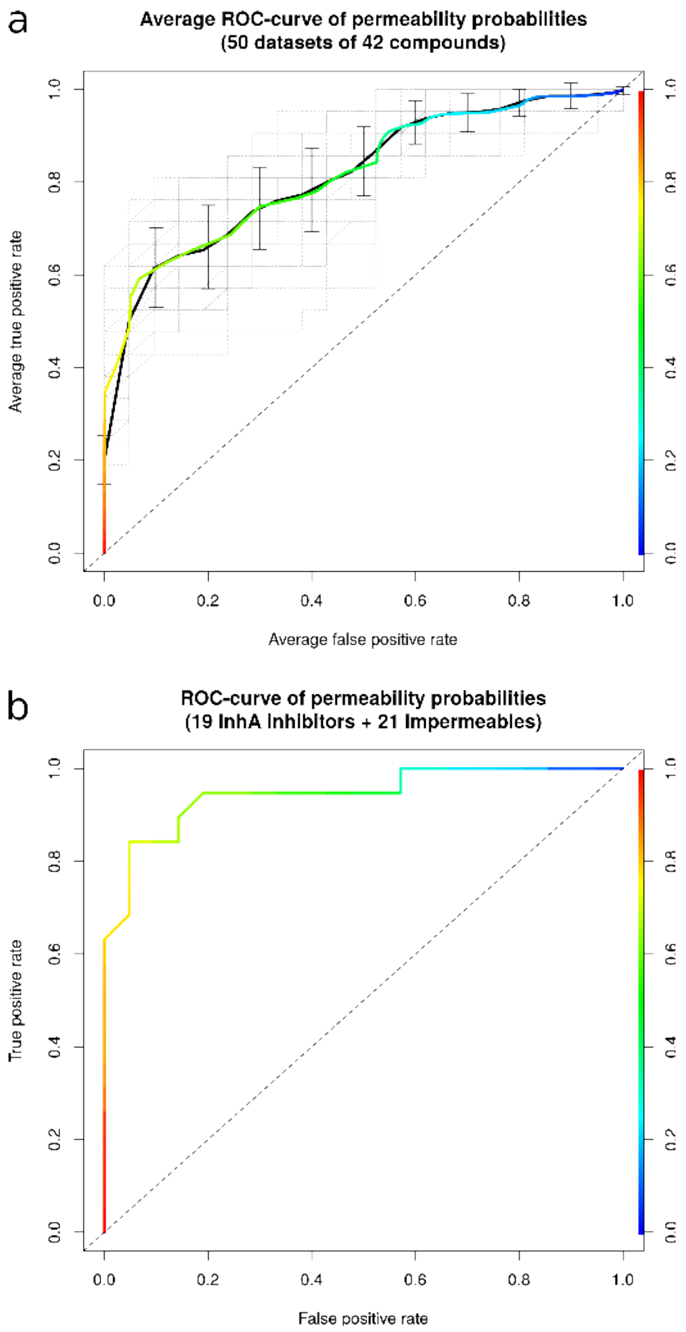


Figure 9.9 **(a)** Multiple ROC-analysis of calculated permeability probabilities for 50 datasets of 21 randomly selected *Permeables* and 21 *Impermeables*. The true positive rate is plotted against the false positive rate for a rising threshold of the calculated permeability probability (indicated by the color-scale). The gray, dashed curves illustrate the single ROC analyses. The thick, black curve shows the ROC-curve averaged by true positive rates, while the thick, colored curve represents the calculated average by threshold. Error bars indicate the standard deviation of the true positives-averaged curve. The dashed angle bisector illustrates a uniform rise of the true positive and false positive rate, equivalent to a random model. **(b)** ROC-analysis of calculated permeability probabilities for the evaluation dataset of 19 InhA inhibitors and 21 *Impermeables*. The true positive rate is plotted against the false positive rate for a rising threshold of the calculated permeability probability (indicated by the color-scale). The dashed angle bisector illustrates the random model. Both ROC-curves show a clear enrichment of permeable compounds at the top of the permeability-ranked list. Reproduced with updated values (MycPermCheck 1.1) from: Merget *et al.*, MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules, *Bioinformatics*, 2013, 29:1, Supplement S3, with permission by Oxford University Press.

randomly distributed drug-like molecules based on five physico-chemical descriptors in a principal component analysis. Based on the resulting first principal component, a logistic-regression model for estimating the permeability probability could be derived. Thereby, instead of hard cut-offs for molecular descriptor interpretation (as, for example, in Lipinski's Rule of 5 [186]), a 'more realistic and gradated description of the continuum of compound quality' is obtained, an advantage recently pointed out by Bickerton and colleagues in the context of their quantitative estimate of drug likeness [191].

MycPermCheck was multiply tested on 50 evaluation datasets of 21 permeable compounds and a set of 21 impermeable compounds. With a standard deviation of true positives of ∼10% for a specificity of both 90 and 75%, the average of the multiple ROC curves shows a robust prediction for randomly selected permeable compounds (cf. Figure 9.5a). Moreover, MycPermCheck was tested on 19 highly active InhA inhibitors and 21 impermeable compounds, leading to a very good enrichment of the permeable compounds in the range of the highest probability values and of the impermeable compounds in the range of the lowest probability values (cf. Figure 9.5b and Table 9.3). In fact, among the top 13 compounds ($0.923 \geq P \geq 0.744$), no false positive is found. Moreover, the 10 lowest ranked compounds ($0.273 \geq P \geq 0.078$) are all true negatives, i.e. *impermeables*. Comparing the molecular structures and descriptor values of these two groups of top-ranked and lowest-ranked compounds indicates that with only few exceptions permeable compounds are characterized by a high PISA to FOSA ratio (i.e. the π-interacting surface area is generally much larger than the hydrophobic surface area), a logP of >4 and an accptHB value <2. In contrast, the 10 lowest-ranked impermeable compounds show frequently larger FOSA than PISA values, have low logP values (often <1) and generally an accptHB value of >5. The compounds with the highest permeability probability show indeed at least two aromatic ring systems to which small to moderately sized hydrophobic substituents and few H-bond acceptors are attached. The impermeable compounds, instead, have often only one (if any) aromatic ring system and–despite a significant hydrophobic surface area–a higher polarity and more H-bond acceptors (cf. compounds IM14-IM21 in Table 9.3).

These observations may provide some guidelines for ensuring mycobacterial permeability of designed compounds. Nevertheless, it is also clear that looking at single parameters only is not sufficient. In fact, simply aiming for descriptor values that are within the 'up' and 'low' borders defined in Table 9.2 does not ensure a high permeability probability. For example, compound IM6 shows three descriptor values within the light-blue range, one within the blue (logP) and only one within the dark-blue range (FOSA), yet the probability is only 0.501, and the compound is indeed impermeable. Conversely, compounds with high probabilities may also show descriptor values in the blue or dark-blue range, as, for example, the top three compounds P1, P2 and P3 (cf. Table 9.3). Thus, a "one-dimensional" view focused at single descriptor values and their univariate statistics is indeed of little value. Instead, the correct combination and relative weighting of molecular properties is essential, as incorporated in the logistic regression model based on the first principal component: PISA should be larger than FOSA, logP not too small (rather >3) and accptHB not too large (rather <5); the descriptor glob plays a minor role in modulating the probability.

To ensure compatibility with open-source software, MycPermCheck was also trained on a combination of five PaDEL descriptors [198]. Comparison of the two models on basis of the AUC leads to the conclusion that the QikProp model performs slightly worse than its PaDEL-counterpart (AUC: 0.786 vs. 0.819). However, early enrichment of true positives is also an important parameter of model quality. Here, the model based on QikProp descriptors shows a higher enrichment of true positives at a specificity of 90%, compared to the PaDEL-model (63.9% vs. 58.4%). Since the difference of 5.5% is, however, well within the standard deviations of ∼10%, the models exhibit a comparable performance.

Although the validation results of MycPermCheck illustrate a high predictivity, it is also clear that an absolute accuracy should not be expected. Considering the false positive IM1, which obtains a probability value >0.7, the lack of permeability cannot be explained in terms of the descriptor values, as they fit the general trends observed for the truly permeable compounds. It should be kept in mind, however, that compounds of the *Impermeables* validation set are actually only assumed to be impermeable because of a lack of antimycobacterial activity despite an inhibitory effect in an *in vitro* target-based assay. Obviously, this lack of activity may also have other reasons than mere impermeability. Examples include the activity of efflux pumps and the *in vivo* degradation/inactivation of a compound. Accordingly, it cannot be ruled out that IM1 is indeed permeable, but inactive for other reasons. Considering false negatives, a few cases are observed as well. Of the 19 permeable compounds in the validation set, P18 and P19 obtain probability values <0.6. Although these compounds have larger FOSA than PISA values and 5–6 hydrogen bond acceptors, they show antimycobacterial activity.

These examples illustrate the limits of the approach, which (i) does not make any distinction with respect to the uptake mechanism and (ii) is not based on a dataset of experimentally proven impermeable compounds. Clearly, a sufficiently large dataset of compounds with known uptake mechanism or confirmed impermeability would be highly advantageous, both for the derivation of improved models as well as for a more reliable validation of the current model. Given the lack of such a dataset, MycPermCheck is an attempt to make best use of the available knowledge base.

Despite these shortcomings, MycPermCheck is expected to be of significant practical value for any (virtual) screening endeavor dedicated to antimycobacterial drug design.

The validation results indicate that a clear enrichment of potentially permeable compounds and a highly reliable filtering of impermeable compounds (with $P < 0.1$) can be achieved with this, to our knowledge, unique approach. Accordingly, MycPermCheck may serve as an additional selection criterion on virtual screening and as a utility for increasing the likelihood of obtaining permeable antimycobacterial compounds.

Chapter 10

Permeability prediction, docking and MD simulations lead to suggestion of potential InhA inhibitors

Findings from both parts of this thesis were considered in conducting a screening for novel potential InhA inhibitors, as described in the following chapter. MycPermCheck 1.1 (cf. Chapter 9) was used for a large-scale permeability prediction against the mycobacterial cell wall of the entire ZINC12 drug-like database. After subsequent filtering steps with respect to ADMET properties predicted with QikProp, docking was carried out to reveal potential InhA inhibitors. Several derivatives of an initial hit compound were analyzed using MD and SMD simulations to assess aspects of protein and ligand stability (cf. Chapters 3 and 6), as well as maximum free energy changes of induced ligand extraction (cf. Chapters 5 and 6). This chapter, thus, shows key aspects of how the methodology presented in both parts of this thesis can support a potential screening campaign for antituberculars.

10.1 Exploration of the *M. tuberculosis* permeability space

10.1.1 Permeability prediction for the ZINC database

The drug-like subset of the ZINC12 [217] database was prepared for descriptor calculation in the process of generation of the MycPermCheck model (cf. Chapter 9.1): after downloading ZINC12 in the SMILES format from http://zinc.docking.org, all 13,205,607 SMILES strings were converted to the SDF format using OpenBabel [230]. Subsequently, the software Corina (available from Molecular Networks GmbH, Erlangen, Germany) [213] was used to convert the 2D molecules into three-dimensional structures. These structures were then protonated at pH 7.0 ± 2.0, stereoisomerized, tautomerized, and energetically minimized with the Schrödinger Maestro tool LigPrep (version 2.3, Schrödinger, LLC, New York, NY, 2009). Lastly, the resulting 19,296,744 3D structures were used for physico-chemical descriptor calculation with Schrödinger QikProp (Version 3.4, Schrödinger, LLC, New York, NY, 2011).

After dismissal of all permanently ionized compounds, 18,988,507 structures with complete QikProp descriptor data remained (negative dataset used in MycPermCheck model generation, cf. Chapter 9.1). The stand-alone version of MycPermCheck 1.1 (Appendix B) was used for permeability prediction for the structures with the option "Calculate Mean of All Isomeric Forms", resulting in permeability prediction data for 13,061,805 unique compounds. The ZINC IDs of all structures with a permeability probability of $\geq$0.900 were saved and again the corresponding SMILES and QikProp data extracted, yielding descriptor data for 21,576 substances.

10.1.2 Distributions of descriptor and ADMET data

The distributions of selected QikProp descriptor data were analyzed for structures classified as permeable. First, the descriptors used by MycPermCheck were considered (Figure 10.1). As expected, the permeable substances tend to have a high logP with a mean of 5.67 and a noticeable π-interacting surface area (PISA, avg. 525.28 Å^2). Accordingly, the hydrophobically interacting surface area (FOSA) is considerably lower (avg. 62.30 Å^2). The number of H-bond acceptors and the globularity (accptHB and glob) show averages of 3.73 and 0.79, respectively. Figure 10.2 shows the distributions of additional descriptors. It is notable that all descriptors show distributions within the range of 95% of known drugs (cf. Table 10.1; QikProp manual, Schrödinger Software Release 2014-1). However, this is not surprising, as the compounds stem from the drug-like subset of ZINC12.

Furthermore, several ADMET parameters were predicted with QikProp to gain insight into the pharmacokinetic and pharmacodynamic (PK/PD) properties of the permeable compounds (Table 10.2 and Figure 10.3). According to QikProp predictions, the compounds have very good ADMET properties regarding the permeability of cells and oral absorption. The only setback is a high predicted average hERG K$^+$ channel blockage.

10.1.3 Correlations in descriptor data

A cross-correlation matrix was generated to examine the inter-descriptor correlations (Figure 10.4). Two distinct groups of descriptors could be identified by heatmapping/-clustering (using complete-linkage) with high internal cross-correlations. Interestingly, these descriptors not only show high correlations within their own cluster, but generally also high negative correlations to descriptors of the respective other cluster.

Cluster 1 contains mostly ADMET parameters describing the oral absorption, $logS$, $logK_p$, hERG channel blockage, MDCK cell permeability, Caco-2-cell permeability, CNS

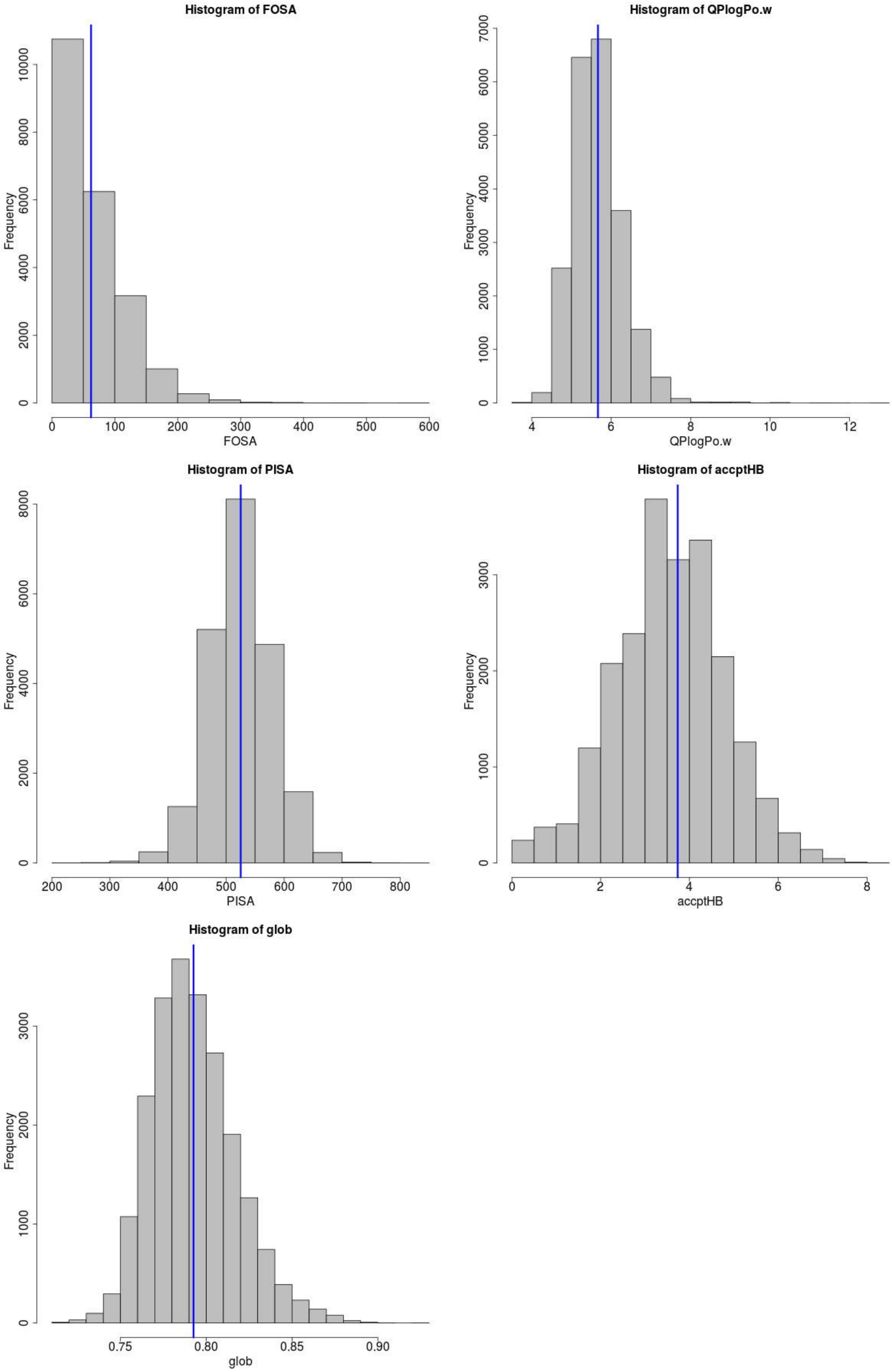


Figure 10.1 Distributions of QikProp descriptors used by MycPermCheck for the 21,576 compounds with predicted permeability probability ≥ 0.900. The blue lines indicate averages.

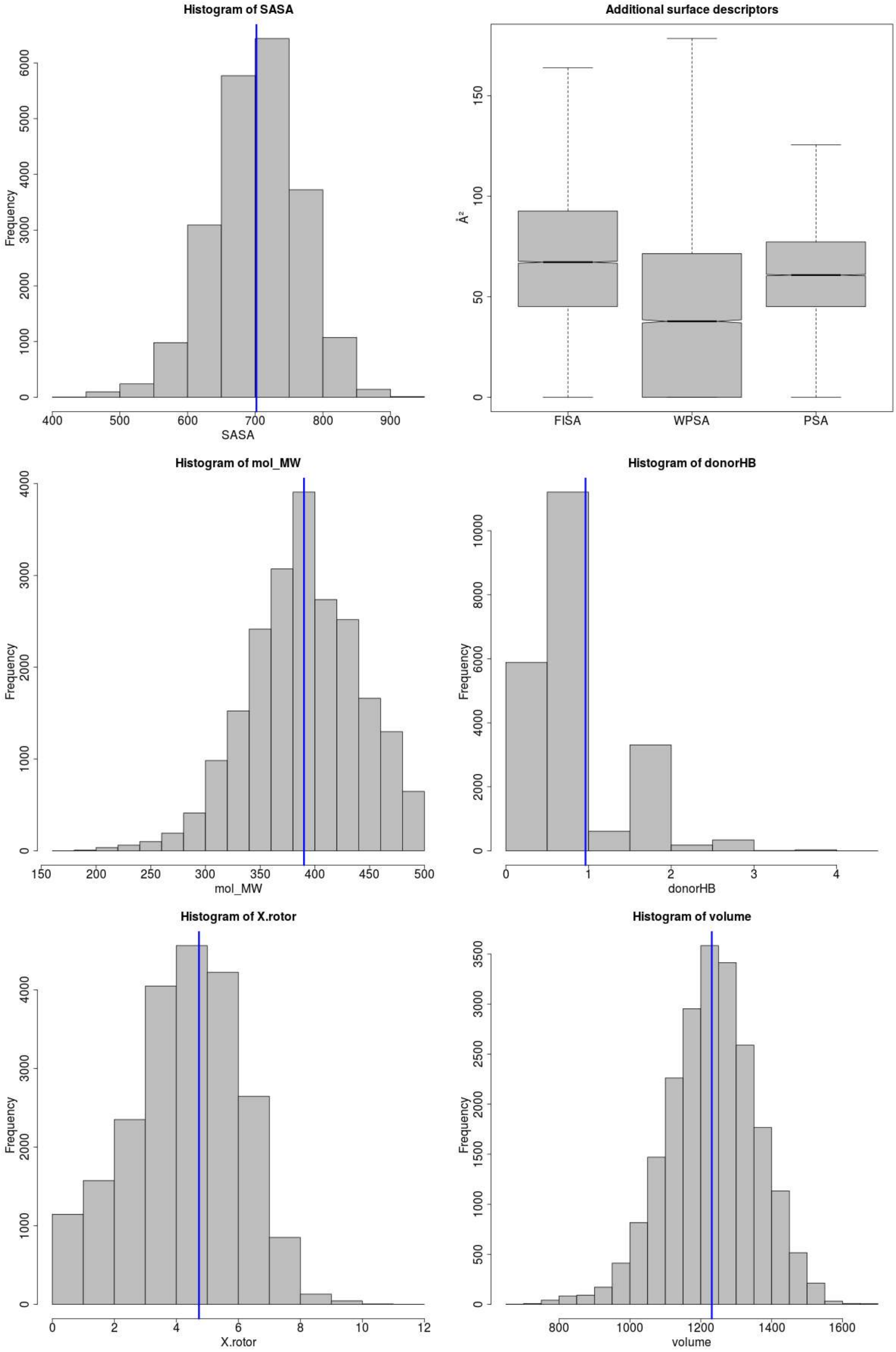


Figure 10.2 Distributions of additional descriptors for the 21,576 compounds with predicted permeability probability ≥ 0.900**.** The blue lines indicate averages. Cf. Table 10.1 for additional property descriptions.

Table 10.1 Additional descriptors. Range is taken from QikProp Manual (Schrödinger Software Release 2014-1).

Property or Descriptor	**Description**	**Range of 95% of known drugs**
SASA	Total solvent accessible surface area in Å^2 using a probe with a 1.4 Å radius	300.0 Å^2 – 1000.0 Å^2
FISA	Hydrophilic component of the SASA (SASA on N, O, H on heteroatoms, carbonyl C)	7.0 Å^2 – 330.0 Å^2
WPSA	Weakly polar component of the SASA (halogens, P, and S)	0.0 Å^2 – 175.0 Å^2
PSA	Van der Waals surface area of polar nitrogen and oxygen atoms and carbonyl carbon atoms	7.0 Å^2 – 200.0 Å^2
mol_MW	Molecular weight of the molecule	130.0 Da – 725.0 Da
donorHB	Estimated number of hydrogen bonds that would be donated by the solute to water molecules in an aqueous solution. Values are averages taken over a number of configurations, so they can be non-integer	0.0 – 6.0
#rotor	Number of non-trivial (not CX$_3$), non-hindered (not alkene, amide, small ring) rotatable bonds	0 – 15
volume	Total solvent-accessible volume in Å^3 using a probe with a 1.4 Å radius	500.0 Å^3 – 2000.0 Å^3

Table 10.2 ADMET descriptors. Range is taken from QikProp Manual (Schrödinger Software Release 2014-1.

Property or Descriptor	**Description**	**Range of 95% of known drugs**
QPlogHERG	predicted $logIC_{50}$ value for hERG K$^+$ channel blockage	concern below -5
QPPCaco	predicted apparent Caco-2-cell permeability to model gut-blood barrier	< 25 poor, > 500 great
QPlogBB	predicted brain/blood partition coefficient	-3.0 – 1.2
QPPMDCK	predicted apparent MDCK cell permeability to model blood-brain barrier	< 25 poor, > 500 great
QPlogKp	predicted skin permeability, $logK_p$	-8.0 – -1.0
PercentHumanOralAbsorption	predicted human oral absorption on 0 to 100% scale	$> 80\%$ is high, $< 25\%$ is poor

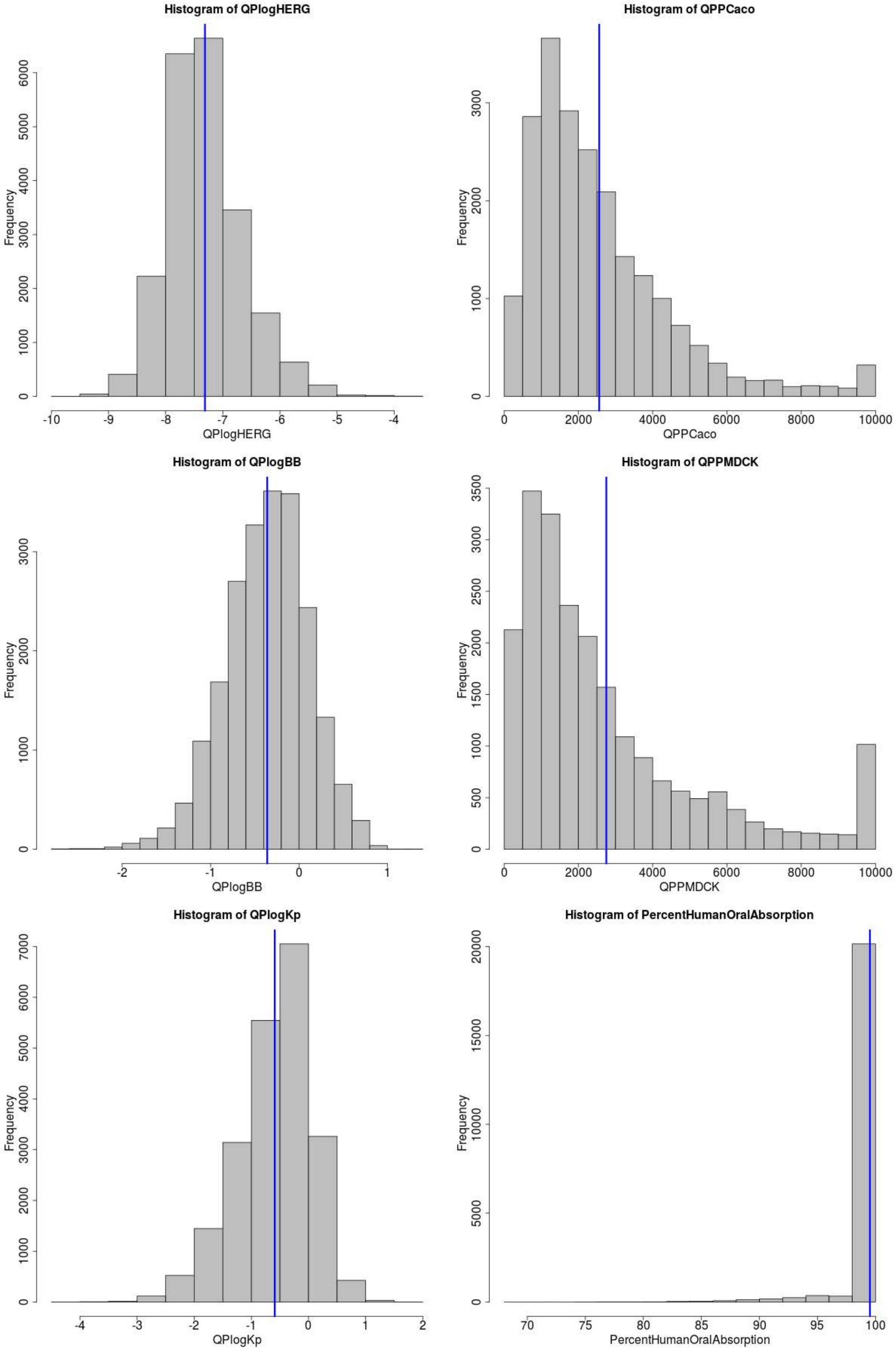


Figure 10.3 Distributions of ADMET descriptors of 21,576 compounds with predicted permeability probability $\geq$ **0.900.** The blue lines indicate the respective averages.

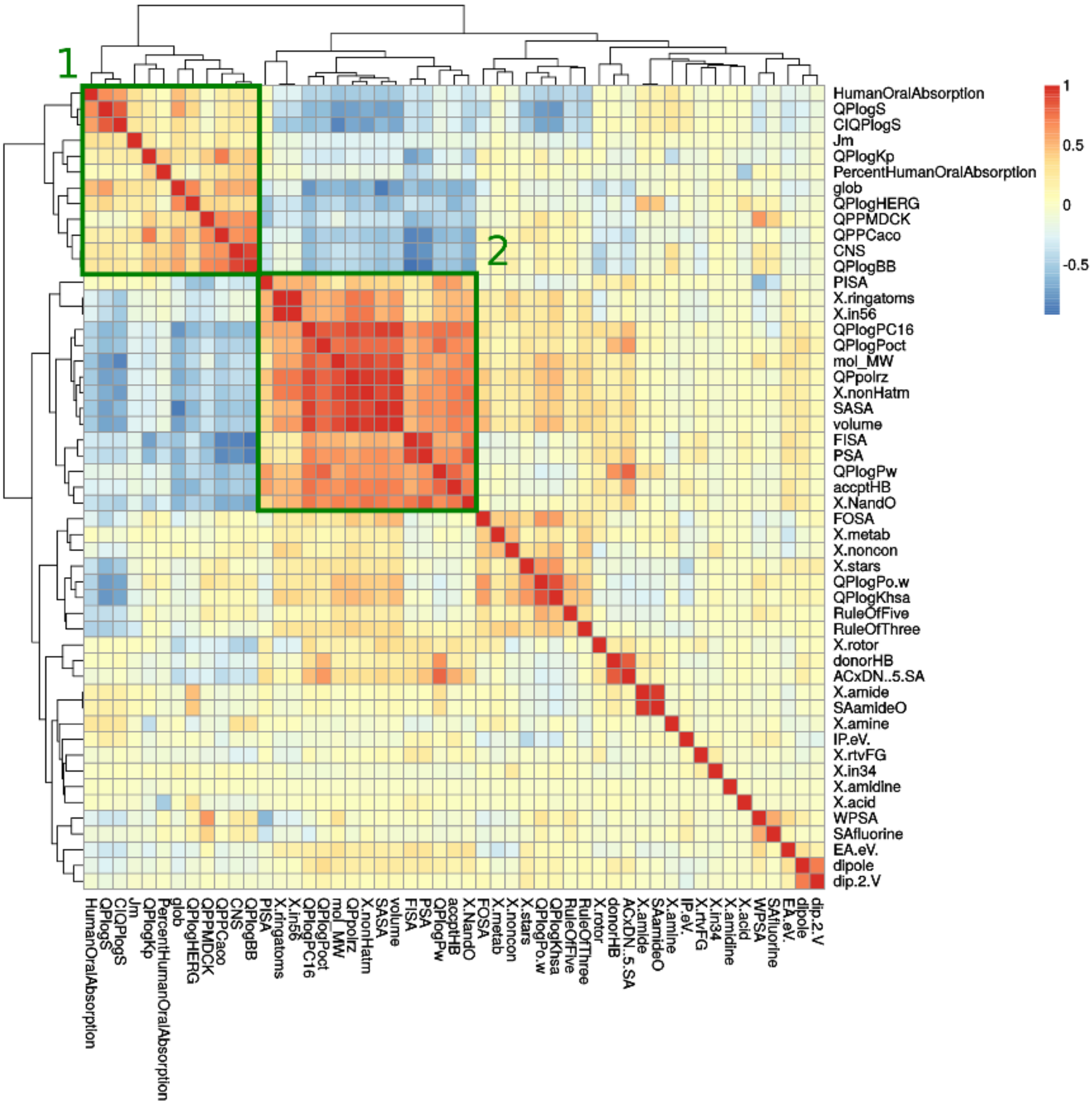


Figure 10.4 Heatmap of cross-correlation matrix of QikProp descriptor data of 21,576 compounds with predicted permeability probability ≥ 0.900**.** Dark green rectangles show descriptor clusters with high internal cross-correlations.

activity and brain/blood partition coefficient (cf. Table 10.2). Cluster 2 contains descriptors about the physico-chemical composition of the compound, e.g. molecular weight, π-interacting surface area, solvent-accessible surface area, volume, polar surface area and number of hydrogen bond acceptors. The highest negative correlations are illustrated as scatterplots in Figure 10.5 and Figure 10.6. The predicted hERG channel blockage correlates to PISA with a Spearman's ρ of -0.56. Although a high π-interacting surface area is desirable for a high *M. tuberculosis* cell wall permeability probability (cf. Chapter 9), it is here also associated with unfavorable toxicity properties (Figure 10.5), which, however, should not be overinterpreted due to poor predictive power of the QPlogHERG descriptor of QikProp [194].

High correlations between clusters 1 and 2 can also be observed for the hydrophilic

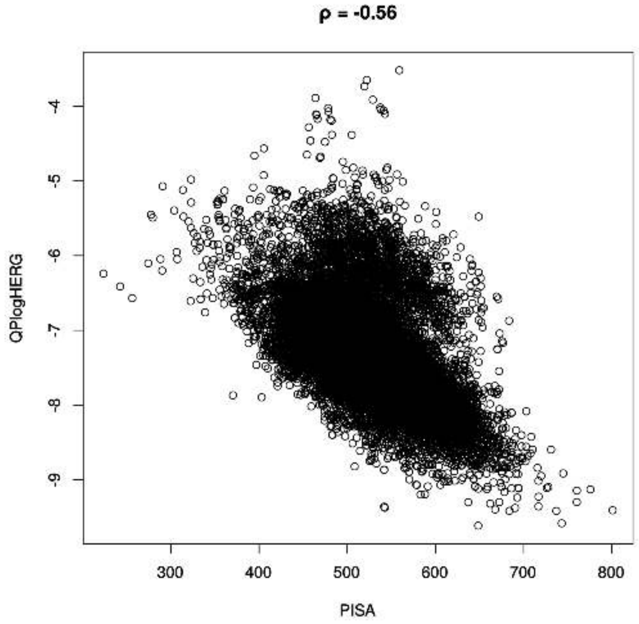


Figure 10.5 Predicted hERG channel blockage ($logIC_{50}$) over π-interacting surface area [Å^2]. The correlation coefficient ρ is calculated using the Spearman method.

surface area (FISA) and the polar surface area (PSA) to the logarithm of the predicted brain/blood partition coefficient and the CNS activity class. With rising surface areas of the related parameters FISA and PSA, the compound is less active in the central nervous system and more abundant in the blood phase (Figure 10.6). Furthermore, the two surface areas are obviously strongly associated in a non-linear way with predicted Caco-2-cell permeability, which models the gut-blood-barrier. With rising hydrophilicity of the compounds the gut-blood-barrier permeability drops rapidly (Figure 10.6), which might hinder absorption of the compound. In the examined dataset, however, over 95% of all compounds still have a predicted Caco-2-cell permeability of $\geq$500 nm/s.

10.1.4 Filtering of ZINC for desirable physico-chemical and ADMET properties and subsequent docking

The molecular weight of the permeable substances has an average of 390 Da. By setting a cutoff at 300 Da, it was possible to reduce the number of compounds to 811. This filtering procedure was continued by considering only molecules with the highest predicted QikProp oral absorption class (3), leaving 629 substances. Next, no reactive groups were allowed in the dataset (#rtvFG = 0), which led to 591 remaining compounds. By considering only molecules with no Lipinski violations, 312 could be dismissed, which led to 279 compounds.

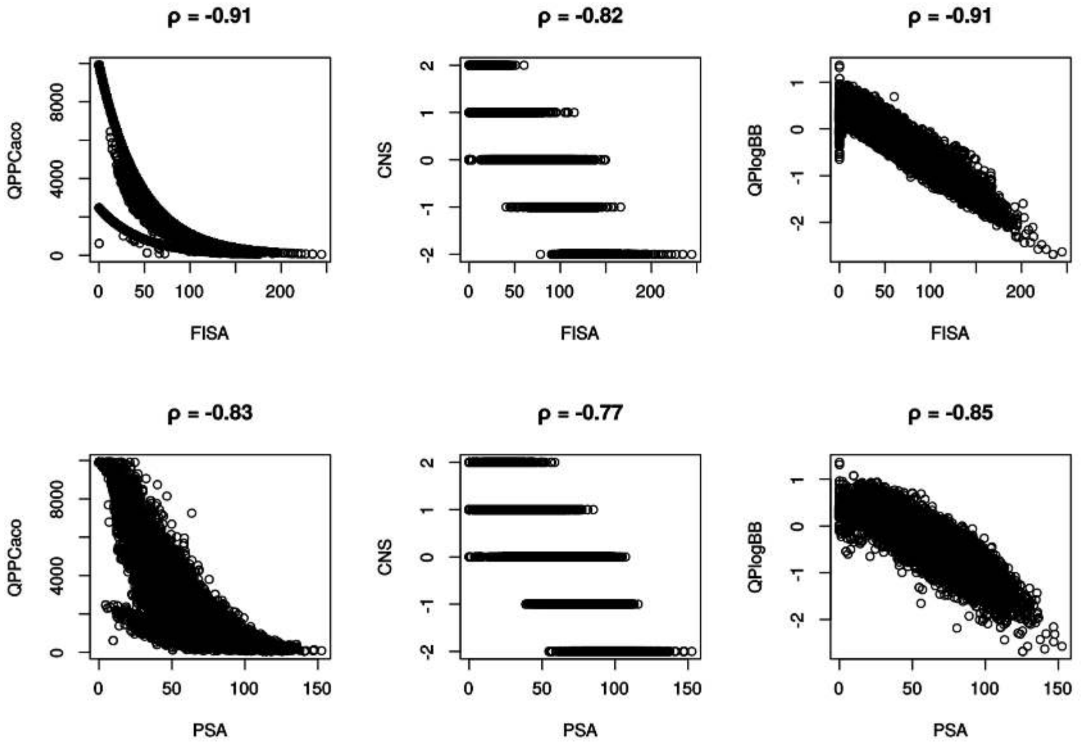


Figure 10.6 Predicted Caco-2-cell permeability [nm/s], CNS activity class and log brain/blood partition coefficient plotted over FISA [Å^2] and PSA [Å^2], respectively, for 21,576 compounds with predicted permeability probability ≥ 0.900**.** ADMET prediction calculated using QikProp. The correlation coefficient ρ is calculated using the Spearman method.

The collection of 279 compounds is the result of a strict filtering procedure for very good predicted permeability and bioavailability properties. However, based on these results, no conclusion about antimycobacterial activity can be drawn. To avoid expensive experimental testing of all substances and an extensive study for target prediction, the target InhA was chosen exemplarily for further investigation. Moreover, by analyzing all 279 substances for potential InhA inhibitors, methods and findings from Part I of this thesis can be considered.

Hence, the screening was continued by docking the 279 compounds with preferable ADMET and permeability properties to InhA using Glide (version 6.2, Schrödinger, LLC, New York, NY, 2014) in extra precision (XP) mode. Ligprep was used to create protomers at pH 7.0 ± 2.0, stereoisomers and tautomers of the 279 compounds, resulting in 390 structures used for docking. Glide was able to generate docking poses for 343 structures (Figure 10.7). The distribution of the docking scores shows a median of -6.299 ± 2.132 MAD. The top 5% of the docking poses was considered for further steps (18 compounds, Figure 10.8).

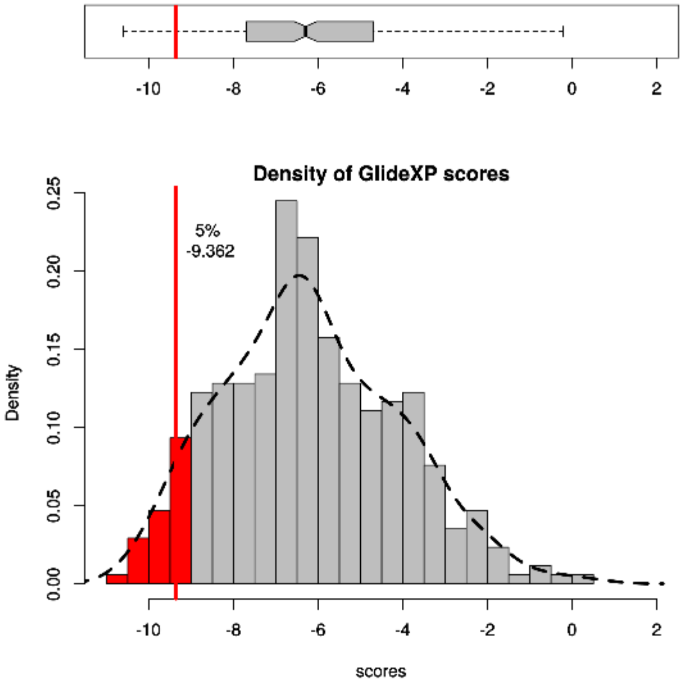


Figure 10.7 Density of GlideXP docking scores as box plot and histogram. The red line illustrates the 5% percentile. The distribution below the 5% percentile is colored in red.

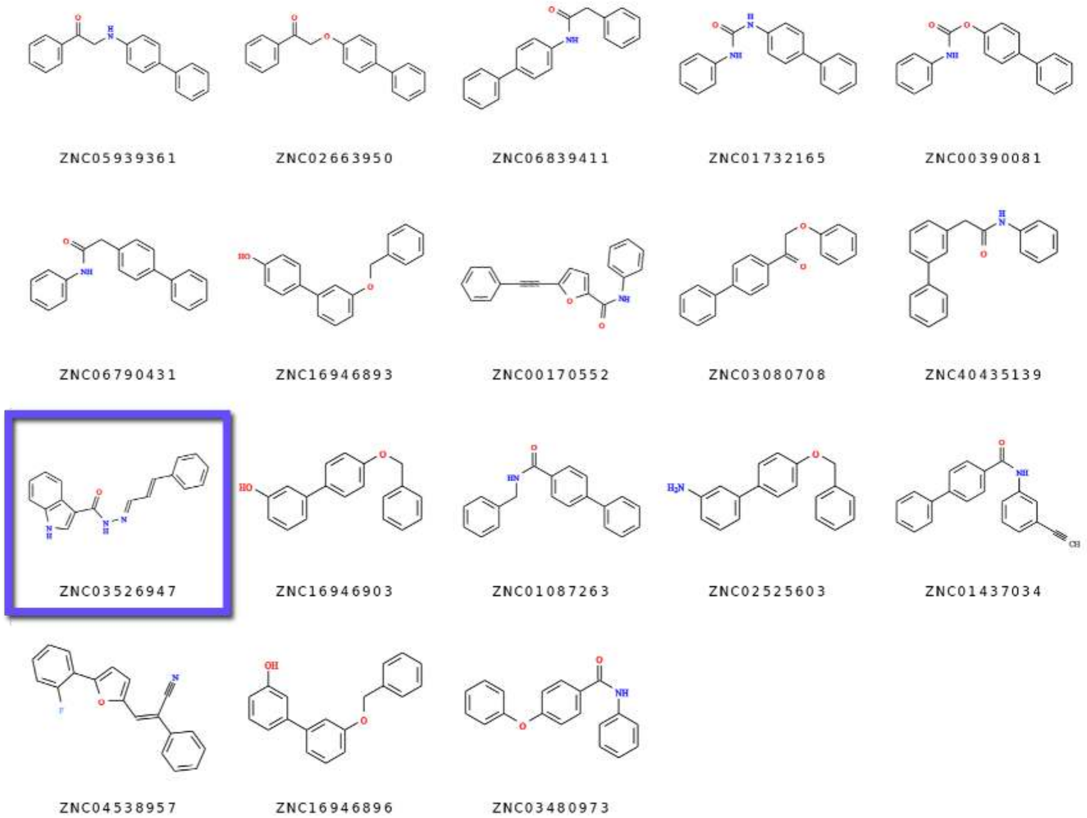


Figure 10.8 Top 5% structures of InhA docking.

A hydrogen bond between the ligand and Tyr158 is a crucial interaction of InhA inhibition [17]. Hence, ligands with a hydrogen bond distance of >3.0 Å in the top docking pose or without hydrogen bond acceptor for Tyr158 were dismissed, leaving 9 candidates. Furthermore, it is known from Chapter 3 that occupation of sufficient space between the cofactor and Ile202/Val203 improves conformational Family 1 stability (cf. **PT70** vs. **6PP**). The structure ZINC03526947 occupies this space with an indolyl moiety embedded in this subpocket, whereas the remaining compounds only have an unsubstituted phenyl group. Thus, compound ZINC03526947 was chosen for further derivatization and evaluation by docking, MD and SMD simulations.

Considerations regarding the binding pocket of InhA and its interactions with potential ligands (cf. Chapter 3) are consistent with the chemical properties of ZINC03526947:
1. the amide oxygen has a high electron density, which is beneficial for forming polar interactions with Tyr158 and NAD^+,
2. a lipophilic, conjugated π-system with terminal phenyl ring might fill the hydrophobic pocket,
3. the terminal indolyl-moiety might embed itself between Ile202 and Val203, which could contribute to stabilizing conformational Family 1 of the binding pocket (cf. Chapter 3),
4. the indolyl-moiety might interact weakly with Met161 via an aryl-methionine interaction (cf. Chapter 4) [158].

Moreover, the compound shows further promising properties, independent of InhA inhibition:
1. the compound has very good predicted ADMET properties,
2. according to calculations with the pKa prediction software MoKa 2.6.0 [231], the pKa of the amide-adjacent nitrogen is 6.18. Hence, at a pH of 7.0, 13.3% of the compound is present in its protonated form, which might improve solubility. Moreover, the formation of a pharmaceutical salt with an acid might further improve solubility and dissolution rate of the compound [232].

10.2 Docking and MD simulations suggest potential InhA inhibitors

10.2.1 Docking of ZINC03526947 derivatives

The top docking pose of ZINC03526947 is illustrated in Figure 10.9. Based on visual inspection of the top binding pose, the ligand was modified to better fit the geometry and interaction sites of the InhA binding pocket. The resulting ligands were again

Table 10.3 Docking results of original and modified ligands. The Glide XP scores of the respective top poses are shown. MycPermCheck values are from version 1.1. MycPermCheck values of **mod7** and **mod8** are not available due to a permanent charge in the molecule.

Ligand	GlideScore XP	MycPermCheck	Reason for ligand modifications
PT70	-10.183	0.736	
ZINC03526947	-9.676	0.902	
mod2	-9.748	0.879	methyl group is assumed to embed between Ile202 and Val203
mod3	-9.489	0.833	cf. **mod2**; N-methyl is assumed to occupy space between ligand and NAD$^+$
mod4	-9.999	0.869	cf. **mod3**, but avoiding potential clash of methyl with Met103
mod5	-10.031	0.866	methylene linker improves geometrical accessibility of amide oxygen to Tyr158 and NAD$^+$ and disrupts the vinylogous semicarbazone moiety
mod6	-9.882	0.853	cf. **mod5**; occupation of space between ligand and NAD$^+$
mod7	-6.044	N/A	cf. **mod5**; guanidyl-moiety introduces permanent positive charge and is supposed to form electrostatic interactions with phosphates of NAD$^+$
mod8	-6.235	N/A	cf. **mod5**/**mod7**; further occupation of space in hydrophobic pocket
mod9	-10.408	0.864	cf. **mod5**; further occupation of space in hydrophobic pocket
mod10	-9.475	0.816	cf. **mod6**; saturation of double bond to reduce reactivity of unsaturated system
mod11	-10.150	0.835	cf. **mod5**; saturation of double bond to reduce reactivity of unsaturated system

docked with Glide (Figure 10.9) in extra precision (XP) mode. Table 10.3 shows the GlideScore XP and MycPermCheck evaluation of the top poses, as well as a detailed description of the performed ligand modifications. Additionally, **PT70** was re-docked into the crystal structure 2X23 as a reference, reaching a GlideScore of -10.183 at an RMSD of 0.68 Å from the crystal structure pose. Of the docked ligands, particularly **mod9** and **mod11** show similar or better docking scores compared to the reference ligand **PT70** (Table 10.3).

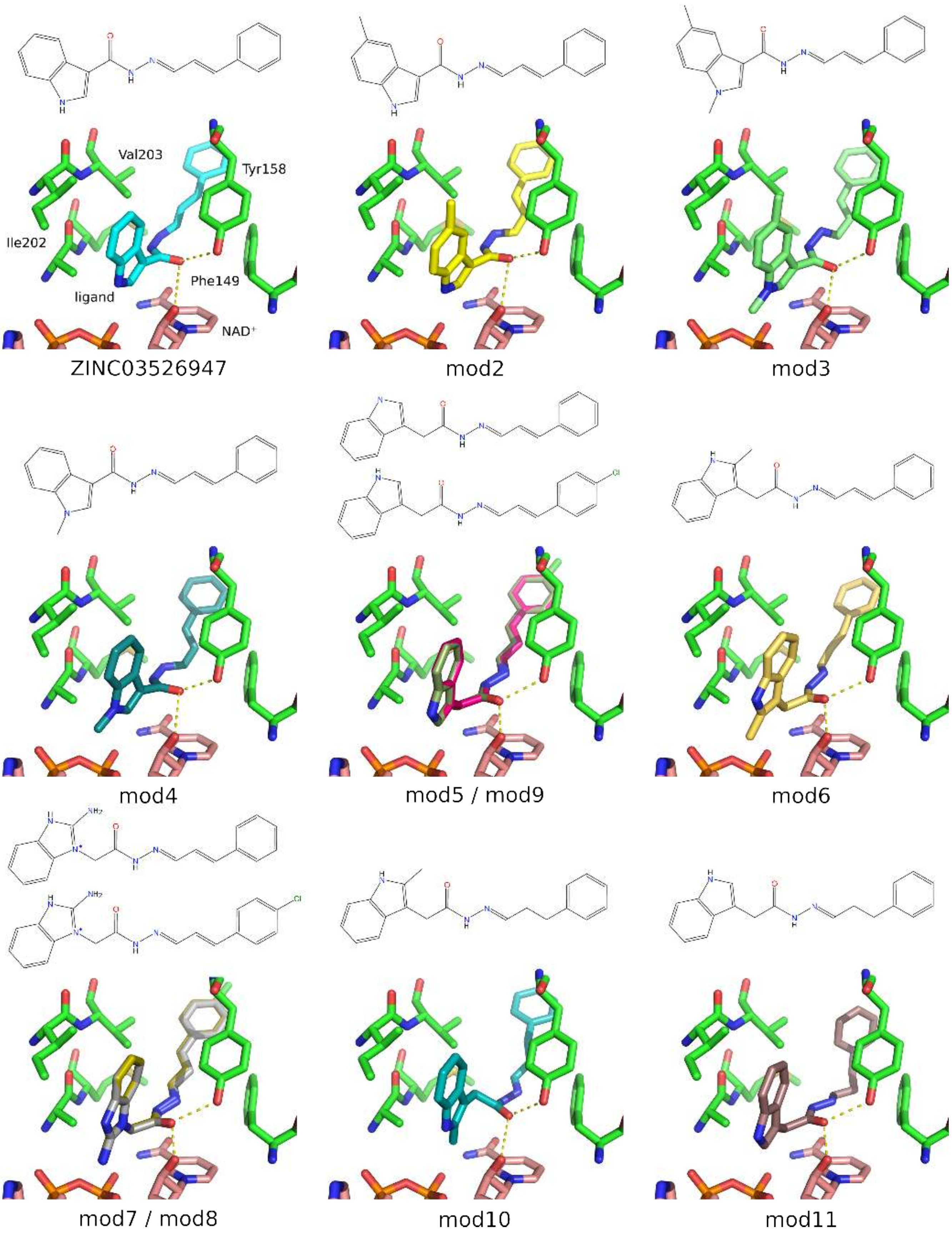


Figure 10.9 Top docking poses of ZINC03526947 and derivatives in InhA. The protein backbone and six pocket residues are illustrated in green, the cofactor NAD$^+$ is depicted in salmon, the ligands in varying colors. Ligand structures are illustrated above.

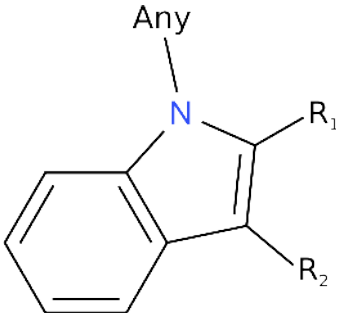


Figure 10.10 Critical PAINS substructure in mod6 and mod10; a 2,3-dialkyl indole.

QikProp was used to assess the physico-chemical and ADMET properties of ZINC03526947 and the designed derivatives except **mod7** and **mod8**. With one minor exception (QPlogKp), all of the previously discussed descriptors are within the range of 95% of known drugs (cf. Tables 10.1 and 10.2; QikProp manual, Schrödinger Software Release 2014-1).

In 2010, Baell and Holloway described several chemical substructures which appear as false positives in many different assays. Compounds containing such substructures are termed *Pan Assay Interference Compounds* or PAINS [233]. Accordingly, a PAINS filter was applied to exclude substructures that result in frequent hitters in bioassays using the PAINS Remover online tool (http://cbligand.org/PAINS/) [233].

The derivatives **mod5** to **mod11** were examined for interfering substructures. One feature of ligand **mod6** (and recurring in **mod10**) was identified as a frequent hitter: a 2,3-dialkyl indole (Figure 10.10).

The overall enrichment of assay hits of this group is, however, comparably low [233]. Hence, the results of this filtering step should not be overrated. On the other hand, it is worth mentioning that *"the broader class of indole-3-acetamides allowing any substituent off the 2 position (including H) yields 138 compounds with a relatively high enrichment"* [233]. This group comprises the ZINC03526947 derivatives **mod5**, **mod6** and **mod9** to **mod11**. While an exclusion from further investigations is not necessarily required, special care must be taken upon an eventual experimental testing of the corresponding compounds.

10.2.2 MD simulations

Since the introduction of a methylene linker between the indolyl and hydrazide was an important modification to reduce the distance between the amide oxygen and Tyr158 as well as NAD$^+$, ligands **mod5** to **mod11** were chosen for 30 ns MD simulations using Amber force field parameters in NAMD 2.9 [101] (see Chapter 5 for simulation protocol details), resulting in 210 ns total sampling time. To assess the stability of the various systems, the following analyses were carried out: distance measurements between amide oxygen of the ligand and phenolic oxygen of Tyr158 as well as ribose-oxygen of NAD$^+$, protein backbone RMSD, RMSD of InhA binding pocket residues Tyr158, Phe148, Ala198, Met199, Ile202 and Val203 (cf. Chapter 3), RMSD of InhA SBL (residues 202 to 218, cf. Chapter 3), RMSD of ligands (Figure 10.11). Additionally, the binding pocket dynamics were compared in terms of their 2D-RMSD (Figure 10.12).

10.2.2.1 RMSD and distances

Tyr158 plays an important role in binding of known inhibitor classes of InhA [17]. Hence, the distances between the amide oxygen of the ligand and the Tyr158-OH atom, as well as the O2D-oxygen of NAD$^+$, were evaluated (Figure 10.11a). Most of the modified ligands show distances well below 3 Å over 30 ns of MD simulation, indicating a stable hydrogen bond. Higher initial distances are resolved in some cases due to attraction of the participating atoms (**mod5**-NAD$^+$, **mod6**-Tyr, **mod7**-Tyr, **mod8**-Tyr, **mod11**-Tyr). The ligands **mod7**, **mod8**, **mod9** and **mod10** lose at least one of the interaction partners during the observed simulation time. Derivatives **mod5**, **mod6** and **mod11** exhibit the strongest hydrogen bond patterns based on the measured distances (Table 10.4).

With RMSD values steadily below 2 Å, the protein backbone is fairly stable in all seven simulations (Figure 10.11b). Again, derivatives **mod5** and **mod6** show the highest stability, as does the backbone of the system containing **mod10** (Table 10.4).

The two most important structural elements that govern slow-onset ligand binding in InhA are the SBL and the binding pocket itself [17, 45, 64], which should show high stability upon ligand association (Figures 10.11c and d). Both regions show the highest stability regarding their RMSD in the simulations of ZINC03526947 derivatives **mod6**, **mod10** and **mod11** with an RMSD steadily below 2 Å. Considering the binding pocket, also **mod5** exhibits very stable behavior. This trend is also observed in the ligand RMSD over 30 ns of simulation. Again, the most stable systems are **mod6**, **mod10** and **mod11** with respect to the starting structure, i.e., the docking pose. However, derivatives **mod10** and **mod11** show higher fluctuations than the remaining complexes after the initial 10 ns. With standard deviations of less than 0.30 Å, **mod5**, **mod6** and **mod7** are the most stable ligands from 10 to 30 ns simulation time, underlining the stability of their conformation adopted during the first 10 ns of simulation time

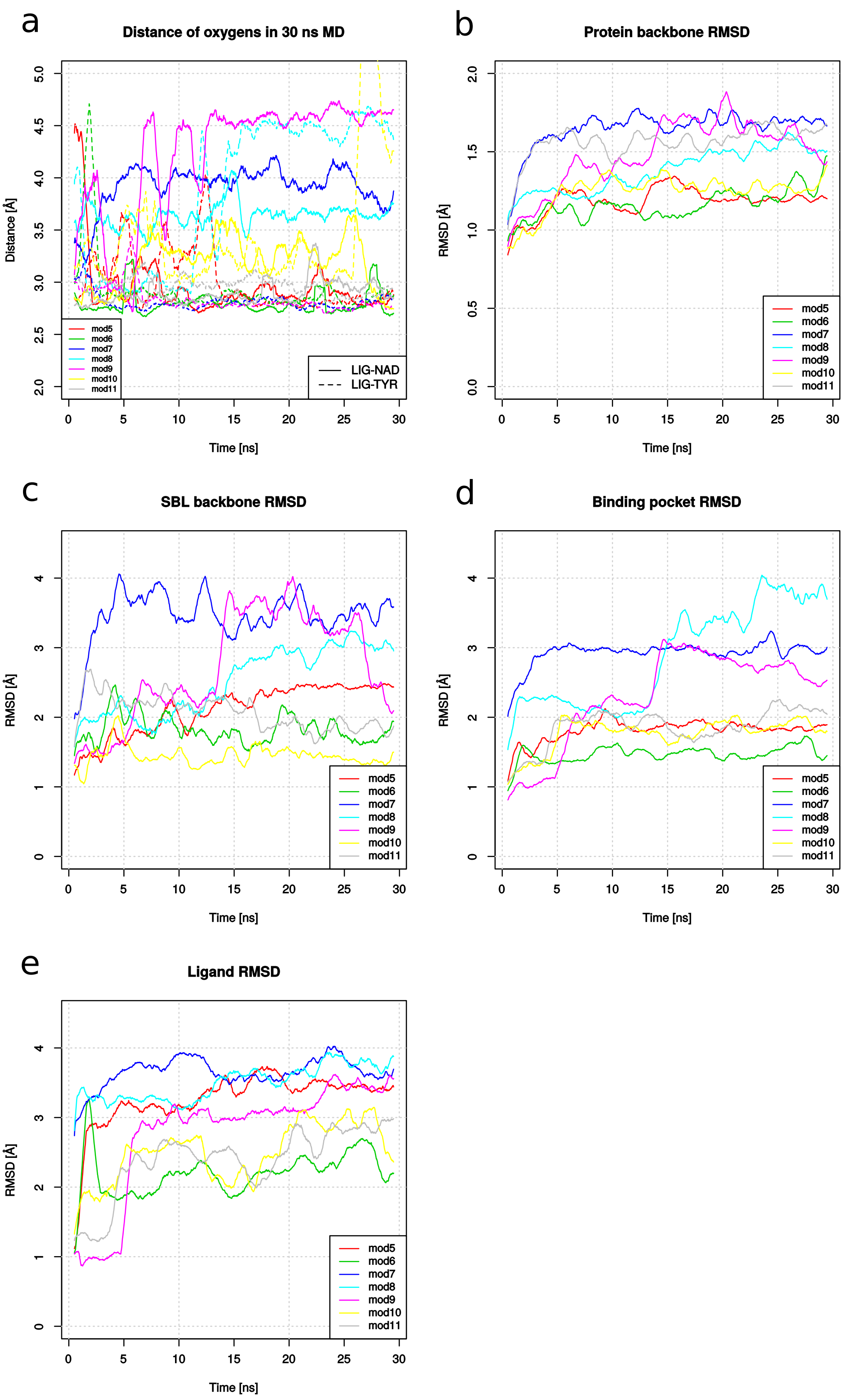


Figure 10.11 RMSD and distance evaluation of 30 ns MD simulations of selected ZINC03526947 derivatives. **(a)** Distances between ligand-O and Tyr-OH or NAD-OH, respectively.**(b)** Backbone RMSD. **(c)** Pocket RMSD. **(d)** SBL RMSD. **(e)** Ligand heavy atom RMSD.

Table 10.4 Average distance and RMSD values over 30 ns of MD simulation in Å.

		mod5	**mod6**	**mod7**	**mod8**	**mod9**	**mod10**	**mod11**
Distance	Avg.	3.00	2.93	2.81	3.87	2.84	3.46	2.97
Tyr158	SD	0.49	0.41	0.20	0.81	0.23	0.86	0.28
Distance	Avg.	2.98	2.80	3.92	3.65	4.23	3.18	2.89
NAD$^+$	SD	0.47	0.29	0.40	0.35	0.72	0.45	0.29
Backbone	Avg.	1.18	1.16	1.64	1.38	1.48	1.24	1.56
RMSD	SD	0.12	0.13	0.16	0.15	0.24	0.14	0.15
SBL RMSD	Avg.	2.10	1.81	3.45	2.51	2.69	1.44	2.06
	SD	0.43	0.36	0.50	0.55	0.84	0.26	0.35
Pocket RMSD	Avg.	1.81	1.46	2.91	2.87	2.29	1.76	1.86
	SD	0.23	0.18	0.27	0.77	0.69	0.29	0.32
Ligand RMSD	Avg.	3.24	2.18	3.64	3.49	2.77	2.48	2.38
	SD	0.53	0.43	0.34	0.38	0.88	0.51	0.56
Ligand RMSD	Avg.	3.45	2.26	3.70	3.62	3.20	2.62	2.59
10 to 30 ns	SD	0.27	0.29	0.26	0.31	0.32	0.45	0.36
Distance	Avg.	5.43	4.97	5.61	5.15	5.14	4.87	5.88
Met161	SD	0.60	0.55	0.78	0.54	0.73	0.49	1.02

(cf. Table 10.4 and Figure 10.11). A stable aryl-methionine interaction between the ligand and Met161 could not be observed in the simulated systems.

10.2.2.2 2D-RMSD analysis

A 2D-RMSD analysis of the six binding pocket residues was used to identify recurring pocket conformations across all seven systems (Figure 10.12). The highest similarities are observed between the pockets bound to ZINC03526947 derivatives **mod5**, **mod6** and **mod10**, as well as the first half of the **mod8**-, **mod9**- and **mod11**-simulation, respectively. Of all observed conformations, this one is closest to the 2X23 crystal structure conformation used in the dockings. The conformation visually resembles the conformational Family 1, as described in Chapter 3 (blue rectangles in Figure 10.12). The **mod7**-system only very briefly populates this Family, before collapsing into a conformation resembling Family 3, with Ile202 flipping over the ligand and Ile202 and Val203 moving deep into the hydrophobic pocket (magenta rectangles in Figure 10.12). This conformation recurs in the second half of the **mod9**-simulation. The second half of the **mod11**-simulation exhibits similarities to all observed binding pocket conformations, except the second half of **mod8**. The similarity in RMSD of the **mod11**-system to both Family 1 and Family 3 conformations strongly suggests a conformation with structural characteristics of both families (purple rectangles in Figure 10.12). In fact, visual inspection reveals a plausible resemblance of the **mod11**-pocket to the conformational Family 2, i.e, only a slight twist of helix $\alpha6$ with a shift of Ile202 towards the ligand and a minor displacement of Val203 towards the hydrophobic pocket (cf. Chapter 3). The second half of the simulation of InhA bound to **mod8** is characterized by a shift of

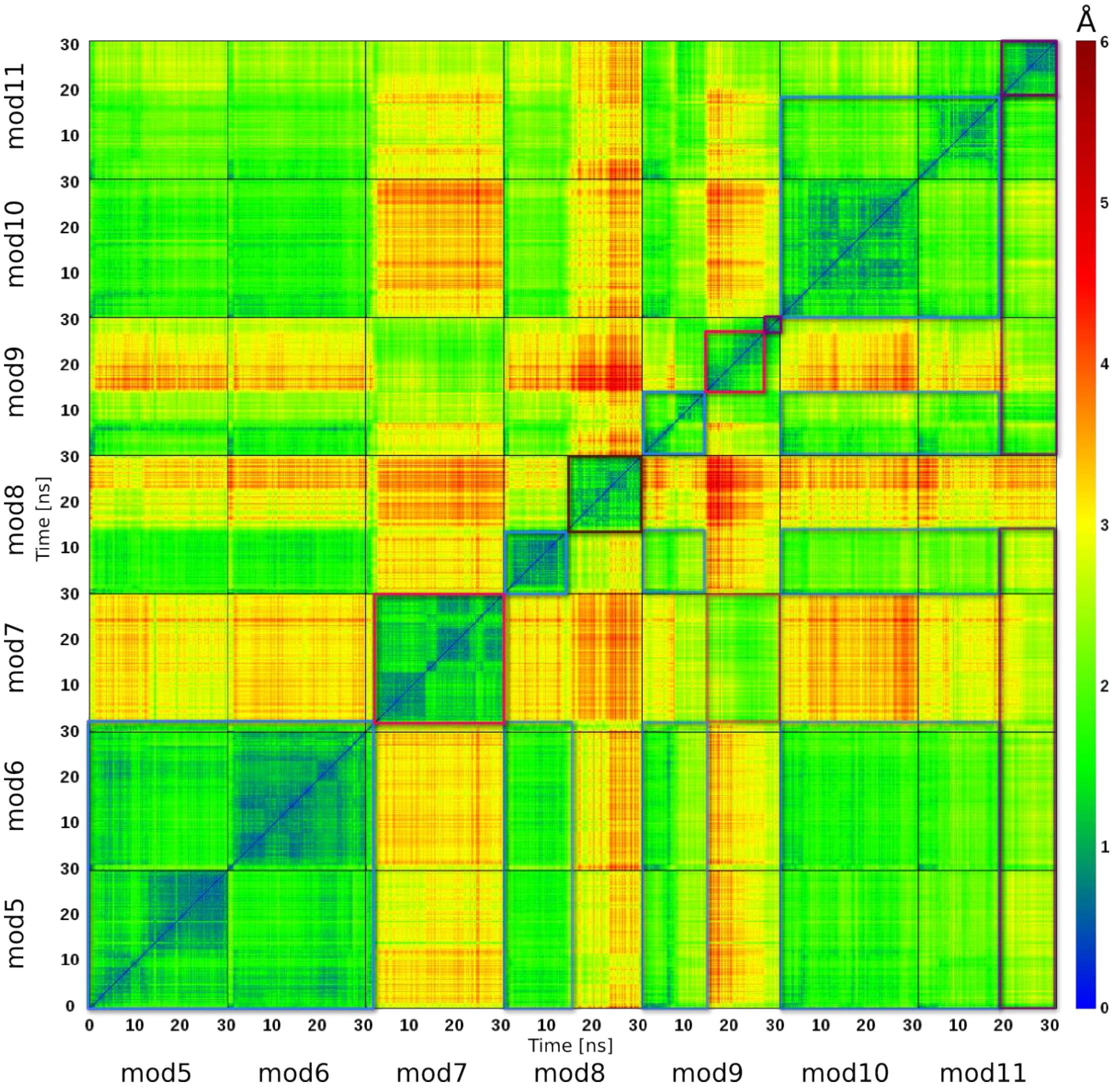


Figure 10.12 7x7 2D-RMSD plot of binding pocket residues over 30 ns of MD simulation of ZINC03526947 derivatives mod5 to mod11. Colored rectangles mark conformational families.

helix α6–in particular Ile202 and Val203–towards the bulk solvent (brown rectangle in Figure 10.12). This feature can visually be best described as a Family 5 conformation (cf. Chapter 3).

10.2.2.3 Clustering analysis

The 2D-RMSD matrix was used as distance metric for a hierarchical clustering analysis using the complete-linkage method in R [130]. At a cutoff of 3.5 Å, the data can be partitioned into three distinct clusters (Figure 10.13).

Comparison of the medoids of the conformational families of InhA (revealed in Chapter 3) to the medoids of the clusters of these MD simulations revealed that, on basis of

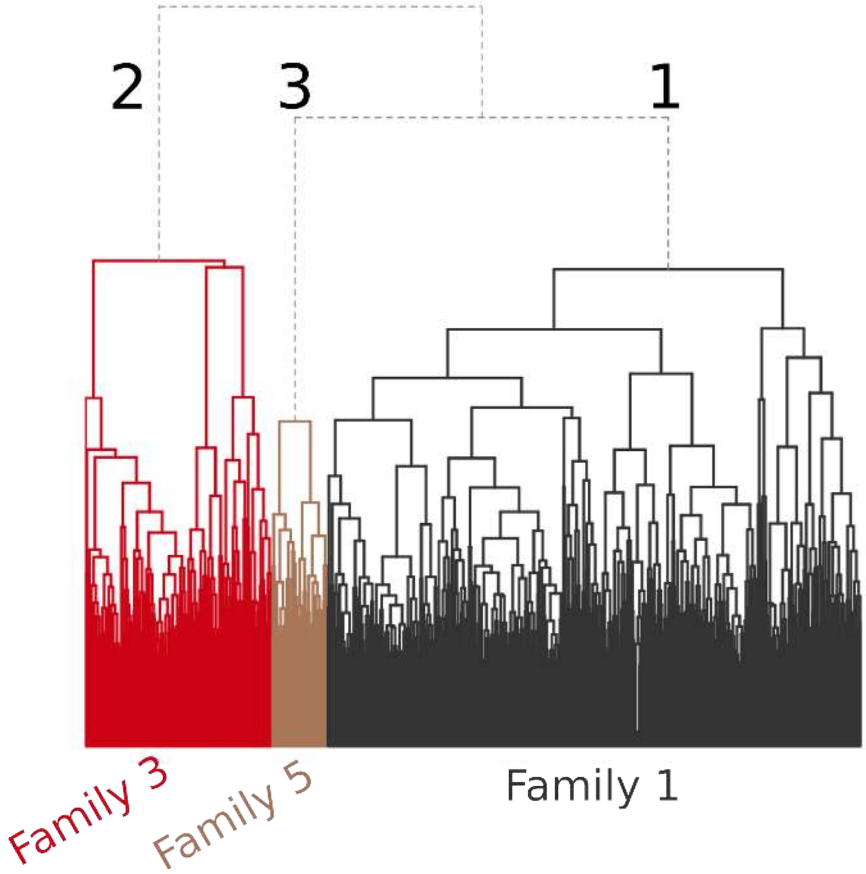


Figure 10.13 Hierarchical clustering of 2D-RMSD data of 30 ns MD simulations bound to ZINC03526947 derivatives mod5 to mod11.

the same cutoff (3.5 Å), no new conformational families could be discovered. In fact, each of the three clusters can distinctly be assigned to one of the Families described in Chapter 3. Thus, (1) the cluster 1 represents a Family 1 conformation, (2) cluster 2 shows a strong resemblance to the Family 3 medoid and especially to the cluster 4 medoid of Family 3, and (3) cluster 3 can be ascribed to Family 5 (Table 10.5 and Figure 10.14).

10.2.2.4 Summary of MD results of derivatives mod5 to mod11

Considering the RMSD analyses, derivatives **mod5** and **mod6**, as well as their saturated counterparts **mod11** and **mod10**, respectively, showed the highest stability. When taking the interaction distance analysis of Tyr158 and NAD$^+$ to the ligands and the fluctuations in the ligand RMSD into account, **mod10** does not fulfill the requirements for stable ligand binding to InhA.

Interestingly, none of the designed ligands is entirely stable in its docking pose. In the cases of **mod5** and **mod6**, the ligands drift at the beginning of the simulation

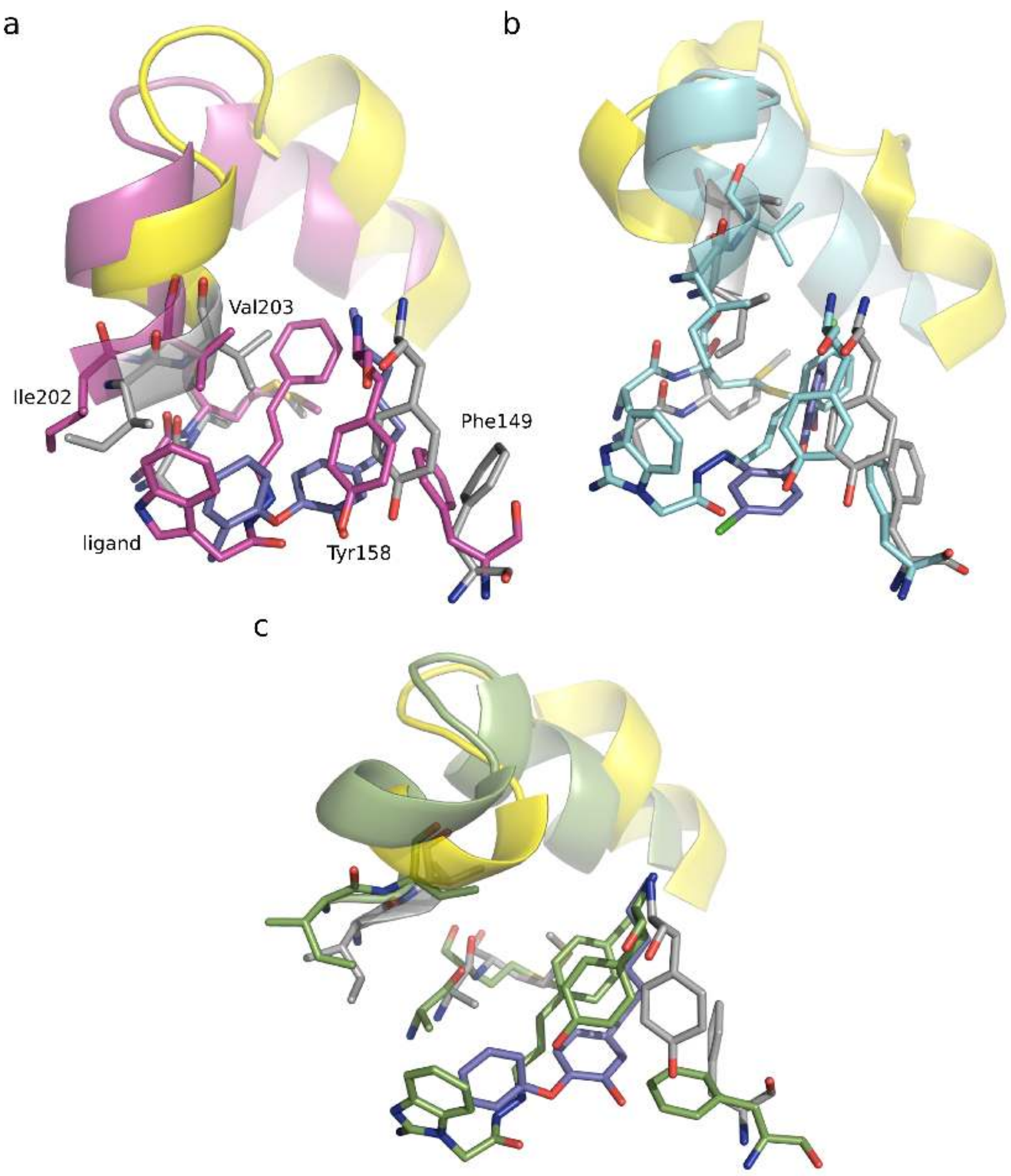


Figure 10.14 Medoid snapshots of clusters 1 to 3 of 30 ns ZINC03526947 derivative simulations, compared to their respective counterparts (cf. Chapter 3). **(a)** Snapshot of **mod5** at 2.3 ns of simulation (pink) compared to the Family 1 medoid, **(b)** snapshot of **mod7** at 8.2 ns of simulation compared to cluster 4 medoid (part of Family 3), **(c)** snapshot of **mod8** at 11.5 ns of simulation compared to the Family 5 medoid.

Table 10.5 RMSD of cluster medoids to conformational Family and cluster medoids defined in Chapter 3 in Å. Minimum RMSD values for each column are highlighted.

	ligand modifications		
	cluster 1	**cluster 2**	**cluster 3**
cluster 1 / Family 1	**1.82**	3.43	3.56
cluster 2 / Family 2	2.71	2.81	3.61
cluster 3 / Family 2	2.60	2.58	3.83
Family 2 medoid	2.81	2.90	3.89
cluster 4 / Family 3	3.80	**1.75**	4.01
cluster 5 / Family 3	3.78	3.06	3.85
cluster 6 / Family 3	3.68	2.84	4.35
Family 3 medoid	3.73	2.71	3.91
cluster 7 / Family 4	2.98	4.82	3.54
cluster 8 / Family 5	3.24	3.96	**2.86**

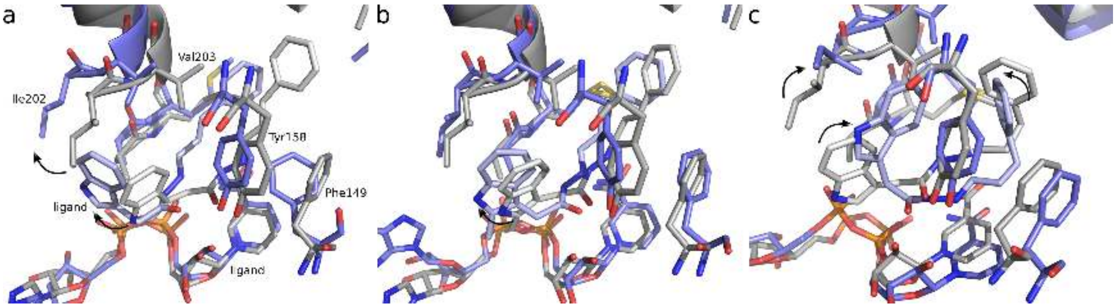


Figure 10.15 Snapshots of MD starting and end structures at 30 ns. (a) Ligand **mod5**; **(b)** ligand **mod6**; **(c)** ligand **mod11**. The starting structure is illustrated in gray, the end snapshot in slate blue. Arrows indicate atomic displacement over 30 ns.

towards the NAD$^+$ ribose-hydroxyl and, thus, move the indolyl-moiety further towards the major portal (Figure 10.15). The new orientation is, however, stable until the end of the simulation. Thus, the conformation shows a slightly altered arrangement of helix $\alpha 6$ compared to the 2X23 crystal structure (cf. Figure 10.14a). The ligand **mod11** has an augmented flexibility due to saturation of the double bond. This results in a unique behavior among the investigated ligands: the termini of the potential inhibitor collapse towards the center of the binding pocket, enabling a Family 2-like conformation. This suggests that **mod11** does not occupy the hydrophobic pocket optimally.

Although the observed clusters based on 2D-RMSD data are very close to the previously described conformational families of *M. tuberculosis* InhA (cf. Chapter 3 and Table 10.5), some conformational differences are detected as well: in particular, the two pocket residues Phe149 and Tyr158, which are not part of the SBL, show severe differences to the Family medoids (cf. Figure 10.14). This behavior likely stems from still inefficient

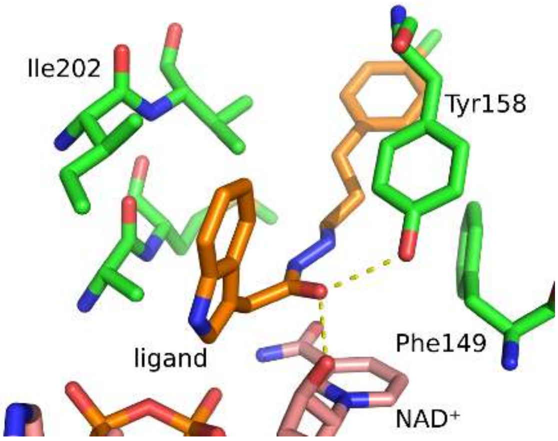


Figure 10.16 Top docking pose of ligand mod12.

occupation of the hydrophobic pocket, which allows Phe149 to move without hindrance. This effect is accompanied by a shift of Tyr158, which follows the drifting of the ligand towards the major portal (cf. Figures 10.14 and 10.15a and b).

10.2.3 An additional ZINC03526947 derivative: mod12

Based on these findings, an additional ligand was evaluated: **mod12** is a structural derivative of **mod11** with an additional *para*-chlorine-substituent at the phenyl-ring. Docking resulted in a binding pose with a GlideXP score of -10.422, which is the best observed GlideScore among all ZINC03526947 derivatives and the reference inhibitor **PT70** (Figure 10.16). The ligand shows a MycPermCheck permeability probability of 0.844.

An additional 30 ns MD simulation was carried out for the InhA-**NAD$^+$-mod12** system. The system shows a very high overall stability based on all previously presented criteria (Table 10.6). All metrics exhibit comparable results to the most stable simulations of derivatives **mod5** to **mod11**. Moreover, the pocket and ligand RMSDs are in fact the lowest observed. The system furthermore seems to be able to stabilize an indolyl-Met161 interaction throughout the whole simulation (cf. Table 10.6).

10.2.4 Steered MD simulations

Steered MD simulations were conducted for the most promising ZINC03526947 derivatives based on the previous MD results according to the protocol introduced in Chapter 5. The exit pathway of PT70 (cf. Chapter 5) was transferred to the investigated ligands for

Table 10.6 Average distance and RMSD values of system **mod12** over 30 ns of MD simulation in Å.

		mod12
Distance	Avg.	2.88
Tyr158	SD	0.25
Distance	Avg.	2.89
NAD$^+$	SD	0.34
Backbone	Avg.	1.39
RMSD	SD	0.29
SBL RMSD	Avg.	1.97
	SD	0.51
Pocket RMSD	Avg.	1.38
	SD	0.30
Ligand RMSD	Avg.	1.68
	SD	0.45
Ligand RMSD	Avg.	2.16
10 to 30 ns	SD	0.32
Distance	Avg.	4.16
Met161	SD	0.58

induced extraction via the major portal of InhA. A total of 360 ns of additional SMD sampling time was produced. The maximum free energy change upon induced ligand withdrawal from the binding pocket was examined with reference to the SMD results of the established inhibitor class of the diphenylethers (cf. Chapter 6). Table 10.7 and Figure 10.17 depict the maximum exponential average ΔG_e and the first and second order cumulant expansions of Jarzynski's equality ΔG_1 and ΔG_2, respectively [33–36]. The primarily considered measure in this analysis is the exponential estimator, due to its validity for small trajectory numbers and robustness with respect to high variance (cf. Chapter 6.3.2) [165].

The maximum free energy changes ΔG_e (trimmed average, Figure 10.17a) of the SMD simulations show a distinct separation between the ZINC03526947 derivatives **mod10**, **mod11** and **mod12**, which reach comparable or–in the case of **mod10**–a much higher maximum than the reference compound **PT70**. Derivatives **mod5**, **mod6** and **mod9** do not exhibit maximum free energy changes comparable to the reference **PT70**. Surprisingly, the lowest energy barrier can be observed for **mod6**, although the previous 30 ns MD simulations showed reasonable SBL and binding pocket stabilization. The exponential average reconstructed from untrimmed work values shows the same trends, although the higher scoring ligands **mod11** and **mod12** are less distinctive from the lower scoring ligands **mod5**, **mod6** and **mod9** (Figure 10.17b). Ligand **mod10** also exhibits the highest maximum ΔG_e using untrimmed work values.

Table 10.7 Maximum free energy change of induced ligand extraction of ZINC03526947 derivatives with standard deviation in kcal/mol.

	trimmed			untrimmed		
	ΔG_e	ΔG_1	ΔG_2	ΔG_e	ΔG_1	ΔG_2
PT70	**89.02 ± 11.31**	106.29 ± 11.32	68.11 ± 2.38	77.53 ± 20.31	107.15 ± 20.32	39.95 ± 6.09
mod5	**78.33 ± 3.00**	80.92 ± 3.00	73.38 ± 3.00	69.47 ± 7.28	81.46 ± 7.31	36.85 ± 7.27
mod6	**69.49 ± 3.93**	75.59 ± 3.96	62.54 ± 3.93	62.32 ± 8.37	75.47 ± 8.37	31.20 ± 3.69
mod9	**73.68 ± 2.28**	76.56 ± 2.30	72.19 ± 2.22	63.97 ± 8.09	76.71 ± 8.09	28.36 ± 7.09
mod10	**112.29 ± 4.62**	117.10 ± 4.62	99.20 ± 4.62	89.92 ± 11.58	116.74 ± 11.58	32.99 ± 4.52
mod11	**84.97 ± 2.93**	89.53 ± 2.95	82.30 ± 2.92	69.85 ± 10.84	89.61 ± 10.94	24.59 ± 5.89
mod12	**92.62 ± 6.10**	102.37 ± 6.10	71.26 ± 6.09	71.49 ± 13.99	101.64 ± 13.98	21.55 ± 3.55

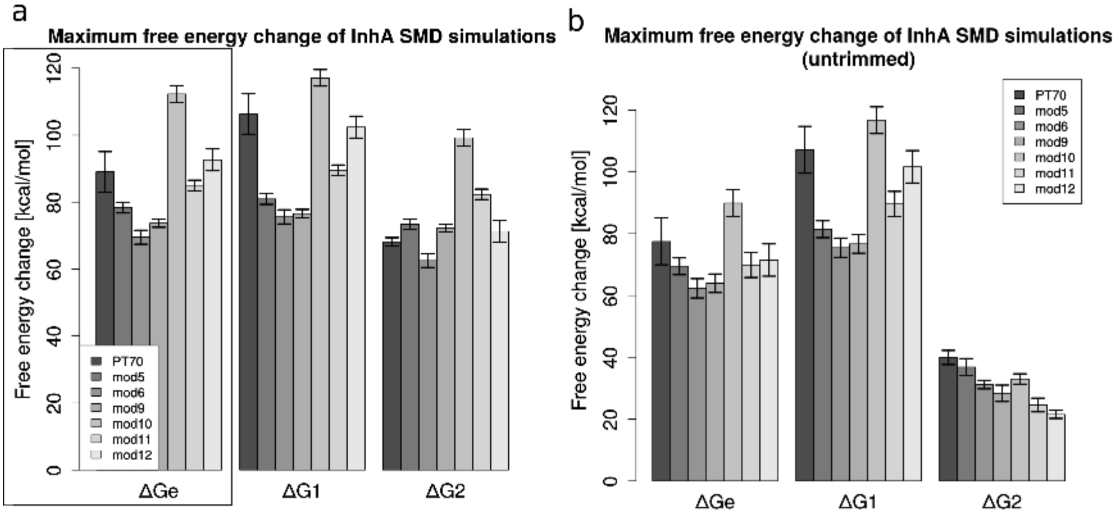


Figure 10.17 Maximum free energy change of induced ligand extraction of ZINC03526947 derivatives from the InhA binding pocket. (a) 25% trimmed average of work at PMF reconstruction, **(b)** untrimmed average of work at PMF reconstruction. Error bars indicate the 95% confidence interval.

Based on the SMD results, the following conclusions and further strategies for ligand optimization can be derived: (i) **mod10** exhibits the highest maximum energy barrier of ligand dissociation along the chosen dissociation pathway, compared to **PT70**; (ii) saturation of the double bond (**mod6** to **mod10** and **mod5** to **mod11**) boosts the maximum ΔG_e; (iii) a **para**-chlorine (**mod11** to **mod12**) boosts the maximum ΔG_e; (iv) thus, the next step in optimization might include combining the structural characteristics of derivatives **mod10** and **mod12**.

10.3 Evaluation of similar compounds

10.3.1 Activity data

To gain information about potential biological activity of the investigated derivatives, a structure search was performed with SciFinder for derivatives **mod5** to **mod12**. Although **mod5** is indeed commercially available, none of the derivatives has yet been the subject of testing in the literature.[1] Thus, the first ZINC03526947 derivative containing the indole-3-acethydrazide scaffold (**mod5**) was chosen for a similarity search against the ChEMBL database [222]. With 88.22% similarity, the compound CHEMBL3210727 (ZINC00221556; PubChem CID 9556992) was the closest match (Figure 10.18). An exhaustive PubChem search [234, 235] revealed assays, in which ZINC00221556 was classified as active (pubchem.ncbi.nlm.nih.gov, accessed March 17, 2015). It shows inhibitory qualities against Human tyrosyl-DNA phosphodiesterase 1 (TDP1), which is proposed as a new anticancer target (PubChem BioAssay: AID 686979). Moreover, activity against Regulators of G protein signaling (RGS) protein 16 (AssayID 1441) could be observed. Against MEK kinase 3 wild-type, the compound showed activating properties (AssayID 1529). A cytotoxicity study using THP1 cells (AssayID 2253) revealed toxicity with cell viability of 69.71% at a concentration of 50 μM.

Furthermore, four screenings were found in which the compound was tested for activity against *Mycobacterium tuberculosis* (PubChem AssayIDs: 1626, 449762, 434955 and 488890). However, it was classified as *inactive* in each assay. None of the mentioned bioassays were explicitly designed for the target InhA.

ZINC00221556 was docked to InhA using Glide with extra precision. Figure 10.18 illustrates the top docking pose with a Glide XP score of -9.216. Except for **mod7** and **mod8**, this is the least favorable docking score of the evaluated structures. Still, the docking pose shows a very similar binding mode to ZINC03526947 and its derivatives. Visual comparison, however, reveals that the hydrophobic pocket is less occupied by this ligand, which might have an unfavorable effect on binding pocket, SBL and ligand stabilization (cf. Chapter 3) and, thus, provides suggestions for the inactivity in antitubercular assays.

10.3.2 MD and SMD simulations

The assumption that possible instabilities, resulting from unideal interactions between ZINC00221556 and InhA, might contribute to the inactivity of the compound against

[1] SciFinder search performed on September 24, 2015 at http://scifinder.cas.org.

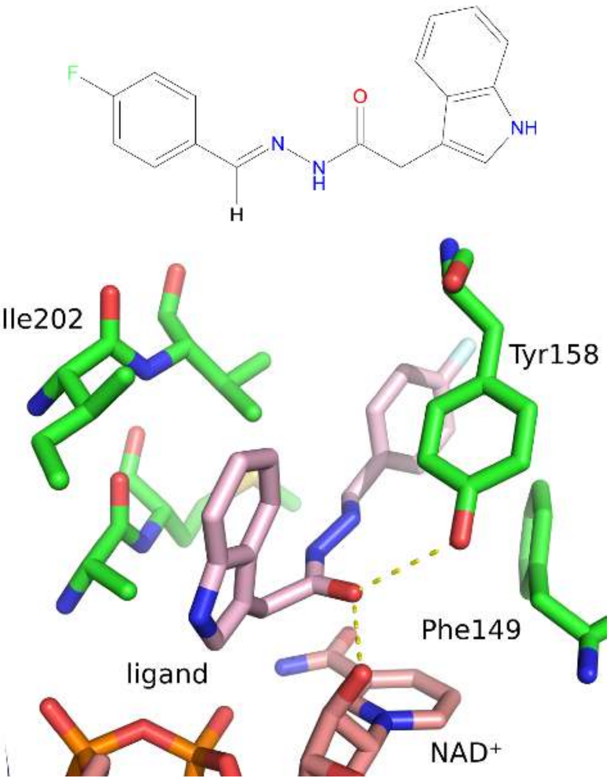


Figure 10.18 Top docking pose of CHEMBL3210727/ZINC00221556 in InhA.

mycobacteria was further investigated in a 30 ns MD simulation. The trajectory was analyzed regarding the previously evaluated distance and RMSD measures (Table 10.8). Whereas the average distance between the ligand and Tyr158 suggests a stable hydrogen bond at 2.81 Å, the average distance between the ligand and NAD$^+$ is by far the highest observed, emphasizing that no stable interaction is maintained between the amide oxygen of the ligand and the ribose hydroxyl group of the cofactor. Interestingly, RMSD calculations of the protein (backbone, SBL backbone and pocket heavy atoms) show stable average values and low fluctuations, similar to those of the rather stable derivatives **mod5**, **mod6**, **mod10** and **mod12**. The ligand RMSD, on the other hand, is rather high with an average RMSD of 2.71 Å, compared to **mod12**, which constitutes the most stable ligand of the evaluated systems. Although the protein behaves stably over 30 ns of simulation time, the ligand is, thus, not entirely stabilized in the predicted top docking pose. Also the heavy atom RMSD of Met161 is higher than the RMSD measured in system **mod12** and rather comparable to the remaining systems.

The maximum free energy change of induced ligand egress was examined using SMD simulations (Table 10.9). With respect to the exponential estimator, the measured maximum free energy change ΔG_e reconstructed using trimmed work profiles is the lowest of the evaluated ligands, except for the derivative **mod6**, which shows a slightly lower

Table 10.8 Average distance and RMSD values of system **ZINC00221556** over 30 ns of MD simulation in Å.

		ZINC00221556
Distance	Avg.	2.81
Tyr158	SD	0.18
Distance	Avg.	6.53
NAD$^+$	SD	0.90
Backbone	Avg.	1.32
RMSD	SD	0.16
SBL RMSD	Avg.	1.78
	SD	0.46
Pocket RMSD	Avg.	1.89
	SD	0.36
Ligand RMSD	Avg.	2.71
	SD	0.43
Ligand RMSD	Avg.	2.87
10 to 30 ns	SD	0.36
Distance	Avg.	4.98
Met161	SD	0.69

Table 10.9 Maximum free energy change of induced ligand extraction of ZINC00221556 with standard deviation in kcal/mol.

	Trimmed avg.	**Untrimmed avg.**
ΔG_e	72.18 ± 2.55	61.12 ± 7.97
ΔG_1	74.51 ± 2.57	74.82 ± 8.00
ΔG_2	69.02 ± 2.55	25.03 ± 7.57

value of ΔG_e. Thus, the simulated maximum free energy change is very low compared to the reference ligand **PT70** and the promising screening candidates **mod10** and **mod12**. Regarding PMF reconstruction using untrimmed work values, ZINC00221556 exhibits the lowest maximum free energy change of all evaluated ligands (cf. Tables 10.7 and 10.9).

10.4 Conclusion

Findings from both parts of this thesis were applied in a screening campaign against the mycobacterial target enoyl-ACP reductase InhA. Based on the docking, MD and SMD results, the presented carbohydrazides might constitute a potential class of InhA inhibitors. In particular, ZINC03526947 derivatives **mod10** and **mod12** are able to stabilize the InhA binding pocket and SBL reliably. Furthermore, these ligands perform very well in SMD simulations with respect to higher maximum free energy barriers ΔG_e

than **PT70**. SMD simulations reveal further that a *para*-chloro-**mod10** might represent a reasonable next step in optimization of the carbohydrazide inhibitors.

Although to date no screening results against *M. tuberculosis* of the similar indole-3-acethydrazide compound ZINC00221556 resulted in proven activity, the examined ligands with enhanced occupation of the hydrophobic pocket of InhA (compared to ZINC00221556) show significantly higher maximum free energy changes in SMD simulations and a more stable hydrogen bonding to the cofactor in MD simulations. Thus, this type of compounds might provide new directions for the development of InhA inhibitors.

Altogether, this Chapter serves as an illustrative example for the combination of knowledge-based permeability classification, docking, MD and SMD simulations to support a virtual screening against *M. tuberculosis* InhA. This approach is generally transferable to other intracellular mycobacterial drug targets.

Chapter 11

Summary – Part II

The largely impermeable cell wall of *M. tuberculosis* is an obstacle in antitubercular drug design. A better understanding of the physico-chemical parameters that contribute to good cell wall permeability would, thus, help define the druggability space of *M. tuberculosis*. Hence, an extensive data mining venture was carried out to find the physico-chemical delimitations of compounds with antimycobacterial activity–which are likely permeable against the *M. tuberculosis* cell wall–from a normally distributed chemical space of drug-like molecules. Based on the molecular descriptor data of both groups, a Principal Component Analysis (PCA) with subsequent logistic regression was conducted. The resulting statistical model was implemented in the free online service MycPermCheck, which can predict the permeability probability of small organic molecules by their physico-chemical composition. Evaluation of the model shows a high predictive power.

In the last chapter, a screening campaign against *M. tuberculosis* was conducted, combining methodologies and findings introduced in both Part I and Part II of this thesis. First, the local stand-alone version of MycPermCheck was used for processing the entire drug-like subset of the ZINC12 database, resulting in a smaller dataset of compounds with high estimated permeability. After multiple further filtering steps based on physico-chemical and ADMET properties and docking, one structure emerged with suitable size, geometry and interaction patterns to bind to the mycobacterial enoyl-ACP reductase InhA. Accordingly, the structure and several derivatives were examined using docking, MD and SMD simulations. Two of the generated derivatives showed promising results in terms of good docking scores, stable interactions in MD simulations and high maximum free energy barriers of induced ligand extraction in SMD simulations. These structures with an indole-3-acethydrazide-scaffold might, hence, constitute a potential class of InhA inhibitors with promising properties, warranting further investigation.

Summary

Molecular Determinants of Drug-Target Residence Times of Bacterial Enoyl-ACP Reductases

Whereas optimization processes of early drug discovery campaigns are often affinity-driven, the drug-target residence time t_R should also be considered due to an often strong correlation with *in vivo* efficacy of compounds. However, rational optimization of t_R is not straightforward and generally hampered by the lack of structural information about the transition states of ligand association and dissociation. The enoyl-ACP reductase FabI of the fatty acid synthesis (FAS) type II is an important drug-target in antibiotic research. InhA is the FabI enzyme of *Mycobacterium tuberculosis*, which is known to be inhibited by various compound classes. Slow-onset inhibition of InhA is assumed to be associated with the ordering of the most flexible protein region, the substrate binding loop (SBL). Diphenylethers are one class of InhA inhibitors that can promote such SBL ordering, resulting in long drug-target residence times. Although these inhibitors are energetically and kinetically well characterized, it is still unclear how the structural features of a ligand affect t_R.

Using classical molecular dynamics (MD) simulations, recurring conformational families of InhA protein-ligand complexes were detected and structural determinants of drug-target residence time of diphenylethers with different kinetic profiles were described. This information was used to deduce guidelines for efficacy improvement of InhA inhibitors, including 5'-substitution on the diphenylether B-ring. The validity of this suggestion was then analyzed by means of MD simulations.

Moreover, Steered MD (SMD) simulations were employed to analyze ligand dissociation of diphenylethers from the FabI enzyme of *Staphylococcus aureus*. This approach resulted in a very accurate and quantitative linear regression model of the experimental $ln(t_R)$ of these inhibitors as a function of the calculated maximum free energy change of induced ligand extraction. This model can be used to predict the residence times of new potential inhibitors from crystal structures or valid docking poses.

Since correct structural characterization of the intermediate enzyme-inhibitor state (EI) and the final state (EI*) of two-step slow-onset inhibition is crucial for rational residence time optimization, the current view of the EI and EI* states of InhA was revisited by means of crystal structure analysis, MD and SMD simulations. Overall, the analyses affirmed that the EI* state is a conformation resembling the 2X23 crystal structure

(with slow-onset inhibitor **PT70**), whereas a twist of residues Ile202 and Val203 with a further opened helix $\alpha 6$ corresponds to the EI state. Furthermore, MD simulations emphasized the influence of close contacts to symmetry mates in the SBL region on SBL stability, underlined by the observation that an MD simulation of **PT155** chain A with chain B' of a symmetry mate in close proximity of the SBL region showed significantly more stable loops, than a simulation of the tetrameric assembly. Closing Part I, SMD simulations were employed which allow the delimitation of slow-onset InhA inhibitors from rapid reversible ligands.

Prediction of *Mycobacterium tuberculosis* Cell Wall Permeability

The cell wall of *M. tuberculosis* hampers antimycobacterial drug design due to its unique composition, providing intrinsic antibiotic resistance against lipophilic and hydrophilic compounds. To assess the druggability space of this pathogen, a large-scale data mining endeavor was conducted, based on multivariate statistical analysis of differences in the physico-chemical composition of a normally distributed drug-like chemical space and a database of antimycobacterial–and thus very likely permeable–compounds. The approach resulted in the logistic regression model MycPermCheck, which is able to predict the permeability probability of small organic molecules based on their physico-chemical properties. Evaluation of MycPermCheck suggests a high predictive power. The model was implemented as a freely accessible online service and as a local stand-alone command-line version.

Methodologies and findings from both parts of this thesis were combined to conduct a virtual screening for antimycobacterial substances. MycPermCheck was employed to screen the chemical permeability space of *M. tuberculosis* from the entire ZINC12 drug-like database. After subsequent filtering steps regarding ADMET properties, InhA was chosen as an exemplary target. Docking to InhA led to a principal hit compound, which was further optimized. The quality of the interaction of selected derivatives with InhA was subsequently evaluated using MD and SMD simulations in terms of protein and ligand stability, as well as maximum free energy change of induced ligand egress. The results of the presented computational experiments suggest that compounds with an indole-3-acethydrazide scaffold might constitute a novel class of InhA inhibitors, worthwhile of further investigation.

Zusammenfassung

Molekulare Determinanten von Wirkstoff-Angriffsziel Verweilzeiten bakterieller Enoyl-ACP Reduktasen

In frühen Phasen der Wirkstoffentwicklung sind Optimierungsprozesse häufig affinitätsgeleitet. Darüber hinaus sollte zusätzlich die Wirkstoff-Angriffsziel Verweilzeit t_R berücksichtigt werden, da diese oft eine starke Korrelation zur *in vivo* Wirksamkeit der Substanzen aufweist. Rationale Optimierung von t_R ist jedoch auf Grund eines Mangels an struktureller Information über den Übergangszustand der Ligandbindung und Dissoziierung nicht einfach umsetzbar. Die Enoyl-ACP Reduktase FabI der Fettsäurebiosynthese (FAS) Typ II ist ein wichtiger Angriffspunkt in der Antibiotikaforschung. InhA ist das FabI Enzym des Organismus *Mycobacterium tuberculosis* und kann durch Substanzen diverser Klassen gehemmt werden. Es wird vermutet, dass Hemmung von InhA durch langsam-bindende ("slow-onset") Inhibitoren mit der Ordnung der flexibelsten Region des Enzyms assoziiert ist, dem Substratbindungsloop (SBL). Diphenylether sind eine InhA Inhibitorenklasse, die eine solche SBL Ordnung fördern und dadurch lange Verweilzeiten im Angriffsziel aufweisen. Obwohl diese Inhibitoren energetisch und kinetisch gut charakterisiert sind, ist noch immer unklar, wie die strukturellen Eigenschaften eines Liganden t_R beeinflussen.

Durch die Verwendung klassischer Molekulardynamik (MD) Simulationen wurden wiederkehrende Konformationsfamilien von InhA Protein-Ligand Komplexen entdeckt und strukturelle Determinanten der Wirkstoff-Angriffsziel Verweilzeit von Diphenylethern mit verschiedenen kinetischen Profilen beschrieben. Anhand dieser Ergebnisse wurden Richtlinien zur Wirksamkeitsoptimierung von InhA Inhibitoren abgeleitet, einschließlich einer 5'-Substitution am Diphenylether B-Ring. Die Validität dieses Vorschlags wurde mittels MD Simulationen nachfolgend analysiert.

Darüber hinaus wurden "Steered MD" (SMD) Simulationen als MD Technik für umfangreicheres Sampling verwendet um die Liganddissoziation von Diphenylethern aus dem FabI Enzym von *Staphylococcus aureus* zu untersuchen. Dieser Ansatz resultierte in einem sehr akkuraten, quantitativen linearen Regressionsmodell der experimentellen Verweilzeit $ln(t_R)$ dieser Inhibitoren als Funktion der berechneten maximalen freien Energieänderung induzierter Ligandextraktion. Dieses Modell kann genutzt werden um die Verweilzeiten neuer potentieller Inhibitoren aus Kristallstrukturen oder validen Dockingposen vorherzusagen.

Die korrekte strukturelle Charakterisierung des intermediären und des finalen Zustandes (EI und EI*-Zustand) eines Enzym-Inhibitor Komplexes bei einem zweistufigen Inhibitionsmechanismus durch langsam-bindende Hemmstoffe ist essentiell für rationale Verweilzeitoptimierung. Daher wurde die gegenwärtige Ansicht des EI und EI*-Zustandes von InhA mittels Kristallstrukturanalyse, MD und SMD Simulationen erneut aufgegriffen. Insgesamt bestätigten die Analysen, dass der EI*-Zustand einer Konformation ähnlich der 2X23 Kristallstruktur (mit langsam-bindenden Inhibitor **PT70**) gleicht, während eine Drehung der Reste Ile202 und Val203 mit einer weiter geöffneten Helix $\alpha 6$ dem EI-Zustand entspricht. Des Weiteren zeigten MD Simulationen den Einfluss naher Kristallkontakte zu Symmetrie-Nachbarn in der SBL Region auf die SBL Stabilität. Dies wird durch die Beobachtung hervorgehoben, dass die Ketten A und B' eines InhA-**PT155**-Komplexes und des angrenzenden Symmetrie-Nachbars, welche in engem Kontakt in der SBL Region stehen, signifikant stabilere SBLs aufweisen, als die Ketten A und B in einer Simulation des Tetramers. Zum Abschluss von Teil I wurden SMD Simulationen angewandt, auf deren Basis es möglich war, langsam-bindende InhA Inhibitoren von schnell-reversiblen ("rapid reversible") Liganden zu unterscheiden.

Vorhersage von *Mycobacterium tuberculosis* Zellwand Permeabilität

Die Zellwand von *M. tuberculosis* erschwert die antimycobakterielle Wirkstofffindung auf Grund ihrer einzigartigen Zusammensetzung und bietet eine intrinsische Antibiotikaresistenz gegenüber lipophilen und hydrophilen Substanzen. Um den chemischen Raum wirkstoffähnlicher Moleküle gegen diesen Erreger ("Druggability Space") einzugrenzen, wurde eine groß angelegte Dataminingstudie durchgeführt, welche auf multivariater statistischer Analyse der Unterschiede der physikochemischen Zusammensetzung eines normalverteilten wirkstoffähnlichen chemischen Raumes und einer Datenbank von antimycobakteriellen – und somit höchstwahrscheinlich permeablen – Substanzen beruht. Dieser Ansatz resultierte in dem logistischen Regressionsmodell MycPermCheck, welches in der Lage ist die Permeabilitätswahrscheinlichkeit kleiner organischer Moleküle anhand ihrer physikochemischen Eigenschaften vorherzusagen. Die Evaluation von MycPermCheck deutet auf eine große Vorhersagekraft hin. Das Modell wurde als frei zugänglicher online Service und als lokale Kommandozeilenversion implementiert.

Methodiken und Ergebnisse aus beiden Teilen dieser Dissertation wurden kombiniert um ein virtuelles Screening nach antimycobakteriellen Substanzen durchzuführen. MycPermCheck wurde verwendet um den chemischen Permeabilitätsraum von *M. tuberculosis* anhand der gesamten ZINC12 Datenbank wirkstoffähnlicher Moleküle abzuschätzen.

Nach weiteren Filterschritten mit Bezug auf ADMET Eigenschaften, wurde InhA als exemplarisches Angriffsziel ausgewählt. Docking nach InhA führte schließlich zu einer Treffersubstanz, welche in darauffolgenden Schritten weiter optimiert wurde. Die Interaktionsqualität ausgewählter Derivate mit InhA wurde daraufhin mittels MD und SMD Simulationen in Bezug auf Protein und Ligand Stabilität, sowie auch der maximalen freien Energieänderung induzierter Ligandextraktion, untersucht. Die Ergebnisse der vorgestellten computerbasierten Experimente legen nahe, dass Substanzen mit einem Indol-3-Acethydrazid Gerüst eine neuartige Klasse von InhA Inhibitoren darstellen könnten. Weiterführende Untersuchungen könnten sich somit als lohnenswert erweisen.

Bibliography

[1] B. Merget and C.A. Sotriffer. Slow-onset inhibition of *Mycobacterium tuberculosis* InhA: Revealing molecular determinants of residence time by MD simulations. *PLoS ONE*, 10(5):e0127009, 2015.

[2] B. Merget and C.A. Sotriffer. An accurate and quantitative prediction model for drug-target residence time of *Staphylococcus aureus* FabI inhibitors based on Steered Molecular Dynamics. Manuscript in preparation.

[3] B. Merget, D. Zilian, T. Müller, and C.A. Sotriffer. MycPermCheck: the *Mycobacterium tuberculosis* permeability prediction tool for small molecules. *Bioinformatics*, 29(1):62–68, 2013.

[4] R.I. Aminov. A brief history of the antibiotic era: lessons learned and challenges for the future. *Frontiers in Microbiology*, 1(134), 2010.

[5] G.S. Bbosa, N. Mwebaza, J. Odda, D.B. Kyegombe, and M. Ntale. Antibiotics/-antibacterial drug use, their marketing and promotion during the post-antibiotic golden age and their role in emergence of bacterial resistance. *Health*, 6(05):410–425, 2014.

[6] J. Davies. Where have all the antibiotics gone? *The Canadian Journal of Infectious Diseases & Medical Microbiology*, 17(5):287–290, 2006.

[7] M. Bassetti and E. Righi. Multidrug-resistant bacteria: what is the threat? *ASH Education Program Book*, 2013(1):428–432, 2013.

[8] E.C. Hett and E.J. Rubin. Bacterial growth and cell division: a mycobacterial perspective. *Microbiology and Molecular Biology Reviews*, 72(1):126–156, 2008.

[9] V. Jarlier and H. Nikaido. Mycobacterial cell wall: structure and role in natural resistance to antibiotics. *FEMS Microbiology Letters*, 123(1-2):11–18, 1994.

[10] M. Daffé. The cell envelope of tubercle bacilli. *Tuberculosis*, 95(Supplement 1): S155–S158, 2015.

[11] D.M. Livermore. Beta-lactamase-mediated resistance and opportunities for its control. *Journal of Antimicrobial Chemotherapy*, 41(Suppl D):25–41, 1998.

[12] C.E. Cade, A.C. Dlouhy, K.F. Medzihradszky, S.P. Salas-Castillo, and R.A. Ghiladi. Isoniazid-resistance conferring mutations in *Mycobacterium tuberculosis* KatG: Catalase, peroxidase, and INH-NADH adduct formation activities. *Protein Science*, 19(3):458–474, 2010.

[13] P.A. Lambert. Cellular impermeability and uptake of biocides and antibiotics in Gram-positive bacteria and mycobacteria. *Journal of Applied Microbiology*, 92 (S1):46S–54S, 2002.

[14] G.L. Archer. *Staphylococcus aureus*: a well-armed pathogen. *Clinical Infectious Diseases*, 26(5):1179–1181, 1998.

[15] World Health Organization. Tuberculosis fact sheet no. 104. `http://www.who.int/mediacentre/factsheets/fs104/en/`, 2015. accessed in March 2015.

[16] A. Koul, E. Arnoult, N. Lounis, J. Guillemont, and K. Andries. The challenge of new drug discovery for tuberculosis. *Nature*, 469(7331):483–490, 2011.

[17] P. Pan and P.J. Tonge. Targeting InhA, the FASII enoyl-ACP reductase: SAR studies on novel inhibitor scaffolds. *Current Topics in Medicinal Chemistry*, 12 (7):672–693, 2012.

[18] E. Nathanson, R. Gupta, P. Huamani, V. Leimane, A.D. Pasechnikov, T.E. Tupasi, K. Vink, E. Jaramillo, and M.A. Espinal. Adverse events in the treatment of multidrug-resistant tuberculosis: results from the dots-plus initiative. *The International Journal of Tuberculosis and Lung Disease*, 8(11):1382–1384, 2004.

[19] H.F. Chambers and F.R. DeLeo. Waves of resistance: *Staphylococcus aureus* in the antibiotic era. *Nature Reviews Microbiology*, 7(9):629–641, 2009.

[20] J. Kluytmans, A. Van Belkum, and H. Verbrugh. Nasal carriage of *Staphylococcus aureus*: epidemiology, underlying mechanisms, and associated risks. *Clinical microbiology reviews*, 10(3):505–520, 1997.

[21] A. Pantosti and M. Venditti. What is MRSA? *European Respiratory Journal*, 34 (5):1190–1196, 2009.

[22] B.A. Diep, H.F. Chambers, C.J. Graber, J.D. Szumowski, L.G. Miller, L.L. Han, J.H. Chen, F. Lin, J. Lin, T.H. Phan, et al. Emergence of multidrug-resistant, community-associated, methicillin-resistant *Staphylococcus aureus* clone USA300 in men who have sex with men. *Annals of Internal Medicine*, 148(4):249–257, 2008.

[23] W. Witte, C. Braulke, C. Cuny, B. Strommenger, G. Werner, D. Heuck, U. Jappe, C. Wendt, H. Linde, and D. Harmsen. Emergence of methicillin-resistant *Staphylococcus aureus* with Panton–Valentine leukocidin genes in central Europe. *European Journal of Clinical Microbiology and Infectious Diseases*, 24(1):1–5, 2005.

[24] H. Lu and P.J. Tonge. Drug-target residence time: critical information for lead optimization. *Current Opinion in Chemical Biology*, 14(4):467–474, 2010.

[25] R.A. Copeland, D.L. Pompliano, and T.D. Meek. Drug-target residence time and its implications for lead optimization. *Nature Reviews Drug Discovery*, 5(9):730–739, 2006.

[26] A.C. Pan, D.W. Borhani, R.O. Dror, and D.E. Shaw. Molecular determinants of drug-receptor binding kinetics. *Drug Discovery Today*, 18(13):667–673, 2013.

[27] R.O. Dror, A.C. Pan, D.H. Arlow, D.W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D.E. Shaw. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(32):13118–13123, 2011.

[28] Y. Shan, E.T. Kim, M.P. Eastwood, R.O. Dror, M.A. Seeliger, and D.E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011.

[29] I. Buch, T. Giorgino, and G. De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10184–10189, 2011.

[30] A. Chang, J. Schiebel, W. Yu, G.R. Bommineni, P. Pan, M.V. Baxter, A. Khanna, C.A. Sotriffer, C. Kisker, and P.J. Tonge. Rational optimization of drug-target residence time: Insights from inhibitor binding to the *S. aureus* FabI enzyme-product complex. *Biochemistry*, 52(24):4217–4228, 2013.

[31] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology*, 11(2):224–230, 2001.

[32] B. Isralewitz, J. Baudry, J. Gullingsrud, D. Kosztin, and K. Schulten. Steered molecular dynamics investigations of protein function. *Journal of Molecular Graphics and Modelling*, 19(1):13–25, 2001.

[33] C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690, 1997.

[34] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56(5):5018, 1997.

[35] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *The Journal of Chemical Physics*, 119(6):3559, 2003.

[36] S. Park and K. Schulten. Calculating potentials of mean force from steered molecular dynamics simulations. *The Journal of Chemical Physics*, 120(13):5946–5961, 2004.

[37] P. Tiwary, V. Limongelli, M. Salvalaglio, and M. Parrinello. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the National Academy of Sciences*, 112(5):E386–E391, 2015.

[38] A.M. Capelli, A. Bruno, A. Entrena Guadix, and G. Costantino. Unbinding pathways from the glucocorticoid receptor shed light on the reduced sensitivity of glucocorticoid ligands to a naturally occurring, clinically relevant mutant receptor. *Journal of Medicinal Chemistry*, 56(17):7003–7014, 2013.

[39] A.M. Capelli and G. Costantino. Unbinding pathways of VEGFR2 inhibitors revealed by steered molecular dynamics. *Journal of Chemical Information and Modeling*, 54(11):3124–3136, 2014.

[40] H. Li, C. Lai, P. Pan, W. Yu, N. Liu, G.R. Bommineni, M. Garcia-Diaz, C. Simmerling, and P.J. Tonge. A structural and energetic model for the slow-onset inhibition of the *Mycobacterium tuberculosis* enoyl-ACP reductase InhA. *ACS Chemical Biology*, 9(4):986–993, 2014.

[41] J. Zhang, Q. Zheng, Z. Li, and H. Zhang. Molecular dynamics simulations suggest ligand's binding to Nicotinamidase/Pyrazinamidase. *PLoS ONE*, 7(6):e39546, 2012.

[42] A.C. Kruse, J. Hu, A.C. Pan, D.H. Arlow, D.M. Rosenbaum, E. Rosemond, H.F. Green, T. Liu, P.S. Chae, R.O. Dror, et al. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature*, 482(7386):552–556, 2012.

[43] Z. Shen, F. Cheng, Y. Xu, J. Fu, W. Xiao, J. Shen, G. Liu, W. Li, and Y. Tang. Investigation of indazole unbinding pathways in CYP2E1 by molecular dynamics simulations. *PloS ONE*, 7(3):e33500, 2012.

[44] S. Decherchi, A. Berteotti, G. Bottegoni, W. Rocchia, and A. Cavalli. The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nature Communications*, 6(6155), 2015.

[45] S.R. Luckner, N. Liu, C.W. am Ende, P.J. Tonge, and C. Kisker. A slow, tight binding inhibitor of InhA, the enoyl-acyl carrier protein reductase from *Mycobacterium tuberculosis*. *Journal of Biological Chemistry*, 285(19):14330–14337, 2010.

[46] P. Pan, S.E. Knudson, G.R. Bommineni, H. Li, C. Lai, N. Liu, M. Garcia-Diaz, C. Simmerling, S.S. Patil, R.A. Slayden, and P.J. Tonge. Time-dependent diaryl

ether inhibitors of InhA: structure–activity relationship studies of enzyme inhibition, antibacterial activity, and *in vivo* efficacy. *ChemMedChem*, 9(4):776–791, 2014.

[47] S.W. White, J. Zheng, Zhang Y., and C.O. Rock. The structural biology of type II fatty acid biosynthesis. *Annual Review of Biochemistry*, 74:791–831, 2005.

[48] H. Lu and P.J. Tonge. Inhibitors of FabI, an enzyme drug target in the bacterial fatty acid biosynthesis pathway. *Accounts of Chemical Research*, 41(1):11–20, 2008.

[49] A. Banerjee, E. Dubnau, A. Quemard, V. Balasubramanian, K.S. Um, T. Wilson, D. Collins, G. de Lisle, and W.R. Jacobs. inhA, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*. *Science*, 263(5144):227–230, 1994.

[50] C.W. Levy, A. Roujeinikova, S. Sedelnikova, P.J. Baker, A.R. Stuitje, A.R. Slabas, D.W. Rice, and J.B. Rafferty. Molecular basis of triclosan activity. *Nature*, 398 (6726):383–384, 1999.

[51] R.J. Heath, M.A. Yu, Y.and Shapiro, E. Olson, and C.O. Rock. Broad spectrum antimicrobial biocides target the FabI component of fatty acid synthesis. *Journal of Biological Chemistry*, 273(46):30316–30320, 1998.

[52] D.A. Rozwarski, G.A. Grant, D.H.R. Barton, W.R. Jacobs, and J.C. Sacchettini. Modification of the NADH of the isoniazid target (InhA) from *Mycobacterium tuberculosis*. *Science*, 279(5347):98–102, 1998.

[53] D.A. Rozwarski, C. Vilchèze, M. Sugantino, R. Bittman, and J.C. Sacchettini. Crystal structure of the *Mycobacterium tuberculosis* enoyl-ACP reductase, InhA, in complex with NAD^+ and a C16 fatty acyl substrate. *Journal of Biological Chemistry*, 274(22):15582–15589, 1999.

[54] S. Rafi, P. Novichenok, S. Kolappan, X. Zhang, C.F. Stratton, R. Rawat, C. Kisker, C. Simmerling, and P.J. Tonge. Structure of acyl carrier protein bound to FabI, the FASII enoyl reductase from Escherichia coli. *Journal of Biological Chemistry*, 281(51):39285–39293, 2006.

[55] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[56] J.S. Oliveira, J.H. Pereira, F. Canduri, N.C. Rodrigues, O.N. de Souza, W.F. de Azevedo, L.A. Basso, and D.S. Santos. Crystallographic and pre-steady-state kinetics studies on binding of NADH to wild-type and isoniazid-resistant enoyl-ACP

(CoA) reductase enzymes from *Mycobacterium tuberculosis*. *Journal of Molecular Biology*, 359(3):646–666, 2006.

[57] The UniProt Consortium. Uniprot: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 2015.

[58] J. Schiebel, A. Chang, H. Lu, M.V. Baxter, P.J. Tonge, and C. Kisker. *Staphylococcus aureus* FabI: inhibition, substrate recognition, and potential implications for *in vivo* essentiality. *Structure*, 20(5):802–813, 2012.

[59] J. Schiebel, A. Chang, B. Merget, G.R. Bommineni, W. Yu, L.A. Spagnuolo, M.V. Baxter, M. Tareilus, P.J. Tonge, C. Kisker, and C.A. Sotriffer. An ordered water channel in *Staphylococcus aureus* FabI: unraveling the mechanism of substrate recognition and reduction. *Biochemistry*, 54(10):1943–1955, 2015.

[60] P. Draper and M. Daffé. The cell envelope of *Mycobacterium tuberculosis* with special reference to the capsule and outer permeability barrier. *In Cole, S.T. (ed), Tuberculosis and the Tubercle Bacillus*, pages 261–273, 2005.

[61] R.J. Heath, J.R. Rubin, D.R. Holland, E. Zhang, M.E. Snow, and C.O. Rock. Mechanism of triclosan inhibition of bacterial fatty acid synthesis. *Journal of Biological Chemistry*, 274(16):11110–11114, 1999.

[62] H. Xu, T.J. Sullivan, J.I. Sekiguchi, T. Kirikae, I. Ojima, C.F. Stratton, W. Mao, F.L. Rock, M.R.K. Alley, F. Johnson, S.G. Walker, and P.J. Tonge. Mechanism and inhibition of *sa*FabI, the enoyl reductase from *Staphylococcus aureus*. *Biochemistry*, 47(14):4228–4236, 2008.

[63] H. Lu, K. England, C. am Ende, J.J. Truglio, S. Luckner, B.G. Reddy, N.L. Marlenee, S.E. Knudson, D.L. Knudson, R.A. Bowen, Kisker C., Slayden R.A., and P.J. Tonge. Slow-onset inhibition of the FabI enoyl reductase from *Francisella tularensis*: residence time and in vivo activity. *ACS Chemical Biology*, 4(3):221–231, 2009.

[64] T.J. Sullivan, J.J. Truglio, M.E. Boyne, P. Novichenok, X. Zhang, C.F. Stratton, H. Li, T. Kaur, A. Amin, F. Johnson, Slayden R.A., Kisker C., and P.J. Tonge. High affinity InhA inhibitors with activity against drug-resistant strains of *Mycobacterium tuberculosis*. *ACS Chemical Biology*, 1(1):43–53, 2006.

[65] E.K. Schroeder, L.A. Basso, D.S. Santos, and O.N. de Souza. Molecular Dynamics Simulation Studies of the Wild-Type, I21V, and I16T Mutants of Isoniazid-Resistant *Mycobacterium tuberculosis* Enoyl Reductase (InhA) in Complex with NADH: Toward the Understanding of NADH-InhA Different Affinities. *Biophysical Journal*, 89(2):876–884, 2005.

[66] M.E. Boyne, T.J. Sullivan, C.W. am Ende, H. Lu, V. Gruppo, D. Heaslip, A.G. Amin, D. Chatterjee, A. Lenaerts, P.J. Tonge, and R.A. Slayden. Targeting fatty acid biosynthesis for the development of novel chemotherapeutics against *Mycobacterium tuberculosis*: evaluation of A-ring-modified diphenyl ethers as high-affinity InhA inhibitors. *Antimicrobial Agents and Chemotherapy*, 51(10):3562–3567, 2007.

[67] K.F.M. Pasqualoto, M. Ferreira, O.A. Santos-Filho, and A.J. Hopfinger. Molecular dynamics simulations of a set of isoniazid derivatives bound to InhA, the enoyl-ACP reductase from *M. tuberculosis*. *International Journal of Quantum Chemistry*, 106(13):2689–2699, 2006.

[68] G. Subba Rao, R. Vijayakrishnan, and M. Kumar. Structure-based design of a novel class of potent inhibitors of InhA, the enoyl Acyl Carrier Protein reductase from *Mycobacterium Tuberculosis*: a computer modelling approach. *Chemical Biology & Drug Design*, 72(5):444–449, 2008.

[69] A. Punkvang, P. Saparpakorn, S. Hannongbua, P. Wolschann, A. Beyer, and P. Pungpo. Investigating the structural basis of arylamides to improve potency against *M. tuberculosis* strain through molecular dynamics simulations. *European Journal of Medicinal Chemistry*, 45(12):5585–5593, 2010.

[70] P. Kamsri, N. Koohatammakun, A. Srisupan, P. Meewong, A. Punkvang, P. Saparpakorn, S. Hannongbua, P. Wolschann, S. Prueksaaroon, U. Leartsakulpanich, et al. Rational design of InhA inhibitors in the class of diphenyl ether derivatives as potential anti-tubercular agents using molecular dynamics simulations. *SAR and QSAR in Environmental Research*, 25(6):473–488, 2014.

[71] N. Homeyer and H. Gohlke. Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. *Molecular Informatics*, 31(2):114–122, 2012.

[72] M. Kaledin, A. Brown, A.L. Kaledin, and J.M. Bowman. Normal mode analysis using the driven molecular dynamics method. ii. an application to biological macromolecules. *The Journal of Chemical Physics*, 121(12):5646–5653, 2004.

[73] W. Balemans, N. Lounis, R. Gilissen, J. Guillemont, K. Simmen, K. Andries, and A. Koul. Essentiality of FASII pathway for *Staphylococcus aureus*. *Nature*, 463 (7279):E3, 2010.

[74] N. Kaplan, M. Albert, D. Awrey, E. Bardouniotis, J. Berman, T. Clarke, M. Dorsey, B. Hafkin, J. Ramnauth, V. Romanov, et al. Mode of action, *in vitro* activity, and *in vivo* efficacy of AFN-1252, a selective antistaphylococcal FabI inhibitor. *Antimicrobial Agents and Chemotherapy*, 56(11):5865–5874, 2012.

[75] D.J. Payne, M.N. Gwynn, D.J. Holmes, and D.L. Pompliano. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature Reviews Drug Discovery*, 6(1):29–40, 2007.

[76] V. Gerusz. Recent advances in the inhibition of bacterial fatty acid biosynthesis. *Annual Reports in Medicinal Chemistry*, 45:295–311, 2010.

[77] S. Escaich, L. Prouvensier, M. Saccomani, L. Durant, M. Oxoby, V. Gerusz, F. Moreau, V. Vongsouthi, K. Maher, I. Morrissey, et al. The MUT056399 inhibitor of FabI is a new antistaphylococcal compound. *Antimicrobial Agents and Chemotherapy*, 55(10):4692–4697, 2011.

[78] V. Gerusz, A. Denis, F. Faivre, Y. Bonvin, M. Oxoby, S. Briet, G. LeFralliec, C. Oliveira, N. Desroy, C. Raymond, et al. From triclosan toward the clinic: discovery of nonbiocidal, potent FabI inhibitors for the treatment of resistant bacteria. *Journal of Medicinal Chemistry*, 55(22):9914–9928, 2012.

[79] C.A. Sotriffer, G. Klebe, M. Stahl, and H. Böhm. Docking and scoring functions/virtual screening. *Burger's Medicinal Chemistry and Drug Discovery*, pages 281–331, 2003.

[80] H. Gohlke and G. Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie International Edition*, 41(15):2644–2676, 2002.

[81] N. Brooijmans and I.D. Kuntz. Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure*, 32(1):335–373, 2003.

[82] M.K. Gilson and H. Zhou. Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure*, 36:21–42, 2007.

[83] B.O. Brandsdal, F. Osterberg, M. Almlof, I. Feierberg, V.B. Luzhkov, and J. Aqvist. Free energy calculations and ligand binding. *Advances in Protein Chemistry*, 66:123–158, 2003.

[84] H. Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935.

[85] A.D. McNaught and A. Wilkinson. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Blackwell Scientific Publications, Oxford, 1997. XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. Last update: 2014-02-24; version: 2.3.3.

[86] V.S. Pande, K. Beauchamp, and G.R. Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

[87] F. Noé and S. Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Current opinion in Structural Biology*, 18(2): 154–162, 2008.

[88] G.R. Bowman, X. Huang, and V.S. Pande. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods*, 49(2):197–201, 2009.

[89] J. Prinz, B. Keller, and F. Noé. Probing molecular kinetics with markov models: metastable states, transition pathways and spectroscopic observables. *Physical Chemistry Chemical Physics*, 13(38):16912–16927, 2011.

[90] P. Bisignano, S. Doerr, M. Harvey, A. Favia, A. Cavalli, and G. De Fabritiis. Kinetic characterization of fragment binding in AmpC β-lactamase by high-throughput molecular simulations. *Journal of Chemical Information and Modeling*, 54(2):362–366, 2014.

[91] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[92] C. Bergonzo, A.J. Campbell, R.C. Walker, and C. Simmerling. A partial nudged elastic band implementation for use with large or explicitly solvated systems. *International Journal of Quantum Chemistry*, 109(15):3781–3790, 2009.

[93] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23 (2):187–199, 1977.

[94] J. Kästner. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942, 2011.

[95] W. Sinko, Y. Miao, C.A.F. de Oliveira, and J.A. McCammon. Population based reweighting of scaled molecular dynamics. *The Journal of Physical Chemistry B*, 117(42):12759–12768, 2013.

[96] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450 (7172):964–972, 2007.

[97] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[98] A.R. Leach. *Molecular modelling: principles and applications.* Pearson Education, Essex, England, 2001.

[99] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.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, 1995.

[100] J.D. Durrant and J.A. McCammon. Molecular dynamics simulations and drug discovery. *BMC biology*, 9(1):71, 2011.

[101] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[102] C.A. Sotriffer. Molecular dynamics simulations in drug design. In *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*, pages 1153–1160. Springer, 2006.

[103] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A.R.H.J. DiNola, and J.R. Haak. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics*, 81:3684–3690, 1984.

[104] A. Brünger, C.L. Brooks, and M. Karplus. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chemical Physics Letters*, 105(5): 495–500, 1984.

[105] R. Kubo, M. Toda, and N. Hashitsume. *Statistical physics II: nonequilibrium statistical mechanics*, volume 31. Springer Science & Business Media, 2012.

[106] G.J. Martyna, D.J. Tobias, and M.L. Klein. Constant pressure molecular dynamics algorithms. *Journal of Chemical Physics*, 101:4177–4189, 1994.

[107] S.E. Feller, Y.H. Zhang, R.W. Pastor, and B.R. Brooks. Constant-pressure molecular-dynamics simulation–the Langevin piston method. *Journal of Chemical Physics*, 103(11):4613–4621, 1995.

[108] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems. *Journal of Chemical Physics*, 98:10089–10092, 1993.

[109] J.W. Ponder and D.A. Case. Force fields for protein simulations. *Advances in Protein Chemistry*, 66:27–85, 2003.

[110] C.I. Bayly, P Cieplak, W. Cornell, and P.A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *Journal of Physical Chemistry*, 97(40):10269–10280, 1993.

[111] W.D. Cornell, P. Cieplak, C.I. Bayly, and P.A. Kollmann. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society*, 115(21):9620–9631, 1993.

[112] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics*, 79:926–935, 1983.

[113] F.H. Stillinger and A. Rahman. Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, 60(4):1545–1557, 1974.

[114] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[115] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, et al. Gaussian 09 Revision C.01, 2009. Gaussian Inc. Wallingford CT 2009.

[116] J. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[117] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, and P.A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[118] B. Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1):275–282, 1995.

[119] J. Kästner and W. Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method:"umbrella integration". *The Journal of Chemical Physics*, 123(14):144104, 2005.

[120] T. Baştuğ, P. Chen, S.M. Patra, and S. Kuyucak. Potential of mean force calculations of ligand binding to ion channels from Jarzynski's equality and umbrella sampling. *The Journal of Chemical Physics*, 128(15):155104, 2008.

[121] A. Laio and F.L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12):126601, 2008.

[122] D. Hamelberg, J. Mongan, and J.A. McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics*, 120(24):11919–11929, 2004.

[123] L.C.T. Pierce, R. Salomon-Ferrer, C. Augusto F. de Oliveira, J.A. McCammon, and R.C. Walker. Routine access to millisecond time scale events with accelerated molecular dynamics. *Journal of Chemical Theory and Computation*, 8(9):2997–3002, 2012.

[124] S.K. Lüdemann, V. Lounnas, and R.C. Wade. How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *Journal of Molecular Biology*, 303(5):797–811, 2000.

[125] H. Vashisth and C.F. Abrams. Ligand escape pathways and (un)binding free energy calculations for the hexameric insulin-phenol complex. *Biophysical Journal*, 95(9):4193–4204, 2008.

[126] L. Rokach. A survey of clustering algorithms. In *Data Mining and Knowledge Discovery Handbook*, pages 269–298. Springer, 2010.

[127] T. Hastie, R. Tibshirani, and J Friedman. *The elements of statistical learning (2nd edition)*, volume 2. Springer, 2009.

[128] J. Shao, S.W. Tanner, N. Thompson, and T.E. Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation*, 3(6):2312–2334, 2007.

[129] G.E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics*, 90(5-6): 1481–1487, 1998.

[130] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

[131] S. Deepayan. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. URL `http://lmdvr.r-forge.r-project.org`.

[132] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2013. R package version 1.14.4.

[133] Raivo K. *pheatmap: Pretty Heatmaps*, 2013. URL `http://CRAN.R-project.org/package=pheatmap`. R package version 0.7.7.

[134] Brenton K. and Curtis S.S. Estimating extensive form games in R. *Journal of Statistical Software*, 56(8):1–27, 2014.

[135] Zhao J.H. gap: Genetic Analysis Package. R package version 1.1-16, http://cran.r-project.org/package=gap, 2015.

[136] Marc J.M. *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*, 2014. URL `http://CRAN.R-project.org/package=AICcmodavg`. R package version 2.00.

[137] D. Adler. *vioplot: Violin plot*, 2005. URL `http://wsopuppenkiste.wiso.uni-goettingen.de/~dadler`. R package version 0.2.

[138] B.J. Grant, A.P.C. Rodrigues, K.M. ElSawy, J.A. McCammon, and L.S.D. Caves. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695–2696, 2006.

[139] L. Skjærven, X. Yao, G. Scarabelli, and B.J. Grant. Integrating protein structural dynamics and evolutionary analysis with bio3d. *BMC Bioinformatics*, 15(1):399, 2014.

[140] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[141] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995.

[142] The PyMOL Molecular Graphics System, Version 1.6.0.0, Schrödinger, LLC.

[143] A.L.P. da Costa, I. Pauli, M. Dorn, E.K. Schroeder, C. Zhan, and O.N. de Souza. Conformational changes in 2-trans-enoyl-ACP (CoA) reductase (InhA) from *M. tuberculosis* induced by an inorganic complex: a molecular dynamics simulation study. *Journal of Molecular Modeling*, 18(5):1779–1790, 2012.

[144] J.D. Durrant, C.A.F. de Oliveira, and J.A. McCammon. POVME: an algorithm for measuring binding-pocket volumes. *Journal of Molecular Graphics and Modelling*, 29(5):773–776, 2011.

[145] H. Böhm and G. Klebe. What can we learn from molecular recognition in protein–ligand complexes for the design of new drugs? *Angewandte Chemie International Edition in English*, 35(22):2588–2614, 1996.

[146] S. Lindert and J.A. McCammon. Dynamics of *Plasmodium falciparum* enoyl-ACP reductase and implications on drug discovery. *Protein Science*, 21(11):1734–1745, 2012.

[147] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, and P.S. Shenkin. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47 (7):1739–1749, 2004.

[148] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, and J.L. Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of Medicinal Chemistry*, 47 (7):1750–1759, 2004.

[149] G. Neudert and G. Klebe. fconv: format conversion, manipulation and feature computation of molecular data. *Bioinformatics*, 27(7):1021–1022, 2011.

[150] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, et al. Amber 10, 2008. University of California, San Francisco.

[151] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, et al. Gaussian 03, revision C. 02, 2008. Gaussian, Inc., Wallingford, CT, 2004.

[152] J. Wang, W. Wang, P.A. Kollman, and D.A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247–260, 2006.

[153] W.C. Still, A. Tempczyk, R.C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112(16):6127–6129, 1990.

[154] J. Srinivasan, M.W. Trevathan, P. Beroza, and D.A. Case. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 101(6):426–434, 1999.

[155] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, et al. Amber 11, 2010. University of California, San Francisco.

[156] J.E. Stone, J.C. Phillips, P.L. Freddolino, D.J. Hardy, L.G. Trabuco, and K. Schulten. Accelerating molecular modeling applications with graphics processors. *Journal of Computational Chemistry*, 28(16):2618–2640, 2007.

[157] J.L. Hintze and R.D. Nelson. Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

[158] C. Bissantz, B. Kuhn, and M. Stahl. A medicinal chemist's guide to molecular interactions. *Journal of Medicinal Chemistry*, 53(14):5061–5084, 2010.

[159] C.C. Valley, A. Cembran, J.D. Perlmutter, A.K. Lewis, N.P. Labello, J. Gao, and J.N. Sachs. The methionine-aromatic motif plays a unique role in stabilizing protein structure. *Journal of Biological Chemistry*, 287(42):34979–34991, 2012.

[160] B.R. Beno, K. Yeung, M.D. Bartberger, L.D. Pennington, and N.A. Meanwell. A survey of the role of noncovalent sulfur interactions in drug design. *Journal of Medicinal Chemistry*, 58(11):4383–4438, 2015.

[161] C.W. am Ende, S.E. Knudson, N. Liu, J. Childs, T.J. Sullivan, M. Boyne, H. Xu, Y. Gegina, D.L. Knudson, F. Johnson, et al. Synthesis and *in vitro* antimycobacterial activity of B-ring modified diaryl ether InhA inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 18(10):3029–3033, 2008.

[162] F. Colizzi, R. Perozzo, L. Scapozza, M. Recanatini, and A. Cavalli. Single-molecule pulling simulations can discern active from inactive enzyme inhibitors. *Journal of the American Chemical Society*, 132(21):7361–7371, 2010.

[163] MOE, Molecular Operating Environment, 2012.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2012.

[164] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, et al. AMBER 12, 2012. University of California, San Francisco.

[165] J. Gore, F. Ritort, and C. Bustamante. Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements. *Proceedings of the National Academy of Sciences*, 100(22):12564–12569, 2003.

[166] P.J. Tummino and R.A. Copeland. Residence time of receptor-ligand complexes and its effect on biological function. *Biochemistry*, 47(20):5481–5492, 2008.

[167] R.D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[168] R.D. Cook and S. Weisberg. *Residuals and influence in regression.* New York: Chapman and Hall, 1982.

[169] K.A. Clarke. A simple distribution-free test for nonnested model selection. *Political Analysis*, 15(3):347–363, 2007.

[170] G.C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, 28(3):591–605, 1960.

[171] B.K. Mai and M.S. Li. Neuraminidase inhibitor R-125489–a promising drug for treating influenza virus: steered molecular dynamics approach. *Biochemical and Biophysical Research Communications*, 410(3):688–691, 2011.

[172] M. Suan Li and B.K. Mai. Steered molecular dynamics–a promising tool for drug design. *Current Bioinformatics*, 7(4):342–351, 2012.

[173] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[174] N. Sugiura. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics-Theory and Methods*, 7(1):13–26, 1978.

[175] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences*, 98(7):3658–3661, 2001.

[176] A.M. Ferrenberg and R.H. Swendsen. Optimized monte carlo data analysis. *Physical Review Letters*, 63(12):1195, 1989.

[177] N. Eswar, B. Webb, M.A. Marti-Renom, M.S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics*, pages 5–6, 2006.

[178] N. Rastogi, C. Frehel, and H.L. David. Triple-layered structure of mycobacterial cell wall: evidence for the existence of a polysaccharide-rich outer layer in 18 mycobacterial species. *Current Microbiology*, 13(5):237–242, 1986.

[179] J. Liu et al. Mycolic acid structure determines the fluidity of the mycobacterial cell wall. *Journal of Biological Chemistry*, 271(47):29545–29551, 1996.

[180] S. Park and A. Bendelac. CD1-restricted T-cell responses and microbial infection. *Nature*, 406(6797):788–792, 2000.

[181] A. Quemard et al. Enzymic characterization of the target for isoniazid in *Mycobacterium tuberculosis*. *Biochemistry*, 34(26):8235–8241, 1995.

[182] X. Hong and A.J. Hopfinger. Construction, molecular modeling, and simulation of *Mycobacterium tuberculosis* cell walls. *Biomacromolecules*, 5(3):1052–1065, 2004.

[183] X. Hong and A.J. Hopfinger. Molecular modeling and simulation of *Mycobacterium tuberculosis* cell wall permeability. *Biomacromolecules*, 5(3):1066–1077, 2004.

[184] S. Ekins et al. A collaborative database and computational models for tuberculosis drug discovery. *Molecular BioSystems*, 6(5):840–851, 2010.

[185] R. Todeschini and V. Consonni. *Handbook of molecular descriptors*, volume 11. Wiley-VCH, Weinheim, 2000. ISBN 3-52-29913-0.

[186] C.A. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1–3):3–25, 1997.

[187] C.A. Lipinski. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, 44(1):235–249, 2000.

[188] D.F. Veber, S.R. Johnson, H. Cheng, B.R. Smith, K.W. Ward, and K.D. Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12):2615–2623, 2002.

[189] P.D. Leeson and B. Springthorpe. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery*, 6(11):881–890, 2007.

[190] J.D. Hughes, J. Blagg, D.A. Price, S. Bailey, G.A. DeCrescenzo, R.V. Devraj, E. Ellsworth, Y.M. Fobian, M.E. Gibbs, R.W. Gilles, et al. Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorganic & Medicinal Chemistry Letters*, 18(17):4872–4875, 2008.

[191] G.R. Bickerton et al. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012.

[192] M.M. Hann. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm*, 2(5):349–355, 2011.

[193] Schrödinger, LCC. QikProp 3.4 User Manual, 2011. URL `http://www.schroedinger.com/`.

[194] L. Ioakimidis, L. Thoukydidis, A. Mirza, S. Naeem, and J. Reynisson. Benchmarking the reliability of qikprop. correlation between experimental and predicted values. *QSAR & Combinatorial Science*, 27(4):445–456, 2008.

[195] P. Artursson and J. Karlsson. Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (caco-2) cells. *Biochemical and Biophysical Research Communications*, 175(3):880–885, 1991.

[196] B. Veronesi. Characterization of the MDCK cell line for screening neurotoxicants. *Neurotoxicology*, 17(2):433–443, 1995.

[197] P. Garberg, M. Ball, N. Borg, R. Cecchelli, L. Fenart, R.D. Hurst, T. Lindmark, A. Mabondzo, J.E. Nilsson, T.J. Raub, et al. In vitro models for the blood-brain barrier. *Toxicology in vitro*, 19(3):299–334, 2005.

[198] C.W. Yap. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7):1466–1474, 2011.

[199] R. Wang, Y. Gao, and L. Lai. Calculating partition coefficient by atom-additive method. *Perspectives in Drug Discovery and Design*, 19(1):47–66, 2000.

[200] P. Ertl, B. Rohde, and P. Selzer. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20):3714–3717, 2000.

[201] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

[202] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987.

[203] J. Shlens. A tutorial on principal component analysis. *Google Research*, 2014. arXiv preprint arXiv:1404.1100, dated: April 7, 2014; Version 3.02.

[204] L.I. Smith. A tutorial on principal components analysis. 2002. URL `http://www.sccg.sk/~haladova/principal_components.pdf`. accessed July 19, 2015.

[205] P.R. Peres-Neto, D.A. Jackson, and K.M. Somers. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4):974–997, 2005.

[206] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874, 2006.

[207] A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[208] J. Oksanen, F.G. Blanchet, R. Kindt, P. Legendre, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M.H.H. Stevens, and H. Wagner. *vegan: Community Ecology Package*, 2013. R package version 2.0-7.

[209] C. Stubben and B. Milligan. Estimating and analyzing demographic models using the popbio package in R. *Journal of Statistical Software*, 22(11), 2007.

[210] Cao et al. ChemmineR: a compound mining framework for R. *Bioinformatics*, 24 (15):1733–1734, 2008.

[211] S. Ananthan et al. High-throughput screening for inhibitors of *Mycobacterium tuberculosis* H37Rv. *Tuberculosis (Edinb)*, 89(5):334–353, 2009.

[212] M. Hohman et al. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov Today*, 14(5–6):261–270, 2009.

[213] J. Sadowski, J. Gasteiger, and G. Klebe. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *Journal of Chemical Information and Computer Sciences*, 34(4):1000–1008, 1994.

[214] V. Jarlier and H. Nikaido. Permeability barrier to hydrophilic solutes in *Mycobacterium chelonei*. *Journal of Bacteriology*, 172(3):1418–1423, 1990.

[215] J. Trias and R. Benz. Permeability of the cell wall of *Mycobacterium smegmatis*. *Molecular Microbiology*, 14(2):283–290, 1994.

[216] M. Laneelle and M. Daffé. Transport assays and permeability in pathogenic mycobacteria. *Methods in Molecular Biology*, 465:143–151, 2009.

[217] J.J. Irwin and B.K. Shoichet. ZINC–a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005.

[218] R.E. Carhart et al. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.

[219] D. Beisser et al. BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*, 26(8):1129–1130, 2010.

[220] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`.

[221] S. Pounds and S.W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003.

[222] A. Gaulton et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 38:D249–D254, 2011.

[223] J.S. Freundlich, F. Wang, C. Vilchèze, G. Gulten, R. Langley, G.A. Schiehser, D.P. Jacobus, W.R. Jacobs, and J.C. Sacchettini. Triclosan derivatives: Towards potent inhibitors of drug-sensitive and drug-resistant *mycobacterium* tuberculosis. *ChemMedChem*, 4(2):241–248, 2009.

[224] X. He, A. Alian, and P.R. Ortiz de Montellano. Inhibition of the *Mycobacterium tuberculosis* enoyl acyl carrier protein reductase InhA by arylamides. *Bioorganic & Medicinal Chemistry*, 15(21):6649–6658, 2007.

[225] X. He et al. Pyrrolidine carboxamides as a novel class of inhibitors of enoyl acyl carrier protein reductase from *Mycobacterium tuberculosis*. *Journal of Medicinal Chemistry*, 49(21):6308–6323, 2006.

[226] M. Muddassar et al. Identification of novel antitubercular compounds through hybrid virtual screening approach. *Bioorganic & Medicinal Chemistry*, 18(18): 6914–6921, 2010.

[227] T. Sing et al. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21 (20):3940–3941, 2005.

[228] J.S. Freundlich et al. Triclosan derivatives: towards potent inhibitors of drug-sensitive and drug-resistant *Mycobacterium tuberculosis*. *ChemMedChem*, 4(2): 241–248, 2009.

[229] T.J. Sullivan et al. High affinity InhA inhibitors with activity against drug-resistant strains of Mycobacterium tuberculosis. *ACS Chemical Biology*, 1(1): 43–53, 2006.

[230] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, and G.R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011.

[231] F. Milletti, L. Storchi, G. Sforna, and G. Cruciani. New and original pk$_a$ prediction method using grid molecular interaction fields. *Journal of Chemical Information and Modeling*, 47(6):2172–2181, 2007.

[232] A.T.M. Serajuddin. Salt formation to improve drug solubility. *Advanced Drug Delivery Reviews*, 59(7):603–616, 2007.

[233] J.B. Baell and G.A. Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010.

[234] E.E. Bolton, Y. Wang, P.A. Thiessen, and S.H. Bryant. Pubchem: integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry*, 4:217–241, 2008.

[235] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B.A. Shoemaker, et al. Pubchem's bioassay database. *Nucleic Acids Research*, 40(D1):D400–D412, 2012.

List of Figures

List of Tables

Abbreviations

ACP	Acyl-Carrier Protein
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
amu	atomic mass unit
AUC	Area under the curve
CNS	Central Nervous System
DMS	Dimethyl sulfide
DNA	Deoxyribonucleic acid
FAS II	Fatty Acid Synthesis II
FEP	Free Energy Perturbation
GBIS	Generalized Born Implicit Solvent
IP	Ionization Potential
IQR	Inter-quartile range
LIE	Linear Interaction Energy
MAD	Median absolute deviation
MD	Molecular Dynamics
MDR-TB	Multidrug-resistant *Mycobacterium tuberculosis*
MRSA	Methicillin-resistant *Staphylococcus aureus*
MSM	Markov State Model
MW	Molecular Weight
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis
PD	Pharmacodynamics
PK	Pharmacokinetics
PMF	Potential of Mean Force
PNEB	Partial Nudged Elastic Band
PSA	Polar Surface Area
QM	Quantum Mechanics
RAMD	Random Accelerated Molecular Dynamics
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuation

RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
SMD	Steered Molecular Dynamics
TB	Tuberculosis
US	Umbrella Sampling
VISA	Vancomycin-intermediate resistant *Staphylococcus aureus*
VRSA	Vancomycin-resistant *Staphylococcus aureus*
XDR-TB	Extensively drug-resistant *Mycobacterium tuberculosis*

Appendix A

SMD simulation protocol summary

The used SMD simulation protocol is straightforward and in principle applicable on any other target and series of ligands. The following summary describes the key steps of the methodology and intends to serve as a how-to for step-by-step reproduction. For details see Chapters 2.3.1.5 and 5.1.

1. Preparation of complexes

   (a) Retrieve protein structure from PDB;

   (b) parameterize ligand and cofactor according to the General Amber Force Field (GAFF) with RESP atom charges;

   (c) assign Amber force field parameters to protein;

   (d) build complex of protein with ligand, cofactor and crystallized water molecules;

   (e) perform a short energy minimization;

   (f) solvate complex and add counter ions to ensure neutrality;

   (g) perform a long energy minimization;

   (h) apply harmonic constraints to all solute atoms (non-water, non-ion);

   (i) heat the system from 100 to 300 K while gradually releasing harmonic constraints over 500 ps in constant-volume box;

   (j) further equilibrate the system for another 500 ps in the NVT ensemble.

2. Random Accelerated MD simulation

   (a) Choose a complex to determine a common exit pathway of ligand series;

   (b) run RAMD simulations in NAMD with constant pressure;

3. Steered MD simulations

   (a) Align the last RAMD snapshot to the last snapshot of each equilibration;

   (b) create a normalized spatial vector for each system from the ligand center of mass of the last snapshot of the equilibrated system to the ligand center of mass of the last snapshot of the RAMD simulation to get the pulling direction;

   (c) apply harmonic constraints on the C_α atoms of the protein to avoid spatial translation;

   (d) apply a large force constant on the SMD spring;

   (e) run several replica simulations of the same complex with constant pressure.

4. Calculation of free energy profiles

   (a) Extract the pulling forces from the NAMD output files;

   (b) integrate the measured forces over the traveled distance of the moving SMD constraint as cumulative sums (CAVE: the NAMD SMD output has the unit pN and uses the conversion factor 1 kcal/mol = 69.479 pN Å);

   (c) apply Jarzynski equality to convert all replica SMD simulations of a complex to the free energy profile;

   (d) extract the maximum value of your free energy profile for each complex.

5. Correlate maximum free energy change to the natural logarithm of the experimental residence time.

Appendix B

Stand-alone version of MycPermCheck

The *perl* source code of the stand-alone version of MycPermCheck 1.1 is illustrated below.

```
#!/usr/bin/env perl
use strict;
#use warnings;

=head1 NAME
    MycPermCheck
=head1 VERSION
    1.1
=cut

=head1 DESCRIPTION
    MycPermCheck predicts the permeability probability of small molecules
    against the Mycobacterium tuberculosis cell wall.

=head1 SYNOPSIS
    ./mycpermcheck1.1 -i <input file> [-s <sort mode> -c <Calculation Mode>]

    sort modes can be n (by name), o (off, order is unchanged) or p
    (by probability, default)

    calculation mode can be a (all, default), f (first molecule of identically
    named), m (average of identically named molecule)
=cut

### Check parameters
my %PAR = @ARGV;
if( !exists $PAR{-i} ) {
    print "MycPermCheck 1.1\nUsage: ./mycpermcheck1.1 -i <input file>
    [-s <sort mode> -c <Calculate Mean of Isomeric Forms>]\n\nsort modes can
    be n (by name), o (off, order is unchanged) or p (by probability,
    default)\n\nCalculate Mean of Isomeric Forms can be y or n (default)\n\n
    For more information use \"perldoc mycpermcheck1.1\"\n\n" and exit;
}

my $isoform = 'a';
$isoform = $PAR{-c} if(exists $PAR{-c} );
if( $isoform ne 'a' and $isoform ne 'f' and $isoform ne 'm' ){
    print "MycPermCheck 1.1\nUsage: ./mycpermcheck1.1 -i <input file>
    [-s <sort mode> -c <Calculate Mean of Isomeric Forms>]\n\nsort modes can
    be n (by name), o (off, order is unchanged) or p (by probability,
     default)\n\nCalculate Mean of Isomeric Forms can be y or n (default)\n\n
     For more information use \"perldoc mycpermcheck1.1\"\n\n" and exit;
}

my $fname2 = $PAR{-i};
if( $fname2 eq '') {
    print "Please specify file!\n\n" and exit;
}

my $sort = '';
$sort = $PAR{-s} if( exists $PAR{-s} );
if( $sort ne 'p' and $sort ne 'n' and $sort ne 'o')  {
    $sort = 'p';
}

### Define descriptor coefficients and centers
my %COEF_qp = ('PISA'      => [7.831499e-05, 250.7896295],
       'FOSA'      => [-5.137187e-05, 236.1673680],
       'QPlogPo/w' => [6.060713e-03, 3.2621865],
       'accptHB'   => [-3.181890e-03, 5.6655750],
       'glob'      => [-7.429219e-02, 0.8191964]);

my %COEF_pad = ('C2SP2'    => [-2.951990e-03, 8.1930000],
'XLogP'    => [-5.353284e-03, 1.8380930],
'TPSA'     => [5.160086e-05, 92.8078396],
'HybRatio' => [5.197880e-02, 0.2705913],
'LOBMAX'   => [-4.130163e-03, 2.1456477]);

=head1 INTERNAL FUNCTIONS
=cut
=over
```

```
=item logreg_qp()
    Method logreg_qp() applies the MycPermCheck logistic regression function
    for a given PC1 coordinate input for the QikProp based model.
=cut

sub logreg_qp {
    my $pc1 = $_[0];
    my $prob = 1/(1+exp(-45.187*$pc1));
    return $prob;
}

=item logreg_pad()
    Method logreg_pad() applies the MycPermCheck logistic regression function
    for a given PC1 coordinate input for the PaDEL based model.
=cut

sub logreg_pad {
    my $pc1 = $_[0];
    my $prob = 1/(1+exp(52.943*$pc1));
    return $prob;
}

=item loadandformat()
    Method loadandformat() reads and checks the input file for file format
    and correctness.
=cut

sub loadandformat {
    my $fname2 = $_[0];

    open( CHECK, "$fname2") or die "File not found!\n\n";
    my $headline = <CHECK>;
    my @HEADER = split(',', $headline);

    my $fileformat = "qikprop";
    if ( $headline =~ /Name/ ) {
$fileformat = "padel";
    }
    elsif( $headline !~ /molecule,/ ) {
print "$fname2 has wrong file format! Please use QikProp or PaDEL CSV
files only.\n" and exit;
```

```
    }
    close CHECK or die "Can't close file! Something went wrong.\n\n";

    ### Check and save indices of wanted descriptors
    my @DESC = ();
    @DESC = qw(FOSA QPlogPo/w PISA accptHB glob) if( $fileformat eq 'qikprop');
    @DESC = qw(LOBMAX TPSA C2SP2 HybRatio XLogP) if( $fileformat eq 'padel');

    my @DESC_ind = ();
    foreach my $desc ( @DESC ) {
push(@DESC_ind, grep { $HEADER[$_] =~ /$desc/ } 0..$#HEADER);
    }
    return ($fileformat,\@DESC_ind,\@HEADER,\@DESC);
}

=item noiso()
    Perform MycPermCheck without averaging of isomeric forms.
=cut

sub noiso {
    my $fname2 = $_[0];
    my $fileformat = $_[1];
    my @DESC_ind = @{$_[2]};
    my @HEADER = @{$_[3]};

    my @OUTPUT = ();
    my %NAMES = ();
    my %DATA_centered = ();

    open( INPUT, "$fname2") or die "File not found!\n\n";
    while( <INPUT> ) {
next if ($_ =~ /molecule/);
next if ($_ =~ /Name/);
my @LINE = split(',', $_);

        if ( $isoform eq 'f' ) {
    next if ( exists $NAMES{$LINE[0]} );
}

next if ( $LINE[1] eq '' );
```

```
my %DATA = ($HEADER[0] => $LINE[0],
    $HEADER[$DESC_ind[0]] => $LINE[$DESC_ind[0]],
    $HEADER[$DESC_ind[1]] => $LINE[$DESC_ind[1]],
    $HEADER[$DESC_ind[2]] => $LINE[$DESC_ind[2]],
    $HEADER[$DESC_ind[3]] => $LINE[$DESC_ind[3]],
    $HEADER[$DESC_ind[4]] => $LINE[$DESC_ind[4]]);

%DATA_centered = ($HEADER[0] => $LINE[0]);
foreach( keys %DATA ) {
    next if ($_ eq 'molecule');
    next if ($_ eq 'Name');
    $DATA_centered{$_} = $DATA{$_}-$COEF_qp{$_}[1] if( $fileformat eq 'qikprop');
    $DATA_centered{$_} = $DATA{$_}-$COEF_pad{$_}[1] if( $fileformat eq 'padel');
}

######################################################################
### MycPermCheck:
### Calculate PC1 coordinate and Calculate logistic regression
my $probability = '';
if( $fileformat eq 'qikprop' ) {
    my $PC1 = 0.1536226 * 0.6510271 * 9.99875 * ($COEF_qp{PISA}
    [0]*$DATA_centered{PISA} + $COEF_qp{FOSA}[0]*$DATA_centered{FOSA} +
    $COEF_qp{'QPlogPo/w'}[0]*$DATA_centered{'QPlogPo/w'} + $COEF_qp{accptHB}
    [0]*$DATA_centered{accptHB} + $COEF_qp{glob}[0]*$DATA_centered{glob});

    $probability = &logreg_qp($PC1);
}
elsif( $fileformat eq 'padel' ) {
    my $PC1 = 0.1649452 * 0.6063332 * 9.99875 * ($COEF_pad{C2SP2}
    [0]*$DATA_centered{C2SP2} + $COEF_pad{XLogP}[0]*$DATA_centered{XLogP} +
    $COEF_pad{TPSA}[0]*$DATA_centered{TPSA} + $COEF_pad{HybRatio}
    [0]*$DATA_centered{HybRatio} + $COEF_pad{LOBMAX}[0]*$DATA_centered{LOBMAX});

    $probability = &logreg_pad($PC1);
}
######################################################################

### Save result
push(@OUTPUT, [$LINE[0],$probability,"$LINE[$DESC_ind[0]],$LINE[$DESC_ind[1]],
$LINE[$DESC_ind[2]],$LINE[$DESC_ind[3]],$LINE[$DESC_ind[4]]"]);
```

```
### Memorize molecule names
$NAMES{$LINE[0]} = 1;
    }
    close INPUT or die "Can’t close file! Something went wrong.\n\n";
    return @OUTPUT;
}


=item iso()
    Perform MycPermCheck with averaging of isomeric forms.
=cut


sub iso {
    my $fname2 = $_[0];
    my $fileformat = $_[1];
    my @DESC_ind = @{$_[2]};
    my @HEADER = @{$_[3]};

    my @OUTPUT = ();
    my @NAMES = ();
    my %DATA = ();
    my %DATA_avg = ();
    my %DATA_centered = ();

    open( INPUT, "$fname2") or die "File not found!\n\n";
    while( <INPUT> ) {
next if ($_ =~ /molecule/);
next if ($_ =~ /Name/);
my @LINE = split(’,’, $_);
next if ( $LINE[1] eq ’’ );

my $name = $LINE[0];
push(@NAMES, $name);
push(@{$DATA{$name}}, {$HEADER[$DESC_ind[0]] => $LINE[$DESC_ind[0]],
       $HEADER[$DESC_ind[1]] => $LINE[$DESC_ind[1]],
       $HEADER[$DESC_ind[2]] => $LINE[$DESC_ind[2]],
       $HEADER[$DESC_ind[3]] => $LINE[$DESC_ind[3]],
       $HEADER[$DESC_ind[4]] => $LINE[$DESC_ind[4]]});

    }
    close INPUT or die "Can’t close file! Something went wrong.\n\n";
    ### Averaging
    foreach my $ligand ( keys %DATA ) {
my $d1 = 0;
my $d2 = 0;
my $d3 = 0;
my $d4 = 0;
my $d5 = 0;
my $nr = (@{$DATA{$ligand}});

foreach my $iso ( @{$DATA{$ligand}} ) {
    $d1 += $iso->{$HEADER[$DESC_ind[0]]};
    $d2 += $iso->{$HEADER[$DESC_ind[1]]};
    $d3 += $iso->{$HEADER[$DESC_ind[2]]};
    $d4 += $iso->{$HEADER[$DESC_ind[3]]};
    $d5 += $iso->{$HEADER[$DESC_ind[4]]};
}
$DATA_avg{$ligand} = {$HEADER[$DESC_ind[0]] => $d1/$nr,
      $HEADER[$DESC_ind[1]] => $d2/$nr,
      $HEADER[$DESC_ind[2]] => $d3/$nr,
      $HEADER[$DESC_ind[3]] => $d4/$nr,
      $HEADER[$DESC_ind[4]] => $d5/$nr};
    }

    my %BEENTHERE = ();

    foreach my $name ( @NAMES ) {
next if ( exists $BEENTHERE{$name} );
foreach ( keys %{$DATA_avg{$name}} ) {
    $DATA_centered{$name}{$_} = $DATA_avg{$name}{$_}-$COEF_qp{$_}[1] if(
     $fileformat eq 'qikprop');
    $DATA_centered{$name}{$_} = $DATA_avg{$name}{$_}-$COEF_pad{$_}[1] if(
     $fileformat eq 'padel');
}

my $probability = '';
if( $fileformat eq 'qikprop' ) {
    my $PC1 = 0.1536226 * 0.6510271 * 9.99875 * ($COEF_qp{PISA}
     [0]*$DATA_centered{$name}{PISA} + $COEF_qp{FOSA}[0]*$DATA_centered{$name}
     {FOSA} + $COEF_qp{'QPlogPo/w'}[0]*$DATA_centered{$name}{'QPlogPo/w'} +
     $COEF_qp{accptHB}[0]*$DATA_centered{$name}{accptHB} + $COEF_qp{glob}
     [0]*$DATA_centered{$name}{glob});
    $probability = &logreg_qp($PC1);
}
elsif( $fileformat eq ’padel’ ) {
    my $PC1 = 0.1649452 * 0.6063332 * 9.99875 * ($COEF_pad{C2SP2}
     [0]*$DATA_centered{$name}{C2SP2} + $COEF_pad{XLogP}[0]*$DATA_centered{$name}
     {XLogP} + $COEF_pad{TPSA}[0]*$DATA_centered{$name}{TPSA} +
     $COEF_pad{HybRatio}[0]*$DATA_centered{$name}{HybRatio} + $COEF_pad{LOBMAX}
     [0]*$DATA_centered{$name}{LOBMAX});

    $probability = &logreg_pad($PC1);
}
#######################################################################

### Save result
push(@OUTPUT, [$name,$probability,"$DATA_avg{$name}{$HEADER[$DESC_ind[0]]},
 $DATA_avg{$name}{$HEADER[$DESC_ind[1]]},$DATA_avg{$name}{$HEADER[$DESC_ind[2]]},
 $DATA_avg{$name}{$HEADER[$DESC_ind[3]]},$DATA_avg{$name}
 {$HEADER[$DESC_ind[4]]}"]);

$BEENTHERE{$name} = 1;
    }

    return @OUTPUT;
}

=item sorting()
    Provides the selectable sorting methods
=cut

sub sorting {
    if( $sort eq ’p’ ) {
$b->[1] <=> $a->[1] or lc($a->[0]) cmp lc($b->[0])
    }
    elsif( $sort eq ’n’ ) {
lc($a->[0]) cmp lc($b->[0]) or $a->[1] <=> $b->[1]
    }
}

###############################################################################
### Start main program
```

```
my ($fileformat,$DESC_ind,$HEADER,$DESC) = &loadandformat($fname2);
if( $isoform eq ’a’ or $isoform eq ’f’) {
    my @OUTPUT = &noiso($fname2, $fileformat, $DESC_ind, $HEADER);
    print "Name,Probability," . join(’,’,@{$DESC}) . "\n";
    foreach( sort sorting @OUTPUT ) {
print "$_->[0]," . sprintf("%.3f", $_->[1]) . ",$_->[2]" . "\n";
    }
}
elsif( $isoform eq ’m’ ) {
    my @OUTPUT = &iso($fname2, $fileformat, $DESC_ind, $HEADER)
     if( $isoform eq ’m’);
    print "Name,Probability," . join(’,’,@{$DESC}) . "\n";
    foreach( sort sorting @OUTPUT ) {
print "$_->[0]," . sprintf("%.3f", $_->[1]) . ",$_->[2]" . "\n";
    }
}
##############################################################################

=back

=head1 TODO
    - "Sort mode off" does not work with "Calculate Mean of Isomeric Forms":
      FIXED in 1.1

=head1 CITE
    Merget et al. (2013) MycPermCheck: The Mycobacterium tuberculosis
     permeability prediction tool for small molecules, Bioinformatics,
     29(1): 62-68.
=cut

=head1 AUTHOR
    Benjamin Merget, 2014
=cut
```

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Fakultät für
CHEMIE UND
PHARMAZIE

Erklärung

Hiermit erkläre ich an Eides statt, dass ich die Dissertation

selbstständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe.

Ich erkläre außerdem, dass diese Dissertation weder in gleicher oder anderer Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Ich habe früher außer den mit dem Zulassungsgesuch urkundlich vorgelegten Graden keine weiteren akademischen Grade erworben oder zu erwerben versucht.

Würzburg, den

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Fakultät für
CHEMIE UND
PHARMAZIE

Erklärung

Hiermit erkläre ich, dass ich in meiner Dissertation

bei Abbildungen aus Journalen das Copyright von den Verlagen bzw. vom Autor eingeholt habe.

Bei Abbildungen aus dem Internet habe ich den entsprechenden Link angegeben.

Würzburg, den