In silico structure-based optimisation of pyrrolidine carboxamides as *Mycobacterium tuberculosis* enoyl-ACP reductase inhibitors

In silico Struktur-basierte Optimierung von Pyrrolidin-Carbonsäureamiden als *Mycobacterium tuberculosis* Enoyl-ACP-Reduktase-Inhibitoren

Dissertation

for

Doctoral degree at the Graduate School of Life Sciences

Julius-Maximilians-Universität Würzburg

Section: Infection and Immunity

Submitted by:

Yogesh Narkhede

from Nasik, India

Würzburg, 2017

Submitted on: ________________________________

Members of the Promotionskomitee:

Chairperson: Prof. Dr. Ulrike Holzgrabe

Primary Supervisor: Prof. Dr. Christoph Sotriffer

Supervisor (Second): Prof. Dr. Caroline Kisker

Supervisor (Third): Prof. Dr. Tanja Schirmeister

Date of Public Defence: ________________________________

Date of Receipt of Certificates: ________________________________

“Live as if you were to die tomorrow. Learn as if you were to live forever.”

-Mahatma Gandhi

All of the presented work was carried out under the supervision of Prof. Dr. Christoph A. Sotriffer at the Institute of Pharmacy and Food Chemistry (University of Würzburg, Germany) between October 2011 and April 2017.

Acknowledgements

To start with, I would like to express my sincere gratitude to my mentor Prof. Dr. Christoph A. Sotriffer for his constant motivation and guidance throughout the entire duration of the project. His guidance consistently allowed me to pursue and explore my ideas while simultaneously taking care that I was not getting lost. Additionally, I would like to thank him for the immense amount of patience and time he has invested in our project and constantly nudging me to bring out the best out of me. A similar word of thanks go to the other members of my thesis supervision committee, Prof. Dr. Caroline Kisker and Prof. Dr. Tanja Schirmeister for their constant support and understanding when the project was going at a steady but nevertheless slow pace.

I am very grateful to Dr. Benjamin Merget and Steffen Wagner for countless scientific (and miscellaneous) discussions. I would also like to thank Dr. Johannes Schiebel for the little talks and sharing of ideas that have mainly culminated in a part of the project. Furthermore I would also like to thank Dr. Thomas B. Adler for his precious suggestions and prudence that was not limited just to science but also arts and history. I would like to express my heartfelt gratitude to Raphael Dives and Steffen Wagner, two of the most helpful and nice guys I have ever met, for always having an open ear to a multitude of problems and for providing valuable suggestions as well as solutions to everybody.

A big round of thanks go to the entire AK Sotriffer – Dr. Michael Hein, Dr. David Zilian, Dr. Sarah Wehle, Dr. Armin Welker, Maximilian Kuhn, Lukas Pason, Dominik Heuler, Matthias Diebold, Daniel Schuster, Thomas Willmes, Manuel Krug, and (again) Raphael Dives, Steffen Wagner, Dr. Thomas B. Adller, and Dr. Benjamin Merget for simply beautiful memories in Würzburg. A special thanks also go to Maximilian Kuhn, Lukas Pason, and Dominik Heuler for the fabulous evening times in Würzburg. The summer BBQ, Weltenburger trip, laser tag, Beer 'n' cheese-wine tasting, and camping trips created a stress free environment for work.

I am also grateful to the Collaborative Research Center (Sonderforschungsbereich; SFB) 630 not only for providing me with a fascinating and interdisciplinary working atmosphere, but also financial backing and allowing me to present my research on numerous occasions.

Finally, I extremely grateful to my parents, my family and all of my friends (Shrikant, Ujjwal, Rahul) for their constant support throughout the past years. Their constant love and support enabled me to travel this far and consistently perform even under stressful conditions without getting lost.

Abstract

The high infection rates and recent emergence of extremely drug resistant forms of *Mycobacterium tuberculosis* pose a significant challenge for global health. The NADH-dependent enoyl-ACP-reductase InhA of the type II mycobacterial fatty acid biosynthesis pathway is a well-validated target for inhibiting mycobacterial growth. InhA has been shown to be inhibited by a variety of compound series. Prominent classes of InhA inhibitors from literature include diaryl ethers, pyrrolidine carboxamides and arylamides which can be subjected to further development. Despite the progress in this area, very few compounds are in clinical development phase. The present work involves a detailed computational investigation of the binding modes and structure-based optimisation of pyrrolidine carboxamides as InhA inhibitors.

With substituents of widely varying bulkiness, the pyrrolidine carboxamide dataset presented a challenge for prediction of binding mode as well as affinity. Using advanced docking protocols and in-house developed pose selection procedures, the binding modes of 44 compounds were predicted. The poses from docking were used in short molecular dynamics (MD) simulations to ascertain the dominant binding conformations for the bulkier members of the series. Subsequently, an activity-based classification strategy could be developed to circumvent the affinity prediction problems observed with this dataset. The prominent motions of the bound ligand and the active site residues were then ascertained using Essential Dynamics (ED). The information from ED and literature was subsequently used to design a total of 20 compounds that were subjected to extensive *in-silico* evaluations. Finally, the molecular determinants of rapid-reversible binding of pyrrolidine carboxamides were investigated using long MD simulations.

Kurzfassung

Hohe Infektionsraten und das Auftreten von multiresistenten Formen von *Mycobacterium tuberculosis* stellen eine große Herausforderung für das globale Gesundheitswesen dar. Die NADH-abhängige Enoyl-ACP-Reduktase des mykobakteriellen Fettsäure-Biosynthesewegs II, InhA, ist ein gut validiertes Target zur Hemmung des mykobakteriellen Wachstums. Es wurde gezeigt, dass InhA durch eine Vielzahl von unterschiedlichen Verbindungsklassen gehemmt wird. Zu den bekanntesten Klassen von InhA-Inhibitoren aus der Literatur gehören Diphenylether, Pyrrolidincarboxamide und Arylamide, die zur weiteren Entwicklung verwendet werden können. Trotz der Fortschritte in diesem Bereich sind sehr wenige Verbindungen in einer klinischen Entwicklungsphase. Die vorliegende Arbeit beinhaltet eine detaillierte computergestützte Untersuchung der Bindungsmodi und die strukturbasierte Optimierung von Pyrrolidincarboxamiden als InhA-Inhibitoren.

Aufgrund von Substituenten mit stark variierendem Raumanspruch stellt der Pyrrolidincarboxamid-Datensatz eine Herausforderung für die Vorhersage von Bindungsmodi und Affinititäten dar. Mit aufwändigen Docking-Protokollen und speziell zu diesem Zweck entwickelten Posen-Auswahlverfahren wurden die Bindungsmodi für 44 Verbindungen vorhergesagt. Die Posen des Dockings wurden in kurzen Molekulardynamik (MD) Simulationen verwendet, um die bevorzugten Bindungskonformationen für die räumlich anspruchsvollen Vertreter des Datensatzes zu ermitteln. Anschließend konnte eine aktivitätsbasierte Klassifizierungsstrategie entwickelt werden, um die in diesem Datensatz beobachteten Probleme in der Affinitätsvorhersage zu umgehen. Die wesentlichen Bewegungen des gebundenen Liganden und der Aminosäuren der Bindetasche wurden daraufhin mit *Essential Dynamics* (ED) ermittelt. Informationen aus der ED-Analyse und der Literatur wurden anschließend verwendet, um insgesamt 20 Verbindungen zu entwerfen, die umfangreichen *in-silico*-Bewertungen unterzogen wurden. Schließlich wurden die molekularen Determinanten der schnell-reversiblen Bindung von Pyrrolidincarboxamiden unter Verwendung von langen MD Simulationen untersucht.

Contents

To my Parents and Grandparents

Chapter 1

Introduction

1.1 Antibiotics discovery and antibiotics resistance

The association of humans and microbes has lasted since time immemorial, that began with the advent of fermentation in the old ages right up to the modern age, where microbes are invaluable for numerous processes. These range from household applications in food (fermentation) and medicine (antibiotics production) to waste management and clean energy (biogas and ethanol production). However, not all of the microbes that play a part in the above processes are beneficial to mankind, with 1% of the microbial population primarily responsible for many diseases not only in humans and animals (e.g., *Mycobacterium sp., Staphylococcus sp., Streptococcus sp.*) but plants as well (e.g., *Agarobacterium sp.*). The loss of life and economic losses caused by bacterial infections runs into billions of dollars per year, with variable global distribution of the respective infectious bacteria [1].

To this end, the treatment of the infections remains a primary point of providing therapy to those affected. Treatment with chemical compounds stood out as a viable means of treating infections. The process of targeting microbial infections with chemical compounds is a more than century-old venture. It began with use of an organoarsenic compound (**Ehrlich 606**), better known as **Arsphenamine** or **Salvarsan** for the treatment of syphilis, in 1910 by Paul Ehrlich and Sahachiro Hata. Salvarsan, also chemically known as 3,3'-Diamino-4,4'-dihydroxy-arsenobenzol, was originally synthesised by Alfred Bertheim, and remained the choice of treatment for syphilis as well as trypanosomiasis [2] until the discovery and emergence of penicillin [3]. The discovery of penicillin heralded the *Golden Age of Antibiotics* (1940s-1960s), when most of the commonly used penicillin related antibiotics were discovered [4–6]. The same era also saw the discovery of diverse antimicrobial agents (β-lactams, aminoglycosides, tetracyclines, macrolides, chloramphenicol, glycopeptides, streptogramines, sulfonamides, and quinolones) [4, 6]. All of these aforesaid classes of antimicrobials encompass a multitude of mechanisms of actions within the microbial cell that can broadly be classified as bacteriostatic or bactericidal.

The golden era on antibiotics also encompassed a common misconception that all of the infections could be treated with antibiotics and that a complete eradication of all diseases could be achieved. However, on the contrary, the irrational and widespread non-therapeutical use in animals culminated in the evolution of antibiotic resistant strains [5, 6]. The problem has been exacerbated by expanded access to all antibiotics as well as incomplete dosage regimens that render bacteria resistant to most of the commonly used antibiotics. In worst cases, the bacteria are deemed **superbugs** that are resistant to antibiotics reserved for serious cases like systemic infections (bacteremia) [7].

An apt case of acquired resistance is *Staphylococcus aureus*, namely the methicillin resistant strain (MRSA), first observed in 1960 [8], which acquired the vanA gene from *Enterococcus faecalis* and became resistant to vancomycin. This strain, vancomycin resistant *S. aureus* (VRSA), represents a significant threat to public health mainly because of its ability to cause nosocomial infections in immunocompromised as well as infect healthy individuals. The infections are also notoriously difficult to treat, since the bacteria no longer respond to the diverse antibiotics of a clinician's arsenal [9, 10]. The widespread marketing and irrational use in lifestock as well as food preservation have certainly made the job of treating bacterial infections even harder [4]. This highlights the danger of spread of these infections globally, and thereby it can be safely said that infectious diseases are no longer limited to developing countries, but becoming a global health concern.

In addition to the acquired resistance, the case of intrinsic resistance also exists mainly in bacteria like *Mycobacterium tuberculosis* and *Pseudomonas aeruginosa.* The former is notorious for exhibiting resistance towards the majority of the antibiotics, owing to the thick ”waxy” mycolic acid coating around the cell wall (cf. Figure 1.1) that acts as a barrier towards most of the lipophilic as well as hydrophilic substances, including putative drugs [11–13]. It also has the ability to exhibit resistance towards antibiotics by means of mutations, the best example being mutations (I21V, I47T, S94A, and M161A) in the catalase-peroxidase (KatG) [14–20], that render it resistant to isoniazide (INH). The latter has a reputation of rapid acquisition of antibiotics resistance even during the course of treatment, with the clinically relevant reasons being cephalosporinase, the extra-membranous porin OprD, and the multi-drug efflux pumps [21, 22]. Other examples of the diverse mechanisms for antibiotics resistance include increased thickness of the cell wall (VRSA), and extension of protein expression profiles (proteome) (MRSA).

1.2 *Mycobacterium tuberculosis*

Mycobacterium tuberculosis and associated *Mycobacterium* species (*M. africanum, M. bovis, Bacillus Calmette Guerin (BCG), M. microti* etc.) are obligatory aerobic and acid fast staining microbes representing a significant global health challenge. Together they form the *Mycobacterium tuberculosis* complex (MTBC), one of the most successful modern

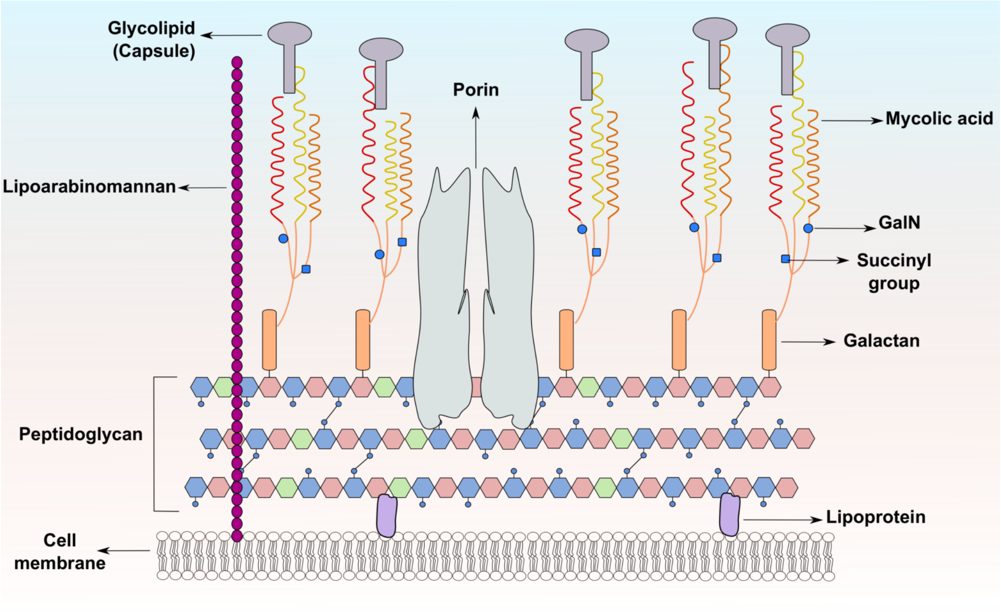


Figure 1.1 Mycobacterial cell wall along with its key components. Figure adapted and redrawn from Kieser, Rubin, 2014 [23].

pathogens, with their rise coinciding with expansion of humans from Africa about 40,000 years ago [24]. It is second only to human immunodeficiency virus (HIV) for mortality resulting from a single infectious agent. The fact that the death rate due to tuberculosis (TB) has fallen over 47% over the past two decades reflects the importance of treating TB [25]. The severity of TB also comes from the fact that 95% of the TB-related deaths were from developing countries, where HIV is also prevalent on a large scale, and that 1 out of 3 HIV related deaths were due to TB. The prevalence of TB infections is another cause of concern, with a total of 9.6 million new cases and 1.2 million deaths in 2014. The co-prevalence of HIV and TB is one of the emerging problems, with HIV infected patients having a 20-30% higher chance of developing TB (including latent TB) than healthy individuals [25]. All of these facts stress the urgent need to identify and treat latent as well as actively diagnosed cases of TB.

The primary anti-tubercular therapy consists of a standard 6 month regimen consisting of four anti-tubercular drugs, namely isoniazide (INH), rifampicin (RFP), pyrazinamide (PZA), and ethambutol (EMB) (cf. Figure 1.2) [26]. However, due to incomplete dosing regimes as well as non-compliance and economic factors, several strains of TB resistant to the frontline drugs have emerged [27]. The diseases due to the resistant strains are multi-drug resistant tuberculosis (MDR-TB or Vank's disease) and extensively drug resistant tuberculosis (XDR-TB). Of these, the former disease is caused by mycobacteria that are resistant to at least two of the frontline drugs, namely isoniazide and rifampicin [28]. An extension of the resistance is manifested in XRD-TB, wherein the mycobacteria are resistant to frontline drugs as well as fluoroquinolones and any injectable anti-tubercular drugs [29]. Furthermore, the most severe form of drug resistant mycobacteria have been

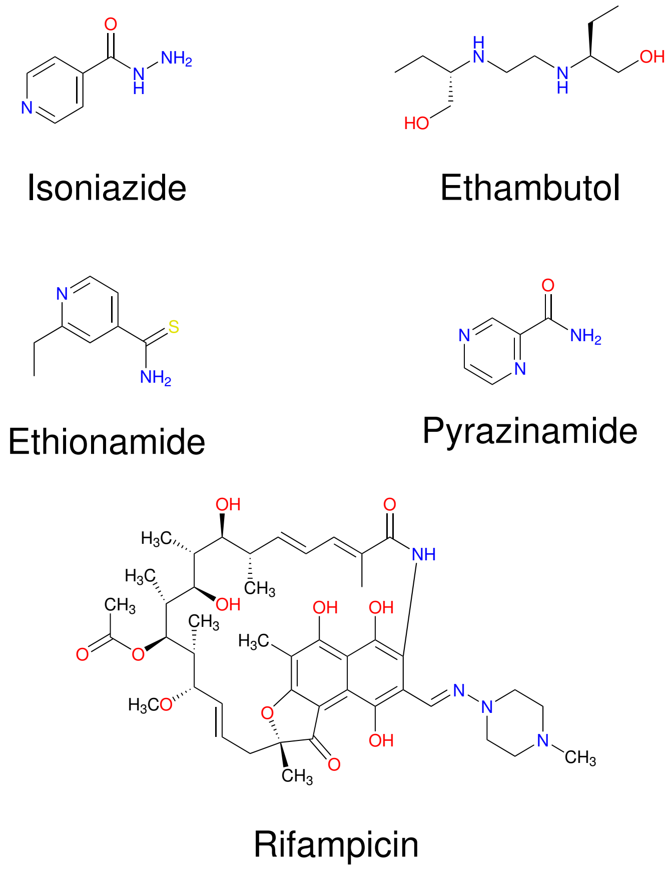


Figure 1.2 Anti-tubercular drugs used primarily under directly observed treatment shortcourse (DOTS).

reported in form of totally drug resistant TB (TDR-TB) [30–32]. The drug-resistant forms of TB are associated with high mortality as well as higher medical costs for treatment. Thus, there is a clear requirement of novel potent anti-tubercular agents that notwithstanding the antibiotics resistance are able to curb the mycobacterial infections.

1.3 Targeting *Mycobacterium tuberculosis*

As mentioned earlier, not only are newer anti-tubercular agents needed for treating TB, but also new molecular targets, mainly to circumnavigate the antibiotics resistance. The central problem lies in the fact that the mycobacteria have an extremely slow growth rate. This in turn reflects a low cellular flux, that makes identification and targeting of essential targets quite challenging. In the current post-genomic area, the advances in whole genomic sequencing and complementary studies have enabled target profiling *en masse*. The collaborative efforts of the Tuberculosis Structural Genomics Consortium (TBSGC) have resulted in a total of 118 unique crystal structures in addition to 139 contributed by the academic and scientific sources [33, 34]. These protein structures were obtained from actively dividing as well as dormant bacteria thereby assisting in the task of inhibiting mycobacterial growth by structure-based and combinatorial chemistry approaches. Some of the principal, known as well as new targets recently profiled are summarised in Figure 1.3.

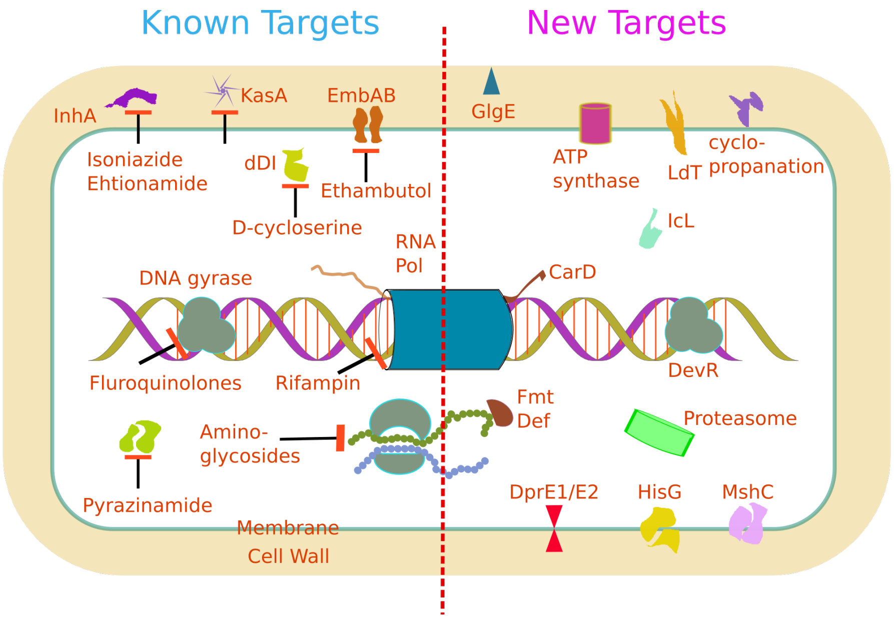


Figure 1.3 Existing and new molecular targets in *Mycobacterium tuberculosis*; figure adapted and redrawn from Lamichhane, 2011 [33].

1.4 Binding Paradigms

One of the central issues any structure-based optimisation protocol faces in its nascency is the accurate representation of the binding modes of new ligands in well defined protein structures. Many of the key parameters associated with protein-ligand association like binding affinity and residence time depend upon the appropriate representation of the binding mode during their computational evaluation. The fundamental assumption behind all such calculations is that the reference structure used represents the bulk of conformations that the system can adopt. Normally, the crystal structure is used in such studies and it is assumed that it represents the dominant conformation for a given system from amongst an ensemble of conformations which it can populate. The binding of ligands to such a conformation usually involves two mechanisms, which are historically well known, namely Fischer's "lock-key" model [35] and Koshland's "induced-fit" model [36], with the kinetics being described in Equations (1.1) to (1.6) [37, 38].

$$\text{Receptor (R)} + \text{Ligand (L)} \underset{\mathrm{k_2}}{\overset{\mathrm{k_1}}{\rightleftarrows}} \text{Receptor-ligand complex (R-L)} \tag{1.1}$$

$$K_d = \frac{[R][L]}{[RL]} = \frac{k_2}{k_1} \tag{1.2}$$

$$\text{residence time} = \frac{1}{k_2} \tag{1.3}$$

$$R + L \underset{\mathrm{k_2}}{\overset{\mathrm{k_1}}{\rightleftarrows}} \text{intermediate R-L'} \underset{\mathrm{k_4}}{\overset{\mathrm{k_3}}{\rightleftarrows}} \text{final R-L*} \tag{1.4}$$

$$k_{off} = \frac{k_2 \cdot k_4}{(k_2 + k_3 + k_4)} \tag{1.5}$$

$$k_2 >> k_3 \text{ and } k_4 \implies k_{off} \approx k_4 \rightarrow \text{residence time} = \frac{1}{k_4} \tag{1.6}$$

More recently, based upon the energy landscape theory of protein structure and dynamics [39–42], a new ”conformational selection” theory of binding has been put forward (Figure 1.4). This theory also takes into consideration the role of metastable states other than the ”native” (low energy state) in driving molecular recognition. The theory proposes that weakly populated ”high energy” conformers are primarily responsible for molecular recognition and binding that is accompanied with a subsequent population shift towards these conformers [43]. Simply put forward, the ”conformational selection” and ”population shift” models suggest that a ligand may interact not only with the ”native” low energy conformation, but also with a singular or multiple high energy conformational substates that are populated in solution. Thus, binding interaction merely does not involve a ”conformational” change as is the case with ”induced-fit” theory, but is rather accompanied by a ”population shift” within the pre-existing conformational substates in solution [44]. The ”lock and key” model thereby cannot represent the molecular recognition process for the ligand binding governed by induced fit or conformational selection processes. Hence, the ligand-receptor binding needs to be studied in much detail (and not solely in a static state) in order to better understand the relationship between the ligand efficacy and its binding kinetics.

Accordingly, the effective sampling of the transition from weakly bound to tightly bound states (induced-fit) and the population shift among the conformational substates (conformational selection) is necessary to depict the protein-ligand association. One of the principal theoretical methods to access the conformational changes in the protein-ligand pair is molecular dynamics (MD). The technical advances in computational hardware (e.g. Anton, GPU acceleration by CUDA) as well as molecular simulation (enhanced sampling, implicit solvent models) have made it possible to sample conformational changes that were previously difficult to achieve [45, 46]. Thereby, classical all-atom explicit MD simulations offer valuable insights into the transition states of biologically important processes like protein folding or even ligand binding [47, 48]. Furthermore, the population

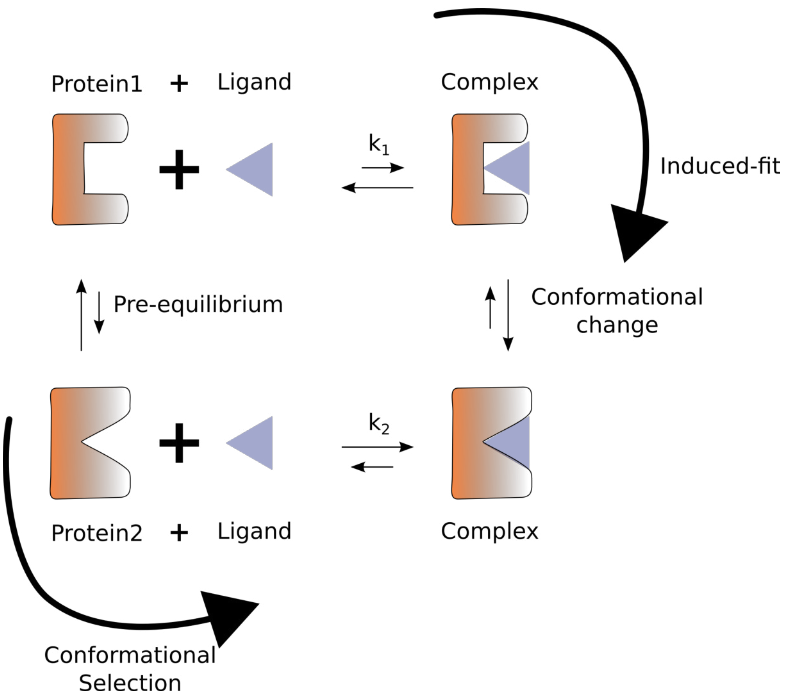


Figure 1.4 A thermodynamic protein-ligand binding cycle describing molecular recognition, with rate constants k_1 and k_2 determining whether ligand binding takes place via conformational selection or induced-fit. Figure adapted and redrawn from Boehr, Nussinov, et al., 2009 [44].

shift can be modelled *in silico* using an ensemble of structures derived from a single structure, using MD, normal mode analysis or ensemble docking [44]. The process of ligand binding can thereby be studied in detail by combining molecular docking, MD simulations and enhanced sampling methods. Alternatively, the redistribution of the protein conformational substates can be directly studied in solution using X-ray crystallography or NMR [49].

1.5 Aims of this work - Binding mode and activity prediction strategies together with Essential Dynamics to drive structure-based optimisation of anti-tubercular molecules

InhA or the mycobacterial FabI (Section 2.1.2) is one of the established targets for inhibiting the growth and proliferation of the mycobacteria. Indeed, InhA is one of the better validated and studied molecular targets with a variety of small molecule inhibitors reported in the literature [50, 51]. Of these InhA inhibitors, pyrrolidine carboxamides and diphenyl ethers are some of the established classes with a sizeable number of compounds and well defined inhibition activities. Typically, the pyrrolidine carboxamides have modest InhA inhibitory activity as compared to the potent and long lasting inhibition for diphenyl ethers. The diphenyl ethers have been thoroughly studied. Abundant structural and binding kinetics information for these are available in the literature.

On the contrary, limited structural information is available for pyrrolidine carboxamides. Just 5 crystal structures for representative compounds have been published up to date [52]. The detailed aspects of their binding to InhA, for e.g., the mechanism of action, binding kinetics etc. have also not been reported so far. Therefore, the aims of this work are:

1. Binding mode prediction for pyrrolidine carboxamides by molecular docking (Chapter 3).
2. Binding affinity prediction and activity-based classification using poses derived from molecular docking (Chapter 4).
3. Dihedral angle analysis and essential dynamics to steer structure-based optimisation of pyrrolidine carboxamides (Chapters 7 to 9).
4. Comparing the dynamic aspects of binding for pyrrolidine carboxamides and diphenyl ethers to reveal conformational changes governing the molecular recognition process (Chapter 10).

1.5.1 Approaches to decipher binding and optimisation of InhA inhibitors

The current work comprises two parts. Each part focusses on the various binding aspects of pyrrolidine carboxamides as putative *Mycobacterium tuberculosis* InhA inhibitors. Despite the fact that InhA is a noteworthy and well validated target for anti-tubercular drug design, even after two decades of research, much of the molecular processes governing the ligand binding and in turn the residence time are unclear. A clear understanding of such processes would effectively support in rational optimisation of the binding affinity as well as residence time. Thus, Part I focusses on the binding mode determination and affinity prediction, activity-based classification with pyrrolidine carboxamides having unknown binding modes. The activity-based classification is expected to provide for a fast separation of binders (actives) from non-binders (inactive/least-actives) in a virtual screening setting.

The diversity in binding and the underlying molecular interactions in case of the various InhA inhibitors suggest different binding kinetics for the respective classes. However, the molecular determinants governing the binding kinetics of pyrrolidine carboxamides as InhA inhibitors have not been widely studied. Accordingly, extensive MD simulations were utilised to reveal the dynamics of the molecular determinants governing binding for pyrrolidine carboxamides. Furthermore, by comparing the dynamical information from the pyrrolidine carboxamides with slow tight-binding inhibitors, the differences amongst intermediate conformations involved in binding (EI and EI*) for the respective classes can be studied in detail (cf. Chapter 10).

Accordingly, Part II of this thesis focusses on revealing the molecular determinants that govern the binding of pyrrolidine carboxamides to InhA. Additionally, a principal component analysis was performed to obtain valuable insights into the direction and extent of maximal variations in the system (cf. Chapter 8). The accumulated information was then utilised together with structural information from slow tight binders to drive the structure-based optimisation of the pyrrolidine carboxamide scaffold. The characteristics of the resultant molecules were thoroughly assessed using molecular docking and MD simulations, along with an evaluation of their mycobacterial cell wall permeability using MycPermcheck 1.2 [53], all while considering results of Part I (cf. Chapter 9).

Part I

Binding Mode Determination and Activity prediction/Activity-based Classification

Chapter 2

Background

2.1 FAS II and Mtb FabI

2.1.1 Mycolic acids and fatty acid synthase system

Mycolic acids are a corresponding series of C_{50}-C_{60} α-alkyl-β-hydroxy fatty acids mainly produced in all bacteria from the family *Mycobacteriaceae*, with shorter chains being produced by *Corynebacterium* and *Nocardia* genera. Though these exist in various forms, the trehalose dimycolate (DTM) or the "cord factor" is the most abundant and toxic lipid found on the cell surface of virulent strains [54]. A significant proportion of mycolic acid is composed of bigger meromycolate (up to C_{56}) and a smaller α-mycolate form (C_{24}-C_{26}), although other forms like methoxy- and keto-mycolates do exist [11] (Figure 2.1). The mycolic acids confer several useful characteristics like protection from harsh chemicals, dehydration resistance, lowered permeability to lipophilic as well as hydrophilic substances that include antibiotics, virulence [55–57], biofilm formation [58], and an ability to persist within host macrophages [55, 59]. The biosynthesis of mycolates comprises two pathways that in turn utilise two types of fatty acid synthesising systems, the type I (FAS I) and type II fatty acid synthases (FAS II). The FAS I system resembles the eukaryotic FAS I system, which is primarily involved in de-novo fatty acid synthesis

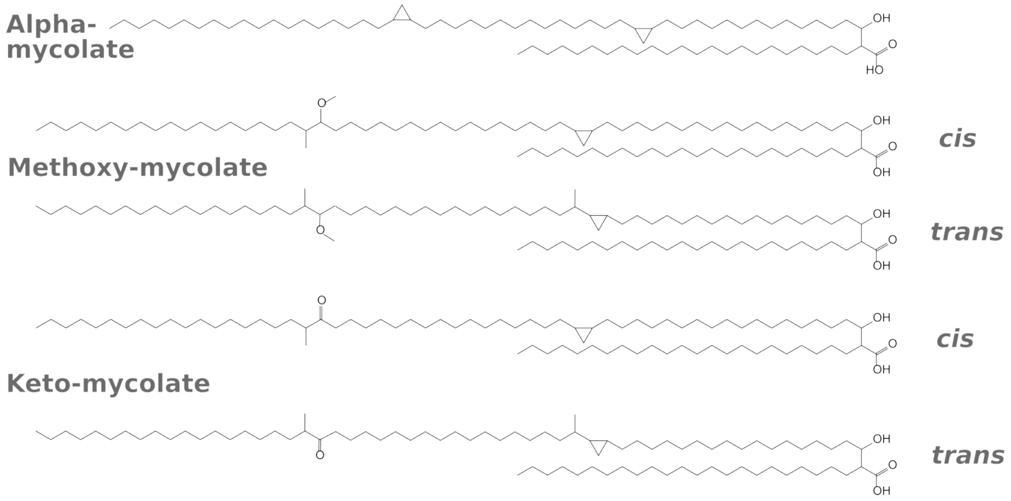


Figure 2.1 Various mycolate forms of α-mycolic acid in *Mycobacterium tuberculosis*, adapted and redrawn from Glickman & Jacobs, 2001 [60].

from acetyl CoA. All mammals including humans, completely lack the FAS II system which is quite prevalent in plants, apicomplexa parasites and bacteria forming an alternate path for fatty acid synthesis [61–63]. This means that targeting the FAS II pathway in mycobacteria would be beneficial owing to lack of target homologues in humans.

The FAS II system of mycobacteria is primarily made up of four dissociable enzymes that act in repetitive succession, resulting in the elongation of the acyl chain by 2 carbon atoms per round. The process initiates with a continual acyl primer activation via a thioester link to the prosthetic group of Coenzyme A (CoA) for FAS-I and acyl carrier protein (ACP) for FAS II. Initially, the malonyl-CoA is converted to malonyl-ACP by the malonyl-CoA:ACP transacylase (FabD). The elongating acyl chain then condenses with malonyl-ACP (from FAS I) forming β-ketoacyl-ACP, in a mtb-FabH (β-ketoacyl-ACP synthase I) catalysed step. FabH mediates the entry of the fatty acids from FASI to FASII. The β-ketoacyl-ACP then undergoes a NADPH dependent-β-ketoacyl-ACP reductase (FabG) catalysed reduction to β-hydroxyacyl-ACP, which undergoes dehydration carried out by β-hydroxyacyl-ACP dehydratase (FabZ). This is followed by an ultimate step wherein the NADH/NADPH-dependent enoyl-ACP reductase (FabI) catalyses the hydrogenation of the substrate to acyl-ACP. The acyl chain now undergoes subsequent rounds of elongation catalysed by the β-ketoacyl-AcpM synthases KasA and KasB, respectively (cf. Figure 2.2).

A key feature of the above cycle is that it is composed of several essential proteins that completely lack homologues in humans, making them attractive targets for drug design. Indeed, targeting the bio-molecular synthesis machinery of the mycolic acids affects the structural integrity as well as permeability of the mycobacterial cell wall. The end effect of this is cell lysis due to a variety of mechanisms [64].

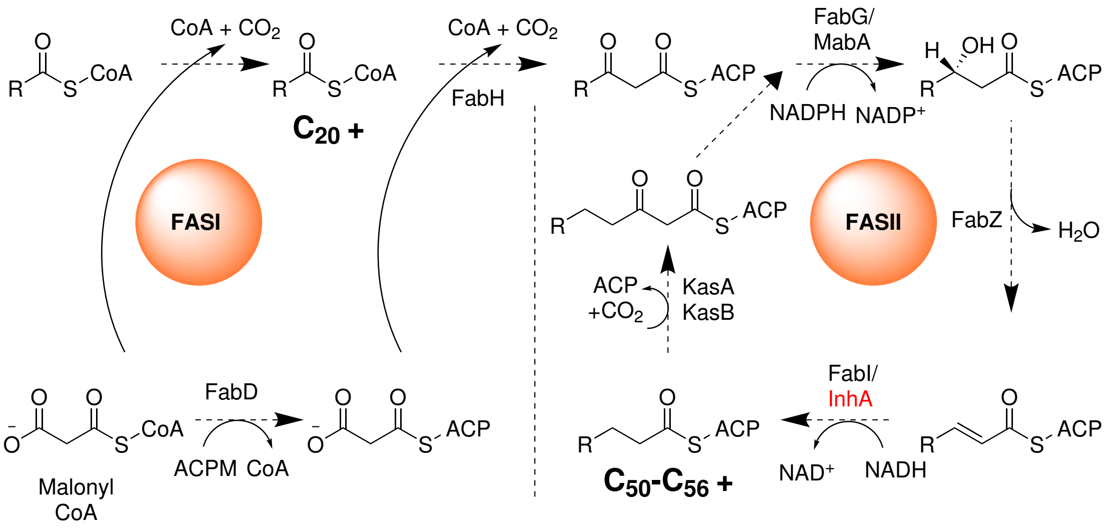


Figure 2.2 Fatty acid synthesis (FAS) II pathway of *Mycobacterium tuberculosis*, adapted and redrawn from Pan, Tonge, 2012 [50].

2.1.2 The Enoyl ACP reductase of *Mycobacterium tuberculosis*-InhA

Of the many targets that FAS II offers, the mycobacterial FabI, commonly referred to as InhA (Figure 2.3), is an attractive and well studied target. Mtb-InhA belongs to the short chain dehydrogenase family of reductases that are a constituent of the dissociated FAS II pathway. It plays a key role in the fatty acid chain elongation process eventually resulting in synthesis of the "waxy" mycolic acids. Mtb-InhA derives its name from Isoniazide (INH), which gets activated in intracellular environment by a catalase-peroxidase system (KatG) to a highly potent and reactive radical that forms the isonicotinic-acyl NADH adduct (K_i 0.75 nM) [16] . This adduct exhibits slow-tight binding with InhA with a very small k_{off} (0.017/min).

InhA is bioactive as a homotetramer, with each of the monomeric units displaying a classical Rossmann fold that binds the cofactor in its reduced state [65]. The binding pocket consists of a catalytic triad of Phe149, Tyr158 and Lys165 along with Phe97 that acts as a gatekeeper apart from residues that lie along the α6 and α7 helices, all of which play a key role in ligand binding. Both α6 and α7 helices constitute a flexible portion of the protein, the substrate binding loop (SBL), which closes upon binding of the substrate (Figure 2.3). The cofactor, which in most crystal structures has been resolved in its oxidised state (NAD^+) lies at the bottom of the binding site, with the nicotinamide part facing the interiors of the pocket and adenine facing the exterior part. The binding pocket has both of its ends exposed to solvent which are referred to as major and minor portal, respectively (Figure 2.4).

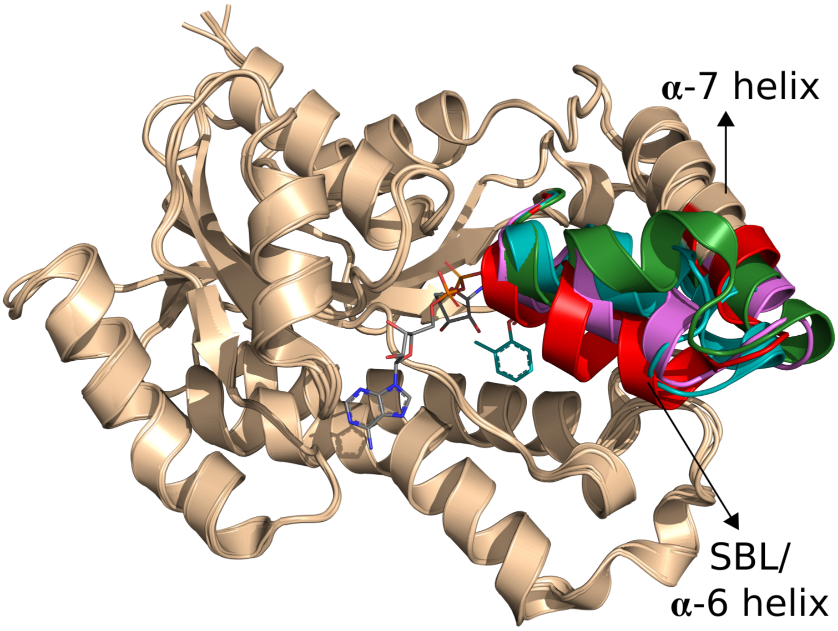


Figure 2.3 C-α aligned structures of *Mycobacterium tuberculosis* InhA, with the substrate binding loop (α6 helix); 2X23 (red), 2NSD (violet), 2H7M/4TZK (teal), and 1P44 (baggy green). Also shown are the cofactor (NAD^+; grey sticks) and PT-70 (cyan) ligand of 2X23 as well as the α7 helix.

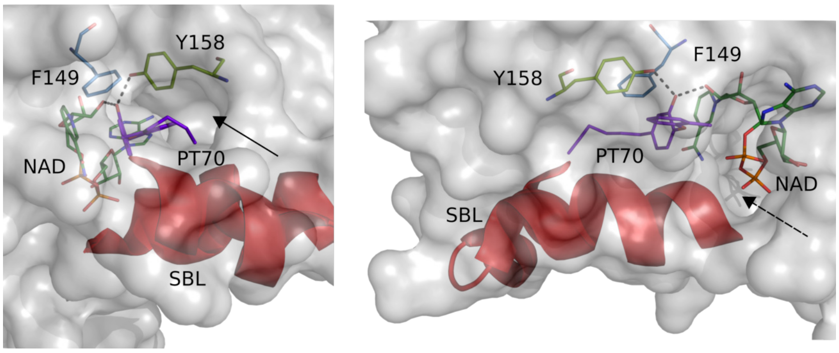


Figure 2.4 Major and minor exit portals of InhA: The picture on the left hand side depicts the minor exit portal (black arrow). The one on right hand side depicts the major exit portal (dashed arrow). The substrate binding loop ($\alpha 6$ and $\alpha 7$ helices) of 2X23 has been coloured red. Also shown are the cofactor (NAD^+; grey sticks), Phe149 (marine sticks), Tyr158 (lime coloured sticks), and the ligand of PDB 2X23 (PT-70; violet). The conserved hydrogen bonds appear as black dashed lines.

Furthermore, InhA has several unique characteristics [66] that set it apart from other bacterial enoyl ACP reductases (ENRs) like:

1. It can handle a variety of enoyl ACP's, with the range of alkyl chain being C18-C56 on average as compared to maximum C18-C20 for other homologous bacterial proteins. This highlights the exceptional flexibility of the protein and thereby the SBL in accommodating very long alkyl chains.

2. The inherent flexibility of InhA also leads it to have a deeper cleft than other ENRs, that enables accommodation of long-chain substrates for mycolic acid biosynthesis.

2.1.3 Inhibiting *Mycobacterium tuberculosis* InhA

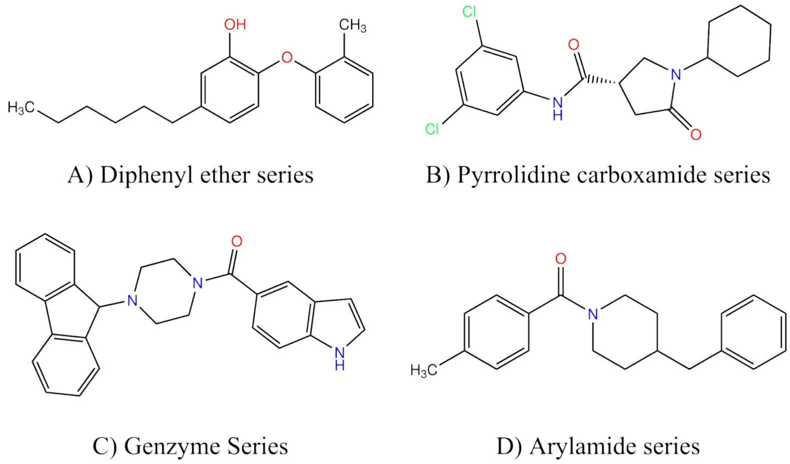
The central theme of InhA inhibitors lies in the disruption and blockade of reduction of the unsaturated precursors of mycolic acid by enoyl ACP reductase. A key feature displayed by majority of InhA inhibitors in the literature is that they sidestep the KatG activation step, unlike isoniazide. Of the many published inhibitors, the focus clearly lies on those which display similar binding characteristics like:

1. Form dual H-bonds with Tyr158 and the oxidised cofactor (NAD^+).

2. Show strong van der Waals contacts with Phe149, Ala198, Met199, Ile202, and Val203.

3. Exhibit some sort of $\pi - \pi$ stacking with the nicotinamide ring of the cofactor as an additional stabilising factor upon ligand binding.

From amongst the several published inhibitors of InhA (Figure 2.5), diphenyl ethers (DPE), 4-hydroxy-2-pyridones (PYR), and pyrrolidine carboxamides (PC) are among the most prominent classes. The former classes have been derived from a widely used anti-microbial agent, triclosan (**TCL**), and they represent some of the most potent inhibitors of InhA till date [50, 67]. The last class has been found not only to be pharmacologically active as moderately potent InhA inhibitor [52], but also exhibiting anti-diabetic/anti-obesity [68], anti-viral [69], anti-cancer [70] and anti-inflammatory effects [71]. A key difference in between the two classes is their ability to bring about the ordering of SBL residues, that ultimately contributes to their binding affinity and residence times. This is evident in states of the SBL in the crystal structures of the respective ligands; **PT70** (PDB 2X23, SBL closed [72]), **TCL** (PDB 2B35, SBL unresolved [73]), **pc-d11** (PDB 4TZK, SBL partially closed [52]). Thus, the SBL conformation in 2X23 can be said to be the final state that most slow tight binders (PT70 and related structures) exhibit, while the latter cases stand for partially ordered/disordered states reflecting a destabilised EI* state [74].



A) Diphenyl ether series

B) Pyrrolidine carboxamide series

C) Genzyme Series

D) Arylamide series

Figure 2.5 Principal chemical series of Mtb-InhA inhibitors considered in molecular docking in the current work.

Although all of the aforesaid facts have been experimentally documented, little is known what triggers the loop ordering or the kinetics behind the same. The structural features of a ligand have a profound effect on the binding kinetics/residence time. This is clearly evident from the case of **PT70** and **6PP**, the former is a slow tight binder with a low K$_i$ and a high residence time, while the latter lags behind the former in both aspects. Given that both ligands differ by a single methyl group [72, 73], the effect of structural features of ligands on their binding kinetics cannot be overlooked in rational modulation of binding affinity.

The pursuit of novel scaffolds with improved binding affinities against InhA has been subject of numerous computational and traditional studies [75–78]. These studies were performed with the apo-enzyme, cofactor-bound protein and a complete protein-cofactor-ligand complex, respectively. These studies assessed the changes in protein conformations and the binding free energies upon ligand binding. The binding kinetics which aid in revealing the protein-ligand associations was however not analysed in these studies, implying their limited utility in describing the process of protein-ligand binding.

The recent studies involving design and optimisation of novel scaffolds against InhA included isoniazide [79], small peptides [80], arylamides [81] and diphenyl ethers [82]. A fairly advanced computational study involving Nudged Elastic Band (NEB) to study the binding energetics and kinetics of diphenyl ethers has recently been performed [74].

2.2 Binding affinity and kinetics and their role in drug discovery

As a general goal of obtaining novel compounds with high binding affinities, drug discovery projects often focus on securing a molecular scaffold with favourable binding characteristics followed by optimisation of its binding affinity. The binding affinity, in thermodynamic aspects, is the free energy change (ΔG) between the bound and unbound states of the ligand and the protein. The binding affinity in itself is a result of structural and thermodynamical changes occurring in both ligand and protein as they associate. In other words, the binding affinity is clearly dependent on the dynamics as well as kinetics that constitute the overall binding phenomenon.

The consideration of structural dynamics and kinetics together in overall binding affinity is clearly visible as a potential gap in the results of *in-vitro* and *in-vivo* assays. In real applications, the central focus lies on accurate description of the kinetics and dynamics of the transition state in addition to the reactants and products. For achieving this purpose, the equilibrium constants like K_i and K_d are inadequate thereby stressing the need for additional parameters like dissociation rate constant (k_{off}) to fairly depict the process of ligand binding [49]. However, the developments in NMR and single molecule spectroscopy have enabled the assessment of structural aspects of transition states (and thereby the kinetics) that were formerly very difficult to capture [83–85].

The modulation of drug-target residence time is an important aspect in the pursuit of highly efficacious drugs with favourable binding kinetics. A molecule with a very low value of k_{off} would interact for an extended duration with its target, i.e. it would exhibit increased residence time. This is the case with the diphenyl ethers **PT70** and **PT92**, with very long residence times and identical kinetics as InhA inhibitors. Both molecules exhibit “slow-onset inhibition” which entails an initial rapid reversible binding (EI) followed by a slow tight binding (EI*) indicated by low k_{off} [37]. The latter stage of binding is expected to proceed as described by induced fit, where the k_{off} is actually composed of numerous microscopic rate constants (cf. Figure 2.6) [86].

The typical characteristics of slow tight binders offer a new avenue for improving their efficacy involving destabilisation of the transition state (between EI and EI*) as opposed to stabilisation of the final bound state for the ligand (EI*). This implies a steering of k_{off} (or k_b; Figure 2.6) that results in increased duration of the ligand in the EI* state. However, this attempt of rational residence time modulation faces multiple barriers owing to the lack of detailed structural information pertaining to the transition state [38].

Moreover, rational modification of drug-target residence time has its own mitigating factors and is quite case specific, with long residence times in some cases being desirable, for e.g. bacterial and viral targets [49], but not in some others (for e.g., Roxifiban an anti-thrombotic [87], D2 receptor antagonists as anti-psychotics [88], and Memantine for treating Alzheimer's disease [89]). Another apt example are engineered antibodies where the association constant (k_{on}) demonstrated a better correlation than k_{off} with measured activity [90]. These cases clearly show that microscopic rate constants of association as well as dissociation are of critical importance in a drug discovery endeavour.

2.3 Computational methods for assessment of structure, dynamics and binding affinity of protein-ligand complexes

The recent strides made in the field of theoretical chemistry and computer-aided drug design have aided in unravelling the key events taking place during protein-ligand binding. Numerous approaches to characterise and quantify the interactions in between a ligand and its target during the entire course of the binding process have been of immense use to drug discovery scientists [91–93]. On parallel terms, the evaluation as well as quantification of protein-ligand interactions mainly in terms of either binding affinity or kinetics has been achieved with techniques like molecular docking and MD simulations. Molecular docking on one end serves for rapid prediction of binding orientations for new ligands, while advanced scoring functions enable prediction of binding affinities. With more improved docking algorithms and scoring functions, the large scale virtual screening of compound libraries has been made possible. However, significant barriers still exist for the explicit use of docking and scoring functions for prediction of binding mode and affinity [94].

Molecular dynamics (MD) aim to bridge the gap in between prediction of binding mode, binding affinity, and binding kinetics/residence time. They provide a means to generate

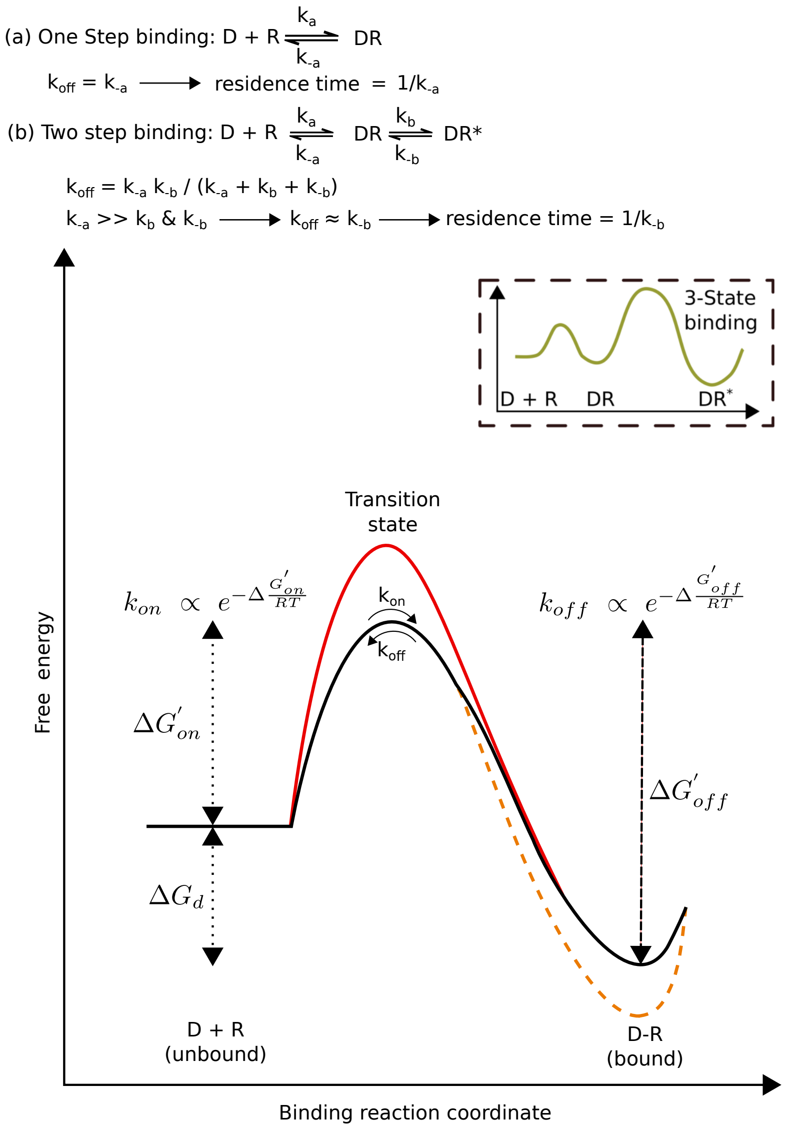
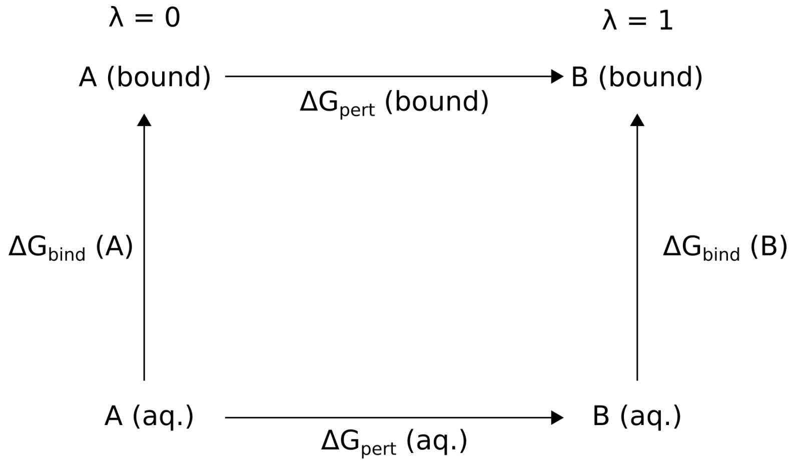

(a) One Step binding: $D + R \underset{k_{-a}}{\overset{k_a}{\rightleftharpoons}} DR$
$k_{off} = k_{-a} \longrightarrow$ residence time $= 1/k_{-a}$
(b) Two step binding: $D + R \underset{k_{-a}}{\overset{k_a}{\rightleftharpoons}} DR \underset{k_{-b}}{\overset{k_b}{\rightleftharpoons}} DR^*$
$k_{off} = k_{-a}\, k_{-b} / (k_{-a} + k_b + k_{-b})$
$k_{-a} >> k_b \,\&\, k_{-b} \longrightarrow k_{off} \approx k_{-b} \longrightarrow$ residence time $= 1/k_{-b}$
3-State binding
D + R
DR
DR*
Transition state
$k_{on} \propto e^{-\Delta\frac{G'_{on}}{RT}}$
k_{on}
k_{off}
$k_{off} \propto e^{-\Delta\frac{G'_{off}}{RT}}$
$\Delta G'_{on}$
$\Delta G'_{off}$
ΔG_d
Free energy
D + R (unbound)
D-R (bound)
Binding reaction coordinate

Figure 2.6 Drug-target binding mechanisms: The upper part describes the kinetics of single step binding followed by kinetics for two step binding. The lower figure describes the free energy profile (black) for one step binding of Drug (D) and Receptor (R), with ΔG_d being the binding affinity that is equal to the difference between bound (DR) and unbound (D + R) states. The modulation of free energy profiles that accompany a decrease in k_{off} are colored red. The solid red line depicts the lowering of k_{off} for a process with destabilised transition state (i.e increasing the barrier height). On the contrary, the dashed line represents a process of increasing the affinity (i.e. stabilising the DR state). The association (k_{on}) and dissociation (k_{off}) rates depend on the free energy differences (Δ G'_{on} and Δ G'_{off}) of the end states and transition state. In the formulas, R is the universal gas constant and T is the temperature. The inset shows a two step binding for a drug-receptor pair; Figure adapted and redrawn from Pan, Borhani, et al., 2013 [37]

an ensemble of structures that represent the structural and energetics changes in the system over time, i.e., a trajectory [95]. It is then possible to perform qualitative as well as quantitative analysis of a trajectory regarding internal motions of a molecule and thereby its function over time. MD simulations also provide for means to assess binding affinity (binding free energy) using numerous methods, each with their own strength and drawbacks. These techniques can be broadly classified into two groups: *rigorous* and *non-rigorous*, which can alternately be called as *free energy pathway* and *end point* methods, respectively. This is because a multitude of binding free energy (affinity) calculations generally pertain to the estimation of relative changes between two equilibrium states, rather than the absolute values. The rigorous methods like Free energy perturbation (FEP) or Thermodynamic integration (TI) are *rigorous* methods that follow Zwanzig's formula (Equation (2.1)) which yields the free energy difference in between two equilibrium states A and B



$$\Delta G_{pert}(bound) - \Delta G_{bind}(B) - \Delta G_{pert}(aq.) + \Delta G_{bind}(A) = 0$$

$$\Delta\Delta G_{A,B} = \Delta G_{bind}(B) - \Delta G_{bind}(A) = \Delta G_{pert}(bound) - \Delta G_{pert}(aq.)$$

Figure 2.7 Thermodynamic cycle describing the relative change in free energy for two ligands A and B bound to the same target. The lower part of the figure describes the calculation of the binding free energy upon protein mutation or ligand modification. The values $\Delta G_{bind}(A)$ and $\Delta G_{bind}(B)$ are readily accessible via free energy perturbation methods, thereby making calculation of the relative binding free energy ($\Delta\Delta G_{A\rightarrow B}$) possible. Figure adapted and redrawn from Sharp, 2012 [96].

$$\Delta G = G_B - G_A = -\beta^{-1} \cdot ln \left\langle exp(-\beta \cdot \Delta V) \right\rangle_A \tag{2.1}$$

where, $\beta = 1/kT$, k is the Boltzmann constant, T the absolute temperature, $\langle\ \rangle_A$ represent a MD or Monte Carlo (MC) ensemble average of $\Delta V = V_B - V_A$, i.e., the change in potential from ligand B to A [97]. The Equation (2.1) yields Gibbs free energy when the configuration sampling is performed in an isothermal-isobaric ensemble (N, P, T conditions), while sampling under N, V, T conditions yields the corresponding Helmholtz free energy.

In actual practice, the transition in between two states (A and B) is split over multiple independent perturbations (m = 1, 2,..., n), each run simultaneously with its own potential (V_m) (Equation (2.2)). The linear combinations of the intermediate potentials ($V_m^{A\rightarrow B}$) gives the potential of either A or B [97, 98]. The free energy change for the process can then be obtained by summation over the change in potential ($\Delta\lambda$) going from A to B (Equation (2.3)).

$$V_m = (1 - \lambda_m)V_A + \lambda_m V_B \tag{2.2}$$

where λ_m varies from 0 to 1, while m stands for the state and V_m for the potential of that state.

$$\Delta G = G_B - G_A = -\beta^{-1} \sum_{m=1}^{n-1} ln \left\langle exp\left[-\beta(V_{m+1} - V_m)\right]\right\rangle_m \tag{2.3}$$

A mathematical alteration of Equation (2.3), splits the path from A to B into very small steps, each with its own λ value followed by integrating them to get the overall free energy change from A to B (Equation (2.4)). This approach is called **Thermodynamic integration (TI)** and is the preferred method in practice. In general use, the free energy change upon ligand binding is investigated in relative sense, as seen from Figure 2.7. However there are certain limitations in regards of the rigorous methods, namely:

1. Convergence of the sampling for the various configurations
2. Time consuming and thereby limited to small perturbations at a time [97].

Additionally, the rigorous methods are unable to offer insights into binding kinetics and residence times since the thermodynamic cycle does not consider transition states of binding/unbinding.

$$\Delta G = \int_0^1 \langle \Delta V \rangle_\lambda d\lambda \tag{2.4}$$

On the other hand, the non-rigorous methods have evolved as popular alternatives to their more rigorous counterparts. These methods (non-rigorous) typically consider the physically relevant portions of the thermodynamic cycle (Figure 2.7), i.e., the bound and unbound states of the ligand. These typically reside at the endpoints of the thermodynamic cycle and hence the alternate name "endpoint methods". The process of binding affinity calculation begins with generation of a thermodynamic ensemble followed by evaluation of the non-bonded interaction energies of each endpoint. Two of the most popular endpoint methods are Linear Interaction energy (LIE) [99–101] and Molecular Mechanics/Generalised Born/Poisson-Boltzmann/Surface Area (MM-/GBSA/PBSA) [102–104]. Both methods are derivatives of the Linear-response approximation (LRA) [105, 106], which estimates changes in electrostatic free energy involved in protein (P)- ligand (L) binding as shown in Equation (2.5).

$$\Delta G = \frac{1}{2}\left(\left\langle E_{ele}^{L-S}\right\rangle_{PL} - \left\langle E_{ele}^{L-S}\right\rangle_{PL'} + \left\langle E_{ele}^{L-S}\right\rangle_{L} - \left\langle E_{ele}^{L-S}\right\rangle_{L'}\right) \tag{2.5}$$

where, E_{ele}^{L-S} is the electrostatic interaction energy between the ligand and its periphery (protein or water), $\langle\ \rangle$ indicate ensemble average, PL and L denote standard MD simulations of bound and free ligand, respectively, while PL' and L' indicate simulations with ligand having null charge.

The LIE method along with its theory and applications in binding affinity prediction for InhA inhibitors will be discussed in Chapter 4. The MM-GBSA/PBSA method [102, 104] utilises a continuum solvent approach for analysing the trajectories, with the "mean" free energy change being calculated by Equation (2.6).

$$\begin{aligned} \Delta G =&\ \langle G_{Complex}\rangle - \langle G_{Protein}\rangle - \langle G_{Ligand}\rangle \\ G =&\ E_{bnd} + E_{el} + E_{vdW} + G_{pol} + G_{np} - TS \end{aligned} \tag{2.6}$$

where the first three terms of Equation (2.6) represent standard molecular mechanics energy terms (bonded and non-bonded), G_{pol} and G_{np} are the polar and non polar contributions to solvation free energies. G_{pol} can be obtained by solving the GB/PB equation, while G_{np} is derived from a linear approximation of solvent accessible surface (SASA). The final term T represents absolute temperature multiplied by the entropy (S), and is calculated by a harmonic analysis of vibrational frequencies [104].

The ensemble averages from Equation (2.6) for the individual components can be calculated using two approaches. The first approach involves separate ensemble generation for protein, ligand and complex, respectively (referred to as 3MM-PBSA). The second approach (1MM-PBSA) makes use of a singular ensemble for the complex from which the corresponding individual ensembles can be obtained by removing the protein or ligand, respectively. In practice, the latter is preferred not only for its speed and ease of implementation but also because it is less error prone [104, 107, 108]. A common drawback of both approaches is the partial capture of entropy that contributes to the total free energy of binding. This is simply because capturing entropies by normal mode analysis or related harmonic analysis of vibrational frequencies is a difficult task [97]. The MM-PB/GBSA also offers no insights into binding kinetics or its determinants.

The extensive analysis of several related systems enables the extraction of structural information pertaining to determinants of ligand binding and residence time. The dimensionality reduction techniques like principal component analysis and related methods additionally aid in revealing the extent of correlated/uncorrelated movements in a given system. The information derived from these techniques can be combined with the structural features of slow tight binders to drive the structure-based optimisation of moderately potent compounds like pyrrolidine carboxamides. This approach will be clearly outlined and discussed further in part II of this thesis.

2.3.1 Basics of Molecular Docking

The past few decades have witnessed a spurt in high-speed synthesis as well as high throughput screening which ultimately led to a revolution in lead discovery against a multitude of molecular targets. This in part has also been due to a parallel increase in number of crystal structures being resolved that cover 125,093 molecular targets and 22,104 ligands [109]. With the passage of time, the targeted HTS methods became increasingly preferred over random library design and HTS screening [110]. Molecular docking and scoring were developed as high throughput alternatives to the conventional biological screening methods. Consequently in the past decade, molecular docking has secured an important place in virtual screening protocols for *in-silico* lead optimisation. The binding energy evaluation for poses from docking using first principle methods is cumbersome and has given way to comparatively faster scoring functions [110]. The scoring functions typically aim for rapid prediction of ligand binding energetics and ranking the same based on the calculated affinities. All of this being performed with a particular emphasis on adequate consideration of the molecular phase space that is critical for correct identification of the protein bound ligand conformation [110, 111].

The process of molecular docking can then be defined as the prediction of structure of receptor-ligand complexes, with the receptor being a protein, DNA or RNA, while the ligand can be a small molecule, oligomer of peptide, DNA or RNA [112]. Pioneering work in docking was performed by Kuntz et al. [113], while several programs with variable approximations have been developed since the original method was implemented [94]. In any lead optimisation or virtual screening endeavour, a docking protocol being used should be able to:

- Predict the binding orientation for a given compound/s while considering full flexibility of protein and ligand together with adequate sampling of their conformational space.
- Accurately predict the binding affinity and rank the compounds based on their ”scores” (predicted binding affinity)

There are two main types of molecular docking depending upon the consideration of partial or full flexibility of either the ligand and the receptor, namely rigid docking and flexible docking [114, 115]. Rigid docking treats both ligand and receptor as frozen while considering geometric complementarity and ignoring binding phenomena like induced fit. Flexible docking usually treats the ligand as fully flexible, while keeping the protein geometry frozen. More recently, flexibility of the protein has also been considered to a limited extent with the process often referred to as induced-fit docking [116, 117].

2.3.1.1 Ligand Placement Algorithms and Conformational Sampling

Addressing the ligand flexibility remains an important endeavour both prior and during the docking process. This is achieved by three main methods: generation of multiconformer libraries, incremental construction in the binding pocket, and stochastic methods, respectively. The multiconformer method uses pre-generated conformers for the rigid docking as is the case with EUDOC [118] and FRED [119], and DOCK [113]. Now-a-days, conformers can be generated on-the-fly using programs like CATALYST (Accelrys Software Inc., San Diego, USA), and OMEGA (Open Eye scientific, Santa Fe, New Mexico, USA). However, a conformer generation software has to ensure that its output retains the bioactive conformer whilst simultaneously giving a reasonable number of conformers that the docking algorithm can handle, mainly to avoid a combinatorial explosion [120].

An apt alternative for the multiconformer generation is the incremental construction algorithm. According to this algorithm, the molecule to be docked is dissected into **"anchor"** and **"fragments"**, usually along the single bonds that are not from a cyclic system. Subsequently, a suitable placement for the anchor fragment is ascertained which is followed by incremental and a step-wise fragment addition that typically yields the receptor bound conformation of the ligand. The placement of the fragments is usually guided by a combination of matching algorithms (superposition of atomic triplets or paired interaction centers) and rules describing their torsional inclinations [120]. This approach is used by FlexX [121] and DOCK 4.0 that employs the anchor-and-grow method [122].

The stochastic methods follow a different path and sample ligand conformations on-the-fly, with two main approaches for the same: Genetic algorithms (GA) and Monte Carlo (MC). The Monte Carlo methods begin with an arbitrary ligand conformation in the phase space followed by random generations of new configurations. The configurations are typically generated by varying the torsional angle and rotational/translational degrees of freedom for the ligand. The selection of favourable configuration is governed by a Metropolis criterion that ensures that low energy configurations are always sampled whilst those of higher energies are only accepted probabilistically (comparison of random number between 0 and 1 with the Boltzmann energy difference between the current configuration and the previously accepted one) [120]. The Monte Carlo technique (coupled with simulated annealing) has been used in ICM [123] and MCDOCK [124].

Another prominent class of stochastic methods are the genetic algorithms (GA's) which are inspired by Darwin's theory of evolution [94]. In docking context, an initial population of poses are generated randomly, with the pose characteristics (dihedral angle and vectors describing global rotation, translation) being coded by chromosomes. The population of chromosomes is then allowed to evolve by performing genetic operations (mutations and crossover of genes) that yield newer generations of the initial poses that exhibit higher fitness (e.g., better docking score) in comparison to their parents. The process continues until a termination criterion is reached (e.g., minimal RMSD cutoff, maximum number of generations, constant optimal fitness value etc.) [120, 125]. The GA based docking programs include AutoDock [126] and GOLD [127].

In addition to the aforementioned techniques, energy-search driven techniques like molecular dynamics enable configurational sampling of the ligand and protein as well with full flexibility. Their applicability has been mostly restricted to pre-processing (ensemble generation for ensemble docking) and post-processing (post-docking energy minimisation). The main barriers to use of MD in docking is its inefficiency in traversing low energy barriers and sampling of multiple low energy minimas [120]. Nevertheless, there are promising techniques like metadynamics [128] that enable the use of MD simulations in docking.

2.3.1.2 Scoring Functions and Binding Affinity Prediction

The need of docking to rapidly and accurately score as well as rank a large number of docking solutions gave rise to scoring functions. Scoring functions are ideally needed to not only correctly predict the binding energy and rank the docking poses, but also differentiate binders from non-binders in a virtual screening endeavour. They are grouped into three main types:

1. **Empirical scoring functions**
   Empirical scoring functions involve regression analysis between the structural descriptors and the experimentally determined affinity to derive an equation for prediction of binding affinity [125, 129]. The term "empirical" comes from the reliance of this group of functions on the experimental data of structures and affinities [120]. The descriptors of the protein-ligand complexes are those that contribute maximally to binding affinity, e.g., hydrogen bonds, buried hydrophobic surface area, number of rotatable bonds. Starting with a large number of protein-ligand complexes (training set), descriptors are calculated followed by correlating them to the experimental affinity by conventional multilinear regression or machine learning methods like random forest [130] and support vector machines [131]. Most of the empirical scoring functions development began with introduction of Böhms function [132], which forms the basis of LUDI score [133] as well as in an adapted form for FlexX [125]. Other noteworthy examples of empirical scoring functions include Glidescore [134], Chemscore [135], and SFCscore [136]. The main drawback of empirical scoring functions is the need to derive the weights for coefficients for each of the weighted terms making their applicability limited to certain cases [137].

This also means that the dependence of empirical scoring functions on limited experimental data creates an ambiguity about the meaning of each term and its error assessment [125]. Furthermore, the applicability of an empirical scoring function as well as its success in the same remains confined by the quality as well as quantity of the data in the training set. Another drawback of empirical functions is the obstacle in quantification of entropy and desolvation terms, both of which contribute significantly to the final binding affinity [125].

2. **Force field/molecular mechanics-based scoring functions**
These scoring functions utilise classical terms from molecular mechanics to estimate the binding free energy for a pose under consideration. These are loosely based on parameters defined in the force fields, which are derived from experimental data (excluding binding affinity) and *ab-initio* Quantum Mechanical (QM) calculations. These scoring functions predict the binding free energy as a sum of van der Waals and electrostatic interactions, with intermittent inclusion of intramolecular strain energy [129]. Some improved force field scoring functions include effects of solvation-desolvation in the final value by considering a continuum solvent approach (MM-PBSA, MM-GBSA) [102, 103] or by considering the bound and unbound states of the ligand (LIE) [101, 138] and related extensions of the same method (LIECE) [139]. Though both LIE and MM-GB/PBSA have been found to be extremely useful in predicting binding affinity for a wide series of protein-ligand complexes, they also come with a fair share of drawbacks. The MM-GB/PBSA technique is particularly sensitive to alterations in protonation states of protein/ligand, change in substitution size on the main scaffold or any such changes that lead to a deviation from the reference scaffold. The LIE method also, by its very definition generates the need for fitted parameters for a particular scaffold making any prediction limited to related molecules or with minor deviations from the reference scaffold. A common drawback of these scoring functions is the inadequate consideration of the entropic contribution to the binding affinity.

3. **Knowledge-based scoring functions**
As the name suggests, this category of scoring functions perform statistical analysis of structural information from protein-ligand complexes from large databases into ”pseudo free energies” for protein-ligand atom pairs [125]. This approach is also referred to as potential of mean force (PMF), where the interaction energy or distance dependent pseudopotentials for a given atom pair can be shown by Equation (2.7), with the term g_{ij} being calculated by estimating the density of occurrence of the atom pair *ij* at distance r in a database, e.g., RCSB PDB.

$$A(r) = -k_b T \cdot ln \cdot g_{ij}(r) \tag{2.7}$$

where i is the protein atom, j is ligand atom, A(r) the protein-ligand interaction free energy, k_b is the Boltzmann constant and T represents the temperature [125].

Finally, the score is a summation of all such interatomic interaction energies in the protein-ligand complex. The most important aspect of this group of scoring functions is their strength in predicting the pose quality that stems from the use of very extensive structural data. This also sidesteps the need for fitting of the experimentally determined binding affinities for the protein-ligand complexes of the training set (like empirical scoring functions). An added advantage of these scoring functions is the implicit treatment of solvation and entropic terms. Prominent examples of these scoring functions include PMFscore [140] and DrugScore [141] alongwith its numerous variants like DrugscoreX [142], Drugscore^{CSD} [143], and Drugscore^{PPI} [144].

2.3.1.3 Consensus Scoring and Tailored Scoring Functions

It is a well known fact that scoring functions represent an Achilles heel for any docking endeavour, given the fact that each scoring function has its own strengths and weaknesses [145]. As a consequence, with the main aims of subverting the main shortcomings of scoring functions, some of the many approaches described in literature will be touched upon. These techniques are as follows:

1. **Appropriate consideration of hydrophobicity and water exclusion**
   This approach has been implemented in the HYDE scoring function by Rarey et al. [146], wherein the binding affinity was derived from two terms representing dehydration of polar and nonpolar groups in the protein-ligand interface alongwith modified treatment of hydrogen bonding. The terms were in turn obtained by atomic level increments in logP that yielded the corresponding hydrogen bonding and dehydration free energies. The inclusion of atomic accessibility ensures that the polar atoms undergoing dehydration in narrow channels or pockets are penalised less than solvent exposed polar atoms, which is consistent with experimental observations [129, 146]. The resulting scoring function is target independent and able to classify ligands at both ends of activity spectrum, with a generalised cutoff for recognising binders. Thus, the inclusion of the above terms led to an improvement in discriminating ability for recognising binders and non-binders in a target independent fashion.

2. **Utilisation of high quality training and test datasets: SFCscore derivation**
   This scoring function developed by Sotriffer et al. [130, 136] under the aegis of the Scoring Function Consortium (SFC) represents another case implying the effect of

a high quality training dataset on the performance of the final scoring function. With a training dataset of over 850 high quality protein-ligand complexes and 60 descriptors representing numerous protein-ligand interaction characteristics, a superior performance was observed for all 8 derived functions on poses generated for 22 different target proteins [136]. However, noticeable improvement over other empirical scoring functions was not observed. The underlying reasons were attributed to lack of consideration for the solvent, flexibility of protein, and to the empirical approach itself. Whilst majority of these issues could be addressed by use of high quality protein-ligand complexes, given the heterogeneous nature of the complexes in the training set, an average prediction error of ≤ 1 pK$_i$ unit is difficult to achieve [145]. This fact merely underscores the prediction limits for empirical scoring functions.

3. **Consensus Scoring**
The use of two different scoring functions should augment their overall performance, which was indeed found to be true in some cases [147, 148]. This concept of using multiple scoring functions in scoring and ranking poses is called "Consensus Scoring" [149]. The use of two or more distinct and high performing scoring functions derived from entirely different methods led to improvements in positive hit rates alongwith a slight decrease of true positives [150]. This necessitates the need for comparison of individual scoring functions versus the combination to assess the performance of the functions across individual targets. Some examples of consensus scoring included GFscore [151], which utilises neural networks and a combination of 5 different scoring functions, and SeleX-CS [152]. In **Chapter 4**, a step-wise consensus scoring based upon GlideScore, Drugscore and SFCscore is used to obtain near native ligand conformations followed by prediction of their binding affinity.

4. **Tailored Scoring functions**
Rather than developing generic scoring functions that perform variably across different targets, functions targeted against a particular type of targets are an attractive undertaking. Such functions are augmented and customised versions of their parent functions using additional terms or filters [145]. One prominent example is **AfMOC** derived by Gohlke and Klebe et al. [153], that is a group of functions derived from DrugScore. TScore [154] uses a CoMFA [155, 156] (Comparison of Molecular Field Analysis) like approach, with molecular fields derived from the target and not just those from ligand delivering superior performance as compared to knowledge based potentials alone, highlighting the success of this customised approach.

2.3.2 Protein Flexibility in Molecular Docking

Chapter 1 describes the fundamental theory behind protein-ligand binding, with induced fit and conformational selection being the major mechanism in binding [36, 44]. Thus, to understand the handling of protein flexibility during docking, it is imperative to consider first the types and zones of flexibilities in a protein and their relation to ligand binding. Generally, loops have maximal mobility with polar solvent exposed residues and residues bordering the loops demonstrating higher movements [157]. It was seen that important residues often undergo conformational changes and subsequent stabilisation upon ligand binding, with a restricted number of side chains actually being significantly affected by the changes [158]. Furthermore, structural measurements classify flexible proteins into 3 main groups: nearly rigid proteins, flexible proteins, and intrinsically unstable proteins based upon the conformational changes upon ligand binding [159].

The handling of protein flexibility during docking is not a one step process but rather a workflow of several processes [129]. Based upon the basic ligand binding mechanism, two main types of processes exist, viz., conformational selection and induced-fit. The computational representation of conformational selection involves ensemble generation prior to docking as a means of introducing protein flexibility. The use of ensemble generating procedures ensures a proper conformer generation of proteins and ligand placement via a complementary fit. The alternate path of induced fit is usually accompanied with noticeable conformational changes in a single protein structure. The Figure 2.8 depicts handling of protein flexibility during numerous docking stages. The process of docking using conformational selection considers the protein as an ensemble of variably populated conformational clusters at equilibrium, with ensemble generation being achieved by experimental techniques like NMR or X-ray crystallography [160]. However given the amount of effort and experimental obstacles for either NMR or X-ray crystallography, MD simulations offer a pragmatic option for ensemble generation. The use of MD simulations in ensemble generation is strongly dependent on the parameters used which are often variable [157]. An example of this approach is *relaxed complex schemes* that yields MD generated snapshots in the initial stage of docking [161, 162].

Alternatively, Monte Carlo simulation based conformational sampling can also be used to describe the equilibrium state of the protein in the form of an ensemble of structures. The ensemble generation can be sped up by using multi-step coarse-grained (CG) simulations or by performing simulations in implicit solvent, at the cost of accuracy [163, 164]. Other approaches for ensemble generation include restricted consideration of relevant protein conformations[165, 166], sampling of rotamers and sidechain of key residues of binding site [167–169], use of Normal mode analysis [170, 171], and targeted productions of ensembles with *a priori* information of binders.

As mentioned earlier, protein flexibility in docking can be considered by two groups of methods: Ensemble docking and Docking with induced fit, of which induced fit shall be covered briefly, while ensemble docking methods will be discussed in a minimalistic way. A summary of various methods amongst both groups is shown in Figure 2.9.

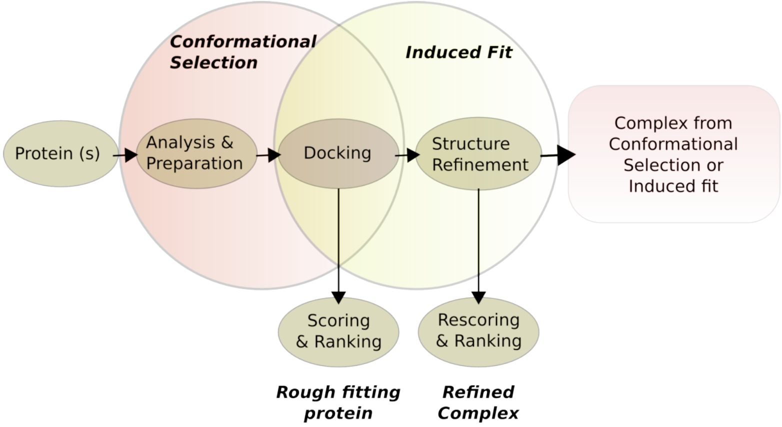


Figure 2.8 Protein flexibility can be considered at numerous stages of docking, namely before, during and in post ligand placement phase. The identification of flexible parts in a protein take place in predocking stage followed by conformational sampling. The docking then samples the ligands conformational space and protein as well. Scoring and pose refinement take place after the docking. Figure adapted and redrawn from Henzler, Rarey et al., 2011 [157].

2.3.2.1 Ensemble Docking

This technique, as the name suggests, considers an ensemble of structures with varying conformations of side chains, loops and backbone arrangements for docking, with a proper selection of the binding pocket prior to ligand fitting. The principal approach would be to dock the ligand in every conformation followed by energetics assessment of a single average structure or of the complex containing the top ranked pose. This approach of sequential docking is clearly cumbersome and time consuming. However, in recent years, several docking protocols that evaluate the structural ensemble in a single run have been developed. These fall into three main categories: use of ensemble generated 3D grids, use of a unified target structure obtained from the ensemble, and improved search strategies considering a selected portion of the conformational ensemble rather than the whole of it.

The 3D grid method utilises precalculated 3D grids of interaction energies for ligand atoms at each grid point enabling fast evaluation of ligand poses [172]. The improved search method relies on a heuristic analysis to find best fits of the ligand to the conformational ensemble rather than dock the ligand across the entire ensemble, e.g. SIMPLEX local optimisation based [157] or MC based docking [173], while the united protein structure method retains the core interacting region, varying the immediate vicinity

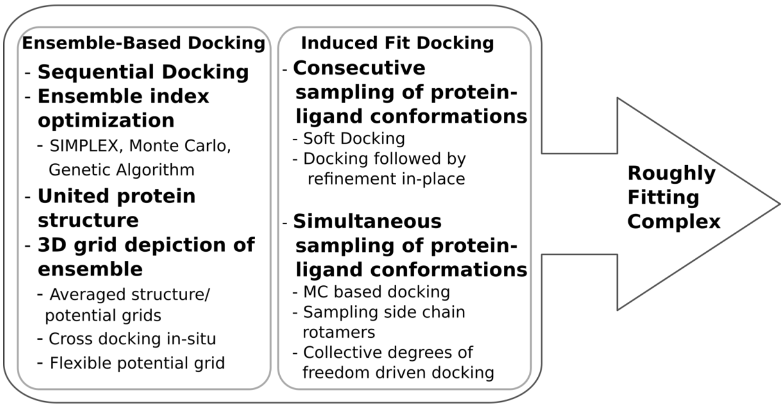

Ensemble-Based Docking
- Sequential Docking
- Ensemble index optimization
- SIMPLEX, Monte Carlo, Genetic Algorithm
- United protein structure
- 3D grid depiction of ensemble
- Averaged structure/ potential grids
- Cross docking in-situ
- Flexible potential grid
Induced Fit Docking
- Consecutive sampling of protein-ligand conformations
- Soft Docking
- Docking followed by refinement in-place
- Simultaneous sampling of protein-ligand conformations
- MC based docking
- Sampling side chain rotamers
- Collective degrees of freedom driven docking
Roughly Fitting Complex

Figure 2.9 Methods utilised for fully flexible docking, Figure adapted and redrawn from Rognan, 2011 [125].

of the surrounding regions, thereby avoiding redundant scoring coupled with ensemble evaluation on the fly. This approach has been implemented in FlexE [174].

2.3.2.2 Induced Fit Docking

Contrary to the ensemble methods, induced fit based methods utilise a single input structure for docking and energetics assessment. The conformational changes are brought about after the ligand placement phase, with the ligand itself experiencing a simultaneous conformational change [129]. This approach of consecutive conformational evaluations avoids an expansion of the underlying phase space. In practice, a quasi-simultaneous conformational analysis of ligand and protein is carried out in a restricted phase space to avoid conformational explosion. It also prevents the need for subsequent energy evaluations for the protein as well as the ligand. Induced fit docking consists of two approaches, the first one dealing with consecutive conformational variations-evaluations of protein and ligand, and the latter performing a simultaneous conformational variation and evaluation.

1. **Consecutive conformational change in ligand and protein**

   During an induced fit docking, numerous changes occur in the binding pocket that initially accommodate the incoming ligand followed by subsequent optimisation of the ligand surroundings. There may be initial steric hindrances to ligand binding, that any induced fit docking protocol has to consider, possibly by ignoring them followed by refinement of the complex structure. There are further two approaches [129] to perform this, namely:

   - **Soft docking followed by external refinement of the complex structure**

     In order to ignore the initial steric hindrance to ligand association in the event of inadequate space for its accommodation in the binding pocket, soft docking methods were proposed. Soft docking allows partial penetration of the proteins surface by the ligand during the initial placement [175]. It makes use of grids and representation of protein and ligand shapes as cubes followed by a matching step in pose evaluation, wherein cube matching is rewarded and cube overlap is penalised [129]. Other alternatives of modelling receptor surface penetration and reduction of repulsion include substitution of Lennard-Jones 6-12 potentials with 9-6 potentials [176]. The main drawback of this method is lack of accuracy and hence soft docking usually precedes complex refinement in initial stages of an induced fit workflow.

- **Consecutive ligand placement and complex optimisation**
  An alternate pathway followed by some commercial programs is to perform an initial placement of the ligand in the binding pocket followed by optimisation of binding site residue side chains (e.g., SLIDE [177]). Another approach by Sherman et al. [116, 117] first identified flexible residues in the binding site, followed by their replacement with alanine and subsequent soft docking with the modified structure. Upon accomplishment of the soft docking, the alanines are resubstituted with the original residues and prediction of their low energy conformations. The pose evaluations then take place by redocking the ligand in this state with the usual potential that rescores the complex. This approach forms the basis for **Docking with induced fit in Glide** and is discussed in detail in Chapter 3.

2. **Simultaneous change in ligand and protein conformations**
   These methods aim to avoid postprocessing of docking solutions with a quasi-simultaneous change in protein and ligand conformations. The prolonged time required for achieving a simultaneous sampling of protein and ligand conformations render this method impractical or confined to a restricted extent with numerous approximations [178]. In the wake of such barriers, the use of heuristic searches together with a reduction in the degrees of freedom are beneficial in enhancing the fully flexible docking methods. Some of the approaches within this group of methods include use of MC driven docking that consider side-chain flexibility (ICM) [123], stochastic tunnelling (FlexScreen [179]), use of binding site residues side chain rotamers [180, 181], and performing docking with selective degrees of freedom [182, 183].

2.3.2.3 Flexible Docking in Virtual Screening

Generally speed is an important factor in a virtual screening endeavour and whilst fully flexible docking can be used in VS, there are many practical considerations. These mainly pertain to the nature and extent of flexibility of the target followed by the time required for complex optimisation. The latter can be minimised by restricted minimisation of the local region around the ligand while keeping most of the target fixed. Even after this, fully flexible docking remains limited to small datasets with approximations that accompany a loss in accuracy [145]. Consequently, a mixture of aforesaid techniques, viz., ensemble and induced fit may provide a means to achieve best of both, provided there are good input structures to begin with. Ensemble docking can provide a case specific restricted or global ensemble while the small fluctuations that separate the members of the ensemble can be considered by induced-fit docking [176]. Other approaches to speed up the process include use of target specific libraries with some *a priori* binding knowledge.

2.3.3 Basics of Molecular Dynamics

The previous sections clearly talk of bimolecular flexibility and computational approaches to consider the same in the drug design process. And though the biomolecular structures provided by X-ray crystallography and NMR are tremendously useful, the structures being depicted as static are actually quite dynamic [184]. The dynamics of biomolecules is quite critical for their function which can be studied by a variety of experimental techniques, all of which have their own shortcomings. An attractive approach to study the dynamics of the biomolecules is to study the evolution of positions and velocities of every atom according to first principle physics (e.g., Newtons 2^{nd} law of motion). The time dependent evolution can be studied computationally via an all-atom simulation pioneered by Alder and McCammon [185, 186]. From a biological and drug design perspective, simulating the process of target-ligand binding provides a very useful avenue to study intracellular signalling (receptors) and to understand the mechanism of action of drugs. In drug design endeavours, it is quite critical to have adequate representation of binding phenomena in order to design and optimise lead compounds, especially so when considering binding kinetics. The conformational changes accompanying the protein-ligand association exhibit a variable time scale that ranges from bond vibration over femtoseconds to protein domain movements that takes place over several microseconds (fast) to seconds (slow) (cf. Figure 2.10).

2.3.3.1 Modelling atomic motions - Molecular Mechanics Force Fields

A key requirement to model and capture atomic movements is the ability to capture and adequately represent the same (quantification). According to the **atomic force field model**, a biomolecule or any physically relevant system can be considered to be

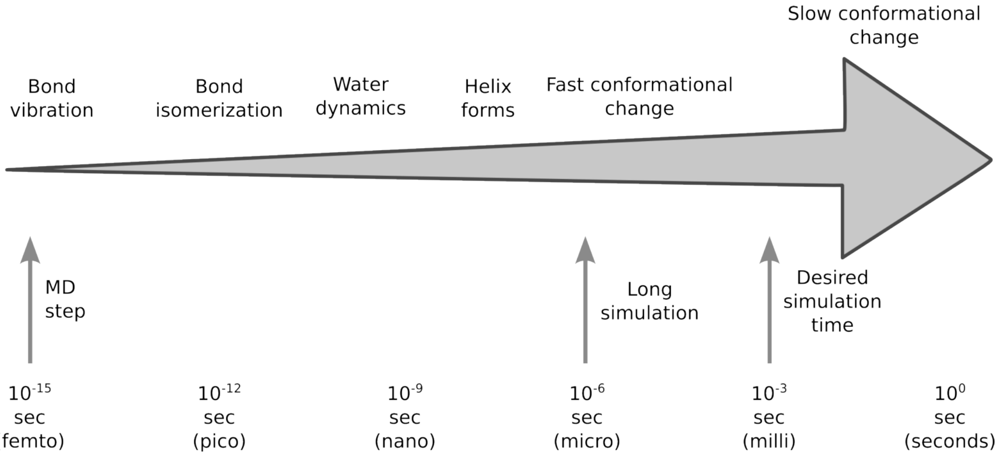


Figure 2.10 Time scale for important events pertaining to proteins with respect to step size of a MD simulation. Figure adapted and redrawn from Ode, Nakashima, et al., 2012 [187].

a mere collection of atoms held together by interatomic forces. This is an approximate description of a physically perceptible system, while the more accurate quantum mechanical model views the system as a group of interacting electrons and nuclei. A good adiabatic approximation (Born-Oppenheimer) of the mass difference in between the nuclei and electrons enables a more accurate quantum description of the system by separation of electronic and nuclear parts [95]. In a highly complex and dynamic system like biomolecules, the application of a quantum mechanical model obviously becomes impractical. A solution for the same comes in form of molecular mechanics force fields (cf. Figure 2.11) which approximate the biological system as a collection of interconnected "balls and springs", with each atom being represented as a ball of fixed radius and charge, while the atomic bond is represented as a spring. The interaction potential of N interacting spheres (atoms) as a function of their positions (a_i=x_i, y_i,z_i) is given by U(a_i, ..., a_N). The force acting on atom i can now be shown by Equations (2.8) and (2.9).

$$F_i = -\nabla_{a_i} U(a_i ... a_N) \tag{2.8}$$

$$F_i = -\left(\frac{\partial U}{\partial x_i}, \frac{\partial U}{\partial y_i}, \frac{\partial U}{\partial z_i}\right) \tag{2.9}$$

A typical force field potential (Figure 2.11) usually consists of bonded terms and non-bonded terms, with the bonded terms describing the energy of deformation for bond lengths, bond angles and torsions. The non-bonded terms represent the van der Waals repulsive and attractive forces in form of a Lennard-Jones 6-12 potential, while electrostatics are covered by the Coulomb potential [188, 189]. It is evident that the atomic interaction potential is a function of the coordinates of each atom, with the intra- and intermolecular potentials together constituting the total potential energy of the system (E_{total}). A key consideration pertaining to force fields is that they are approximations and empirical in nature. Since commonly used force fields describe an atom as a sphere of

$$V(r^N) = \sum_{bonds} k_b(l-l_0)^2 + \sum_{angles} k_a(\theta-\theta_0)^2 + \sum_{torsions}\sum_{n} \frac{V_n}{2}[1+cos(n\omega-\gamma)] + \sum_{j=1}^{N-1}\sum_{i=j+1}^{N} f_{ij}\left\{\epsilon_{ij}\left[\left(\frac{r_{0ij}}{r_{ij}}\right)^{12} -2\left(\frac{r_{0ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right\}$$

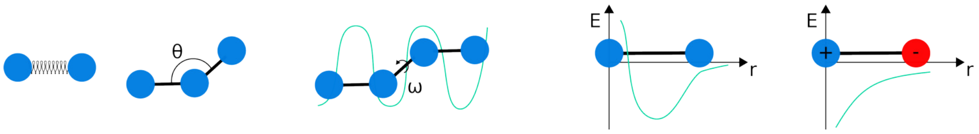


Figure 2.11 Components of a force field according to Durrant, McCammon, 2011 [189]: The potential consists of functions describing equilibrium bond lengths, angles, torsions (bonded interactions), and those describing non-bonded interactions (Lennard-Jones potential and Coulomb potential). The $V(r^N)$ denotes energy, k_a, k_b represent force constants, V_n represents the barrier to rotation around a bond, while θ and ω represent bond and dihedral angles, respectively. r denotes the equilibrium distance in between two atoms i and j, while γ symbolises dihedral phase, n the multiplicity of the dihedrals, while $r_{0_{ij}}$ encloses the Lennard-Jones potential at which the potential is zero. The term q denotes the charge of i and j, while ϵ signifies the dielectric constant.

unit mass and fixed point charge, it becomes imperative that the atomic parameters being used should resemble experimentally determined values to get a more realistic modelling of any event being mimicked [189]. Furthermore, the different levels of approximation are needed for different scenarios to be modelled, for example atomic parameters describing an amide bond of the protein backbone cannot be used to describe a C-N bond in nucleosides or sugars. Another thing to be considered in case of molecular force fields, is that choice of a particular force field is driven by factors like property of the system to be simulated, desired accuracy levels, and duration of the simulation [190]. Accordingly, the literature contains many references to force fields that are classified on the basis of factors described earlier [190].

2.3.3.2 Parameterisation of Force Fields

As mentioned before, the force fields are purely empirical in nature with a certain level of approximations involved. This very nature of force fields requires the description of atomic parameters to a high level of accuracy. Quite often, the parameters in force fields are derived either from *ab initio* calculations or by fitting to experimental data. Of the many force fields used conventionally, a few are noteworthy since they are used by academia and industry globally, e.g., CHARMM [191, 192] (**C**hemistry **A**t **H**arvard **M**olecular **M**echanics), AMBER [193] (**A**ssisted **M**odel **B**uilding with **E**nergy **R**efinement), GROMOS [194] (**GRO**ningen **MO**lecular **S**imulation), and OPLS [195] (**O**ptimised **P**otentials for **L**iquid **S**imulations). It must be noted that the aforesaid names stand for a complete family of force fields, with each force field designed for specific purposes.

Of the aforesaid force fields, the AMBER force field and associated force fields present an attractive way to perform all atom simulation with explicit treatment of solvent. In this force field, the electrostatic potentials obtained from quantum mechanical calculations at the Hartree-Fock (HF) 6-31G* level are used to assign atomic charges using the restrained electrostatic potential (RESP) method [193, 196]. The bond and angle parameters were derived by fitting parameters for small molecular fragments of proteins and nucleic acids with their structural and vibrational frequency data. The dihedral parameters were derived by obtaining dihedrals for a representative small molecule dataset followed by calibration for bigger molecules [193]. Finally, the van der Waals parameters for known atoms were ascertained by performing Monte Carlo simulations for simple molecules (e.g., methane, benzene) followed by fitting to their atomic densities and vaporisation enthalpies.

The most widely used Amber force field for condensed phase all-atom simulations is the Amber *ff99SB* which was developed to improve the descriptions of secondary structures as well glycine residues as compared to previous force fields [197]. A key feature that is lacking in this force field (and majority of force fields) is the consideration of polarisability of molecular orbitals, making simulations in variably charged environments difficult to assess [95]. Furthermore, the process of fitting charges to the potentials at HF-6-31G* level led to overestimated values of bond-dipoles as compared to their gas state, i.e., in other words, "overpolarisation". Overpolarisation although beneficial for simulations in explicit solvent, suffers some drawbacks in the sense that accurate modelling of a molecular response to variations in the dielectricity of the surrounding medium becomes difficult, particularly in case of folding of proteins and binding affinity estimation [198]. The problem has been solved by use of polarisable force fields, although they are computationally intensive to use apart from being difficult to parameterise [199].

In addition to developing force fields for macromolecules, the need for developing automated parameterisation of small molecule was also felt [197]. As a result, Wang et al. developed the General Atom Force Field (GAFF) that describes the parameters for common atoms like H, C, N, O, S, P, and halogens as well as ions [200]. Here, just like AMBER *ff99SB*, the atomic charges are ascertained by RESP fitting [196] of electrostatic potentials derived from quantum mechanical calculations at HF/6-31G* level of theory.

2.3.3.3 Integrating Equations of Motion

Molecular dynamics is strictly a statistical mechanics method that has its roots in first principle physics [190]. In essence, it follows Newton's second law of motion; $F_x = m_x \cdot a_x$, where F_x the force acting on a particle x, is the product of its mass m_x and the acceleration a_x [188]. In context of MD simulations, the time dependent evolution for a set of interacting atoms can be derived by numerical integration of the basic equation mentioned earlier. With respect to time t and a new position $r_x(t_1)$ for the next time step $t_1 = t_0 + \Delta t$, the force F_x can be calculated as depicted in Equation (2.10), while the time dependent evolution has been summarised in Figure 2.12.

$$F_x = m_x \frac{d^2\ r_x(t)}{dt^2} \tag{2.10}$$

The force evaluation can be performed using several algorithms which assume that the atomic positions and dynamic properties can be appropriately represented as a Taylor series of expansions. The most commonly used algorithm for force evaluation is the Verlet algorithm [201], which calculates the new positions based upon the acceleration and coordinates of previous positions. Other related algorithms include the Velocity-Verlet [202, 203] and Leap-Frog algorithm [202], both of which are modified versions of the Verlet algorithm. In case more accurate velocity consideration is needed during the simulation, the Beeman algorithm [204] can be used.

A typical MD simulation starts with selection of a model system of “N” particles followed by solving Newtons equation for all of the atoms until the system properties do not exhibit any significant change, i.e., they get equilibrated. The majority of data collection then takes place in the post-equilibration phase, depending upon the quantity to be measured. Generally, the “N” particles are assigned random velocities determined by a Maxwell-Boltzmann distribution that gives probabilities to ascertain the velocity (v_i) of an atom (a_i)) with mass (m_i) at a defined temperature T [95]. The initial configuration fed to the Maxwell-Boltzmann distribution can be obtained from experiments, theoretical models or a combination of both. Another important parameter related to the simulation directly affecting the force evaluations as well as system evolution is the **time step**, whose value has to be selected after careful consideration. A large timestep results in difficulties in numerical integration of the forces, while a smaller step size increases data output and simulation time at the cost of covering only a small portion of the phase space [95]. A proper selection of the time step enables appropriate numerical integration of forces and by introducing random alterations in the same results in a time dependent evolution of a system in an unbiased manner. The recommended times for capturing key atomic properties have been summarised in Table 2.1

Another thing to be considered in case of MD simulations is the ability to change the timestep without affecting the physical result of the simulation. In general, if the system was an ideal system (i.e., rigid body), changing the timestep would not change the results. This is, however, not possible in practice especially with biomolecules, thereby necessitating a constraint on bonds and angles to avoid any detrimental effect on accuracy of the simulation following a change in timestep. This simply implies that in case of biomolecules the low frequency (physically relevant) movements are coupled [95], while the high frequency motions (e.g. bond vibrations) are usually independent. A well known example of atomic constraints is SHAKE, which has been further improved by Tobias

Table 2.1 Recommended timestep for simulating different types of atomic motions according to Leach, 2001 [95]

System	Motions exhibited	Timestep
Atoms	Translation	10 fs
Rigid Molecules	Translation, Rotation	5 fs
Rigid bonds in flexible molecules	Translation, Rotation, Torsion	2 fs
Flexible bonds in flexible molecules	Translation, Rotation, Torsion, Vibration	1/0.5 fs

and Brooks [95, 205, 206]. SHAKE uses holonomic constraints on atoms which implies that the coordinates of the constrained atoms are connected and thereby the equations describing their motion [95]. Some of other popular constraints are RATTLE [207] and SETTLE [208].

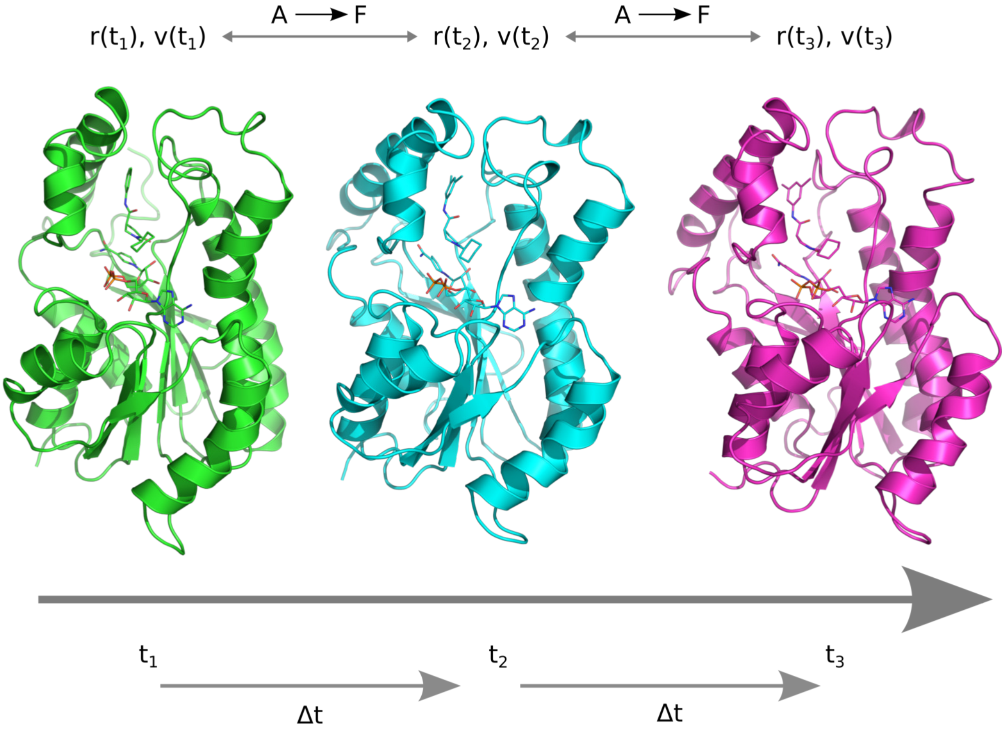


Figure 2.12 Schematic representation of structural evolution during an MD simulation. For every time step denoted as Δt, the position (r(t$_i$)) and velocity (v(t$_i$)) are evaluated for each atom *i*. The molecular mechanics force field function (A) is used to derive the underlying force F$_i$. Figure adapted and redrawn from Sotriffer, 2006 [209].

2.3.3.4 Thermodynamic ensemble sampling - Ensembles in MD simulations

The process of MD simulation usually deals with N particles that are represented as a set of isolated systems, each with a unique state of energy range (E, E + δE). In statistical mechanics, such a collection is referred to as an ensemble, which are of three types namely, **micro-canonical, canonical and grand-canonical**. In case of the micro-canonical ensemble, also called the **NVE** ensemble, the number of particles (N), the total volume V, and the potential E, remain constant. The constant energy signifies no heat exchange with surroundings, thereby implying that the NVE ensemble corresponds to an adiabatic process, making it unsuitable for general energetic equilibration of dynamic systems like biomolecules [95]. The equilibration of the isolated system can be achieved by weak coupling to a heat bath [210] or by Langevin dynamics [211], thereby conserving N, V, and $\boldsymbol{T}$, i.e. the temperature instead of the energy. This system of constant NVT is referred to as the canonical ensemble. Once the total energy ($\mathbf{E}_{tot}$) and the temperature have been stabilised, constant pressure simulations (**NPT**) or isothermal-isobaric can be initiated. Such simulations usually require pressure management, with some prominent examples like Nosé-Hoover barostat [212] and Berendsen barostat [210].

Furthermore, the simulation of the system takes place in a unit cell which can have various shapes like sphere, cubic, octahedral etc. The presence of artefacts at the boundary of the cell is sidestepped by means of periodicity (periodic boundary conditions [PBC]) that creates infinite copies of the system by translation [213]. Hence, particles of the system that exit the cell boundary get replaced by an identical copy on the opposite side [213]. This approach cannot be utilised in case of simulations involving spherical cells (spherical boundary conditions) [214]. However, the use of periodicity also leads to increased computation time. Moreover, since non-bonded interactions exist in between particles of the system, estimating them for all particles is impractical. Although the estimation of van der Waals interactions is much easier because of their rapid decay with distance, the estimation of long-range electrostatics is more difficult since they decay slowly (with r^{-1}). As a result, the long range electrostatics in systems with periodic boundary conditions can be calculated with the particle-mesh Ewald (PME) method that utilises fast fourier transformed Ewald summation of the entire system [215]. Other methods to calculate long-range electrostatics include local reaction field [216] and group-based truncation [217].

2.3.3.5 Simulating long durations - Enhanced Sampling

One of the key considerations while performing molecular dynamics is its duration, since the extent up to which the important parts of the configurational space get sampled depends on the sampling algorithm. An ideal algorithm must exhibit ability to overcome a multitude of energy barriers common to macromolecules. This is, inspite of the progresses in computational power as well as simulation algorithms [218]. Additionally, important biological events like protein folding and even comparatively large conformational transitions (e.g. R to T haemoglobin) take tens of microseconds to even minutes, for the former place an upper limit on length of simulations [188]. Hence, as a means to overcome this prime limitation of sampling, several theoretical methods referred to as **enhanced sampling methods** have been developed [128, 219]. This group of methods mainly aim to avoid Boltzmann statistics while retaining the current distribution of states in the conformational ensemble.

In regards of current work and in general protein-ligand binding, the free energy or potential of mean force (PMF) [220] upon ligand binding is the key parameter being calculated. Free energy or PMF are related to collective variables (CV) that are probability density functions describing a particular event. Collective variables are also functions of cartesian variables of the system [219] (e.g. angles, bonds). Some of the notable equilibrium methods used to perform enhanced sampling include Umbrella Sampling [221, 222] and Accelerated molecular dynamics [223, 224]. The notable examples of non equilibrium methods for enhanced sampling include Replica Exchange Hamiltonian Metadynamics (h-REMD) [225] and Steered Molecular Dynamics [226].

2.3.4 Analysis of Docking and MD simulations

There are numerous possibilities for analysing and interpreting the information emanating from molecular docking and MD simulations. Some of the principal metrics to evaluate the molecular docking include the docking and rescoring. Additionally, one can quantify the quality of the pose placement by investigating the RMSD (including substructure) as compared to a reference structure. The RMSD can be simply shown as Equation (2.11):

$$RMSD = \sqrt{\frac{1}{N}\sum_{j=1}^{N} d_j^2} \tag{2.11}$$

where d_j represents the distance in between each of the N atomic pairs consisting of equivalent atoms. The analysis of molecular dynamics on the other hand is quite varied and ranges from the simple RMSD calculations to advanced collective variable (CV) calculations and conformational analysis. The theory and application of the respective methods for MD analysis will be discussed briefly in part II of this thesis.

Chapter 3

Binding Mode Prediction for Pyrrolidine carboxamides: Molecular Docking and Induced-fit

3.1 Introduction

Molecular docking forms one of the main stays for modern HTS procedures like virtual screening (VS) as well as structure based design with affinity prediction often being one of the main criterion's driving the process [129]. It has been recently observed that kinetics and residence time also play an important role in determining the efficacy of InhA inhibitors [49, 72, 74, 227]. Given the fact that loop ordering ability of a molecule plays a critical role in determining its affinity as well as residence time (t_R), the importance of structural features, ligand placement in the binding pocket and its subsequent effect on the substrate binding loop (SBL) ordering remains one of several key criteria to be studied during rational optimization of InhA inhibitors. The ligand placement and its effect on the ordering of the SBL can be assessed *in-silico* with numerous techniques, namely molecular docking and ensuing MD simulations based upon the ligand orientations emanating from docking.

Molecular docking aids in predicting the initial ligand binding orientations that can be evaluated thoroughly via the more intensive molecular dynamics simulations. Since docking is a fast and "approximate" technique aimed at binding mode prediction, certain aspects have to be carefully considered. These are related to the quality of the poses, the ranking ability of the docking program and incorporation of receptor flexibility during docking. The latter is more important in case of InhA, since the flexibility of the substrate binding loop and its ordering upon ligand association play a central role in determining the inhibitory potential of the ligand. This can be seen in case of potent and long acting InhA inhibitors (diphenyl ethers) which bind via a two-step mechanism also referred to as "slow-tight binding". Prime examples of slow-tight binders are PT-70 (PDB 2X23) and PT-92 (PDB 4OHU). On the contrary, several other classes of moderately potent InhA inhibitors lack the ability to bring about loop ordering and closure, for example pyrrolidine carboxamides, arylamides, and triclosan (PDB 2H7M, 2NSD, 2B35, respectively). These inhibitors (with sub-nanomolar affinity) are assumed to exhibit

rapid reversible binding that is in line with their experimentally determined affinity and "apparent" residence time (cf. Chapter 2).

MD simulations provide an attractive way to elucidate the dynamics and kinetics of binding for InhA inhibitors exhibiting slow-tight as well as rapid-reversible binding. Prior to the dynamics investigation, well defined binding modes are an essential prerequisite for molecules with unknown binding orientation. In case of congeneric series of ligands, the binding modes can be assumed to resemble related crystal structure ligand orientations. Moreover, the conformation of a protein-ligand complex as seen in a crystal structure can be viewed as a representative of all low energy conformations that can be visited by the protein and ligand alike. Molecular docking provides for a means to ascertain the aforesaid assumptions.

Accordingly, this study has the following aims:

1. Establishing a binding orientation prediction protocol (docking procedure) that accommodates the substrate binding loop (SBL) flexibility during docking and subsequent scoring, with a reasonable accuracy. The efficiency of the docking protocol should be evident in form of high enrichment in a typical ROC (receiver operating characteristic) analysis [228].

2. Provide for a means of clear separation in between active and inactive compounds, i.e., activity-based classification.

The receptor flexibility can be accommodated during docking by various means described earlier (cf. Chapter 2 and section 2.3.2). With a focus on SBL flexibility, across many cases, a simple redocking in structures representing varying conformations of SBL was performed [229–237]. In rare cases, the cross-docking and ensemble docking methods were utilised to consider the flexibility of the SBL alongwith subsequent pose quality computations [78, 238, 239]. Whilst most docking studies focussed on pose prediction of known inhibitors with unknown binding mode, quite few dealt with prediction of binding modes for novel molecules [240, 241]. Thus, the focus of the current study also aimed at refinement of the binding protocol in order to predict the binding mode for novel molecules in addition to the aims mentioned earlier.

While considering the previously mentioned facts and the aims, the outline of the study is as follows: We first focus on validating the performance of the ligand placement algorithm and subsequent scoring functions in binding mode prediction and pose enrichment. To this aim, four different PDB structures in their monomeric form (1P44, 2H7M, 2NSD, and 2X23; chain A) were extensively used. These structures represent various conformations of the substrate binding loop (SBL) as well as four principal series of InhA inhibitors, namely: Genzyme series (arylamide derivative, 1P44); Arylamides (2NSD), Diphenyl ethers (2X23) and Pyrrolidine carboxamides (PDB 2H7M). Starting with the crystal structure ligands covering all of the aforementioned series of InhA inhibitors, a cross docking was performed across all four proteins using Glide single precision (SP) mode [134]. This docking mainly aided in preliminary evaluation of the binding mode reproduction by the docking algorithm. The post-docking analyses consisted of RMSD calculation with respect to the respective crystal structure ligand. A rescoring with DrugScoreX [142] was also performed to assess the **"nativity**" of the docked poses. These post-docking analyses are quite useful in ascertaining the performance of the ligand placement algorithm and the subsequently utilised scoring function.

Subsequently, an InhA inhibitor dataset covering the aforementioned chemical series (N=113, Appendix A) was constituted from literature sources [50]. A key criterion satisfied by all molecules of this dataset was well defined InhA inhibitory activities (as IC_{50}) measured with the same assay. This dataset was docked across all four proteins with **Glide SP** followed by the post-docking analyses. This step primarily aided in evaluating the performance of the docking algorithm in pose prediction for different series of InhA inhibitors with variable molecular weight and sizes.

Upon conclusion of the pose prediction, the ranking and pose enrichment capability of docking can be evaluated. In other words, the ability of the protocol in differentiating binders from non-binders and subsequent ranking of the poses according to their experimentally determined activity. This can be achieved through the use of viable and statistically derived decoy datasets (DUD, DUD-E) [242, 243] that assist in the pose enrichment. Another approach for achieving pose enrichment is the use of scoring functions tailored for pose enrichment. In current context, the Glide scoring functions (Gscore and XPscore) were used for initial pose selection and enrichment. This approach was further refined by using DrugScoreX to ascertain the quality of the poses, thus mimicking a **step-wise consensus based scoring**.

Since the outreach of this work is structure-based optimization of pyrrolidine carboxamides, the InhA inhibitor dataset was subsequently docked across the four representative proteins using **Glide extra precision (XP)** mode [244] and induced fit [116]. This was followed by rescoring with DrugScoreX [142] and SFCscore [130, 136]. However, only the pyrrolidine carboxamides were analysed extensively in line with the aims of this thesis. The pose selection was primarily aided by the **Glidescore XP** or **XPscore**, a "semi-empirical scoring" function especially developed for pose enrichment [244]. GlideXP docking is meant to be used only if GlideSP is successful in yielding a reasonable docked pose. In the current work, the resultant poses were selected by a consensus based scoring using three different scoring functions derived by different approaches that mainly serve to validate the results of XPscore. The stepwise pose selection and rescoring approach is discussed further in Section 3.1.3.

In order to incorporate flexibility of the SBL during docking, a new procedure termed *docking with induced fit* [116, 117] or **Induced-fit docking** (IFD) was tried upon for the entire pyrrolidine carboxamide dataset, with a focus on pose prediction and correct affinity-based ranking for the bulkier members of the dataset. The bulky pyrrolidine carboxamides were expected to cause issues when docking in a narrow binding pocket exhibited by PDB 2X23 or even 2NSD. The pose selection and rescoring strategy derived from earlier steps was also continued in this approach. The ensuing sections describe the theory of the ligand placement algorithm and scoring functions used in Glide as well as Induced fit docking. Accompanying them is a description of the pose selection strategy using a RMSD-consensus based scoring approach followed by results of the docking and lessons learnt from the same.

3.1.1 Docking in Glide

The overall docking process including rescoring and pose selection for all of the InhA inhibitor dataset is outlined in Figure 3.1. There are two main approaches for docking in Glide, each suited for a different purpose, namely, single precision mode (SP) (for virtual screening and initial binding mode prediction) and extra precision mode (XP) (for pose enrichment in post-virtual screening/initial docking via GlideSP)

The basic methodology for ligand placement and scoring is depicted in Figure 3.2, highlighting the thorough systematic approach of Glide in sampling the conformational, positional and orientational space for small molecules [134]. The program utilises a series of filters in a step-wise manner to probe the ligand orientations in the receptor's active site. The process initiates with preliminary coarse placement of the ligand in the active site. The ligand placement is guided by numerous fields that describe the receptor properties (shape, orientation of the amino acids etc.) and progressively enable proper ligand placement within the binding site [134].

An exhaustive conformational sampling of the ligand follows the initial placement. This process is further refined by a torsionally flexible energy minimization using a molecular mechanics force field (MMFF; not to be confused with Merck Molecular Force Field with same initials) like OPLS-AA [195, 245] together with a distance dependent dielectric field to consider polarization effects. The ligand sampling typically considers a comprehensive catalogue of conformations around the torsion-angle space of the ligand. The final few poses (usually 3-6) emanating from the flexible energy minimization undergo a Monte-Carlo simulation to evaluate the minima around the ligand torsions [134] yielding the final poses that are scored by the GlideSP score (Gscore) and ranked by a hybrid molecular mechanics-empirical scoring function (Emodel).

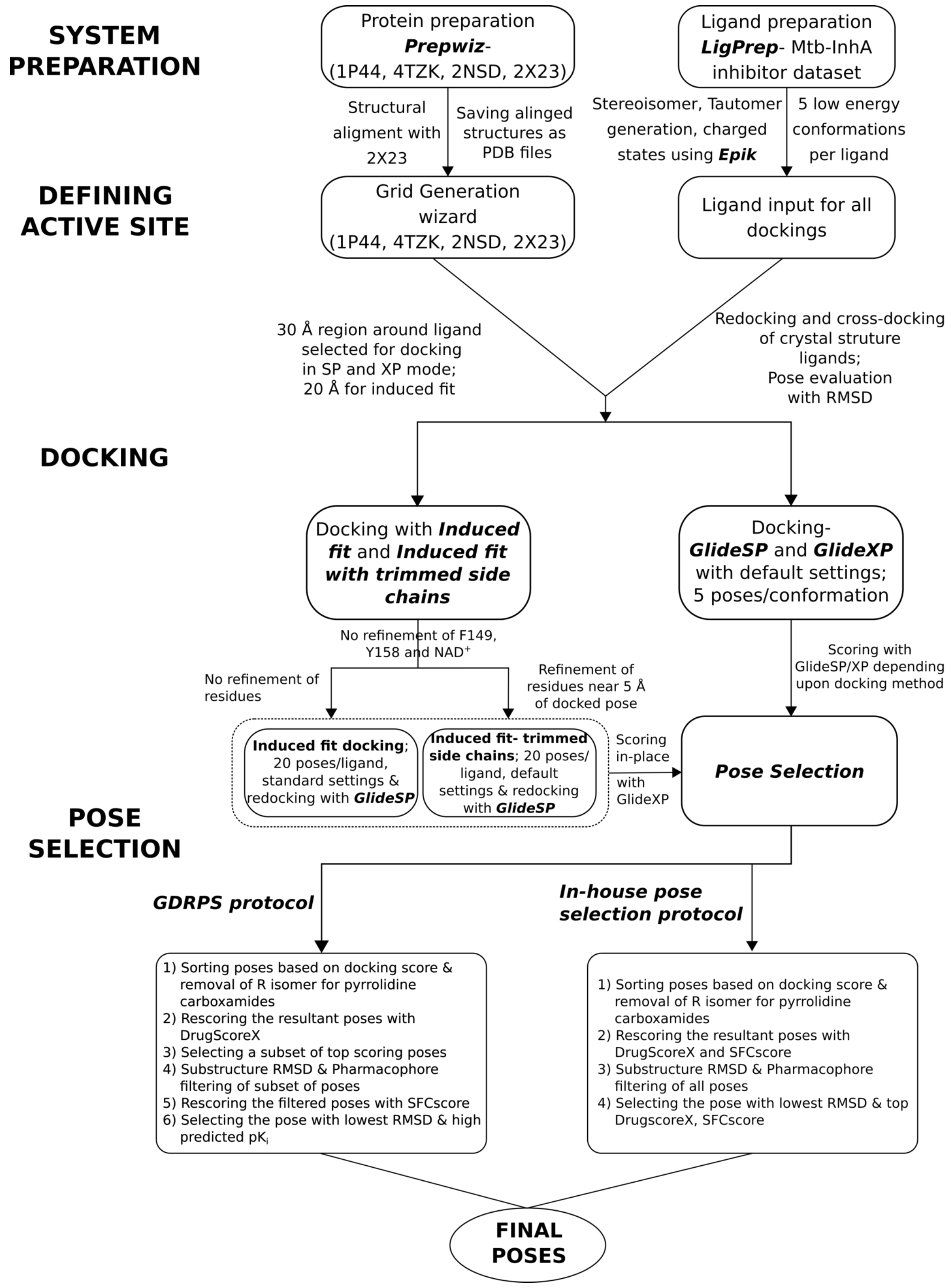

SYSTEM PREPARATION
Protein preparation
Prepwiz-
(1P44, 4TZK, 2NSD, 2X23)
Ligand preparation
LigPrep- Mtb-InhA inhibitor dataset
Structural aligment with 2X23
Saving alinged structures as PDB files
Stereoisomer, Tautomer generation, charged states using Epik
5 low energy conformations per ligand
DEFINING ACTIVE SITE
Grid Generation wizard
(1P44, 4TZK, 2NSD, 2X23)
Ligand input for all dockings
30 Å region around ligand selected for docking in SP and XP mode; 20 Å for induced fit
Redocking and cross-docking of crystal struture ligands; Pose evaluation with RMSD
DOCKING
Docking with Induced fit and Induced fit with trimmed side chains
Docking- GlideSP and GlideXP with default settings; 5 poses/conformation
No refinement of F149, Y158 and NAD$^+$
No refinement of residues
Refinement of residues near 5 Å of docked pose
Scoring with GlideSP/XP depending upon docking method
Induced fit docking; 20 poses/ligand, standard settings & redocking with GlideSP
Induced fit- trimmed side chains; 20 poses/ ligand, default settings & redocking with GlideSP
Scoring in-place with GlideXP
Pose Selection
POSE SELECTION
GDRPS protocol
In-house pose selection protocol
1) Sorting poses based on docking score & removal of R isomer for pyrrolidine carboxamides
2) Rescoring the resultant poses with DrugScoreX
3) Selecting a subset of top scoring poses
4) Substructure RMSD & Pharmacophore filtering of subset of poses
5) Rescoring the filtered poses with SFCscore
6) Selecting the pose with lowest RMSD & high predicted pK$_i$
1) Sorting poses based on docking score & removal of R isomer for pyrrolidine carboxamides
2) Rescoring the resultant poses with DrugScoreX and SFCscore
3) Substructure RMSD & Pharmacophore filtering of all poses
4) Selecting the pose with lowest RMSD & top DrugscoreX, SFCscore
FINAL POSES

Figure 3.1 Overall docking process and pose selection protocol: The protocol described here includes three different docking approaches to consider receptor flexibility. The first two utilise different crystal structure representatives of the different conformations of the substrate binding loop. The latter approach generates the conformations on-the-fly using the procedure mentioned in Sherman, Day, et al., 2006 [116]. The post docking process consists of using RMSD and visualization coupled with consensus scoring to get desired binding orientations.

3.1.1.1 Scoring of Poses - Glide SP score and XP score

A key component as well as a desirable ability of the docking program/protocol is its efficiency in binding affinity prediction and ranking the poses. A key consideration

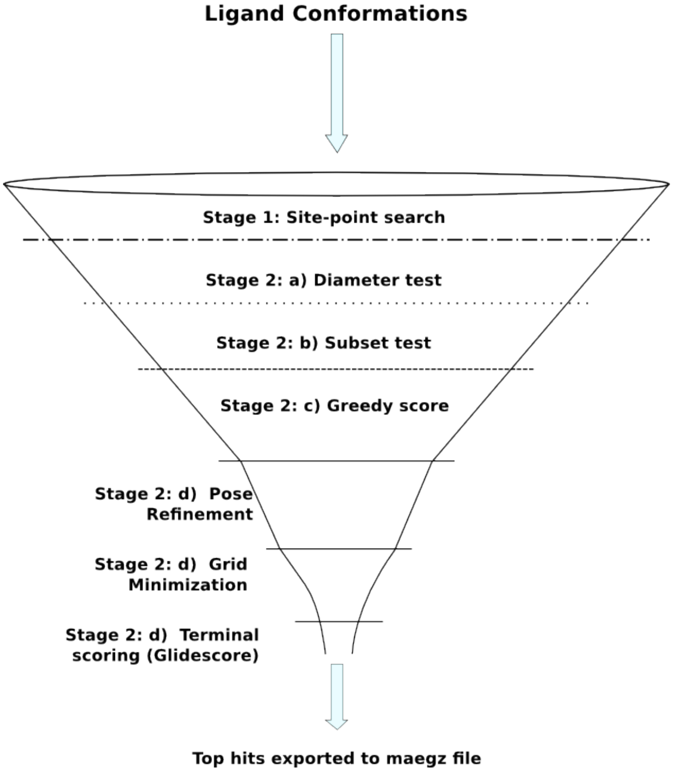

Ligand Conformations
Stage 1: Site-point search
Stage 2: a) Diameter test
Stage 2: b) Subset test
Stage 2: c) Greedy score
Stage 2: d) Pose Refinement
Stage 2: d) Grid Minimization
Stage 2: d) Terminal scoring (Glidescore)
Top hits exported to maegz file

Figure 3.2 Overall docking process in Glide. Figure adapted and redrawn from Friesner, Banks, et al., 2004 [134].

for scoring poses in Glide is the need for modification of molecular mechanics terms (primarily non-bonded), since the input structure of the protein is not optimized for a particular ligand [134]. Mimicking modest "induced fit" and thereby better fitting of protein and ligand is then achieved by scaling of the van der Waals component [246]. The scoring function implemented in Glide is empirical in nature and is based on another function, i.e. Chemscore [135] (cf. Equation (3.1)).

$$\begin{aligned}\Delta G_{bind} = C_0 + C_{lipo} \sum f(r_{lr}) + C_{hbond} \sum g(\Delta r) h(\Delta \alpha) \\ + C_{metal} \sum f(r_{lm}) + C_{rotb} H_{rotb}\end{aligned} \tag{3.1}$$

where C$_{lipo}$ expands over entire ligand-atom/receptor-atom pairs deemed lipophilic, C$_{hbond}$ extends over all ligand-receptor hydrogen-bonding interactions. The *f, g*, and *h* are functions that give full score (1.00) for normal bonds and angles, while partial scores (1.00-0.00) are awarded to those outside normal limits but within a larger threshold. The GlideScore 2.5 (or GScore) (Equation (3.2)) is a modified and expanded version of Chemscore, mainly because of its inadequacy in scoring and ranking ligands with varying net charges.

$$\begin{aligned}\Delta G_{\text{bind}} = {} & C_{\text{lipo-lipo}} \sum g(\Delta r)\ h(\Delta\alpha) + \\ & C_{\text{hbond-neut-neut}} \sum g(\Delta r)\ h(\Delta\alpha) + \\ & C_{\text{hbond-neut-charged}} \sum g(\Delta r)\ h(\Delta\alpha) + \\ & C_{\text{hbond-charged-charged}} \sum g(\Delta r)\ h(\Delta\alpha) + \\ & C_{\text{max-metal-ion}} \sum f(r_{lm}) + C_{rotb} H_{rotb} + \\ & C_{\text{polar-phob}} V_{\text{polar-phob}} + \\ & C_{coul} E_{coul} + C_{vdW} E_{vdW} + \text{solvation terms}\end{aligned} \tag{3.2}$$

The salient features of the Glide scoring function (simply **Gscore**) as compared to Chemscore include an improved hydrogen bonding term that is partitioned into individual weighted terms dependent on the charged state of the donor-acceptor pair. The anion-metal interactions are accounted by considering best metal-anion ligation in event of multiple metal ligations and charge dependent addition or suppression of preference for anionic ligands [134]. The presence of a polar non H-bonding species situated in hydrophobic pocket is rewarded and is incorporated in Gscore by terms from SiteMap [247–249]. Other notable additions to Gscore include improved description of ligand-receptor non-bonded interactions and incorporation of solvation effects by a empirical solvation model [134].

Furthermore, GlideSP docking and **Gscore** are a part of a procedure adept at recognising ligands with a propensity to bind to a given receptor even in cases of imperfect binding [134]. This makes GlideSP/Gscore suitable mainly for pilot virtual screening and initial pose ensemble generation that serves as an input for the ensuing stringent GlideXP docking. GlidescoreXP or **XPscore** (cf. Equation (3.3)) on the contrary to Gscore, is a scoring function aimed at *"semi-quantitatively ranking the ability of candidate ligands to bind to a specified conformation of the protein receptor"* [244]. XPscore enforces several strict penalties for non-conforming poses (for e.g. clashing poses, those with unfavourable torsions or unusual geometry etc.) to obtain better scoring poses that eventually result in a much higher pose quality (enrichment) as compared to Gscore.

$$\begin{aligned}\text{XPscore} &= E_{\text{coul}} + E_{\text{vdW}} + E_{\text{bind}} + E_{\text{penalty}} \\ E_{\text{bind}} &= E_{\text{hyd-enclosure}} + E_{\text{hb-mn-motif}} + \\ &\quad E_{\text{hb-cc-motif}} + E_{\text{PI}} + E_{\text{hb-pair}} + \\ &\quad E_{\text{phobic-pair}} \\ E_{\text{penalty}} &= E_{\text{desolv}} + E_{\text{ligand-strain}}\end{aligned} \tag{3.3}$$

The implementation of GlideXP docking and scoring is essentially the same as GlideSP (cf. Section 3.1.1) with a few caveats. Firstly, GlideXP explicitly employs a broader sampling of ligand conformations to ensure maximum diversity amongst the structures to be docked [244]. Secondly, the performance of GlideXP is tied with that of GlideSP in correctly docking the core part of the ligand in the binding pocket. Subject to satisfying this prime condition, GlideXP extensively uses the anchors to build up the entire molecule systematically and derive better scoring poses from their initial states (ligand fragments). This is achieved by clustering the anchors followed by selection of suitable cluster representative, stepwise build-up of the molecule by addition of rotamer groups and coarse scoring of the intermediate ligands to select nascent high resolution conformations (up to 4 degrees per rotatable bond) [244]. This stepwise sampling addresses difficult docking scenarios while avoiding penalties for the selected conformation.

The molecules emanating from the previous step are then selected on basis of their scores and lack of steric clashes with amino acid side chains. The selected poses are then extensively minimised by a Glide-specific total energy function employing a distance-dependent dielectric to assess their electrostatic interactions. The minimised poses are then ranked with the Emodel pose selection function of Glide [134]. Finally, a subset of top ranking structures are assessed with a grid-based water addition method followed by calculation of penalties and computation of the final XPscore.

3.1.2 Docking with induced fit

All of the aforesaid docking procedures employed pre-generated conformations (PDB structures) to consider receptor flexibility during docking. This section describes a procedure that employs consecutive ligand placement and complex optimisation to account for the effect of receptor flexibility on the final poses (cf. Section 2.3.2.2). The overall process of induced fit (IFD) employed in Glide consists of four main steps [116]:

1. Ensemble generation via docking into rigid receptor with softened vdW potentials (”Soft-docking” phase).
2. Sampling low energy protein conformations for individual poses from previous step (Mutation and protein minimisation phase using Prime).
3. Redocking of the poses in the low energy protein conformations (Docking with normal non-bonded potentials).
4. Composite scoring of the final poses that takes into account the binding energy (Gscore) together with solvation terms and receptor strain (Prime energy minimisation terms).

Each step is faced by its respective challenges beginning with prediction of reasonable ligand poses in the first step followed by approximation of the protein structure for the predicted pose. This is followed by a problem in guessing the correct low energy conformation for the ligand that fits to the protein structure from the previous step followed by ranking of the predicted pose. The entire procedure is repeated in the event that top ranked scores from the penultimate step have similar scores [116]. In such cases, the top ranked poses are redocked into the optimised protein structure in absence of a soft potential [117]. The entire process of docking with induced fit has been summarised in Figure 3.3.

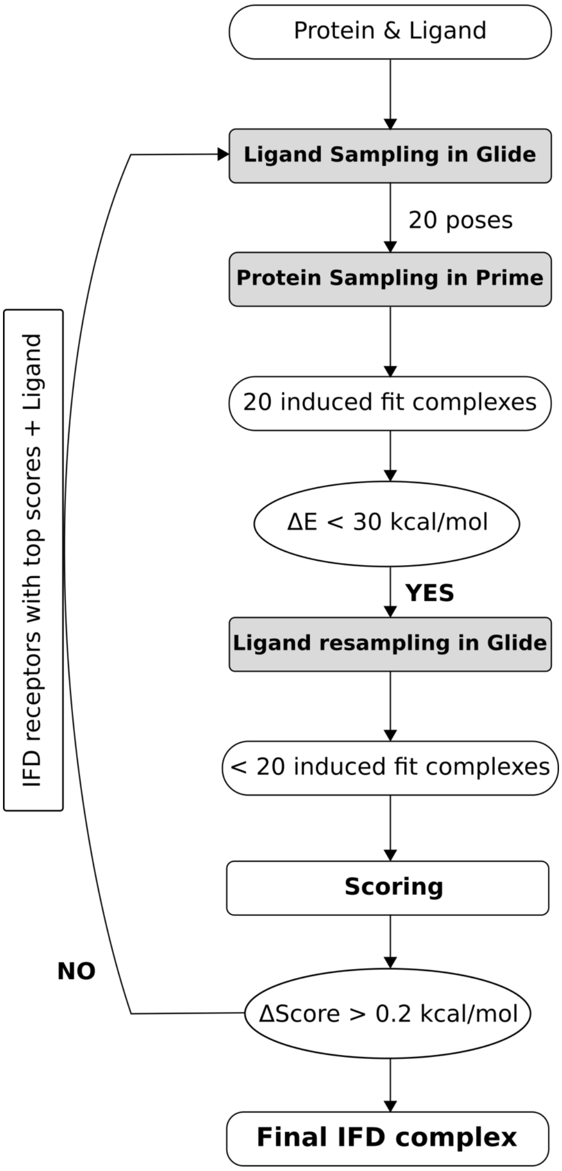


Figure 3.3 Basic workflow of docking with induced fit: The most time-consuming steps of the workflow have been highlighted, with ΔE representing the energy difference between the pose in consideration and the pose with lowest energy. Figure adapted and redrawn from Sherman, Day, et al., 2006 [116].

It is amply clear that the most time consuming steps in the workflow pertain to the conformational sampling of both ligand and protein (in ascending order). As a corrective measure, in the initial ligand sampling, the van der Waals radii of both protein and ligand are scaled down to 50% of their initial values that results in reduction of steric clashes. Subsequently, flexible residues of the binding pocket are mutated to alanine and back upon the conclusion of ligand placement [116]. The most time-consuming step of the workflow, i.e., receptor sampling, only considers the top 20 poses from the previous step to save time, during which each of the 20 protein-ligand complexes undergoes a full minimisation in Prime [250–253] using the OPLS-AA force field [195, 245] and a surface Generalized Born implicit solvation model [254, 255]. The optimised ligand structures from the first step are now redocked into the minimised protein structures with default vdW scaling for all atoms followed by scoring of the 20 protein-ligand complexes emanating from the procedure [116].

3.1.2.1 Scoring of induced fit poses-Induced fit score

The process of scoring the final poses for each of the 20 protein-ligand complexes emanating from the IFD procedure is performed with the aid of a compound scoring function (IFDScore) that takes into account the protein-ligand interaction energy and the prime minimisation energy scaled by a factor of 0.05 [116]. This factor was considered adequate to weed out any unusual protein structures from the second step of the workflow. The large energy difference (30 kcal/mol) employed as a filter ensures that aberrant protein-ligand conformations do not contribute to the overall noise in the final score [117]. The "similarity" of two top ranked poses is judged by considering the overall energy difference in between them. For values below 0.20 kcal/mol, the entire procedure mentioned in Figure 3.3 is repeated [116].

3.1.3 Pose selection

The process of pose selection is an interesting and challenging procedure as far as docking is concerned. Scoring functions when used solitarily or in combination have some advantages as well as shortcomings [94, 129]. The post-docking analysis and subsequent choice of top pose is largely user dependent. Often, the choice of a particular pose is based on the reasonable assumption that the scoring function used for ranking demonstrates a satisfactory accuracy. However, it is clear that in many cases, use of a solitary scoring function in scoring and ranking poses is seldom enough [129, 145]. As a consequence, one of the primary aims of this study was the development of a step-wise pose selection protocol that utilised a combination of scoring functions, the root mean square of deviation (RMSD) and a pharmacophore model in pose selection (cf. Figure 3.1).

The pose selection procedures of this study were aimed at reliable estimation of the binding mode followed by reasonably accurate affinity prediction and affinity-based ranking. A key desirable feature of the protocol was minimal need of manual intervention which should appear at the very end for visualization of the ranked poses. The current study made use of two *”hierarchical consensus scoring”* pose selection procedures, depending upon the **consideration of the docking score** and **stage of pharmacophore model use** in pose selection. The term *”hierarchical consensus scoring”* stems from the fact that a combination of semi-empirical (Glidescore), knowledge-based (DrugScoreX) and empirical scoring functions (SFCscore) were utilised in stepwise manner alongwith RMSD and pharmacophore filtering to achieve reasonable binding orientations ranked on basis of their predicted binding affinity.

The first of the two procedures abbreviated as **”GDRPS”** (**G**lidescore **D**rugScore **R**MSD **P**harmacophore **S**FCscore) considers the contribution of Glidescore and DrugScoreX in determination of the pose quality. The subset of poses being assigned top ranks by both scoring functions were then subjected to further evaluations that yielded the desired pose. The GDRPS procedure (Figure 3.1) consists of the following steps:

1. **Step 1:**
   In this step, the poses get scored and ranked (affinity-based) by the semi-empirical scoring functions Gscore, XPscore, or IFDscore depending upon the docking protocol utilised. The ranked poses are sorted based on their docking score, hydrogen bond count and stereochemistry in decreasing order of hierarchy. The sorted poses are collectively exported to a multimol2 file for further processing. In case of pyrrolidine carboxamides, the crystal structure ligand exists exclusively in the *S* configuration [52], while all other molecule series (diphenyl ethers, Genzyme and arylamides) were nearly devoid of stereoisomers. In order to evaluate the discriminating ability of the docking algorithm (binders vs. non-binders), the R isomers were also docked across all four proteins. In the current case, the R isomers of pyrrolidine carboxamides being inactive served as ”non-binders”. Hence, in case of pyrrolidine carboxamides, all R isomers were simply not considered for the post-docking analysis.

2. **Step 2:**
   This step consisted of rescoring the combined poses from the preceding step with DrugScoreX. At this stage, a small subset of top scoring poses (2-3) were manually chosen as follows: Starting with top 5 ranking poses from docking (for each compound), a rescoring with DrugscoreX was performed. From these 5 molecules, irrespective of the docking score, a total of 2 to 3 top ranking poses with high DrugscoreX values were combined into a multimol2 file and forwarded to the next stage. The central assumption in performing the trimming of the rescoring output was that top scoring poses from Glide should also show up within the top few poses of DrugScoreX.

3. **Steps 3 and 4**

   The combination of steps 3 (substructure RMSD in fconv [256]) and 4 (pharmacophore filtering in MOE [257]) constituted a phase aimed at obtaining the final poses that were subjected to an additional affinity prediction with SFCscore [130, 136]. The prime motive of using pharmacophore and RMSD evaluation is as follows:

   - The RMSD calculation usually refers to the substructure RMSD of molecules other than in case of crystal structures. The substructure RMSD signifies the deviations of the maximum common structure (i.e. scaffold) in the top ranked pose with respect to the corresponding crystal structure of its series. A low value (< 1.5 Å) of the substructure RMSD indicates appropriate placement of the core scaffold in the binding pocket while a higher value indicates improper placement of the scaffold (cf. Figure 3.4).

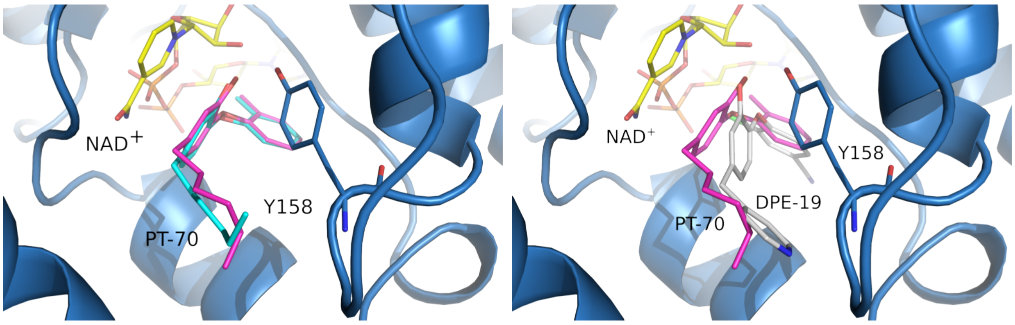


Figure 3.4 Utility of substructure RMSD: The picture on the left depicts a pose with a low substructure RMSD indicating a better placement in the binding pocket. On the contrary, the one on the right has a much higher RMSD and is clearly improperly placed.

   - The pharmacophore model derived from the PDB codes 1P44, 2H7M, 2NSD and 2X23 ensures adherence to the spatial location and orientation of the hydrogen bond donor-acceptor pair, i.e., the catalytic Tyr158-OH group and cofactor (NAD$^+$) ribose oxygen. The pharmacophore filtering serves as a strict criterion by ensuring removal of poses not satisfying the conditions laid down in the pharmacophore model. Thus, poses emanating from pharmacophore filtering conform to the bonding pattern displayed by the crystal structure ligands.
   - The RMSD and pharmacophore filtering carried out in parallel was expected to yield poses with better scores as well as reasonable binding modes.

4. **Steps 5 and 6:**

   The final steps of the GDRPS protocol involved the affinity prediction of the final top ranked poses with SFCscore followed by selection of the pose with the highest affinity. SFCscore is a composite group of eight empirical scoring functions that predict the binding affinity of a given pose as pK_i units. Manual intervention was

> kept to minimal levels and was only needed at two stages, namely, combining all poses (including removal of R isomers of pyrrolidine carboxamides) into a single file and creating a subset of top ranked poses from Glidescore and DrugScoreX. All of the aforesaid steps were achieved by sequential use of scripts originally written by Steffen Wagner, University of Würzburg and modified for an additional rescoring step with SFCscore.

The other pose selection procedure developed in-house by Steffen Wagner, University of Würzburg (abbreviated **I**n-**H**ouse **P**ose **S**election (IHPS)), exclusively made use of DrugScoreX score for selecting binding poses, followed by RMSD-Pharmacophore and rescoring with SFCscore for additional pose evaluation (cf. Figure 3.1). In comparison with GDRPS, IHPS does not consider the contribution of Glide scoring function (Gscore/XPscore/IFDscore) in the overall binding pose selection procedure. The steps involved in IHPS are essentially the same as those mentioned previously except in **step 2**. Whereas, the *GDRPS* scheme considered a subset of top scoring poses from Glide and DrugScoreX for further evaluation, the IHPS procedure submitted the entire output of DrugScoreX and SFCscore to ensuing assessments, primarily because this pose selection procedure was aimed at selecting binding modes for novel molecules with unknown binding modes.

3.2 Methods

3.2.1 Protein Selection

Consideration of receptor flexibility during docking is quite important in case of InhA, mainly because of the underlying induced fit mechanism that drives the protein-ligand association. Other factors to be considered include the structure and resolution of the input protein structures [94]. In the current case, a completely resolved SBL in the crystal structure was a key requirement to understand the process of ligand binding to InhA. As a consequence, four high quality crystal structures of InhA (PDB 1P44, 2H7M, 2NSD, and 2X23) that exhibited the substrate binding loop (SBL) in decreasing order of “openness” (cf. Figure 2.3) were chosen for docking.

Of these structures, the low resolution of 1P44 (2.7 Å) indicates a higher uncertainty in the resolved structural features. In spite of this, it was used as a reference for the ”open” form of the SBL since it exhibits a better resolution than PDB 1BVR (resolution 2.8 Å) which also exhibits the SBL in open form. In case of the pyrrolidine carboxamides, the older crystal structures were refined and superseded by newer codes (Table 3.1). Upon comparing the old and new PDB codes, negligible differences in the binding site residues were observed (cf. Table 3.2). Another noteworthy change in the new crystal structures was that the bound ligand did not exhibit an aberrant conformation for the amide bond linking the rings B and C (Figure 3.5). Throughout this study, all crystal structures from the pyrrolidine carboxamides will be referred with their updated codes. Furthermore, the structural aberrations of the bound ligands in the obsolete PDB codes were resolved prior to the docking stages (Section 3.2.3.4).

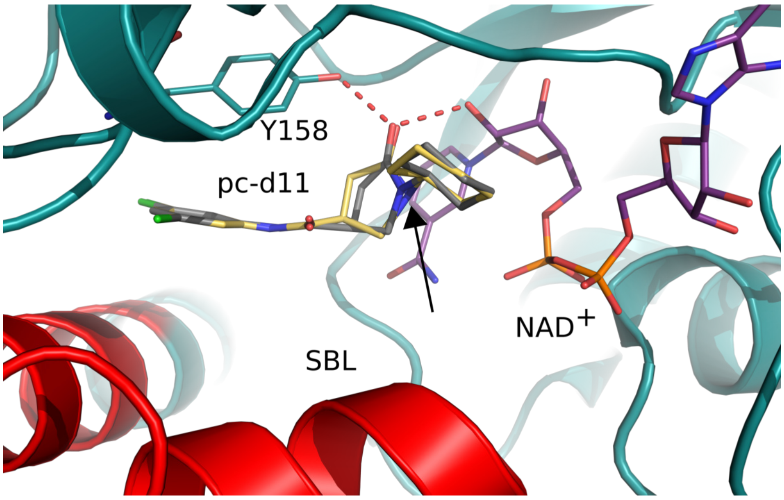


Figure 3.5 The C-α aligned structures of old (2H7M) and new (4TZK) representative crystal structures for pyrrolidine carboxamides. The ligand in yellow represents the old structure while the one in grey is the refined one. The arrow denotes the corrected amide bond, while the SBL has been colored in red.

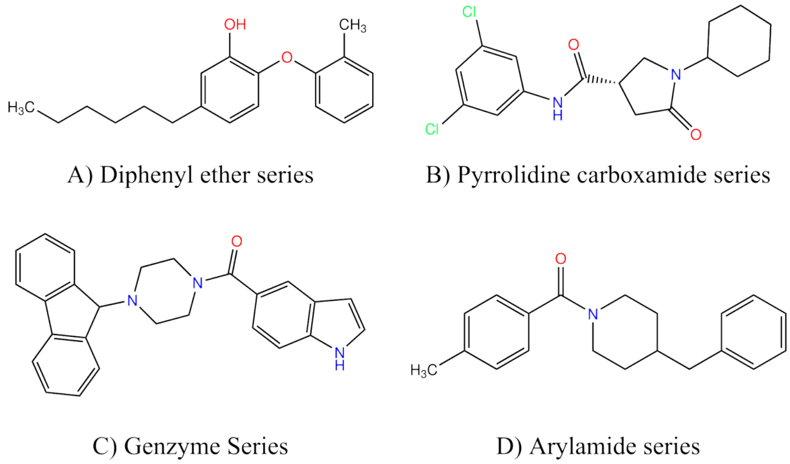
Upon comparing the crystal structures of diphenyl ethers with PDB 4TZK, noticeable differences were observed in case of the SBL residues namely, M199, I202, V203 and L207 (Table 3.2). This merely signifies the different states of SBL in all these structures as compared to 4TZK. Since the PDB 2X23 represented the most complete and well resolved SBL, it was chosen as a reference structure for diphenyl ethers.

Table 3.1 Old and new PDB codes for the pyrrolidine carboxamides: The overall C-α RMSD difference in between old and new crystal structures of pyrrolidine carboxamides. The RMSD values were obtained by aligning the old and new structures based on the least squares fit of the C-α atoms of the protein.

Old-PDB	New-PDB	RMSD difference (Å)	Resolution (Å)
2H7I	4U0J	0.05	1.62
2H7L	4TRJ	0.06	1.73
2H7M	4TZK	0.05	1.62
2H7N	4U0K	0.06	1.90
2H7P	4TZT	0.08	1.86

3.2.2 Docking of InhA inhibitors - The dataset

A total of 113 molecules spanning four different chemical classes of Mtb-InhA inhibitors was obtained from literature [50] (cf. Figure 3.6 and Appendix A). The dataset comprises 15 molecules from Genzyme Inc., 25 diphenyl ethers, 24 arylamides and 49 pyrrolidine carboxamides [50], with activity being reported in terms of IC_{50}. For some diphenyl ethers, the inhibition constant (K_i) was also reported [16, 72, 74, 258]. The actual number of molecules in each dataset is much higher than the one used for docking, because molecules whose activity was not determined or having no IC_{50} values were simply excluded. The four chemical classes were chosen because of well described binding modes corresponding to a total of 12 PDB entries (cf. Figure 3.7), with the SBL in different states (open to nearly closed, Figure 3.7). The activities (pIC_{50}) of the molecules ranged from 9.69 to 4.13 (0.2 nM to 73.58 μM) that nearly traversed 6 orders of magnitude. Since the primary aim of the study was to obtain reasonable binding orientations for molecules with unknown binding modes, the entire dataset was considered for docking initially. In the latter stages, a pIC_{50} based separation model as well as ROC analysis [228, 259, 260] of the docking protocol was performed only for the pyrrolidine carboxamide subset.



A) Diphenyl ether series

B) Pyrrolidine carboxamide series

C) Genzyme Series

D) Arylamide series

Figure 3.6 Representative molecules of principal chemical series of InhA inhibitors utilized in molecular docking.

Table 3.2 C-α RMSD for binding site residues (chain A only) from different crystal structures of InhA; residues with RMSD > 3 Å have been marked in bold. The RMSD values were obtained by aligning the A chains of all protein structures with the chain A of PDB 4TZK. The protein alignment was based on the least squares fit of their C-α atoms.

Residues	RMSD vs. 4TZK (Å)														
	2H7M	2H7I	2H7L	2H7N	2H7P	1P44	2NSD	2X23	3FNE	3FNF	3FNG	3FNH	4TRJ	4U0J	4U0K
Ile21	0.05	0.08	0.11	0.09	0.11	0.20	0.16	0.26	0.16	0.18	0.29	0.38	0.07	0.06	0.08
Ile95	0.09	0.09	0.10	0.12	0.10	0.14	0.14	0.15	1.62	1.70	1.75	1.88	0.03	0.03	0.05
Gly96	0.04	0.09	0.07	0.09	0.07	0.11	0.10	0.18	0.09	0.16	0.19	0.48	0.06	0.06	0.06
Phe97	0.09	0.16	0.13	0.18	0.13	0.48	0.16	0.22	0.12	0.27	0.45	0.30	0.10	0.14	0.10
Met98	0.08	0.09	0.11	0.09	0.11	0.22	0.18	0.12	1.15	0.21	1.17	0.99	0.07	0.06	0.11
Pro99	0.05	0.08	0.06	0.07	0.06	0.16	0.36	0.19	0.13	0.12	0.29	0.46	0.06	0.09	0.13
Met103	0.16	0.38	0.20	0.32	0.20	0.61	1.44	0.47	0.55	0.51	0.36	0.79	0.14	0.34	0.29
Asp148	0.09	0.13	0.05	0.20	1.10	0.08	0.16	0.10	0.13	0.18	0.13	0.17	1.09	0.06	0.15
Phe149	0.07	0.08	0.10	0.32	0.10	0.32	0.23	0.18	1.46	1.46	1.46	1.47	0.06	0.06	0.18
Met155	0.14	0.41	0.25	0.28	0.25	0.40	0.50	0.27	0.62	0.77	0.60	0.87	0.26	0.41	0.11
Pro156	0.18	0.22	0.18	0.18	0.18	0.46	0.20	0.42	0.38	0.40	0.33	0.56	0.12	0.16	0.13
Tyr158	0.10	0.16	0.14	0.25	0.14	0.42	0.22	0.57	1.42	1.43	1.41	1.43	0.06	0.13	0.22
Lys165	0.05	0.08	0.06	0.05	0.06	0.28	0.10	0.10	0.23	0.24	0.11	0.17	0.05	0.07	0.06
Thr196	0.14	0.15	0.15	0.24	0.15	0.29	1.01	0.96	1.50	1.63	1.84	1.82	0.12	0.14	0.18
Leu197	0.07	0.22	0.21	0.26	0.13	0.74	1.60	**3.16**	2.32	2.63	**3.46**	3.39	0.16	0.21	0.30
Ala198	0.10	0.24	0.20	0.36	0.20	1.31	2.27	**3.73**	**3.47**	**3.63**	**3.61**	**3.59**	0.15	0.21	0.39
Met199	0.13	0.30	0.18	0.46	0.18	1.52	2.43	2.15	**3.56**	**3.05**	**3.27**	2.50	0.19	0.26	0.48
Ile202	0.75	0.16	0.14	0.75	0.14	2.44	2.92	**4.65**	**3.82**	**3.97**	**3.36**	**3.38**	0.10	0.16	0.19
Val203	0.12	0.54	0.19	0.42	0.19	2.83	**3.38**	**6.15**	**3.66**	**3.78**	2.99	**3.46**	0.15	0.52	0.46
Leu207	0.12	0.21	0.15	0.22	0.15	**3.57**	**5.75**	**3.50**	**9.89**	**9.72**	**9.12**	**8.44**	0.09	0.19	0.14
Ile215	0.13	0.27	0.30	0.33	0.30	1.36	2.80	2.48	1.87	1.95	1.57	2.44	0.20	0.27	0.38
Leu218	0.11	0.09	0.37	0.17	0.37	0.96	2.47	1.12	1.69	2.31	1.23	2.18	0.29	0.61	0.26

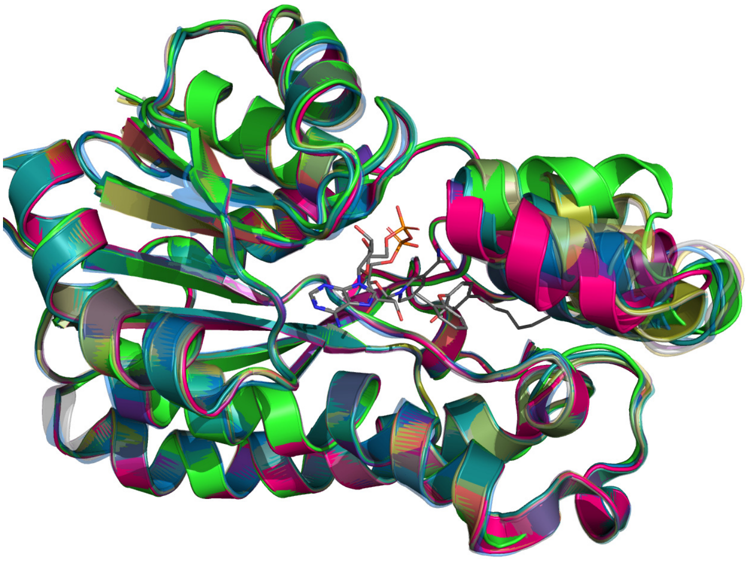


Figure 3.7 Globally aligned PDB structures of *Mycobacterium tuberculosis* InhA (N=16) covering the four chemical classes of InhA inhibitors [cf. Figure 3.6], with the SBL in various states, ranging from open (1P44, green) to nearly closed (2X23, pink).

3.2.2.1 Choice of pyrrolidine carboxamides for docking

The pyrrolidine carboxamides with a rather modest potency represent an activity range that spans only somewhat more than two orders of magnitude (pIC_{50} range 4.13 to 6.85). The following reasons led to the choice of the pyrrolidine carboxamide series to be considered for docking and further assessments:

1. The pyrrolidine carboxamides represent a large number of compounds with uniformly measured activity values (i.e. using the same assay conditions).

2. The crystal structures for pyrrolidine carboxamides represent only moderate to weakly active compounds. The potent compounds of this compound series have unknown binding modes, mainly due to lack of crystal structures. Hence, one of the main aims of this study included the development of a protocol for binding mode prediction of pyrrolidine carboxamides with unknown binding modes.

3. The narrow activity range of this series represents a challenging test case for adjudging the performance of docking and binding affinity prediction.

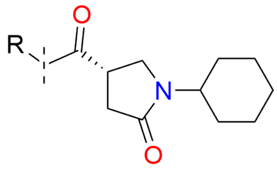
4. The docking poses served as the basis for a binding affinity prediction model using the Linear interaction energy (LIE) method. For any model derivation, an adequate sample size of reasonable quality is essential. The pyrrolidine carboxamide series satisfies both of these criteria with a total of 49 ligands whose activity was measured in uniform way under similar conditions. From the total 49 molecules, 44 were simulated and assessed (cf. Chapter 4).

5. The determinants of slow-onset binding for diphenyl ethers and related compounds have been recently elucidated [74, 261]. Comparatively, the molecular determinants governing the apparent ”rapid reversible” binding of pyrrolidine carboxamides is not clearly known. The only information available is the contribution of the several residues to the overall binding energy by MM-PBSA [262]. Elucidation of the conformational dynamics for these residues can aid in revealing the process of binding of pyrrolidine carboxamides to InhA. A comparative evaluation of the pyrrolidine carboxamides and diphenyl ethers with respect to the conformational dynamics of the molecular determinants can aid in structure based optimisation of putative Mtb-InhA inhibitors. This forms the basis of part II of this thesis.

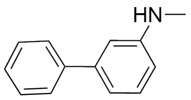
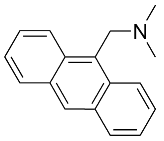
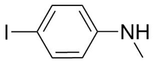
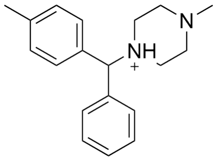
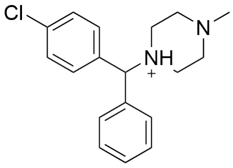
3.2.2.2 The pyrrolidine carboxamide dataset

A mentioned earlier, the pyrrolidine carboxamides represent a suitable class for a structure-based optimisation endeavour. The subset of 49 pyrrolidine carboxamides was initially classified into **”light”** (N=29) and **”bulky”** (N=20), depending upon the nature of the A ring and its substituents. The binding modes for the light pyrrolidine carboxamides has been elucidated in the form of 5 crystals structures [52] (Table 3.1). Comparatively, none of the ”bulky” pyrrolidine carboxamides got crystallised with the InhA-NAD$^+$ adduct and hence the lack of their binding modes. Another important aspect pertaining to bulky pyrrolidine carboxamides is that noticeable number of ligands exhibit marked potency against InhA that can be seen from their IC$_{50}$ values (Table A.4). This dataset of 49 compounds was subjected to some refinements prior to affinity prediction which led to a total of 46 compounds. These molecules could be utilised for binding affinity prediction model generation and other assessments mentioned in Chapter 4. Additionally, from these 46, two molecules (pc-p27 and pc-p37) were further omitted that led to a final pyrrolidine carboxamide dataset of 44 compounds (Tables A.2 to A.4). The details of the excluded molecules and the reasons for their exclusion are as follows:

Table 3.3 Molecules excluded from MD simulations and subsequent binding affinity calculations due to various reasons.



Core pyrrolidine carboxamide scaffold

Compound	**R**	Reason for exclusion
pc-p24, pc-c6a1		Duplication
pc-p21, pc-c6a2		Duplication
pc-s9		Parameterisation of iodine
pc-p27		Structural issues leading to MD simulations failure
pc-p37		Structural issues leading to MD simulations failure

1. There were two instances wherein bulky pyrrolidine carboxamides with identical structures (and activities) were assigned different identification codes (Table 3.3). The removal of duplicate molecules brought down the number of molecules from 49 to 47.

2. One molecule from the light pyrrolidine carboxamide subset (pc-s9, Table 3.3) had an iodine substituent on the A ring that prevented its proper QM parameterisation using the conventional Hartree-Fock 6-31G* method. The lack of proper parameters hampered its MD simulations in Amber [263] and subsequent affinity prediction using the LIE method. The number of molecules in the dataset now stood at 46.

3. Finally, in case of two bulky pyrrolidine carboxamides, pc-p27 and pc-p37, reasonable binding modes could be obtained via docking. However, in the ensuing

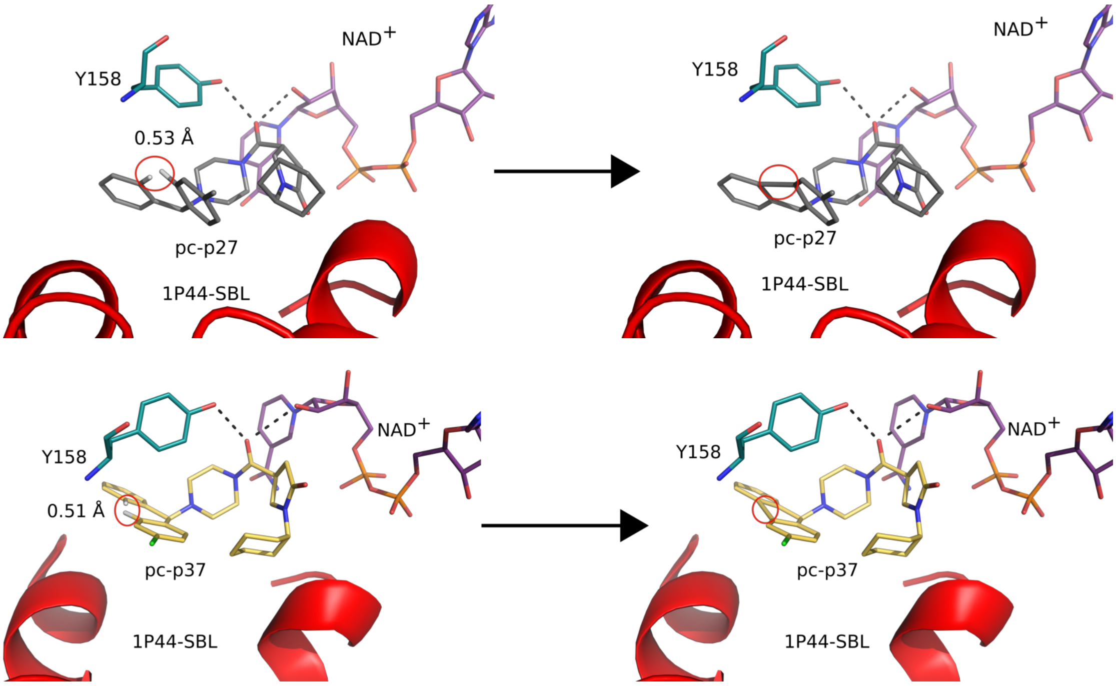


Figure 3.8 Intramolecular close contacts in pc-p27 and pc-p37: The close contacts of the rings from the docked pose of pc-p27 (left hand side top picture) and pc-p37 (left hand side bottom picture) that leads to the MD simulation program linking the two rings together (right hand pictures; both top and bottom). The parameterisation of the aberrant bond (red circles) was not carried out leading to stability issues during MD simulations.

molecular dynamics simulations, both molecules failed because of close intramolecular contacts. In both cases, the hydrogens of the phenyl rings that together constituted the A ring system were too close. As a result, the system linked both of the rings together with a bond that was not covered in the parameterisation (cf. Figure 3.8). To sidestep this issue, prior to parameterisation of the ligand, a short minimisation was carried out in MOE. Inspite of this, the wrong bonding issue persisted. Hence, both of these molecules were excluded from MD simulations.

Thus, the number of pyrrolidine carboxamides came down to 44 which was the final number of molecules from this subset to be considered for analysis of docking as well as MD simulations. Accordingly, the final number of molecules in the Mtb-InhA inhibitor dataset decreased by 5 to 108.

3.2.3 System Preparation

3.2.3.1 Protein Preparation

All protein structures were initially loaded into Schrödinger Maestro 9.3 [264] and prepared using the Protein preparation wizard (Prepwiz) using the default settings. During the protein preparation, water molecules that were > 3 Å away from the ligand

and which exhibited ≤ 2 H-bonds were removed, while the terminal amino acids were capped and missing side chains were added using Prime. The protonation states for polar amino acid residues like histidine, lysine etc. were automatically assigned using PROPKA 3 [265, 266] at pH of 7.0.

Thereafter, a structural alignment was performed with PDB 2X23 (chain A) as a global reference system, followed by saving the translated crystal structures as individual PDB files. In case of the PDB 2NSD, the oxygen atoms of the central phosphate group in the cofactor (NAD^+) were found to exhibit an unusual conformation (cf. Figure 3.9). The aberration was solved by initial protonation of the protein-ligand complex with Protonate3D of MOE 2012.10 [257]. Thereafter, the protonated complex was loaded in Schrödinger Maestro and the oxygen atoms were fixed using the build module of the Prepwiz followed by a restrained minimization using Macromodel 9.9 [267] employing the OPLS2005 force field [195, 245, 268], and the default settings.

During the minimization, the "fixed" oxygen atoms were free to move, while the remainder of the system remained frozen. It was seen that the restrained minimization of the phosphate group solved the aberration, with a flip of the oxygens away from each other consistent with electrostatic repulsions in between the two. Thereafter, the protein-ligand complex was loaded in the workspace for the protein structure alignment step mentioned earlier. In all cases, the prepared protein system was loaded separately in Schrödinger Maestro and subsequently utilized for grid generation in the ensuing step.

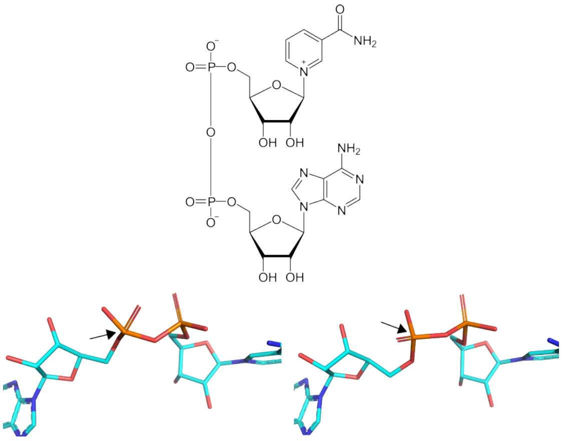


Figure 3.9 Perspective of odd and corrected structures of NAD^+, with the flip of phosphate oxygens being denoted by arrows.

3.2.3.2 Grid Generation

The prepared protein-ligand complex was loaded in Schrödinger Maestro followed by a grid generation using the *Grid generation wizard* of Schrödinger Maestro 9.3. The docking region or "grid" was defined as a cube of 30 Å region from the centroid of the bound crystal structure ligand. Furthermore, the hydrogen atoms from the hydroxyl groups of the ribose ring of the cofactor and Y158 were kept flexible, while remaining hydroxyl groups of the protein and cofactor remained frozen. The resultant grid files (total 4; one for each of the four representative structures) were then utilized subsequently for the molecular docking studies. The resultant grid files in a compressed (.zip) format contain critical information pertaining to the shape and properties of the receptor. This information is depicted on the grid by several fields that progressively enable efficient scoring of the poses [134].

3.2.3.3 Conformer generation

Glide uses a systematic and exhaustive conformational search of the ligand to yield the most probable binding modes within the binding site. This conformational search is sped up with help of a probing screen that weeds out high energy (unfavourable) conformations, e.g., cis amide conformation. The conformer generation protocol of Glide divides the ligand into a core (conserved) region and a suitable number of rotatable groups termed *"rotamers"*. The rotamers are rigid groups attached to the core region by a rotatable bond. During the conformer generation, a certain number of conformations (up to 500) represent the core region to which the rotamers are attached. The individual conformations of every rotamer group are annotated to enable scoring in latter stages. The combination of core region and all conformations of the rotamers are then docked in the receptor region followed by the steps described in Figure 3.2. In the current case, the conformer generation takes place in two stages, the first one during the ligand preparation with LigPrep and the latter during the actual docking as described earlier.

3.2.3.4 Ligand Preparation

Initially, the entire dataset excluding the crystal structure ligands was sketched in Sybyl-X 1.2 [269], with the MMFF94 partial atomic charges being assigned to all atoms. The individual members of the dataset were then minimized for 1000 steps to a RMSG convergence criterion of 0.05 kcal/mol·Å with the default settings. The molecules were then saved to a multi-mol2 file. Subsequently, the dataset was loaded into Schrödinger Maestro 9.3 [264] followed by ligand preparation being performed using LigPrep 2.5 [270]. Using the default settings, various possible states (tautomers and stereoisomers alike) were generated at a pH range of 7 ± 2. From these various states, a total of **5 low energy ring conformations per molecule** were exported in the final output (maegz format) and used for subsequent molecular docking. The use of 5 low energy ring conformations is expected to account for small minima around the lowest energy conformer that can be attained prior to the conformational selection via docking.

During the ligand preparation phase, another aberration was observed in crystal structure ligands of pyrrolidine carboxamides and arylamides, wherein all bound ligands exhibited a non-planar configuration for the amide bond (bent) (cf. Figure 3.10). The amide bonds were subsequently corrected according to a procedure similar to that used for correcting the cofactor of 2NSD. A point to be noted is that the updated crystal structures of pyrrolidine carboxamides do not exhibit any bent amide bond for the bound ligands (Table 3.1).

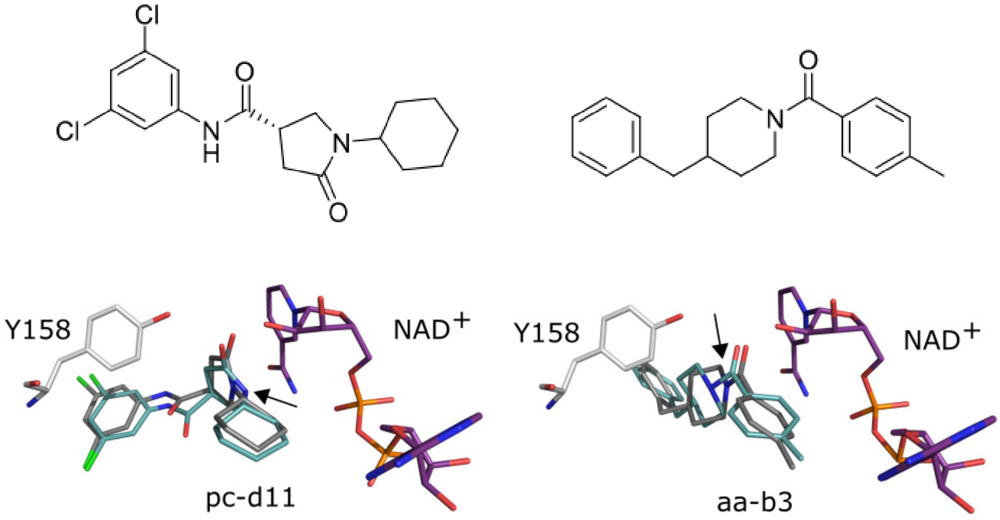


Figure 3.10 2D structures of ligands from PDB 2H7M (now 4TZK) and 2NSD; structures of the same ligands before and after minimization in Macromodel, with arrow denoting corrected amide bonds.

3.2.4 Docking approaches

In order to determine the probable binding orientations of putative InhA inhibitors with no crystal structures available, two main approaches were followed:

- **Single precision docking:**
  The initial step in docking with Glide SP mode consisted of a simple redocking and cross-docking of the crystal structure ligands across all four representative structures. The basics of the docking with SP can be read in Section 3.1.1. The redocking/cross-docking with Glide SP involved extraction of the ligands from the crystal structures followed by their preparation in LigPrep and docking the LigPrep output (.maegz) in respective structures using the default settings. During the docking, the hydroxyl groups of the ribose of the cofactor (NAD$^+$) and Y158 were flagged as rotatable to enable non-covalent bonding with the ligand. The number

of poses per input conformation was set to 5 in order to enhance the likelihood of observing the native binding mode (global minimum). The higher number of poses per ligand was also expected to aid in ascertaining the native pose within a particular binding orientation. This stage was followed by RMSD assessment of the top scoring pose with respect to the crystal structure ligands. The protocol was then subsequently utilised for the entire InhA inhibitor dataset, with the pose selection protocol (*GDRPS*) being utilised in post-docking stage that facilitated the thorough assessment of the placement algorithm in regards of correct placement of the key scaffold. Subsequently, the pyrrolidine carboxamide dataset was subjected to docking across all four proteins using the Extra precision mode (XP) employed in Glide.

- **Extra precision docking:**
  The docking with GlideSP enabled the assessment of the ability of the placement algorithm in correctly placing the key scaffold in the binding pocket. However, as seen from Section 3.1.1, the main aim and purpose of GlideSP is to provide for an initial placement of the ligand and not pose enrichment or ranking. Hence, with an aim of achieving better pose enrichment (quality) the pyrrolidine carboxamide dataset was docked across all four representative proteins. The protocol for docking using GlideXP remained identical to that of GlideSP, i.e, using default settings and number of poses per conformation of ligand set to 5. In the post-docking phase, the pose selection was performed with both *GDRPS* and *IHPS* protocols.

- **Induced fit docking**
  With an aim to incorporate receptor/protein flexibility during docking, the pyrrolidine carboxamide dataset was also docked into each of the four proteins using the *"Induced fit workflow"* [117] implemented in Schrödinger Maestro. The basics of the methodology have been explained earlier in Section 3.1.2. The induced fit workflow provides the user flexible options to define the docking region, with the grids generated in earlier steps being used for defining the docking region. Alternatively, the user can load the prepared and translated crystal structures from earlier steps and simply define the docking region as a **20 Å** region from the centroid of the selected residue, which is usually the bound ligand. In the current case there are two main approaches to incorporate receptor flexibility, namely plain Induced fit abbreviated as **IFD** and Induced fit with trimmed side chains abbreviated as **IFD-trim**.

  1. **Normal Induced fit docking (IFD):**
     For plain induced fit docking, the pre-prepared grids were simply loaded in Schrödinger Maestro followed by launching the Induced fit workflow. The LigPrep output for the pyrrolidine carboxamide dataset was selected as an

input for docking. The ligands were set to be docked according to default settings (rigid docking with no sampling of ring conformations for the “soft docking” followed by redocking with GlideSP). The amide bond for receptor as well as ligand was set to planar and the planarity of the conjugated systems being considered on an enhanced scale to better monitor $\pi - \pi$ stacking interactions. The reason behind choosing rigid docking of the ligands in the “soft docking” phase is that their conformations get sampled on-the-fly during the redocking phase thereby sidestepping the need to generate conformations prior to the soft docking. In the post-soft docking phases, a total of three key residues namely Phe149, Tyr158 and the cofactor (NAD$^+$) were not subjected to extensive Prime refinement during the step of receptor conformational sampling. The prime reason was that the pose selection procedures employing pharmacophore filtering mandated maintenance of fixed distances in between the two H-bond donor groups, that would be impossible to satisfy had the Prime refinement for the same been turned on. Furthermore, the terms in Glide redocking were also set to default values, since the main aim of the IFD protocol was to assess the performance of the placement algorithm in correctly sampling and reporting the poses with correct orientation within the binding pocket. In the post-docking phase, the poses that were ranked by the IFDscore were subjected to both GDRPS and IHPS protocol together with rescoring the selected poses ”in-place” with the strict GlideXP scoring function. The entire procedure was scripted with a bash script to ensure correct step-wise pose selection and ranking. Manual intervention was only needed after the conclusion of docking to submit the docking output to the script as a multi-mol2 file.

2. **Induced fit docking with Trimmed side chains (IFD-trim):**
   The IFD-trim procedure differs from the IFD in the sense that in addition to mutating the key active site residues to alanine in the ”soft docking” phase, the procedure allows ”trimming” or removal of side chains of those amino acids that hinder proper placement of the ligand. The procedure for performing IFD-trim is very much similar to IFD except that the user has to select the ”Trim side chains” from the “Prime refinement” tab depending upon their B-factors or by manually specifying a list of residues. The residues to be ”trimmed” for the current case included those located at a distance of 5 Å from ligand. In short, the docking with IFD-trim consists of the following steps:

   – Loading of the grid/s in Schrödinger maestro followed by selecting the induced fit workflow

- Selecting the prepared pyrrolidine carboxamide dataset as an input for docking
- Setting default options for sampling of planar ring conformations, enforcing planarity of amide bonds, activating enhanced sampling of conjugated systems
- Excluding Phe149, Tyr158 and cofactor (NAD^+) from Prime refinement and providing a list of residues that lie within 5 Å from the crystal structure ligand
- Keeping default options for redocking the ligand in minimised protein after the Prime minimisation and conformational sampling step
- Once the docking is over, perform pose selection and "in-place" rescoring of selected poses with GlideXP scoring function.

3.2.5 Post-docking analysis

Once the docking was concluded, in addition to the substructure RMSD and rescoring performed during pose selection, the pose quality (enrichment) was additionally assessed using receiver operating characteristic (ROC) analysis [228, 271, 272] and visualization (cf. Figure 3.11 and equation (3.4)). The ROC analysis is a graph based technique that aids in visualization, organization and overall performance of the classifiers. In the current study, the scoring functions serve as a classification (prediction) means that predict whether the given pose (instance) is active or inactive based on their score. In conceptual terms, a classifier covers matching of instances to their predictions [271], that yield four possible outcomes given for a instance and classifier, namely:

1. **True Positive:** When the prediction and instance are both positive.
2. **True Negative:** When the prediction and instance are both negative.
3. **False Negative:** When the prediction is negative and instance is positive.
4. **False Positive:** When the prediction is positive and instance is negative.

		Actual Activity	
		t	**f**
Predicted Activity	**Y**	True Positives	False Positives
	N	False Negatives	True Negatives
Total columns:		**T**	**N**

Figure 3.11 ROC basics: A confusion matrix denoting common performance criteria calculated from it, T refers to total number of molecules accurately identified; N refers to the number of molecules incorrectly identified, t and f represent actual actives and inactives; Y and N refers to activity predictions of the model as active or inactive. Figure adapted and redrawn from Fawcett, 2006 [271]

$$
\begin{aligned}
False\ Positive\ Rate\ (FPR) =& \frac{Incorrectly\ classified\ negatives}{Total\ negatives} = \frac{FP}{N} \\
True\ Positive\ Rate(TPR) =& \frac{Correctly\ classified\ positives}{Total\ positives} = \frac{TP}{P} \\
=& \ recall\ rate\ or\ hitrate = Sensitivity \\
Specificity =& \ \frac{True\ negatives}{False\ positives + True\ negatives} = 1 - fp\ rate \\
Precision =& \ \frac{True\ positives}{False\ positives + True\ positives} \\
=& \ Positive\ predictive\ value \\
Accuracy =& \ \frac{True\ positives + True\ negatives}{Total\ positives + Total\ negatives} \\
F - measure =& \ \frac{2}{\frac{1}{precision} + \frac{1}{recall}}
\end{aligned}
\tag{3.4}
$$

A ROC curve is a simple plot of TPR versus the FPR (cf. Figure 3.12) depicting the relative tradeoff's between the true positives and false positives. The point (0,0) denotes the origin which signifies no classification of data, since there are no true or false positives being predicted. The point (1,1) reflects unconditional prediction of only true positives, while the point (0,1) is an ideal classification. Ideally, the more the ROC curve extends to the north-west corner of the plot, the better the classification, while the more it goes towards the south-west of the plot, worse is the classification. A key identifier signifying the quality of the ROC curve is the area that it covers, i.e. the *Area under the curve* (AUC). The value of AUC ranges from 0 to 1 (modified from 0 to 100%) [228]. The more the value of AUC approaches 1.0, the better the classification, with a value of 0.5 (along the diagonal) signifying a random model.

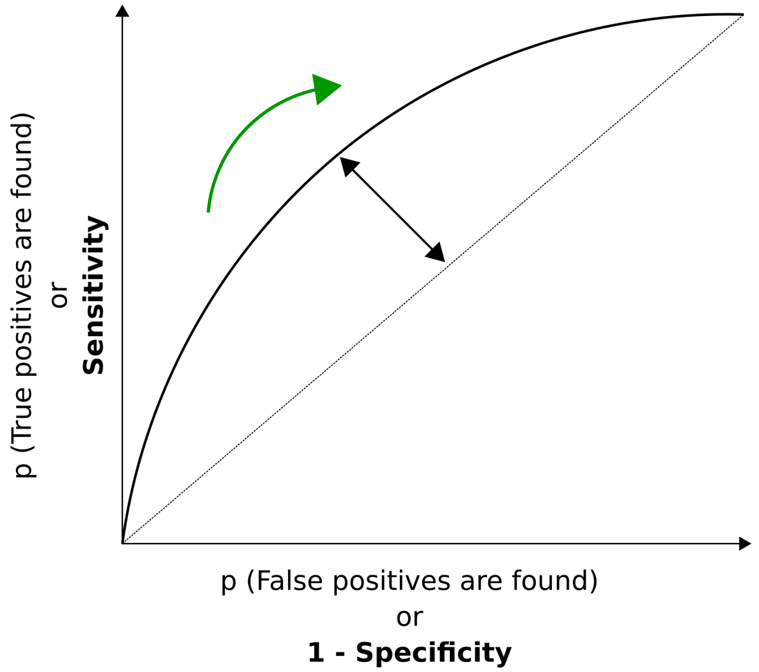


Figure 3.12 A simple ROC curve depicting the sensitivity (TPR) and specificity (FPR/ 1-Specificity). The quality of the ROC (thick line) as measured by the Area under the curve (AUC) increases as it moves more away from the diagonal towards the Y axis (increase in length of the bidirectional arrow). On the contrary, a ROC curve that is diagonal or below it (towards the X axis) implies poor predictive power of the model being represented by the ROC curve.

In the current study, the ROC analysis was performed in R [273] using the ROCR [274], pROC [259] and enrichvs [275] packages. The substructure RMSD calculations were performed using fconv [256], while pose visualization in post-docking phase as well as figures in results section have been obtained from Pymol 1.8 [276]. All plots were traced using the ggplot2 package [277] of R.

3.3 Results

3.3.1 Preliminary redocking

In the initial stages of the docking, i.e. in the redocking phase, the selected pose was by default the top scoring pose. Since GlideSP is adept at identifying the probable binding modes, the docking was expected to reproduce the binding modes of all representative ligands from the crystal structures. Indeed, the substructure RMSD for the top ranked pose of the redocked ligand with respect to the crystal structure ligand was found to be within the permissible limits (Table 3.4). In Table 3.4, the comparatively higher RMSD value for PT-70 is mainly due to the flexible alkyl chain on the A ring (Table A.8).

For analysing the preliminary redocking and cross-docking of crystal structure ligands, the RMSD value as calculated from fconv took priority over the docking score (GScore) in ascertaining the performance of the docking algorithm. The Tables 3.4 and 3.5 depict the score and RMSD of the top poses emanating from the redocking and cross-docking of the crystal structure ligands.

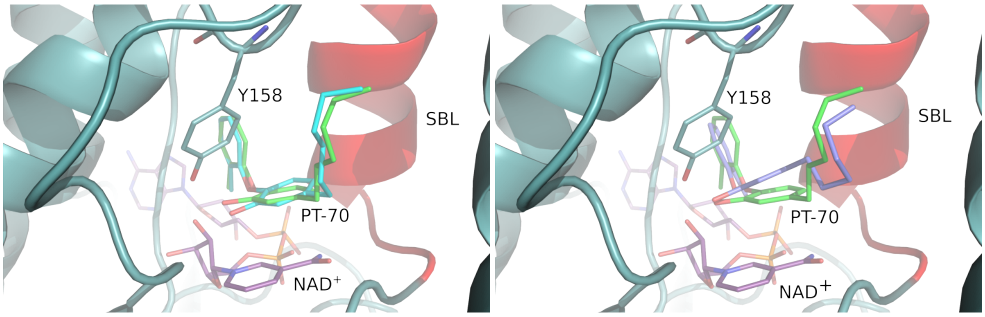
Table 3.4 Heavy atom RMSD and docking score (GlideSP) of top scoring poses for **redocked** crystal structure ligands.

Sr. No.	Protein	RMSD Å				Gscore (kcal/mol)
		Gz-10850	pc-d11	aa-b3	PT-70	
1	1P44	0.24	-	-	-	-9.60
2	4TZK	-	0.50	-	-	-9.51
3	2NSD	-	-	0.10	-	-10.85
4	2X23	-	-	-	0.97	-7.55

From Tables 3.4 and 3.5 and figure 3.13, it is clear that the placement algorithm of GlideSP works reasonably well in pose reproduction for most of the ligands except the ligand of PDB 1P44 (Gz-10850). The size of the ligand hampered its proper placement and scoring in the narrow active site of 2X23, resulting in no docking pose appearing in the output. A similar case was also observed for Gz-10850 when docked in 4TZK. The scaffolds (substructure) for the Genzyme and arylamide series being quite similar resulted in Gz-10850 getting high scores in PDB 1P44 and 2NSD, respectively. Overall, the RMSD of the docked/cross-docked ligand did not exceed 1.5 Å, barring a few cases. This meant that the placement algorithm employed in Glide was able to reproduce the binding modes of the crystal structures in both instances of redocking and cross-docking. The InhA inhibitor dataset was subsequently docked across all four proteins (1P44, 4TZK, 2NSD, and 2X23).

Table 3.5 Heavy atom RMSD and docking score (GlideSP) of top scoring poses for **cross-docked** crystal structure ligands; Gz-10850 (1P44 ligand) owing to its size did not get docked in 4TZK and 2X23.

Protein	Gscore (kcal/mol)				RMSD (Å)			
	Gz-10850	pc-d11	aa-b3	PT-70	Gz-10850	pc-d11	aa-b3	PT-70
1P44	**-9.60**	-8.25	-8.08	-6.47	0.24	1.50	1.23	1.14
4TZK	-	**-9.51**	-6.40	-6.65	-	0.50	2.11	1.15
2NSD	-11.80	-9.10	**-10.85**	-8.96	0.20	1.34	0.10	1.30
2X23	-	-8.62	-9.01	**-7.55**	-	1.90	2.05	0.97



(a) Redocked pose of PT70 (blue; in 2X23) with respect to crystal structure ligand (green).

(b) Cross-docked pose of PT70 (violet; in 2NSD) with respect to crystal structure ligand (green).

Figure 3.13 Redocking and Cross-docking results of PDB 2X23 ligand PT70.

3.3.2 Docking with GlideSP

The docking of the InhA inhibitor dataset across the four representative proteins was performed with the default settings that included no post-docking minimisation. These settings (5 poses/input conformation) together with the use of ”soft” potentials, lack of strict penalties for poses with unusual bonds/torsions, and no post-docking minimization led to a large number of poses in the final output file. Ascertaining the most probable binding mode for each individual compound from amongst a variety of poses in the output thereby assumed critical importance. To this end, the pose selection procedures (cf. Figure 3.1 and section 3.1.3) were extensively used that yielded the putative binding modes for molecules in the InhA inhibitor dataset.

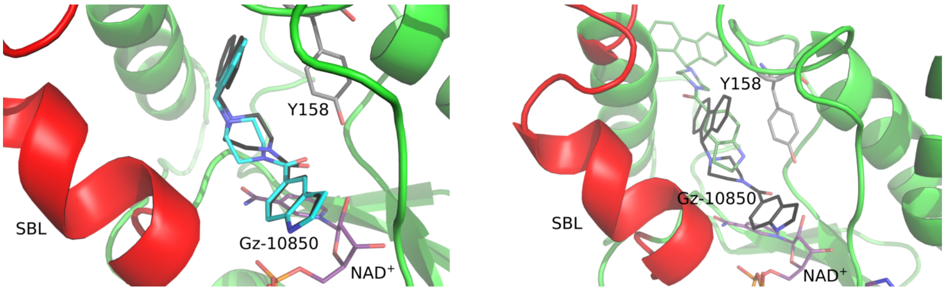
The GlideSP docking performed with different receptor conformations to incorporate receptor flexibility also resulted in a particular pattern of compound series towards a particular receptor. This can be seen in Table 3.6, which depicts the mean deviations of the combined top poses of a particular ligand series with respect to their native crystal structure ligand. From Table 3.6, it can be seen that bigger ligands prefer wide open pockets (e.g., of 1P44) while small sized ligands selectively get docked in tight binding pocket (e.g., of 2X23). Thus, structurally similar arylamides and Genzyme series prefer 1P44 and 2NSD and to a negligible extent 4TZK, while the Genzyme series seldom gets docked in 2X23. Comparatively, the small sized pyrrolidine carboxamides and diaryl ethers prefer the binding pockets of 4TZK and 2X23, respectively (cf. Figures 3.14 and 3.15).

The low substructure RMSD values for the poses selected by the *GDRPS* protocol indicate that GlideSP is able to reasonably predict the binding modes for majority of the compounds. This laid the ground for docking the InhA inhibitor dataset with the more stringent GlideXP docking. This is primarily because of two reasons, first the success of GlideSP in achieving reasonable scaffold placement automatically translates to a similar

Table 3.6 Substructure RMSD for combined top poses of each series in Glide SP protocol, * signifies RMSD was calculated for a single compound getting docked from amongst entire compound series; ** signifies no compound from the series was getting docked and selected.

Protein	GlideSP			
	RMSD (Å)			
	Genzyme (N=15)	Pyrrolidine carboxamides (N=44)	Arylamides (N=24)	Diphenyl ethers (N=25)
1P44	0.57 ± 0.17	2.99 ± 1.67	1.43 ± 0.23	2.05 ± 0.39
4TZK	7.40*	0.98 ± 0.86	2.23 ± 0.88	2.79 ± 2.18
2NSD	2.44 ± 2.69	3.40 ± 2.38	1.05 ± 0.22	1.34 ± 0.21
2X23	**	2.33 ± 0.20	1.85 ± 0.34	0.73 ± 0.10

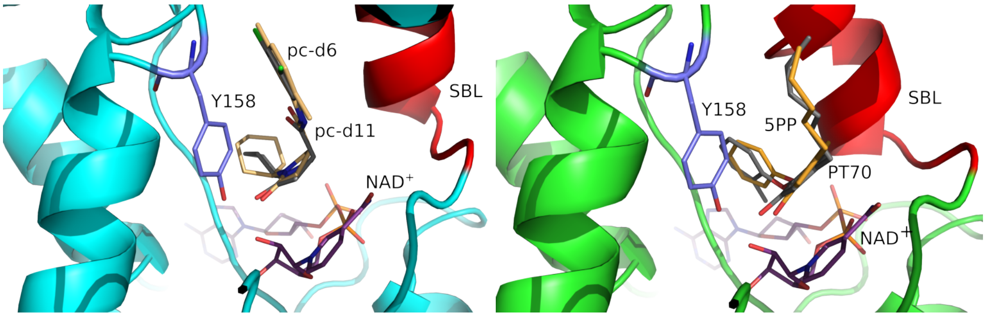
case for GlideXP. Second, GlideSP docking is adept at ”pose identification” and not ”pose discrimination” as opposed to GlideXP which is tailored for the latter. The results and analysis of the GlideXP docking is focussed on the pyrrolidine carboxamides.



(a) Redocked pose of Gz-10850 (blue; in 1P44) with respect to crystal structure ligand (Gz-10850, black) and SBL in red.

(b) Cross-docked pose of Gz-10850 (baggy green; in 2X23) with respect to crystal structure ligand (Gz-10850, black) and SBL in red.

Figure 3.14 GlideSP redocking and cross-docking results for Genzyme series.



(a) Docked pose of pc-d6 (wheat; in 4TZK) with respect to crystal structure ligand (pc-d11, grey), with 4TZK-SBL in red.

(b) Cross-docked pose of 5PP (orange; in 2X23) with respect to crystal structure ligand (PT70, grey) with 2X23-SBL in red.

Figure 3.15 GlideSP docking results for pyrrolidine carboxamides and diphenyl ethers.

3.3.3 Docking with GlideXP

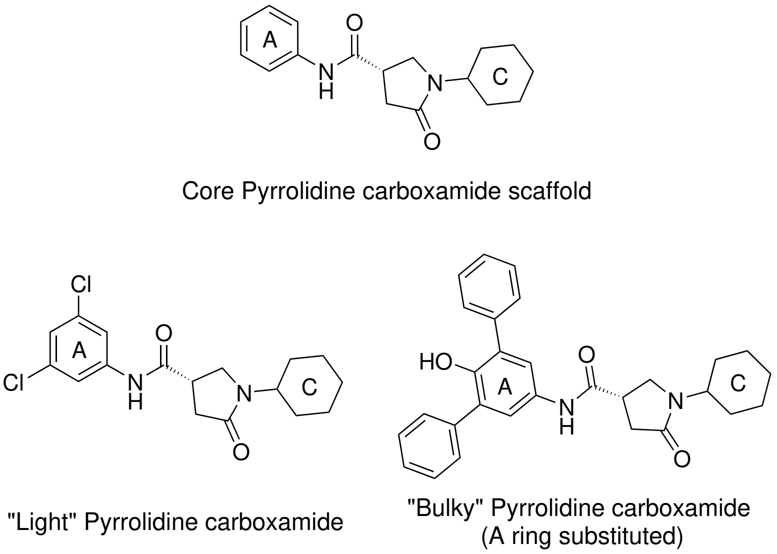
The docking with GlideXP was also performed with default settings in a way similar to GlideSP. The following are notable additions to the docking with GlideSP:

1. Extensive minimization of the poses in post-docking phase with Macromodel and OPLS2005 force field, followed by rescoring with GlideXP scoring function
2. Dual use of *GDRPS* and *IHPS* pose selection procedures to guide pose selection. In the current study, if GDRPS protocol did not yield any pose, the top ranking pose for the respective compound was chosen from the pool selected by *IHPS*, with the pose having lowest RMSD and highest DrugScoreX being selected.

The extensive post-docking minimisation and the stringent XP scoring function are the principal factors that differentiate GlideSP and GlideXP docking, respectively. This is evident in the small number of poses per ligand emerging from the XP docking and pose selection procedure (Table A.10). The small number of poses also comes from the strict filtering by the pharmacophore model that allows only a fraction of poses from the XP docking to pass through. Together the XP scoring function and the pose selection protocols were expected to deliver poses of high quality. The quality of poses can be ascertained by ROC analysis and its associated AUC value (Section 3.3.6).

In order to ease the evaluation of the poses of pyrrolidine carboxamides emerging from GlideXP docking, the pyrrolidine carboxamide dataset was split into two groups. Each of the subgroups was analysed individually, while the performance of the SFC scoring functions in an activity-based classification of the entire pyrrolidine carboxamide dataset was assessed by correlation and ROC analysis [228, 271, 272].

- **”Light” pyrrolidine carboxamides:** This group consists of those compounds with a mono/di-substituted Ring A (N=28, cf. Figure 3.16 and tables A.2 and A.3). Since the light pyrrolidine carboxamides vary slightly from the crystal structure ligands, their binding orientations are expected to be identical to that of the reference ligand (pc-d11, PDB 4TZK).
- **”Bulky” pyrrolidine carboxamides:** This group consists of 18 ligands characterised by replacements of either rings A and C by bicylic, heterocyclic or even multi-ring aromatics systems (cf. Figure 3.16 and table A.4). The name ”bulky” arises from the fact that all members of this series exhibit bulky A rings. As discussed earlier (Section 3.2.2.2), this group has no reference binding conformation due to lack of crystal structures. Additionally, many of these compounds are more potent by one order of magnitude as compared to light pyrrolidine carboxamides.



Core Pyrrolidine carboxamide scaffold

"Light" Pyrrolidine carboxamide

"Bulky" Pyrrolidine carboxamide
(A ring substituted)

Figure 3.16 Pyrrolidine carboxamide scaffold along with representative structures for ”light” and ”bulky” pyrrolidine carboxamides.

Hence, evaluating the performance of GlideXP in identification as well as correct ranking of poses for this dataset is worth investigating.

3.3.3.1 Substructure RMSD

The primary criterion of evaluating the performance of ligand placement in GlideXP is the substructure RMSD (cf. Table 3.7). When Table 3.6 and Table 3.7 are compared, the superiority of GlideXP in appropriate placement of the core scaffold becomes evident. The low values of substructure RMSD and accompanying standard deviation indicate that GlideXP reasonably predicted the binding modes of InhA inhibitors. This was accompanied with a much better scaffold placement as compared to GlideSP. Just like in case of GlideSP docking (cf. Section 3.3.2), the preference of a particular ligand series towards a protein with specific SBL conformation is clearly evident. Once the validity of GlideXP in reasonable placement of the scaffold was ascertained, the substructure RMSD analysis was performed in detail for the segregated pyrrolidine carboxamide subsets described in Section 3.3.3.

1. **”Light” Pyrrolidine carboxamides:** In case of the ”light” pyrrolidine carboxamides, the GlideXP was found to yield poses with reasonable substructure RMSD values (PDB 4TZK ligand as a reference) as obtained from fconv [256] (cf. Table 3.8 and figure 3.17). However, a small number of light pyrrolidine carboxamides (11/28, cf. Table 3.10) poses were deemed aberrant upon visual inspection. In almost all cases, the substituents on the A ring were pointing in the wrong direction when compared against the crystal structure ligand. Furthermore, in case of pc-d1, a

Table 3.7 Average substructure RMSD for combined top poses of each series in Glide XP protocol, The asterisk (*) signifies RMSD was calculated for a single compound getting docked from amongst entire compound series, ** signifies no compound from the entire ligand series was getting docked and selected.

Protein	GlideXP			
	RMSD (Å)			
	Genzyme (N=15)	Pyrrolidine carboxamides (N=44)	Arylamides (N=24)	Diphenyl ethers (N=25)
1P44	0.57 ± 0.27	3.75 ± 1.66	1.10 ± 0.39	1.32 ± 0.51
4TZK	**	0.77 ± 0.52	1.57 ± 0.39	0.76 ± 0.08
2NSD	4.44	2.67*	1.11 ± 0.18	1.26 ± 0.29
2X23	**	1.83 ± 0.43	1.59 ± 0.14	0.34 ± 0.19

flipped binding mode was observed that can be attributed to wrong placement of the ligand. On the contrary, such a flipped binding mode was not observed for any other light pyrrolidine carboxamide. In such a scenario, the aberrant conformations can be assumed because of:

- **Placement algorithm:** It is possible that the core scaffold was reasonably placed in the binding site. However, the conformational sampling of the ligand would have yielded a low energy conformation with significant deviations as compared to the orientation of the crystal structure ligand (pc-d11). This would manifest in form of clashes with the side chains of the active site and a huge penalty by the stringent XPscore. Such a pose would not show up in the final result at all, given the stringent nature of the XPscore and the pharmacophore model. Such aberrations were seldom observed in case of GlideSP docking. This indicates that the placement algorithm is not the source of the aberrant poses.
- **Extensive post-docking minimization and pose filtering:** The extensive post-docking minimisation appears to be a likely reason behind the aberrant conformations, primarily because the placement algorithm worked in GlideSP and up to a certain extent in GlideXP docking. Assuming the correct placement of the core scaffold, the incorrect orientation of the A ring substituents can be attributed to the rotation of the phenyl ring around the C-N bond (connecting the A ring to rest of the molecule) during the minimisation whilst leaving the key scaffold in its place. This gives rise to a low energy conformation which is quite identical to the orientation of the reference ligand as seen in the crystal structure with the exception of the A ring and its substituents. The said conformation passes through the pharmacophore model filtering and ends up getting the top score by both XPscore and DrugScoreX.

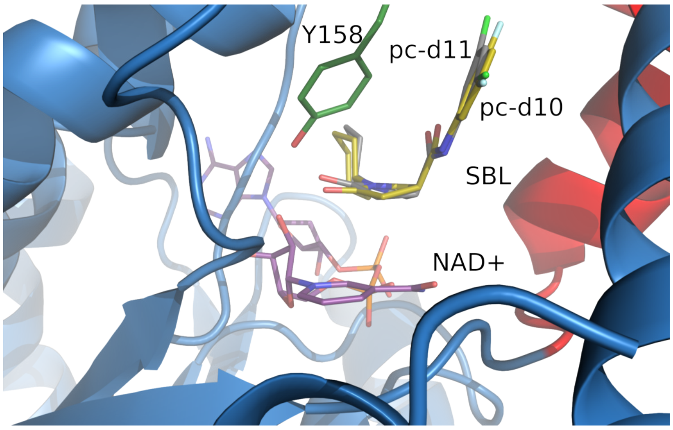


Figure 3.17 GlideXP docking results for pyrrolidine carboxamides: The top pose of pc-d10 (olive; in 4TZK) with respect to crystal structure ligand (pc-d11, grey); SBL depicted as red helix.

The molecules with aberrant conformations mostly had para substituents on the A ring, for example, in case of pc-d1 and pc-d9 (cf. Figure 3.18). The problem was solved with an approach explained in Section 3.3.4.

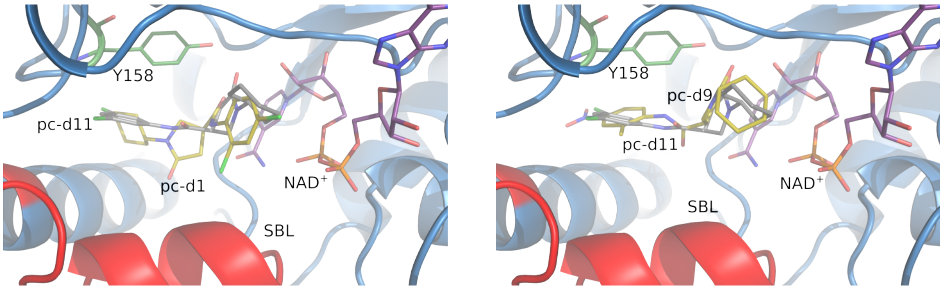


Figure 3.18 Aberrant poses of pyrrolidine carboxamides subjected to in-situ mutation: Left figure depicts docked pose of pyrrolidine carboxamide-d1 (olive; in 4TZK) with respect to crystal structure ligand pyrrolidine carboxamide-d11 (grey); Mutated pose of pyrrolidine carboxamide-d9 (olive; in 4TZK) with respect to crystal structure ligand pyrrolidine carboxamide-d11 (grey); SBL in red.

2. **"Bulky" Pyrrolidine carboxamides:** As seen from earlier examples, GlideXP encountered difficulties with delivering poses that conform to the orientation of the crystal structure. The post-docking minimisation performs its mandated job but at the cost of aberrant poses that demonstrate the rotation of the A ring. This suggested that the problem might be aggravated in case of the bigger sized "bulky" pyrrolidine carboxamides that will clash with the side chains of binding pocket residues in addition to the ring A flips described earlier. Indeed, when the poses of this subset were evaluated in terms of their substructure RMSD (cf. Table 3.9), it became amply evident that GlideXP docking worked reasonably with small sized ligands, but not the bigger "bulky" pyrrolidine carboxamides.

   As seen from Table 3.9, the preference of the big ligands towards the more open binding pocket (1P44) becomes evident, with a fraction favouring 2X23 and 4TZK,

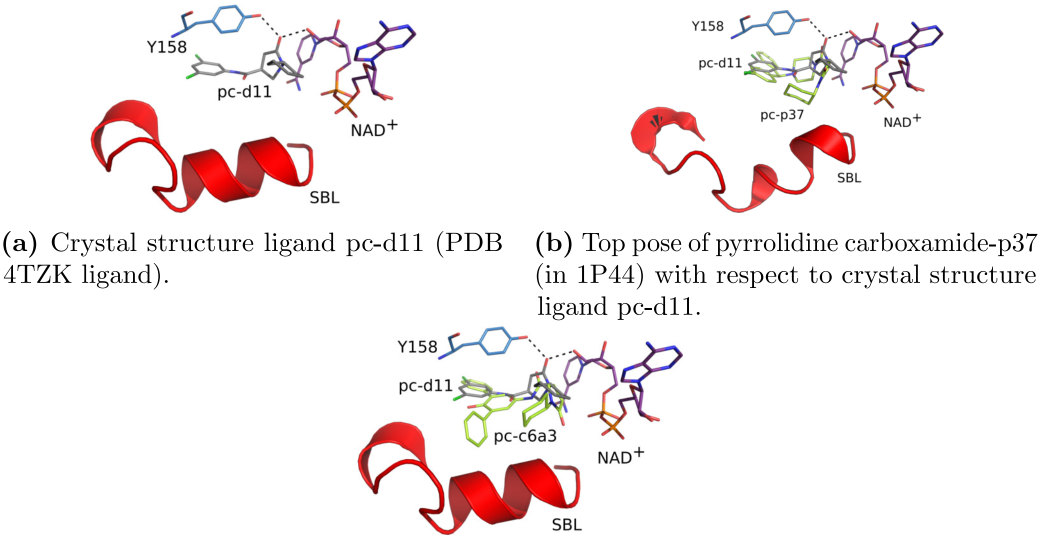
Table 3.8 Substructure RMSD and XPscore for top ranked poses of light pyrrolidine carboxamides (Tables A.2 and A.3) derived from GlideXP docking and GDRPS pose selection protocol across the four representative proteins 1P44, 4TZK, 2NSD, and 2X23; The aberrant poses appear in bold face.

Pyrrolidine carboxamide	**Pose selected from**	**Docking Score (kcal/mol)**	**Substructure RMSD (Å)**
s1	4TZK	-9.30	0.69
s2	4TZK	-9.29	0.77
s4	4TZK	-9.25	0.76
s5	4TZK	-8.77	0.91
s6	4TZK	-10.49	0.71
s10	4TZK	-10.15	0.74
s11	4TZK	-10.32	0.54
s12	4TZK	-9.22	0.74
s15	4TZK	-10.36	1.26
s17	4TZK	-9.43	0.87
d1	1P44	-9.37	7.26
d2	4TZK	-9.87	0.49
d3	4TZK	-9.91	0.42
d4	2X23	-7.45	1.29
d6	4TZK	-9.79	0.39
d7	4TZK	-9.89	0.58
d8	4TZT	-9.81	0.52
d9	4TZK	-8.04	1.39
d10	4TZK	-10.10	0.51
d11	4TZK	-10.09	0.37
d12	4TZK	-10.73	0.38
d13	4TZK	-9.90	0.65
d14	2X23	-8.59	1.24
d15	4TZK	-9.38	0.82
d16	4TZK	-9.96	0.81
3a	4TZK	-9.31	0.55
3i	4TZK	-9.27	0.71
3j	4TZK	-8.87	0.89

respectively. The reason behind the bulky pyrrolidine carboxamides getting docked only in 1P44 was that more space was available for the ligands during post-docking minimisation. This manifests in diverse binding modes (cf. Figure 3.19) and higher substructure RMSD primarily due to the availability of more space in the open binding pocket of 1P44. In other words, the approach of incorporating protein flexibility in docking by using different protein structures works reasonably for small sized ligands but not for large sized ligands. The solution to this problem was incorporating protein flexibility to an increased extent by sampling the protein conformation simultaneously during the ligand placement to achieve reasonable pose placement.

Table 3.9 Substructure RMSD's and protein from which top ranked poses for bulky pyrrolidine carboxamides were selected using GlideXP docking and combination of GDRPS and IHPS pose selection protocols.

Compound	Pose selected from	XPscore (kcal/mol)	Substructure RMSD (Å)
pc-r7	2X23	-8.17	2.00
pc-p9	1P44	-10.23	1.94
pc-p20	1P44	-12.01	1.89
pc-p21	2X23	-10.36	2.08
pc-p24	2X23	-10.49	2.20
pc-p28	1P44	-11.52	4.54
pc-p31	2X23	-9.87	2.19
pc-p33	2X23	-9.77	2.26
pc-p36	1P44	-11.48	4.47
pc-c1a1	1P44	-9.76	2.43
pc-c1a2	1P44	-12.56	2.29
pc-c6a3	1P44	-11.40	6.40
pc-c7a2	4TZK	-10.49	3.13
pc-c7a3	1P44	-10.60	3.33
pc-c8a2	1P44	-11.83	2.65
pc-c8a3	1P44	-10.03	3.34



(a) Crystal structure ligand pc-d11 (PDB 4TZK ligand).

(b) Top pose of pyrrolidine carboxamide-p37 (in 1P44) with respect to crystal structure ligand pc-d11.

(c) Top pose of pyrrolidine carboxamide-c6a3 (in 4TZK) with respect to crystal structure ligand pc-d11.

Figure 3.19 Representative varying binding orientations of bulky pyrrolidine carboxamides obtained from Glide XP protocol.

3.3.4 In-situ mutation - Correction of aberrant poses from "light" pyrrolidine carboxamides

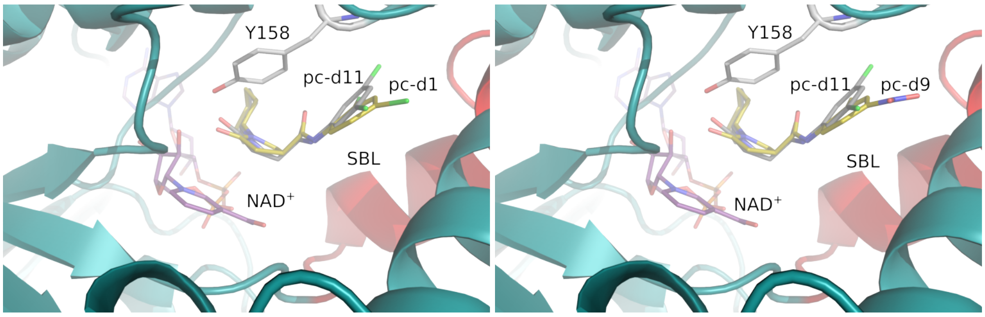
As mentioned in Section 3.3.3.1, a small subset of "light" pyrrolidine carboxamides (N=11, boldface ligands in Table 3.8) posed problems for the GlideXP binding pose prediction.

In order to ascertain the probable binding orientation for such compounds, the use of both GlideSP and XP was ruled out, since neither would yield appropriate poses. The solution for the same was manifested in form of in-situ mutation of the reference crystal structure ligand for the pyrrolidine carboxamides, i.e. pc-d11 (4TZK ligand). The central assumption behind this solution was simple; these molecules do not differ significantly from the reference ligand **AND** the majority of the related ligands demonstrates a “crystal structure-like” binding orientation. Thus, it is safe to assume that these ligands should also bind to Mtb-InhA in an orientation similar to the reference ligand (pc-d11, PDB 4TZK).

The process consisted of mutating the crystal structure ligand (pc-d11; PDB 4TZK) followed by a short minimization using Macromodel 9.9 and scoring in-place using the GlideXP scoring function (XPscore). The resultant poses were then subjected to rescoring with DrugScoreX and SFC scoring functions alike their ”normal” counterparts.

Table 3.10 Docking, DrugScoreX and substructure RMSD values of ”mutated” poses before (suffix _old) and after in-situ mutation (suffix _new). The old_RMSD refers to the substructure RMSD of the pose obtained from docking with GlideXP. The New_RMSD refers to the substructure RMSD of the in-situ mutated pose after the short minimisation. Both substructure RSMD values were calculated in fconv with respect to the reference ligand (pc-d11). Units for substructure RMSD are in Å, XPscore in kcal/mol.

Compound	pIC_{50}	DSXscore_old	DSXscore_new	XPscore_old	XPscore_new	Old_RMSD	New_RMSD
pc-s4	6.05	-115.35	-116.90	-9.25	-9.60	0.61	0.38
pc-s6	5.86	-110.48	-117.03	-10.49	-9.64	0.47	0.40
pc-s10	4.77	-124.06	-119.17	-10.14	-9.48	0.51	0.39
pc-s11	5.45	-132.41	-135.74	-10.32	-10.34	0.29	0.38
pc-s12	4.97	-117.23	-125.02	-9.22	-8.92	0.51	0.40
pc-s15	5.25	-134.47	-133.95	-10.36	-9.85	0.87	0.47
pc-d1	4.25	-115.65	-117.94	-9.37	-9.67	0.97	0.56
pc-d4	4.43	-148.29	-118.05	-7.45	-9.48	1.29	0.47
pc-d9	4.50	-103.24	-126.05	-8.04	-8.78	1.39	0.57
pc-d14	5.44	-128.86	-156.08	-7.45	-10.76	1.60	0.38
pc-d15	5.79	-128.15	-121.52	-9.38	-8.31	1.15	0.69



(a) Mutated pose of pyrrolidine carboxamide-d1 (olive; in 4TZK) with respect to crystal structure ligand pyrrolidine carboxamide-d11 (grey); SBL in red.

(b) Mutated pose of pyrrolidine carboxamide-d9 (olive; in 4TZK) with respect to crystal structure ligand pyrrolidine carboxamide-d11 (grey); SBL in red.

Figure 3.20 Mutated poses of pyrrolidine carboxamides with reference to the crystal structure orientation.

Table 3.11 SFC229p and SFC290p scores of pyrrolidine carboxamides whose poses were generated by in-situ mutation of crystal structure ligand pc-d11.

Compound	pIC_{50}	SFC_229p		SFC_290p	
		new	old	new	old
pc-s4	6.05	8.05	7.45	7.99	7.86
pc-s6	5.86	8.02	7.30	7.91	7.63
pc-s10	4.77	8.17	7.27	7.83	7.40
pc-s11	5.45	8.44	7.11	8.40	7.53
pc-s12	4.97	8.39	7.29	8.34	7.64
pc-s15	5.25	8.46	7.33	8.21	7.73
pc-d1	4.25	7.80	6.07	7.88	6.26
pc-d4	4.43	8.01	7.68	7.96	7.66
pc-d9	4.50	8.13	7.31	8.07	7.22
pc-d14	5.44	8.91	7.76	8.98	8.05
pc-d15	5.79	8.24	7.53	8.31	7.83

From the Tables 3.10 and 3.11 and figure 3.20, it is amply evident that generating poses for the problematic ligands was largely successful. In case of pc-s11, the substructure RMSD increased marginally that was revealed upon comparison of the pose from docking and the one obtained from the in-situ mutation of pc-d11. For this solitary compound, mutation was necessary, since the pose from GlideSP had a much bigger substructure RMSD while the pose from GlideXP depicted the wrong orientation of the A ring substituents. The downside to the in-situ mutation was that some of the new poses demonstrated slightly lower values of DrugScoreX than their “docked” counterparts (for e.g., pc-d4). For such compounds, it implied that the mutated poses were less favourable as compared to that from docking in GlideXP, since more negative values of DrugScoreX means more favourable orientation.

3.3.5 Docking of Pyrrolidine carboxamides with induced fit

The Section 3.3.3.1 highlights the shortcomings of GlideXP in predicting reasonable binding modes for the bulky pyrrolidine carboxamides (N=18). This was reflected in the high substructure RMSD values for the entire group. This can be attributed to the lack of receptor flexibility during docking with GlideXP that makes ligand placement and subsequent scoring difficult in tight binding pockets. The receptor flexibility during docking was considered by docking the bulky pyrrolidine carboxamides with the induced fit (IFD) and associated induced fit with trimmed side chains (IFD-trim) protocols followed by pose selection in usual manner (GDRPS/IHPS protocols). The procedure and theory of docking with induced fit has been discussed earlier in Sections 2.3.2.2 and 3.1.2.

The pose evaluation criterion for the bulky pyrrolidine carboxamides was quite similar to that of the smaller pyrrolidine carboxamides, which consisted of evaluating the substructure RMSD and assessment of pose quality via ROC analysis. Of this, the substructure RMSD will be discussed here, while the latter part which relates to pose enrichment will be discussed in Section 3.3.6.

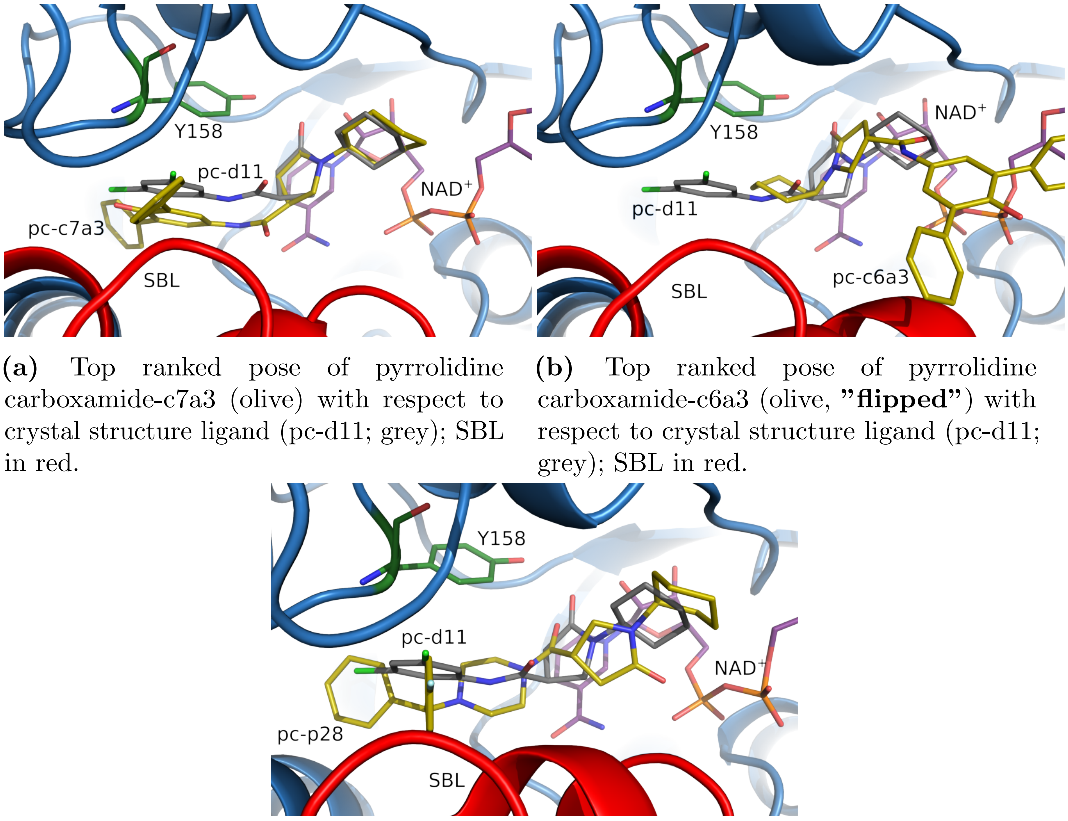
3.3.5.1 Pose preference

From the Table 3.13, it becomes amply clear that most of the bulky pyrrolidine carboxamides favour the active site of 2NSD followed by 4TZK and 2X23. Going by the size of the binding pocket available across each protein, the pose preference was somewhat surprising. The bulky molecules clearly favoured 2NSD as opposed to 4TZK for the light pyrrolidine carboxamides. Some molecules also got docked in 2X23, which has a very tight binding pocket. However, these molecules did not get docked in 2X23 at all using the GlideXP docking. This merely underscores the impact of incorporating receptor flexibility on the overall docking result. Furthermore, most of the docked ligands got selected from the IFD-trim protocol indicating that extra space in addition to the incorporated flexibility was necessary for pose generation.

3.3.5.2 Pose evaluation - Substructure RMSD and binding mode evaluation

Table 3.13 depicts the comparison of various evaluation parameters for bulky pyrrolidine carboxamides whose poses were obtained from induced fit and GlideXP docking protocols. There was a marked improvement in the overall XPscore and the values of DrugScoreX and SFCscore. However, the substructure RMSD values were still quite high. This underscores the problems of accurate pose prediction while simultaneously incorporating receptor flexibility.

One interesting observation from both GlideXP and induced fit docking was the predicted binding mode of pc-c6a3 (IC_{50} 140 nM). In both instances, the substructure RMSD values for the poses were quite identical ($\approx$ 6.40 Å), simply because of a **"flipped"** binding mode (cf. Figure 3.21). In case of some compounds (pc-p27, p28, p36, and p37) with substructure RMSD $>$ 3 Å, an atypical binding mode termed as **"flipped2"** was also seen (cf. Figure 3.21). In case of the "flipped2" mode, the secondary carbonyl group (near the A ring) forms the dual hydrogen bonds with Y158 and cofactor as opposed to the primary (from the B ring) as seen in crystal structures and the majority of the pyrrolidine carboxamides. The atypical binding mode can be attributed to the placement algorithm since the ligand conformation is not expected to change drastically during the protein conformational sampling. Nevertheless, a majority of the pyrrolidine carboxamide dataset exhibited a typical binding mode resembling the crystal structure ligands.



(a) Top ranked pose of pyrrolidine carboxamide-c7a3 (olive) with respect to crystal structure ligand (pc-d11; grey); SBL in red.

(b) Top ranked pose of pyrrolidine carboxamide-c6a3 (olive, **"flipped"**) with respect to crystal structure ligand (pc-d11; grey); SBL in red.

(c) Top ranked pose of pyrrolidine carboxamide-p28 (olive, **"flipped2"**) with respect to crystal structure ligand (pc-d11; grey); SBL in red.

Figure 3.21 Representative varying binding orientations of bulky pyrrolidine carboxamides obtained from induced fit protocol.

The diverse binding modes observed for a small number of compounds (N=5, pc-c6a3, pc-p27, p28, p36, and p37) warranted further investigation. The scaffold placement for the rest of the molecules was in line with the crystal structure ligand conformation albeit with higher RMSD values. In case of these 5 compounds, extensive molecular dynamics simulations were performed to assess the stability of such atypical binding modes. The results of the same are discussed in part II of this thesis. Considering the results of both GlideXP and induced fit, it can be seen that reasonable binding orientations for bulky pyrrolidine carboxamides could be obtained. The binding stability of these poses was further evaluated using molecular dynamics simulations.

3.3.6 Pose Ranking

The previous sections briefly described the numerous approaches for binding mode prediction of InhA inhibitors, especially pyrrolidine carboxamides. A critical aspect of docking pending evaluation was its ranking capability. In other words, the ability of the docking protocol to correctly rank the compounds based on their affinity. In pursuit of this goal, the top ranking poses of pyrrolidine carboxamides from GlideXP and Induced fit protocols (N=44, cf. Tables A.2 to A.4) were extensively evaluated. The starting point was to observe the correlation in between the experimentally determined activity (pIC_{50}) and the scoring functions (XPscore, DrugScoreX, and SFCscore). Table 3.12 depicts the correlations of various scoring functions with the experimental activity. From Table 3.12, it can be inferred that XPscore performs best from amongst the various scoring functions utilised for scoring and ranking of the selected poses from the pyrrolidine carboxamide subset and the InhA inhibitor dataset as well. Furthermore, DrugScoreX and SFC229/290 scoring functions also exhibited modest correlations with the experimental activity in case of the pyrrolidine carboxamide dataset. A contrasting case was observed for the InhA inhibitor dataset, where the aforementioned scoring functions correlated poorly with the experimental activity.

Table 3.12 Pearson's and Spearman's correlation values in between various scoring functions and pIC_{50} for entire pyrrolidine carboxamide dataset as well as the Mtb-InhA inhibitor dataset.

	pIC_{50}			
Scoring Function	Pyrrolidine carboxamide dataset (N=44)		InhA inhibitor dataset (N=108)	
	Pearson's R	Spearman's R	Pearson's R	Spearman's R
XPscore	-0.55	-0.61	-0.66	-0.50
DrugScoreX	-0.35	-0.34	-0.13	-0.12
SFC229p	0.49	0.50	0.31	0.38
SFC290p	0.41	0.40	0.27	0.30
SFC290m	0.46	0.49	0.29	0.39
SFC_RF	0.05	0.09	0.28	0.34
SFC_ser	0.14	0.10	-0.04	-0.05
SFC_met	-0.10	-0.15	-0.49	-0.45
SFC_frag	0.16	0.11	0.10	0.09
SFC_855	0.04	-0.10	-0.26	-0.25

Table 3.13 Substructure RMSD, SFC229p and DrugScoreX values of top ranked poses for bulky pyrrolidine carboxamides derived from GlideXP and Induced fit protocols. IFD stands for normal induced fit while IFDT denotes induced fit with trimmed side chains; the protein and protein-ifd procedure columns refer to the PDB from which the top ranking pose was selected. The induced fit and GlideXP docking did not yield any satisfactory pose for pc-c6a3 (marked in bold). Hence the best pose from GlideSP was selected and used subsequently.

		GlideXP protocol					Induced Fit protocol				
Pyrrolidine carboxamide	pIC$_{50}$	Protein	XPScore (kcal/mol)	DSX_score	SFC229p	RMSD (Å)	Protein-IFD procedure	XP-rescore (kcal/mol)	DSX_score	SFC229p	RMSD (Å)
r7	5.29	2x23	-8.17	-120.91	7.99	2.00	2x23-IFD	-9.34	-161.13	8.06	2.21
p9	5.46	1p44	-10.23	-134.65	7.37	1.94	2nsd-IFD	-10.79	-191.44	9.18	1.56
p20	6.12	1p44	-12.01	-152.34	8.50	1.89	2nsd-IFD	-12.37	-191.80	9.37	1.40
p21	6.39	2x23	-10.36	-175.88	8.41	2.08	2nsd-IFDT	-5.72	-175.21	8.73	3.59
p24	6.41	2x23	-10.49	-132.49	8.52	2.20	2nsd-IFDT	-8.67	-184.24	9.07	1.42
p27	5.28	1p44	-11.23	-139.35	7.62	4.41	2nsd-IFDT	-9.93	-199.52	9.35	3.16
p28	5.13	1p44	-11.52	-147.89	7.62	4.54	2nsd-IFDT	-7.70	-214.93	9.57	2.76
p31	5.86	2x23	-9.87	-149.81	7.79	1.92	2x23-IFDT	-10.79	-181.75	8.54	1.92
p33	5.59	2x23	-9.77	-158.97	8.30	2.26	2nsd-IFDT	-5.25	-177.32	9.83	3.07
p36	5.25	1p44	-11.48	-148.66	7.66	4.47	2nsd-IFDT	-8.81	-208.24	9.39	3.15
p37	5.34	1p44	-11.22	-125.15	7.54	3.88	2nsd-IFD	-8.80	-212.26	9.76	3.03
c1a1	6.33	1p44	-9.76	-110.29	6.56	2.43	2x23-IFDT	-11.11	-180.02	9.13	3.13
c1a2	6.07	1p44	-12.56	-164.77	8.29	2.29	2nsd-IFD	-6.26	-184.48	9.98	1.37
c6a3	6.85	1p44	-11.40	-127.38	7.77	6.40	4tzk-rigid	-5.17	-120.99	8.33	6.46
c7a2	6.49	1p44	-10.49	-106.80	7.96	3.13	2nsd-IFDT	-12.06	-183.65	9.06	1.89
c7a3	6.56	1p44	-10.60	-125.42	7.54	3.33	4tzk-IFDT	-9.18	-181.86	10.24	1.07
c8a2	6.20	1p44	-11.83	-154.98	7.42	2.65	2x23-IFDT	-12.69	-194.72	9.13	1.62
c8a3	5.88	1p44	-10.03	-134.20	8.24	3.34	2nsd-IFD	-7.83	-213.19	10.97	1.03

The moderate correlation of XPscore with the experimental activity can be understood from the fact that it is tailored for pose enrichment. Moreover, the poses selection protocols were intended to deliver poses of reasonable quality. Comparatively, DrugScoreX showed moderate correlation with pIC_{50} which worsened for the entire InhA inhibitor dataset. This merely indicated the utility of DrugScoreX in correctly recognising proper binding orientations that is independent of affinity prediction. The data also suggests that the SFC229p/SFC290p scoring functions show a moderate correlation with pIC_{50}, especially in case of pyrrolidine carboxamides (cf. Table 3.12). Hence, the four scoring functions, XPscore, DrugScoreX, SFC229p and SFC290p were subsequently evaluated for their ranking/activity-based distinguishing ability in an activity-based separation endeavour.

The ranking/activity-based distinguishing ability of the scoring functions was subsequently evaluated by setting activity-based cutoffs and then evaluating the affinity prediction of each individual scoring function via ROC analysis. To this end, the activity of the pyrrolidine carboxamide dataset (pIC_{50}: 4.13-6.85) was evenly split into three parts (cf. Figure 3.22). The first part (pIC_{50}: 4.13-5.04) represents the least active pyrrolidine carboxamides, followed by the "moderately active" (pIC_{50}: 5.04-5.94) and "highly-active" (pIC_{50}: 5.94-6.85) pyrrolidine carboxamides, respectively. The predictions of the individual scoring functions were then mapped onto the experimental activity via binomial logistic regression [278]. These models have been summarised in Table 3.14. The logistic regression models can be compared on the basis of their AIC and AICc values. The Akaike information criterion and its finite sample size correction term (AICc) [279], are indicators of the amount of information lost during the model generation. Thus, a lower value of AIC/AICc indicate a better model, and thereby a better model. However, both AIC and AICc are not the indicators of absolute quality of the model. The Section 4.3.3.1 briefly describes the theory of AIC. From Table 3.14 and figures A.1 and A.2, it can clearly be seen that the XPscore based "high" model has the lowest AIC and AICc value and hence is expected to have a better predictive power than that of other models based on scoring functions.

Table 3.14 Binomial logistic regression models generated using the pyrrolidine carboxamide dataset (N=44); p represents the SFC290p value, o is the SFC229p value, g the XPscore value, and d the DrugScoreX value.

Binomial Logistic regression			
Model	**Equation**	**AIC**	**AICc**
XPscore **mod**	$z = f(g) = -1.62 \cdot g - 15.20$	46.32	44.61
DrugScoreX **mod**	$z = f(d) = -0.08 \cdot d - 10.09$	49.24	49.53
SFC229p **mod**	$z = f(o) = 2.80 \cdot o - 20.36$	49.82	50.11
SFC290p **mod**	$z = f(p) = 2.16 \cdot p - 16.01$	51.89	52.18
XPscore **high**	$z = f(g) = -1.52 \cdot g - 16.23$	44.57	44.86
DrugScoreX **high**	$z = f(d) = -0.03 \cdot d - 5.44$	55.49	55.78
SFC229p **high**	$z = f(o) = 2.34 \cdot o - 18.71$	50.94	51.23
SFC290p **high**	$z = f(p) = 2.12 \cdot p - 17.43$	51.18	51.47

This was followed by assessment of the predictions of each of the scoring functions in correctly ranking molecules from each part. The ROC analysis of the predictions can be seen in Figures 3.23 and 3.24. It should be noted that the aforementioned ROC analysis represents a modification of the original method that is commonly used in virtual screening endeavours. In ROC analysis of a virtual screening, the true positive corresponds to a "binder" (active molecule) while the true negatives are represented by "non-binders" (decoys). The ROC curves fall into two main categories as follows:

- **MOD:** This group of curves assess the performance of the four aforesaid individual scoring functions in correctly recognising least active molecules. In other words, these curves signify the performances of the scoring functions in separating the least active pyrrolidine carboxamides from moderately and highly active compounds (separating molecules with $\mathrm{pIC}_{50} < 5.04$ from those with $\mathrm{pIC}_{50} > 5.04$). These curves hereafter will also be referred to as **Mod curves** for comparison purposes.
- **HIGH:** This group of ROC curves assess the performance of the four aforesaid individual scoring functions in correctly recognising highly active pyrrolidine carboxamides. These curves assess the performance of the scoring functions in separating the highly active pyrrolidine carboxamides from the remaining two parts, i.e., least active and moderately active. These curves hereafter will also be referred to as **High curves**.

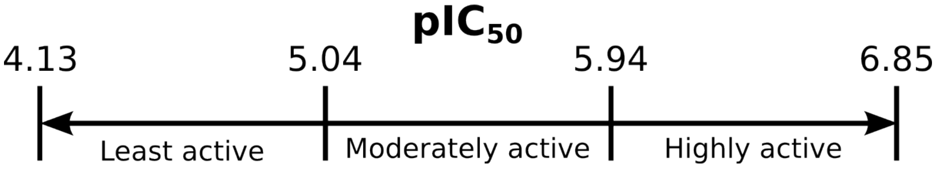


Figure 3.22 pIC_{50} range of pyrrolidine carboxamide dataset used for "activity class" generation.

The comparison of ROC curves and thereby the performance of individual scoring functions in activity-based separation can be judged by the *Area under the curve* or simply AUC value, which denotes the probability of the scoring function in ranking a randomly chosen highly active molecule (true positive) higher than randomly chosen weakly active molecule (true negative). The AUC values for Mod curves in Figure 3.23 clearly demonstrate the ability of XPscore in correctly differentiating moderately and highly active pyrrolidine carboxamides from the least active compounds. This is relevant from the early enrichment or the initial steep slope of its ROC curve. This was not surprising, since XPscore and docking with GlideXP (enhanced sampling) is tailored to provide a better separation of "inactive" compounds from "active" ones [244].

Furthermore, the AUC value for the mod curve of DrugScoreX was slightly better than that of either SFC229p or SFC290p. On the other hand, the performance of XPscore in separating the least and moderately active pyrrolidine carboxamides from highly active ones (High curves) was slightly changed as can be seen in Figure 3.24. This is supported by the drop in the early enrichment that was accompanied with a negligible change in the overall AUC of the curve. A similar trend was observed in the case of the SFC229p function as well. Both of these can be attributed to the fact that there are no ”inactives” as such in case of pyrrolidine carboxamides. Moreover, there is an equal number of weakly active and moderately active compounds in the pyrrolidine carboxamide dataset (N=15 for each group) with the remaining 14 constituting the highly active compounds. The performance and AUC values of SFC290p remained unchanged across both predictions indicating its robustness in providing an activity-based separation. On the contrary, the High curve for DrugscoreX showed an increase in the initial enrichment alongwith a drastic decrease in AUC value (0.63 vs. 0.80 for Mod curve), indicating an independent behaviour in correctly recognising either weakly or highly active pyrrolidine carboxamides.

A simple comparison of the AUC values of the mod and high curves indicate that it is much easier to recognise weakly active pyrrolidine carboxamides by any of the scoring functions as compared to the highly active ones. Going by the early enrichment trends across both mod and high curves, a combination of XPscore and any of the SFC scoring functions would be expected to yield a better result in providing an activity-based separation as compared to the individual scoring functions. This approach has been explored and analysed in detail in Chapter 4. The important factors to be considered for this approach is the narrow activity prediction range, small sample size and correlation in between the predicted and actual affinities for the pyrrolidine carboxamide subset.

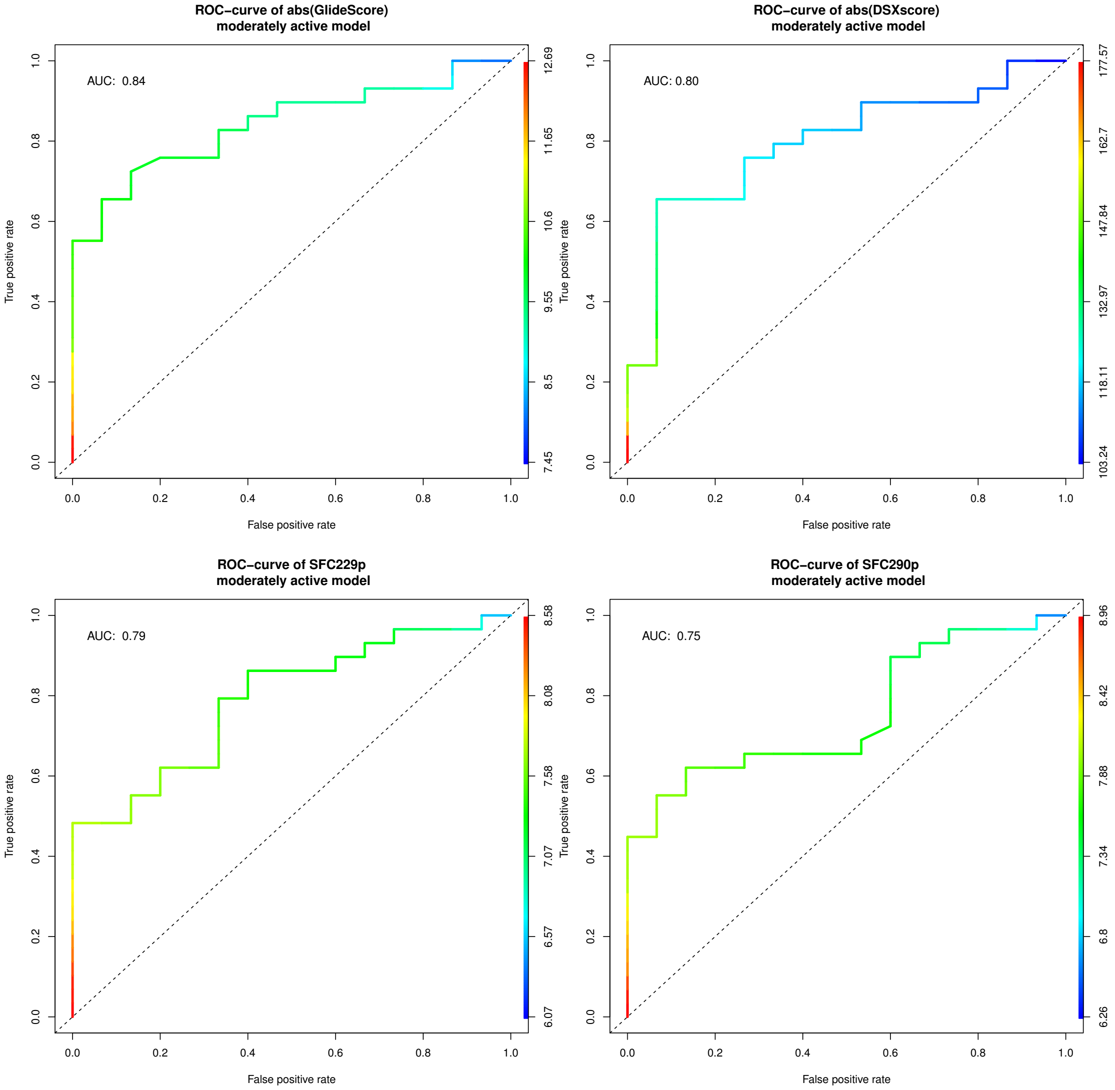


Figure 3.23 ROC curves denoting the performances of XPscore, DrugScoreX, SFC229p and SFC290p scoring functions in separating least active molecules from moderately-highly active molecules of pyrrolidine carboxamide dataset (N=44).

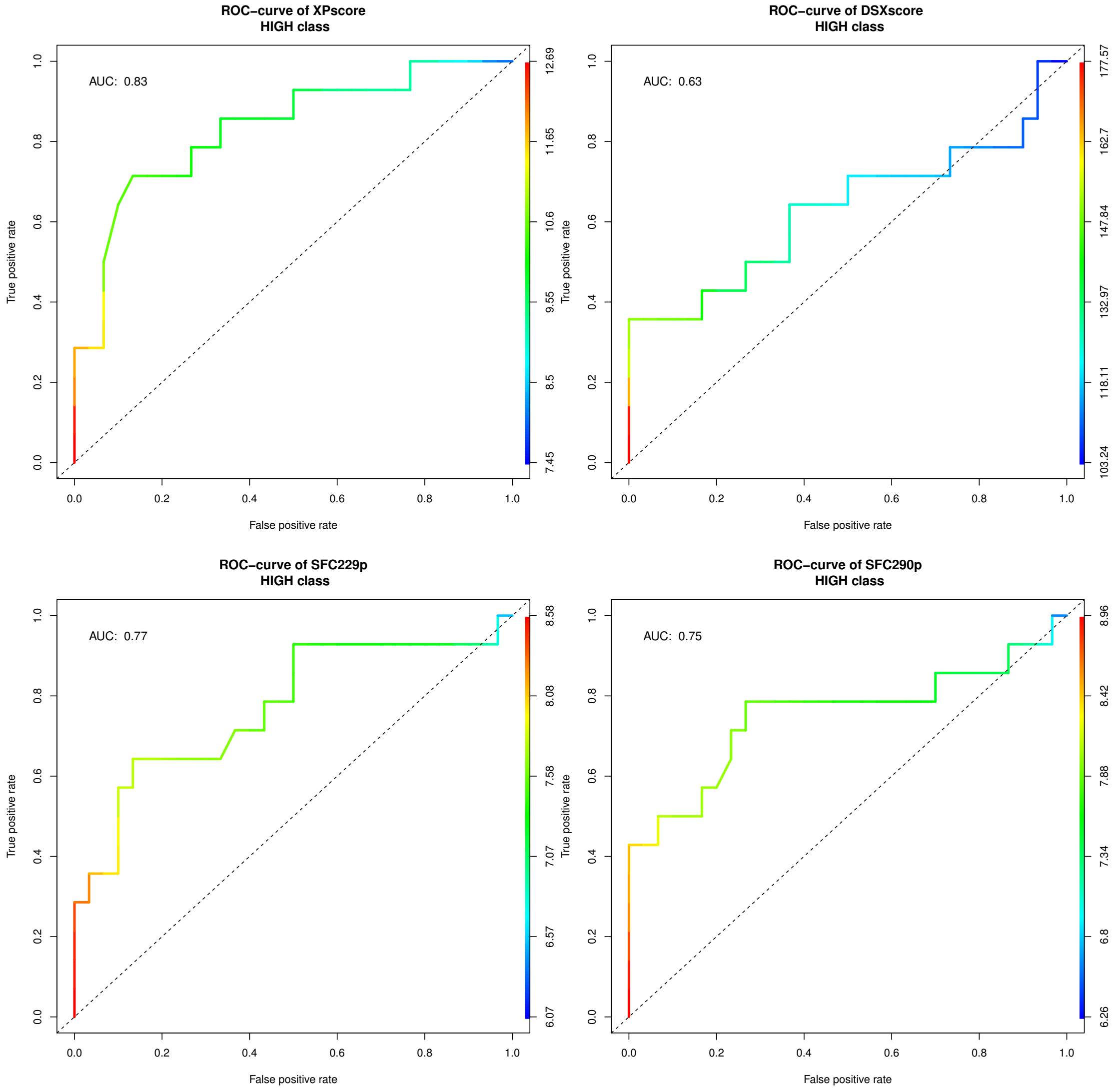


Figure 3.24 ROC curves denoting the performances of XPscore, DrugScoreX, SFC229p and SFC290p scoring functions in separating highly active molecules from moderately and least active molecules of pyrrolidine carboxamide dataset (N=44).

3.4 Conclusions - Implications for docking of moderate to bulky sized ligands in InhA

Using molecular docking, it was possible to generate reasonable binding orientations for pyrrolidine carboxamides that will serve as an input for binding free energy calculations using the Linear interaction energy method (cf. Chapter 4). The utilisation of Glide docking protocols that consider receptor flexibility demonstrated the pros and cons of the Glide docking methodology in docking diverse ligand series into InhA. The docking results clearly demonstrated the need for extended inclusion of protein flexibility while docking large sized ligands in InhA, while small to medium sized ligands can be docked with the extra precision mode with ease.

Moreover, the docking results especially from GlideXP and Induced fit emphasised upon the preference of specific inhibitor series towards their native protein that was clearly linked to the size of the binding pocket, size of the ligand, and chemical similarity across the inhibitor series. Accordingly, the Genzyme series favoured the wide open pocket of 1P44 followed by 2NSD, given that arylamides and Genzyme series share a large common substructure. This trend was also observed in smaller members of pyrrolidine carboxamides which preferred 4TZK to a major extent followed by 2NSD, 2X23 and 1P44. The diphenyl ethers preferred 2X23 and 2NSD over 4TZK and 1P44.

On the contrary, many of the bulky pyrrolidine carboxamides got docked selectively in 2NSD, which was surprising since it represents a much tighter binding pocket as compared to 4TZK. However, the fact that all of the poses were obtained via induced fit/induced fit with trimmed side chains indicate that a substantial degree of receptor flexibility is needed to dock large ligands in tight binding pockets. Accordingly, it can be safely concluded that for small to moderate sized ligands, GlideXP should suffice provided GlideSP is successful in yielding reasonable binding modes. In the event of problems in binding mode prediction with GlideXP and for large sized ligands, the computationally intensive Induced fit protocol is recommended.

The evaluation of the four scoring functions, XPscore, DrugScoreX, SFC229p and SFC290p highlighted their abilities in providing activity-based separation. The narrow activity range of pyrrolidine carboxamides in addition to the uniform distribution of molecules across all three activity classes (weakly active, moderately active, and highly active) presents a challenge for reasonable activity-based prediction. Finally, the pose selection protocol that served as the basis for the pose enrichment analysis seemed to perform reliably without much need for human intervention. The pose selection procedure mentioned in this work can thereby be utilised for pose selection in structure based optimisation protocols for other series of InhA inhibitors.

Chapter 4

Binding affinity prediction and activity-based classification of pyrrolidine carboxamides

4.1 Introduction

The docking performed in Chapter 3 aided in predicting the putative binding conformations for pyrrolidine carboxamides, with a particular focus on the potent members of the series. Given the approximate nature of docking, critical parameters like receptor flexibility, solvent contributions in binding and change in entropy upon protein-ligand association are truncated or neglected altogether. All of these terms actually warrant in-depth investigation to accurately predict the binding affinity of the ligand under consideration. The approximation of the protein-ligand binding becomes evident in the final interaction energy value or ”score”. This score is subject to further refinements by improved scoring functions or their combinations that progressively yield values that approach the actual or experimental binding affinity. An alternate way to evaluate the ligand-receptor interactions that incorporate receptor flexibility as well as solvation effects is Molecular dynamics/Monte Carlo simulations [280]. Accurate estimation of protein-ligand binding affinities is a quite challenging task with many methods available to assess the same [97]. The theory and pitfalls of these methods have briefly been covered in Chapter 2.

A central aim of this work was to develop methods that could aid in the structure-based optimisation of pyrrolidine carboxamides. In this context, methods that could aid in the classification or identification of promising molecules based on the structural or activity information of known molecular series were desirable. A key metric to classify or identify promising lead molecules is their activity that can be expressed as binding affinity or experimentally measured value like IC_{50}. This chapter mainly deals with the potential methods that can aid in the structure-based optimisation of pyrrolidine carboxamides.

The current work describes the utility of the Linear Interaction Energy (LIE) [99] method in generation of a binding affinity prediction model using poses of pyrrolidine carboxamides derived by docking. This method will be covered in detail in Section 4.1.1. The principal aim of this model was to focus on the structure-based lead optimisation of molecules deemed promising based on their calculated affinity values (ΔG_{calc}). In other words, this approach consisted of identification of promising molecules from traditional lead optimisation endeavour by using apparent binding affinity as a filter or classifier. In this context, the binding affinity prediction model was generated using a ligand series like pyrrolidine carboxamides with a reasonable number of molecules with well defined activity values. The resultant model could subsequently be used as a classifier to identify potential lead molecules as InhA inhibitors. A prime requirement for the binding affinity prediction model is ensemble averaged non-bonded interaction energies that were obtained via MD/MC simulations. The ensemble generation additionally aided in ascertaining the binding stability and the dynamic interactions of pyrrolidine carboxamides whose poses were obtained via docking.

An alternate approach for identification of promising leads would be to evaluate the combination of scoring functions in providing an activity-based separation. This is a modification of the approach in Section 3.3.6, wherein individual scoring functions were evaluated for their ability to correctly identify and assign a docking pose to an activity class based upon its score. In other words, a regression model built up from a combination of scoring functions was used for an activity-based separation endeavour of the entire pyrrolidine carboxamide dataset. In contrast, the current approach used a combination of scoring functions on a much smaller dataset containing compounds with stable binding as verified from the MD simulations. In context of the structure-based optimisation of pyrrolidine carboxamides, the XPscore and rescoring values (DrugScoreX, SFCscore) as well as the non-bonded interaction energy terms of selected pyrrolidine carboxamides were used to derive activity-based classification models using logistic regression. The underlying theory for this approach has been enshrined in Section 4.1.3.

4.1.1 Binding affinity prediction - The Linear Interaction Energy (LIE) method

There are numerous methods currently available for computational estimation of ligand-binding affinities. These range from simple scoring functions (empirical or knowledge-based) to the rigorous and time consuming alchemical perturbations (TI/FEP). The **Linear Interaction Energy** (LIE) method developed by Åqvist et al. [99, 138] offers a reasonable compromise between speed and accuracy for binding free energy prediction [101]. Moreover, the accuracy of this method [99, 138] has been reported to be much better than that of scoring functions. The prime requirement of the method is the ensemble representation of the bound and unbound states of the ligand from which the binding free energy is calculated according to Equation (4.1). As described earlier, the ensemble generation provides a means of assessing the process of protein-ligand binding and conformational changes associated with it.

The basis of the LIE method is the linear response approximation (LRA) [99, 281], which considers the endpoints of the thermodynamic cycle, i.e. bound and unbound forms to calculate the electrostatic free energy change upon binding. The method was adapted and generalized by Åqvist et al., wherein the binding free energy of a ligand is denoted as the change when a ligand gets transferred from aqueous (unbound state) to the protein environment (bound state). A thermodynamic cycle (Figure 2.7) can then be constructed to calculate the binding free energy given by Equation (4.1).

$$\begin{aligned}\Delta G_{bind} &= \Delta G_{bind}^{el} + \Delta G_{bind}^{vdW} \\ &\approx \beta \cdot \langle V_{bound}^{el} - V_{free}^{el} \rangle + \alpha \cdot \langle V_{bound}^{vdW} - V_{free}^{vdW} \rangle + \gamma\end{aligned} \tag{4.1}$$

where V_{bind}^{el} and V_{bind}^{vdW} are the differences in the non-bonded interaction energies for the ligand in bound and free states, respectively. $\alpha, \beta,$ and γ are empirical parameters, while $\langle\ \ \rangle$ denote ensemble averaged force field energies.

It is important to note that in order to predict the binding affinity, only the physically relevant states of the ligand (i.e. bound and unbound) are sampled using MD/MC procedures. This means that the bound and unbound states of the ligand have to be simulated separately in order to achieve appropriate ensemble sampling of the respective states. This is markedly different from rigorous methods like TI/FEP which construct several non-physical intermediate states in between the aforesaid forms to calculate the binding free energy. On the contrary, statistical methods like scoring functions predict the binding affinity using descriptors derived solely from the bound form.

A key aspect of Equation (4.1) are the empirical parameters which scale the contribution of the non-bonded interaction energies to the overall binding affinity. The empirical parameters (α, β) had an equal value of 0.50 stemming from the LRA with no scaling. However, scaling of these parameters along with addition of a new parameter γ was necessary in order to accommodate the various approximations being made while predicting absolute binding free energy [138]. For example, the value of α was refined to be 0.18 based on a set of 18 diverse protein ligand complexes in order to predict binding affinities for a wide variety of ligands [100, 282]. The offset parameter (γ) is system dependent and shows values that range from 0 kcal/mol for charged complexes to -7 kcal/mol for hydrophobic binding pockets (e.g. CYP450, retinol) [282]. Furthermore, the LIE method has several other variants that have been successfully utilised in binding affinity prediction [139, 283, 284] for diverse ligand series binding to numerous receptors. The electrostatic scaling factor was the last one to be refined, with optimal values being depicted in Table 4.1 according to Hansson et al. [285].

Table 4.1 Optimal values for β according to the chemical nature of ligand. Table adapted from Hansson, Marelius, et al., 1998 [285]

β	Chemical nature of ligand
0.50	Charged
0.43	Neutral
0.37	Neutral compound with a single hydroxyl group
0.33	Neutral compound with two or more hydroxyl groups

4.1.2 Sampling of MD simulations

The binding affinity prediction using the LIE method requires sampling of the bound and unbound forms of the ligand. Some key considerations while performing the MD sampling and subsequent affinity predictions are as follows:

1. The original method makes use of spherical boundary conditions (SBC) implemented in a surface constrained all-atom solvent (*SCAAS*) model [286] to solvate the bound and unbound states of the ligand, respectively. The *SCAAS* model represents a solvent sphere and its main aim was consistent treatment of polarisation effects at the surface in classical all-atom MD simulations. The spherical shape of the bulk solvent in contrast to other shapes like cubic, dodecahedron etc. [287, 288] implies a smaller number of solvent molecules thereby maximising speed and efficiency.

2. Normally, the same ligand conformation is used for the bound and free simulations of the ligand under SBC conditions [101]. This is primarily done to track the conformational changes as well as the associated energetics of an identical starting conformation in bound and free states, respectively.

3. The size of the solvent sphere/cube should be adequate so as to ensure proper solvation of the ligand. This in turn, avoids the lack of dielectric screening. Normally, a buffer of 10-15 Å around the ligand is sufficient to maintain a balance in between speed and accuracy.

4. Polar residues in the vicinity of the spherical edge (3-5 Å) and those outside the sphere boundary must be explicitly modelled as neutral. This is primarily because of the lack of dielectric screening in the simulations under SBC. On the contrary, no such modifications are necessary under periodic boundary conditions (PBC) because often it is used in conjunction with the particle mesh Ewald method that ensures appropriate consideration of dielectric effects.

5. The simulations of both forms of the ligand need to be performed under similar boundary conditions. This includes the center of the sphere and its size in both simulations [101]. Again this requirement is restricted to simulations under SBC [101].

6. The original LIE method accounted for the long range electrostatics by local reaction field approximation [216]. On the contrary, most modern MD simulation engines utilise the Particle Mesh Ewald method [215] for evaluation of long range electrostatics.

4.1.3 Activity-Based classification

The Section 3.3.6 describes in short the approach of using scoring functions for discrimination of binders from non-binders via logistic regression and subsequent ROC analysis. This section is a modification of that approach, wherein a combination of scoring functions (as opposed to a singular scoring function in Chapter 3) were used to achieve an activity-based separation of the pyrrolidine carboxamide dataset. The model emanating from the logistic regression of scoring functions and the experimentally determined InhA inhibitory activity assigns a probability value to each test molecule of being **least active or highly active**. This value was dependent on the value of the scores assigned to the molecule by the scoring functions used in the model generation. Based on the probability value, the molecule could be classified as least active, moderately active, or highly active. The logistic regression method was extensively used to generate the models. This section describes the theory and logic behind the model generation.

4.1.3.1 Principal Component Analysis (PCA)

Prior to the model generation, it was essential to determine which of the scoring functions would result in maximal separation of highly active from the least active ones. Furthermore, a case where the aforementioned separation is achieved using a minimum possible number of scoring functions would be desirable. A key point here would be to use scoring functions that correlate nicely with the experimentally determined activity while being of different type e.g., empirical, knowledge-based etc. Hence, it was necessary to ascertain the least possible number of scoring functions enabling maximal activity-based separation of pyrrolidine carboxamides.

In pursuit of this aim, a multivariate data reduction technique- Principal Component Analysis (PCA) was utilised. PCA primarily aims at simplification of the multiple dimensions of data to a few degrees and highlighting the common patterns that otherwise would remain concealed [289]. PCA was first described by Karl Pearson in 1901 and forms the basis for modern analysis of multivariate data [290].

PCA primarily achieves the simplification of data by orthogonal transformation of the data that yields the principal components (PC), with the first component (PC1) describing maximal variance of the system (cf. Figure 4.1). Each PC is a linear combination of the underlying variables that describe the data. Other principal components are derived along the same lines except that they remain orthonormal to PC1 and to each other as well. Mathematically the PC's are also referred to as eigenvectors of the covariance matrix used to perform the PCA [289]. A critical consideration while performing PCA is the minimal number of PC's used to extract meaningful information from the dataset. Although there is no straightforward answer to this question [291], the main point is that whilst performing multivariate dimensionality reduction, the first few PCs are the most important.

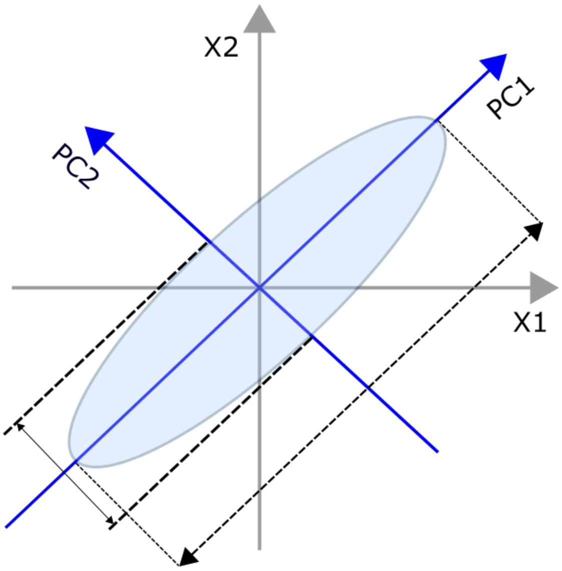


Figure 4.1 Basics of principal component analysis (PCA); X1 and X2 are the original variables describing the data in the shaded oval region. PCA orthogonalises and rotates the data yielding new axes PC1 and PC2. The PC1 represents the principal component describing maximal variance in the data.

4.1.3.2 Logistic Regression

Logistic regression is a statistical method used to estimate the probability of a binary response (outcome) based on one or more predictor variables. It was first described by David Cox in 1958 [278], wherein the logistic regression describes the relationship between the dependent and independent variables in form of probabilities using a logistic function ($\sigma(t)$) (cf. Equation (4.2)), where t is any real input and t ∈ R, where R is the range of the input. For such an input, the output always lies between 0 and 1, thereby interpretable as a probability.

$$\sigma(t) = \frac{e^t}{1+e^t} = \frac{1}{1+e^{-t}} \tag{4.2}$$

where t represents a linear combination of explanatory variables (y, y_1 ..., y_n). t can be expressed as:

$$t = \beta_0 + \beta_1 y_1 + ... \tag{4.3}$$

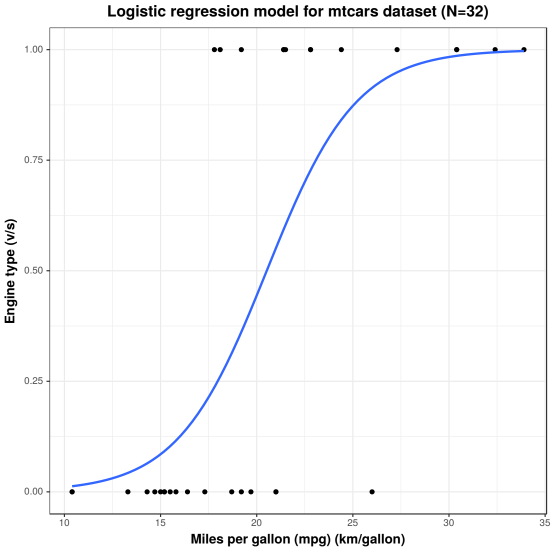


Figure 4.2 Logistic regression analysis of sample data mtcars from R: The dataset consists of 32 data points with 2 variables, type of engine (vs) and mileage (mpg). The blue curve represents the logistic regression model with the equation *1/ 1 + exp [- (0.43 - 8.83 . x)]*, p $\ll$ 0.05.

The logistic function can now be shown as:

$$F(y) = \frac{1}{1 + e^{-\beta_0 + \beta_1 y_1}} \tag{4.4}$$

The gradual nature of *F (y)* avoids the hard cutoffs (negative/positive) inherent to a binary classification scheme. A shift in the distribution of the explanatory variables manifests as a change in the slope of the function which is proportional to the shift, i.e., a large shift in the distribution of the explanatory variables results in a steep logistic regression function. The transition from either binary outcomes is usually denoted by a smooth path (cf. Figure 4.2).

The Figure 4.2 was generated in R using the inbuilt *mtcars* dataset that comprises of fuel consumption and 10 miscellaneous aspects of automobile design and performance for 32 automobiles [292]. The logistic regression function (blue line) was built with *vs* as the outcome variable and *mpg* as the continuous predictor, which explains the relationship in between the type of engine and its mileage. The term *vs* refers to the type of engine (vertical (v) or straight (s)) and *mpg* is miles per gallon of fuel. The model (blue curve) *1/1 + exp [- (0.43 - 8.83 . x)]* (x= *mpg*) models the shift in fuel efficiency when switching from a vertical engine (v) with a value of 0 to a straight engine (s) with a value of 1 (cf. Figure 4.2).

The term *"binomial" logistic regression* refers to those cases with a single continuous predictor variable and a dichotomous binary outcome (1 or 0) like in the aforementioned example [293]. When both predictor variable and outcomes are dichotomous, it is referred to as *multinomial logistic regression.* In such cases, only a single model yields the probabilities for the possible outcomes as a function of the independent variable, with the sum total of all probabilities equal to 1. The logistic regression can also be expressed in a generalised linear form [294], with the Equation (4.5) depicting the probability calculations, with $K = 3$ for the three activity classes as seen in Figure 3.22.

$$\begin{aligned} P(Y_i = K) &= \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}} \\ P(Y_i = 1) &= \frac{e^{\beta_1 \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}} \\ P(Y_i = K-1) &= \frac{e^{\beta_{K-1} \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}} \\ P(Y_i = K) + P(Y_i = 1) &+ P(Y_i = K-1) = 1 \end{aligned} \tag{4.5}$$

where Y$_i$ is the categorical outcome with K possible values (probability). The term k represents the outcome of an observation i, while the vector of the explanatory variables that describe i is denoted by X_i. Furthermore, since the logistic regression is nothing but a logit distribution, it can be used on output of PCA, with the model being trained on the coordinates of the PC subspace.

4.2 Methods

Figure 4.3 describes the methodology followed for the binding affinity prediction and activity-based separation models.

4.2.1 Dataset Used

The input for both affinity prediction and activity-based classification models were the top ranked poses obtained by molecular docking in Glide. A total of 44 compounds (cf. Tables 3.8 to 3.10) were docked, rescored with DrugScoreX and SFCscore and ranked. The entire dataset was subjected to molecular dynamics simulations both in bound (complex form) and unbound form (solvated in water box). From the "bound" simulations, molecules exhibiting stable binding were then ascertained by setting a "stability" criterion. The criterion set a cutoff of C-α RMSD ≤ 1.30 Å for the protein and ≤ 1.00 Å for the heavy atom RMSD of the ligand in the bound state.

A total of 23 molecules (cf. Table 4.2 and section 4.3.1) passed the cutoff and were subsequently used as a "training" set for both binding affinity prediction and activity-based separation models. It can be seen that about half of the starting compounds did not pass the stability criterion.This fact coupled with the narrow activity range of the pyrrolidine carboxamides, represents a challenge in generating an affinity prediction

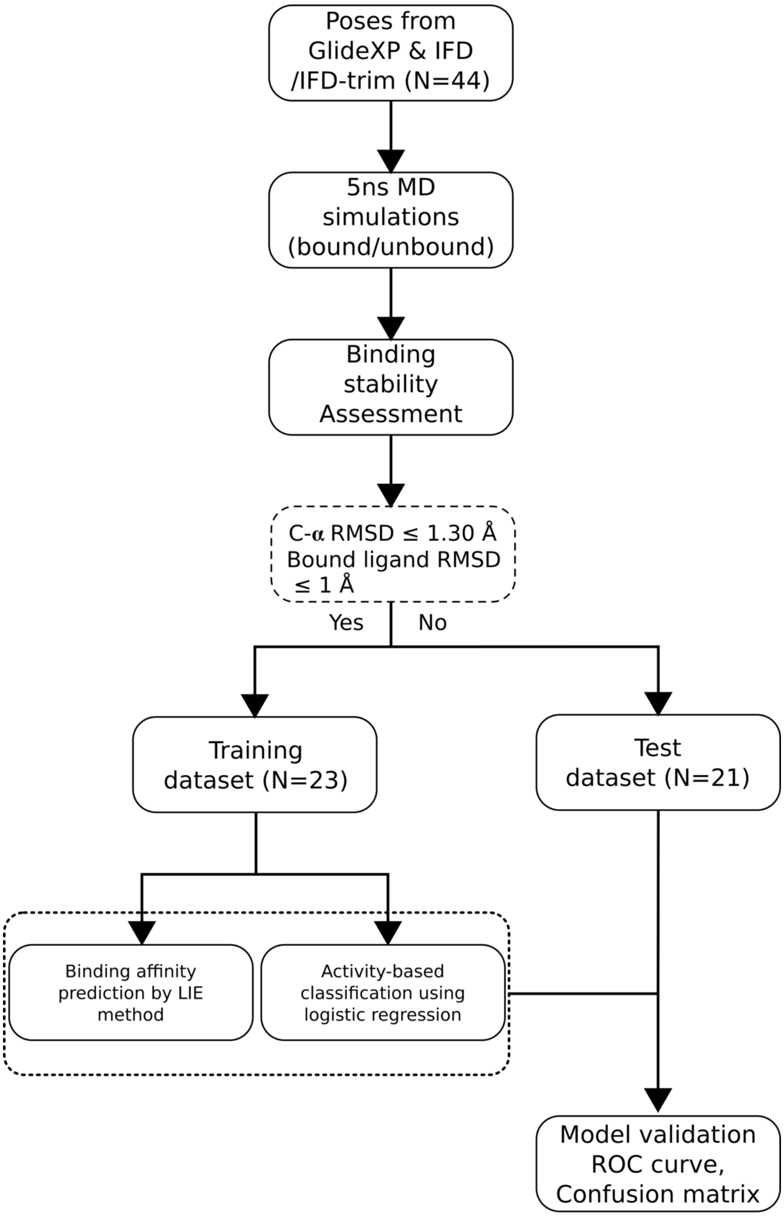

Poses from GlideXP & IFD /IFD-trim (N=44)
5ns MD simulations (bound/unbound)
Binding stability Assessment
C-α RMSD ≤ 1.30 Å Bound ligand RMSD ≤ 1 Å
Yes
No
Training dataset (N=23)
Test dataset (N=21)
Binding affinity prediction by LIE method
Activity-based classification using logistic regression
Model validation ROC curve, Confusion matrix

Figure 4.3 Workflow followed for derivation and validation of binding affinity and activity-based classification models

model. Moreover, the remaining molecules (N=21) (cf. Table 4.3) served as the test set for model validation (again both models). These molecules primarily served as an internal validation before the extensive process of model validation could be performed.

4.2.2 System Preparation

The LIE method requires sampling of the bound and unbound states of the ligand. Prior to setting up the system, all protein-ligand complexes with bound ligands and cofactor (NAD^+) were aligned to PDB 2X23 in MOE (v2012.10, Chemical Computing Group

Inc., Montreal, QC, Canada, 2012). The translated protein-ligand complexes were saved as PDB files for subsequent use. This was done because the poses for light pyrrolidine carboxamides were obtained from proteins that were aligned to 2X23. On the contrary, the poses of bulky pyrrolidine carboxamides came from original PDB structures. As a result, for consistent treatment of all systems, the C-α alignment was performed. Subsequently, the ligand and cofactors were extracted from the aligned PDB files and saved as individual mol2 files. The apo-protein resulting from this step was further stripped of hydrogen atoms and saved as a PDB file for subsequent system preparation. This step yielded a total of 44 apo protein files and equivalent number of ligand and cofactor files in mol2 format.

Subsequently, the tleap module of Amber10 [295] was utilised to assign the parameters of the Amber ff99SB force field to the protein. The RESP [196] charges for the ligands as well as cofactors were calculated based on the single point HF/6-31G* electrostatic potentials obtained from Gaussian 03 [296]. The missing parameters for the ligands and cofactors according to the General Amber force field were calculated using the parmchk [297] module of Amber10, while atom and bond types for the entire system including protein, ligand and cofactor were assigned using Antechamber [298]. The entire protein-ligand complex was subsequently regenerated in tleap and used for MD simulations.

4.2.3 Molecular dynamics simulations

The protein ligand complex was first subjected to a short minimisation of 2000 steps performed using a Born implicit solvent model [299–301] implemented in the sander MD engine of Amber11 [295]. Subsequently, the entire protein-ligand complex was solvated in a rectilinear water box with a buffer of 10 Å around the protein, with the solvent being defined by the TIP3P model [287]. The rectilinear shape represents one of the commonly used shapes by MD simulation programs, besides being ideal for implementation of the periodic boundary conditions (PBC). The PBC mainly aids in prevention of finite system effects, for example, the difficulties in reasonable estimation of the long range electrostatics.

Initially, the implicit water molecules from the crystal structure that were deemed important in ligand binding were retained followed by solvation of the protein-ligand complex. An adequate number of sodium ions was added to maintain the neutrality of the system. The resulting systems on average contained 42,000 atoms with dimensions of 76Å · 77Å · 78Å. At this point, the bound ligand was also transferred to a TIP3P rectilinear water box whilst retaining the bound conformation. An important point to be noted here is that the buffer region for the ligand was also equal to 10 Å, though the box size was much smaller and thereby contained a significantly reduced number of solvent molecules. Subsequently, both the solvated systems (protein-ligand complex and free ligand) were equilibrated using the same approach.

During equilibration, in order to relax the systems (solvated protein-ligand complex and solvated ligand) prior to the production runs, the system was heated up from 100K to 300K (over 20 ps) and cooled back to 100K (over 5 ps) under constant volume and temperature conditions, i.e. NVT ensemble using a Berendsen weak coupling algorithm [210] with time constant equating 0.5 ps. During the NVT run, the water molecules and ions were mobile whilst protein-ligand complex/ligand remained rigid by application of strong constraint of 10 $\frac{kcal}{mol \cdot \AA}$.

Subsequently, the temperature of the systems was gradually increased to 300K over a period of 25 ps (under NVT conditions), with all atoms of the system being mobile. Thereafter, the system was allowed to equilibrate further for a period of 50 ps under constant pressure and temperature conditions (NPT ensemble). During the simulation, the SHAKE algorithm was utilised to constrain the covalently bound hydrogen atoms while a time step of 2 fs was deemed suitable to ensure appropriate force and thereby energy evaluations of the system. The NPT simulations were run using Particle Mesh Ewald-MD (PMEMD) [302, 303] under periodic boundary conditions, which is a faster implementation of SANDER (**S**imulated **A**nnealing of **N**MR **D**erived **E**nergy **R**estraints). The constant pressure conditions were maintained by a Nosé-Hoover Langevin piston [212] while constant temperature was achieved by Langevin dynamics. For evaluating the non-bonded interactions, a cutoff of 12 Å was used, while the long range electrostatics were treated using the particle mesh Ewald method [215].

Subsequently, all of the systems were subjected to 5 ns production run with PMEMD whilst trajectory snapshots being saved every picosecond. For trajectory analysis and interaction energy calculation every frame of the 5 ns production run was considered, whilst omitting the equilibration frames. Once the production runs concluded, the Amber trajectories were imaged and saved in a binary format (.dcd) for further analysis. Given the number of molecules in the pyrrolidine carboxamide dataset subjected to MD simulations a total simulation time of $44 \cdot 5 \cdot 2 = 440$ ns was achieved. The trajectories were then subjected to routine methods of trajectory analysis.

4.2.4 Trajectory Analysis

4.2.4.1 RMSD analysis

The analysis of RMSD forms a basic and routine exercise when evaluating a trajectory. The RMSD analysis of the trajectories for the protein-ligand complexes (bound ligand state) was achieved by fitting the systems to backbone atoms (C, C-α, N, O) of a translated PDB 4TZK which was aligned to PDB 2X23 (chain A) in MOE (v2012.10). In effect, the chain A of PDB 2X23 served as an indirect global reference system, whilst the PDB 4TZK served as immediate reference for all simulated protein-ligand complexes. The RMSD analysis for the complexes was carried out in terms of C-α atoms and heavy atoms of the ligand in the bound state. In the free state, the ligand heavy atoms were used, with the ensemble averages being calculated over 5 ns for each ligand.

4.2.4.2 Calculating ensemble averaged interaction energies

The ensemble averaged interaction energies for the ligand in bound and free state were calculated with a cutoff of 12 Å. For calculating the pairwise non-bonded interaction energies, the pairs were defined as follows:

1. In case of the solvated protein-ligand complex, the ligand formed one object while the rest of the system, including water and ions, formed the other part of the pair.
2. For the solvated ligand, the above approach was applied analogously, with the ligand forming one object whilst the water molecules formed the other pair.

The pairwise non-bonded interaction energies were calculated for all ligands of the pyrrolidine carboxamide dataset, while the averaged interaction energies of ligands exhibiting stable binding were utilised to derive the binding affinity prediction models. In each case, the averaged interaction energies were obtained by averaging over the entire 5 ns trajectory (5000 frames).

4.2.4.3 Analysis Tools

The trajectory analyses were carried out in VMD [304] and cpptraj [305]. The RMSD analysis was performed using the *RMSD Trajectory Tool* of VMD, while the interaction energy analysis was carried out using the NAMD energy plugin of VMD and cpptraj_lie, respectively. All subsequent statistical analyses were performed in R [273] and associated packages, while structural visualizations were performed with PyMOL [276].

4.3 Results

4.3.1 Selection of dataset for training models

A key consideration while deriving both affinity prediction and activity-based classification models was that the molecules being considered should exhibit stable binding. In the absence of stable binding, reasonable convergence (ensemble average standard deviation ± 2 kcal/mol) of the interaction energies would not be achieved making the process of binding affinity prediction difficult. On a parallel consideration, docking poses (ligands) which exhibited stable binding can be considered to be reasonable and validated via the extensive conformational sampling via molecular dynamics. The **"stability"** cutoff (cf. Section 4.2.1) aided in identification of the stable binders (cf. Table 4.2) amongst the pyrrolidine carboxamide dataset.

Upon application of the stability criterion, a "training" set of 23 molecules (cf. Table 4.2) was obtained. From Table 4.2, it can be seen that the training dataset primarily consists of light pyrrolidine carboxamides, while only 4 bulky pyrrolidine carboxamides exhibited stable binding. The training dataset contained within itself two subsets of pyrrolidine carboxamides that were primarily used to ascertain the overall change in affinity prediction with a change in number of molecules in the training set.

Table 4.2 Training dataset used for generation of a linear regression model using the LIE method. The RMSD values are averaged over 5 ns production runs (bound state). The values of the empirical parameters $\alpha, \beta,$ and γ from the best performing LIE model (α=0.17, β=0.03, γ=-3.13) were used to arrive at ΔG_{calc}. The experimental binding free energy (ΔG_{exp}) has been obtained from the IC_{50} by using the equation $\Delta G = -RT\ ln(IC_{50})$ = -2.303 · R · T · pIC_{50} = -1.354 · pIC_{50} (using the SI value of R and T = 296 K). Units for RMSD are in Å and for binding free energy in kcal/mol.

Compound	pIC_{50}	C-α RMSD	Ligand RMSD	ΔG_{calc}	ΔG_{exp}
pc-s2	4.45	1.27	0.34	-7.11	-6.03
pc-s4	6.05	1.30	0.29	-7.41	-8.19
pc-s5	4.55	1.21	0.69	-6.53	-6.16
pc-s6	5.86	1.16	0.67	-6.84	-7.93
pc-s10	4.77	1.09	0.34	-6.87	-6.46
pc-s11	5.45	0.96	0.74	-6.83	-7.38
pc-s12	4.97	1.16	0.55	-7.43	-6.73
pc-s15	5.25	1.30	0.75	-7.01	-7.11
pc-d2	4.24	1.13	0.35	-7.07	-5.75
pc-d4	4.42	1.17	0.45	-7.03	-6.00
pc-d8	4.63	1.16	0.39	-6.87	-6.28
pc-d9	4.50	1.15	0.61	-6.80	-6.09
pc-d11	6.40	1.11	0.49	-7.05	-8.68
pc-d12	6.07	1.15	0.78	-7.09	-8.22
pc-d13	5.88	1.22	0.81	-7.66	-7.96
pc-d14	5.43	1.09	0.83	-7.57	-7.35
pc-d15	5.79	1.21	0.53	-7.36	-7.84
pc-d16	4.82	1.21	0.47	-6.98	-6.53
pc-3i	4.86	1.26	0.75	-6.51	-6.59
pc-r7	5.29	1.22	0.71	-7.04	-7.16
pc-p31	5.85	1.12	0.75	-7.42	-7.93
pc-c7a3	6.56	1.22	0.81	-8.17	-8.89
pc-c8a2	6.49	1.29	1.00	-9.38	-8.78

4.3.2 Binding affinity prediction-LIE

Binding affinity prediction models were trained using the molecules in training set (Table 4.2) and the entire pyrrolidine carboxamide dataset, respectively. The model generation and subsequent statistical analysis was performed in R [273]. Table 4.4

Table 4.3 Test set used for validation of binding affinity prediction and activity-based classification models. The ΔG_{calc} values were calculated with the empirical parameters $\alpha, \beta,$ and γ obtained for the training set (α=0.17, β=0.03, and γ=-3.13, cf. Table 4.4). See Table 4.2 for more details.

Compound	**pIC$_{50}$**	**C-α RMSD**	**Ligand RMSD**	Δ**G**$_{calc}$	Δ**G**$_{exp}$
pc-s1	4.97	1.44	0.32	-6.36	-6.73
pc-s17	4.13	1.29	1.06	-7.25	-5.59
pc-d1	4.25	1.61	0.92	-6.77	-5.75
pc-d3	6.01	1.46	0.54	-6.54	-8.14
pc-d6	4.99	1.31	0.67	-6.83	-6.76
pc-d7	5.50	1.79	0.32	-7.19	-7.45
pc-d10	5.82	1.44	0.57	-6.89	-7.88
pc-3a	5.40	1.39	0.82	-7.00	-7.31
pc-3j	4.53	1.36	0.88	-7.02	-6.13
pc-p9	5.46	1.41	0.98	-7.89	-7.40
pc-p20	6.12	1.59	1.41	-7.53	-8.29
pc-p21	6.38	1.53	1.83	-7.97	-8.64
pc-p24	6.40	1.48	1.19	-7.63	-8.68
pc-p28	5.19	1.74	1.13	-8.67	-7.03
pc-p33	5.59	1.75	1.64	-8.44	-7.56
pc-p36	5.25	1.62	1.16	-8.07	-7.12
pc-c1a1	6.07	1.36	0.69	-7.32	-8.22
pc-c1a2	6.33	1.52	0.71	-8.90	-8.58
pc-c6a3	6.85	1.30	1.50	-6.86	-9.28
pc-c7a2	6.20	1.44	1.18	-7.88	-8.40
pc-c8a3	5.88	1.47	1.19	-9.56	-7.97

and Figure 4.4 depict the binding affinity prediction models and statistical parameters associated with them. From Table 4.4, it can be seen that binding affinity prediction models represent a modest performance that is evident from the moderate Pearson correlation coefficient (R) of 0.61 for the best model. The low correlation can at least in part be attributed to the fact that the activity range is quite narrow (2.32 pIC$_{50}$ for the training set that corresponds to 3.1 kcal/mol). The residual standard error of the two models ranged from 0.85 to 0.92 kcal/mol, which may be confronted with the

Table 4.4 Binding affinity prediction models generated with the LIE method

Parameters	Linear Regression	
	”Training” (N=23)	”Whole” (N=44)
Multiple R^2	0.37 (0.61*)	0.22 (0.47)
Adjusted R^2	0.31	0.18
F-statistics	5.98	6.01
Residual Standard error	0.81	0.89
P-value (F-statistics)	9.10*10^{-3}	5.13*10^{-3}
Degrees of Freedom	20	41
Alpha (p-value)	0.17 (2.94*10^{-3})	0.09 (8.32*10^{-3})
Beta(p-value)	0.03 (0.31)	-0.009 (0.69)
Gamma(p-value)	-3.13 (0.01)	-4.90 (2.97*10^{-7})

Asterisk corresponds to Pearson’s R value

Training refers to the training set of 23 compounds (Table 4.2) used to derive the best affinity prediction model.

Whole refers to the entire pyrrolidine carboxamide dataset used for MD simulations.

The values of the empirical parameters from the best LIE model (the one from training set) were used to calculate the predicted relative binding free energy (ΔG_{calc}) in Tables 4.2 and 4.3, respectively.

standard errors of other affinity prediction methods, that typically range between 0.5 and 2 kcal/mol [306]. Considering the aforesaid facts, the performance of the generated models can be understood, since a small deviation in the prediction leads to a noticeable change in the correlation.

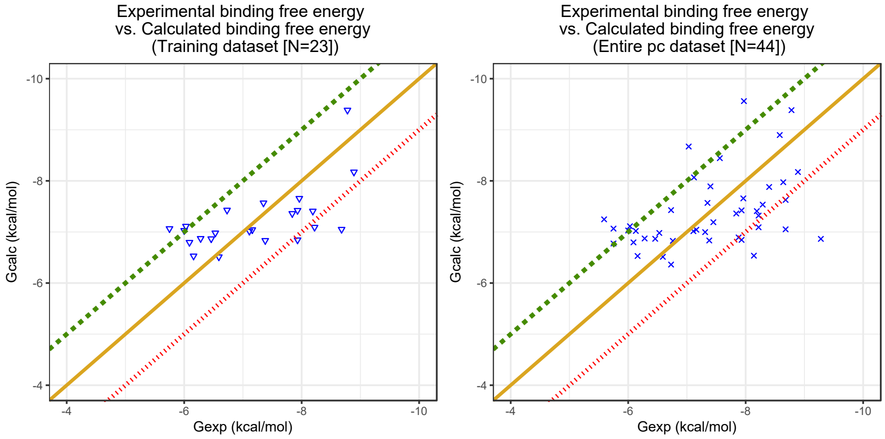


Figure 4.4 Experimental binding free energy (G_{exp}) vs. calculated binding free energy (G_{calc}) for the training and the entire pyrrolidine carboxamide dataset, respectively. The units of energy are kcal/mol. The dotted lines denote a range of ± 1 kcal/mol for the predicted binding affinity from the diagonal line that passes through the origin.

4.3.2.1 Testing of the LIE model

In order to assess the performance of the best LIE model in binding activity prediction, it was used to predict the activities of the molecules of test set and plotting the predicted affinity versus the actual activity (cf. Figure 4.5). From Figure 4.5, modest performance of the best LIE model is evident, with a sizeable number (5/21 or 24%) of the molecules being deemed as outliers visually. Furthermore, a total of 7 molecules (33%) were very close to either of the upper and lower affinity limits (± 1 kcal/mol). These facts merely suggest the affinity prediction challenge using the pyrrolidine carboxamide dataset. As described earlier, the model was trained with limited number of molecules with a narrow activity range. The test dataset had an even lesser number of molecules (N = 21) and a slightly wider activity range (2.71 pIC_{50} units) than that of the training set. However, it contained within it the most active (pc-c6a3; IC_{50}: 140 nM) and least active pyrrolidine carboxamide(pc-s17; IC_{50}: 73,580 nM) alongwith a large number of bulky pyrrolidine carboxamides, all of which were quite potent than the light pyrrolidine carboxamides against InhA. Nevertheless, the abject performance of the affinity prediction using pyrrolidine carboxamide dataset had a silver lining. Upon inspecting the predicted binding affinities for the molecules of the test set, it was seen that the best LIE model actually predicted the binding affinity for majority of the molecules within the 1 kcal/mol error range. This can be seen in Figure 4.5, where 5 molecules were wrongly predicted beyond the error range.

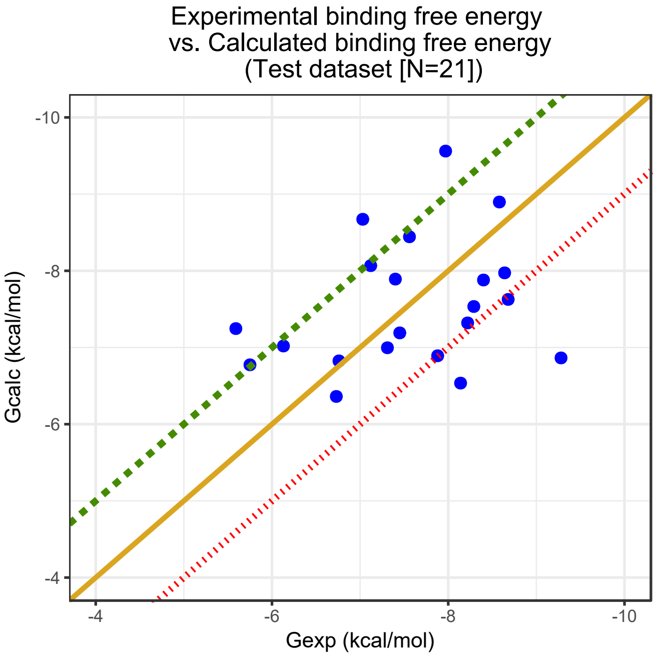


Figure 4.5 Calculated binding affinity (G_{calc}) versus the experimental affinity (G_{exp}) for the test set (N=21) according to the best LIE model (using the training set, cf. Table 4.4); The dashed lines represent the range (± 1 kcal/mol) for the predicted binding affinity using the best LIE model. The bold diagonal line passes through the origin.

The selection of the test dataset was quite a tricky affair, not just because of the aforesaid facts, but also because, in general, the test set should contain diverse molecules with well defined activity values measured against the same target under similar assay conditions. From this point of view, the molecules that were used in training the affinity models would represent and ideal choice. In other words, the pyrrolidine carboxamides and other InhA inhibitors would fit the aforementioned criteria. Though all of the molecules reported in Appendix A, have well defined activities under similar assay conditions, the error range for the IC_{50} values for each molecule exhibited a variable range. Given the experimental uncertainty in the IC_{50} measurements for each classes, only the pyrrolidine carboxamides not considered in binding affinity prediction model generation were used for model testing. Added to this, is the need for ensemble averaged interaction energy values for individual molecules. Considering the sheer size of the InhA inhibitor dataset, performing MD simulations for individual molecules in bound and free states would be cumbersome, if not difficult. These facts merely led to the use of the present test set for ascertaining the best affinity prediction model. The apparent performance of the binding affinity prediction models stressed the need of alternate methods for activity prediction or activity-based separation.

4.3.3 Activity-based separation - The activity ”class” approach

The activity-based separation was primarily devised as a means of classifying a test molecule as active or inactive without the need for the time consuming MD simulations. In short, the main aim was to provide an efficient method that was both fast and robust in separating the molecules of interest from the unimportant (inactive/least active) ones. To this end, the models were trained with a combination of scoring functions using logistic regression methodology (cf. Section 4.1.3.2). This was quite identical to the approach in Chapter 3 (cf. Section 3.3.6), where the ability of individual scoring functions was assessed in providing an activity-based separation of the poses obtained from docking.

Prior to training the logistic regression models (hereinafter referred to as *logreg*), the scoring functions of importance that were to be used in the training needed to be identified. These were identified by using a combination of correlation analysis and a subsequent principal component analysis (PCA) of the ”training” dataset (cf. Figure 4.6). From the correlation analyses (Figure 4.6), it can be seen that none of the scoring functions show a strong linear relationship with the experimental activity. Accordingly, in order to obtain the scoring functions that would aid in an activity-based separation, an PCA of the scoring functions was performed. The plot on the right hand side depicts the PCA of the molecules from the training set (Table 4.2) used in the affinity model generation. Each of the molecules are colour coded according to the pIC$_{50}$ as least (black), moderately (red), and highly (green) active. In addition to this, the principal components (PC1 and PC2) can be explained as a linear combinations of the parameters multiplied by their respective eigenvalues.

A combination of correlation and PCA analyses revealed that XPscore and one of the SFC scoring functions (SFC290p/229p) resulted in maximal activity-based separation of the pyrrolidine carboxamides along the first principal component (PC1). The separation of pyrrolidine carboxamides along the second principal component (PC2) was less as compared to that along PC1. Nevertheless, DrugscoreX and the electrostatic interaction energy term (ele) represented noticeable activity-based separation of the training set compounds along PC2. The correlation analyses, on the other hand, depicted a strong correlation amongst the SFC scoring functions, and a moderate correlation with pIC$_{50}$ as well as XPscore and the rest of the scoring functions, except for SFC290m and DrugScoreX. Both of these scoring functions, surprisingly showed a very high negative correlation. Going by the absolute correlation values and the results from RA, the following scoring function combinations were probed in an activity-based separation endeavour:

1. **Scoring function based models:** As the name implies, this group of models was purely trained with combinations of scoring functions. The various combinations of

the scoring functions were: XPscore-SFC290p, DrugscoreX-SFC290p, and XPscore-DrugscoreX. Additionally, the performance of the individual scoring functions in activity-based separation of the training set was also assessed.

2. **Force field based models:** This group of models was based on the non-bonded interaction energy terms derived from MD simulations. These terms were already available first hand for the training set from the LIE method.

All of the aforementioned *logreg* models could either predict the probability of test molecules being classified as active or inactive (binomial) or as least active, moderately active, and highly active (multinomial), respectively. Accordingly, binomial and multinomial models were generated, while the model quality was being assessed with help of AIC/AICc, respectively. For the model generation, the R packages lattice [307], nnet [308], pROC [259], nlme [309], and AICc [310] were utilized.

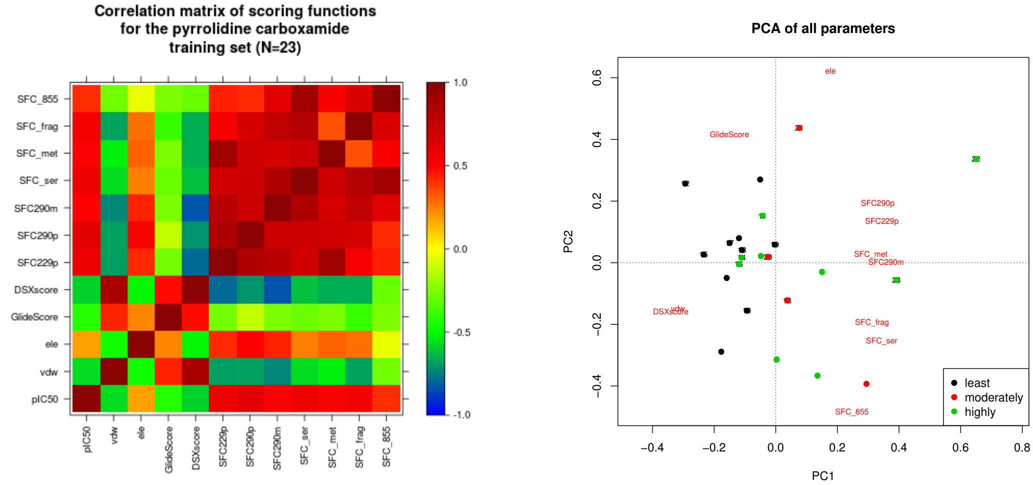


Figure 4.6 Correlation matrix and principal component analysis of experimental activity (pIC_{50}) as well as scoring values for molecules from the "final" set (N=23)

An activity-based range (cf. Figure 3.22) was used to partition the dataset into least active, moderately active and highly active. Using this range as a starting point, prediction models based on linear or logistic regression can be constructed to identify molecules of interest. Both regression strategies are fundamentally different (ordinary least squares versus maximum likelihood) with their own strengths and weaknesses. Logistic regression was preferred over linear one, simply because the aim of the work was to predict an outcome variable that was categorical (least, moderately active or highly active) based upon predictor variables that are continuous (rescoring values). Having a categorical outcome variable is not possible in case of linear regression since it violates the assumption of linearity. Furthermore, linear regression requires the dependant variable to be continuous (no categories/groups).

This can be better understood as follows: In the current activity-based separation endeavour, a model based upon linear regression would take in continuous variables as input (rescoring values) and would predict the affinity of the molecule (continuous). Logistic regression, on the contrary, would take in the rescoring values of a test molecule (continuous) and the outcome would be least, moderately active or highly active (categorical), with the sum of probabilities for the molecule being assigned to the individual class equal to 1.

In order to facilitate the activity-based classification using logistic regression, a class based approach was utilised. It consisted of deriving logistic regression models that belonged to two main classes namely,

- **Moderately active prediction:** This group of models (abbreviated as "Mod") had a pIC_{50} cut-off at 5.04. This model aimed at separating the least active molecules from moderately and highly active molecules.
- **Highly active prediction:** This group of models (abbreviated as "High") had a higher pIC_{50} cut-off of 5.94 and thus it was tailored to identify only highly active molecules, thereby separating the highly active molecules from moderately and least active molecules.

4.3.3.1 Model Characterisation:

Because of the intricacies of the underlying methods, the logistic regression models cannot be directly compared unlike in the case of linear regression where generally the square of Pearson's correlation (R^2) or the standard error and related statistics suffice. Hence, the *Akaike information criterion* [279] (AIC) was utilized for estimating the relative qualities of generated models. The AIC can be represented as:

$$AIC = 2k - 2\ln(L) \tag{4.6}$$

where, k = number of parameters estimated in the model and L = Maximum Likelihood of the model.

The maximal likelihood is a function of the parameters used in training of the model. In informal contexts, the maximum likelihood can be considered as "probability". In statistics, *probability* is used to describe the possible future outcomes given a fixed value of the parameter. *Likelihood* is used to describe a function of a parameter given a fixed outcome. This can be better understood by the following example: let p be the probability that a coin lands heads up (H) when tossed. So, a probability of getting two heads when the coin is tossed twice is p. If p=0.50, then the probability of seeing two heads is 0.25. Thus, the *likelihood* that p=0.50 given the observation HH = 0.25, is:

$$L(p_H = 0.50|HH) = P(HH|p_H = 0.50) = 0.25 \tag{4.7}$$

where, L is the likelihood. The AIC has its foundations in the information theory, and AIC offers a relative estimate of the information lost during model generation, representing the process that yields the data. Thus, a low AIC value indicates lesser loss of information during model generation and hence better predictive power. A critical point to note is that AIC offers no information about the absolute quality of the model, while it performs only as a discriminating metric for model comparison. Furthermore, since the sample size (number of pyrrolidine carboxamides) used for the model generation was finite, the 2^{nd} order AIC referred to as AICc [310, 311] (AIC with correction) was used for model comparison. The AICc can be put forward as follows:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{4.8}$$

where, n = number of samples and k = number of parameters.

4.3.3.2 Binomial logistic regression models

A brief comparison of numerous binomial models (cf. Figures 4.7, B.1 and B.2) is summarised in Table 4.5. From the Table 4.5 and Figures B.1 and B.2, it is evident that the binomial XPscore-SFC290p based ”mod” model exhibits the lowest AIC/AICc value and thereby least loss of information during its generation. As discussed earlier (Section 4.1.3.2), the quality of a *logreg* model can be assessed visually by inspecting its slope. The binomial XPscore-SFC290p based ”mod” model exhibits an almost vertical slope indicating a flawless separation of least active pyrrolidine carboxamides from the moderately active and highly active molecules. Thus, this model can be considered to be performing best from amongst all the binomial logistic models generated so far.

In addition to the binomial models, the XPscore-SFC290p combination was also assessed in providing a detailed activity-based separation via a multinomial model (cf. Figure B.3). Similarly, multinomial logistic regression models were constructed for the individual scoring functions and their combinations. The quality of these models was assessed in a way similar to the binomial models.

Table 4.5 Binomial logistic regression models generated using the training set of 23 pyrrolidine carboxamides exhibiting stable binding; p represents the SFC290p value, g the Glide XPscore value, and d the DrugScoreX value

Binomial Logistic regression			
Model	**Equation**	**AIC**	**AICc**
SFC290p + XPscore **mod**	$z = f(p,g) = 21.51 \cdot p - 8.54 \cdot g - 255.47$	14.67	15.94
SFC290p **mod**	$z = f(p) = 5.99 \cdot p - 47.79$	25.12	25.72
XPscore **mod**	$z = f(g) = -1.98 \cdot g - 18.98$	29.36	29.96
SFC290p **high**	$z = f(p) = 4.00 \cdot p - 32.81$	27.52	28.12
XPscore **high**	$z = f(g) = -1.03 \cdot g - 10.60$	31.87	32.48
SFC290p + XPscore **high**	$z = f(p,g) = -1.41 \cdot g + 4.15 \cdot p - 47.91$	27.28	28.54
SFC290p + DrugScoreX **mod**	$z = f(p,d) = -0.11 \cdot d + 4.03 \cdot p - 46.02$	22.05	23.32
DrugScoreX **mod**	$z = f(d) = -0.17 \cdot d - 21.00$	21.91	22.51
DrugScoreX **high**	$z = f(d) = -0.04 \cdot d - 5.29$	31.58	32.18
SFC290p + DrugScoreX **high**	$z = f(d,p) = -3.70 \cdot\ 10^{-3} \cdot d + 4.14 \cdot p - 33.51$	29.51	30.77
XPscore + DrugScoreX **mod**	$z = f(d,g) = -0.27 \cdot d - 3.30 \cdot g - 64.25$	18.30	19.57
XPscore + DrugScoreX **high**	$z = f(d,g) = -0.03 \cdot d - 0.90 \cdot g - 13.26$	32.24	33.50

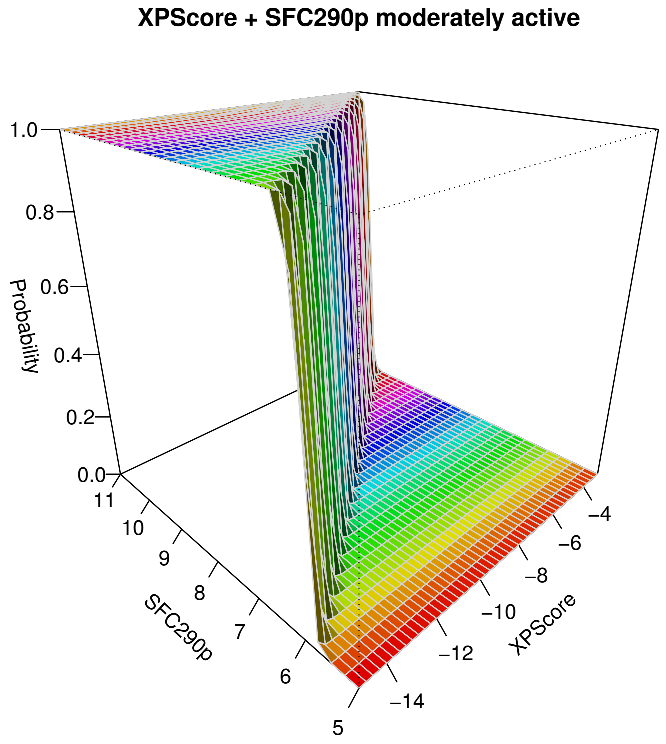


Figure 4.7 Binomial logistic regression model of XPscore and SFC290p to detect moderately active compounds; derived using ”final” dataset (N=23)

4.3.3.3 Multinomial logistic regression models

The multinomial logistic regression models that are based upon XPscore-SFC290p combination are depicted in Figures 4.8 and B.3. The comparison of the models is summarised in Table 4.6. Considering the AICc values of the respective models, it becomes evident that the XPscore-SFC290p combination works best in activity-based classification.

Table 4.6 Multinomial logistic regression models for XPscore-SFC290p combinations with their respective AIC, AICc values

Multinomial Logistic regression		
Model	**AIC**	*AICc*
SFC290p + XPscore	37.67	42.92
XPscore	49.35	51.57
SFC290p	44.23	46.45
DrugScoreX	41.87	44.09
SFC290p + DrugScoreX	42.08	47.28
XPscore + DrugScoreX	40.14	45.39
XPscore + DrugScoreX + SFC290p	42.02	48.03

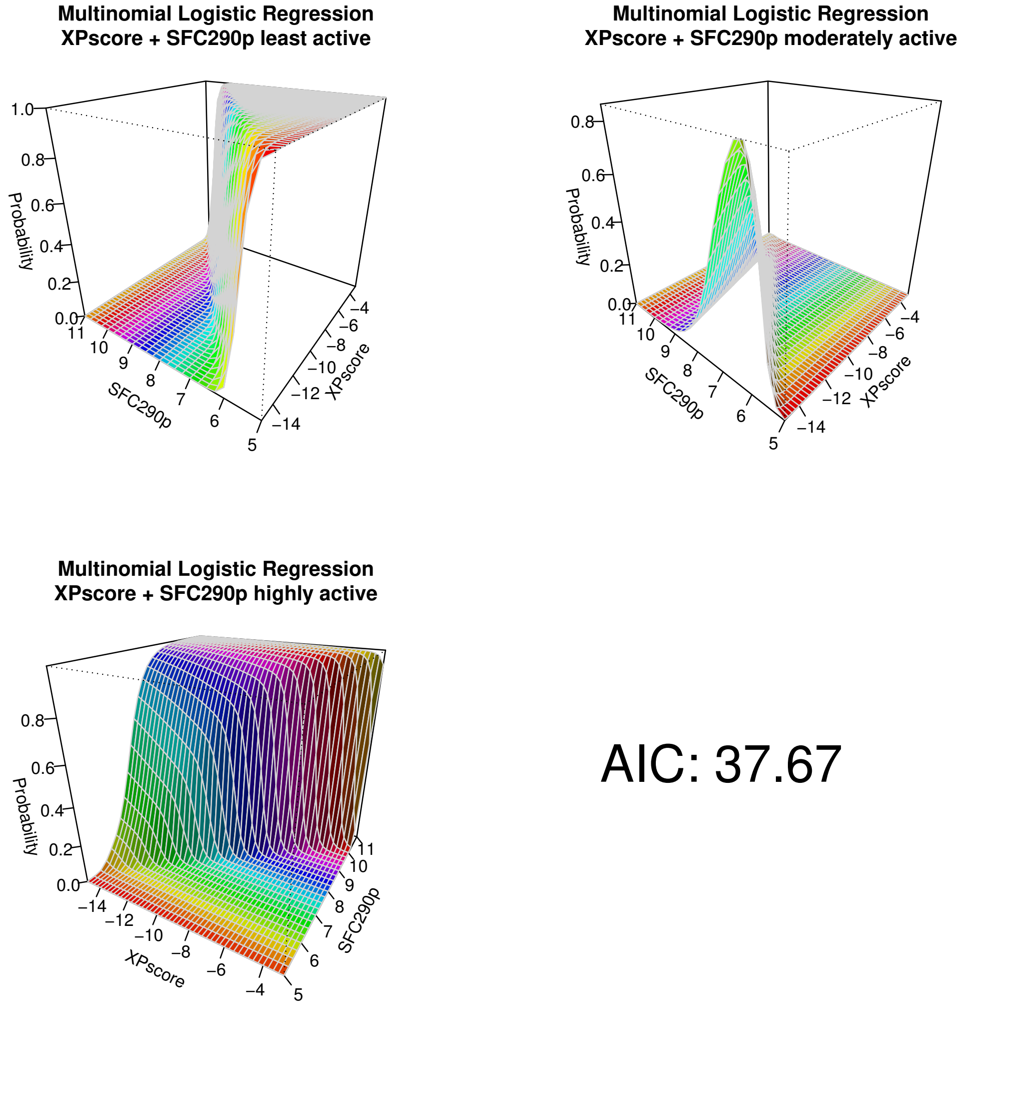


Figure 4.8 Multinomial logistic regression models of XPscore and SFC290p to classify compounds as least active, moderately active and highly active.

4.3.3.4 Comparison of force field terms with scoring functions

The previous sections described the utility of XPscore and SFC290p in providing for an activity-based classification. An extension of the previous approach was to compare the performance of the scoring function terms with that of the non bonded interaction energy terms derived from the MD simulations of molecules from the training set (N=23). The AIC and AICc values for "force field" based binomial and multinomial *logreg* models are summarised in Tables 4.7 and 4.8 and figure 4.9. The models themselves are depicted in Figures B.4 and B.5. The AIC and AICc values of the models clearly indicate the superiority of the XPscore-SFC290p combination over the force field terms combination in providing an activity-based classification.

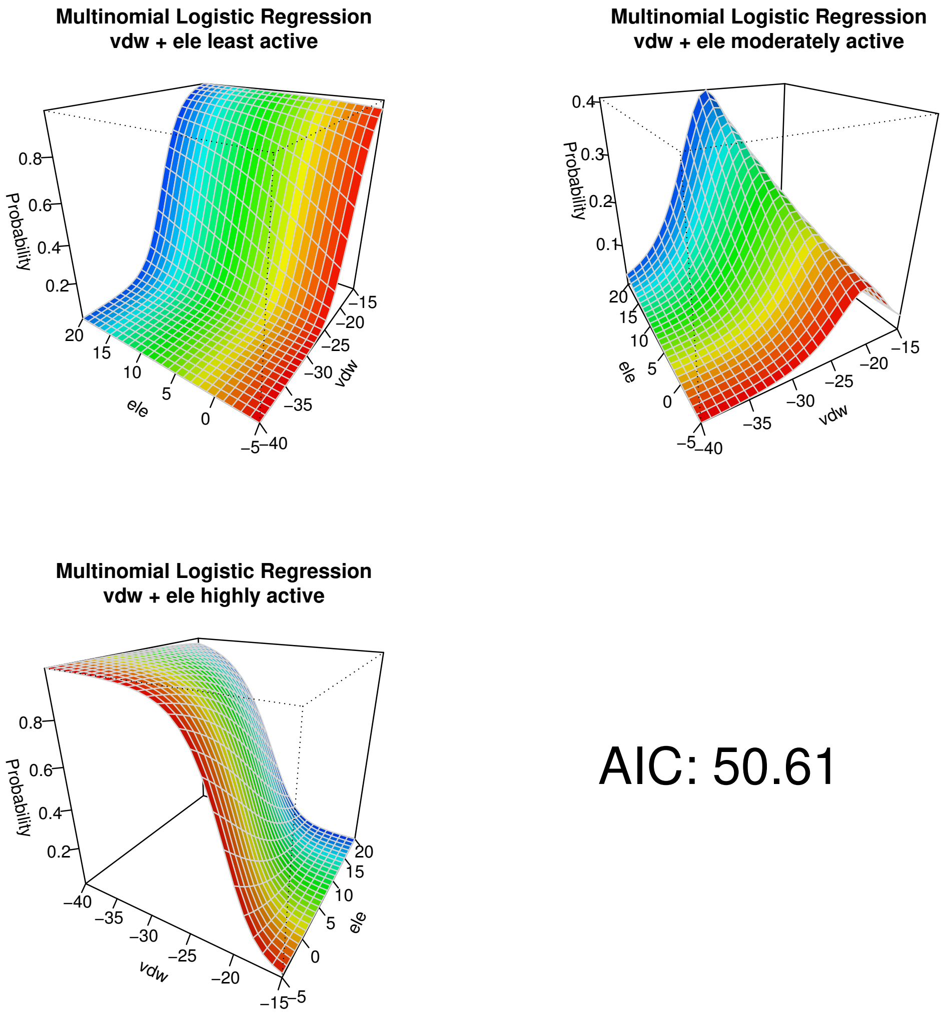


Figure 4.9 Multinomial logistic regression models of force field terms to classify compounds as least active, moderately active and highly active.

Table 4.7 Binomial logistic regression models based on force field terms, where e is electrostatic interaction energy and v is the van der Waals interaction energy

Binomial Logistic Regression			
Model	Equation	AIC	AICc
Elec + vdW "mod"	$z = f(e, v) = -0.10 \cdot ele - 0.53 \cdot vdW - 11.44$	30.28	31.54
Elec "mod"	$z = f(e) = 0.06 \cdot ele - 0.08$	35.01	35.61
vdW "mod"	$z = f(v) = -0.45 \cdot v - 9.90$	28.86	29.50
Elec + vdW "high"	$z = f(e, v) = -0.14 \cdot ele - 0.48 \cdot vdW - 10.92$	29.69	30.95
Elec "high"	$z = f(e) = 0.03 \cdot ele - 0.64$	34.63	35.23
vdW "high"	$z = f(v) = -0.33 \cdot v - 8.36$	29.02	29.60

Table 4.8 Multinomial logistic regression models based on force field terms and their combinations with their respective AIC, AICc values

Multinomial Logistic regression		
Model	**AIC**	**AICc**
Elec + vdW	50.61	55.86
Electrostatics	55.05	57.27
van der Waals	47.87	50.09

4.3.3.5 Validation of XPscore-SFC290p based binomial "mod" model

The test set that was used for validation of the LIE model was also utilised for validation of the XPscore-SFC290p "mod" binomial model. The test set was subjected to activity classification (cf. Table 4.9) followed by a confusion matrix analysis of the predictions (cf. Figure 4.10).

The true power of the XPscore-SFC290p "mod" binomial model can be clearly seen from the analysis of the validation results (cf. Table 4.10). These values clearly indicate the predictive power of the model in providing an activity-based classification. There are some caveats to this, mainly:

1. The limited sample size of the test set used in model validation.
2. The remarkable structural similarity in between the molecules used to derive the model and those used to validate it.

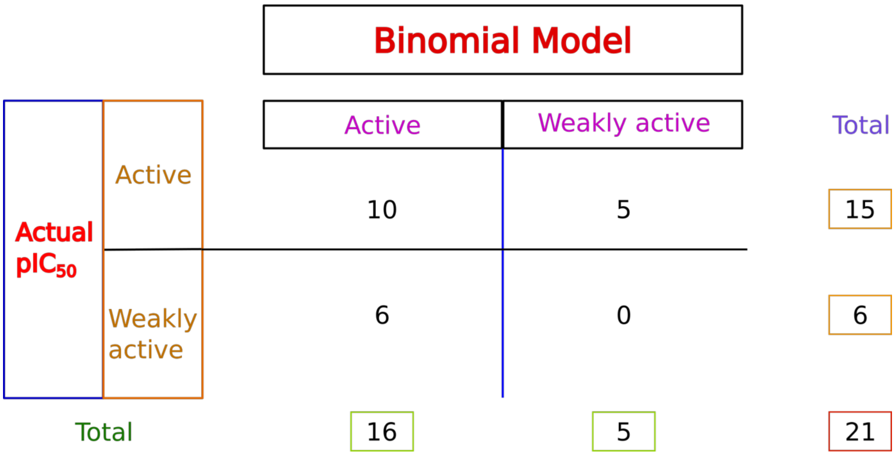
Inspite of the aforesaid facts, the AICc values and the absence of false positives in the final predictions do indicate the overall utility of the model. The performance of the binomial "mod" model can also be seen in the case where a logistic regression model (cf. Figure 4.12) was trained with a linear combination of XPscore and SFC290p scoring functions (PC1) as opposed to individual scoring functions like in the case of Figures 4.7 and 4.11. The PC1 was obtained from an RA analysis of Xpscore and SFC290p scoring functions. The steep slope of the "mod" model (left figure of Figure 4.12) is identical to the slope of XPscore-SFC290p based "mod" binomial model, merely indicating the

Table 4.9 The activity-based classifications of the pyrrolidine carboxamide test set (**Activity classification** column) according to the XPscore-SFC290p based binomial ”mod” *logreg* model. The values in round brackets denote the pIC_{50} values of the compounds. The last column depicts the actual activity class of the compound based on Figure 3.22. In the table, the terms ”least active” refers to the compounds that are actually least active ($\mathrm{pIC}_{50} < 5.04$, cf. Figure 3.22) as well as classified least active (model prediction < pIC_{50} 5.04). The term ”highly active” refers to the compounds that are highly active ($\mathrm{pIC}_{50} < 5.94$) and are classified as highly active. Finally, the term ”moderately active” refers to the compounds with actual pIC_{50} values in between 5.04 and 5.94. Given, the binomial *logreg* model had a binary output, the ”moderately active” compounds were clubbed together with the ”highly active” molecules by the model, i.e., the model treats moderately active molecules as highly active. The values in square brackets indicate whether the prediction was true (T) or false (F) as compared to the actual activity.

Pyrrolidine carboxamide	XPscore (kcal/mol)	SFC290p	**Probability** (ρ)	**Activity classification**	**Actual activity**
s1 (4.97)	-9.30	7.05	0.0000	Least active [T]	Least active
s17 (4.13)	-9.43	8.00	0.0620	Least active [T]	Least active
d1 (4.25)	-9.37	6.26	0.0000	Least active [T]	Least active
d3 (6.01)	-9.91	8.01	0.8323	Highly active [T]	Highly active
d6 (4.99)	-9.79	7.61	0.0003	Least active [T]	Least active
d7 (5.50)	-9.89	7.43	0.0000	Least active [F]	Moderately active
d10 (5.82)	-10.10	7.41	0.0001	Least active [F]	Moderately active
3a (5.40)	-9.31	7.44	0.0000	Least active [F]	Moderately active
3j (4.53)	-8.87	6.92	0.0000	Least active [T]	Least active
p9 (5.46)	-10.79	9.49	1.0000	Highly active [T]	Moderately active
p20 (6.12)	-12.37	9.42	1.0000	Highly active [T]	Highly active
p21 (6.39)	-5.72	8.63	0.0000	Least active [F]	Highly active
p24 (6.41)	-8.67	9.17	1.0000	Highly active [T]	Highly active
p28 (5.13)	-7.70	9.72	1.0000	Highly active [T]	Moderately active
p33 (5.59)	-5.25	9.56	0.0075	Least active [F]	Moderately active
p36 (5.25)	-8.81	9.59	1.0000	Highly active [T]	Moderately active
c1a1 (6.33)	-11.11	9.13	1.0000	Highly active [T]	Highly active
c1a2 (6.07)	-6.26	10.21	1.0000	Highly active [T]	Highly active
c6a3 (6.85)	-5.17	7.63	0.0000	Least active [F]	Highly active
c7a2 (6.49)	-12.06	8.99	1.0000	Highly active [T]	Highly active
c8a3 (5.88)	-7.83	10.69	1.0000	Highly active [T]	Moderately active

Table 4.10 Results from the confusion matrix analysis for the validation of the XPscore-SFC290p binomial *logreg* model

Parameter	Value
True Positives	10
True Negatives	5
False Positives	0
False Negatives	6
Sensitivity	63%
Specificity	1
Accuracy	71%
Precision	1



		Binomial Model		
		Active	Weakly active	Total
Actual pIC_{50}	Active	10	5	15
	Weakly active	6	0	6
Total		16	5	21

Figure 4.10 Confusion matrix analysis of the validation of the XPscore-SFC290p binomial*logreg* model carried out using the test of 21 pyrrolidine carboxamides excluded from model generation

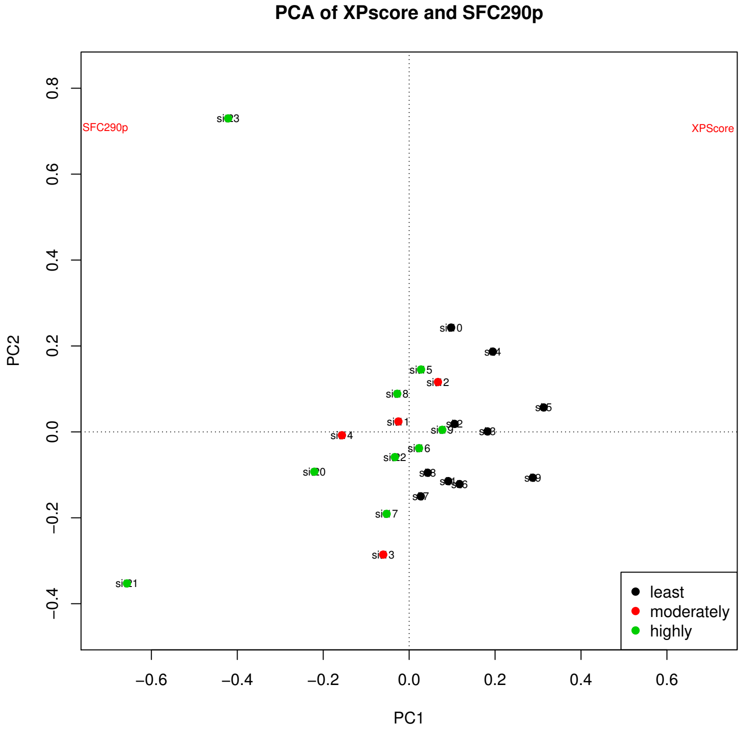


Figure 4.11 PCA of XPscore and SFC290 in an activity-based separation endeavour.

applicability of these two scoring functions in activity-based separation. Furthermore, in an applied scenario, the model requires the output of docking in form of a simple text file containing tab separated values of XPscore and SFC290p scores. Since the latter is easy to achieve, the use of the XPscore-SFC290p based ”mod” binomial model in a large scale virtual screening becomes appealing. However, an important caveat pertaining to all the logreg models is that all of them were trained with a small number of active compounds from the pyrrolidine carboxamide dataset. The narrow activity range of the pyrrolidine carboxamides together with limited number of compounds in training set limit the applicability of these models.

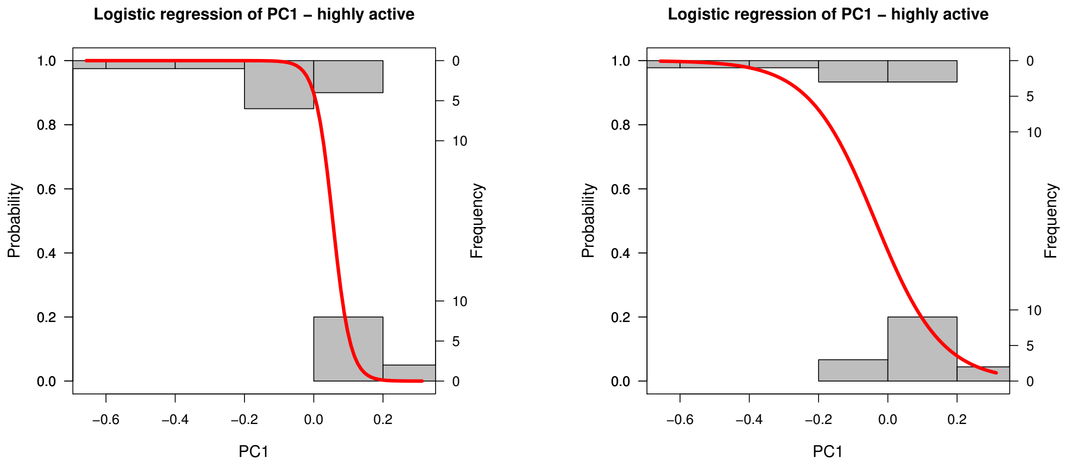


Figure 4.12 ”Mod” and ”high” logreg models trained with PC1 obtained from the PCA of XPscore and SFC290 (cf. Figure 4.11), respectively.

4.4 Discussion

The results section, particularly the binding affinity prediction underscores the challenge in predicting the affinity using the pyrrolidine carboxamide dataset for training the model. The LIE method [99, 101, 138] was used to generate the binding affinity prediction model. As a prime requirement of this method, the ensemble averaged interaction energies for the ligands of the pyrrolidine carboxamide dataset were needed. These energies were obtained by performing extensive MD simulations of the ligand in bound and free states, respectively. The stability criterion (C-α RMSD: $\leq$ 1.30 Å for protein and $\leq$ 1.00 Å for ligand) clearly pointed out that almost half of the dataset exhibited unstable binding. A majority of the ligands were from the bulky pyrrolidine carboxamides, that merely underscored the problems of pose prediction for bulky molecules. Thereafter, affinity prediction models were generated with molecules exhibiting stable binding that constituted the training set (Table 4.2). Additionally, specific subsets of the training dataset as well as the entire pyrrolidine carboxamide dataset was used to derive affinity prediction models.

A close inspection of the statistical summaries of the models revealed modest Pearson's R values, with the best LIE model having an Pearson's R equal to 0.61. This figure dropped to 0.47 when the entire pyrrolidine carboxamide dataset was used in model derivation. The situation remained unchanged even after careful choosing of select subsets from the training set of 23 pyrrolidine carboxamides(Table 4.2). The first subset contained 14 molecules and another one was made up of 11 molecules (hereinafter referred to as **”mutated”**). The latter exclusively contained light pyrrolidine carboxamides whose poses were generated by in-situ mutation of the reference ligand (pc-d11). The former subset (N=14, hereinafter referred to as **"original"**) was diverse and contained only those molecules that passed the stability criterion and did not include the poses from in-situ mutation of pc-d11. As mentioned earlier, no improvement in the R was seen, with the original dataset yielding a Pearson's R of 0.47. However, the mutated dataset yielded R of 0.61 that was equal to that of the best LIE model. Nevertheless, these facts suggest that binding affinity models trained with molecules having stable and uniform binding are expected to perform better than those models trained with molecules with stable but diverse binding.

The testing of the best LIE model, as expected revealed its modest performance in predicting affinity for a small test of 21 pyrrolidine carboxamides. These 21 compounds did not pass the stability criterion described earlier. The reasons for choosing the test set have already been explained in Section 4.3.2.1. Nevertheless, a small proportion of the test set was predicted wrongly and showed up as outliers, while a majority of the dataset was predicted within the narrow 1 kcal/mol margin (Figure 4.5). The performance of the model can be attributed to a variety of factors. Of these, the obvious reasons are the low numbers of molecules in the training as well as test set and the narrow activity range of pyrrolidine carboxamide dataset itself.

The abject performance of the LIE method in case of InhA was surprising given that it has been used to successfully predict the relative as well as absolute binding affinities for a variety of targets [100, 283, 312, 313]. In addition to the aforementioned causative issues, there are several differences in the implementation of the original LIE method and the current case that may be the reason behind the poor performance of the best LIE model. The following is a short discussion of the performance issues:

1. Almost all of the LIE studies cited earlier used weakly polar to non-polar ligands, which is quite identical to the current case. However **none** of them contained a cofactor except for MurD (ADP as cofactor) [312] and Cytochrome 1A2 (CYP1A2, heme as a cofactor) [313, 314]. In both cases, treatment of charged groups and cofactors poses issues in affinity prediction. This merely means that affinity prediction was bound to be affected in the current case as well. Furthermore, since the latter two cases are quite identical to the current study, all comparisons will be done with respect to these two systems.

2. A fundamental issue that might contribute significantly to the performance of the LIE model for InhA as opposed to other cases is the conditions and the software employed. As seen from the methods section (Section 4.1.2), there are fundamental differences among the implementation of the original LIE method and the current case. Of these, the nature of solvent, its buffer region and most importantly, the

approximation of electrostatics are expected to play a crucial role in the performance of the LIE model.

The work of Capoferri et al. [313] comes closest to the current LIE implementation (in this work) with respect to the conditions and protocol employed, although the simulation duration (2.5 ns vs. 5 ns for bound and free states) as well as programs differed (GROMACS vs. AMBER). A major difference to all other works [100, 312] is that they utilised a different solvent shape as well as calculation of long range electrostatics in comparison to the current work. The accurate estimation of long range electrostatics is one the fundamental issues that give rise to the standard deviation in the pairwise electrostatic interaction energy. In both current case and Capoferri et al. (see supporting information in [313] and Table B.1), there was a wide variation in the both absolute value and the standard deviation of the electrostatics as compared to the other non-bonded interaction energy, i.e. van der Waals interaction energy. This is in stark contrast to Perdih et al. [312], where the electrostatics in both bound and free states did not exhibit any noticeable difference.

These facts mainly underscore the fact that type of solvation model and shape, simulation conditions, and most importantly the calculation of long range electrostatics plays a critical role in the final affinity prediction using the LIE method.

3. The term γ denotes the hydrophobicity of the target. In order to properly account for this term, a modification of LIE termed LIE$_{SASA}$ was born [315]. This method evaluated the change in the solvent accessible surface area (SASA) of the ligand in the bound and unbound states during MD simulations and used the difference to arrive at the value of γ. However, in the current case, a similar approach with the training set demonstrated no improvement over (Pearson's R = 0.31). This again highlighted the challenge of affinity prediction while using the pyrrolidine carboxamide dataset.

Nevertheless, the Root Mean Square Error (RMSE) of the prediction of the current method (0.81 kcal/mol for best LIE model) is comparable to that of Capoferri et al. (4.1 kJ/mol $\approx$ 1 kcal/mol) [313] and Perdih et al. (0.90 kcal/mol) [312]. Considering these facts, the modest performance of the best LIE model can be justified.

Given the modest performance of the binding affinity prediction models using the pyrrolidine carboxamide training set, the possibility of achieving an activity-based separation was probed. Logistic regression proved quite valuable in this purpose. The redundancy (RA) and correlation analyses of the scoring functions used to rescore the docking poses revealed that Glide XPscore and SFC290p scoring functions resulted in the best possible results. The XPscore-SFC290p binomial model trained on the training set demonstrated the best performance amongst the various *logreg* models generated. The same was ascertained by ROC and confusion matrix analyses. However, the applicability of the model remains limited primarily due to the limited information that was used in the model generation. Nevertheless, the model can be applied as a filter in a large scale virtual screening endeavour, given the simple nature of its requisite input and output.

4.5 Conclusions

The binding affinity prediction and activity-based classification models built with poses obtained from molecular docking showed certain trends within the pyrrolidine carboxamide dataset. The binding affinity models generated using molecules exhibiting stable binding demonstrated a modest performance, given the limited sample size and narrow prediction range of pyrrolidine carboxamides. Hence, an activity-based separation endeavour was embarked upon. The principal component and correlation analyses prior to model generation showed that the XPscore-SFC290p combination was found to yield the best separation of actives and least active compounds. The binomial XPscore-SFC290p based "mod" model that separated the least active pyrrolidine carboxamides from the moderately and highly active compounds was deemed as the best performing by its AICc value as well as by a confusion matrix analysis. Furthermore, the ease of application and speed of prediction makes this model appealing for application in large scale virtual screening protocols. The applicability of the model remains limited primarily because of the low number and lack of chemical diversity amongst the molecules in the training set. Nevertheless, in a pilot virtual screening, this model can be used initially to identify potential hits.

Furthermore, the extensive conformational sampling of pyrrolidine carboxamides in bound and free states aided in the identification of pyrrolidine carboxamides exhibiting stable binding. Starting with the poses obtained from GlideXP and induced fit, it was assumed that the ligands exhibiting an orientation similar to the crystal structure ligand would be stable. However, MD simulations clearly proved that a sizeable proportion of the pyrrolidine carboxamide dataset exhibited binding instability even after conforming to the crystal structure conformation. The qualitative nature of the movements cannot be ascertained by inspecting the drifts of RMSD values alone. The next part of the thesis focuses on qualitative and quantitative aspects of ascertaining the dynamics that lie behind the apparent binding instabilities.

Chapter 5

Summary - Part I

Of lately, the importance of binding kinetics and their relation to binding affinity has warranted increased focus, given their intricately tied relation with overall efficacy of the drug [49]. However, the constraining factor behind the elucidation of the affinity-kinetics relation is the lack of structural information for the transition state. This is moresoever important in case of mycobacterial enoyl ACP reductase inhibitors which need to demonstrate a two step (slow-tight) binding in order to exhibit longer duration of inhibition and thereby their efficacy. The recent advances in computational resources as well as methods are expected to aid in ascertaining the mode and mechanism of binding of long acting inhibitors of enoyl ACP reductase an important target for anti-mycobacterial drug design [50].

The enoyl ACP reductase of *Mycobacterium tuberculosis* (InhA) is known to be inhibited by a wide variety of chemical scaffolds, all of which share a well-defined mode of interaction with InhA [50], but varying activity values as well as residence times. The ability of the molecule to bring about the ordering and subsequent closure of the substrate binding loop of InhA has already been shown to have a profound effect on the overall residence time and thereby the binding affinity [72]. This ability to bring about loop ordering and closure is quite intricately tied with the manner in which the molecule binds to the protein. This in turn is governed by conformational changes that happen over a period of time. MD simulations provide an attractive way of sampling the conformational changes that take place on a microsecond scale. For initiating MD simulations, reasonable starting structures were obtained by means of docking the InhA inhibitors across proteins demonstrating several conformations of the substrate binding loop. Furthermore, the receptor flexibility was also addressed by using a protocol [116] (Induced fit) that performs a truncated conformational sampling of the protein-ligand complex on the fly. Since the main aim of the study was structure based optimisation of pyrrolidine carboxamides [52], the computational methodologies used had a clear focus on the pyrrolidine carboxamides with a substantial attention being paid to the binding orientations being predicted by the docking and their quality of the selected poses thereof.

The pose selection process using in-house protocols yielded reasonable starting structures whose validation was performed by extensive MD simulations. The non-bonded interaction energy terms covering the ligand and its surrounding media were used to train binding affinity prediction models according to the Linear interaction energy approach [99, 101, 138]. The pose validation step followed prior to the model derivation, wherein pyrrolidine carboxamides exhibiting stable binding were ascertained by using stability cutoffs and subsequently used for model derivation while the excluded molecules formed the validation set. It was during this step the real ”dynamic” picture came to the front clearly highlighting the advances as well as shortcomings of the GlideXP and Induced fit protocols. The MD simulations clearly demonstrated that while the poses obtained from docking were of reasonable quality, there was still much more to the ”static” picture being painted by molecular docking. Nevertheless, the molecules exhibiting stable binding did buttress the notion that molecules exhibiting a crystal structure ligand like orientation should exhibit stable binding.

Additionally, the binding affinity prediction models were found to exhibit a not so satisfactory correlation in between the predicted affinity and the actual affinity (best model Pearson $R^2 = 0.37$). However, a thorough investigation of the model revealed causative reasons behind the poor performance of the model. The model though was found to exhibit high sensitivity and a moderate specificity. This was expected since the overall prediction range was quite narrow. The low number of molecules used to train the model as well as lack of chemical diversity in the training set limited its applicability to affinity prediction of only pyrrolidine carboxamides, which prompted use of other methods to achieve the activity prediction/classification. The information emanating from docking as well as MD simulations was subsequently used to train models that aided in activity-based classification. For this purpose, logistic regression was extensively utilised that partially did away the need of activity cutoffs for classification of a test molecule as active or inactive. The scoring function based *logreg* models were found to outperform the force field based binomial/multinomial models, with the XPscore-SFC290p based model clearly being deemed best amongst the numerous molecules derived. The validation of the same model via confusion matrix analysis as well as a principal component analysis of the ”predictions” (activity classifications) of the model aided in verifying the utility of the model. The model applicability was furthered by its incorporation in a script that accepts the docking score (XPscore) and SFC290p values (rescoring) as an input and yields easy-to-visualise colour coded probabilities of the molecule being active or inactive (least active).

In summation, the molecular docking as well as MD techniques aided in ascertaining the dominant binding conformations for pyrrolidine carboxamides with unknown binding modes. Moreover, the molecules exhibiting stable binding aided in deriving models that provide for rapid activity-based classification. The relative ease of application, accuracy as well as speed of prediction of the *logreg* models can certainly contribute to the virtual screening endeavours targeting search of mycobacterial enoyl ACP reductase inhibitors.

Part II

Conformational Analysis and Essential Dynamics of Pyrrolidine carboxamides to ascertain determinants of rapid reversible binding and their application to molecular design

Chapter 6

Introduction to Conformational analysis, Essential Dynamics, and determinants of rapid reversible binding

6.1 Pyrrolidine carboxamides and their diverse binding modes

Part I of this thesis depicted the utility of molecular docking in providing plausible binding orientations for pyrrolidine carboxamides. It also provided rapid activity-based classification models which have applicability in large scale virtual screening. The use of MD simulations in validating the results of molecular docking also aided in identifying the compounds exhibiting stable binding and those who do not. The RMSD values do not offer detailed structural insights into the molecular recognition process of those pyrrolidine carboxamides exhibiting unstable binding. Furthermore, it is unclear which key interactions are poorly established or even lacking in the case of the bulky pyrrolidine carboxamides.

Molecular docking performed in part I of this thesis clearly yielded 3 different binding modes for the potent pyrrolidine carboxamides (cf. Figure 3.21), most of which have a bulky A ring substituent. The first chapter of part II focuses on revealing the dynamics of these three binding modes with an aim to reveal the dominant binding orientation. For this purpose, an extensive dihedral analysis was performed for the bound ligands in the 5 ns MD simulations of protein-ligand complexes. The information emanating from this process was utilised to ascertain the nature of the movements exhibited by each of the binding modes. This information was also used in conjunction with the structural deviations (RMSD) to determine the dominant binding orientation of the bulky pyrrolidine carboxamides.

Additionally, the correlation between structural drifts of the protein-ligand complex and the distribution of the dihedrals for each class of the pyrrolidine carboxamides in the bound state was analysed. This mainly aided in ascertaining whether the pose fluctuated ”in-place” or was mobile within the binding pocket. The entire methodology and application of the same is explained in Chapter 7.

6.2 Essential Dynamics and Dynamic Cross Correlation analysis

The dihedral analysis in conjunction with the RMSD values and information available from the literature [52] aided in determining the dominant conformation for the bulky pyrrolidine carboxamides. However, both dihedral as well as the RMSD analysis cannot pinpoint the **direction** along which the fluctuations take place. Additionally, both analyses are unable to shed light on the protein-ligand fluctuations that are often correlated. Furthermore, these ”instantaneous” observable fluctuations cannot be extrapolated on a extended timescale to better understand the binding of pyrrolidine carboxamides to InhA.

Given the complex nature of MD simulations, it becomes quite difficult to locate and analyse the functionally important movements of the protein-ligand pair. This task is made simpler by dimensionality reduction techniques like Principal Component Analysis (PCA). Performing PCA on a trajectory yields **”modes”** or principal components that depict the variance in the movements. An interesting fact regarding PCA of MD simulations is that a major portion of protein dynamics is efficiently represented by a few number of collective ”modes” [316]. These modes appear to be essential for protein function and hence the dynamics of the subspace being represented by the main modes is referred to as ”Essential Dynamics” (ED) [316]. In the current scenario, essential dynamics was performed on the heavy atoms of the bound ligand as well as the key residues of the active site and the SBL (trajectory analyses) to reveal important movements of the residues that play a critical role in ligand binding and thereby its potency.

The most common application of ED has been in quasi-harmonic analysis of mass weighted coordinates of protein atoms to construct the covariance matrix of atomic movements [317, 318]. This approach is subject to various limitations just like the related Normal Mode Analysis (NMA), especially in case of proteins exhibiting large conformational changes taking place over a long time span [319]. Hence, essential dynamics was used in conjunction with dynamic cross correlation (DCC) [320, 321] to reveal the functionally important movements in the protein as well as the ligand. The theory and application of essential dynamics together with dynamic cross correlation has been enshrined in Chapter 8.

6.3 Design and analysis of optimised pyrrolidine carboxamides

The combination of essential dynamics and dynamic cross correlation aided in revealing the functionally important movements taking place in InhA upon binding of pyrrolidine carboxamides. This information was then coupled with the vast array of structural and dynamical information of slow-tight binders (diphenyl ethers and 4-hydroxy-2-pyridones) available in literature [37, 50, 67, 322] to design a series of optimised pyrrolidine carboxamides. These were subjected to extensive in-silico evaluations pertaining to their overall activity and binding dynamics. For this purpose, the previously established docking protocol was used to identify the probable binding modes. The poses emanating from molecular docking were subjected to short 5 ns simulations just like the published pyrrolidine carboxamides, followed by assessment of their binding stability. The trajectories were subjected to the entire post-MD analyses which were performed on their precursors.

Additionally, the designed molecules were assessed for their predicted in-silico activity using the XPscore-SFC290p "mod" logreg model followed by their mycobacterial permeability assessment using MycPermCheck 1.2 [53]. All of the aforementioned analyses aided in ascertaining the top hits from the optimised pyrrolidine carboxamides. These "hits" together with a small number of potent bulky pyrrolidine carboxamides and the reference crystal structure ligands were subjected to extended molecular dynamics simulations (150 ns per ligand) to ascertain their binding stabilities on extended time scale. Finally, the novelty of the designed molecules was ascertained by performing a "molecular" cross-check with the scaffold search tool implemented in SciFinder® [323]. The entire approach is depicted in Chapter 9.

6.4 Molecular determinants of rapid reversible binding

The conformations that the ligand and thereby the binding site residues attain plays a critical role in the overall nature of ligand binding, i.e. slow tight binding or rapid reversible binding. The detailed structural and conformational information governing the binding of slow-tight binders like diphenyl ethers is well known [261]. This information is clearly lacking in case of moderately potent InhA inhibitors like pyrrolidine carboxamides, mainly due to the lack of crystal structures for the more potent bulky pyrrolidine carboxamides. Elucidation of the structure and dynamics behind the binding of bulky pyrrolidine carboxamides can thereby reveal molecular determinants of rapid reversible binding which can drive the structure-based optimisation of this series of InhA inhibitors.

In order to reveal the molecular determinants driving rapid reversible binding of pyrrolidine carboxamides, the extensive MD simulations (150 ns) of selected pyrrolidine carboxamides that totalled 1.35 μs were subjected to clustering based on the 2D RMSD matrices of residues of the active site and the SBL. This was followed by analysis and comparison of the conformations of the cluster representatives with the information published in literature [261]. The clustering algorithm and atomic selections were kept similar to the ones used by Merget et al. [261] in order to compare the critical conformational differences in between the dominant cluster representatives. The entire approach for the above is enshrined in Chapter 10.

Chapter 7

Conformational analysis of pyrrolidine carboxamides

7.1 Pyrrolidine carboxamides and their binding modes

The molecular docking of pyrrolidine carboxamides, especially the bulky members predicted diverse binding modes for the respective compounds. A sizeable number of bulky (15/18) and light pyrrolidine carboxamides (6/26) exhibited binding instabilities, given that they did not satisfy the stability criterion (C-α RMSD $\leq$ 1.3 Å; bound ligand RMSD $\leq$ 1 Å, cf. Table 4.3). The binding instabilities for the bulky compounds can be attributed to their overall size. However, this argument cannot be applied for the light pyrrolidine carboxamides, which are considerably smaller and get docked in a way identical to the reference ligand (pc-d11). Hence, the RMSD analysis of the 5 ns MD simulations warranted a more thorough assessment of the binding orientations with respect to the reference ligand. In the case of bulky pyrrolidine carboxamides, in particular, the dominant binding conformation needed to be ascertained along which the subsequent analyses could be based and interpreted. This was not the case for light pyrrolidine carboxamides, where the preferential orientation can be safely assumed to be identical to the reference ligand (pc-d11). The X-ray crystallographic analysis provides a direct way to ascertain the dominant binding conformations for bulky pyrrolidine carboxamides. The lack of crystal structures for bulky pyrrolidine carboxamides [52] stressed the need for methods that would provide reasonable binding modes for these compounds. A major obstacle to solving this issue was lack of crystal structures for the respective compounds.

An intensive computational method to predict plausible binding modes within the active site is the Replica Exchange Molecular dynamics (REMD) or parallel tempering (in case of MC simulations) [128, 324]. However, due to the complex and time consuming nature of REMD and its derivatives, simpler methods that aided in qualitative comparisons of the different binding conformations of pyrrolidine carboxamides were preferred. The methods and the approach to determine the dominant binding conformation for bulky pyrrolidine carboxamides can be seen in the methods section of this chapter.

An important aspect is the qualitative nature of the movements observed for pyrrolidine carboxamides exhibiting stable and unstable binding, respectively. The nature and extent of the movements exhibited by the ligand are determined by its structure and binding orientation, i.e. the nature of the molecule itself and the way it binds to its target. Thus, a pyrrolidine carboxamide demonstrating stable binding should exhibit particular traits that should be conserved amongst other molecules showing the same orientation/binding stability whilst differing from those exhibiting binding instabilities. Thus, by comparing the movements of the light and bulky pyrrolidine carboxamides with that of the reference ligand (pc-d11), a generalised compound/class specific conformation can be ascertained. Ultimately, this information can aid in better understanding the dynamic behaviour of the pyrrolidine carboxamides that will aid in structure-based optimisation of this series.

7.2 Methods and data analysis

7.2.1 Dihedral analysis

The motions of the ligand within the binding pocket can be ascertained in a qualitative fashion by examining the time dependent distribution of specific ligand torsions coupled with conventional RMSD analysis. Hence, a simultaneous analysis of the dihedral angles and RMSD provides a means of studying the conformational (i.e., internal) as well as translational and rotational motions of the bound ligand with respect to the protein. The dihedral angles encompassing the bonds that connect the central B ring to the other rings were chosen for analysis (cf. Figure 7.1). The time dependent variations and distributions of the dihedral angle α (blue) and β (red) were then plotted simultaneously with the RMSD values to assess the overall movements of the pyrrolidine carboxamides. For a small subset of bulky pyrrolidine carboxamides (cf. Figure 7.2), the location of the dihedral angle α was altered primarily because of a change in the atoms that constituted the respective angle.

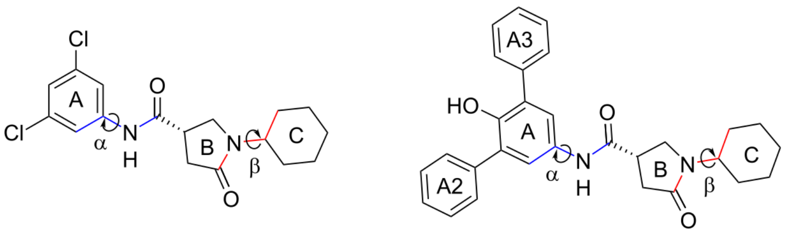
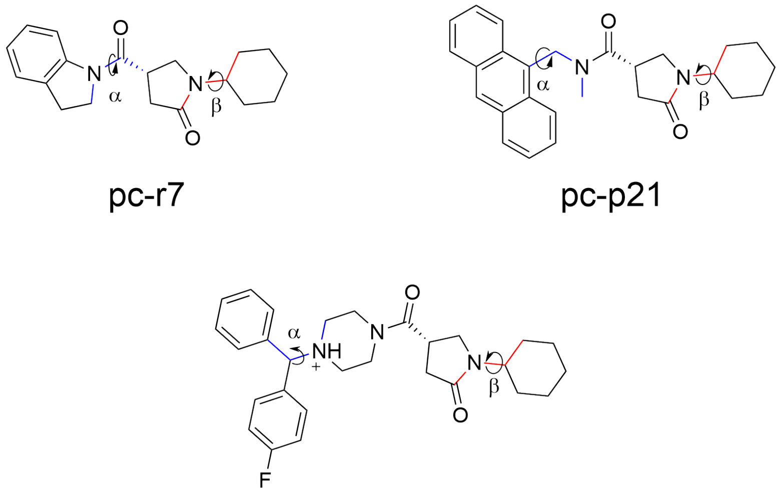


Figure 7.1 Dihedral angles α (blue) and β (red) for "light" and "bulky" pyrrolidine carboxamides, respectively.

For ease of analysis, the light and bulky pyrrolidine carboxamides were analysed separately throughout this entire chapter, with the distribution and time dependent variations in ligand 4TZK (pc-d11) serving as a global reference for comparison. Furthermore, the light pyrrolidine carboxamide subset was segregated into those having single substituents on ring A and those which have a di-substituted phenyl ring.



pc-r7

pc-p21

pc-p28

Figure 7.2 Dihedral angles α (blue) and β (red) for selected "bulky" pyrrolidine carboxamides with altered amide bond linking the rings A and B.

7.2.2 Distance and H-bond analysis

A key factor driving the binding stability of the pyrrolidine carboxamides are the intermolecular non-bonded interactions between the protein and the ligand. These interactions are mainly distance dependent. In the current context, one of the primary factors determining the binding stability is the presence of dual H-bonds between the ligand and the catalytic residue (Y158) as well as the cofactor (NAD$^+$). For the purpose of this analysis, the criteria [325, 326] for hydrogen bonds were set as follows:

1. The distance between the donor (D) and the acceptor atom (A) must be less than 3 Å.
2. The angle spanning D-H-A must be more than 90°.

In case of pyrrolidine carboxamides, there are two carbonyl groups whose oxygen atoms have almost identical chemical micro-environment and can, at least in principle, both form the dual H-bonds mentioned earlier (Figure 7.3). However, as seen from the crystal structures, only the primary carbonyl oxygen designated as 1° (from the B-ring) forms the dual H-bonds, while the other carbonyl group (designated 2°) forms weak and transient H-bonds with residues located near the ligand. In PDB 4TZK, the 2° carbonyl oxygen is quite far from either Y158 or NAD$^+$ donor hydroxyl groups (distances 4.88 Å and 6.73 Å; Figure 7.3) and is thus unable to form H-bonds with either donors. However, if the binding mode exhibits instabilities and thereby movements, the probability of an alternate H-bonding conformation increases. Hence, monitoring the time-dependent variations in distances of both 1° and 2° carbonyl oxygen atoms to the donor atoms represents an ideal supplementary method to analyse the conformational dynamics of the pyrrolidine carboxamides, especially the bulky ones.

In practice, the **hydrogen bond occupancy** or the fraction of the total simulation for which H-bonding was observed was obtained from VMD-1.9.1 with a binary output for each frame. In other words, if an H-bond is observed between any of the donor and acceptor atoms, then a value of 1 is assigned to that frame. In all other cases, a value of zero is assigned followed by computation of the H-bond occupancy. In addition to the H-bond occupancy, the overall nature of the H-bond can be better studied by monitoring the time-dependent changes in the distances of the donor-acceptor atoms. For ease of visualisation, the moving average (designated as MAV; over a sliding window of width 20 frames) of the donor-acceptor atom distances was plotted to effectively visualise the time-dependent shifts in the distances. On parallel lines, the output of the H-bond analysis was also averaged over a sliding window of 20 frames to better visualise the trends in H-bond formation with respect to time.

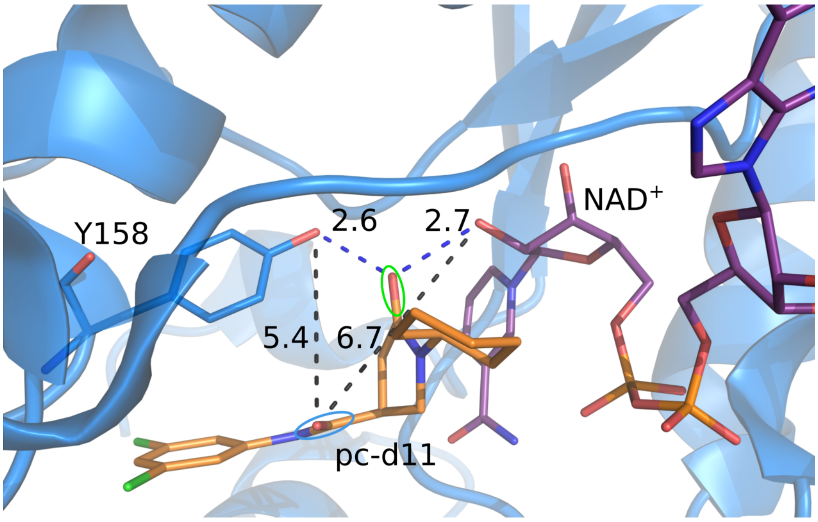


Figure 7.3 Primary (1°) (green circled) and secondary (2°) carbonyl (blue circled) groups for representative pyrrolidine carboxamide-d11, alongwith their distance in Å to the donor atoms of Y158 and NAD$^+$, respectively.

Moreover, the light and bulky pyrrolidine carboxamides were analysed separately, mainly to come up with a compound class specific picture pertaining the overall nature of binding. The primary focus lied upon the determination of the dominant binding conformation for the bulky pyrrolidine carboxamides from amongst the ones obtained from molecular docking as well elucidating the determinants driving the molecular motions. Furthermore, correlation of the structure-activity relationship (SAR) of representative pyrrolidine carboxamides with the fluctuations in dihedral angles as well the RMSD values was also performed. For all of the aforesaid analyses, the 5 ns MD simulations of each protein-ligand complex were extensively utilised. For the crystal structure complexes (N = 5), the GPU accelerated simulations were also analysed simultaneously along their normal counterparts. The raw data for dihedral angles, H-bond occupancies, and donor-acceptor distances were obtained from VMD-1.9.1 [304], while plotting was performed with R [273] and associated ggplot2 and Cairo packages [277]. The molecular visualization and figures were ray-traced with PyMol 1.8 [276].

7.3 Results

The SAR of light pyrrolidine carboxamides has already been amply discussed in literature [50, 52], whilst the dynamics of binding for the entire series has not been published so far. The following are the salient observations of the SAR for light pyrrolidine carboxamides (Figure 7.4), as pointed out by He et al. [52]:

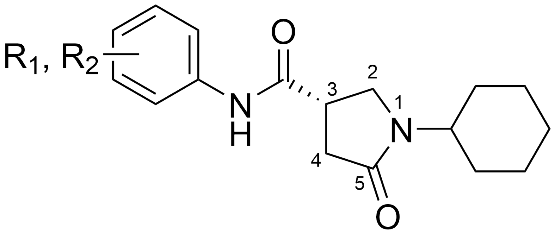


Figure 7.4 Light pyrrolidine carboxamide with annotated B ring heavy atoms.

1. In case of pyrrolidine carboxamides with mono-substituted phenyl ring (A) (s-series, pc-s1 to pc-s17), the order for favourable positional substitutions decreases in the order meta > ortho > para. At both ortho and meta positions, halogens and electron-withdrawing groups, e.g. -CF_3, are well tolerated, while substitution at para position leads to a loss of activity.

2. In case of the di-substituted ring A pyrrolidine carboxamides (d-series, pc-d1 to d16), the favourable positional substitutions decrease in the order meta, meta (3,5 di-substituted) > ortho, meta (2,5 di-substituted only) > ortho, para (2,3 di-substituted) and meta, para (3,4 di-substituted). In all cases, only small electron withdrawing substituents and halogens (F to Br) resulted in activity augmentation, while in all other cases, especially with para substitutions, a significant drop in potency was observed.

3. In regards of the central B ring, the substitution or unsaturation of the bond in between C2-C3 or C3-C4 led to a complete loss of InhA inhibitory activity.

In line with the SAR of pyrrolidine carboxamides, the dihedral angle and the accompanying RMSD analysis as well as the hydrogen bonding analysis mainly attempted to correlate the observable fluctuations of the compounds with their structure and activity, in addition to the original aim of qualitative analysis of their movements.

7.3.1 Dihedral angle and RMSD analysis of light pyrrolidine carboxamides

Before initialising the process of correlating the positional and dihedral drifts of the ligands with their structure and in-vitro activity, it is important to understand the fluctuations exhibited by the reference ligand, i.e. PDB 4TZK ligand (pc-d11). Figure 7.6 depicts the positional and rotational fluctuations of pc-d11 as derived from CPU (pmemd) simulations using AMBER 12. In Figure 7.6, the left graph depicts the narrow distribution of the dihedral angles α and β. The relatively low number of values (small red dots) that lie outside the grey density contours suggest sporadic movements of both A and C rings of pc-d11. The dihedral density distribution plot does not reflect its time dependence. Hence, a simple line plot (middle plot) was utilised to depict the time dependent fluctuations of the individual dihedral angles, while the corresponding structural drifts of the ligand and the protein are highlighted in the RMSD plot (right graph).

The line plots for the dihedral angles clearly highlight the lack of significant movements, most likely owing to the dual H-bonds with Y158 and the cofactor that stabilise the binding mode in addition to the van der Waals interactions with residues of the active site. The stable binding of pc-d11 is additionally highlighted by lack of any noticeable fluctuations in C-α and bound ligand RMSD values. Taking a cue from the reference ligand, compounds exhibiting stable binding must show a narrow distribution of dihedral angles as well as low drifts in RMSD values. The next sections discuss the specific examples and outliers for light as well as bulky pyrrolidine carboxamides alongwith a generalised correlation amongst their observed fluctuations and their measured activity in-vitro.

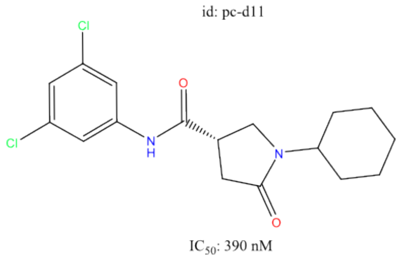


Figure 7.5 Structure of pc-d11.

7.3.1.1 Mono-substituted light pyrrolidine carboxamides

In case of the mono substituted pyrrolidine carboxamides, the ligand from PDB 4UOJ (pc-s1) (Table A.2) with its unsubstituted phenyl ring (ring A) was chosen as a reference to highlight the effects of substitution on the overall activity and binding dynamics. Figure 7.7 depicts the dihedral angle distribution and RMSD drifts for the 5ns MD simulation of

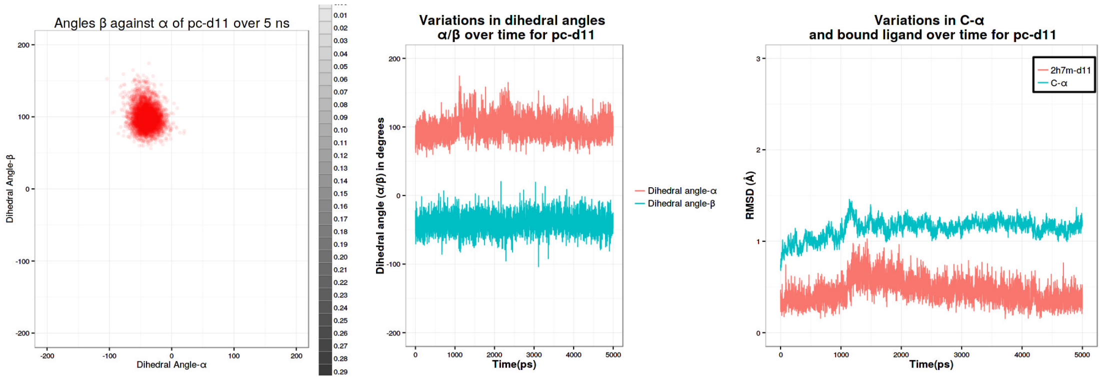


Figure 7.6 Dihedral angle distribution and RMSD fluctuations for pc-d11 over 5ns NPT CPU (pmemd) simulation.

pc-s1 (pmemd simulation). By comparing Figure 7.6 with Figure 7.7, the stabilising effect of two halogens at the meta position becomes evident, although this has clearly been ruled out by He et al. [52]. A closer inspection of the crystal structure poses revealed that the H-bonding strength could play an important role in stabilising the binding mode.

Indeed, this was the case as can be seen later (cf. Section 7.3.3), wherein pc-d11 retained its dual H-bonds for a longer period of time as compared to pc-s1 (Table 7.1). The net effect is evident in the dihedral RMSD plots, wherein pc-s1 exhibits a noticeable change in conformation at around 1ns (phenyl ring rotation) that leads to a cascading effect and finally a large change in the overall C-α RMSD. Surprisingly though, the bound ligand RMSD remains stable throughout the simulation indicating that there are several factors in addition to hydrogen bonds that play a role in overall stabilisation of the binding mode. For a group-wise assessment of light pyrrolidine carboxamides based on the position of their substituents and observable fluctuations in RMSD as well as dihedral angles, the compounds were segregated into the following groups:

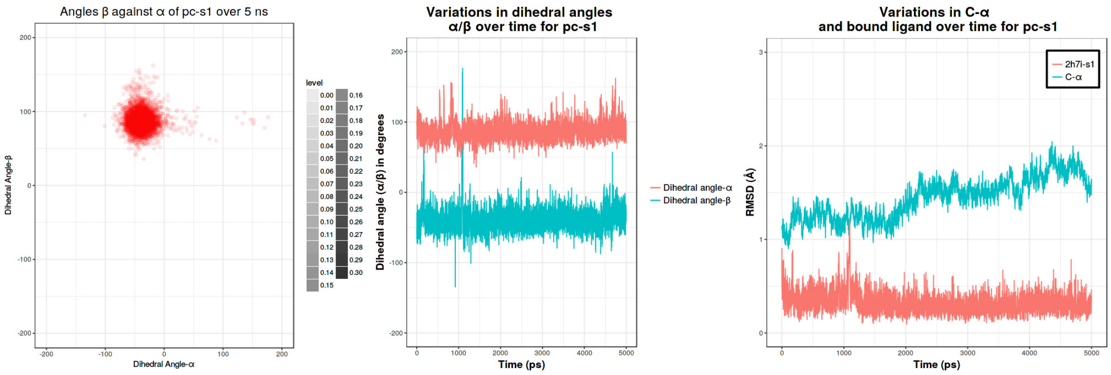


Figure 7.7 Dihedral angle distribution and RMSD fluctuations for pc-s1 over 5ns NPT simulation.

1. **2' (ortho) substituted pyrrolidine carboxamide:**
   This group of monosubstituted light pyrrolidine carboxamides is represented by pc-s2 (2-$COOCH_3$, IC_{50}: 34.88 μM) and pc-s3 (2-Br, IC_{50}: $>$ 100 μM). It clearly highlights the unfavourable effects of a large substituent on the ortho position of ring A. The bulky ortho substituent clashes with the side chains of the pocket residues leading to increase in the C-α RMSD (Table A.2). Nevertheless, the boxplots for the dihedral angles and RMSD for the bound ligand suggest a stable "in-place" movement of pc-s2 (Figures 7.8 and 7.14).

2. **3' (meta) substituted pyrrolidine carboxamides:**
   The meta substituted light pyrrolidine carboxamides represent favourable effects in regards of InhA inhibitory activity and binding stability. Of all electron withdrawing substitutions, halogens (except iodine) and trifluoromethyl (-CF_3) group are well tolerated [52], while increasing the bulkiness of the meta substituent leads to a decrease in activity up to two orders of magnitude (Table A.2). In regards of the dynamic aspects, the crystal structure ligand pc-s4 (PDB 4TRJ) serves as a reference for this subgroup and highlights the overall low movements of the ligand within the binding pocket (Figure 7.9). Nevertheless, the C-α RMSD still rises to 2 Å.

   Moreover, as the bulkiness and nature of substituent varies, changes are observed in the distribution of the dihedral angles as well as the C-α and bound ligand RMSD drifts (Figures 7.8, 7.13 and 7.14). Thus, pc-s4 with 3-bromo substituent on ring A barely shows any movement of the ligand (Figure 7.9). This is also reflected in the respective box plots for the dihedral angles (Figure 7.8) and RMSD values (Figures 7.13 and 7.14) for pc-s4. However with an increase in the rotatable bonds in the substituent, as was the case with pc-s11 (3-CF_3) and pc-s15 (3-$CH(CH_3)_2$), marked changes in both dihedral angles and bound ligand RMSD can be observed.

   However, with an increase in the size of the substituent, as was the case with pc-s11 (3-CF_3) and pc-s15 (3-$CH(CH_3)_2$), marked changes in both dihedral angles and bound ligand RMSD can be observed. Furthermore, the overall low values of the bound ligand RMSD ($\leq$ 1 Å) imply low movements of the ligands within the binding pocket. Most of the observed motions for the bound ligands stem from the rotation of the phenyl ring around the covalent bonds (C-N) linking the rings A and B as well as B and C, respectively.

   The compound pc-s6 that despite a favourable 3-chloro substituent shows marked departure from the dihedral distribution of meta monosubstituted pyrrolidine carboxamides. Here, the entire molecule exhibits destabilised binding though it satisfies the stability criterion as described in Chapter 4. The destabilised in-place motions of the bound ligand stem from the ring movements described earlier (i.e.,

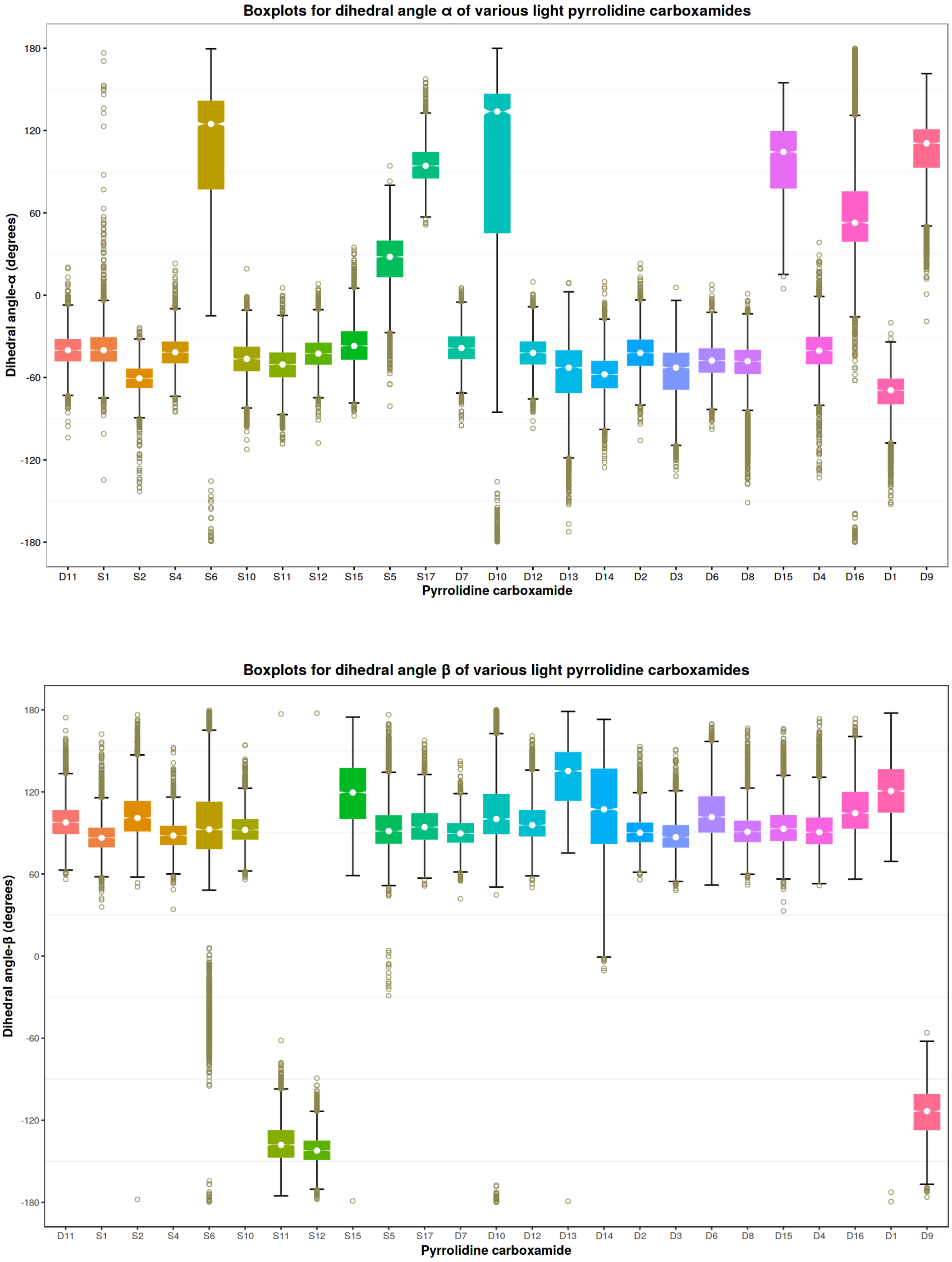


Figure 7.8 Boxplots depicting the **(a)** dihedral angle-α (top), and the **(b)** dihedral angle-β (bottom) for light pyrrolidine carboxamides, with the reference ligand d11 at the pole position. The median for the distribution of the dihedral angles is depicted as a white circle. The compounds are sorted according to the position of the substituent on A ring, with the monosubstituted pyrrolidine carboxamides (S suffix) preceding the disubstituted ones (D suffix). Furthermore, the order for the monosubstituted compounds is ortho (s2), meta (s4-s15), and finally para (s5,s17) substituted compounds. On the other hand, the disubstituted compounds begin with meta disubstituted (d7-d14) followed by ortho-meta (d2-d15), meta-para (d4, d16), and finally ortho-para (d1, d9) compounds.

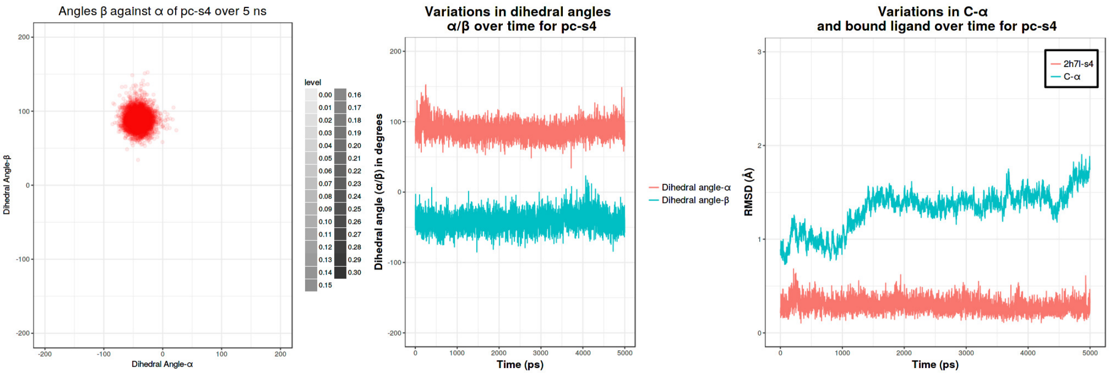


Figure 7.9 Dihedral angle distribution and RMSD fluctuations for pc-s4 over 5ns NPT simulation.

rotation of ring A). This also can be seen in Figure 7.10, where a steep rise in the bound ligand RMSD can be observed at the end of the simulation. On the other hand, for compounds pc-s11 (3-CF_3) and pc-s12 (3-NO_2), the destabilised ”in-place” motions stem from the movements of the ring C. This leads to marked ”in-place” motions that are supported by the deviations of the dihedral angle β (Figure 7.8) as well as bound ligand RMSD distribution (Figure 7.14). Moreover, another feature of this group is that with the exception of pc-s4 (3-Br), the binding orientations of pc-s6 (3-Cl), pc-s10 (3-CH_3), pc-s11, pc-s12 (3-NO_2) and pc-s15 that were subjected to MD simulations were obtained by in-situ mutation of pc-d11 (Section 3.3.4).

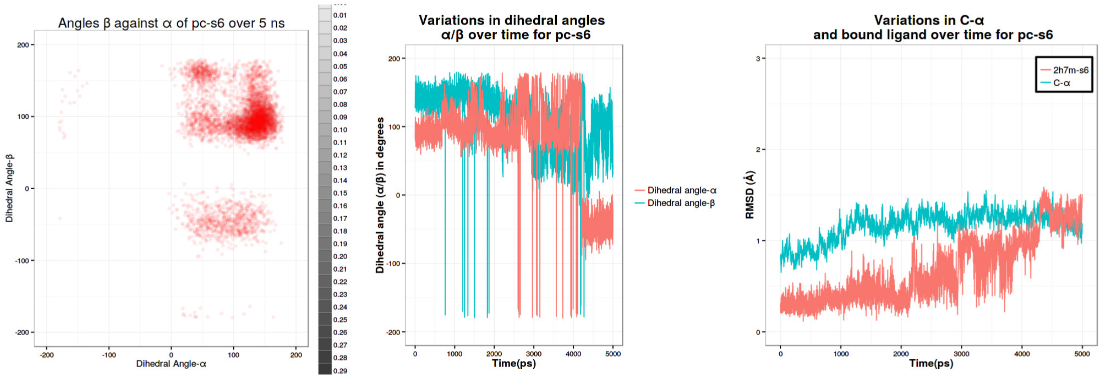


Figure 7.10 Dihedral angle distribution and RMSD fluctuations for pc-s6 over 5ns NPT simulation.

3. **4’ (para) substituted pyrrolidine carboxamides:**

The para substitution of the A ring (phenyl) clearly highlights the detrimental effects of the substitution, irrespective of the nature of the substituent, that is manifested in a notable decrease in InhA inhibition (e.g.: pc-s17, 4-Ac, IC_{50}: 73.58 μM). In regards of the docking pose, the para substituent comes in close contact with the backbone of P156 (Figure 7.12) which is involved in stabilisation of the ligand binding primarily via van der Waals interactions. Increased close contacts like the aforementioned case primarily result in unstable binding that is clearly visible in the huge drifts of dihedral angles as well as RMSD values (Figures 7.8, 7.11, 7.13 and 7.14). A similar trend is seen in case of pc-5 (4-Br) and pc-s9 (4-I), although pc-s9 was not considered for MD simulations.

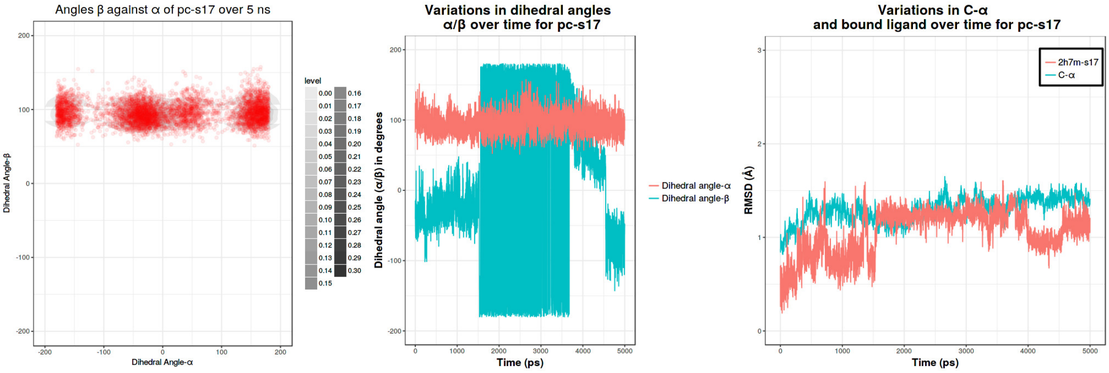


Figure 7.11 Dihedral angle distribution and RMSD fluctuations for pc-s17 over 5ns NPT simulation.

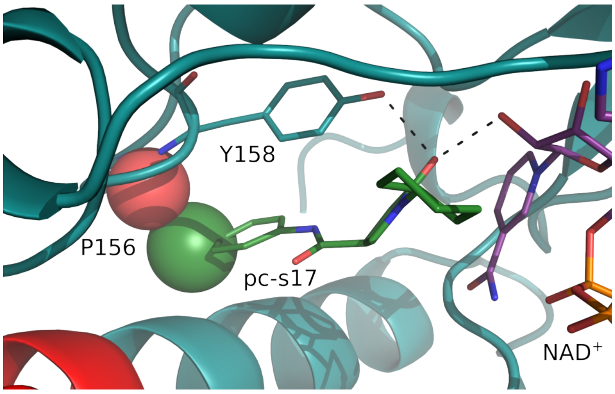


Figure 7.12 Close contact of the carbon atom from the acetyl substituent (-C=O group) of pc-s17 (green sphere) with the backbone oxygen of P156 (red sphere).

7.3.1.2 Di-substituted light pyrrolidine carboxamides

The di-substituted light pyrrolidine carboxamides encompass some of the most potent compounds from the entire series, including the reference compound pc-d11 (3, 5-Cl). Just like in case of their mono-substituted counterparts, the disubstituted compounds can be sorted in decreasing order of activity and a concurrent order of binding stability as follows:

1. **meta-meta disubstituted pyrrolidine carboxamides:**
   The substitution of the A ring at positions 3 and 5 has a strong favourable effect on the measured InhA inhibitory activity that is evident in the low IC_{50} of all

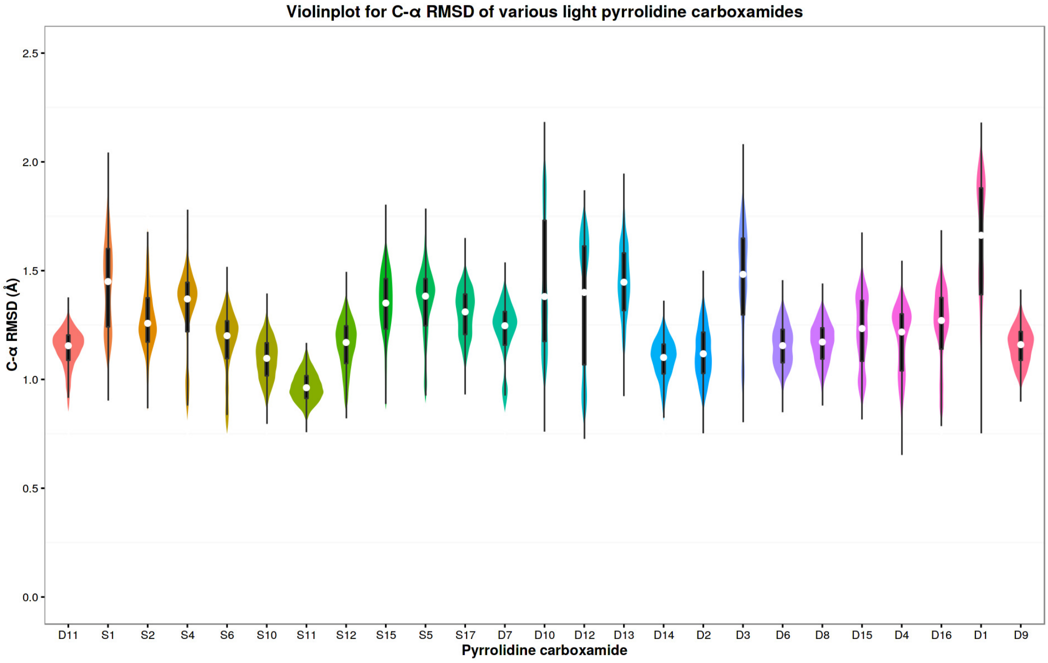


Figure 7.13 Violin plot depicting the kernel density distribution of C-α RMSD for light pyrrolidine carboxamides, with the order of compounds being described in Figure 7.8. The thick central bar depicts the interquartile range, white circle denotes the median for the distribution. Furthermore, a dip in the smooth shape of the violin indicates a steep change in the C-α RMSD values.

members of this subgroup (Table A.2). Upon close inspection of the dihedral angle distribution for this group (d7 - d14, Figure 7.8), it can be seen that d10 and d14 are the outliers, whilst the dihedral angles for other compounds remain more or less around the median value of pc-d11 (dihedral angle α = -40° and β = 80°). The compound pc-d10 (3, 5 fluoro) behaves quite similar to pc-s6 and likewise, the free rotation of ring A rather than ring C contributes to the comparatively higher RMSD values (Figures 7.13 and 7.14). On the contrary, for compounds with bulky meta substituents, i.e., pc-d13 (3-OCH_3, 5-CF_3) and pc-d14 (3,5-CF_3) the C ring movement contributes to the overall ligand RMSD. If one observes the structure of these two molecules, it becomes evident that there is a size limit on the group that can be substituted at the 3' position, while having a halogen like chlorine or bromine and groups like trifluoromethyl at 5' position is always beneficial.

2. **ortho-meta disubstituted pyrrolidine carboxamides:**

   The ortho-meta substituted light pyrrolidine carboxamides (d2 - d15) point out the favourable substitutions on the A ring, with almost all molecules exhibiting low IC_{50} values (Table A.2). The exception to this are the compounds pc-d2 (IC_{50}: 56.50 μM) and pc-d6 (IC_{50}: 10.05 μM). In both cases, both dihedral angle distribution and RMSD values do not provide any explanation for their low InhA inhibitory activity. However, He et al. suggest that the ortho and para positions

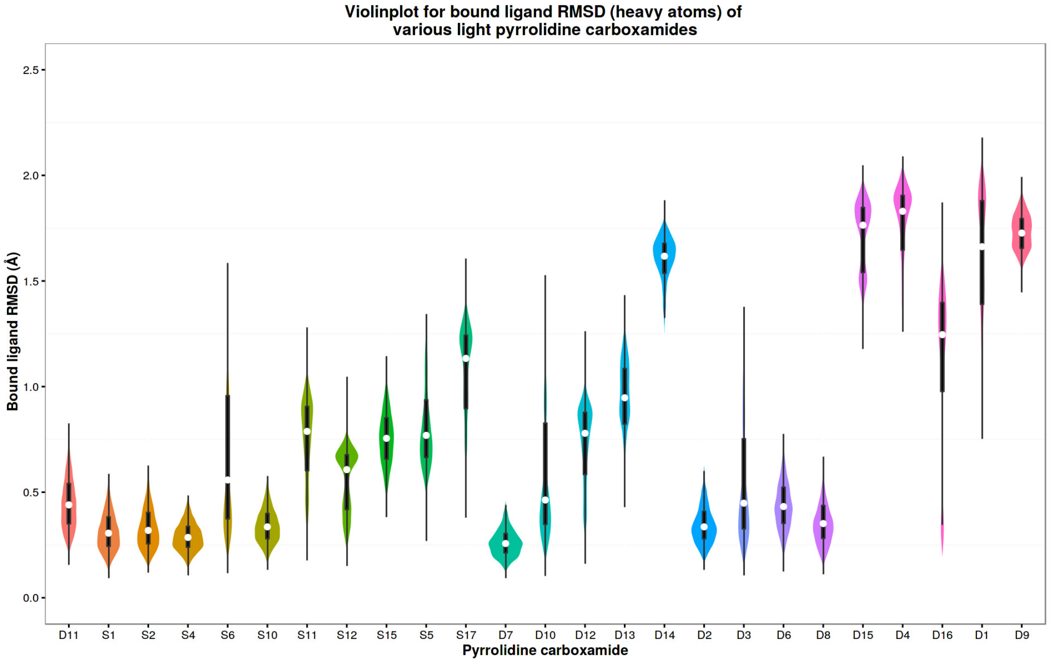


Figure 7.14 Violin plot depicting the kernel density distribution of bound ligand RMSD (heavy atoms only) for light pyrrolidine carboxamides, with the order of compounds being described in Figure 7.8. The thick central bar depicts the interquartile range, white circle denotes the median for the distribution. Furthermore, a dip in the smooth shape of the violin indicates a steep change in the ligand RMSD.

are generally unfavourable for substitution [52]. This can explain the low activity for these compounds.

Moreover, the destabilising effect of a bulky ortho substituent can be clearly seen in case of pc-d15 (2-OCH$_3$, 5-Cl, IC$_{50}$: 1.69 μM). It exhibits an almost two-fold reduction in the InhA inhibitory activity as compared to pc-d3 (PDB 4UOK ligand; 2-CH$_3$, 5-Cl; IC$_{50}$: 0.97 μM). Both molecules differ by only a single oxygen atom (of the -OCH$_3$) at 2' position. The destabilised binding of pc-d15 is clearly evident in its distribution of the dihedral angle α (cf. Figure 7.8) and bound ligand RMSD (cf. Figure 7.14). Going by the activity trends, a halogen substituent (preferably chlorine) or a methyl/trifluoromethyl group at 5' position are associated with increased InhA inhibitory potential.

3. **meta-para and ortho-para disubstituted pyrrolidine carboxamides:** Molecules that represent this subgroup of light pyrrolidine carboxamides (d4 - d9) clearly highlight the detrimental effects of para substituent as well as placing bulky groups near each other (ortho-meta) that results in diminished InhA inhibitory activity (Table A.2). The ortho-para disubstituted compounds (d1 and d9), in particular, not only exhibit significantly diminished InhA inhibitory potential, but also destabilised binding, as seen from the distribution of their dihedral angles (α and β) as well as bound ligand RMSD (cf. Figures 7.8 and 7.14). The compound pc-d9 (2-CH$_3$, 4-NO$_2$) and its destabilised binding primarily arise from the motions

of the phenyl ring rather than the C ring, which results in a distinct distribution of dihedral α (cf. Figure 7.8). The distribution of the dihedral angle β for pc-d9 is also distinct with respect to other pyrrolidine carboxamides, and is quite identical to pc-s11 and pc-s12. In all these three cases, the pseudo symmetry of the ring C results in a different value for the dihedral angle β, inspite of the same atoms defining the angle.

The meta-para disubstituted pyrrolidine carboxamides (d4 and d16) represent another case of unfavourable substitution that results in significant loss of InhA inhibitory activity. The presence of a bulky methyl group at 3' together with a big halogen (bromine) atom at the 4' position of the A ring (pc-d4) results in close contacts. The presence of a bromine atom at the para position results in clashes with Pro156 much like pc-s17 (Figure 7.12). The net result is destabilisation of the ligand binding and a four-fold decrease in InhA inhibitory activity as compared to pc-d11. Moreover, a common fact pertaining to all these molecules (except pc-d16) is that their poses were predicted by in-situ mutation of pc-d11 (cf. Section 3.3.4).

From the aforesaid observations, the following things can be said regarding SAR as well as the qualitative nature of the dynamics of light pyrrolidine carboxamides:

1. The meta positions of the A ring represent the most favourable for substitution with small electron withdrawing groups and halogen atoms except iodine. The meta position is followed by ortho while para substitution leads to diminished activity. A 5-chloro substituent was observed to exert favourable effects on the overall activity as well as binding stability that is in line with the general observation made by He et al. [52].

2. The overall dynamics being exhibited by the light pyrrolidine carboxamides closely follows the nature and position of substituents on the A ring, with a clear size limit on the substituent. This can be seen in the cases of light pyrrolidine carboxamides with noticeable destabilised binding. For example, pc-d10 (3,5-F) exhibits more fluctuations than pc-d11 (3,5-Cl). Furthermore, pc-d13 (3-OMe, 5-CF_3) exhibits more fluctuations than pc-d11. In the former case, pc-d10 with its 2 fluorine substituents forms much weaker halogen bonds with the active site residues as compared to the chlorine substituents of pc-d11. The halogen bonds represent a barrier to rotation for the phenyl ring which being weak in case of pc-d10 lead to a free rotation of the phenyl ring as compared to pc-d11. In the latter case (pc-d13), both chlorine atoms at 3 and 5 positions of pc-d11 are replaced by methoxy and trifluoromethyl groups. The methoxy group is quite bulky as compared to a chlorine atom and lacks the ability to form halogen bonds. Furthermore, the trifluoromethyl group at 5' position can form a very weak halogen bond with the neighbouring

active site residues. A combination of these two is manifested as decreased InhA inhibitory potential and increased destabilisation in binding as compared to pc-d11.

3. Summing up, all of the aforesaid analyses shed light on the movements of light pyrrolidine carboxamides in their bound state. In case of compounds with low InhA inhibitory activity, a low barrier to rotation of the ring A was associated with increased destabilisation. This gave rise to characteristic dihedral angle distribution as well as high RMSD values for the ligand and protein alike. On the contrary, for compounds with high InhA inhibitory activity, the ring A motions were mostly subdued as evident from the dihedral angle distribution. This was associated with low drifts in both C-α and bound ligand RMSD, respectively.

7.3.2 Dihedral angle and RMSD analysis of bulky pyrrolidine carboxamides

As seen from Chapter 3, a total of three binding modes were reported for bulky pyrrolidine carboxamides, with most of the ligands exhibiting a crystal structure like conformation. Of the 18 bulky pyrrolidine carboxamides, a total of 5 molecules (pc-c6a3, pc-p27, pc-p28, pc-p36, and pc-p37) exhibited alternate binding modes (Figure 3.21), while an additional 4 molecules (pc-r7, pc-p31, pc-c7a3, and pc-c8a2) (Table 4.2) exhibited stable binding. A common feature amongst the molecules exhibiting stable binding was that they got docked in an orientation similar to the reference ligand viz. pc-d11, and consequently exhibited very low fluctuations in either of the dihedral angles and the RMSD values (Figures 7.15 to 7.17), although with some exception as seen in Section 7.3.2.2.

On the contrary, the rest of the bulky pyrrolidine carboxamides exhibited both "in-place" movements as well as within the binding pocket. The observed drifts in both bound ligand RMSD as well C-α RMSD stayed well below 2 Å, indicating limited movements within the binding pocket. The drifts in C-α RMSD and bound ligand RMSD variably coincide that warrant a more thorough assessment of the underlying binding modes. Accordingly, the three binding modes for bulky pyrrolidine carboxamides (Figure 3.21), their associated dihedral angle distribution as well RMSD values were assessed to derive "binding mode" specific movements that in turn correlated with the structure of the molecule. This would aid in better understanding the binding dynamics for this series of pyrrolidine carboxamides.

7.3.2.1 "Reference ligand binding mode"

This binding mode of the reference ligand (pc-d11) is exhibited by the majority of the bulky pyrrolidine carboxamides (all except pc-c6a3, pc-p27, pc-p28, pc-p36, and pc-p37)

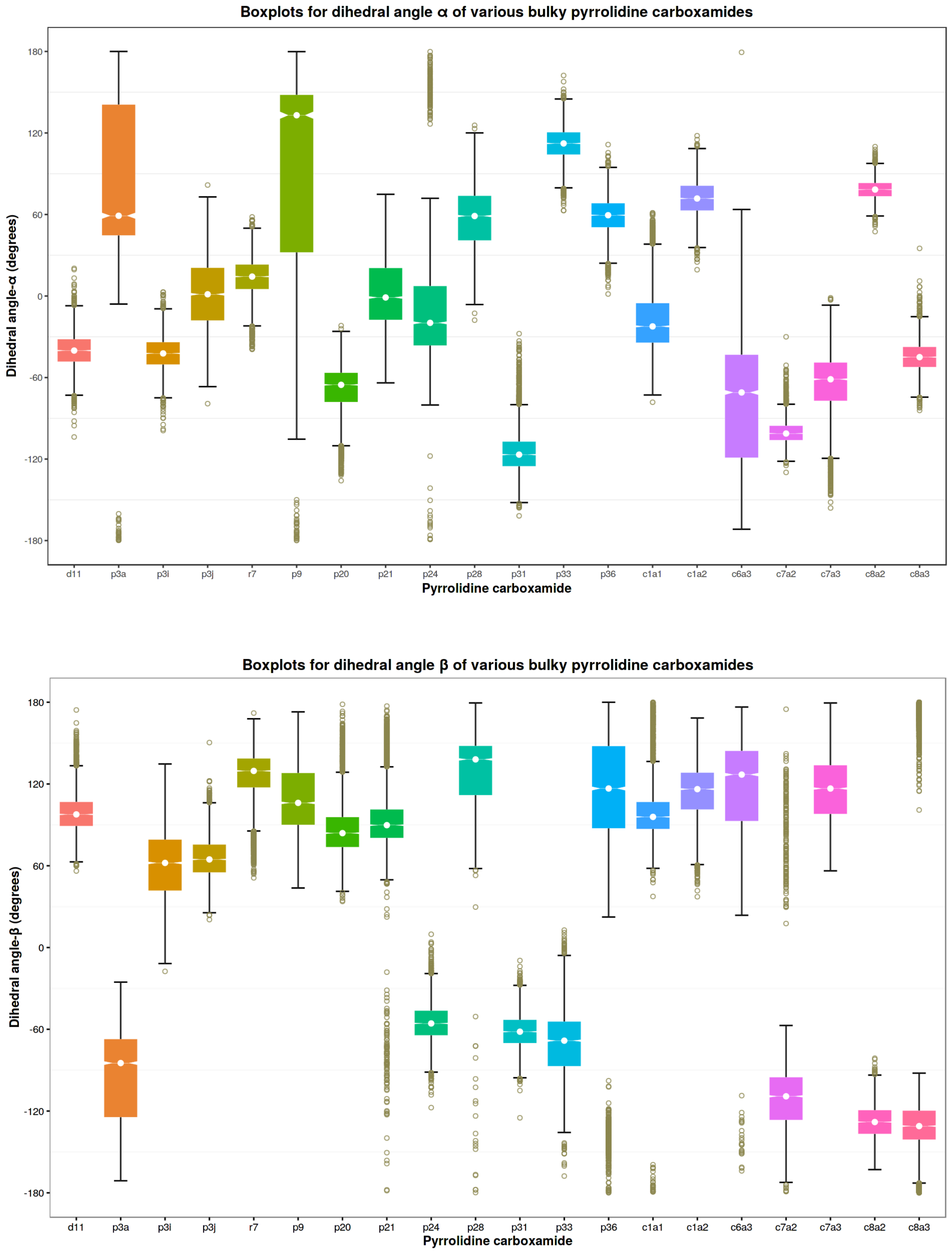


Figure 7.15 Box plots depicting the **(a)** dihedral angle-α (top), and **(b)** dihedral angle-β (bottom) for bulky pyrrolidine carboxamides, with the reference ligand d11 at the pole position. The median for the distribution of the dihedral angles has been depicted as a white circle.

and has been depicted in Figure 3.21. It signifies proper placement of the ligand and thereby increased interactions with the active site residues that stabilise the binding mode. The dihedral angle distribution and the RMSD fluctuations are summarised in Figure 7.15.

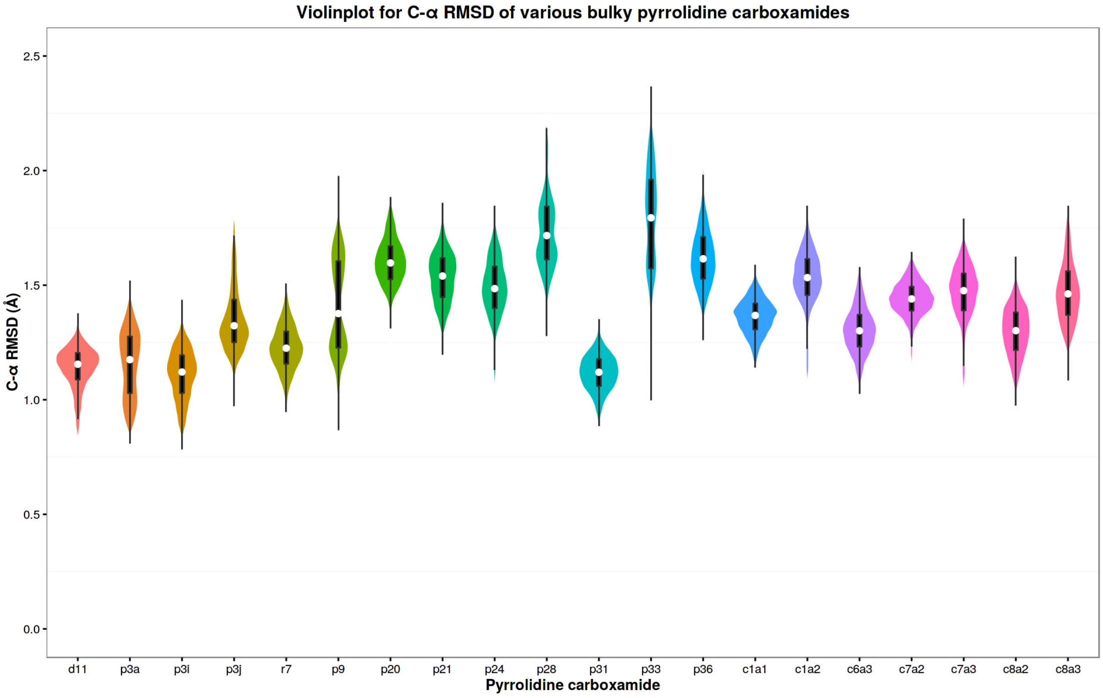


Figure 7.16 Violin plot depicting the kernel density distribution of C-α RMSD for bulky pyrrolidine carboxamides, with the order of compounds being described in Figure 7.15. The thick central bar depicts the interquartile range, white circle denotes the median for the distribution. Furthermore, a dip in the smooth shape of the violin indicates a steep change in the bound ligand RMSD value.

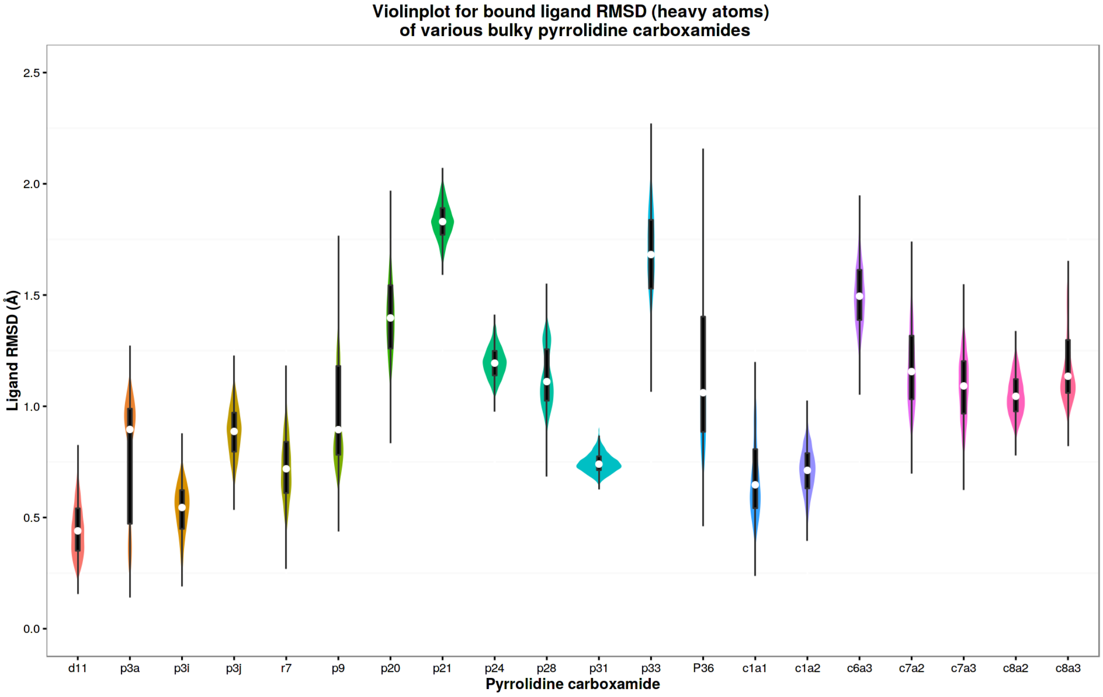


Figure 7.17 Violin plot depicting the kernel density distribution of bound ligand RMSD (heavy atoms only) for bulky pyrrolidine carboxamides, with the order of compounds being described in Figure 7.15. The thick central bar depicts the interquartile range, white circle denotes the median for the distribution. Furthermore, a dip in the smooth shape of the violin indicates a steep change in the bound ligand RMSD value.

In regards of the dihedral angle α, a wide variety in the distribution was observed for the entire bulky pyrrolidine carboxamide subset. The wide variety in distributions indicates a differential barrier to rotation of the A ring system in the respective compounds. The wide variation in α can be attributed to the numerous types of heterocyclic systems replacing the phenyl ring of light pyrrolidine carboxamides. Some of the important factors to be considered while analysing these variations are the structure and binding mode of the ligand under inspection. In case of the other dihedral angle, i.e. β, the variation was not as wide as in case of α. This suggests that the majority of the structural motions in case of bulky pyrrolidine carboxamides happens in the A ring system, much like in case of light pyrrolidine carboxamides.

For the representative bulky pyrrolidine carboxamide-c7a3 and other compounds pc-c1a1, pc-c1a2 and pc-c6a3 to pc-c8a3, there is a huge barrier to rotation owing to the size of the A ring. This is evident from the narrow distribution of α for these compounds (Figure 7.15). The only exception to this is the compound pc-c6a3 which got docked in an alternate conformation (Figure 3.21) and will be discussed separately in Section 7.3.2.3. On the other hand, the dihedral angle β for the respective compounds did not exhibit any noticeable variation. The change in median values of β for pc-c8a2 and pc-c8a3 can be attributed to the pseudo-symmetry of the cyclooctyl ring. However, in case of pc-c7a2, a ring flip leads to the change in the median value. These results and the information from literature clearly suggest that a cyclohexyl or a phenyl ring is most favourable on the side of the ligand facing the substrate binding loop.

For a total of three compounds, i.e., p3a, p3i, and p3j, the cyclohexyl (C ring) was replaced with a mono/di-substituted phenyl ring. As such these compounds are light pyrrolidine carboxamides, but are discussed here mainly because of the changed C ring that makes comparisons with both bulky and light pyrrolidine carboxamides easier. As a result of the C ring replacement, both ends of these molecules were expected to have free rotations. This was indeed the case, though p3i exhibited an almost identical distribution of α as compared to pc-d11. For the remaining bulky pyrrolidine carboxamides, i.e., pc-r7 to p36, there were noticeable changes in the dihedral angle distributions. Of these, 4 compounds (pc-p27, pc-p28, pc-p36, and pc-p37) exhibited another atypical binding conformation (Figure 3.21) and will be discussed separately in Section 7.3.2.2.

Of the remaining compounds, i.e., r7, p9, p20, p21, p24, p31, and p33, the distribution of dihedral angle α for r7 and p9 differed significantly. For r7 (Figure 7.2), the dihedral angle had to be modified. However, the barrier to rotation of the indole ring around the C-N bond was much higher than for a simple phenyl ring. Moreover, the binding mode appeared to be stabilised by its interactions with the active site residues. This resulted in a narrow distribution for the dihedral angle α with a median value higher than that of pc-d11. The compound pc-p9, on the other hand, exhibited a 1-benzyloxy-benzene system replacing the normal phenyl ring of light pyrrolidine carboxamides. Inspite of the increase in the overall size of the A ring, the increased range of the dihedral distribution indicates lowering of the barriers to rotation of the A ring system. Furthermore, the pyrrolidine carboxamide-p31, with a fluorene ring replacing the A ring, showed a high barrier to rotation that was evident from its dihedral angle distributions.

On the other hand, the dihedral angle β did not vary noticeably for the aforesaid compounds. All of the aforementioned compounds contain a cyclohexyl ring just like the light pyrrolidine carboxamides. The visible outliers for the dihedral angle β are the pyrrolidine carboxamides p3a, p24, p31, and p33. In all cases, the trajectories did not reveal any ring flips and consequently, the varying distribution of the dihedral β can be attributed to the pseudosymmetry of the C ring.

Apart from the hydrophobic interactions, hydrogen bonds also play an important role in the stabilisation of the binding mode. This is more evident in case of compounds with 4-amino-2,6-diphenylphenol replacing the A ring (compounds with a3 suffix, e.g., pc-c7a3; Table A.4), wherein the hydroxy group of the central phenol ring was found to form an additional H-bond with Proline-156 (Figure 7.19). This additional H-bond (the other two being with Y158 and the cofactor) can only be formed if the ligand binding mode resembles that of the reference ligand (Figure 3.21). For ligands exhibiting alternate binding modes, the lack of ability in forming additional H-bonds is clearly manifested in the decreased InhA inhibitory activity apart from exhibiting noticeable drifts in dihedral angles and RMSD. All of these facts support the notion that this binding mode might be the preferred by bulky pyrrolidine carboxamides over the other two modes depicted in Figure 3.21.

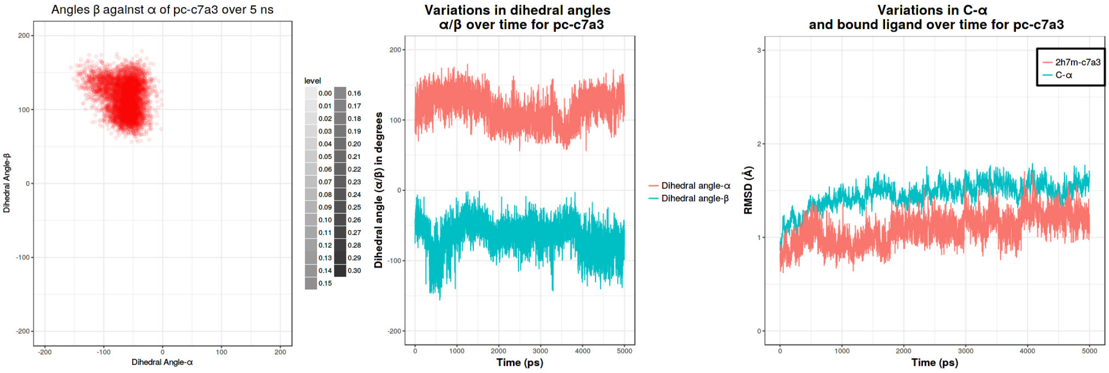


Figure 7.18 Dihedral angle distribution and RMSD fluctuations for pc-c7a3 over 5ns NPT simulation.

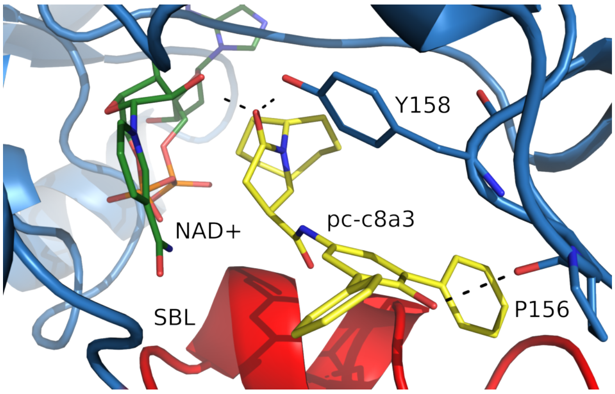


Figure 7.19 Pyrrolidine carboxamide-c8a3 binding in reference ligand like conformation; H-bonds are shown as black dashes while the substrate binding loop is shown in red and cofactor as dark green sticks. Viewpoint from the protein interior (minor exit portal) to the major exit portal (protein exterior).

7.3.2.2 ”Alternate binding mode-1”

This binding mode was observed for a small number (N=4) of structurally similar molecules (pc-p27, p28, p36, and p37; Table A.4). All of these molecules exhibit a piperazine ring that is connected to the B ring containing the 1° carbonyl group and a substituted bulky 1,1 diphenyl ethane moiety. For these compounds the 2° carbonyl group forms the dual H-bonds instead of the carbonyl from the B ring, with the B ring being accommodated in the binding pocket (Figure 7.20). Furthermore, the tertiary amine nitrogen connecting the piperazine ring to the 1,1 diphenyl ethane system is non-protonated at physiological pH as reported by He et al. [52] as well as PROPKA [265, 266] and MoKa 2.6.0 [327]. However, given the slightly acidic pH within the mycobacteria containing phagolysosome (pH 4.0-6.0) [328], the possibility of these ligands being converted to their protonated form is much higher (80% at pH 5.5, Figure 7.21). Hence, their protonated form was also considered and docked in InhA. The protonated forms got selected via the pose selection process elucidated in Chapter 3. Of the aforementioned four molecules, pc-p27 and pc-p37 had to be excluded from the MD simulations owing to structural issues (intramolecular close contacts) that lead to instabilities and failure of the trajectory output.

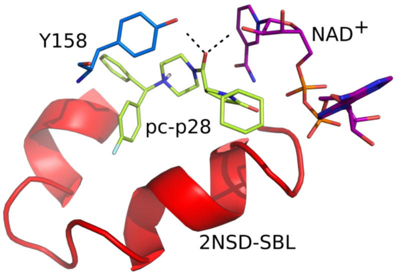


Figure 7.20 Alternate hydrogen bonding conformation of pc-p28. The hydrogen bonds formed by the 2° carbonyl group appear as black dashes while the substrate binding loop is coloured red.

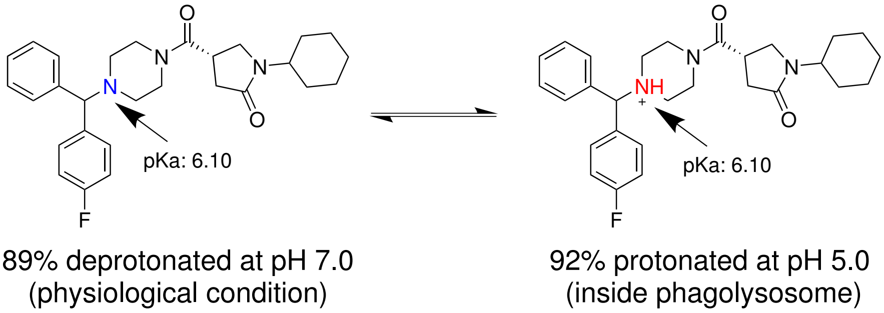


Figure 7.21 Protonated and deprotonated states as predicted by MoKa for pc-p28 at pH of 5.0 and 7.0, respectively. The nitrogen prone to protonation has been shown in blue and red, respectively.

As described in the methods section, the position and the atoms involving dihedral angle α were changed as compared to the rest of the bulky pyrrolidine carboxamides. The alternate compounds exhibiting the aforementioned mode were associated with:

1. **Increased RMSD:**
   Both RMSD values, i.e., bound ligand RMSD and C-α RMSD tend to be on the higher side with much more noticeable jumps in both as compared to the molecules exhibiting a crystal structure like binding orientation (cf. Figure 7.17). The instability associated with this binding mode can be seen in Figure 7.22, where there is a rise in C-α RMSD that coincides with a sharp drop in ligand RMSD for pc-p28 as well as a sharp drop in the dihedral angle α. The rising C-α RMSD as well as the unique dihedral distribution suggests significant motions taking place in the binding site as well as the bulky portion of the ligand.

2. **Low variation in dihedral α:**
   The alternate binding mode represents a non-optimal orientation of the ligand and thereby many of the stabilising interactions with the active site residues are expected to be disrupted. This is accompanied alongwith increased motions of the bound ligand. However, as seen from Figure 7.15, both pc-p28 and pc-p36 exhibit a narrow distribution for the dihedral α. This indicates subdued rotations of the phenyl rings from the substituted 1,1 diphenyl ethane moiety. The subdued motions of the phenyl rings can be attributed to the multipolar interactions of the fluorine substituents with the residues of the active site and the hinge region. Moreover, there is a possibility of weak halogen bonding coming into play.

3. **Wide variation in the dihedral β:**
   The non-optimal orientation of the ligand also leads the ring C to occupy a place that is seldom occupied by it. As a result, the ring C of pc-p28 and p36 alike exhibits wide variations in its motions that manifests a broad range of dihedral angle β alongwith noticeable increase in the number of outliers (Figure 7.15).

4. Summing up, this atypical binding mode is likely to be less stable as compared to the crystal structure pose as evident from the dihedral angle distributions and the RMSD values for the representative compounds exhibiting this mode.

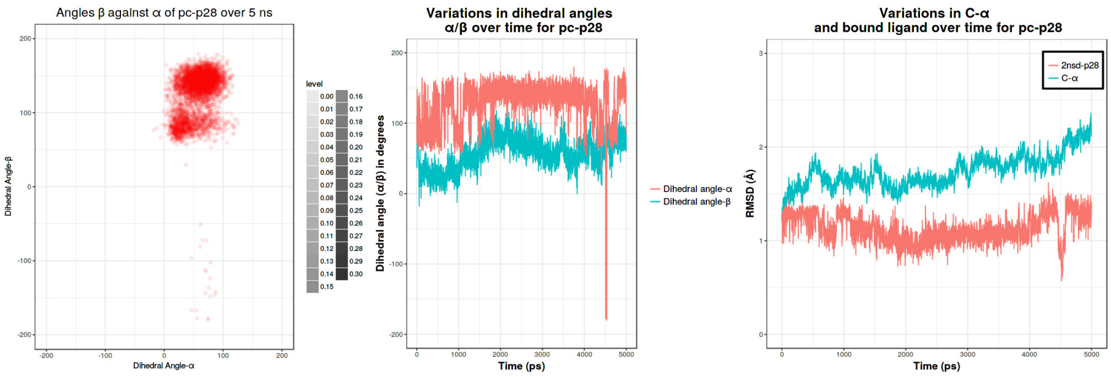


Figure 7.22 Dihedral angle distribution and RMSD fluctuations for protonated form of pc-p28 over 5ns NPT simulation.

7.3.2.3 ”Alternate binding mode-2”

This binding mode with a complete inversion of the crystal structure orientation (Figure 3.21) was solely observed for the most potent pyrrolidine carboxamide-c6a3 (IC_{50}: 140 nM). In this mode the bulky A ring faces the exterior, while the C ring partially penetrates into the binding pocket (Figure 7.23). The 1° carbonyl group forms the usual dual hydrogen bonds with the cofactor and Y158, respectively. The binding mode representing partial penetration of the ligand in the binding pocket, gives rise to the most distinct and diverse dihedral/RMSD distribution as compared to any other bulky pyrrolidine carboxamide (Figures 7.15, 7.17 and 7.24).

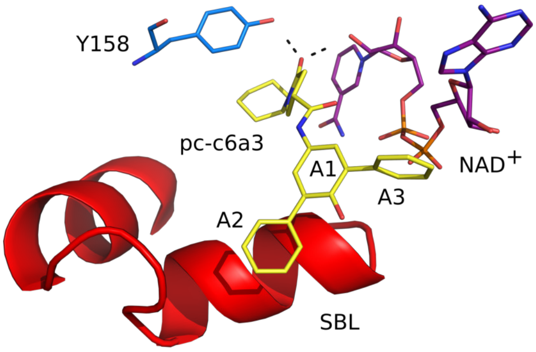


Figure 7.23 Inverted binding mode of pc-c6a3 (docked pose from GlideSP) alongwith annotated A ring system. The hydrogen bonds with the catalytic Tyr158 and cofactor (NAD^+) appear as black dashes, while the substrate binding loop is coloured in red.

In Figure 7.24, the four widely spaced dihedral contours and the corresponding line plots for the dihedral angles clearly delineate the free rotations of the ring C within the binding pocket. A similar observation holds true for the A ring system that faces the exterior of

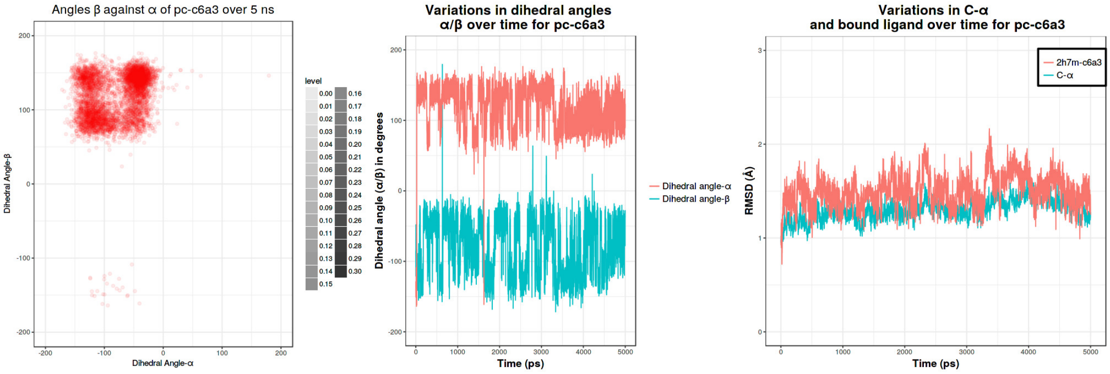


Figure 7.24 Dihedral angle distribution and RMSD fluctuations for pc-c6a3 over 5ns NPT simulation.

the protein. While the ring C motions arise from a lack of interactions with active site residues, the A ring system encounters solvent that surrounds the protein. Furthermore, the additional H-bond that pc-c7a3 and pc-c8a3 form with Pro156 (Figure 7.19) is not possible in this orientation. The additional H-bond contributes to the overall barrier to rotation for the A ring system. A direct effect of the ring C motions are the weak and transient H-bonds of the ligand with the cofactor and Tyr158. In absence of stabilising interactions, pc-c6a3 in its flipped state exhibits a classical case of destabilised binding mode, which is also reflected in its bound ligand and C-α RMSD values. In summary, the following can be said of this binding mode:

1. This binding mode with a weak H-bonding and lack of stabilising interactions does little to justify its potency (IC$_{50}$: 0.14 μM). Thus, it can be considered as an artefact from docking because other closely related molecules, such as pc-c7a3 and pc-c8a3 that differ in C ring by one and two carbon atoms, respectively, bind like the reference ligand with comparatively low fluctuations in dihedral angles and RMSD values.

2. A key feature of pc-c6a3 is that its binding mode was always predicted as flipped irrespective of the docking protocol (GlideXP/induced fit), implying that the ligand pose might get trapped in local minima and consequently features in top reported poses that score rather poorly. As a result, this binding mode can be considered as artefactual.

In summary, from Figures 7.15 to 7.17, the following can be said regarding the bulky pyrrolidine carboxamides:

1. A major fraction of bulky pyrrolidine carboxamides exhibit wide variation in the distribution as well as fluctuation irrespective of the binding mode when compared to the reference ligand viz. pc-d11.

2. In case of the bulky pyrrolidine carboxamides exhibiting stable binding, three got docked and selected in PDB 2X23 (pc-r7, pc-p31 and pc-c8a2), while only one ligand (pc-c7a3) was selected from 4TZK. The former three molecules exhibit marked stability with markedly less fluctuations in RMSD values (Figures 7.16 and 7.17) and dihedral angles (cf. Figure 7.15). On the contrary, pc-c7a3 (Figure 7.18) exhibits noticeable shifts in dihedral angles as well as bound ligand RMSD although the overall RMSD values stay well below 1 Å.

3. The alternate binding conformations as observed for a total of 5 compounds (pc-p27, p28, p36, p37, and pc-c6a3) were found to exhibit higher fluctuations in both dihedrals and RMSD values indicating that the conformation of the crystal structure ligand was most plausible for bulky pyrrolidine carboxamides.

7.3.3 H-bond and Distance analysis

This section briefly discusses the nature and quality of the H-bond interactions in between pyrrolidine carboxamides and InhA. The role of H-bonds in stabilising the binding orientations was probed by individual assessment of each ligand and then coming up with a generalised description of class-specific hydrogen bonding. The dependence of hydrogen bond on the distance in between the 1° carbonyl group of the ligand (acceptor) and the Tyr158 (-OH group), cofactor (-O13 of the ribose) warranted equal attention. The distance assessment for the donor-acceptor atom pairs would enable one to follow the motions of the ligand (and in turn hydrogen bonds it forms) with respect to time. Accordingly, a donor-acceptor atom distance analysis was carried out with respect mainly to the three binding modes observed for pyrrolidine carboxamides, with overall comparison being performed for light and bulky pyrrolidine carboxamides, respectively.

7.3.3.1 H-bond analysis of light pyrrolidine carboxamides

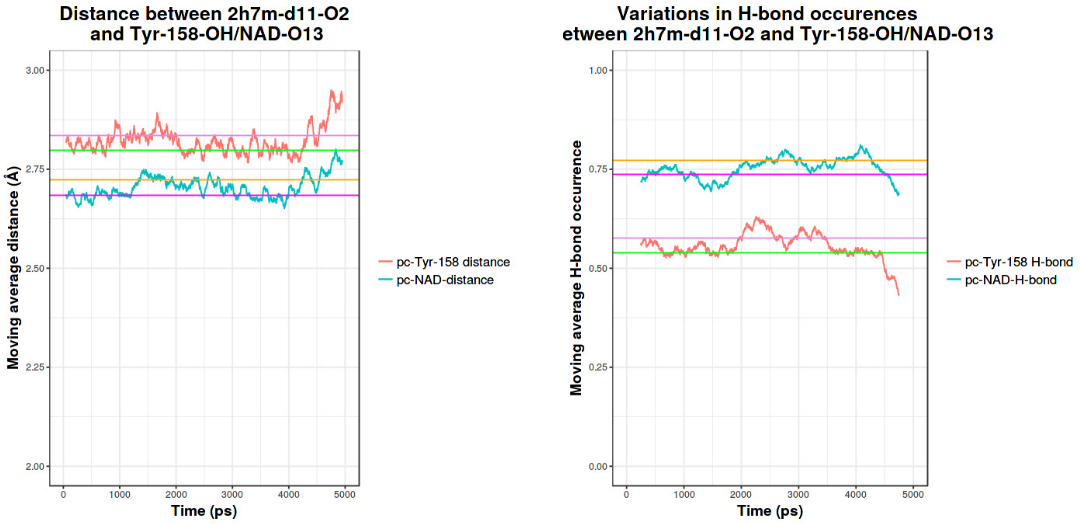
For the reference ligand, pc-d11, the analysis of the hydrogen bonds and the distances in between the 1° carbonyl group-Tyr158-OH as well as 1° carbonyl group-NAD-ribose-OH groups were initially assessed in VMD-1.9.1. As discussed in Section 7.2.2, the moving averages for the hydrogen bond occurrence and distances were considered for ease of visual analysis. The moving average is especially important for the hydrogen bond occurrence mainly because of the binary nature of output from VMD which makes visual interpretation of the results difficult. The fluctuations in the hydrogen bond occurrences and distances for the reference ligand are summarised in Figure 7.25. From Figure 7.25, the stable binding of pc-d11 is revealed with very low fluctuations in the distances between the key donor acceptor atom pairs viz, d11-O2 and Y158-OH/NAD-O13, and thereby the H-bond.

Table 7.1 Hydrogen bond occupancy values for ligand-Y158 (atom O1 of ligand and OH of Y158) and ligand-cofactor (atom O1 of ligand-O13 of ribose ring of the cofactor) donor-acceptor atom pairs of light pyrrolidine carboxamides, with ligands in bold face signifying crystal structure ligands.

Pyrrolidine carboxamide	**pIC$_{50}$**	**C-α RMSD (Å)**	**Ligand RMSD (Å)**	**H-bond occupancy (%)**	
				Ligand-Y158	**Ligand-cofactor**
s1	4.97	1.44	0.32	56.68	41.64
s2	4.45	1.27	0.34	44.98	68.08
s4	6.05	1.30	0.29	38.74	76.44
s5	4.55	1.21	0.69	69.36	22.00
s6	5.86	1.16	0.67	24.10	31.10
s10	4.77	1.09	0.34	66.18	73.60
s11	5.45	0.96	0.74	56.32	76.50
s12	4.97	1.16	0.55	58.78	76.54
s15	5.25	1.30	0.75	53.20	54.90
s17	4.13	1.29	1.06	43.02	73.58
d1	4.25	1.61	0.92	23.84	28.72
d2	4.24	1.13	0.35	37.26	70.86
d3	6.01	1.46	0.54	14.10	28.44
d4	4.43	1.17	0.45	53.06	45.16
d6	4.99	1.31	0.67	62.22	60.42
d7	5.50	1.79	0.32	50.04	79.12
d8	4.63	1.16	0.39	65.60	1.90
d9	4.50	1.15	0.61	57.72	71.32
d10	5.82	1.44	0.57	47.28	44.74
d11	6.40	1.11	0.49	54.86	74.54
d12	6.07	1.15	0.78	53.22	73.90
d13	5.88	1.22	0.81	14.22	21.00
d14	5.44	1.09	0.83	45.90	74.36
d15	5.79	1.21	0.53	53.18	65.86
d16	4.82	1.21	0.47	49.60	35.72

A key index which provides for the overall comparison of the H-bonds amongst the various pyrrolidine carboxamide subgroups is the **H-bond occupancy**, which in percentage indicates the overall duration for which the H-bond was observed. For example, if the H-bond occupancy for ligand-Y158 bond is 53%, then it means that over 53% of the 5 ns duration, an H-bond in between the two moieties was observed using the H-bond criterion described earlier. It also provides an indirect estimate about the stability of the binding orientation since binding instability is closely associated with rupture of H-bonds and "in-place" motions as well as ligand drifts within the binding pocket.

Given the fact that light pyrrolidine carboxamides bind to InhA like the reference ligand, they should exhibit moderate to high values of H-bond occupancy, which indeed is observed (Table 7.1). From Table 7.1, it can be seen that the following ligands exhibit low H-bond occupancies for both ligand-Tyr158 and ligand-cofactor pairs: s6, d1, d3, d8, d13, and d14. Of these, d3 and d8 are crystal structure ligands (PDB 4U0K and 4TZT,



(a) Moving average of distance between pc-d11 and Y158-NAD$^+$ donor atoms.

(b) Moving average of occurrences of H-bonds for pc-d11.

Figure 7.25 Moving average plots for 5ns simulation of pc-d11, green and magenta lines indicate lower quartile while light pink and goldenrod lines indicate the upper quartile of the moving average for respective distance/H-bond.

respectively). Both pc-d3 and pc-d8 exhibit significant differences in the ligand-Tyr158 distance as compared to pc-d11 (Figure 7.26), with pc-d3 exhibiting a wide variation in the said distance and pc-d8 having a noticeable number of outliers. Similar arguments hold true for the compounds pc-s6, d1, and d13, wherein there was a marked deviation in the distributions of both ligand-Tyr158 as well as ligand-cofactor distances.

Furthermore, in case of pc-d8, in addition to the weak H-bonds and thereby increased destabilisation, the possibility of alternate H-bonds being formed increases. This phenomenon involves a twist in the entire molecular structure of the ligand, with the primary carbonyl group (1°) of the ligand retaining its H-bond with the cofactor while the secondary carbonyl group (2°) moving towards Y158 and forming another H-bond (Figure 7.27). The H-bonding formed in this conformation is weak and transient, apart from being observed for a very short duration of the sampled 5 ns MD simulation. Thereby, this conformation can be considered to be metastable.

From Table 7.1 and Figure 7.26, the following can be said about light pyrrolidine carboxamides:

1. As a generalisation, potent light pyrrolidine carboxamides exhibit higher H-bond occupancies than those with lower InhA inhibitory potential. There is an indirect relationship in between the bound ligand RMSD and the nature of the H-bonds formed in between the ligand-Tyr158 and ligand-cofactor, respectively.

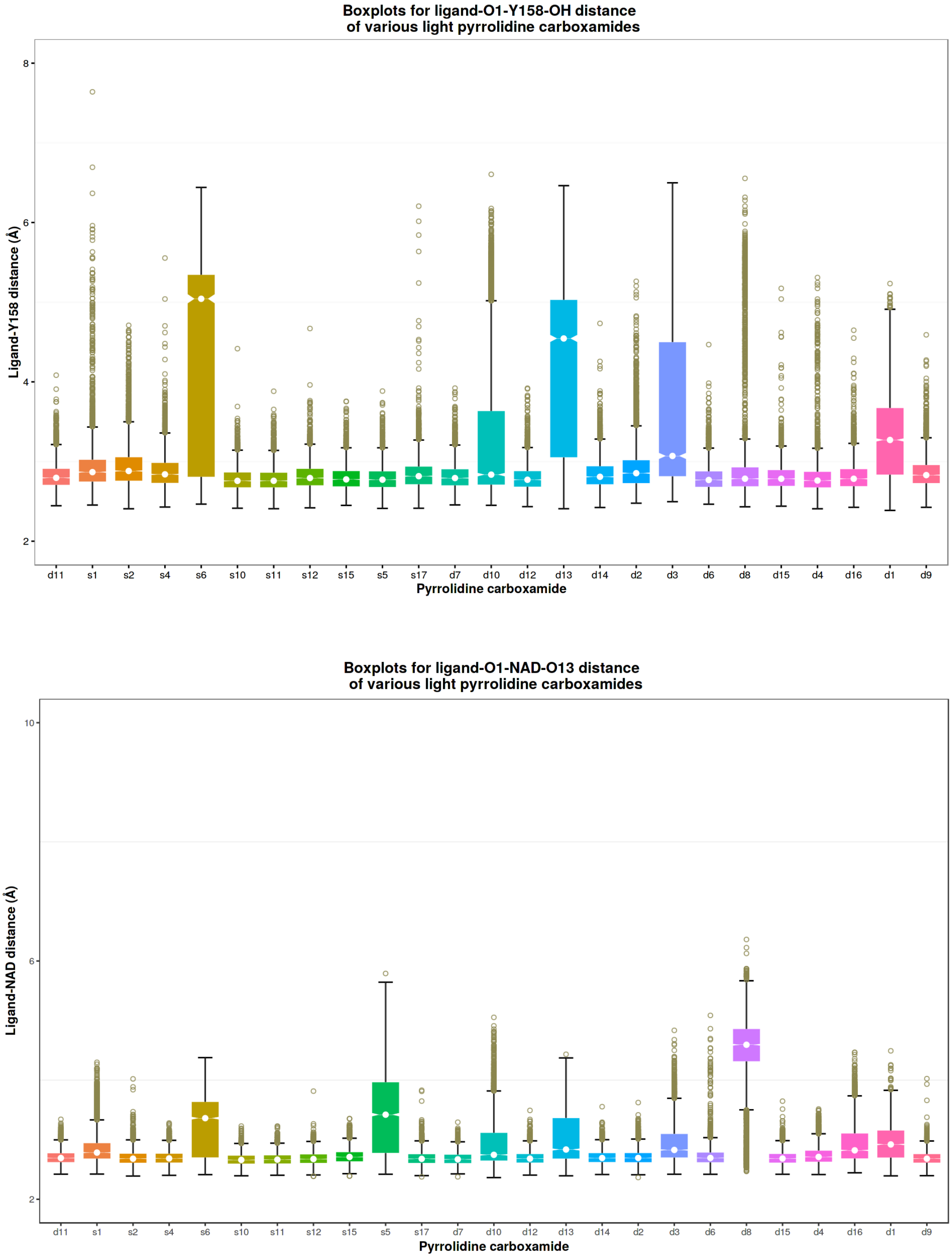


Figure 7.26 Boxplots depicting the distribution of donor-acceptor atom distances for light pyrrolidine carboxamides-NAD$^+$-Y158, with the ligands being sorted in the same way like Figure 7.8. Furthermore, the outliers are bronze circles while the median is depicted as a white circle.

2. The ligand-cofactor H-bond occupancies were found to be higher than those for the ligand-Tyr158 bond, meaning that ligands remained bound to the cofactor for a longer period as compared to the catalytic residue, i.e., Tyr158.

3. From the MD simulations of all light pyrrolidine carboxamides, it was observed that the ligand-Tyr158 H-bond was the first to rupture, primarily due the destabilisation

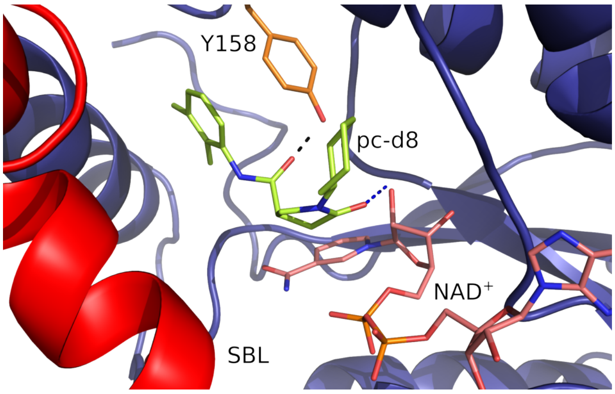


Figure 7.27 Metastable conformation of pyrrolidine carboxamide-d8 (green sticks), with alternate H-bonding; SBL is shown as red helix, Tyr158 as orange sticks, cofactor as salmon coloured sticks and H-bonds as dashed lines.

induced by the freely moving phenyl ring located adjacent to the B ring that harbours the 1° carbonyl group.

4. In case of pyrrolidine carboxamides with weak H-bonding and observable destabilisation, the chances of a metastable H-bonding conformation being observed increases, as can be seen in case of the PDB 4TZT ligand (pc-d8).

7.3.3.2 H-bond analysis of bulky pyrrolidine carboxamides

As seen from Section 7.3.2, the bigger and bulkier members of the pyrrolidine carboxamide dataset exhibit divergent binding modes as well as increased destabilisation within the binding pocket. The destabilised binding is closely associated with plausible alternate binding orientations for few molecules (pc-p27, pc-p28, pc-p36, pc-p37, and pc-c6a3) (Sections 7.3.2.2 and 7.3.2.3). However, even in case of ligands exhibiting a binding mode like the reference ligand, destabilizing movements of the A ring together with weak H-bonding contribute to a major extent for their observable motions. This means that weak H-bonding precedes destabilised movements exhibited by bulky pyrrolidine carboxamides. Accordingly, the bulky pyrrolidine carboxamides should exhibit donor-acceptor atom distances reaching the upper permissible limits for an H-bond. This was indeed the case, as can be seen from Table 7.2 and Figures 7.28 and 7.29. Moreover, since the bulky pyrrolidine carboxamides were observed to bind in three different conformations, the bonding analysis for each of the binding modes can complement the dihedral angle analysis in revealing the dominant binding conformations for this subgroup of the dataset.

Table 7.2 Hydrogen bond occupancy values for ligand-Y158 and ligand-cofactor donor-acceptor atom pairs of bulky pyrrolidine carboxamides, with the ligands in bold face exhibiting stable binding.

Pyrrolidine carboxamide	**pIC$_{50}$**	**C-α RMSD (Å)**	**Ligand RMSD (Å)**	**H-bond occupancy (%)**	
				Ligand-Y158	**Ligand-cofactor**
3a*[1]	5.40	1.39	0.82	17.66	16.00
3i*	4.86	1.26	0.75	59.66	28.76
3j*	4.53	1.36	0.88	35.18	53.14
r7	5.29	1.22	0.71	42.64	6.28
p9	5.46	1.41	0.98	25.18	57.06
p20	6.12	1.59	1.41	51.58	0.00
p21	6.39	1.53	1.83	17.84	33.78
p24	6.41	1.48	1.19	43.66	49.88
p28	5.13	1.74	1.13	32.20	9.16
p31	5.86	1.12	0.75	57.44	54.18
p33	5.59	1.75	1.64	28.20	3.42
p36	5.25	1.62	1.16	59.22	18.24
c1a1	6.33	1.36	0.69	20.42	37.76
c1a2	6.07	1.52	0.71	28.44	46.90
c6a3	6.85	1.30	1.50	0.00	20.18
c7a2	6.49	1.44	1.18	5.08	9.06
c7a3	6.56	1.22	0.81	61.50	17.98
c8a2	6.20	1.29	1.00	64.90	64.72
c8a3	5.88	1.47	1.19	34.84	56.70

[1] Although 3a-3j are light pyrrolidine carboxamides, they appear here simply because of C ring replacement.

1. **"Reference ligand binding mode":**
   The reference ligand like binding mode is demonstrated by the majority of the bulky pyrrolidine carboxamides, including pc-c7a3, that should translate to better H-bonding and bond occupancies. However, as seen from Table 7.2 and Figures 7.28 and 7.29, this is far from true. The compound pc-r7 presents an intriguing case amongst all bulky pyrrolidine carboxamides. The dihedral β (representing cyclohexyl ring motions) showed slightly wider distribution as compared to the dihedral α. This means that the motions of the C ring contributed more to the observed RMSD than those of the A ring. This should translate to a weaker ligand-cofactor bond which indeed was the case. In this case, it differed from the other bulky pyrrolidine carboxamides exhibiting stable binding (pc-p31, pc-c7a3, and

pc-c8a2). Hence in the current case, the hydrophobic interactions are expected to stabilise the binding of the ligand in absence of stabilising H-bonding. Additionally, all of the bulky pyrrolidine carboxamides except pc-r7 (pc-p31, pc-c7a3, and pc-c8a2) that exhibited stable binding showed low to moderate H-bond occupancies for both H-bonds of the bound ligand, i.e., with Tyr158 and the cofactor. From these observations, the following can be said of bulky pyrrolidine carboxamides binding like pc-d11:

- The majority of the bulky pyrrolidine carboxamides (11/18) exhibit low to moderate H-bonding occupancies ($\leq$ 50%) mainly for the ligand-Y158 bond, while the H-bond occupancy for the ligand-cofactor was mostly low. The exception to this observation were the compounds that exhibited stable binding (pc-p31, pc-c7a3, and pc-c8a2) and pc-p9, pc-c8a3. For these three compounds (pc-p31, pc-c7a3, and pc-c8a2), a weak trend in between the H-bond occupancy and the binding stability (as seen from the RMSD values) can be observed. Although trivial, this was quite opposite to light pyrrolidine carboxamides, which exhibited a higher ligand-cofactor H-bond occupancy.
- A major portion of bulky pyrrolidine carboxamides exhibiting stable binding got docked in PDB 2X23 (pc-r7, pc-p31, and pc-c8a2) with the exception of pc-3i and c7a3 (both 2H7M), respectively. In the former compounds, the tight binding pocket of 2X23 leads to increased interactions with the ligand that can explain the stable binding as well as moderate H-bonding occupancies for these compounds. In case of the latter (pc-3i and pc-c7a3), the hydrophobic interactions along with the chlorine mediated halogen bonds (especially for pc-3i) stabilise the binding mode and give rise to the observed H-bond occupancy.
- A common observation pertaining to all of the aforementioned compounds is that all of them exhibited an almost converging donor-acceptor atomic distance ($\leq$ 3 Å) (for both ligand-cofactor and ligand-Tyr158). From this, it can be inferred that the crystal structure orientation is optimal in the case of bulky pyrrolidine carboxamides, given the co-existence of stable binding and moderate to high H-bond occupancies.

2. **"Alternate binding mode-1"**

This binding mode is exhibited by four bulky pyrrolidine carboxamides (pc-p27, pc-p28, pc-p36, and pc-p37), all of which have a similar type of hydrogen bonding. Furthermore, only pc-p28 and pc-p36 will be discussed since pc-p27 and pc-p37 were not subjected to MD simulations due to reasons described earlier. Both pc-p28 and pc-p36 exhibit weak to moderate H-bonding with Tyr158 which differs significantly from the ligand-cofactor H-bond occupancy. This transient H-bonding

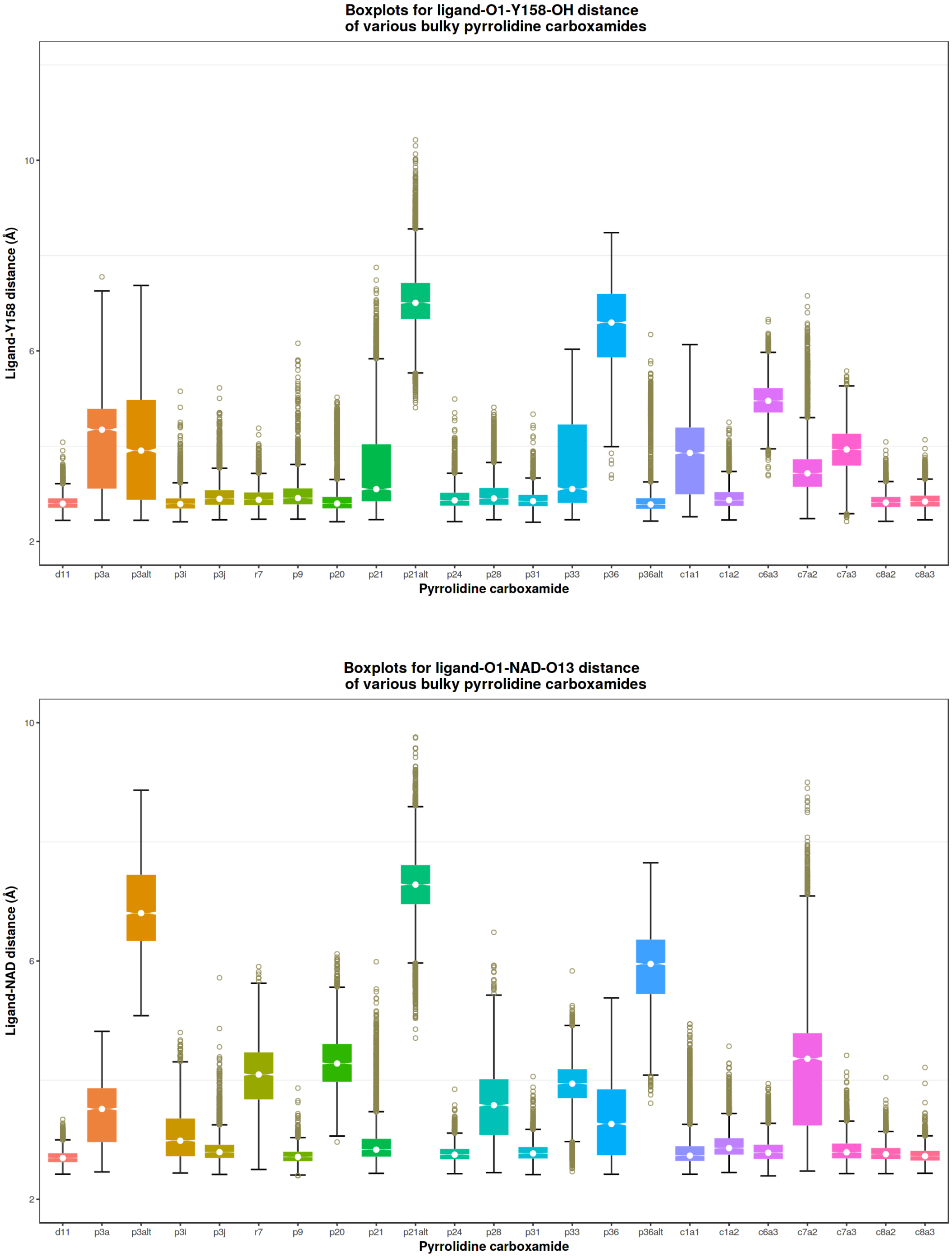


Figure 7.28 Boxplots depicting the distribution of donor-acceptor atom distances for bulky pyrrolidine carboxamides-NAD$^+$-Y158, with the ligands being sorted in a way similar like Figure 7.15. Furthermore, the outliers are bronze circles while the median is depicted as a white circle.

increases the chances of metastable conformations being sampled during the MD simulations like pc-d8. Indeed, this conformation was observed prominently in both compounds, with an increase in the ligand-1° carbonyl group-Y158-OH distance that coincides with a decrease in the distance between ligand 2° carbonyl and Y158-OH (Figures 7.30 and 7.31). This metastable conformation arises mostly from

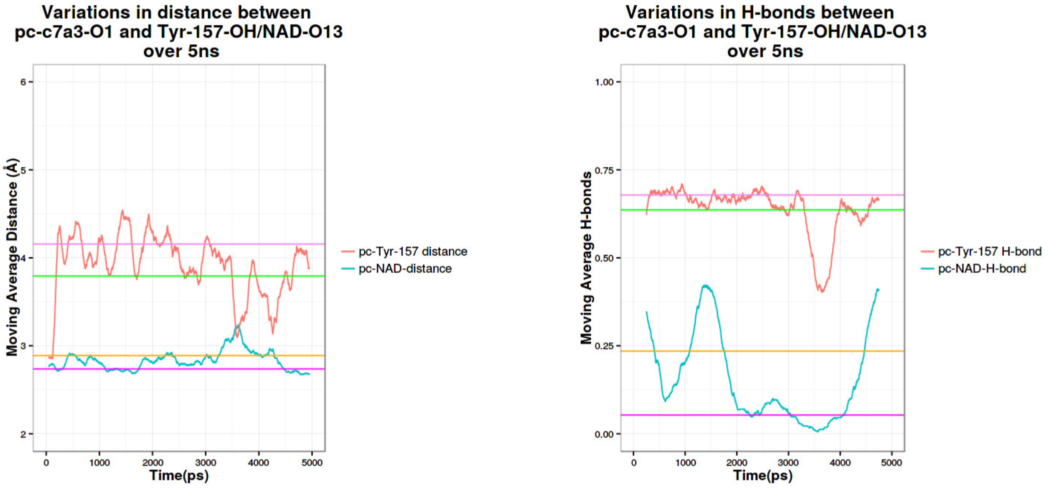


Figure 7.29 Moving average plots for (a) pc-c7a3 and Y158-NAD$^+$ donor atom distances and (b) H-bond occurrence for pc-c7a3.

"in-place" motions (twisting of the ligand). Furthermore, the metastable H-bonding conformation is a direct result of weak hydrogen bonding which in turn arises from the non-optimal orientations predicted for pc-p28 and pc-36, respectively. Going by the results of the dihedral angle analysis as well the hydrogen bonding, it can be safely concluded that this binding mode is far from stable as compared to the standard binding conformation seen in the crystal structures.

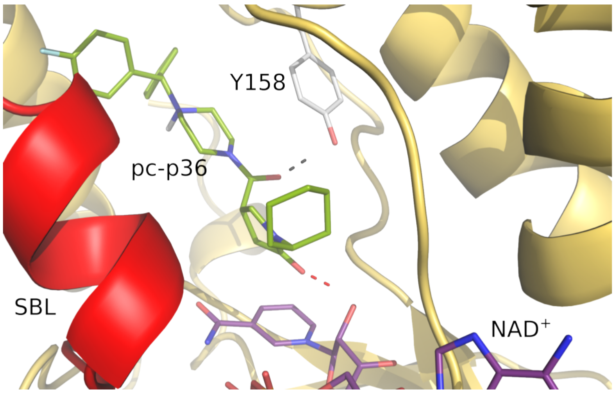


Figure 7.30 Transient conformation of pyrrolidine carboxamide-p36 (green sticks), with alternate H-bonding; SBL is shown as red helix, Y158 as white sticks, cofactor as violet sticks and H-bonds as dashed lines.

3. **"Alternate binding mode-2"**

This inverted binding mode of pc-c6a3 represents inadequate penetration of the ligand in the binding pocket (Figure 3.21). As a direct consequence the H-bonding is naturally expected to be weak and susceptible to rupture. Indeed this was the case, wherein the H-bond between the ligand and Y158 was altogether missing while the H-bond in the ligand and the cofactor is not strong either as seen from the H-bond occupancy of the same (Table 7.2). This results in free movements of the ligand in place and consequently within the binding pocket as seen from the

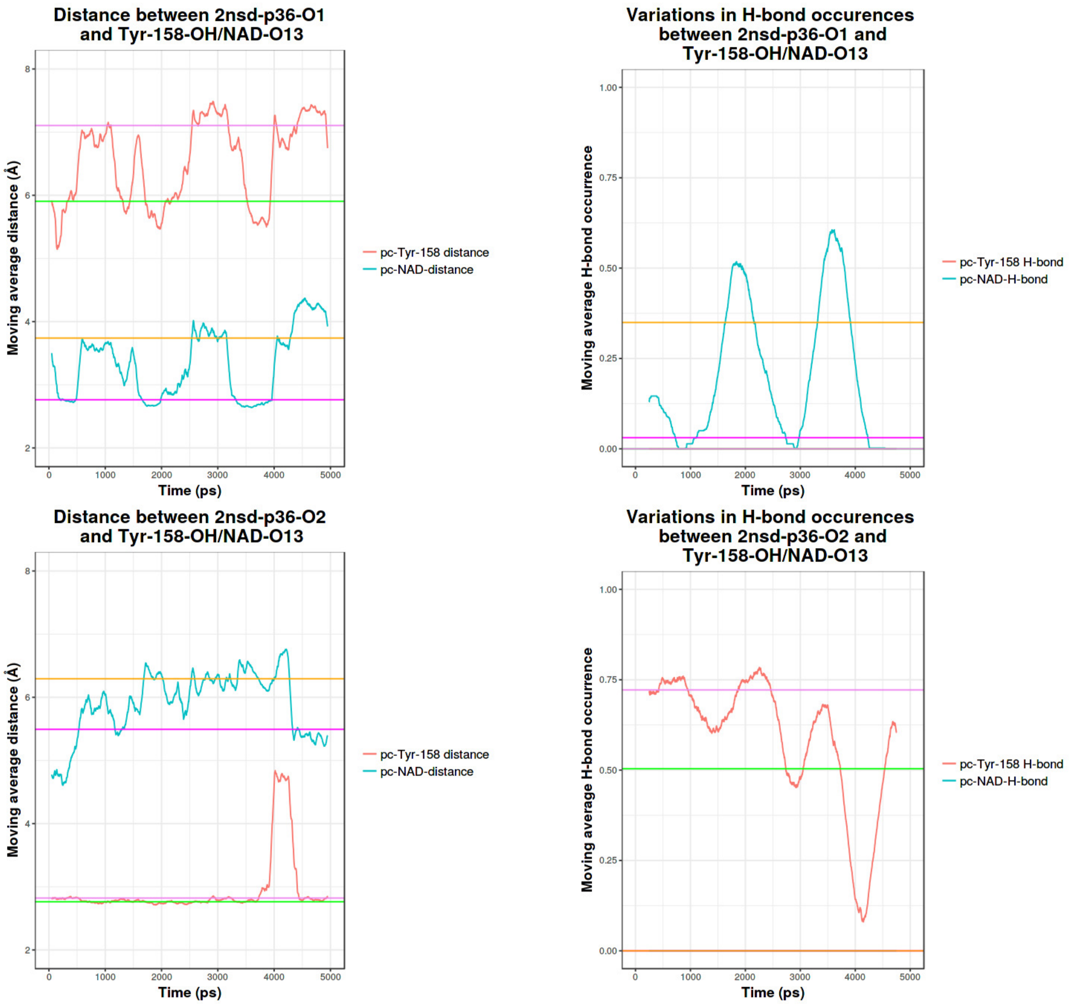


Figure 7.31 Moving average plots for ligand-Tyr158 and ligand-cofactor distances and H-bond occurrences of 1° carbonyl oxygen (top pictures) and 2° carbonyl oxygen (bottom pictures) from pc-p36

characteristic dihedral distributions and the RMSD drifts of the ligand. In short, the H-bonding analysis offers additional indirect support to the notion that the inverted binding mode is an artefact from molecular docking.

Considering all of the observations emanating from the H-bond and distance analysis of bulky pyrrolidine carboxamides, the following statements can be safely made in terms of their dominant binding conformation as well as the overall binding stability:

1. All of the indirect ways attempted to reveal the most plausible binding mode for the bulky pyrrolidine carboxamides clearly pointed out that these compounds bind to InhA in a similar conformation like the crystal structure ligands.
2. A small number of ligands bind in an alternate conformation (Sections 7.3.2.2 and 7.3.2.3) that is much more susceptible to destabilisation and exhibit random movements both in place and within the binding pocket.

3. As a generalisation, stable binding is associated with low drifts in RMSD, donor (Tyr158 and cofactor)-acceptor (ligand-1°/2° carbonyl group) distance. Furthermore, stable binders exhibit moderate to high H-bond occupancies for the dual H-bonds of the ligand.

4. The following statement can be generalized for **pyrrolidine carboxamides irrespective of their class, viz., light or bulky substituents**: The presence of dual hydrogen bonds with cofactor (NAD^+) and Ty158 is **absolutely a must** for exhibiting stability within the binding pocket.

5. Finally, the inverted binding mode as exhibited by pc-c6a3 can be considered an artefact from docking, given the instabilities and movements it exhibits in stark contrast with the rest of the group. Furthermore, all of the observations from the dihedral as well as H-bond and distance analysis clearly support this statement.

7.4 Discussion

The central aim of the current analyses was ascertaining the dominant binding modes for bulky pyrrolidine carboxamides, mainly due to lack of crystal structure. The orientation of the molecule supplied as an input to MD simulations as well as the substituents on the main scaffold largely determine its binding stability as well as strength of the interactions. Hence, correlating the dynamics and the structure activity relationship was an additional task performed during the analyses.

The dihedral angle analysis aided greatly in shedding light on the SAR-dynamics relationship, the conclusions of which fell in line with the SAR of light pyrrolidine carboxamides as reported by He et al. [52]. The analysis also revealed that in case of the light pyrrolidine carboxamides, a majority of ”in-place” movements arise from the free rotations of the A (phenyl) ring. The determinants of movement within the binding pocket were, however, hard to visualise since the fluctuations in the dihedral angles hardly coincide with that of bound ligand RMSD/C-α RMSD. However, potent light pyrrolidine carboxamides did follow the trend as put forward by He et al. [52], with meta and 3,5-disubstituted (meta, meta) phenyl rings (e.g., halogens, small electron withdrawing groups) exhibiting much better stability than monosubstituted (ortho/para) and 2,3- or 2,4-disubstituted compounds. The non-preferred para substitution as mentioned by He et al. [52], was confirmed in the dihedral angle-RMSD analysis.

One of the most important things revealed from the dihedral angle analysis was that there is more space within the binding pocket that could be exploited for stabilising the ligand binding via increased van der Waals interactions. The support for this comes from the following: firstly, all of the ring A of pyrrolidine carboxamides not getting docked/selected from PDB 2X23 exhibited more fluctuations as compared to those who got docked in PDB 2X23 (e.g., pc-r7, pc-p31, and pc-c8a2). The PDB 2X23 represents a much tighter binding pocket and, thus, stabilises the binding of molecules that hardly fit inside the pocket, primarily via H-bonds and increased van der Waals interactions. On the other end, PDB 2NSD and 4TZK offer comparatively more space to accommodate the bulky ligands, making it possible for the observed movements to take place. Thus, the movements of the ligands within the binding pocket can be stabilised by targeting this ”extra” space by suitable substitutions with an aim to increase the stabilising non-bonded interactions.

The bulky pyrrolidine carboxamides were designed with this exact aim, with bulky structures replacing the solitary phenyl ring of the light pyrrolidine carboxamides. However, with their increased size came the problems of predicting the binding conformation reasonably. The docking with induced fit yielded three different conformations, whilst the majority of the compounds were found to retain the crystal structure ligand conformation. However, a sizeable number of compounds (Sections 7.3.2.2 and 7.3.2.3) that also included the most potent pyrrolidine carboxamide (pc-c6a3, IC_{50}: 140 nM) exhibited a different binding mode. The dihedral angle-RMSD analyses together with analysis of the H-bond occupancies and distances in between the ligand-Tyr158 and ligand-cofactor revealed that the crystal structure orientation is the most plausible binding mode for bulky pyrrolidine carboxamides. For some compounds (e.g. pc-d8, pc-p36), alternate binding conformations were also sampled during the MD simulations. The alternate binding mode also shows a completely different H-bonding pattern, although the donor and acceptor moeties remain the same. Moreover, the artefactual nature of the inverted binding mode of pc-c6a3 was evident from the high fluctuations in dihedral angles and the RMSD values. This also coincided with weak and transient H-bonds with the cofactor as well as Tyr158.

The H-bond occupancy and distance analyses clearly supported the aforesaid observations. Additionally, the ligand-Tyr158 H-bond occupancy was lower than that of the ligand-cofactor bond for light pyrrolidine carboxamides. It was exactly opposite in case of the bulkier members of pyrrolidine carboxamides. Additionally, with the exception of pc-c7a3 which got docked in 4TZK, all bulky pyrrolidine carboxamides exhibiting stable binding (N=3, pc-r7, pc-p31, and pc-c8a2, Table A.4) were found to get docked in 2X23, irrespective of their size. The reason for this can be attributed to the almost closed binding pocket of 2X23 that enables increased van der Waals interactions for the protein-ligand pair that ultimately results in stabilised binding, inspite of the bulky size of the ligands. The increased binding stability for these ligands is also manifested in the high H-bond occupancies for the ligand-Y158 and ligand-cofactor H-bond, respectively, alongwith low fluctuation in the respective donor-acceptor atom distances. All of these analyses also pointed out the stabilsing effect of the dual H-bonds that the ligands forms with Tyr158 and the cofactor.

7.5 Conclusion

From the information available about the SAR of pyrrolidine carboxamides and the findings from the dihedral angle analysis as well as the H-bond/distance analysis, the following things can be said about the pyrrolidine carboxamides in general:

1. The trend in between the overall potency and observed binding holds true strongly for smaller members whilst not so strongly in case of bigger members of the bulky pyrrolidine carboxamides.

2. From the perspective of SAR, the presence of halogens or small electron withdrawing groups at meta position on the A ring (for light pyrrolidine carboxamides) and bulky aromatic ring systems (for bulky pyrrolidine carboxamides) works well.

3. In regards of binding stability, the majority of in place motions for both light and bulky pyrrolidine carboxamides are driven by phenyl ring flips that suggested that there was additional space within the binding pocket to accommodate the ring flips. The support for this comes indirectly from PDB 2X23 with a narrow binding pocket and the bulky pyrrolidine carboxamides that got docked in it. All of the compounds exhibited remarkable binding stability in a way similar to the crystal structure ligands (except pc-d8).

4. In regards of the dominant binding conformation for the pyrrolidine carboxamides, especially the bulky pyrrolidine carboxamides, the crystal structure like conformation was indirectly deemed to be the dominant and most likely binding mode by all analyses performed in this study. The alternate binding mode can exist, but exhibits much higher instabilities whilst the inverted binding mode is a docking artefact.

5. In regards of H-bonding and distance analysis, the pyrrolidine carboxamides were generally found to exhibit moderate to weak H-bonding that was in line with the potency of the compounds under investigation. Furthermore, the ligand-Y158 H-bond was deemed weaker than the ligand-cofactor H-bond mainly due to the A ring movements that push the acceptor atom (primary carbonyl) away from Y158 towards the cofactor.

6. Weak and transient H-bonding of the ligand (as seen in both light and bulky pyrrolidine carboxamides) was associated with atypical binding orientations. Such a conformation was also observed in case of PDB 4TZT ligand (pc-d8) upon longer

simulation durations. A direct manifestation of the weak bonding was the transient nature of the atypical binding orientations themselves, rendering them metastable.

Chapter 8

Essential dynamics analysis of MD simulations of pyrrolidine carboxamides

8.1 Introduction

The dihedral angle and hydrogen bond analyses from Chapter 7 provided useful insights into the binding of light and bulky pyrrolidine carboxamides to InhA. They also aided in elucidating the nature of the movements and especially shed light onto the atoms contributing maximally to these movements. All of these analyses were made against the background of the binding modes obtained from molecular docking. The binding modes of light pyrrolidine carboxamides were usually conserved (including the in-place mutation) and did not cause any issues. On the contrary, the bulky pyrrolidine carboxamides posed a problem for accurate pose prediction. Despite of their considerable bulkiness, 3 different binding modes were predicted by docking. From amongst these three binding modes, MD simulations aided in ascertaining one of them (crystal structure ligand-like) as the dominant pose.

A critical factor pertaining to the MD corroboration was its simulation length and the extension of observed movements at prolonged time scales. Additionally, the direction of the global movement of the bound ligand and the correlation in between the movements of ligand and protein atoms remained unrevealed. Whereas certain aspects of the dynamic binding of pyrrolidine carboxamides arose from the analysis of the dihedral angles and the distances (cf. Chapter 7, Sections 7.3 and 7.3.3), a more reliable description of the motions upon binding of pyrrolidine carboxamides to InhA at extended time scales can be obtained from essential dynamics [329, 330]. This technique especially provides information about the correlation between critical patterns of atomic movements in the molecular recognition process. The essential dynamics also aids in focussing upon the movements that play a crucial role in ligand binding.

Dimensionality reduction techniques come handy in filtering out the motions of interest from the vast multitude of movements typically observed in a MD simulation. Principal Component Analysis (PCA) is such a technique that aids in ascertaining the motions with maximal variance that are often closely related with the molecular recognition process [329]. In the current context, PCA was performed on the bound ligand as well as the active site residues in order to study the functional motions of the ligand upon binding. The background of this method is explained in Section 8.2.

A critical aspect in binding is whether the protein/ligand motions exhibit any sort of correlation (intramolecular/intermolecular). These motions are important for studying the biomolecular functions of proteins, which are difficult to assess experimentally [331]. Although MD simulations provide a useful approach to study these motions, assessment of their relationship from PCA remains restricted to linearised correlations. Since MD simulations sample events that often are not a result of linear correlations, the information from non-linearly correlated motions that also contribute to protein function remains invariably unaccounted in PCA [331]. Hence, a different approach that describes also a generalised correlation term can be utilised in a complementary fashion to gain an overall description of the correlated protein-ligand motions. Dynamic Cross Correlation (DCC) [331] was chosen for this purpose in this thesis.

Both principal component analysis and dynamic cross correlation analyses were performed on the trajectories of bound ligands to observe the dominant movements occurring in both protein and ligand. The complementary information emanating from principal component analysis and dynamic cross correlation analyses can then be utilised to strengthen traditional structure-based drug design efforts in order to yield new pyrrolidine carboxamides with improved binding stability.

8.2 Materials and Methods

The following section elaborates on the theory of the methods described earlier, with a particular focus on the elucidation of both linearised and non-linearised correlated motions of protein and ligand alike.

8.2.1 Essential Dynamics

The process of Principal Component Analysis can be better explained using Figure 8.1, with the original dimensions being X1 and X2, respectively. The shaded region represents the subspace of the complete ensemble that is populated by points that represent all motions including the functionally important ones. PCA diagonalizes and transforms the original dimensions (X1, X2) to new uncorrelated axes PC1 and PC2 that are linear combinations of the original points. The PC1 represents the first principal component and as seen from the diagram, represents maximal variance of the system, followed by a 2^{nd} component (PC2) that is orthogonal to the first accounting for as much of the remaining variance as possible.

Generally, a system with N atoms has $3N$-*6* degrees of freedom that span its cartesian coordinate space. The motions of the system (protein) within this space can be reduced to a few functionally relevant ones using principal component analysis (PCA). For example, in the case of lysozyme, where $N \approxeq 2000$, the degrees of freedom representing important motions can be reduced to 5% of the total degrees of freedom upon careful selection of 300 collective coordinates spanning the cartesian subspace [332]. It important to note that at physiological temperature, atomic fluctuations within such a subspace have a major contribution to the overall fluctuations of the system. Furthermore, the major modes of collective fluctuations are essentially anharmonic in nature [316, 333]. A PCA performed on a trajectory typically aids in ascertaining the major collective fluctuations in the system.

In addition to PCA, several other techniques yielding similar results exist, e.g. Normal Mode Analysis (NMA) [171, 334, 335] and its adapted method, the Elastic Network Model (ENM) approach [336, 337]. PCA differs from these techniques in the sense that it is purely anharmonic (all other methods assume harmonicity in the system). Furthermore, the aforementioned techniques generate an ensemble of structures from a singular input structure to ascertain the collective fluctuations within the system as opposed to PCA, which is typically performed on an ensemble of structures derived from a MD simulation. PCA usually assesses protein dynamics as a diffusion along short shallow minima on small spatial scales and large scale anharmonic motions between multiple deep minima [171].

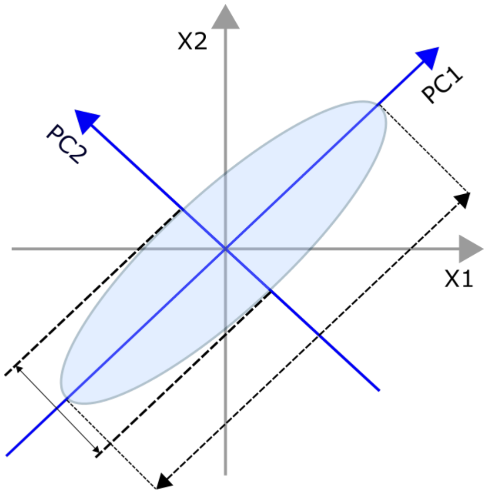


Figure 8.1 Principal Component Analysis basics for two functionally important dimensions X1 and X2. The shaded region represents the essential subspace covering the important movements, whilst PC1 and PC2 represent the principal components (axes).

In general practice, the PCA of a trajectory consists of the following steps:

1. **Fitting the ensemble of structures onto a single reference structure:** This step consists of a least squares fit of the configurations (cartesian coordinates)

from the trajectory onto a reference structure, mainly to remove the rotational and translational fluctuations of the protein/ligand atoms.

2. **Construction of a variance–covariance matrix (C) and its orthogonal transformation:**

   A covariance matrix $\boldsymbol{C}$ generated with the "fitted" trajectory represents the correlation in between the atomic motions:

$$\begin{aligned} C = cov(x_i) = & \left\langle \Delta x_i \cdot \Delta x_i^T \right\rangle, \\ \Delta x_i = & x_i - x_i^{ref} \\ x_i - \langle x_i \rangle = & T\vec{x_i} \text{ or } \vec{x_i} = T^T \left(x_i - \langle x_i \rangle \right) \end{aligned} \tag{8.1}$$

   where $\langle \ \rangle$ denotes an average over time, x_i^{ref} is an arbitrary reference value of x_i which represents the coordinates from the trajectory and Δx_i^T the transpose of x_i. In other words, x_i^T represents the orthogonal coordinate transformation (T) of Δx_i. This transforms $\boldsymbol{C}$ into a diagonalised correlation matrix Λ of eigenvalues (λ_i) where:

$$\begin{aligned} \Lambda = & \left\langle aa^T \right\rangle \\ C = & T\Lambda T^T \\ \Lambda = & T^T C T \end{aligned} \tag{8.2}$$

   Λ is a symmetric matrix (like $\boldsymbol{C}$) whose i^{th} column is an eigenvector with an eigenvalue (λ_i). The eigenvalues represent the co-variances of the atomic displacements relative to their respective averages for each pair of atoms in the cartesian coordinate subspace [171, 338]. It also contains the variances of individual atomic displacement (mean square fluctuations) along the diagonal of Λ. These eigenvectors can be ordered with respect to their eigenvalues corresponding to the variances being depicted by the eigenvectors [332]. The motions of atoms according to these vectors are referred to as **"modes"**.

   In general, if a system has *"N"* atoms, then **C** will be a *3N*×*3N* matrix, with *3N-6* degrees of freedom for the cartesian coordinate subspace which can be shown as:

$$\begin{aligned} \text{Covariance Matrix } C = & \left[m_{ij} \right] \\ m_{ij} = & \left(\frac{1}{S} \right) \sum_t \left(x_i\left(t\right) - \langle x_i \rangle \right) \left(x_j\left(t\right) - \langle x_j \rangle \right) \end{aligned} \tag{8.3}$$

   where S denotes the total number of configurations from the trajectory, t is the time in picoseconds, i refers to the ith coordinate, where (i = 1,2,...,3N) and N is the number of atoms. The term $\langle x_i \rangle$ is the averaged value for x_i over all of the sampled configurations [316].

3. **Projection of the eigenvectors onto the original coordinates:**
In the final step of the PCA, the trajectory is projected frame by frame onto the coordinate space also referred to as the **principal coordinate space**. This subspace is often adequately approximated by the first few large amplitude modes [316]. Since the large amplitude modes also describe anharmonic motions that contribute to protein function, this space is also called "essential" space, mainly because it describes motions critical for a protein function. Thus, the process of projecting the trajectory on a few collective modes spanning the "essential" subspace is called as **"Essential Dynamics"** (EDA). A key term pertaining to the "modes" is the **Relative positional fluctuation** (RPF) (Equation (8.4)) [339] that denotes the amount of the movements that are associated with the "essential" subspace spanned by the first n eigenvalues.

$$RPF\ (n) = \frac{\sum_{i=1,n} \lambda_i}{\sum_{i=1,3N} \lambda_i} \tag{8.4}$$

where λ_i is the ith eigenvalue, n is the number of eigenvalues, and 3N equals the coordinates of N atoms in the system under study.

8.2.2 Dynamic Cross Correlation

The *dynamic cross correlation* as a generalised correlation measure highlights especially the non-linear correlated motions amongst non-covalently bonded atoms/protein residues when used complementarily with essential dynamics. The general correlation of movements in MD simulations can be obtained from calculation of a normalised covariance matrix according to Equation (8.5). Each matrix element is calculated in the same way as the *Pearson correlation coefficient* of the individual atomic displacements [331]. The correlated motions being described by essential dynamics only hold true subject to the conditions that the x_i and x_j are two colinear vectors (i.e., along the same plane) (Equation (8.5)). In other words, essential dynamics does not describe the correlations amongst the non-colinear vectors [340] (i.e., not occur along the plane of the eigenvector being studied). In case of linearly correlated movements, a clear separation of the positively and negatively correlated motions would be beneficial in describing the molecular interactions during ligand binding.

$$C_{ij} = \frac{\langle (r_i - \langle r_i \rangle) \rangle - \langle (r_j - \langle r_j \rangle) \rangle}{\sqrt{\left(\langle r_i^2 \rangle - \langle r_i \rangle^2 \right) \left(\langle r_j^2 \rangle - \langle r_j \rangle^2 \right)}} \tag{8.5}$$

where C_{ij} denotes the cross correlation amongst atoms i and j, r_i and r_j are the coordinates of the atoms at a given time point and $\langle r_i \rangle$ and $\langle r_j \rangle$ are their averaged values over the entire trajectory.

Dynamic cross correlation (DCC) is capable of achieving both (i.e. separation of positively and negatively correlated movements) since it is based on the theory of *Shannon mutual information* [341]. This theory enables direct comparison of the generalised correlation with the output of Pearson correlation coefficient [331]. The discussion of *Shannon mutual information* is out of context for this work. The reader is advised to refer to literature [331] for more details.

8.2.3 Atom selections for analysis

The following section focusses on revealing the positively and negatively correlated motions of the protein-ligand pair in order to understand the events taking place after protein-ligand association. The essential dynamics and the dynamic cross correlation analyses were carried out for the following atom selections:

- Heavy atoms of the bound ligand.
- Active site residues: C-α atoms of the residues lying within a 5 Å radius of the ligand, viz., Phe97, Met98, Pro99, Met103, Asp148, Phe149, Met155, Tyr158, Lys165, Leu197, A198, Met199, Ser200, Ile202, Val203, and Leu207.
- Substrate binding loop residues: C-α atoms of residues with indices 195 to 212.

To perform the essential dynamics, the solvated protein-ligand complex before equilibration (of PDB 4TZK) followed by the parameter file and the 5ns trajectory were loaded in VMD 1.9.1 [304]. Using the RMSD trajectory tool implemented in VMD, a C-α atom based least squares fitting of the trajectory to the reference structure was performed. The Normal Mode Wizard (NMWiz) GUI of VMD was then invoked which allows the user to depict, animate and perform comparative analysis of normal (in the current case, principal) modes. The GUI is a front end for ProDy [342], that actually performs the principal component analysis on the trajectory. NMWiz takes in the user input as an atom selection described earlier followed by generation of the C matrix and its diagonalisation in the background. Using the default settings, a total of 10 principal modes were exported to a normal mode (.nmd) file for subsequent visualisation and analyses.

Although the analyses were performed for the entire pyrrolidine carboxamide dataset, the individual results are generalized for the light and bulky pyrrolidine carboxamides, respectively. Of the 10 principal modes exported by default, observations and conclusions thereof were derived from the top 5 modes, since they are expected to contain the maximal variance of the system across all cases [316]. Accordingly, all figures for the results of essential dynamics analysis depict the information derived from the inspection of top 5 principal modes for the individual atom selections.

8.3 Results

8.3.1 Essential Dynamics

The summary of the essential dynamics for the aforementioned atom selections can be seen in Table 8.1, while the ensuing sections describe the same briefly.

8.3.1.1 Principal Modes for ligand heavy atoms

The information from the first 5 principal modes of the heavy atoms of the ligand can be collated to depict the collective motions of the ring. In other words, the atomic fluctuations from the principal modes were combined to allow for a generalised description of the relative ring motions for light and bulky pyrrolidine carboxamides, respectively. The reference ligands pc-d11 (light pyrrolidine carboxamides) and pc-c7a3 (bulky pyrrolidine carboxamides) were chosen as suitable representatives. Their complexes were utilised to depict the collective motions of the residues constituting the active site and the substrate binding loop, respectively. Figure 8.2 depicts the mobile rings for "light" and "bulky" pyrrolidine carboxamides, respectively. Although the mobile atoms are being discussed, it is the collective motions of the rings and their direction that is critical to highlight the difference amongst light and bulky pyrrolidine carboxamides.

In case of both light and bulky pyrrolidine carboxamides, majority of the motions stem from the rotation of the rings A, C and the secondary (2°) carbonyl group around the C-C and C-N covalent bonds linking the rings A, B, and C, respectively. The ring B exhibits low mobility mainly due to the dual H-bonds of the 1° carbonyl group with Tyr158 and NAD^+. The mobility of the ring B is closely related with the binding stability of the ligand. This can be seen from the decreased ring B mobility in cases where a strong hydrogen bond in between the ligand-cofactor and ligand-Y158 was established. This also affects the overall potency of the molecule.

The detailed movements of light pyrrolidine carboxamides as seen for the representative ligand pc-d11 can be summarised as follows:

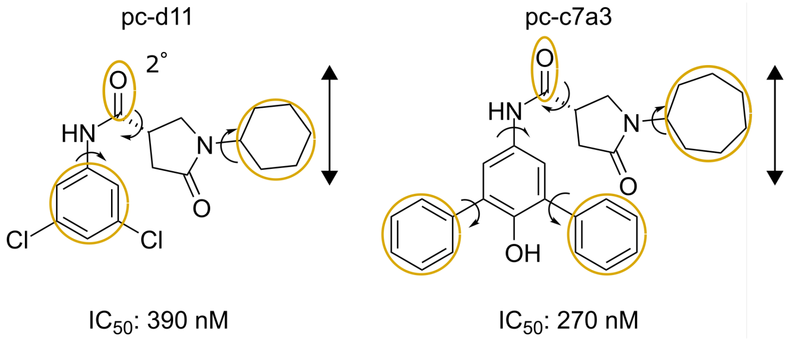


Figure 8.2 Representative structures of light and bulky pyrrolidine carboxamides alongwith mobile rings circled in yellow (pc-d11 and pc-c7a3). The rotations of the annotated rings have been marked by arrows. The primary and secondary carbonyl groups have been depicted by 1° and 2°, respectively. The vertical arrows indicate the additional motion of ring C in both molecules. Furthermore, the barrier to rotation of ring A1 (central phenolic ring) in case of bulky pyrrolidine carboxamides has been depicted as a half arrow.

1. **Ring A:**

   The rotation of ring A (phenyl ring) (cf. Figure 8.2) was found to decrease in ligands with the substitution pattern in the following order: para > ortho > meta. The motions of substituents on the phenyl ring were always directed towards the key residues of the minor exit portal and the mid to end region of the SBL (Met155, Met161, Met199, Ile202, and, Val203) (cf. Figure 8.3). The mobility of the 2° carbonyl group was an indicator of the shallow potential energy surface on which the rings A and B move. In Figure 8.3, the arrows merely denote the direction of the rings, while their length is purely symbolic and does not represent the actual length (eigenvalues) of the principal modes.

2. **Ring B:**

   The movement of ring B is rather dampened as a results of the dual H-bonds (between the 1° carbonyl group and ligand, cofactor). This movement is characterised by the ring B moving slightly away from Y158 and cofactor, rupturing the ligand-Y158 H-bond in the process. This phenomenon characterises the weak Y158-ligand bond, which is commonly observed in all pyrrolidine carboxamides. Thereafter, the stabilisation of the B ring is exclusively through favourable van der Waals interactions in between it and the nicotinamide ring of the cofactor. A geometrically non-optimal CH-π interaction [343–345] can be seen as already being established (Figure 8.4).

3. **Ring C:**

   The cyclohexyl ring atoms facing towards the exterior of the protein has plenty of space to move freely. As a direct consequence, the entire ring exhibits high mobility during which it frequently comes in contact with Phe97, Met98 and the phosphate group of the cofactor (cf. Figure 8.3).

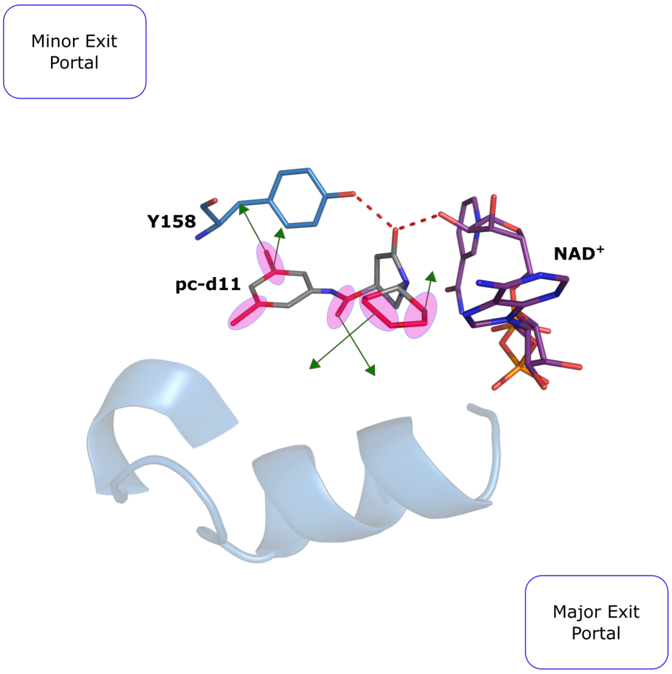


Figure 8.3 Collective motions of the maximally mobile atoms of reference ligand pc-d11 (red sticks). The green arrows denote the direction of the collective motions of the rings and are purely symbolic and do not represent the actual length (eigenvalues) of the principal modes.

4. The aforementioned facts underscore the direction of the collective motions for the rings that make up the ligand. Collectively, the major motions of the light pyrrolidine carboxamides as deduced from their principal modes can be visualised as rotations for ring A. They are initiated by the movement of secondary (2°) carbonyl group towards the substrate binding loop, whilst the C ring shows independent motions that result due to the availability of space and lack of stabilising interactions.

The situation changes when the bulky pyrrolidine carboxamides are considered primarily because of the replacement of rings A and C by larger and bulkier rings. Assuming that all bulky pyrrolidine carboxamides bind like the reference ligand (pc-d11), the compound pc-c7a3 served as a model system to highlight the generalised collective motions of bulky pyrrolidine carboxamides. In order to efficiently depict the individual movements of the A ring system, the respective rings were named as A1 (central ring), A2 (which lies near the SBL), and A3 (which lies near the cofactor and Phe97) (cf Figure 8.5).

The motions for bulky pyrrolidine carboxamides being represented by pc-c7a3 (Figure 8.5) can then be summarised as follows:

1. **2° carbonyl group:**

   This group exhibits the maximal mobility in the bound ligand (pc-c7a3) primarily because of the rotation around the C-C bond that connects the 2° carbonyl group

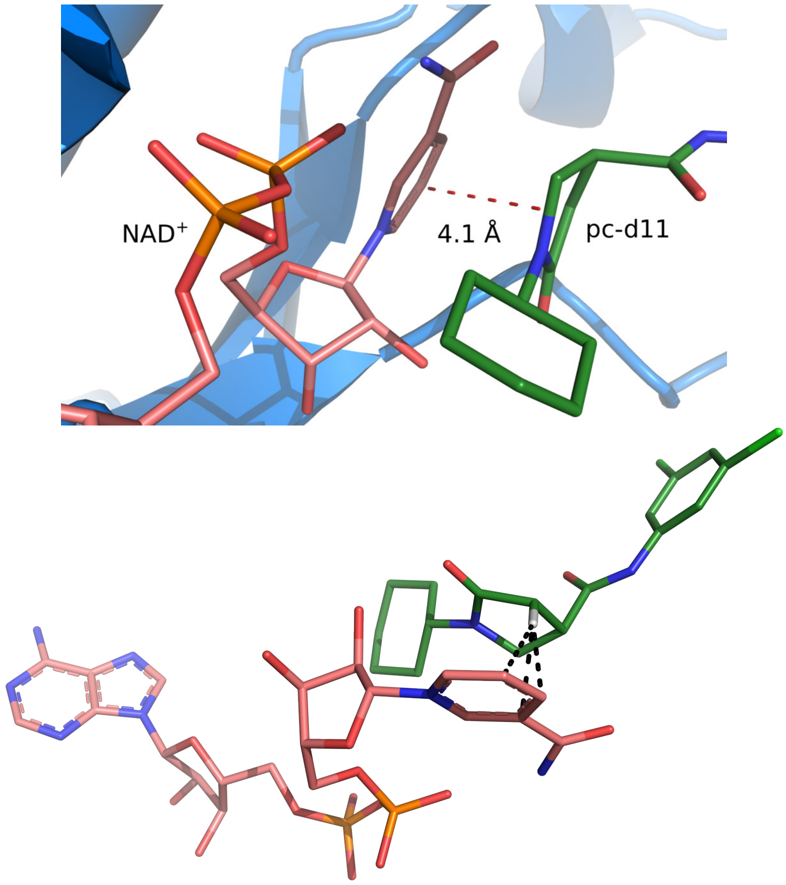


Figure 8.4 A CH-π interaction of the central B ring and the nicotinamide ring of the cofactor (NAD$^+$) shown in red dashed lines. The ring centroid distance has also been indicated. The lower figure depicts the non-optimal pointing of the ligands hydrogen atom to the nicotinamide ring of the cofactor.

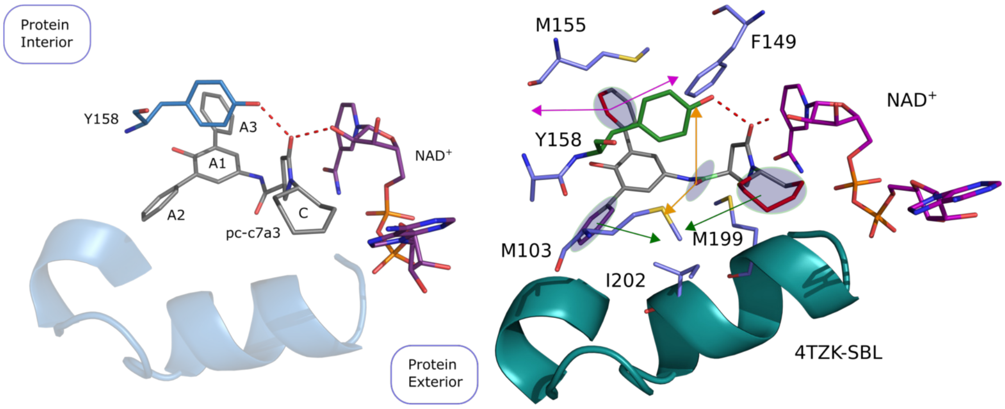


Figure 8.5 The left figure depicts pc-c7a3 along with its annotated rings of the bulky A ring system and the cofactor (violet sticks) and the substrate binding loop. The right figure depicts the Collective motions for maximally mobile atoms (red, violet sticks) of pc-c7a3 along with the direction of the ring motions (green arrows).

to the ring B. The strong movements of the 2° carbonyl group stretches from Phe149 side chain (minor exit portal) to Ile202 situated in the substrate binding loop.

2. **Ring A:**
   The A ring system faces towards the interior of the binding pocket and collectively exhibits varying amplitudes of movements depending on the nature of interactions (and thereby binding) being exhibited by the ligand. The per ring collective motions of the A ring system are as follows:

   - The central ring (A1) exhibits low mobility primarily due the steric hindrance of the rings A2 and A3 that represent a barrier to rotation (cf. Figure 8.2 (small half arrow around ring A1). This hindrance is lacking in case of light pyrrolidine carboxamides that exhibit free rotation of the ring A around the C-N bond.
   - The ring A3 exhibits free rotations around the C-C bond linking it to the central A1 ring. During these motions, it comes in close contacts with the backbone of Phe149 and Asp148.
   - Likewise the ring A3, A2 exhibits ring rotations but much more dampened since it interacts extensively with a key residue lining the hinge region, i.e., Leu207 which moves away from the A2 thereby imparting some movement to its adjacent residues. This results in a motion of the hinge region (a short loop connecting $\alpha 6$ and $\alpha 7$ helices) that ultimately accommodates the approaching 2° carbonyl group.

3. **Ring B:**
   The B ring of bulky pyrrolidine carboxamides exhibits a noticeably increased mobility that sets it apart from light pyrrolidine carboxamides. This indicates a lack of the stabilising van der Waals interactions with the cofactor alongwith an already weak bond with the catalytic residue, i.e., Tyr158. The atoms of this ring indicate a preferential interaction with the sidechain of Met103 (not shown in Figure 8.5).

4. **Ring C**:
   The C ring of bulky pyrrolidine carboxamides exhibits motions that were demonstrated in case of light pyrrolidine carboxamides as well. The ring C demonstrates extensive motions that stretch from the ribose ring to Ile202 of the substrate binding loop. Its the high mobility of this ring that concurrently leads to movement of the key residues of the substrate binding loop as seen in Section 8.3.1.3.

5. In summary, the net movement of the "bulky" pyrrolidine carboxamides can best be described as a **pincer's open-close** motion (Figure 8.5) that varies slightly for each compound of this group.

From the above observations, it can be seen that the 2° carbonyl group and the entire C ring are most mobile for both light and bulky pyrrolidine carboxamides, respectively.

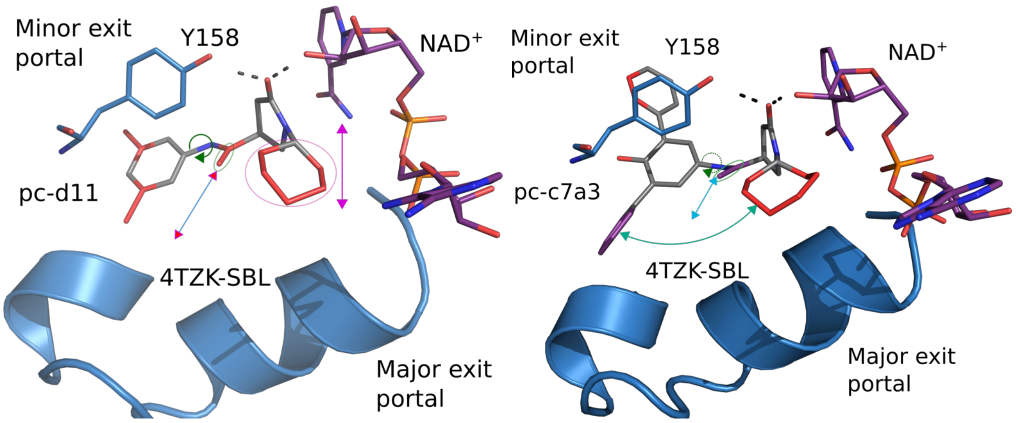


Figure 8.6 Collective motions of the representative ligands for light (pc-d11) and bulky (pc-c7a3) pyrrolidine carboxamides: The ligand pc-d11 exhibits ring flips (green circular arrow) whilst the 2° carbonyl moves towards the substrate binding loop (blue arrow), and the C ring moves up and down (magenta arrows). For pc-c7a3, the 2° carbonyl exhibits a conserved motion (blue arrow) while ring A2 and ring C move simultaneously towards each other (teal arrow). The hindered rotation of the central ring A1 is depicted as dashed green arrow

Additionally, the A ring(s) also exhibit ring rotations, which are more pronounced in the smaller pyrrolidine carboxamides than the larger and bulkier ones, primarily due to less steric hindrance in the former case. Moreover, the collective motions of the representative molecules (Figure 8.6) clearly reveal the differences in the motions of each class of pyrrolidine carboxamides, that are associated with different interactions with the residues of the active site (cf. Section 8.3.1.2) and substrate binding loop respectively (cf. Section 8.3.1.3).

8.3.1.2 Principal modes for active-site residues

The differential movements amongst the individual ligands in the case of light and bulky pyrrolidine carboxamides were reflected in the pronouncedly different collective motions of the key residues. These are as follows:

1. **Light pyrrolidine carboxamides:**

   For the protein-ligand complexes of light pyrrolidine carboxamides, maximal collective motions were observed in the side chains of key residues already known to be critical in stabilising ligand binding i.e., **Ala198, Met199, Ile202 and Val203**. Additionally, the rotation of the phenyl ring and subsequent motions of the substituents on the phenyl ring led to noticeable movements in the sidechain of Phe149 and the entire residue Met155. Amongst all of the aforesaid residues, Met155 was found to move away from the ligand, while Met199 moved in the direction of the 2° carbonyl group (Figure 8.7). The observations suggest that the light pyrrolidine carboxamides induce noticeable collective motions for key residues located at both ends of the molecule, i.e., Met155 and Phe149 from the minor exit portal and Met199 from the major exit portal, respectively.

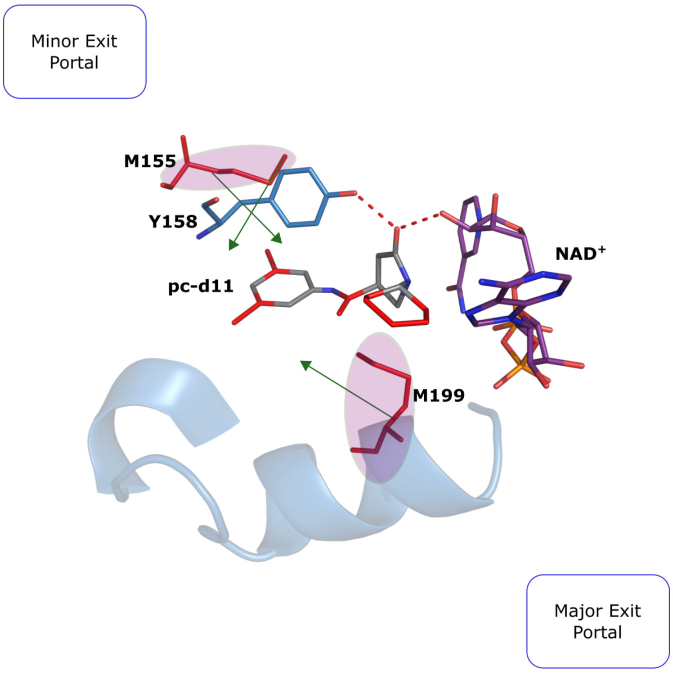


Figure 8.7 Collective motions for maximally mobile active site residues (red sticks) of PDB 4TZK along with their directions (green arrows). The maximally mobile atoms of the ligands have been highlighted as red sticks.

2. **Bulky pyrrolidine carboxamides:**

   In contrast to the light members of the pyrrolidine carboxamide dataset, the active site residues exhibit different collective motions for bulky pyrrolidine carboxamides (Figure 8.8). Figure 8.8 highlights the key residues involved in binding, i.e., Ile202 and Val203 move towards the ligand (2° carbonyl group) as opposed to Met199, which is part of the major exit portal. A key observation in case of bulky pyrrolidine carboxamides was that residues in close proximity of rings A2 and A3 were quite stable that was reflected in their low mobilities in essential dynamics analysis. This clearly differs from the light pyrrolidine carboxamides. These observations suggest that in case of bulky pyrrolidine carboxamides, maximal collective motion was observed in the key residues situated in the substrate binding loop as compared to residues of both minor and major exit portal for the light pyrrolidine carboxamides. From the principal modes, it was also seen that there was a slight movement of the substrate binding loop towards the ligand as indicated by the direction of movement for Ile202 and Val203 (Figure 8.8), respectively.

The essential dynamics analysis for the active site residues of the light and bulky pyrrolidine carboxamides stressed the differences in their collective motions (sampled over a

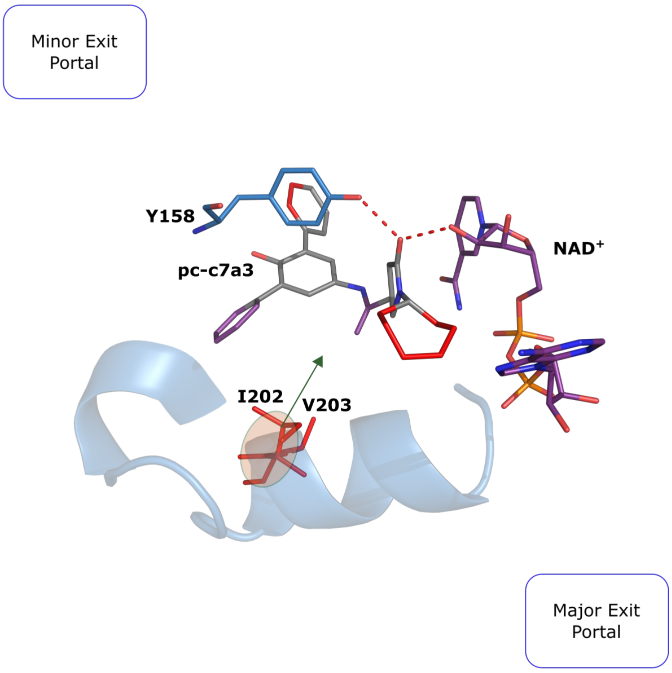


Figure 8.8 Collective motions of maximally mobile active site residues (red sticks) of 4TZK-c7a3 complex along with their directional motion (green arrow). The mobile atoms of the ligand have been depicted as red and violet sticks, respectively.

period of 5 ns per protein-ligand complex). It was seen that key residues from the substrate binding loop, that also constitute a part of the active site (Met199, Ile202, and Val203), collectively move towards the ligand's 2 ° carbonyl group. However, the collective movement (and direction) of the substrate binding loop can only be understood upon examining the collective motion of the entire stretch of residues that span its length. This is been discussed in Section 8.3.1.3.

8.3.1.3 Principal modes of the residues of the substrate binding loop

The collective motions of the residues in the substrate binding loop can be summarised as follows:

1. **Light pyrrolidine carboxamides:**

   Figure 8.9 depicts the collective motions of the residues of the substrate binding loop for PDB 4TZK. It can be seen that the residues exhibiting maximal movement as deemed by the first 5 principal modes are **Arg195, Leu197, Met199, Ile202, Val203, Leu207, Glu209, and Glu210**, respectively. The characteristics of the motions of Met199, Ile202 and Val203 can be confirmed and agree with the discussion in section 8.3.1.2. The side chain of a pair of leucine residues (Leu197 and Leu207) that lie at the beginning and the end of the α-6 helix exhibit anti-parallel motions, with the Leu197 side chain moving away from the binding pocket. Simultaneously, the side chain of Leu207 moves towards the binding pocket. Additionally, the side chain of Glu209 moves into the same direction as that of Leu207, while Glu210 situated at the end of the hinge region (connecting the $\alpha6 - \alpha7$ helices) moves upwards in the direction of the ligand. Collectively, these motions indicate that a part of the substrate binding loop spanned by residues 195-198 moves away from the ligand, whilst the residues 199-209 show a movement towards the ligand.

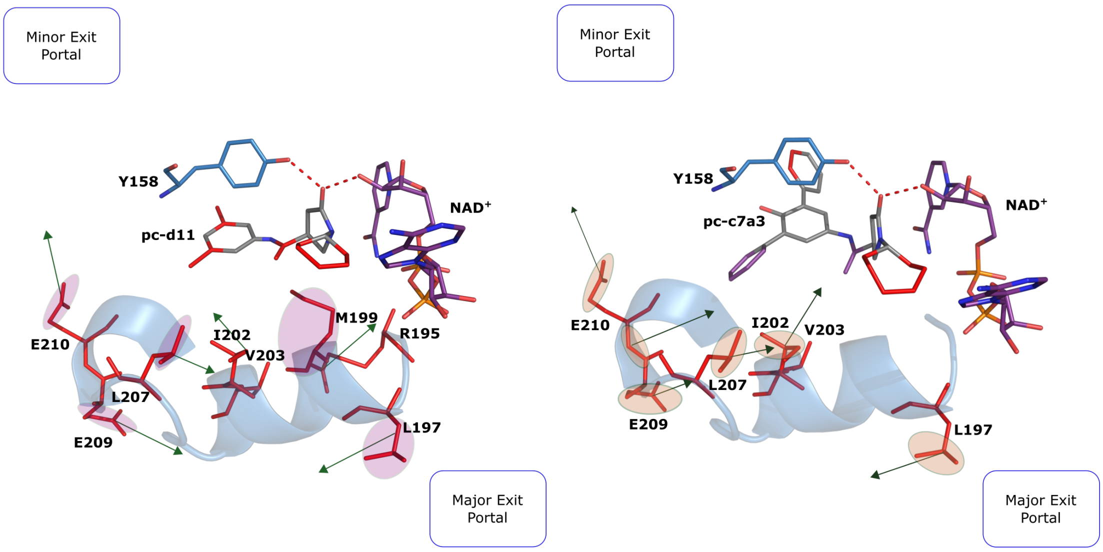


Figure 8.9 Collective motions for maximally mobile substrate binding loop residues of (a) PDB 4TZK (left) and (b) 4TZK-c7a3 (right) along with their directions (green arrows). The mobile atoms of the ligand have been depicted as red (for pc-d11) and as red and violet sticks (for pc-c7a3), respectively.

2. **Bulky pyrrolidine carboxamides:**

   The essential dynamics performed for the substrate binding loop resulted in the following residues being maximally mobile: **Leu197, Ile202, Val203, and L207-Ala211**. The collective motions and directions of these residues can be seen in Figure 8.9. It is clear that the collective motion of these residues are conserved when compared to the case of light pyrrolidine carboxamides. A noticeable difference is the extent of movement of the hinge region that connects the substrate binding loop α-6 and α-7 helices. In case of light pyrrolidine carboxamides, marginal movement of the side chain of the hinge residues (Leu207-Ala211) is observed, as opposed to a strong backbone movement in case of bulky pyrrolidine carboxamides. This strong movement of the hinge region and a part of the substrate binding loop is primarily in response to the characteristic motion of bulky pyrrolidine carboxamides as described in Section 8.3.1.1.

From the above observations, the following conclusions can be drawn:

- The residues of the substrate binding loop show mainly conserved collective motions (towards the ligand) in case of light and bulky pyrrolidine carboxamides, though the residues of the hinge region for the latter exhibit an increased movement towards the ligand.

- A pair of leucines i.e., Leu197 and Leu207 exhibit anti-parallel collective motions that coincide with the movement of the substrate binding loop towards the ligand.
- The role of Met199, Ile202 and Val203 in ligand binding and stabilisation can be corroborated by the essential dynamics analysis.

So far, the previous sections (Sections 8.3.1.1 to 8.3.1.3) portrayed the ”essential” motions for the 3 different atom selections (ligand, active site, and substrate binding loop) for the light and bulky pyrrolidine carboxamides, respectively. These collective motions merely depict the ”linearly” correlated movements of these atom selections upon protein-ligand binding. A significant portion of non-linearly correlated movements that also contribute to the overall motions observed during a MD simulation is missing. In order to account for the non-linearly correlated movements in the protein-ligand complex, the dynamic cross correlation was carried out for the entire pyrrolidine carboxamide dataset, with the same 3 atom selections that were utilised for the essential dynamics analysis.

Table 8.1 Synopsis of findings from Essential Dynamics (EDA) and Dynamic cross correlation (DCC). The right hand side of the table summarises the correlated and anti-correlated motions for light and bulky pyrrolidine carboxamides as exemplified by representative ligands pc-d11 and pc-c7a3, respectively. The upper half of the left hand side describes the maximally mobile parts of the ligand and residues. The lower half of left hand side describes the direction of movement for the maximally mobile parts of the ligand and key residues as deduced from the top 5 principal modes.

	Essential Dynamics			Dynamic cross correlation		
compound class	**mobile ligand groups**	**interacting active site residues**	**SBL**	**nature of ligand movement**	**nature of active site movements**	**SBL movement**
	Maximally mobile ligand parts					
Light	ring C, A ring substituents, 2° carbonyl group	F149, M155, M199	L197, M199, I202, V203, L207, E209, E210	independent and anti-correlated movement for ring C, highly correlated for ring A and B	highly correlated for F97-P99 and M199-V203, anti-correlated amongst all other residues	highly correlated (intra-group) for two subgroups; R195-M199 & S200-G212
bulky	ring C, ring A2, ring A3, 2° carbonyl group	I202, V203*	L197, I202, V203, L207, E209, E210	independent anti-correlated for ring C, weakly correlated for ring A and B	highly correlated for F97-P99 and S200-V203, anti-correlated amongst all other residues	highly correlated for I202-G205, moderately correlated for A198-A201, independent & anti-correlated amongst all other residues

continued on next page

* Both of these residues form a part of the substrate binding loop as well as the active site atom selections.

	Essential Dynamics			**Dynamic cross correlation**		
compound class	**mobile ligand groups**	**interacting active site residues**	**SBL**	**nature of ligand movement**	**nature of active site movements**	**SBL movement**
	Direction of movement					
light	ring C: towards F97-P99, ring A: towards Met155 & I202-V203, 2° carbonyl group : towards I202	M155: away from ligand, M199: towards ligand	I202 & V203: towards ligand, L207: towards ligand & L197 in opposite direction			
bulky	2° carbonyl group: towards M103, Ring A2: towards L207, Ring C: towards I202	I202 & V203*: towards ligand	I202 & V203: towards ligand, L207: towards ligand & L197 in opposite direction			

* Both of these residues form a part of the substrate binding loop as well as the active site atom selections.

8.3.2 Dynamic cross correlation

The light and bulky pyrrolidine carboxamides were analysed separately, with the non-linearly correlated motions being explained for the ligands pc-d11 and pc-c7a3, respectively.

8.3.2.1 Dynamic Cross Correlation for light pyrrolidine carboxamides

1. **Ligand heavy atoms:**

   Figure 8.10 depicts the dynamic cross correlations for pc-d11. The strongly correlated movements of rings B and C can be seen, while ring A can be said to exhibit motions that are independent of ring C and vice versa. The motions of ring B and ring A are weakly correlated. Atoms of the amide group connecting the respective rings exhibiting a moderately correlated motion with ring A. The rest of the atoms making up ring B including the 1° carbonyl group shows almost no correlation with motions of ring A.

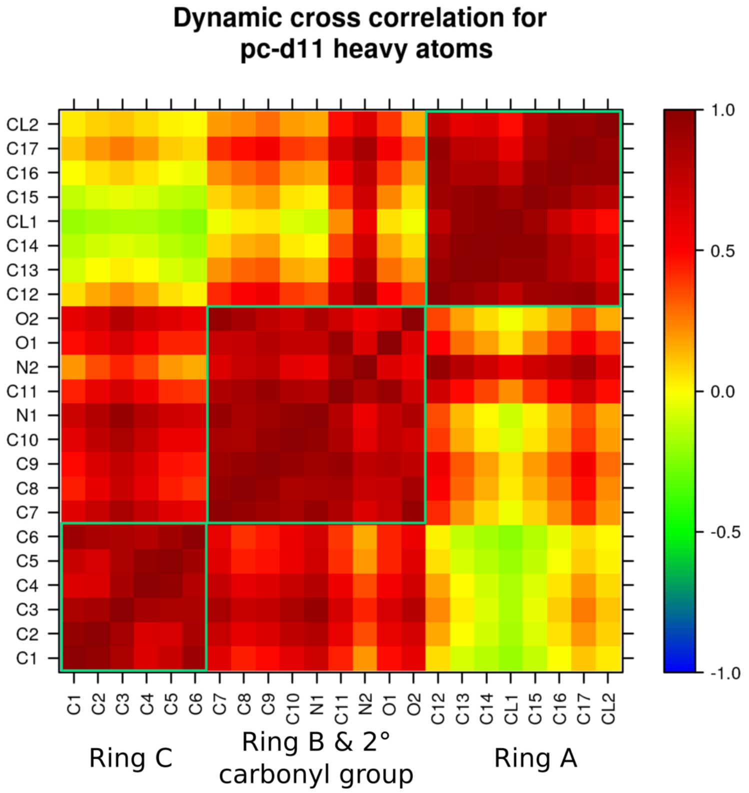


Figure 8.10 Dynamic Cross Correlation for the reference ligand pc-d11: The range of atoms belonging to the respective rings has been marked as blue boxes, with the scale on the right side indicating the dynamic cross correlation range.

2. **Active site residues:**

   Figure 8.11 depicts the dynamic cross correlations for the C-α atoms of active site residues for PDB 4TZK (pc-d11). The dynamic cross correlation and essential dynamics analysis of the active site residues allow the following observations and conclusions to be drawn:

- The residues Phe97 to Pro99, which interact mainly with ring C, show positively correlated motions. With respect to other active site residues, the correlations are weakly negative. The same residues were analysed to exhibit low mobility in the essential dynamics, and hence do not show up in Figure 8.7.
- The residues Met155 and Met199 that have been categorised as maximally mobile in the essential dynamics analysis, exhibit independent motions with respect to Tyr158 as well as each other (Figure 8.7). Furthermore, Met199 shows strong positive correlation with Ser200, Ile202, and Val203 (and vice versa), implying that they move collectively towards the ligand.
- Residues of the catalytic triad (Phe149, Tyr158, and Lys165) show a moderately correlated motion with respect to each other. This merely bespeaks their role in ligand binding. A similar correlation was observed for the residues Met199 to Val203 from the major exit portal. They are mainly involved in stabilising the ligand binding, as already known from literature [50, 52].
- Finally, the pair of leucine residues situated at the beginning and end of the α-6 helix, i.e., Leu197 and Leu207 exhibit moderately anti-correlated movements with respect to each other. This merely supports the observation derived from Figure 8.9.

3. **Substrate binding loop residues:**

The dynamic cross correlation for the substrate binding loop residues of light pyrrolidine carboxamides (Figure 8.12) provides a rather interesting picture, with strongly positively correlated motions observed within two subgroups of residues that span the entire length of the substrate binding loop: residues Arg195 to Met199 (subgroup I) and Ser200-Gly212 (subgroup II), respectively. However, the two groups just exhibit weakly correlated motions when compared with each other. Furthermore, the residues situated at either ends of the substrate binding loop (Arg195 and G212) exhibit no correlation amongst each other. Collecting the information from essential dynamics and dynamic cross correlation, it can be seen that the part of the substrate binding loop (Arg195-Met199) near the C ring moves away. It makes space for the highly mobile C ring atoms. On the contrary, the remaining part of the substrate binding loop being spanned by Ser200 to Gly212 moves slightly towards the ligand as seen from the projection vectors (arrows) of the essential dynamics analysis.

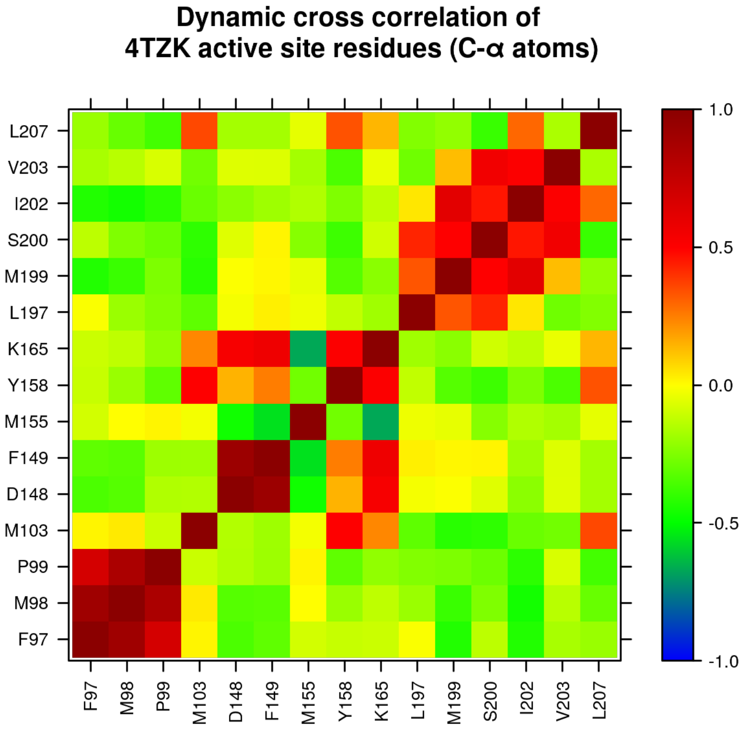


Figure 8.11 Dynamic Cross Correlation for the active site residues of 4TZK with the scale on the right side indicating the dynamic cross correlation range.

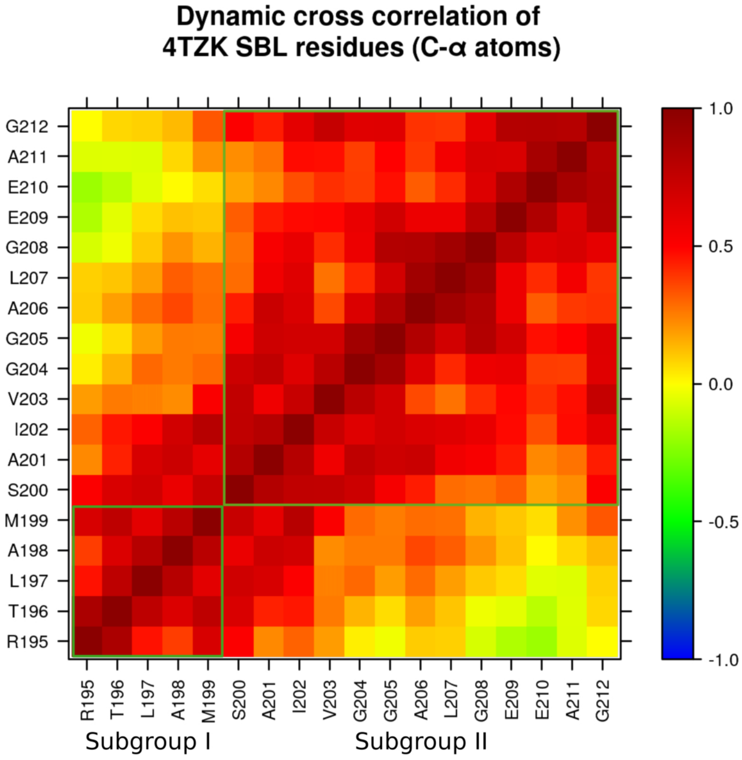


Figure 8.12 Dynamic Cross Correlation of the SBL residues of 4TZK with the scale on the right side. The subgroup of residues exhibiting different direction of movement has been marked and annotated accordingly.

Summing up the observations from the DCC and the ED for light pyrrolidine carboxamides, the collective motions of ligand and active site residues become evident. The most important aspect for light pyrrolidine carboxamides is the noticeable collective motion of key residues like Ile202 and Val203 towards the ligand. This motion is evident from the projection vectors (arrows) as well as the high positive correlation amongst these residues. Furthermore, a slight movement of a part of the substrate binding loop away from the ligand is also evident from the same analyses, apparently in order to avoid the clashes with the highly mobile C ring atoms. Support for this comes from Figure 8.11, where Leu197 exhibits moderately anti-correlated motion with respect to residues Met199 to Val203 as well as Leu207. The effect of ligand binding upon the substrate binding loop can be rationalised as Arg195-Met199 moving away from the ligand while the rest of the residues move slightly towards the ligand.

8.3.2.2 Dynamic Cross Correlation for bulky pyrrolidine carboxamides

1. **Ligand heavy atoms:**
   The dynamic cross correlations of the heavy atoms for the representative bulky pyrrolidine carboxamide, pc-c7a3 are depicted in Figure 8.13. The complex motions of pc-c7a3 and bulky pyrrolidine carboxamides in general become evident, given the varying degree of correlations that the heavy atoms of ring A and C exhibit with respect to each other. The following observations can be made pertaining to the ring movements:

   - **Ring B:** The atoms of ring B (blue box) exhibit anti-correlated motions with respect to the rest of the ligand, with the exception of ring A1 and A2, with which they exhibit very weak or no correlation. Furthermore, ring B and the amide group (atoms C5, N2, O2) that connects it to the A ring system exhibit modestly correlated motions, which signifies that the collective displacements of ring B affect that of the amide group containing the 2 ° carbonyl group.
   - **Ring A:** The ring A system consisting of three aromatic rings exhibit the most varying correlations with respect to other rings, and this is characteristic for all bulky pyrrolidine carboxamides. This distinguishes all bulky pyrrolidine carboxamides from their lighter counterparts. Rings A1 and A2 exhibit moderately correlated motion with respect to each other, while both exhibit moderately anti-correlated motions to rings A3 and C. Furthermore, the motions of ring A1 are mostly limited due to the steric hindrance (from the protein) while the rings A2 and A3 exhibit the characteristic rotations around the C-C bond (as seen from the dihedral analysis). Surprisingly, the movements of ring C and A3 were found to be moderately correlated although the rings are separated by a sizeable number of atoms. The anti-correlated

motions of ring A2 and C merely lend support to the ”pincer open-close” motions of pc-c7a3 as described in in Section 8.3.1.1.

- **Ring C:** Ring C atoms just like in case of light pyrrolidine carboxamides exhibit maximally anti-correlated motions with respect to the rest of the system except the ring A3 (moderate positive correlation). The correlation in the collective motions of the rings A3 and C can be attributed to the overall direction of movements for both rings as seen from Figure 8.5. During the characteristic pincer open and close cycle, both of them move in the same direction, although ring A3 simultaneously exhibits ring rotations, thus resulting in modestly positive correlations.

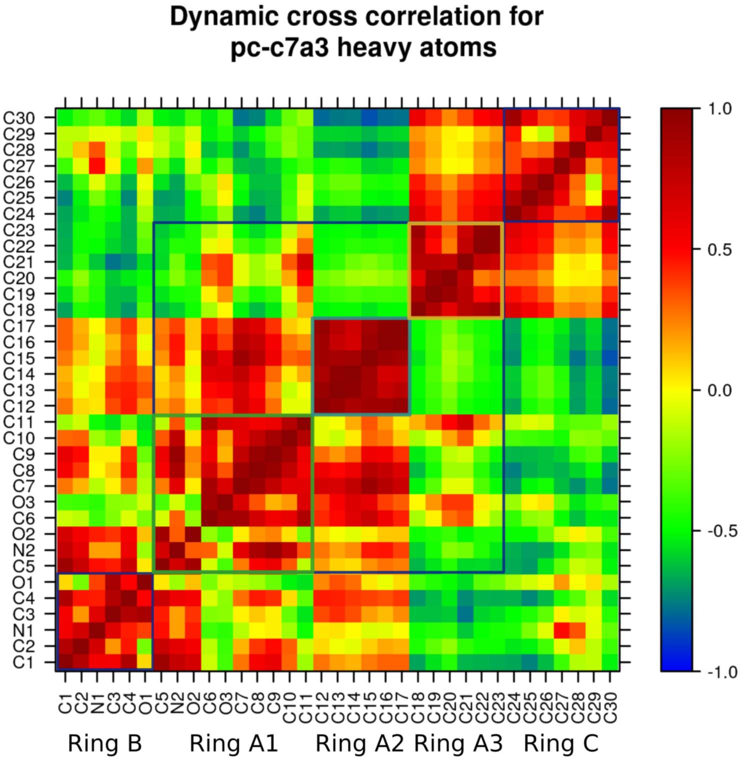


Figure 8.13 Dynamic Cross Correlation for the bulky pyrrolidine carboxamide-c7a3: The range of atoms belonging to the respective rings has been marked as blue boxes while for the A ring system, atoms belonging to different rings have been marked and annotated in different boxes. The scale on the right side indicates the dynamic cross correlation range.

2. **Active site residues:**

The complex movements of the bound ligand (pc-c7a3) feature markedly different correlations amongst the active site residues, which are depicted in Figure 8.14. It can be seen that, with the exceptions of the subgroups of residues Phe97-Pro99 and Met199-Val203, all other residues exhibit varying degrees of weakly anti-correlated motions with respect each other. The motions of these residue subgroups can be explained in better manner upon consideration of the information from the essential dynamics of active site residues. During the pincer "closing" motion of the ligand, ring C atoms come in close vicinity of Phe97-Pro99, which move away to avoid clashes. However, these motions were not shown to be significant in the essential dynamics analysis of the active site. During the pincer "open" phase, ring C moves back, while simultaneously Met199-Val203 move towards the ligand. The Phe97-Pro99 subgroup then returns back to their original position. The anti-correlated movements amongst the Phe97-Pro99 and Met199-Val203 subgroups arise purely from the directions along which they move during the pincer "open-close" cycle of the ligand.

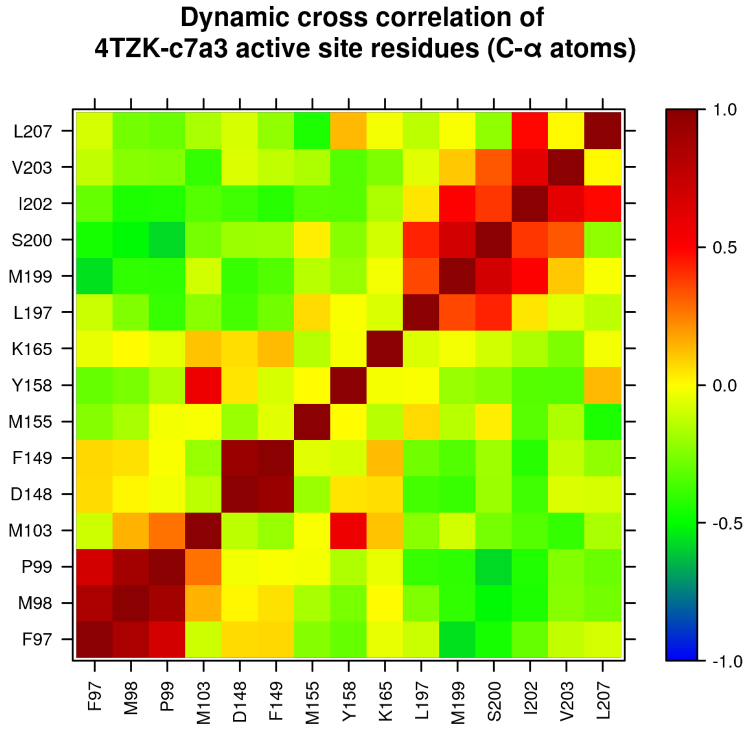


Figure 8.14 Dynamic Cross Correlation for the active site residues of 4TZK-c7a3 with the scale on right side indicating the dynamic cross correlation.

3. **Substrate binding loop residues:**

The dynamic cross correlations of the substrate binding loop residues for bulky pyrrolidine carboxamides exhibit a completely different pattern of correlations when compared against light pyrrolidine carboxamides (Figure 8.15). With a notable exception of Met199 to Val206, all other residues of the substrate binding loop mostly exhibit moderately anti-correlated (negatively correlated) motions that signify that the residues move in opposite directions with respect to each other. The dynamic cross correlations for the SBL merely lend support to the observation that the residues Met199 to Val203 move slightly towards the ligand (pc-c7a3) (Figure 8.9). Furthermore, the anti-correlated (negative correlation) motions for the remaining residues of the SBL can be attributed to the characteristic pincer ”open-close” motions of the bound ligand (pc-c7a3).

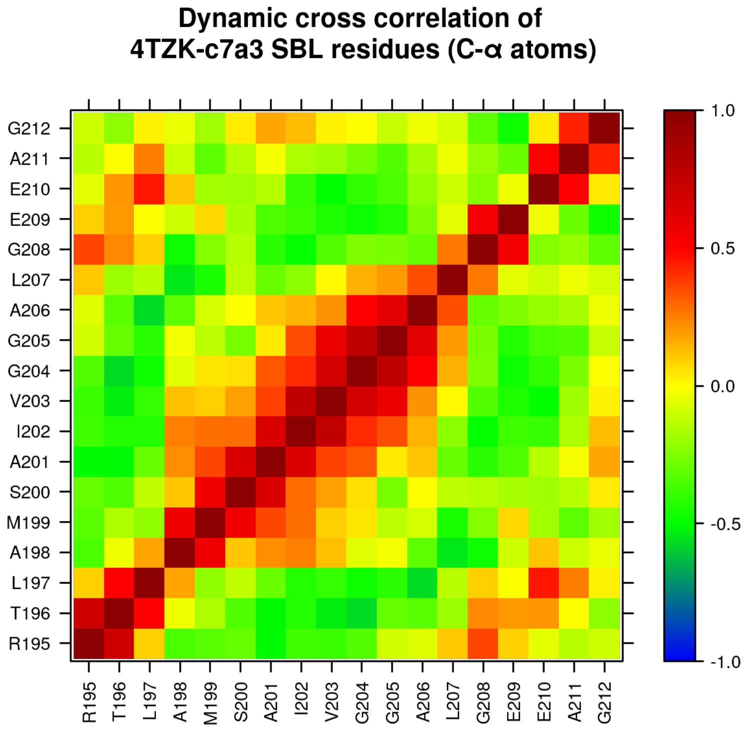


Figure 8.15 Dynamic Cross Correlation for the SBL residues of 4TZK-c7a3 with the ramp on the right side indicating the dynamic cross correlation.

In summary, for bulky pyrrolidine carboxamides, it can be said that the observed motions for the representative ligand (pc-c7a3) appear to be more complex than for the light pyrrolidine carboxamides. This manifests in higher RMSD as can be seen from the RMSD analysis of their MD simulations. Moreover, essential dynamics and dynamic cross correlation of pc-c7a3 and residues of the active site and the SBL revealed that only the substrate binding loop was affected by the ligand movements. On the contrary, for pc-d11, residues of the active site and the SBL simultaneously exhibited movements in response to the ligand motions. Considering this, it can be safely said that bulky pyrrolidine carboxamides mainly bring about the motions for the residues of the substrate binding loop as opposed to both active site and SBL in case of light pyrrolidine carboxamides.

8.4 Discussion

The extensive molecular dynamics simulations carried out for the pyrrolidine carboxamides contain a plethora of structural information about the important movements that are critical in understanding their binding. Studying these movements can aid in optimising the binding of the ligands to InhA and drive structure-based design. However, a myriad of motions spanning from a lower ("essential") to a higher frequency (vibrational) occur during a typical MD simulation. The former type of motions are often slow (occur with a slow frequency), but contribute maximally to the overall protein motions as well as protein function.

In case of light pyrrolidine carboxamides, the 2° carbonyl group as well as the C ring were found to be the maximal contributors followed by the A ring. However, in the case of bulky pyrrolidine carboxamides, given their sheer size and multiple rings, a different pattern of movement was observed that resembled a pincer's open and close cycle. The maximal contributors to the ligand movement in the latter case were the rings A2 and A3. The 2° carbonyl group and ring C atoms exhibited maximal mobility that was found to be conserved in both light and bulky pyrrolidine carboxamides.

The differences in the characteristic motions of light and bulky pyrrolidine carboxamides were manifested in a conserved pattern of correlated/anti-correlated movements for the active site residues. A completely different pattern was observed in case of the substrate binding loop residues. For active site residues of light pyrrolidine carboxamides, Met155 stood out as a solitary residue that exhibited anti-correlated movement with respect to the rest of the system. Additionally, Met199 from the central region of the substrate binding loop was found to approach the ligand. This implies adjustments in the SBL upon its association with a ligand. The corresponding essential dynamics analysis of the SBL for bulky pyrrolidine carboxamides revealed that Ile202 and Val203 moved towards the ligand instead of Met199. Upon comparison of the dynamic cross correlation heatmaps for pc-d11 and pc-c7a3, it follows that light pyrrolidine carboxamides induce modest to strongly correlated motions at both ends, i.e., in the minor exit portal near the catalytic triad (Phe149, Tyr158, Lys165) as well the major exit portal (region around the C ring including the SBL). The bulky pyrrolidine carboxamides owing to their characteristic motions affect only the residues of the major exit portal.

These conclusions are supported by the findings stemming from the dynamic cross correlation for the substrate binding loop residues. A strong positively correlated motion in the direction of the ligand was observed for light pyrrolidine carboxamides as opposed to only a slight movement for bulky pyrrolidine carboxamides. From the same analyses, it also emerged that the central part of the α-6 helix (residues Met199 to Ala206) always moves towards the ligand for both light and bulky compounds from the pyrrolidine carboxamide dataset. Finally, a pair of leucine residues (Leu197, Leu207) situated at the beginning and end of the α-6 helix showed conserved motions with opposite directions. Leu197 moved away from the ligand (ring C) while Leu207 moved in the opposite direction (towards ring A2). The collective description of all these motions merely suggests a modest movement of the SBL towards the ligand.

8.4.1 Implications of essential dynamics for driving structure-based drug design

The main question pertaining to the aforementioned analyses was whether the information could be utilised to drive the structure-based optimisation of the pyrrolidine carboxamide scaffold, particularly the bulky pyrrolidine carboxamides, given their higher potency. In the above section, it was concluded that the rings A2, A3, and C are the maximally mobile ones, with the former two also exhibiting ring rotation simultaneously as they move. The stabilisation of these rings via increased interactions with the residues of the binding pocket can be expected to bring about favourable changes that include the increased substrate binding loop movement towards the ligand (i.e. loop closure). This should directly affect the time for which the ligand interacts with the protein and thereby its activity.

Accordingly, using the information from essential dynamics analysis, dynamic cross correlation, as well as from literature [50, 346], several substitutions can be performed on the bulky pyrrolidine carboxamide scaffold aimed at increased protein-ligand interactions. For example, in order to stabilise the C ring motions, it can be replaced with a m-chloro/bromo-substituted phenyl ring that is in conformity with results of Kumar et al. [346]. Similarly, the motions of ring A2 and A3 can be attenuated by suitable substitutions that increase the H-bonding interactions or steric hindrance to their ring rotations. Finally, the logP and the synthesizability of the molecules have to be taken care of whilst choosing the substitutions. The resultant molecules and the evaluations of the same are described briefly in Chapter 9.

Chapter 9

Design and analysis of new pyrrolidine carboxamides

9.1 Introduction

The Chapters 7 and 8 highlighted the comparatively “rigid” and “mobile” parts of the ligand as well as the key residues along with the direction of their maximal variance, respectively. In the Chapter 7, the relationship in between the structure of the ligand and its observed dynamics for pyrrolidine carboxamides was thoroughly evaluated. The essential dynamics of bulky and light pyrrolidine carboxamides alongwith the residues of the active site and the substrate binding loop highlighted their principal motions in MD simulations. An important aspect of the analyses performed in the aforementioned chapters was their utility in the structure-based optimisation of pyrrolidine carboxamides, especially the bulky ones. In line with the information gathered from Chapters 7 and 8 and the SAR of pyrrolidine carboxamides from literature [50, 52], the design of new pyrrolidine carboxamides was steered in line with the following direction:

1. Since most potent pyrrolidine carboxamides come from the bulky subgroup especially with the 3,5 diphenyl phenol system as the A ring, it was decided initially to retain the entire ring and perform substitutions on ring A2 and A3, especially at the meta positions.

2. The central B ring containing the 1° carbonyl group (Figure 7.3) as well as the adjacent amide group were retained mainly because:

   - Substitution of the B ring at 3’ and 4’ positions or introduction of a double bond in between these carbon atoms led to nearly complete abolishment of activity [52].
   - The 1° carbonyl group or any group capable of forming dual H-bonds with the cofactor and Y158 is critical as seen from numerous InhA inhibitors published in literature [50].
   - Upon comparison of pc-c7a3 and pc-p28, one can observe their structural similarities, similar natures of the A and B rings. The cycloalkyl rings of the two differ by a single carbon atom (Figure 9.1). However, inspite of

this, docking predicted different binding modes for the same (Figure 3.21). Moreover, pc-p28 is less potent than pc-c7a3 by an order of magnitude. This highlights the importance of a simple amide group as a linker for the two rings. As a result, the amide group linking A ring system to the B ring too was left untouched.

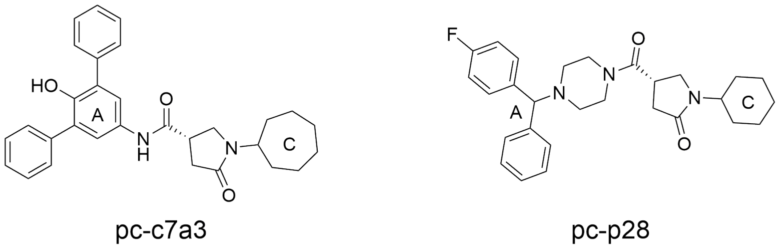


Figure 9.1 Structures of pc-c7a3 and pc-p28, with the A and C rings labelled individually, while the conserved pyrrolidin-2-one substructure represents the B ring.

label

3. In regards of the C ring, as mentioned in Section 8.4.1, the cyclohexyl ring or phenyl ring suffices, though an increase or decrease in the number of carbon atoms (and thereby ring size) led to decreased InhA inhibitory activity. The C6/C7/C8 cycloalkyl rings were found to be quite mobile, mainly because they faced the exterior of the protein. In order to decrease the motions of ring C and increase its interactions with the residues of the substrate binding loop, the ring C was replaced with a unsubstituted or a meta-substituted phenyl ring (all halogens except iodine, -CF_3, and -CH_3 group).

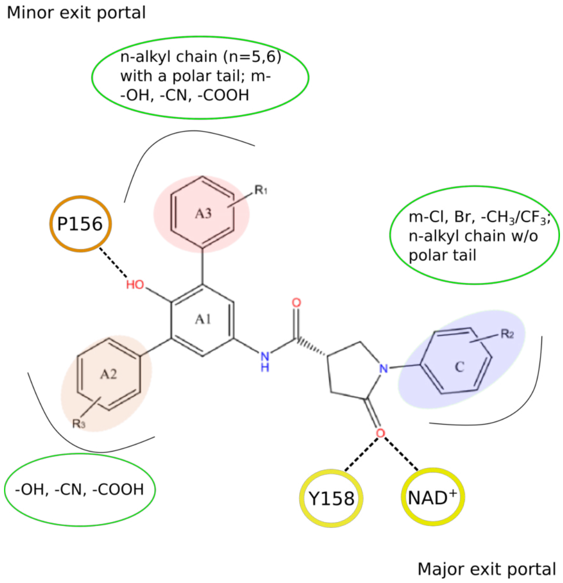

Minor exit portal
n-alkyl chain (n=5,6) with a polar tail; m- -OH, -CN, -COOH
P156
A3
R1
HO
A1
O
N H
m-Cl, Br, -CH3/CF3; n-alkyl chain w/o polar tail
A2
R3
N
C
R2
O
-OH, -CN, -COOH
Y158
NAD+
Major exit portal

Figure 9.2 Core bulky pyrrolidine carboxamide scaffold with various substitution possibilities being highlighted.

4. Moreover, from slow tight binding inhibitors of InhA, e.g., PT-70 and PT92 [50], it was seen that the C5-C8 n-alkyl chain stabilised the ligand binding orientation by increased van der Waals interactions with the active site residues. Accordingly, the A3 ring or C ring were replaced by n-alkyl chain with 6-8 carbon atoms, primarily to stabilise the ligand binding. Additionally, to further stabilise the position of the alkyl chain within the binding pocket, polar groups/rings were added to increase H-bonding with binding pocket residues.

The end result of the above scheme were a total of 20 compounds whose structure and a general synthesis scheme for a representative compound has been depicted in Figures 9.3 to 9.5, with the molecules b1-b11 being initially designed and analysed with various *in-silico* methods (Section 9.3) in order to ascertain their activity as well as validate their structure-based design. Thereafter, based on the initial results of the 11 compounds, the structure-based optimisation was iterated to yield another 9 putative InhA inhibitors (b12 - b20).

Figure 9.6 describes the overall workflow for the current analyses. Initially, the binding orientations for b1 to b11 were obtained by mutating pc-c7a3 and pc-c6a3 to the respective compounds followed by scoring them in-place with XPscore, with pc-b1 to pc-b7 obtained from pc-c7a3, while the binding orientations for the remaining molecules (b8 - b11) being obtained from pc-c6a3. These compounds (and later b12 to b20) were subsequently subjected to docking in InhA using induced fit considering their molecular size, followed by rescoring with DrugScoreX and SFCscore. The docking (and subsequent MD simulations) was mainly performed to test the hypotheses that the alternate binding observed for pc-c6a3 was artefactual in nature. If the binding orientations for any of the molecules gets predicted as inverted much like pc-c6a3, then these ligands must exhibit the same behaviour as that of pc-c6a3 in all analyses, i.e., low docking scores, high RMSD values and weak to almost negligible H-bonding. On the contrary, if a majority of the new ligands get docked like pc-d11 (reference ligand), then the inverted binding mode can be considered as an artefact, underscoring the problems faced by the docking algorithm in accurate scaffold placement. The detailed results of all analyses will be discussed in the Section 9.3.

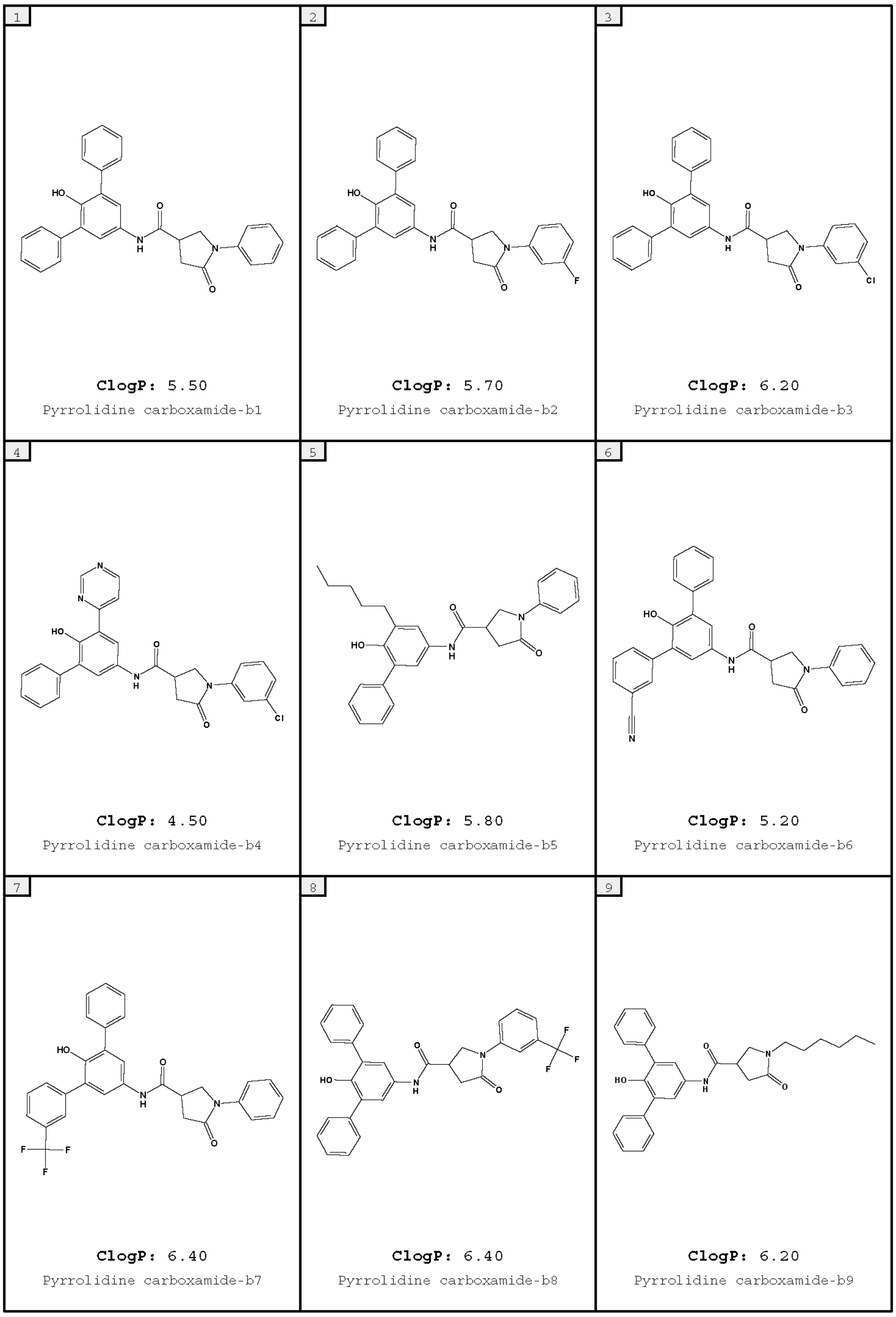

1
ClogP: 5.50
Pyrrolidine carboxamide-b1
2
ClogP: 5.70
Pyrrolidine carboxamide-b2
3
ClogP: 6.20
Pyrrolidine carboxamide-b3
4
ClogP: 4.50
Pyrrolidine carboxamide-b4
5
ClogP: 5.80
Pyrrolidine carboxamide-b5
6
ClogP: 5.20
Pyrrolidine carboxamide-b6
7
ClogP: 6.40
Pyrrolidine carboxamide-b7
8
ClogP: 6.40
Pyrrolidine carboxamide-b8
9
ClogP: 6.20
Pyrrolidine carboxamide-b9

Figure 9.3 New pyrrolidine carboxamides designed with help of information from literature and essential dynamics analysis of MD simulations.

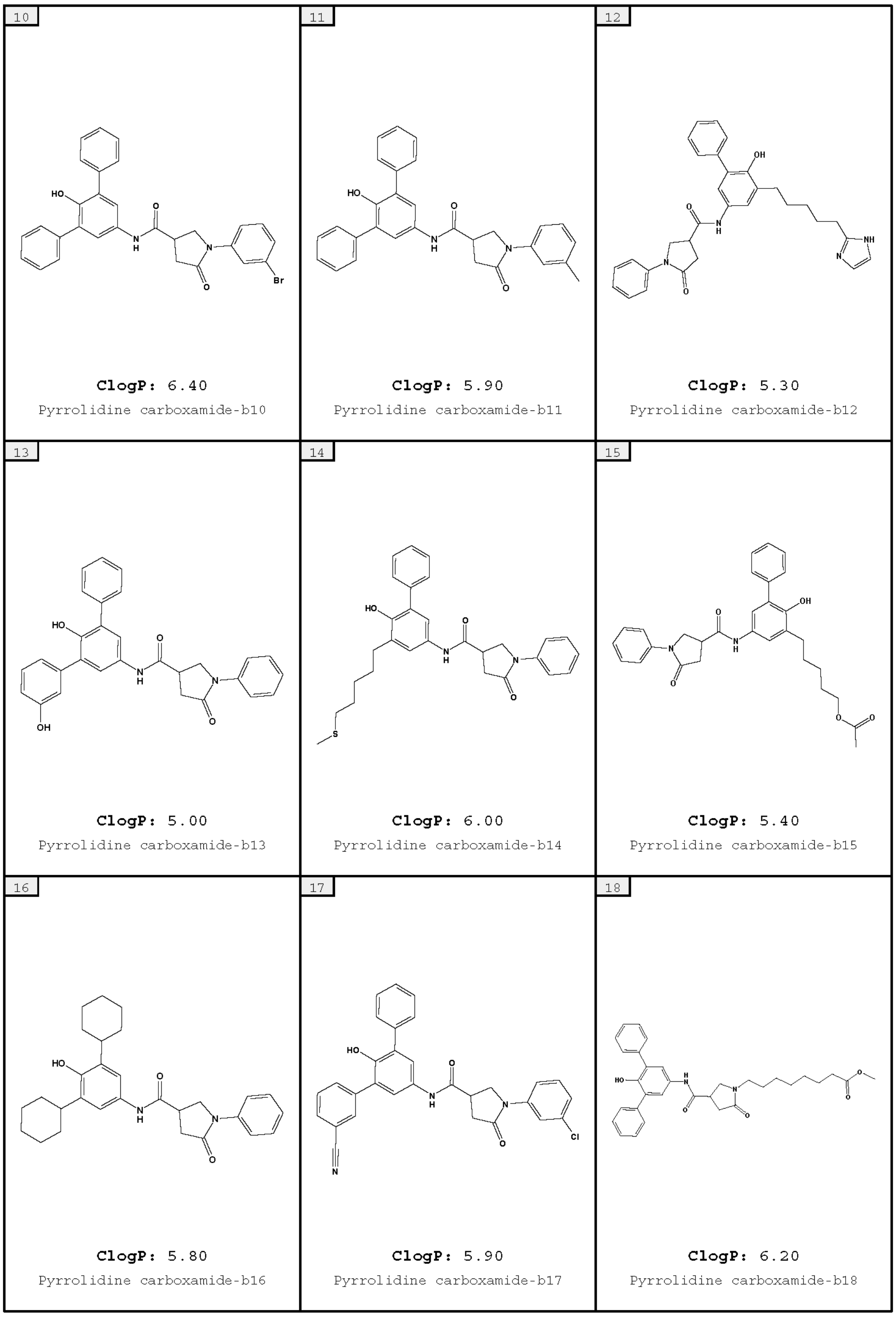

10
ClogP: 6.40
Pyrrolidine carboxamide-b10
11
ClogP: 5.90
Pyrrolidine carboxamide-b11
12
ClogP: 5.30
Pyrrolidine carboxamide-b12
13
ClogP: 5.00
Pyrrolidine carboxamide-b13
14
ClogP: 6.00
Pyrrolidine carboxamide-b14
15
ClogP: 5.40
Pyrrolidine carboxamide-b15
16
ClogP: 5.80
Pyrrolidine carboxamide-b16
17
ClogP: 5.90
Pyrrolidine carboxamide-b17
18
ClogP: 6.20
Pyrrolidine carboxamide-b18

Figure 9.4 New pyrrolidine carboxamides designed with help of information from literature and essential dynamics analysis of MD simulations.

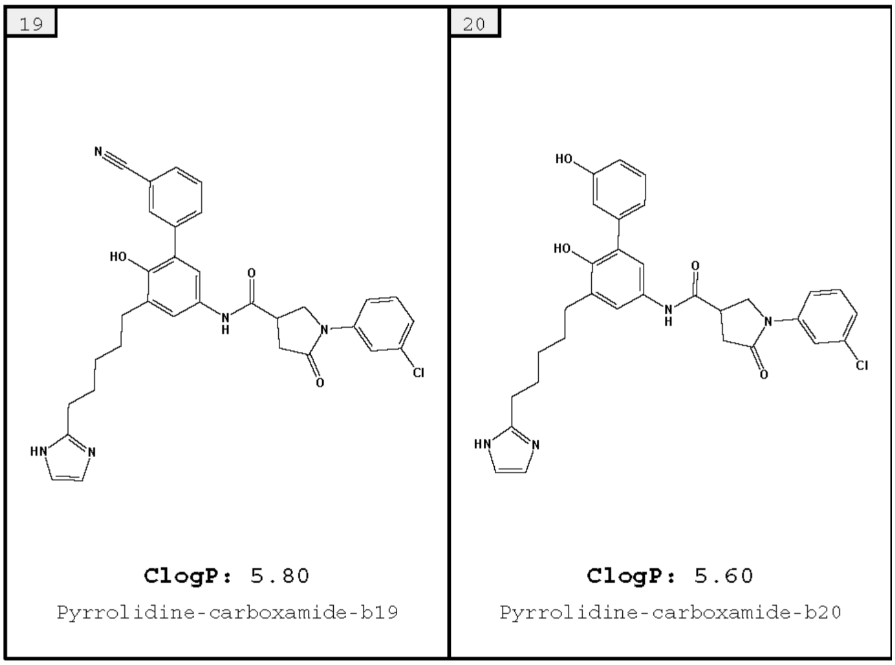


Figure 9.5 New pyrrolidine carboxamides designed with help of information from literature and essential dynamics analysis of MD simulations.

An advantageous prospect of the said molecules is that for a majority of the molecules, the corresponding building blocks are available commercially, making their synthesis and testing relatively easier. Finally, the novelty of the designed molecules was ascertained by performing a thorough scaffold search in PubMed and Scifinder®, with the core pyrrolidine carboxamide scaffold as a search query (cf. Section 9.4).

9.2 Materials and Methods

As mentioned earlier, the designed compounds were subjected to a variety of *in-silico* evaluations in order to ascertain an improvement in terms of various parameters, over their parent compounds (pc-c7a3/c6a3) and the reference compound for all pyrrolidine carboxamides, i.e. pc-d11. Accordingly, the compounds pc-b1 to b20 were subjected to the following analyses:

1. **Mutation and scoring in place using XPscore**[1]
2. **Activity prediction using LIE**[2]
3. **Docking and Rescoring**
4. **Activity classification using XPscore-SFC290p based logreg model**

[1] only for compounds b1 to b11

[2] only for compounds b1 to b11

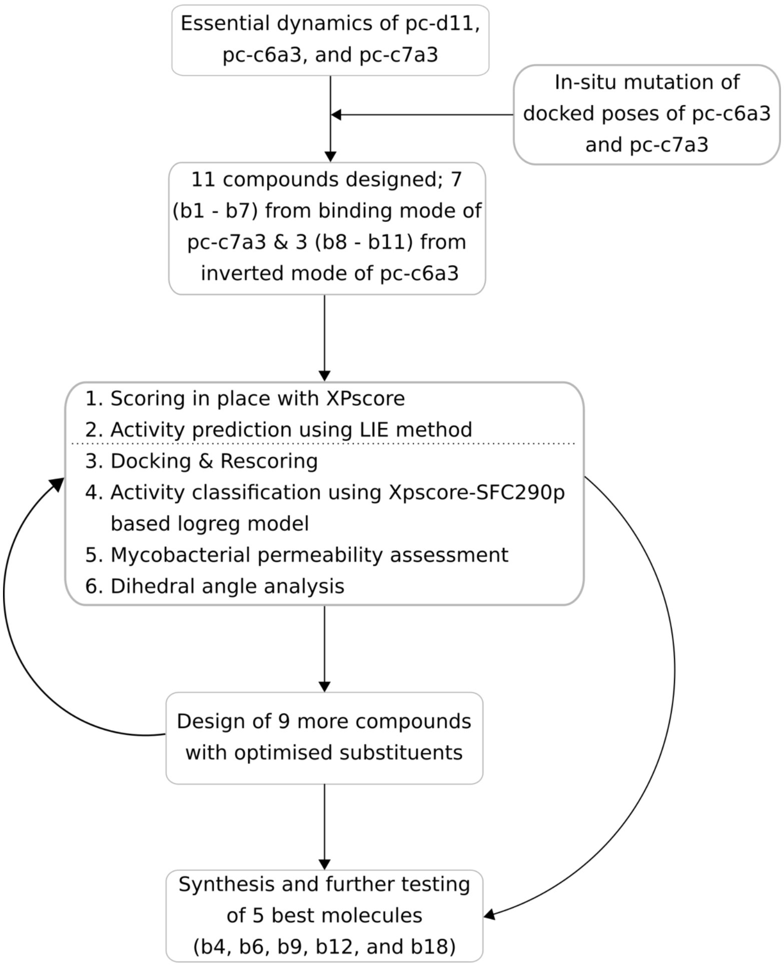

Essential dynamics of pc-d11, pc-c6a3, and pc-c7a3
In-situ mutation of docked poses of pc-c6a3 and pc-c7a3
11 compounds designed; 7 (b1 - b7) from binding mode of pc-c7a3 & 3 (b8 - b11) from inverted mode of pc-c6a3
1. Scoring in place with XPscore
2. Activity prediction using LIE method
3. Docking & Rescoring
4. Activity classification using Xpscore-SFC290p based logreg model
5. Mycobacterial permeability assessment
6. Dihedral angle analysis
Design of 9 more compounds with optimised substituents
Synthesis and further testing of 5 best molecules (b4, b6, b9, b12, and b18)

Figure 9.6 Workflow for the analysis of the molecules derived via structure-based optimisation of the pyrrolidine carboxamides-c6a3 and c7a3, respectively.

5. **Mycobacterial permeability assessment**

6. **Dihedral angle and distance analysis**

9.2.1 Mutation and scoring in place

This was a preliminary exercise to ascertain the effects of the various substitutions on the core pyrrolidine carboxamide scaffold, evident from the comparison of their scores with that of the reference ligand (pc-d11) as well as their parent compounds i.e., pc-c7a3 and pc-c6a3, respectively. The rationale for the comparison was simple: a favourable substitution should result in a noticeable increase in the rescoring values, whilst unfavourable substitutions should get a lower score. Accordingly, the binding orientations of compounds b1 to b11 were obtained via mutation of docked poses of pc-c6a3 and pc-c7a3, respectively (Figure 9.6). The activity prediction using LIE method followed this endeavour. In order to obtain reliable binding orientations for the compounds b1 to b11 (and later b12 to b20), molecular docking with induced fit was performed.

Accordingly, the mutated poses of b1 to b11 were not considered for any other analyses other than docking and rescoring with DrugscoreX and SFCscore. The results and discussion pertaining to this exercise has been covered in Section 9.3.

9.2.2 Affinity prediction using LIE method

For a small subset of designed compounds (N=11; b1 to b11), the affinity prediction was carried out using the LIE method. The values of empirical parameters $\alpha,\ \beta, \text{and}, \gamma$ were derived from the best LIE model described earlier in Chapter 4. The affinity prediction using the LIE method served two purposes:

1. Assessment of the beneficial/unfavourable effects of the substituents on the overall binding free energy.
2. Rigorous assessment of the orientation predicted by docking by evaluating bound state in a 5 ns MD simulation.

9.2.3 Docking and rescoring

In order to obtain reliable binding orientations for the deigned pyrrolidine carboxamides, the docking and pose selection protocol from Chapter 3 was extensively put to use. The pose selection and parameters used for docking were exactly identical to those used for bulky pyrrolidine carboxamides, as was the rescoring scheme. An additional purpose behind the molecular docking was to further ascertain the hypotheses pertaining the artefactual nature of the inverted binding mode as observed for pc-c6a3. Starting from inverted input conformations, if the new molecules got docked in a crystal structure like conformation, then it would signify that pc-c6a3 gets trapped in a local minimum during its placement in the binding pocket. Since it is unable to escape the minima, the inverted conformation gets selected as a top pose. This rationale holds true because the predicted binding mode for any compound is **independent** of its input conformation. Moreover, use of induced fit would aid in ascertaining the dominant binding mode for altogether new compounds that share a common scaffold with molecules of known activity.

Furthermore, the induced fit docking and subsequent rescoring would enable one to:

1. Ascertain the authenticity of the poses generated by mutation *in-situ.*
2. Investigation of any new interactions with the active site residues of InhA.

The results of the molecular docking and rescoring are depicted in Tables 9.1 and 9.2, respectively.

9.2.4 Activity classification using XPscore-SFC290p based logreg model

The process of classifying the designed molecules as active/least active closely followed molecular docking, since the activity-based classification model requires the docking score and SFC290p values as input. Accordingly, the newly designed molecules were subjected to an activity-based classification using the XPscore-SFC290p based model depicted in Figure 4.7 and table 4.5. This approach was primarily used to validate whether the designed molecules were highly active and thereby represented improvement over their parent compounds and the reference ligand or not. The results of this analysis are depicted in Table 9.3.

9.2.5 Mycobacterial permeability assessment

One of the main barriers for any compound that may exert anti-tubercular action are the thick and waxy mycolic acids that form a part of the mycobacterial cell wall. MycPermcheck [53], is a tool that provides for rapid and reliable estimation of the mycobacterial cell wall permeability for a given test molecule. MycPermcheck takes in various descriptors of the test molecule as calculated by QikProp or Padel [347, 348], and gives the probability for the molecule permeating the mycobacterial cell wall.

In the current case, the requisite physiochemical descriptors of the designed molecules were calculated using Qikprop and their mycobacterial permeability was assessed using MycPermcheck 1.1. The results of the same are depicted in Table 9.3. Given the fact that all of the new molecules were based of active bulky pyrrolidine carboxamides, the chance that all of them were permeable across the mycobacterial cell wall was quite high. Furthermore, from this analysis, one could observe in a qualitative fashion the effects of various substituents on the overall mycobacterial permeability.

9.2.6 Dihedral angle and distance analysis

As seen from Chapter 7, the light and bulky pyrrolidine carboxamides exhibit a wide variety in movements that was clearly dependent on their binding modes and the substituents on the A ring, with both factors being closely tied with the overall potency of the molecules. Accordingly, the potent molecules exhibited low fluctuations in both dihedral angles as well as the RMSD values (C-α and bound ligand RMSD). There was a close association of the binding mode and the overall nature and strength of hydrogen bonding, which was depicted by the ligand-cofactor and ligand-Tyr158 distances, respectively. These analyses provided a means of assessing the overall nature of binding for the designed ligands. Accordingly, the newly designed molecules were subjected to the dihedral angle and donor-acceptor atom distance analysis. For this purpose, the 5 ns MD simulations of the ligands bound to InhA were utilised, with the raw data being obtained with VMD-1.9.1 and results being plotted with the statistical framework R [273] and associated packages. The results of this analysis are covered in Section 9.3.4.1.

9.3 Results

9.3.1 Docking and rescoring of new pyrrolidine carboxamides

The results of the molecular docking are depicted in Table 9.1. From the table, certain trends were observed that can be enlisted as follows:

1. All of the molecules that were scored "in-place" with XPscore showed a noticeable increase in predicted binding energy as compared to their parent compounds (pc-c6a3, pc-c7a3) and reference compound (pc-d11), respectively.

2. For a sizeable number of compounds (b4, b5, b7, b8, b10, and b11), the docking scores were markedly less than the scores of their poses obtained via in-situ mutation (Table 9.1). Of these, all molecules except b10 got docked in an inverted conformation like pc-c6a3 (Figures 7.23 and 9.7). The occurrence of inverted binding orientations for more than 50% of the initial compounds (b1-b11) suggests a general problem of the docking algorithm in proper placement of the bulky ligands. These molecules presented a perfect test for evaluating the hypothesis concerning the inverted binding mode of pc-c6a3.

3. The molecules with inverted binding modes (b4, b5, b8, and b11) exhibit lower DrugscoreX and SFC rescoring values as compared to molecules with a reference-ligand-like binding mode. Simultaneously, these molecules demonstrate a noticeable difference to the DrugscoreX and SFC scores of pc-c6a3. A general observation pertaining to all molecules is that they exhibit improved docking and rescoring values as compared to pc-d11 and pc-c6a3 and pc-c7a3 up to some extent. As compared to pc-c7a3, the molecules with inverted binding modes fare poorly, while the *"in-situ"* mutated poses of the same compounds exhibited a contrasting behaviour, i.e., improvement over pc-c7a3. After considering the trends from docking of bulky pyrrolidine carboxamides, the "in-situ" mutated poses for molecules getting docked in an inverted mode may very well be the "correct" binding mode. However, the fact that almost 50% of the designed ligands were predicted to bind in an inverted mode warranted further investigation.

Table 9.1 Docking and rescoring values for mutated (*in-situ*) and docked poses of the designed pyrrolidine carboxamides-b1 to b11. Molecules which got docked in an inverted binding mode have docking scores in bold. Furthermore, the Δ G_{calc} values (in kcal/mol) were calculated using LIE method. The values of the empirical parameters α, β, and γ were obtained from the best performing affinity prediction model from Table 4.4. For DrugscoreX and SFC scoring functions, with the exception of pc-d11, pc-c6a3, and pc-c7a3, the top row consists of values for the *"in-situ"* mutated poses, while the lower row depicts the values of the docked poses.

Parameter	**d11**	**c6a3**	**c7a3**	**b1**	**b2**	**b3**	**b4**	**b5**	**b6**	**b7**	**b8**	**b9**	**b10**	**b11**
ΔG_{exp}[3]	-8.68	-9.28	-8.89	-	-	-	-	-	-	-	-	-	-	-
ΔG_{calc}	-7.05	-6.86	-8.17	-9.37	-9.54	-9.18	-7.44	-8.70	-8.36	-9.57	-7.81	-9.16	-9.20	-8.36
XPscore (in-place scoring)	-9.15	-5.17	-9.18	-14.14	-14.34	-14.28	-13.74	-12.78	-13.89	-13.66	-10.38	-9.43	-10.37	-10.33
XPscore (Docking)	-10.09	-5.17	-12.87	-14.23	-10.89	-15.30	**-3.54**	**-5.35**	-15.01	-8.16	**-5.45**	-11.57	-6.73	**-5.23**
RMSD[4]	0.37	6.46	1.81	1.20	1.19	1.12	6.29	5.45	1.16	1.14	6.42	0.82	1.17	6.42
DrugScoreX	-120.79	-120.99	-181.86	-179.55	-182.33	-184.04	-172.30	-172.09	-179.34	-180.40	-159.63	-144.94	-150.28	-153.43
				-219.45	-214.83	-211.30	-175.04	-130.46	-215.84	-189.69	-183.90	-174.79	-211.02	-169.80
SFCscore[5]														
sfc_229p	7.82	8.25	9.38	9.44	9.59	9.88	9.50	9.03	10.14	10.20	9.27	8.64	9.10	9.09
				10.40	10.49	10.38	9.06	7.82	10.83	9.38	9.57	8.93	10.69	9.24
sfc_290p	8.09	8.04	9.01	9.09	9.26	9.50	9.17	8.64	10.08	10.25	8.62	8.20	8.42	8.36
				10.06	10.27	10.18	8.39	7.34	10.55	8.95	8.91	9.04	10.71	8.68
sfc_rf	7.50	7.73	8.72	8.53	8.74	8.80	8.78	8.19	8.71	8.72	8.24	7.92	8.29	7.93
				9.19	8.96	8.92	8.51	7.24	9.44	8.81	8.63	8.95	9.01	8.17

[3] units: kcal/mol

[4] Substructure RMSD with respect to pc-d11, units Å; only for docked poses

[5] In above table, wherever one encounters a multiple row, the first row corresponds to the mutated poses, which have been scored in place, while the second row refers to the poses obtained from docking using the induced fit protocol.

4. The LIE method demonstrated a slight improvement of the designed ligands over their parent compounds (pc-c6a3 and pc-c7a3) irrespective of their binding mode. The exceptions to this are the molecules b4 and b8 whose binding affinity was predicted to be lower than that of pc-c7a3.

5. In regards of the molecular docking, the appropriate placement of the 1° carbonyl group of the B ring was achieved for all of the designed pyrrolidine carboxamides, irrespective of the binding mode. The proper placement of the B ring can be attributed to the pharmacophore that filters out the poses which do exhibit the 1° carbonyl group (and thereby the B ring) in proper place ensuring the dual H-bonds with Y158 and cofactor, respectively. The substructure RMSD values for majority of the poses from molecular docking were always in the range of 0.80-1.20 Å (Tables 9.1 and 9.2). The compounds with inverted binding modes (b4, b5, b8, and b11), on the contrary, exhibited high substructure RMSD values (around 6 Å).

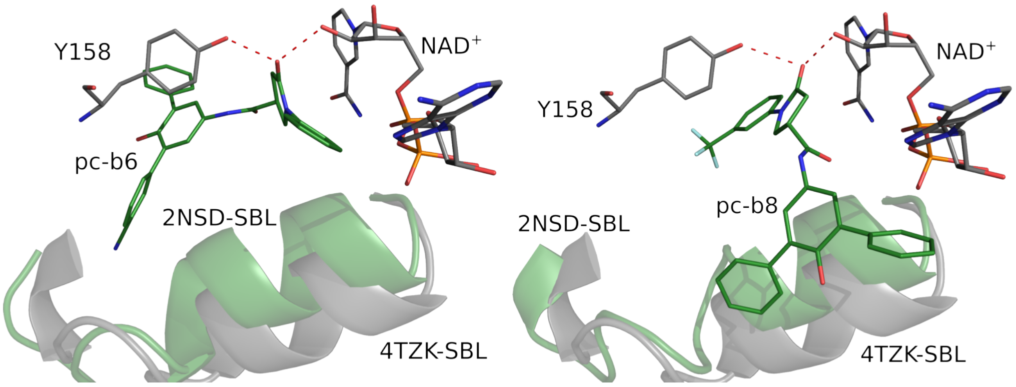


Figure 9.7 The picture on the left depicts pc-b6 with correct placement in the binding pocket. On the contrary, the one on the right has pc-b8 with an inverted binding mode and thereby wrong placement of the ligand.

Going by the results depicted in Table 9.1, it was inferred that the proposed replacements and substitutions of the rings A and C were proceeding in the right direction. Hence, an additional 9 molecules (b12 - b20) were designed. These molecules represented optimal ring replacements and substitutions, derived from the analysis of the earlier compounds. Furthermore, while designing the compounds b12 - b20, attempts to optimise the calculated logP (calculated log of octanol/water partition coefficient, abbreviated a clogP) were also performed. The resulting molecules were thoroughly assessed in a manner similar to their precursors except for the prediction of their binding affinity using the LIE method. This was primarily done to thoroughly investigate the binding modes for the new compounds before the time consuming MD simulations and their affinity prediction using the LIE method. The results of the docking and rescoring for the newest 9 molecules are depicted in Table 9.2.

Upon inspecting the Tables 9.1 and 9.2, the following observations hold true for the compounds b12 to b20:

1. Modest improvements in XPscore as compared to pc-c7a3, which also translates to better InhA inhibitory potential when compared against pc-d11 or pc-c6a3. The compounds b14, b15, b16, b18, and b19 show noticeable decrease in the docking scores. Of these, the compound b16 is a clear outlier since it not only exhibits a big jump in the substructure RMSD (3 Å) but also when compared to the others (substructure RMSD $\ll$ 3 Å). In the case of b18 and b19, the marked decrease in docking scores can be attributed to a scaffold placement that differs significantly as compared to pc-d11 (or pc-c7a3). The deviant poses for the aforementioned compounds still form the required interactions with Y158 and cofactor because of the pharmacophore which ensures the proper scaffold placement.

2. All of the compounds show a definite improvement in DrugscoreX and SFCscore values when compared against pc-d11 and pc-c6a3. When compared against pc-c7a3, a similar trend was observed except for b19, whose sfc290p value (8.65) is slightly lower than that of pc-c7a3 (9.01). In general, the high values of DrugscoreX alongwith the pharmacophore filtering ensure a correct placement of the new ligands, while the rescoring with SFC predicted a rise of 1 pK$_i$ unit for most of the compounds.

In general, for all compounds, various ring replacement and substitutions perform well resulting in modest improvements in averaged apparent InhA inhibitory potential (pK$_i$). The averaged SFC290p difference for b1 to b11 was 0.36 pK$_i$ units as compared to pc-c7a3 including the scores of compounds with inverted binding modes. Upon swapping the inverted poses with **”in-situ”** mutated poses, a marginal increase in the averaged difference in SFC290p was seen (w.r.t. c7a3: 0.50 pK$_i$ units). The averaged difference in SFC290p values rose to 1.30 (including inverted binding modes) and 1.20 (replacing inverted binding modes with *”in-situ”* mutated poses) pK$_i$ units when compared to pc-d11. A similar trend was observed in case of the compounds b12 to b20, with an averaged difference (increase) of 1.03 pK$_i$ units as compared to pc-c7a3 and an increase by 1.95 pK$_i$ units when compared with pc-d11.

Furthermore, it can be seen that the various ring replacement and substitutions perform well (Figure 9.8) as seen from docking as well as rescoring analysis performed with DrugscoreX and SFCscore. In regards of scaffold placement, four molecules (b14, b16, b19, and b20) exhibit substructure RMSD in range of 2 to 3 Å. This indicates the problem of proper scaffold placement even when protein flexibility was taken into account. Moreover, b19 and b20 have long alkyl chains replacing the ring A3 (cf. Figure 9.5) and get docked in PDB 2X23. This was surprising, given the huge size of the ligands.

Nevertheless, the high substructure RMSD can be attributed to the bulky size of the ligands as well as the tight binding pocket of PDB 2X23. For the compound b14 and b16, a similar argument holds true.

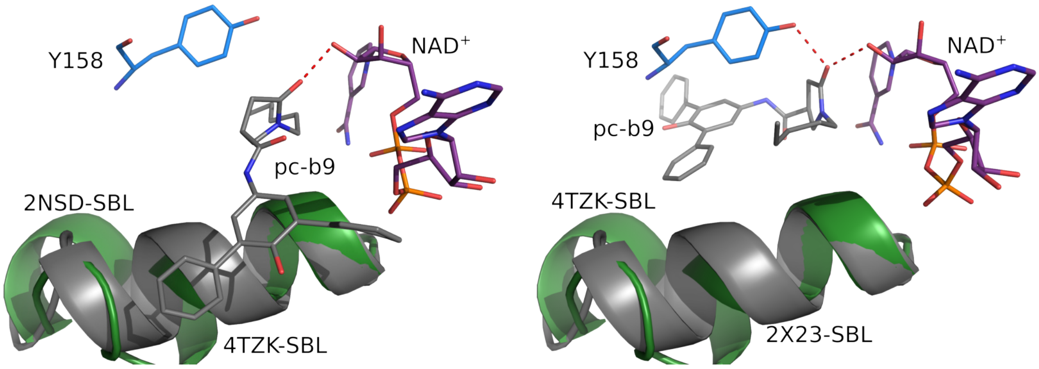


Figure 9.8 The picture on the left depicts the "in-situ" mutated pose of pc-b9 obtained from pc-c6a3 (inverted binding mode). On the contrary, the one on the right depicts the docked pose of pc-b9 with a correct placement of the ligand.

9.3.2 Mycobacterial cell wall permeability of new pyrrolidine carboxamides

The designed pyrrolidine carboxamides were assessed for their ability to permeate the mycobacterial cell wall, with the results being depicted in Table 9.3. The results clearly indicate that all of the designed molecules exhibit satisfactory mycobacterial cell wall permeability when compared against pc-d11 or their parent compounds (pc-c6a3 and pc-c7a3). Consequently, all of the designed molecules must be able to permeate through the mycobacterial cell wall in in-vitro activity assays.

9.3.3 Activity classification of new pyrrolidine carboxamides

As an additional *"in-silico"* assessment of the new pyrrolidine carboxamides, the activity classification of the new molecules was carried out using the XPscore-SFC290p based logistic regression model depicted in Figure 4.7 and table 4.5. The results were in line with the expectations that most of the compounds will be deemed as highly active (Table 9.3). This is partly because almost all of the molecules exhibited noticeable improvements over their starting compounds in terms of docking as well as rescoring. The underlying model that enables the activity-based classification of pyrrolidine carboxamides owes majority of its predictive power to the SFC290p scoring function. Since, all of the new pyrrolidine carboxamides demonstrated moderate to high SFC290p values, the outcome of this exercise was in line with the expectations.

However, as seen from Table 9.3, the compound pyrrolidine carboxamide-b19 has been classified as least active. The primary reason behind this is the divergent binding of the ligand, which stems from its bulky size that causes problems for proper scaffold placement during docking. This results in a strict penalty by the Glide XP scoring function that assigns a low XPscore to the docked pose and thereby the molecule is predicted as less active. Upon consideration of the "in-place" rescoring values for b19, a contrasting result is obtained. Moreover, a similar behaviour is also observed for pyrrolidine carboxamide-b18, but it is not predicted as least active primarily because of its high SFC290p score. In summary, the *in-silico* assessments performed so far demonstrated the beneficial effects of the substitutions on the core pyrrolidine carboxamide scaffold.

Table 9.2 Docking and rescoring values for docked poses of the designed pyrrolidine carboxamides-b12 to b20. The substructure RMSD was with respect to pc-d11, the ligand in PDB 4TZK.

Parameter	**d11**	**c6a3**	**c7a3**	**b12**	**b13**	**b14**	**b15**	**b16**	**b17**	**b18**	**b19**	**b20**
XPscore[6]	-10.09	-5.17	-12.87	-14.23	-13.37	-11.95	-11.73	-10.65	-15.26	-8.84	-6.15	-12.80
RMSD[7]	0.37	6.46	1.81	0.79	1.74	2.18	1.13	3.00	1.58	1.51	2.20	2.90
DrugscoreX	-120.79	-120.99	-181.86	-224.20	-207.67	-204.57	-192.80	-218.98	-214.94	-213.20	-208.74	-211.67
sfc229p	7.82	8.25	9.38	10.37	10.36	10.85	9.64	9.77	10.97	9.87	9.09	9.76
sfc290p	8.09	8.04	9.01	9.98	10.28	10.82	9.67	9.94	10.73	9.89	8.65	10.37
sfc_rf	7.50	7.73	7.82	9.42	8.97	8.88	8.78	8.47	8.84	8.80	8.72	8.49

9.3.4 Dihedral angle and H-bond analysis of new pyrrolidine carboxamides

The previous sections dealt with static *in-silico* assessment of the binding affinity, activity and mycobacterial permeability of the new compounds. In order to shed more light on the binding of these new compounds to InhA, the binding dynamics of the new compounds has to be studied and understood in detail. Likewise in Chapter 7, the dihedral angle and hydrogen bond analysis was performed on the 5 ns MD simulations of the ligands in bound state. The entire dataset of 20 compounds were simulated under NPT conditions, with conformational changes being stored in the trajectory at every pico second. Thus, these analyses enable one to extensively follow the conformational changes and thereby shed light on the binding dynamics of the new compounds. The ensuing sections highlight the nature of binding for the designed compounds, while the last section will briefly discuss the binding modes of some promising compounds from the designed molecules.

[6] units: kcal/mol

[7] Substructure RMSD, units Å

Table 9.3 Mycobacterial cell wall permeability and activity classification for the reference compound (pc-d11), parent compounds (pc-c6a3, pc-c7a3), and the designed pyrrolidine carboxamides, respectively. The ligands marked in boldface are the ones with inverted binding modes. For these compounds, the "in-place" XPscore and SFC290p have been used.

Compound	clogP	Permeability	XPscore	SFC290p	Activity
d11	4.98	0.35	-10.09*	8.09	Highly active
c6a3	5.72	0.74	-9.80	8.04	Highly active
c7a3	6.28	0.74	-14.48	9.01	Highly active
b1	5.78	0.90	-14.14	9.09	Highly active
b2	5.93	0.90	-14.34	9.26	Highly active
b3	6.50	0.90	-14.28	9.50	Highly active
b4	4.02	0.78	**-13.74**	**9.17**	Highly active
b5	6.76	0.77	**-12.78**	**8.64**	Highly active
b6	5.39	0.85	-13.89	10.08	Highly active
b7	6.80	0.91	-13.66	10.25	Highly active
b8	6.67	0.91	**-10.38**	**8.62**	Highly active
b9	6.60	0.74	-9.43	8.20	Highly active
b10	6.65	0.91	-10.37	8.42	Highly active
b11	6.28	0.88	**-10.33**	**8.36**	Highly active
b12	4.84	0.73	-14.23	9.98	Highly active
b13	5.12	0.85	-13.37	10.28	Highly active
b14	5.85	0.81	-11.95	10.82	Highly active
b15	5.19	0.71	-11.73	9.67	Highly active
b16	7.45	0.48	-10.65	9.94	Highly active
b17	5.89	0.85	-15.26	10.73	Highly active
b18	4.50	0.69	-8.84	9.89	Highly active
b19	6.10	0.75	-6.15	8.65	Least active
b20	5.82	0.77	-12.80	10.37	Highly active

* units for docking in kcal/mol

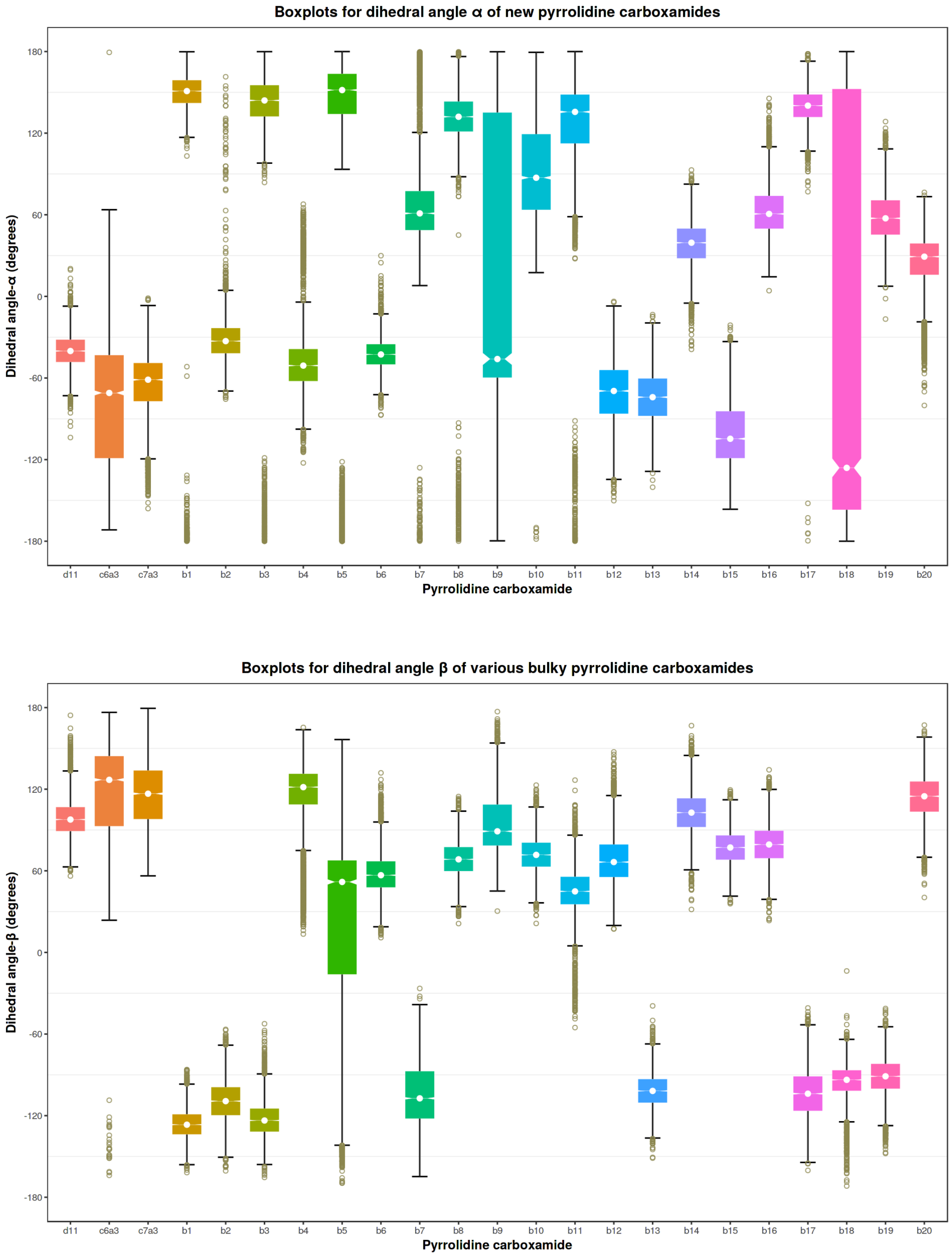
The docking scores (XPscore and SFC290p) in boldface indicate they have been taken from in-situ mutated poses

The clogP for all compounds was calculated using MoKa-2.6.0 [327].

The input conformations for the MD simulations were derived primarily from docking with induced fit. The dihedral angle and H-bond analysis, much like Chapter 7 provide a means for testing the hypotheses pertaining the artefactual nature of the inverted binding mode.

9.3.4.1 Dihedral angle analysis

Figure 9.9 depicts the dihedral angle distribution for the new compounds, with the reference compound, i.e. pc-d11, occupying the pole position followed by the parent compounds (pc-c6a3, and pc-c7a3) in order to facilitate comparison of the distributions. Similarly, Figure 9.12 depicts the RMSD fluctuations of the designed compounds, with the order being identical to that in Figure 9.9.



Boxplots for dihedral angle α of new pyrrolidine carboxamides

Boxplots for dihedral angle β of various bulky pyrrolidine carboxamides

Figure 9.9 Boxplots depicting the **(a)** dihedral angle-α (top), and the **(b)** dihedral angle-β (bottom) for new pyrrolidine carboxamides, with the reference ligand d11 at the pole position. The median for the distribution of the dihedral angles has been depicted as a white circle

From Figure 9.9 , the following observations were made:

1. In terms of dihedral angle α and thereby the ring A system, the compounds pc-c6a3, pc-b9, and b18 show a noticeable and simultaneously wide variation in the distribution implying that there is low barrier for rotation of ring A1 around the C-N bond linking the B ring to the A ring system. In case of pc-c6a3, the inverted

binding mode with the bulky portion facing the exterior region of the protein makes possible the free rotation and thereby the wide variation in dihedral α. In case of pc-b9, the wide variation in the dihedral angle arises due to the rotation of the 2° carbonyl group at around 1.7 ns (Figure 9.10) that leads to a huge change in the dihedral α. However, after this the binding mode is fairly stable with no further changes as observed from the 5ns MD simulation. However, pc-b18 does not show a complete flip of the 2° carbonyl group. Nevertheless, the wide variation in the dihedral α for b18 arises from in-place motions of the central A ring and 2° carbonyl group, respectively. The in-place motion of the A ring system is further enhanced due to the rupture of the hydrogen bond of the central ring with Pro156. A similar observation holds true for the compounds pc-b2 and pc-b7, with a m-CF_3 substitution on ring A2.

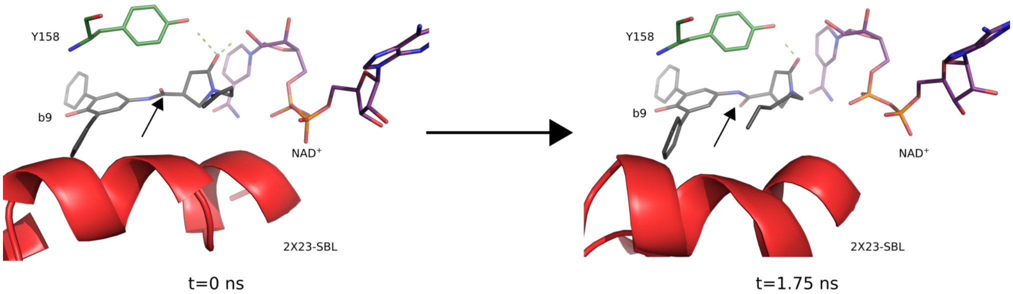


Figure 9.10 Flip of 2° carbonyl group after 1.7 ns: The arrows denote the flipped 2° carbonyl group. The substrate binding loop is coloured red, while the hydrogen bonds are shown as lime coloured dashes.

2. There is a wide variation in the distributions of the dihedral angle α, though a majority of the ligands get docked in a conformation like pc-d11. For example, the compounds b4, b5, b8, and b11 exhibit a wide distribution as well as increased number of outliers than rest of the designed compounds. The slightly wider variation in the dihedral α for these compounds is in line with the observation made for the parent compound pc-c6a3, whose A ring system faces the exterior and moves freely. The comparatively narrow distribution of dihedral angle α for the remaining molecules stems from their proper placement within the binding site, where interactions with the active site residues stabilise the motions of the A ring system. This merely underscores the conclusions from Chapter 7, that the crystal structure ligand conformation is the dominant binding conformation for bigger pyrrolidine carboxamides.

3. The dihedral angle β, which represents the ring C rotations also exhibited a wide variations as far as the compounds b1 to b11 were concerned. Likewise in the case of dihedral angle α, the compounds getting docked in an inverted binding mode (b4, b5, b8, and b11) exhibited a wider variation in the distribution as well as more number of outliers as compared to rest of the compounds. The compound

b5 is clearly an exceptional example with its inverted binding mode that signifies improper placement of the ligand in the binding pocket. The planar phenyl ring (ring C) points into the space normally occupied by the bulky A ring system and thereby moves freely. This gives rise to the wide variation in β for the respective compound. Similar observations hold true for the other compounds (b4, b8, and b11) that get docked in an inverted fashion.

4. The compounds b12 to b20 present an intriguing case as far as the motions of the ring C are concerned. All of these compounds show a comparative narrow distribution as compared to all other molecules, except the reference molecule (i.e., pc-d11) and pc-b1. The (un)-/substituted phenyl ring for the aforementioned compounds was found to interact with the cofactor and key residues (Met199, Ile202, and Val203) that the line the substrate binding loop. Assuming that the MD simulations sample almost real physical states of the protein, the geometrical arrangements around the C ring are expected to give rise to C-H-π, lone pair-π and sulphur-arene ring interactions [349] (Figure 9.11). While the force field does not specifically describe any of these interactions, somehow these might be at least implicitly accounted for by the van der Waals terms. Furthermore, the drastic changes in the median values of dihedral β, for example, the median value for the compounds pc-b1, b2, b3, b4, b13, b17, and b19 is around -120°, while that of rest of the system varies in between 60° to 120°. These changes can be attributed to the pseudo-symmetry of the C ring that results in the respective dihedral angle distribution. Finally, in the case of compound pc-b18, with a n-heptyl ester group as the C ring, the stabilisation of the flexible alkyl chain can be attributed exclusively to the van der Waals interactions with the residues that line the substrate binding loop.

5. Inspection of the bound ligand RMSD values for the designed compounds (Figure 9.12) revealed that the majority of them exhibited RMSD in the range 0.8 - 2.2 Å. The visible outliers are pc-b4, pc-b8, pc-b14 and pc-b16, respectively. The high RMSD values for pc-b4 and pc-b8 can be attributed to their inverted binding modes. The compound pc-b16 represents a notable exception, since it gets docked in a crystal structure ligand like conformation. The compound b16 is quite hydrophobic and consequently, van der Waals interactions play a greater role in stabilising its binding mode as compared to hydrogen bonds. The compound itself forms weak and transient bonds (cf. Table 9.4) that offer a very low barrier to the motions of the ligand within the binding pocket. As a result, upon rupture of the solitary hydrogen bond with Tyr158, this ligand moves within the binding pocket resulting in a higher median RMSD value. Furthermore, the higher than normal median RMSD values for the majority of the designed ligands can be attributed to their bulky nature and the rearrangements within the binding pocket through

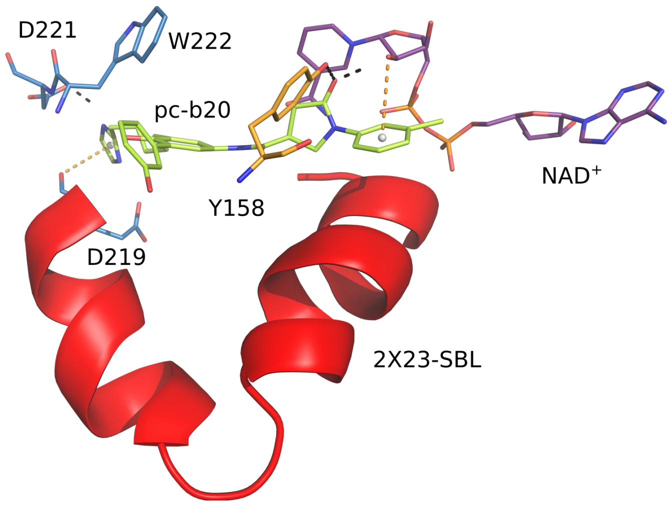
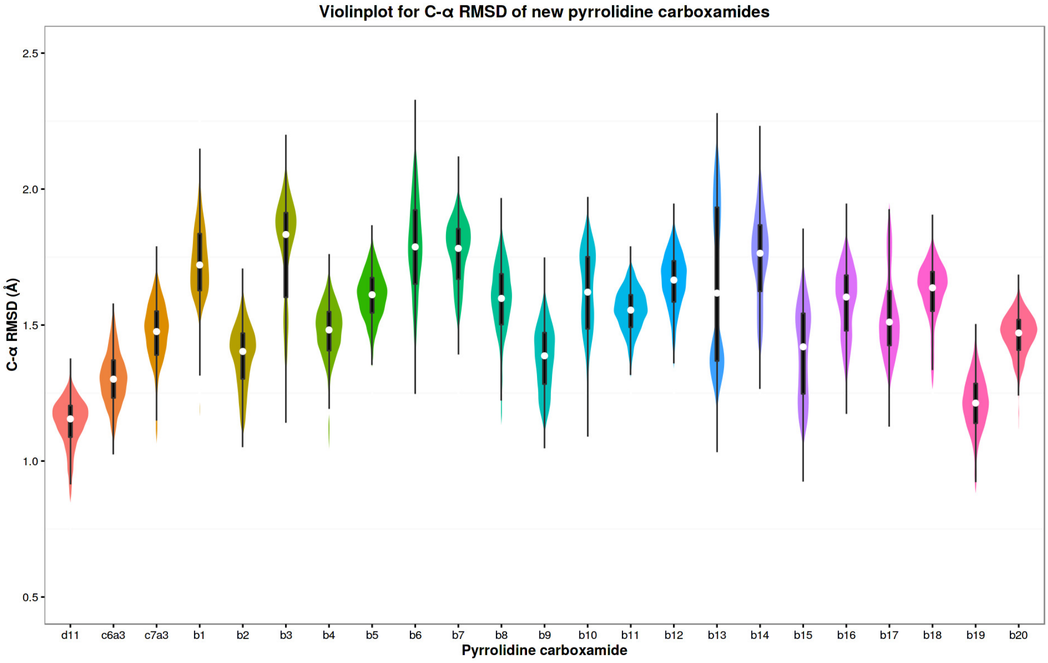


Figure 9.11 Various non-bonded interactions for a representative designed pyrrolidine carboxamide-b20: The hydrogen bonds appear as black dashes. The lone pair-π interactions are depicted as orange dashes. The ligand has been depicted as lime coloured sticks, cofactor as purple coloured sticks and the substrate binding loop as a red coloured α helix.

the MD simulations. In the case of pc-b14, with an methylsulfanylbutyl group as the A3 ring, the higher RMSD values can be attributed to two factors, namely the difficulty faced by docking in placing the ligand correctly in the binding pocket; and the aforementioned group that is quite flexible which gives rise to the binding instability in the said compound.

6. Finally, the C-α RMSD distributions of the new ligands depicted an expected observation, where the majority of the median RMSD values were above 1.5 Å. The comparatively higher median C-α RMSD values for the designed compounds is hardly surprising, given the bulky nature of the ligands and the flexible nature of InhA. The compound b13 exhibits a quite narrow and unique RMSD distribution alongwith b15. The small dips in the smooth shape of their RMSD distributions signify noticeable fluctuations in the motions of the residues in the immediate vicinity of the ligand. This can be observed in Figure 9.13, where the dihedral angles indicate an ”in-place” motion of the ligand that is accompanied alongwith a systematic rise in the C-α RMSD. A similar observation holds true for b15 as well, that suggest initial adjustments within the binding pocket followed by stabilisation of the binding mode.

From the above observations, it can be clearly seen that optimisation of the pyrrolidine carboxamide scaffold worked in the right direction, with majority of the ligands exhibiting stable binding, although most of them exhibited both higher than normal median RMSD values. Nevertheless, the information from the dihedral angle analysis clearly depicts remarkable stability for about half of the designed compounds. A sizeable number of the remaining compounds exhibited inverted binding modes that have been associated with



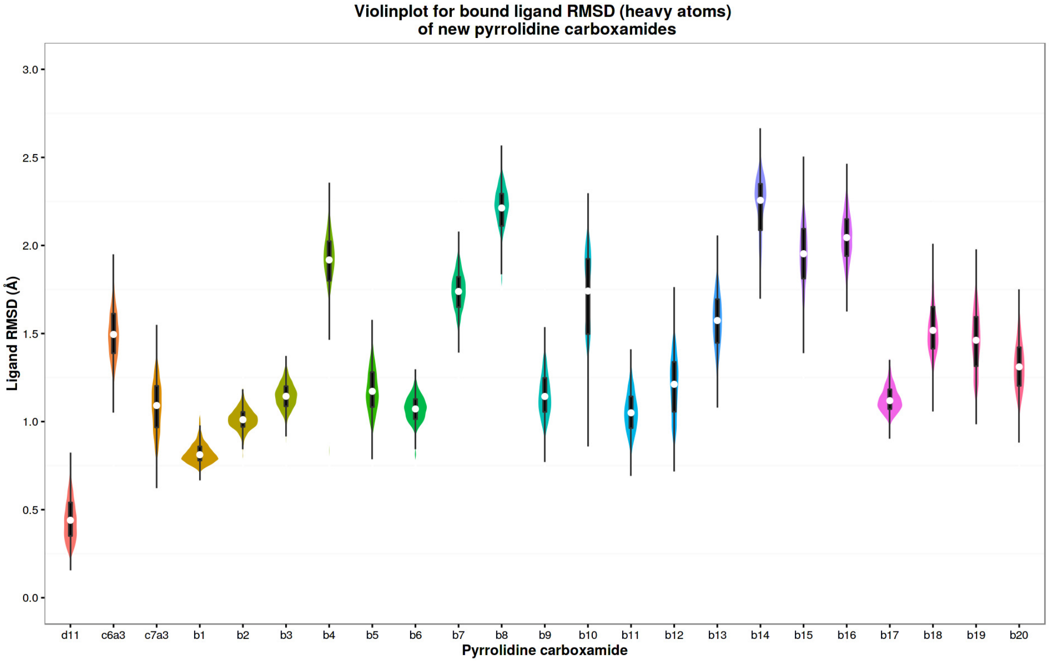


Figure 9.12 Violinplot depicting the kernel density distribution of C-α RMSD and bound ligand RMSD (heavy atoms only) for new pyrrolidine carboxamides, with the order of compounds being described in Figure 9.9. The thick central bar depicts the interquartile range, white circle denotes the median for the distribution. Furthermore, a dip in the smooth shape of the violin indicates a steep change in the RMSD distribution

increased binding instabilities and weak H-bonding. Considering the fact that stable binding was associated principally with the crystal structure ligand conformation, the artefactual nature of the inverted binding modes is lent further support.

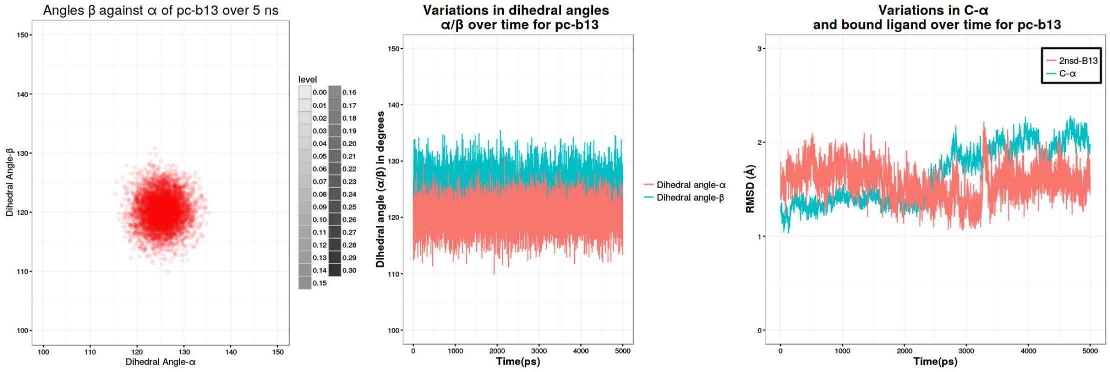


Figure 9.13 Dihedral angle distribution and RMSD fluctuations for pc-b13 over 5ns NPT simulation

9.3.4.2 H-bond analysis

The quality of the binding as visible in the H-bond occupancy can be seen in Table 9.4, with the ligands marked in bold showing the alternate H-bonding conformation as seen in Chapter 7 for pc-d8 (4TZT ligand; Figure 7.27). As opposed to pc-d8, the alternate conformation of the designed compounds is not transient, but a stable one, as can be seen in case of pc-b12 (Figure 9.15). The transition from the crystal structure ligand like conformation to the alternate H-bonding conformation takes place early in the simulation (around 200 ps) and stays as is till the end of the sampling period, i.e., 5 ns. The presence of an alternate H-bonding conformation is further lent support from the donor-acceptor atom distance plots (Figure 9.14) that clearly suggest that in case of some compounds, the alternate H-bonding conformation is indeed sampled during the MD simulations. Table 9.4 also helps to reinforce the following observations already made in Chapter 7:

1. The new compounds also exhibit weak to moderate hydrogen bonding with Tyr158 and NAD$^+$.

2. The compounds exhibiting inverted binding modes (b4, b5, b8, and b11) show weak and transient H-bonding as compared to the other designed compounds which bind to InhA like the reference ligand (pc-d11). This can be seen from Table 9.4 in form of very poor to poor H-bond occupancies with either Tyr158 and NAD$^+$. This merely lends support to the fact that the inverted binding mode results in non-optimal h-bonding. It is the lack of proper H-bonding that manifests in noticeable fluctuations in dihedral angles and RMSD values during the MD simulations.

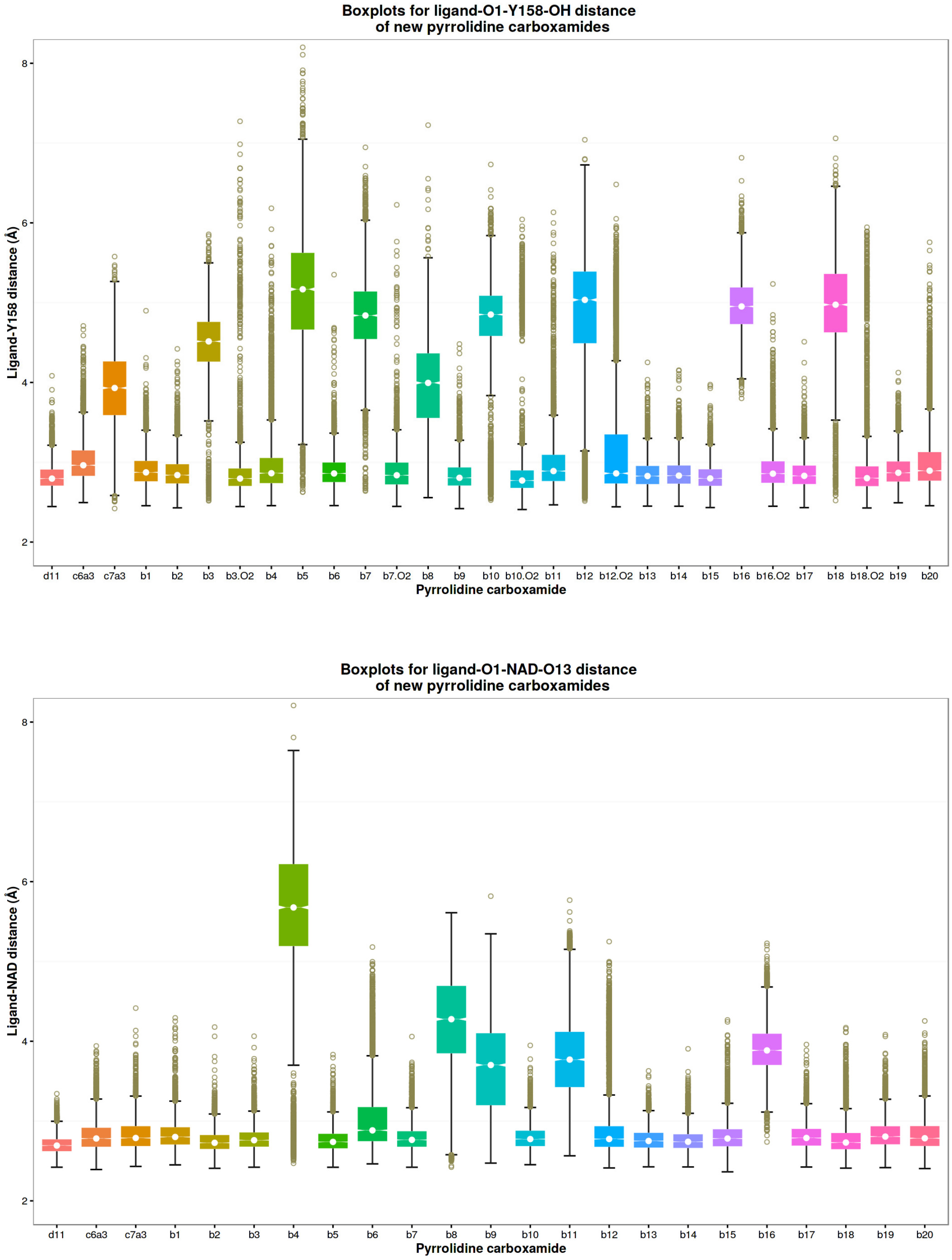
3. In case of the ligand-Tyr158 distance, a wide variation in the distances can be seen. The visual outliers are the compounds b4, b5, b8, b10, b11, b12, b16, and b18. Of these, the ligands with inverted binding mode, i.e., b4, b5, b8, and b11 exhibit a wide variation in the donor-acceptor atom distance and a copious number of

Table 9.4 The H-bond occupancies highlighting the H-bonding quality for ligand-Y158 and ligand-cofactor H-bond. The molecules marked in bold exhibit the alternate H-bonding conformation.

Compound	**H-bond occupancy (in % over 5 ns)**	
	Ligand-Y158	**Ligand-NAD**
b1	36.26	55.64
b2	48.16	61.38
b3	**42.96**	45.52
b4	21.62	2.78
b5	0.14	43.32
b6	48.66	38.72
b7	32.60	34.72
b8	3.50	3.74
b9	59.26	8.52
b10	**59.30**	40.50
b11	21.80	1.50
b12	**45.88**	21.60
b13	47.10	52.52
b14	45.28	35.96
b15	56.70	53.70
b16	**49.72**	0.00
b17	46.92	54.88
b18	**50.68**	36.86
b19	43.82	52.48
b20	37.22	46.66

outliers. This also supports the destabilised binding of the respective compounds. For the rest of the compounds, i.e., b12, b16, and b18, all exhibit the alternate H-bonding conformation that explains the wide variation in distributions of their donor-acceptor distances.

4. Furthermore, in case of the ligand-NAD$^+$ distance, the compounds b4, b8, b9, b11, and b16 can be classified as outliers by visual inspection. Of these, b4, b8, and b11 exhibit the inverted binding mode, which again points to their artefactual nature. In case of b16, the transition from the crystal structure like conformation to the alternate H-bonding one can be attributed to a higher median value of the ligand-NAD$^+$ distance. The ligand b9 presents an interesting case, since it binds like the reference ligand and moreover gets docked in PDB 2X23. A combination of the aforementioned factors should lead it to exhibit stable binding, which it does. However, the disruption in the ligand-NAD$^+$ bond can be attributed to the free motion of the n-alkyl chain that replaces the cyclohexyl ring from bulky pyrrolidine carboxamides.



Boxplots for ligand-O1-Y158-OH distance of new pyrrolidine carboxamides

Boxplots for ligand-O1-NAD-O13 distance of new pyrrolidine carboxamides

Figure 9.14 Boxplots depicting the distribution of donor-acceptor atom distances for new pyrrolidine carboxamides-NAD$^+$-Y158, with the ligands being sorted a way similar like Figure 9.9. Furthermore, the outliers are bronze circles while the median is depicted as a white circle. The whiskers depict the interquartile range for the ligand specific distance distribution

9.4 Discussion

The present chapter described in detail the approaches for thorough *in-silico* evaluation of the new pyrrolidine carboxamides whose design was inspired by the findings from the essential dynamics (Chapter 8), dihedral angle analysis (Chapter 7), and SAR

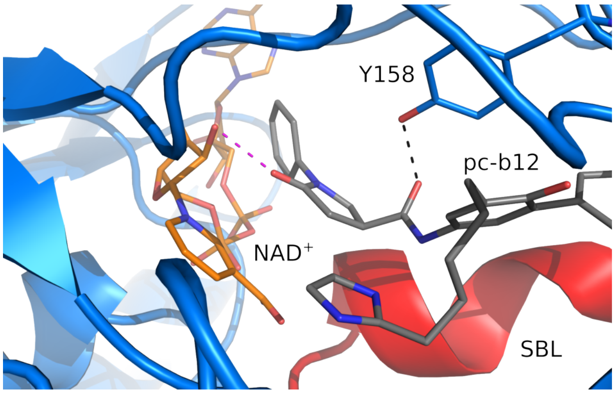


Figure 9.15 Alternate H-binding conformation of pyrrolidine carboxamide-b12 (grey sticks); SBL is shown as red helix, Y158 as marine sticks, cofactor as orange sticks and H-bonds as dashed lines (black dashed line for secondary carbonyl-Y158 bond; magenta dashed line for primary carbonyl NAD bond).

of InhA inhibitors derived from literature [50]. The main aim behind the design of these compounds was optimisation of the binding by means of increased non-bonded interactions with the protein, particularly increasing the number of H-bonds that the ligand forms with the protein. This was in addition to the dual H-bonds that are expected to stabilise the binding to a large extent. Accordingly, a series of 11 molecules were initially designed followed by assessment of their binding modes and affinity by molecular docking and the LIE method. This was closely followed by another round of structure-based optimisation, wherein the beneficial substitutions from the first round were applied on the core pyrrolidine carboxamide scaffold taking the total number of designed molecules to 20.

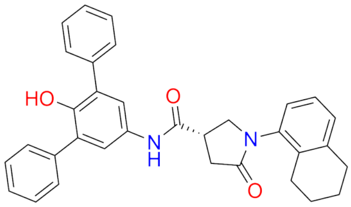
The induced fit docking was able to satisfactorily predict the binding modes for a majority of the compounds, whilst inverted binding modes were reported for 20% of the total compounds. As a generalised observation, compounds with inverted binding mode scored poorly as compared to rest of the dataset. The difference in between the docking score for the better scoring ligands and their parent compounds (pc-c6a3 and pc-c7a3) as well as the reference compound (pc-d11) ranged from 0.5 kcal/mol to as high as 6 kcal/mol.

On the contrary, the inverted binding modes clearly lagged behind the parent as well as the reference compound. The rescoring exercise that followed the induced fit docking clearly showed an improvement in the average predicted affinity by 0.5-1.3 pK$_i$ units when compared against pc-c7a3 and 1.2-1.95 pK$_i$ units when compared against the reference compound (pc-d11). Likewise, during docking, the compounds with inverted binding mode got lower scores than their counterparts getting docked like the reference ligand (pc-d11).

The dihedral as well as H-bond analysis suggested that the inverted binding modes were possibly artefactual. For a sizeable number of ligands, the input conformation for docking was inverted whilst the docked pose resembled the reference ligand orientation (e.g. pc-b9). The same analysis also highlighted the remarkable stability of the new pyrrolidine carboxamides barring a few exceptions (pc-b7, pc-b16), which were purposefully designed to serve as a sort of litmus test for favourable/unfavourable pattern of substitution. The H-bond analysis also revealed that 20% of the designed molecules (Table 9.4) exhibited alternate hydrogen bonding conformation which were stable and demonstrated their own characteristic distribution of dihedral angles and donor-acceptor atom distances. The complementary H-bond and distance analysis provided support to the aforesaid observations, while suggesting the artefactual nature of inverted binding mode. The designed compounds, much like their parent pyrrolidine carboxamides, exhibited moderate H-bonding.

The additional *in-silico* tests that evaluated the mycobacterial permeability as well as predicted the activity class for the new compounds yielded expected results. Across all of these tests, almost all of the molecules exhibited noticeable improvement over the reference ligand and their parent compounds. The beneficial effects of the substitutions can be seen in case of the new pyrrolidine carboxamides with polar substituents (pc-b4, b6, and b18), all of which exhibited high mycobacterial permeabilities (Table 9.3). A similar trend was observed in case of activity prediction using the XPscore-SFC290p based logistic regression model, wherein all except one molecule (pc-b19) were deemed highly active. This prediction can be attributed to the low docking score assigned to the compound during docking, given the reliance of the logistic regression model on both XPscore and SFC290p. Nevertheless, the supplementary evaluations clearly pointed out the favourable nature of the new compounds [52].

Moreover, the novelty of the designed molecules was confirmed, when a substructure query on Scifinder® using the core pyrrolidine carboxamide scaffold returned only a solitary hit (Figure 9.16). The solitary molecule was synthesised and tested by He. et al., who are the original authors behind the published pyrrolidine carboxamides. Additionally, the synthesis scheme for a representative compound has been enshrined in Appendix C.



(*S*)-*N*-(2'-hydroxy-[1,1':3',1''-terphenyl]-5'-yl)-5-oxo-1-(5,6,7,8-tetrahydronaphthalen-1-yl)pyrrolidine-3-carboxamide

Figure 9.16 Solitary compound returned from Scifinder® using the core pyrrolidine carboxamide scaffold as a query.

A interesting finding from the MD simulations was the stabilisation of the ligand binding, which was previously difficult to achieve with the bulky pyrrolidine carboxamides, although the bound ligand RMSD for all of the ligands was much higher than their parent compound. The stable binding of the new pyrrolidine carboxamides alongwith the role of the dual H-bonds in binding mode stabilisation is explained via extensive MD simulations (150 ns in bound state) of representative compounds (b6, b9, and b12). All three compounds get docked like pc-d11 and exhibit remarkable stability. This is evident in negligible change in the binding modes for b6 and b9, respectively. In case of b12, a change in binding mode much similar to pc-p36 and pc-d8 (Figures 7.27 and 7.30) was seen. However, in contrast to the aforementioned compounds, the binding mode of b12 changes from crystal structure ligand like to the one depicted in Figure 9.15 and stays put through the 150 ns MD simulation. This implies the stability of the alternate H-bonding conformation of b12 as compared to metastable in case of pc-p36 and pc-d8.

Moreover, it was quite important to choose the best performing ligands from amongst the designed ones for further synthesis and testing. Since, all of the compounds represented modest improvement over pc-d11, pc-c6a3, and pc-c7a3, appropriate compounds for further synthesis and testing were selected on the basis of the following criteria:

1. The difference in between the SFC290p score of the ligand and pc-d11 as well as pc-c7a3. For the former case, a difference of ≥ 1 pK$_i$ unit and in the latter case $\geq$ 0.5 pK$_i$ was used as a filter.
2. Thereafter, the bound ligand RMSD values of the ligands were considered. For this purpose a upper limit of 1.30 Å was set, following which the ligands with low logP values were considered .
3. It is a known fact that compounds with higher logP values tend to cause significant formulation and testing issues during *in-vitro* and *in-vivo* phases. However, this rule has several exceptions [350]. Considering this, an upper limit of logP 5 ± 0.70 units was set as a filter. Furthermore, the actual logP values for the designed molecules are not available and hence the respective logP values (abbreviated as **clogP**) were calculated using MoKa-2.6.0 [327].

Applying the aforesaid filters on the designed compounds revealed the following compounds worth investigating further: pc-b4, pc-b6, pc-b12, and pc-b18.

9.5 Conclusion

In conclusion, the following points can be safely considered as worth mentioning:

1. The structure-based optimisation of pyrrolidine carboxamides performed using information from essential dynamics and literature was deemed effective as relevant from the various *in-silico* assessments as well as rigorous MD simulations.

2. The dihedral angle as well as H-bond analysis revealed the possibility of the inverted binding modes being artefactual. The support for the same comes from very weak hydrogen bonds with either Tyr158 or cofactor as well as high values for C-α and bound ligand RMSD. Verification of the inverted binding modes remains to be ascertained given the lack of crystal structures for bulky pyrrolidine carboxamides.

3. The 150 ns MD simulations for select compounds pc-b6, b9, and b12, shed light on the the existence of alternate H-bonding conformations that are quite stable. The said mode was observed for 20% of the designed ligands.

4. Inspite of higher median values for both C-α and bound ligand RMSD, a majority of the ligands exhibited remarkable binding stability.

5. On an overall basis, all of the compounds, barring b7 and b16, appear promising. After taking into consideration the *in-silico* evaluations and the extensive MD simulations, the compounds pc-b4, pc-b6, and pc-b12 can be put forward as promising candidates for further evaluation and testing.

Chapter 10

”Rapid reversible binding” of pyrrolidine carboxamides: Revealing molecular determinants by MD simulations

10.1 Introduction

The chapters 8 and 9 exemplify the utility of MD simulations and essential dynamics in driving the structure-based optimisation of pyrrolidine carboxamides. The Chapter 9 also reveals the dominant binding conformations of the new pyrrolidine carboxamides along with promising candidates for further *in-vitro* evaluation. The important residues involved in the binding of light pyrrolidine carboxamides have already been known. However, this information is lacking in the case of bulky pyrrolidine carboxamides as well as the new ligands due to a lack of crystal structures. Revealing the key residues involved in their binding along with their conformations would aid in a better understanding of the binding of pyrrolidine carboxamides to InhA. The corresponding information in case of the slow tight binding inhibitors of InhA (diphenyl ethers) is well known [67, 261]. A brief comparison of the conformational changes of key residues for these two classes of InhA inhibitors can qualitatively reveal the causative factors behind the nature of their binding to InhA.

The current work focusses on elucidating the detailed molecular determinants driving the apparent ”*rapid reversible*” binding of pyrrolidine carboxamides to InhA. Loop ordering and the associated conformational changes of the SBL and active site upon ligand binding are decisive factors in the context of both slow-tight and ”rapid reversible” binding. An effective ligand is able to bring about an ordering (and closure) of the SBL which is closely associated with a two-step association (cf. slow-tight binders) [74]. As seen from Li et al. [74], the open and closed states of the SBL correspond to the EI and EI* states, respectively (cf. Figure 10.1). The EI state can be observed in the case of **PT155** (2-pyridone, rapid reversible inhibitor; PDB 4OXK/4OXN), the substrate analogue (**C16-NAC**, PDB 1BVR) and **pyrrolidine carboxamides** (Figure 10.2). The EI* state, characterised by an ordered (and closed) SBL, has been observed for the slow-tight binding diphenyl ethers **PT70** and **PT92**. These observations suggest that pyrrolidine carboxamides might exhibit rapid reversible binding and thereby lacking the ability to bring about an closure of the SBL.

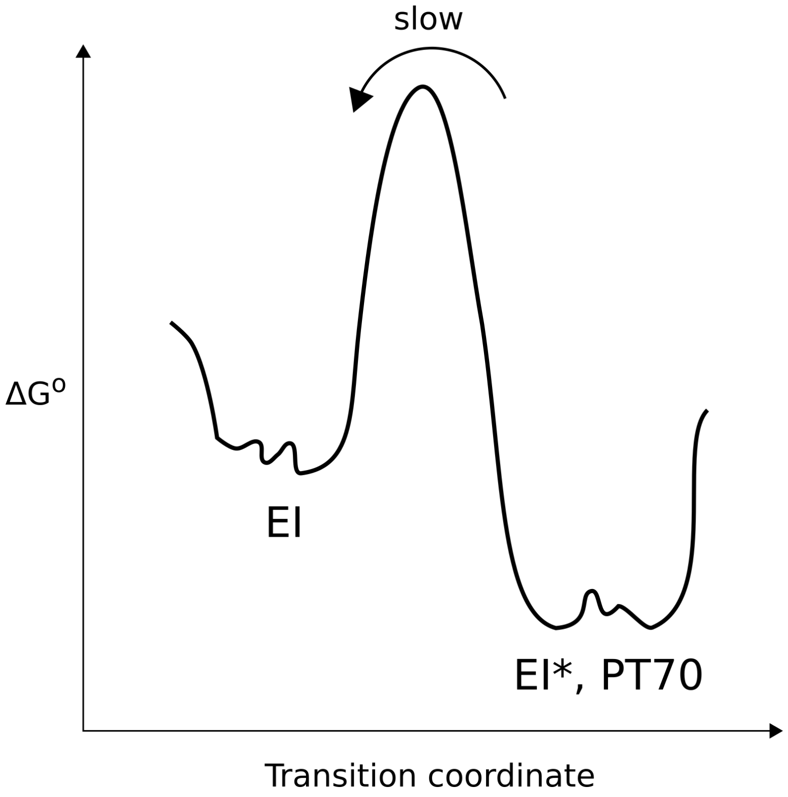


Figure 10.1 Representative free energy profile for a slow-tight binding InhA inhibitor (PT70). The macrostates EI and EI* comprise several microstates (conformations). Figure adapted from [261].

Considering the aforementioned research results pertaining InhA, this chapter is based on the following assumptions: 1) The ternary structure of InhA-NAD$^+$-PT70 represents the EI* state with a closed SBL conformation. The discussions in the recent literature strongly support this assumption [67, 74]. 2) The open conformation of the SBL as observed in the PDB structures 4OXK/4OXN, 1BVR, and 1P44 (genzyme series) and 4TZK (pyrrolidine carboxamides) corresponds to the EI state. 3) Because of the open conformation of the SBL in case of 4TZK, all pyrrolidine carboxamides can be considered as *rapid-reversible inhibitors*. However, the conformations of the SBL in the 4OXK/4OXN, 1BVR, 1P44, and 4TZK vary slightly (Figure 10.2). The SBL states for PDB 4OXK/4OXN and 4TZK can be considered as one of the several microstates that populate the EI macrostate. The conformational changes between EI and EI* happen on a time scale that is hardly accessible by classical unbiased MD simulations. However, the conformational changes within the macrostate EI (represented by PDB 2NSD, 4TZK, and 1BVR etc.) and the EI* state (represented by PDB 2X23) can easily be analysed via long MD simulations. Valuable mechanistic insights into the conformational dynamics of the aforementioned macrostates can be obtained from the analysis of these long MD simulations. In line with these assumptions, conformational clustering techniques have been extensively used in the current chapter to unveil the similarities and differences amongst the closed and open conformations of the SBL which are visible in PDB 2X23 and the PDB structures 2NSD and 4TZK, respectively.

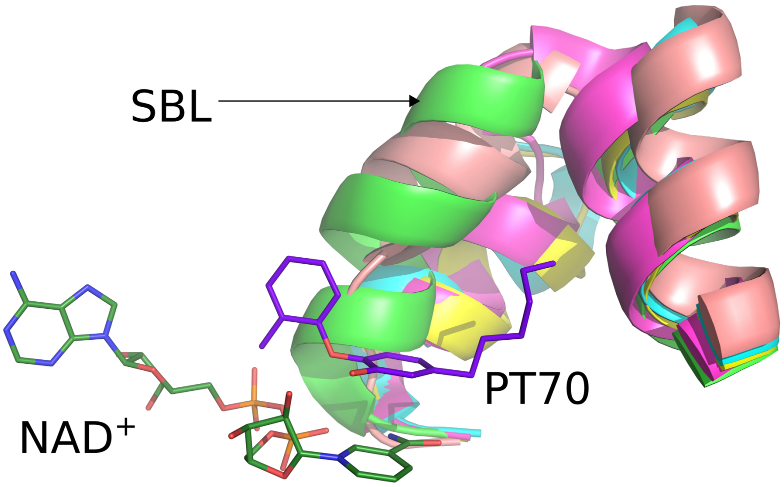


Figure 10.2 Various conformations of the substrate binding loop starting with the closed state (EI*, 2X23; green) as opposed to all the remaining structures that show a progressive opening of the SBL signifying a destabilised state. The other conformations of the SBL are in the following order: salmon (4OXK; PT155), magenta (4TZK, pc-d11), yellow (1BVR, C16-NAC (C16-thioester)), and cyan (1P44, GEQ). The bound ligand of PDB 2X23 has been depicted as violet sticks. All structures correspond to chain A of the respective PDB files.

10.2 Materials and Methods

This section elaborates on the atom selections defining the active site and the SBL, the theory behind the clustering methods used, and particularly focusses on revealing the conformational differences of the active site residues in the case of pyrrolidine carboxamides in contrast to PDB 2X23 that represents a closed SBL (EI*) state.

10.2.1 Atom selections

As seen in Chapter 8, the essential dynamics analysis was centered around the residues of the active site and the SBL. However, the definitions of the active site and SBL as used by Luckner et al. [72] and Merget et al. [261] differ from the ones used in the current work. In the aforementioned works, the following residues constituted the active site: Phe149, Tyr158, Ala198, Met199, Ile202 and Val203. In the current work, however, all residues within a radius of 5 Å around the bound ligand were considered to form the active site (cf. Chapter 8). Clearly, our new definition considers a greater number of residues in the subsequent analyses.

Additionally, according to Luckner et al. [72], the definition of the SBL includes residues Ile202 to Ile218, while Li et al. defined the SBL from Leu197 to Glu210 [74]. However, the current work defines the SBL as the stretch of residues from Arg195 to Gly212. In order to account for the 6 residues 213 to 218, that have not been considered in the new definition of the SBL described so far, these missing 6 residues were added to the earlier definition of the active site to give rise to a new atom selection termed *"extended active site"*.

- **Active site:**
  The definition of active site was set to the one used by Luckner et al. [72].
- **Extended active site:**
  Coined mainly to account for the discrepancies between the current SBL definition (residues 195 to 212) and that made by Luckner et al. [72]. Corresponds to residues situated within a 5 Å radius of the bound ligand.

10.2.2 Systems analysed

In order to comprehensively portray the effect of pyrrolidine carboxamides (including the designed ones) on InhA and to compare the conformational changes in its active site residues with respect to the published results, a total of 9 protein-ligand complexes has been analysed. Of these, two represent light pyrrolidine carboxamides as well as the crystal structures (PDB 4TZK and 4TZT), two represent the most potent compounds from the bulky class (4TZK-c6a3 and 4TZK-c7a3) while the remaining 5 are the compounds whose design was driven by essential dynamics. These 5 compounds represent the most favourable (2NSD-b3, 2NSD-b6, 2X23-b9, and 2NSD-b12) as well as one unfavourable (2NSD-b7) substitution on the pyrrolidine carboxamide scaffold. In order to capture the slow dynamic changes in the conformations of the active site residues for the aforementioned ligands, long MD simulations (150 ns per protein-ligand complex) were utilised. The total length of these simulations was expected to yield insights into the intermediate conformations that populate the EI state. All of the 150 ns simulations were extensions (under the same conditions) of the 5 ns MD runs performed for the binding affinity prediction in Chapter 4.

10.2.3 Clustering

In order to elucidate the dominant conformational families of the binding site residues in a ligand bound state, conformational clustering techniques were used. Cluster analysis is an unsupervised technique adept in ascertaining similar patterns in complex data as for those obtained from MD simulation. For example, a cluster is a set of data points in such a manner that an individual point of that bunch is closer (more similar) to every other point of that bunch than any other set of points (dissimilar) [351–353]. There are various types of clusters, with the reader being referred to Tan et al. [351] for more details. The hierarchical agglomerative approach will be discussed in more detail since it yields similarly sized clusters with readily interpretable results. An added value of the hierarchical approach is the dendogram which can be partitioned at a specific level yielding a partitional clustering. Another advantage of this technique is that there is no need to assume any particular number of clusters [351].

10.2.3.1 Agglomerative hierarchical clustering:

This approach belongs to a group of techniques that begin with numerous singular clusters and iteratively merges the singletons to their nearby neighbours until all objects form a single cluster [351]. By using this approach on an MD trajectory, similar conformations can be grouped together whilst separating the distinct ones, thereby yielding the distinct conformations that have been sampled. There are different implementations of this method (cf. Tan et al. [351], Wolf et al. [353], and Torda et al. [352]): *single-linkage*, *complete-linkage*, *average-linkage*, *wards method*, and *centroid*.

In our work, agglomerative hierarchical clustering was used primarily because it is deterministic, i.e., it allows reproducibility of the resulting clusters. This is one of the key advantages over K-means or even the divisive approach. The dendogram obtained from this method effectively portrays the relationships between clusters and their sub-clusters. Furthermore, this method also highlights the order in which the dendogram was built. In order to minimise the time and space requirements for clustering of the MD simulations, a 2D-RMSD matrix that aids in visual analysis of the conformational changes was subjected to clustering.

10.2.3.2 Cluster validation metrics

Clustering being an unsupervised learning technique is quite difficult to assess, primarily due to the lack of a native evaluation criterion. Nevertheless, there are several metrics that offer a generalised indication of cluster quality, each with their strengths and drawbacks [351, 352, 354]. Of these metrics, the most commonly used ones are the following:

- **Calculation of optimal cluster number:**
  A straightforward way to calculate the number of clusters that would portray the

patterns in the input data is calculating the SSR/SST ratio: the sum of squares regression (SSR) divided by the total sum of squares (SST) [353].

- ***Pseudo F-statistics*:**
  An additional metric based on the theory of ANOVA (analysis of variance) was introduced by Caliński and Harabasz [355] termed the *pseudo F-statistic* (Equation (10.1)). This metric is also referred to as Caliński-Harabasz index (CH index/ criterion). It usually indicates the degree of "tightness" or "proximity" for the clusters, with high values indicating better nested clusters. In the current work, the quality of clustering was evaluated with the aid of the CH index.

$$pFS \;=\; \frac{\left(\frac{SSR}{K-1}\right)}{\left(\frac{SSE}{N-K}\right)} \tag{10.1}$$

- **Davies-Bouldin index:**
  An alternative internal evaluation metric is the *Davies-Bouldin index* (DB) [356] that provides a dataset dependent value for evaluating the clustering results.

10.3 Results

A 2D RMSD matrix that represents the conformational changes over the course of a trajectory is used as input for the clustering. Before initialising the clustering, it is important to ascertain the differences in the conformations of the residues that have already been involved in binding of pyrrolidine carboxamides and diphenyl ethers. Table 10.1 shows the RMSD values for important residues that form either a part of the active site or the SBL. From this table, it is apparent that maximal changes occur in the residues that are located in the SBL, with especially large changes being observed for Ile202 and Val203, respectively. Both of them have already been known to be important molecular indicators of the inhibitor's ability to bring about ordering of the SBL.

In regards of the molecular indicators involving diphenyl ethers, it is obvious that the stretch of residues from Ala198 to Val203 determines their binding [261]. However, as seen from He et al., there are several other residues that interact primarily with pyrrolidine carboxamides [52]. Hence, the focus of the clustering was to reveal the conformational changes in these additional residues together with Met199, Ile202, and Val203, respectively. In order to reveal the conformational changes occurring in the ligand-bound state, a 2D RMSD plot of all 9 complexes (Section 10.2.2) against each other and themselves has been drawn for all atom selections as defined in Section 10.2.1. The 2D RMSD matrix was then subjected to a K-means clustering to get an appropriate number of clusters as defined by the Caliński-Harabasz index. Subsequently, the hierarchical agglomerative clustering was

Table 10.1 Comparison of RMSD values for key residues of the binding pocket and substrate binding loop (Met199, Ile202, and Val203) for chain "A" of selected InhA crystal structures with respect to chain A of PDB 2X23 upon C-α atom alignment in MOE 2015.10. The residues that form part of the active site (top 4) are separated from those of the SBL (bottom 4) by a horizontal line. The crystal structures are ordered by an increasing degree of "open" state for the SBL. The RMSD is given in Å.

Residue	4OXK	2NSD	4TZK	1BVR	1P44
Phe149	0.28	0.15	0.20	0.15	0.11
Met155	0.44	0.34	0.26	0.49	0.30
Tyr158	0.42	0.36	0.55	0.63	0.55
Lys165	0.29	0.28	0.13	0.42	0.17
Ala198	1.11	1.88	3.76	3.57	2.81
Met199	3.05	0.52	2.18	2.62	2.73
Ile202	4.18	3.54	4.70	7.73	8.09
Val203	4.84	4.50	7.54	8.70	9.53

performed with the optimal number of clusters as suggested by the Caliński-Harabasz index.

10.3.1 Active site clustering

The 2D RMSD matrix of the active site defined by Luckner et al. (Figure 10.3) allows a comparison of the conformational changes in each of the individual protein ligand complexes over 150 ns. From Figure 10.3, several interesting trends can be summarised as follows:

1. With the exception of the protein-ligand complexes of pc-c63, pc-c7a3 and pc-b3 (2NSD-b3), all other complexes exhibit marked deviations compared to PDB 4TZK (pc-d11). This is not surprising considering the different conformations of the SBL in PDB 4TZK and the rest of the complexes which are mostly PDB 2NSD like.

2. Compound pc-d8 corresponding to PDB 4TZT represents the most peculiar and distinguishing trait of unstable binding even amongst the supposedly rapid reversible binders. The destabilising effects of the A ring substitutions are manifested as the ligand nearly exiting the binding pocket (Figure 10.4). This is accompanied by a marked destabilisation of the SBL, particularly in the region spanning Met199-Val203 that is relevant from their lower average secondary structure propensities (Table 10.2) as calculated by cpptraj [305]. The induced instability in the binding pocket manifests as high RMSD value of 4TZT with respect to the rest of the protein-ligand complexes (> 5 Å).

3. The protein-ligand complexes of pc-b6 to pc-b12 all show conformational changes in the active site residues compared to pc-d11, pc-c6a3, and pc-c7a3. Furthermore,

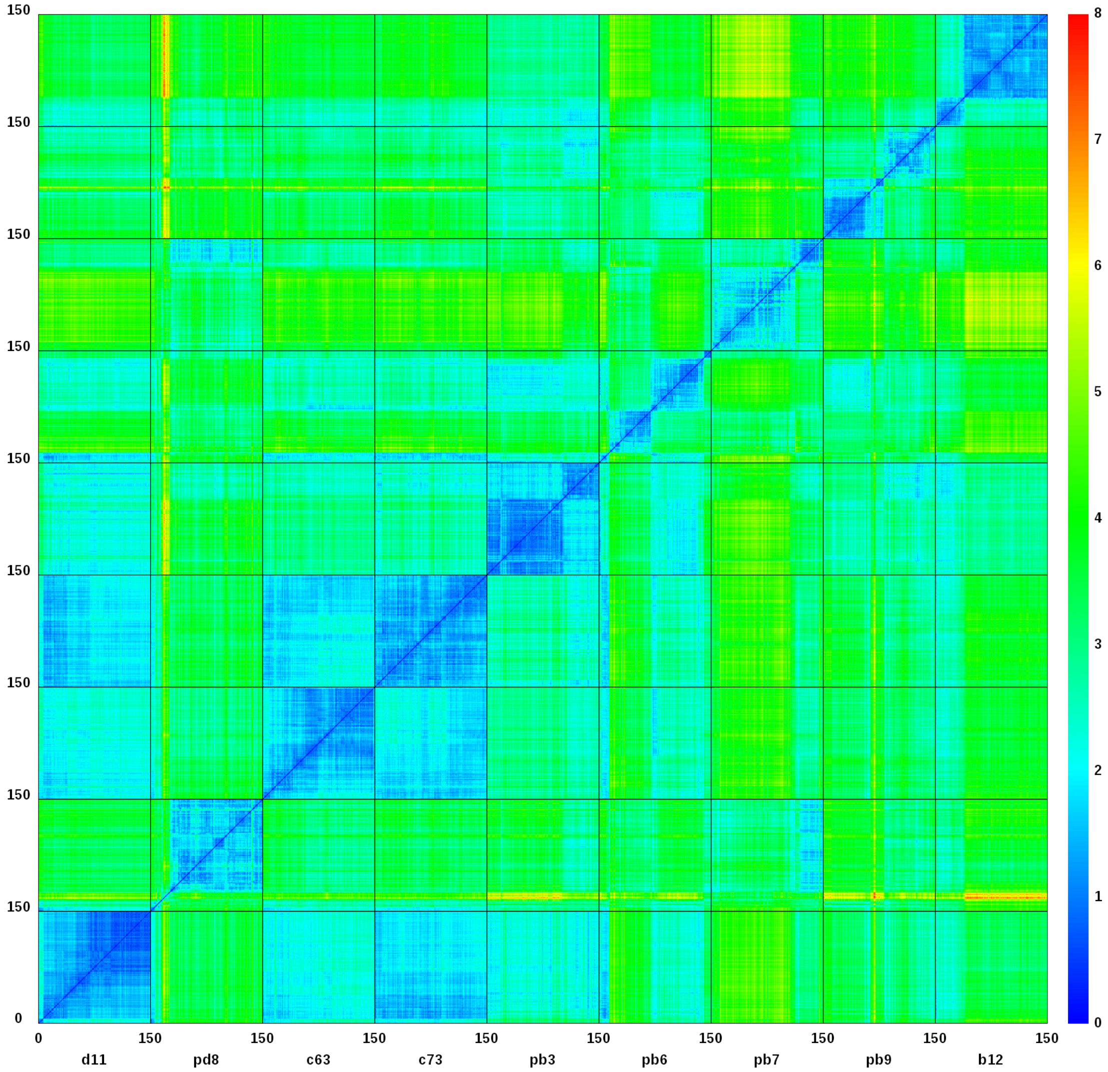


Figure 10.3 A (9×9) 2D-RMSD matrix for the active site residues (heavy atoms of Phe149, Tyr158, Ala198, Met199, Ile202, and Val203) of chain A from PDB 4TZK, 4TZT and pyrrolidine carboxamides-c6a3, c7a3, b3, b6, b7, b9, and b12. RMSD values in between each frame are illustrated by the color scale on the left. Each small box corresponds to a 150 ns simulation of a monomer and represents a comparison of the conformational changes (snapshots) either within the system (boxes along the diagonal) or with other systems (off-diagonal boxes).

all protein-ligand complexes of the new pyrrolidine carboxamides exhibit noticeable conformational changes compared to PDB 4TZK.

The K-means clustering performed on the 2D RMSD matrix reveals that the conformational changes occurring throughout the combined simulations (1.35 μs in total) can be represented as 6 clusters (Figure 10.5). Starting with the total number of clusters set to 6, the hierarchical agglomerative clustering was performed. The 6 clusters obtained by using a cutoff of 6 Å can be subsumed into 3 ”monophyletic” conformational families. The subsuming of the 6 clusters to the three families has been achieved via visual inspection while sidestepping the need for increasing/decreasing the RMSD cutoff. This was

Table 10.2 Average secondary structure propensities (in percentages) for key residues of the SBL in case of PDB 4TZT (pc-d8) over a 150 ns MD simulation. The terms para and anti refer to parallel and anti-parallel beta sheet, while 3_{10}, Alpha and Pi correspond to the various types of helical content in the protein-ligand complex, respectively. Values in the lower half are for the reference protein-ligand complex, i.e., PDB 4TZK.

Residue	Para	Anti	3_{10}	Alpha	Pi	Turn
Ala198	0%	0%	23%	47%	0%	30%
Met199	0%	1%	23%	48%	0%	28%
Ile202	0%	0%	4%	4%	0%	4%
Val203	0%	0%	3%	3%	0%	5%
Ala198	0%	0%	1%	96%	0%	3%
Met199	0%	0%	1%	98%	0%	1%
Ile202	0%	0%	1%	34%	0%	65%
Val203	0%	0%	0%	27%	0%	30%

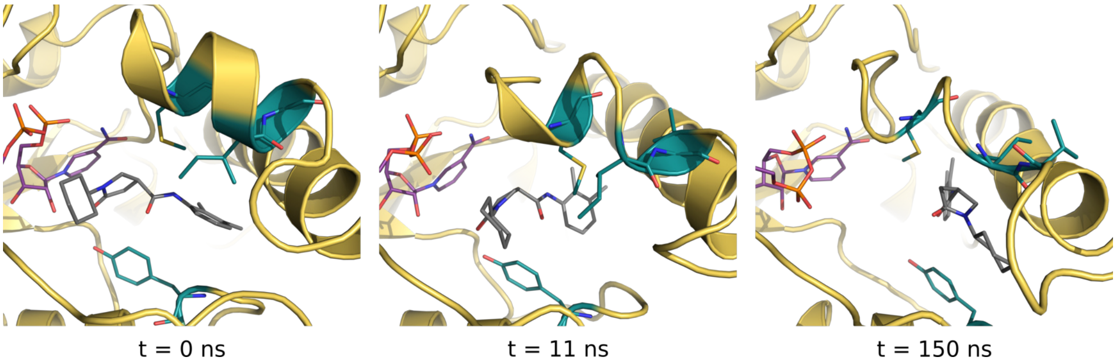


Figure 10.4 The exit pathway for pc-d8, the ligand of PDB 4TZT (grey sticks) as observed at t = 0 ns (left most), t = 11 ns (center) and t = 150 ns (right). Note the conformational changes in the SBL from an α helical structure to a turn and finally a distorted mixture of a 3_{10} helix and a coil.

mainly because the increase/decrease in the RMSD cutoff can lead to an overestimation of minor backbone movements whilst simultaneously overlooking the important side chain movements. These families will be referred to as families 1 to 3 (cf. Figure 10.7) hereinafter:

1. **Family 1** (based on clusters 1, 2, and 4): Represents a dominant conformational substate observed throughout the total simulation duration of 1.35 μs. It represents an intermediate conformation between those of the substrate binding loop of PDB 4TZK and 2NSD. This family is characterised by a slight but noticeable shift of Ile202 towards the ligand while Val203 moves away from the ligand turning to the outside. Quantitatively speaking, this conformational family accounts for 81% of all observed conformations across all simulations (Table 10.3 and figure 10.7). When compared to the work of Merget et al., family 1 has a remarkable resemblance to the open conformation of the SBL (Family 3) [261].

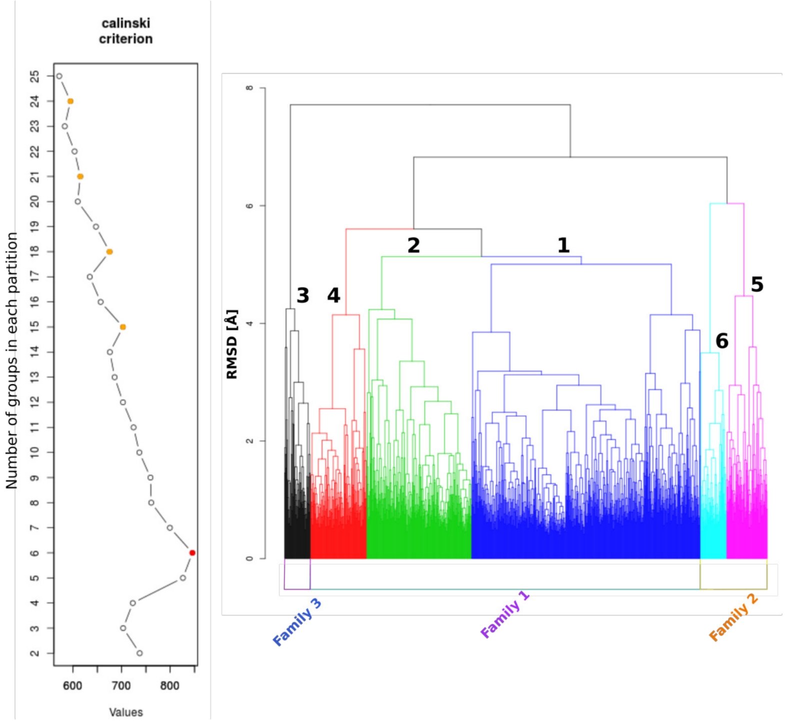


Figure 10.5 Hierarchical clustering analysis of binding site conformations for various protein-ligand complexes from the pyrrolidine carboxamides d11, d8, c6a3, c7a3, b3, b6, b7, b9, and b12 based on the 2D RMSD matrix in Figure 10.3. (a) The elbow plot on the left depicts the optimal number of clusters (red circle) as deemed by the K-means clustering and the Caliński-Harabasz index. (b) The dendogram on the right shows the annotated clusters and their conformational families upon subsumption.

2. **Family 2** (based on clusters 5 and 6): Represents a smaller conformational family which features a slight shift of Val203 away from the ligand alongwith Ile202 moving towards the binding pocket. The conformations of Ile202 and Val203 from this family are intermediate to those observed for the same residues in crystal structures 2NSD and 2X23. When compared against the clustering results of Merget et al. [261], this family was found to be identical to the second most populated cluster (family 2).

3. **Family 3** (based on cluster 3): A small standalone family that is characteristic for the loop destabilisation seen in the case of PDB 4TZT. This family is represented by the ligands pc-b6 and pc-b7, respectively. Its characteristic feature is a near complete transition from an helix to a loop that is accompanied with Ile202 occupying the position of Val203 while Val203 is completely pushed out by the meta substituent (-CN for pc-b6 and -CF_3 for pc-b7) on ring A2 (cf. Figure 10.7). In both cases, Ile202 is pushed away from the ligand momentarily due to the motions of the residues Leu207-Ala212 of the hinge region. The residues Leu207-Ala212 move in response to the motions of the A2 ring and its meta substituent. However, in the case of pc-b6, Ile202 moves in and out of the active site as opposed to

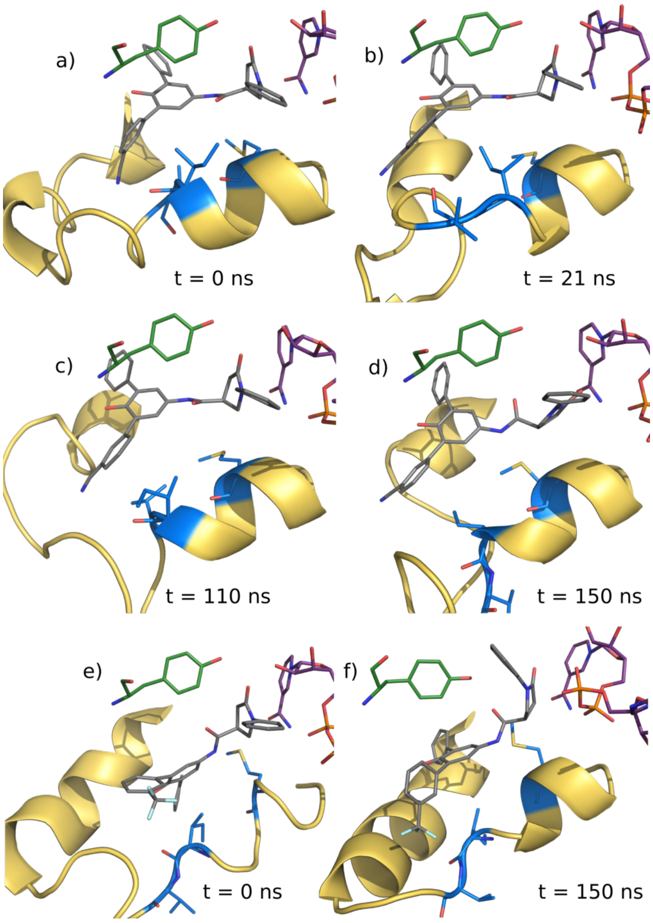


Figure 10.6 Conformational changes in the key residues of the SBL, i.e., Met199, Ile202, and Val203 (blue sticks) throughout a 150 ns MD simulation of 2NSD-pb6 (a) to d)) and 2NSD-b7 (e) and f)), respectively.

pc-b7, where it is just situated outside since the start of the 150 ns simulation (Figure 10.6).

10.3.2 Extended active site clustering

A clustering of the extended active site has been performed mainly to ascertain the role of amino acids other than Met199, Ile202, and Val203 in a more stable ligand binding. The clustering of the 2D RMSD values for the extended active site (cf. Figure 10.8) yielded a total of 9 clusters (cf. Figure 10.9) at a cutoff of around 4.8 Å. Upon subsumption, the 9 clusters yielded 2 major and one comparatively smaller conformational family (cf. Figures 10.10 and 10.11) as follows:

1. **Family E1:** This superfamily comprises clusters 1, 2, 3, and 7 (Figure 10.9). This family contains the dominant (cluster 1, N = 463 frames) as well as the least visited conformations (cluster 2, N = 12 frames) from amongst all of the simulated protein-ligand complexes. Likewise, in the case of active site clustering,

Table 10.3 Occurrence frequencies (in %) of the various conformational families of the InhA active site in the MD simulations of various pyrrolidine carboxamides. The values denote the fractional occurrences of the conformational families in an individual trajectory (150 · 1 ns). Additionally, the sum (in %) represents the average assignments of all 9 simulated protein-ligand complexes to the conformational families.

	Family 1	**Family 2**	**Family 3**
4TZK	100	0	0
4TZT	90	0	10
4TZK-c6a3	100	0	0
4TZK-c7a3	100	0	0
2NSD-b3	100	0	0
2NSD-b6	57	6	37
2NSD-b7	28	39	33
2X23-b9	51	49	0
2NSD-b12	100	0	0
Sum	81	11	8

the superfamily 1 accounts for the bulk of conformations sampled/adopted by the protein-ligand complexes over a period of 150 ns. The medoid that represents this conformational super family corresponds to the complex 4TZK-c6a3. A comparison of the medoid snapshot (t = 35 ns) and the active site of PDB 2X23 reveals noticeable changes in almost all residues of the SBL. Additionally, substantial changes are also observed for Met155 that is situated near the minor exit portal (Figure 10.10). The marked changes in the SBL as well as of Met155 can be attributed to the inverted binding mode of pc-c6a3 that pushes Ile202 and Val203 further away from itself. Simultaneously, Leu207 moves away from the pocket while Ile215 and Leu218 move towards the pocket, primarily because the cyclohexyl ring (C ring) occupies the position of the bulky A ring system. This conformation corresponds to a wide open state of the $\alpha6$ helix, and hence corresponds to conformational family 3 of Merget et al. [261].

2. **Family E2:** This is a minor conformational family comprised of clusters 6 and 9 (Figure 10.9). This family is represented by 2NSD-b3, with the medoid snapshot corresponding to t = 6 ns. It is characterised by large changes in the positions of Tyr158, Met155, Met161, Leu207, Ile215, and Leu218 (Figure 10.11). Furthermore, both Ile202 and Val203 are shifted slightly away from the ligand. Upon comparison of the SBL of these two complexes, it can be inferred that the mid to lower portion of 2NSD-b3 moves away from the ligand while the starting portion of the SBL largely remains stable. It can also be seen that this conformation is intermediate to that of families 3 and 5 from the clustering performed for long MD simulations of diphenyl ethers.

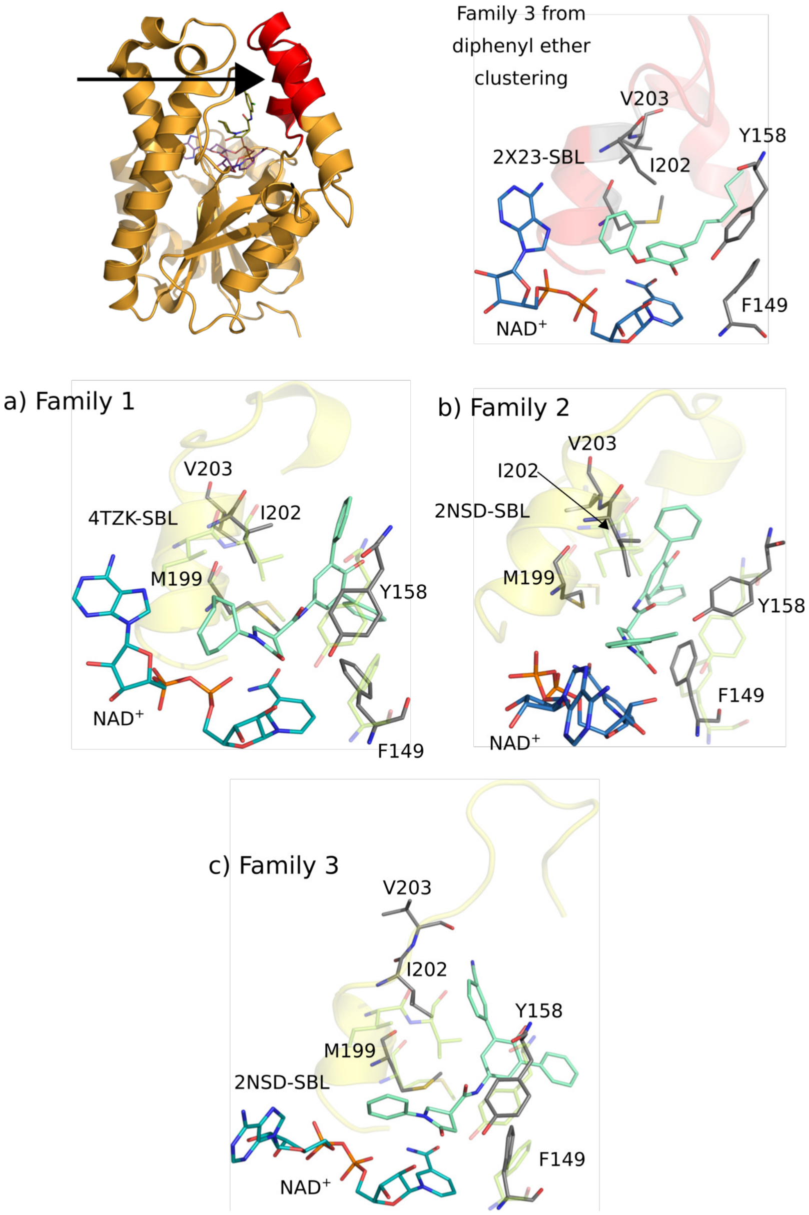


Figure 10.7 Dominant conformational families for pyrrolidine carboxamides: The 6 clusters from hierarchical clustering analysis were subsumed to 3 conformational families. Subsequently, a Partitioning Around Medoids (PAM) was performed in R around each conformational family that yielded medoids or cluster representatives. The top left figure depicts the entire chain A of PDB 4TZK, while the substrate binding loop is coloured red. The arrow denotes the viewpoint for subsequent images. The top right figure depicts the conformation of family 3 as seen from Merget et al. [261]. a) Family 1: Conformation of 4TZK-c7a3 after 122 ns of MD simulation. The important residues involved in binding have been depicted as grey sticks and labelled. The SBL has been coloured yellow and annotated. Furthermore, the same residues of PDB 2X23 have been depicted as transparent green sticks; cofactor heavy atoms as purple sticks. The shift of Ile202 and Val203 (of PDB 2X23) away from the bound ligand is clearly visible in figure a). b) Family 2: Corresponds to 2NSD-b3 after 38 ns of MD simulation. The slight shift of Ile202 and Val203 along with a major shift of Y158 can clearly be seen. c) Family 3: Corresponds to 2NSD-b6 after 56 ns of MD simulation and shows a major shift of Ile202 and Val203 further away from the bound ligand. The α-helix to coil conversion is also visible.

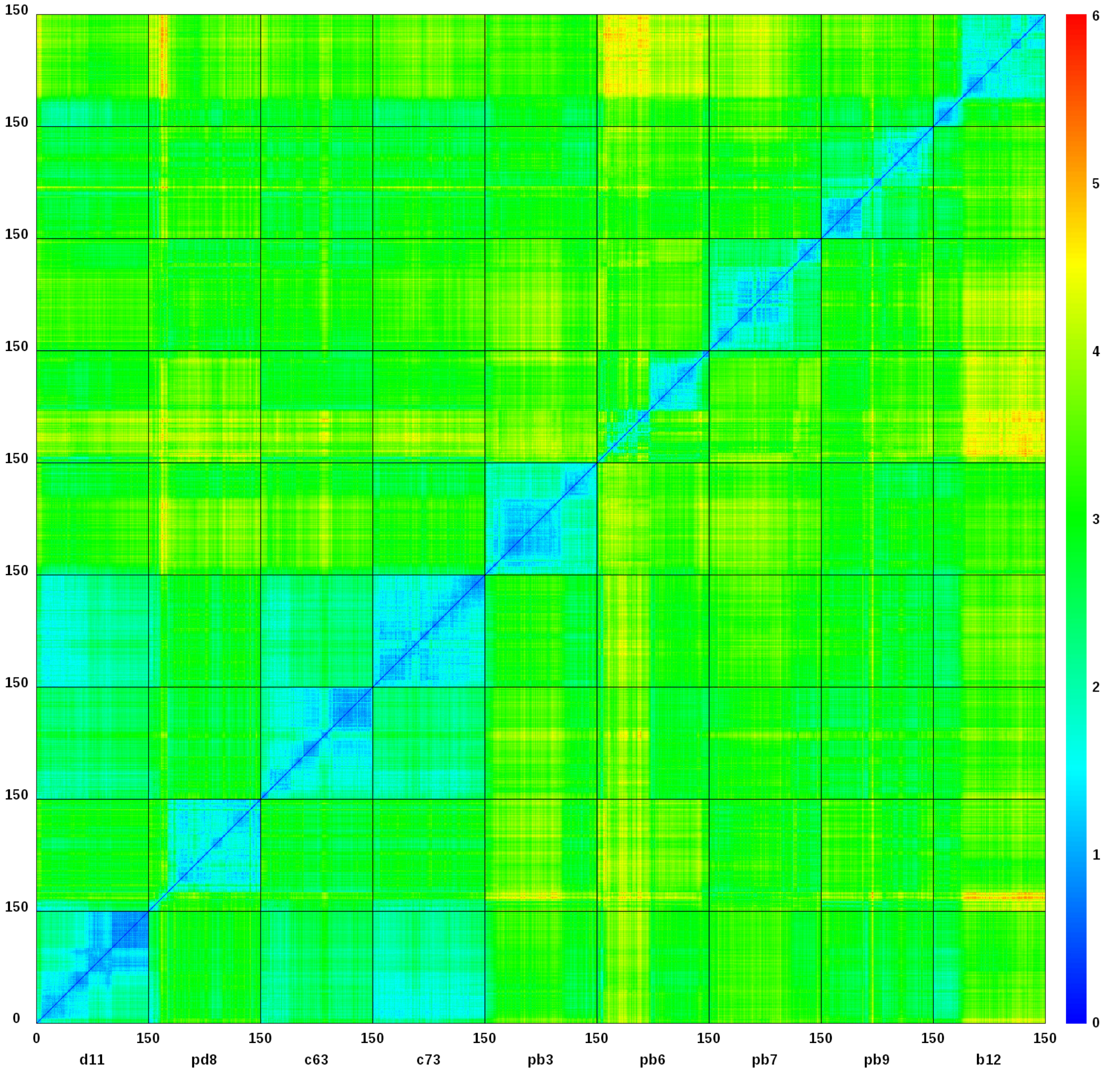


Figure 10.8 A 9 × 9 2D-RMSD matrix for the extended active site residues (residues within 5 Å of the bound ligand) for monomers (chain A) of PDB 4TZK, 4TZT and pyrrolidine carboxamides-c6a3, c7a3, b3, b6, b7, b9, and b12, respectively. The RMSD values in between each frame are illustrated by the color scale on the left. Each small box corresponds to a 150 ns simulation of a monomer and represents a comparison of the conformational changes (snapshots) either within the system (boxes along the diagonal) or with other systems (off-diagonal boxes).

3. **Family E3:** The superfamily 3 represents another dominant conformation (cf. Figure 10.11) and is comprised of three clusters, namely 4, 5, and 8. The medoid snapshot for this conformational family belongs to 2NSD-b6 after a period of 123 ns. This superfamily corresponds to a conformation where Ile202 is pushed away from the ligand and Val203 takes its place. This occurs primarily due to the transition of an α6 helix to a coil and later to a loop that results in dramatic movements of residues 203 to 218 (Figure 10.11). The movement of the SBL can be viewed as follows: α6 portion away from the ligand along with a simultaneous motion of the α7 helix towards the A1 ring of the ligand. This superfamily corresponds to family

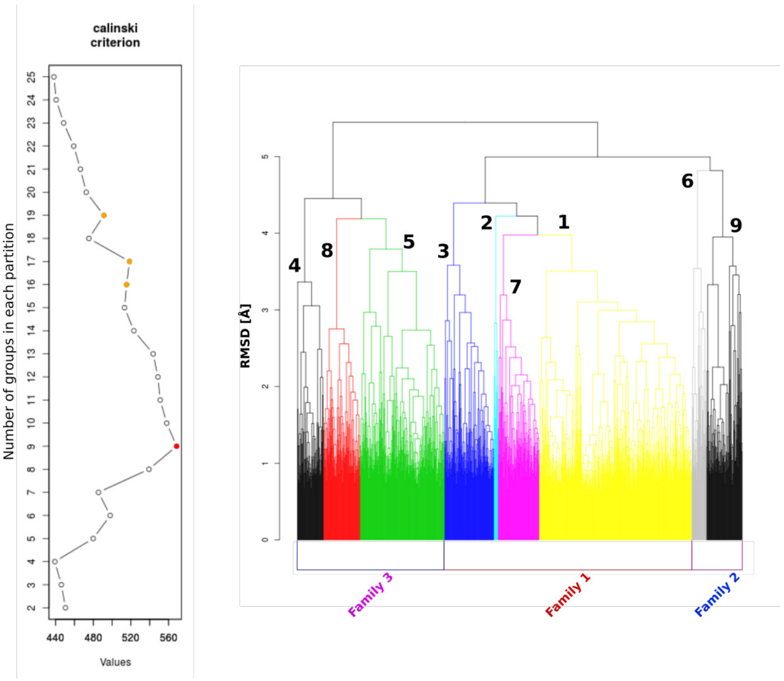


Figure 10.9 Hierarchical clustering analysis of binding site conformations for various protein-ligand complexes from pyrrolidine carboxamides-d11,d8, c6a3, c7a3, b3, b6, b7, b9, and b12 based on the 2D RMSD matrix from Figure 10.8. (a) A typical elbow plot depicting the optimal number of clusters (red circle) as deemed by the K-means clustering and the Caliński-Harabasz index. (b) The dendogram on the right depicts the annotated clusters and their conformational families upon subsumption.

5 (a very wide open α6 helix) from the literature [261].

Qualitatively speaking, family E1 accounts for 66% of the conformations sampled during the MD simulations followed by family E3 (24%) and family E2 (11%) (cf. Table 10.4). A comparison of the clustering of the residues of the active site and the extended active site revealed the following:

1. In both cases, the most dominant conformation corresponds to an **open** state of the α6 helix and is observed in PDB 4TZK and 2NSD. This state has also been observed in the long MD simulations (150 ns) of triclosan (TCL) bound to PDB 2X23. Moreover, a majority of the simulated protein-ligand complexes features no substantial motions in the SBL.

2. The other dominant conformation observed in both cases corresponds to that of 2X23-b9 and 2NSD-b12 specifically. Both are accompanied with a slight shift of Ile202 and Val203 away from the ligand but not completely away as in the case of pc-b6 or pc-b7. This conformation is intermediate to 2X23 on one end and 4TZK-2NSD on the other. In short, both 2X23-b9 and 2NSD-b12 feature less SBL movement than the other pyrrolidine carboxamides.

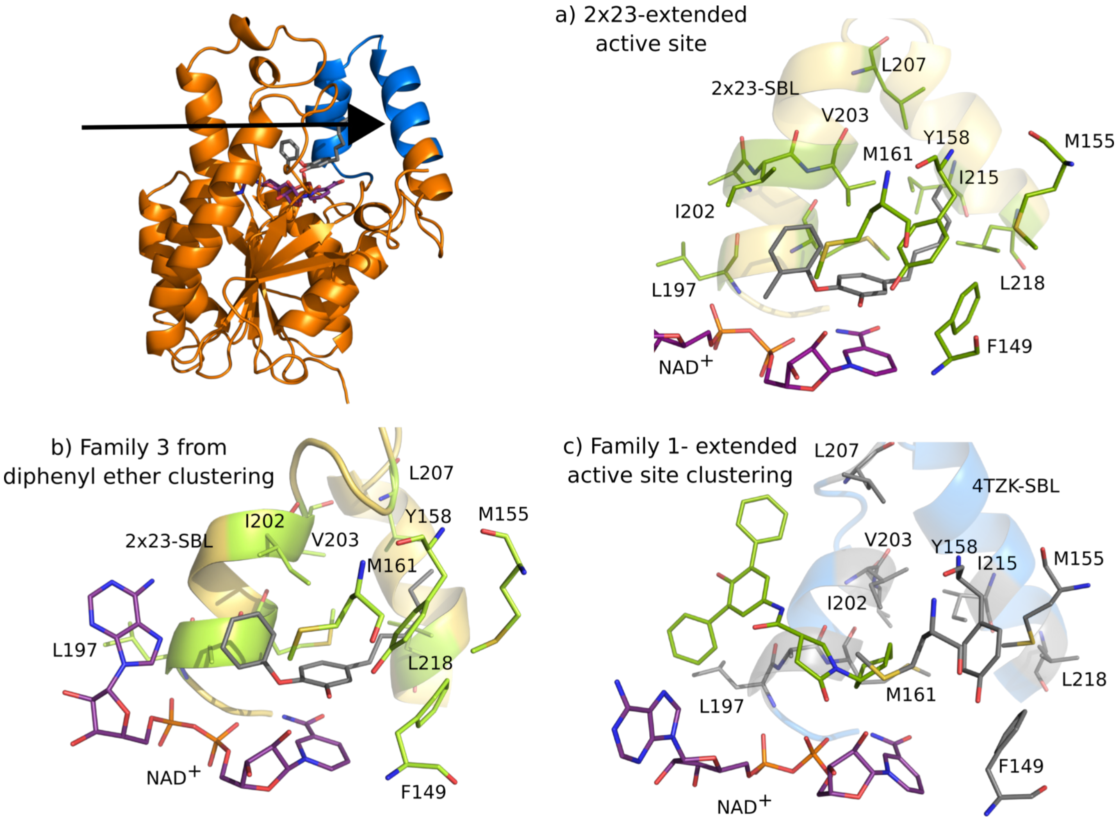


Figure 10.10 Dominant conformational families for pyrrolidine carboxamides: The 9 clusters from hierarchical clustering analysis were subsumed to 3 conformational families. Subsequently, a Partitioning Around Medoids (PAM) was performed in R around each conformational family that yielded medoids or cluster representatives. The top left figure depicts the entire chain A of PDB 2X23, while the substrate binding loop is coloured blue. The arrow denotes the viewpoint for subsequent images. a) The extended active site for PDB 2X23. b) The conformation of family 3 as seen from Merget et al. [261]. c) Family E1: Conformation of 4TZK-c6a3 after 35 ns of MD simulation. The important residues involved in binding have been depicted as grey sticks and labelled. The SBL has been coloured blue and annotated. Furthermore, the cofactor heavy atoms have been coloured as purple sticks. The shift of Ile202 and Val203 (of PDB 2X23) away from the bound ligand is clearly visible in figure c).

3. The least visited and populated conformations were observed in 2NSD-b6 and 2NSD-b7. The ligands pc-b6 and pc-b7 bring about a helix to loop transformation of a large interacting portion of the SBL. This is evident from the observation that both of the aforementioned compounds push Ile202 and Val203 away from the binding pocket that characterises the helix to loop transformation. However, in case of pc-b6, Ile202 oscillates in between positions that are near and far from the bound ligand after a period of 21 ns (cf. Figure 10.6). This is not the case for pc-b7 which consistently pushes Ile202 and Val203 away from the binding pocket since the start of the simulation. In both cases, it can be clearly seen that **loop destabilisation** occurs as opposed to the desired loop ordering, which is associated closely with slow-tight binders. A similar observation holds true for the crystal structure 4TZT, where the ligand was found to nearly exit the binding pocket at the end of the 150

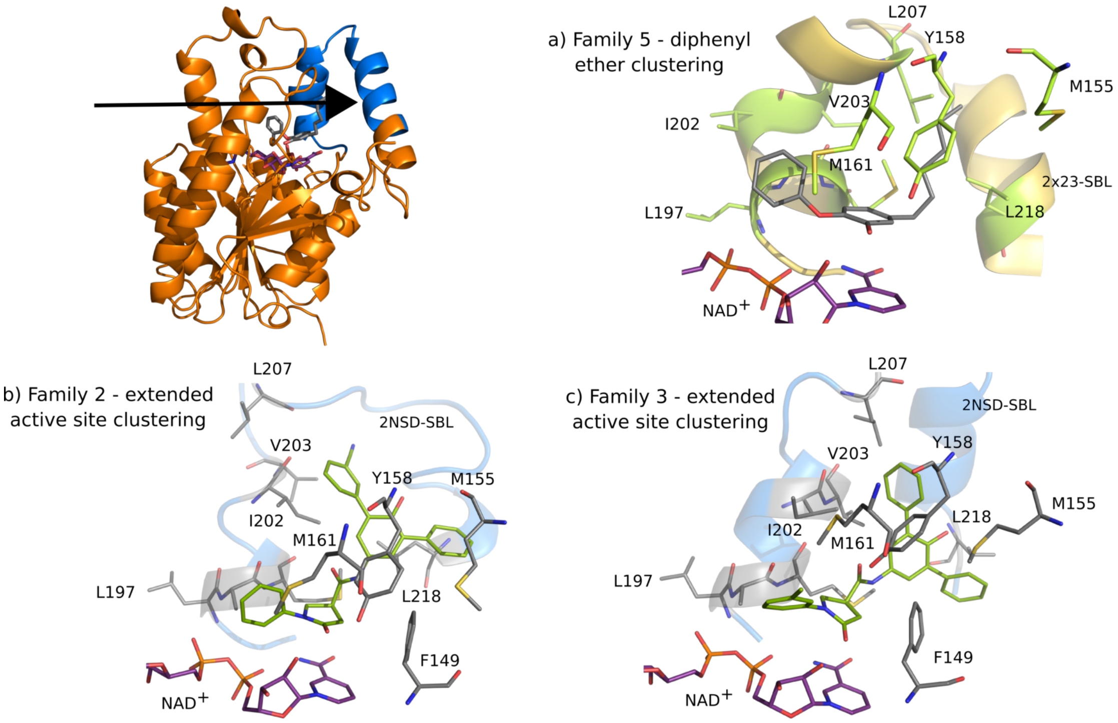


Figure 10.11 Dominant conformational families for pyrrolidine carboxamides: a) The conformation of extended active site residues for family 5 as seen from Merget et al. [261]. b) Family E2: Conformation of 2NSD-b6 after 6 ns of MD simulation. The important residues involved in binding have been depicted as grey sticks and labelled. The SBL has been coloured blue and annotated. Furthermore, the cofactor heavy atoms have been coloured as purple sticks. The shift of Ile202 and Val203 (of PDB 2X23) away from the bound ligand is clearly visible in figure b). c) Family E3: Conformation of 2NSD-b3 after 123 ns of MD simulation. As compared to the family E2, family E3 exhibits a lesser shift of Ile202 and Val203 away from the bound ligand.

ns trajectory, resulting in destabilisation of the SBL (Figure 10.4).

10.3.3 Analysis of SBL dynamics and its secondary structure

The **ordering and subsequent closure of the SBL is a decisive aspect of slow-tight binding inhibition** of InhA. Thus, the dynamics of the SBL and its closure upon ligand binding warrants special attention. The changes in the SBL can be ascertained by studying backbone deviations or C-α RMSD's of the simulated complexes with respect to a reference system. In the current case, the backbone RMSD values were calculated using the chain A of PDB 2X23 as reference. Expectedly, all of the ligand bound systems exhibited higher than normal average backbone RMSD values that ranged from 1.78 Å (lowest, PDB 4TZK-c7a3) to 2.20 Å (highest; 2NSD-b7). Nevertheless, the median values of all systems remained below 2.5 Å, implying a reasonable degree of stability. A comparison against the 150 ns monomer simulation of PT70 [261] clearly revealed the differences in the backbone distributions of PDB 2X23 (EI*) and all of the other complexes (EI macrostate).

Table 10.4 Occurrence frequencies (in %) of the various conformational families of the InhA active site (extended) in the MD simulations of various pyrrolidine carboxamides. The values denote the fractional occurrences of the conformational families in an individual trajectory (150 · 1 ns). Additionally, the sum (in %) represents the average assignments of all 9 simulated protein-ligand complexes to the conformational families.

	Family E1	**Family E2**	**Family E3**
4TZK	100	0	0
4TZT	90	0	10
4TZK-c6a3	100	0	0
4TZK-c7a3	100	0	0
2NSD-b3	99.33	0.66	0
2NSD-b6	0	100	0
2NSD-b7	100	0	0
2X23-b9	0	0.66	99.33
2NSD-b12	0	0.66	99.33
Sum	66	11	23

When the backbone RMSD of the SBL is considered, the loop movements become amply clear. All of the simulated systems show an average backbone RMSD exceeding 3 Å. In this case, 2X23-b9 exhibited the lowest average backbone RMSD (3.39 Å), while 2NSD-b6 showed maximal change in the average backbone RMSD value (4.33 Å). Surprisingly, the average backbone RMSD value for 2NSD-b6 was much lower (1.92 Å). The higher values of the backbone RMSD for 2NSD-b6 and-b7 merely signify the helix to loop transformation (and thereby a more open conformation) induced by these two compounds. A similar observation can be seen for the PDB 4TZT ligand (pc-d8). When compared to the work of Merget et al. [261], again it was seen that highest flexibility of InhA resides in the SBL. This can be deduced from the fact that the average backbone RMSD for the SBL of PDB 2X23 (from 150 ns monomer simulation) was around 2.2 Å as opposed to > 3 Å for all of simulated protein-ligand complexes. Figure 10.12 depicts the RMSD distributions for the simulated protein ligand complexes.

Since there is a close relationship between ligand binding, the ordering and subsequent closure of the SBL, a secondary structure analysis of the SBL for the each of the simulated complexes was performed using cpptraj. Cpptraj calculates the secondary structural propensities for the backbone atoms of the residues of the SBL (residues 195 to 217) using the DSSP method as proposed by Kabsch and Sander [357]. According to the DSSP method, seven secondary structural motifs can be assigned to the backbone atoms namely: (1) none, (2) parallel β-sheet, (3) anti-parallel β-sheet, (4) 3-10 helix, (5) α-helix, (6) π-(3-14) helix, and (7) turn. The SBL of PDB 2X23 consists entirely of α-helix and 3-10 helix as opposed to that of PDB 4TZK and 2NSD whose SBL's are comprised of a coil and an α helical structure exclusively. In the case of the simulated complexes, the average percentage of the aforesaid secondary structure motifs were calculated over 150

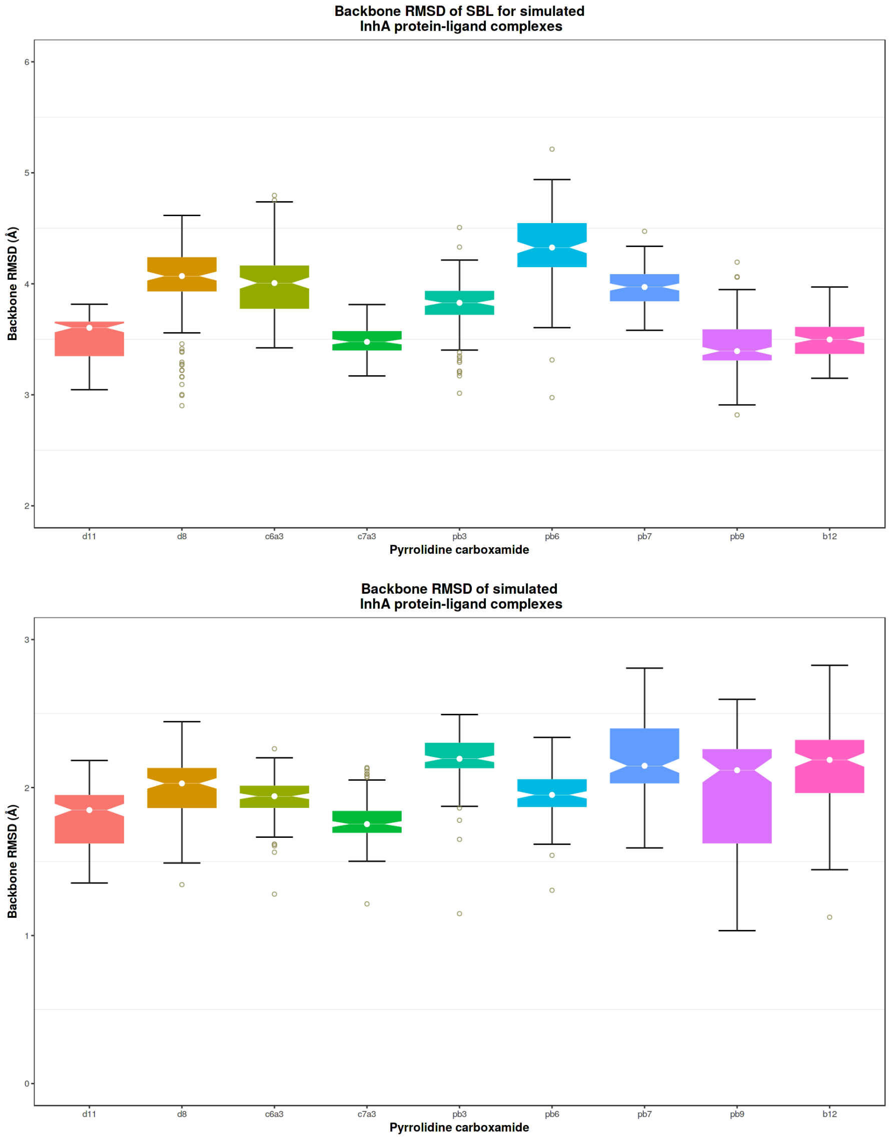


Figure 10.12 Distribution of backbone (atoms C, CA, N, and O) RMSD for the (a) substrate binding loop (residues 195 to 218, top), and (b) entire protein (bottom) of the simulated protein ligand complexes. Each monomer was simulated for 150 ns and fitted individually onto PDB 2X23 chain A (reference structure). The boxes represent the interquartile range for each monomer. The white circles denote the median value of the RMSD's, while outliers are shown as khaki coloured circles.

frames, with each frame corresponding to 1 ns (150 ns = 150 frames).

From work of Merget et al. [261], the SBL of PDB 2X23 shows an average of 68% α-helix and 3-10 helix motifs. This is followed by two rapid reversible inhibitors **6PP** (63%) and **triclosan (TCL)** (47%), while the lowest proportion of aforementioned motifs were observed for the apo-proteins (32%). The progressive decrease in the average percentage of helical motifs for 6PP and TCL is parallel to the decrease in their InhA inhibitory activity. This buttressed the notion that adequate occupation of the binding pocket goes hand in hand with retention of the helical motifs that represent the EI* state. Table 10.5

depicts the average percentage for the secondary structural motifs in the simulated systems. All of the systems clearly show a marked change in the average percentage of either α-helix or 3-10 helix motifs as compared to that of PDB 2X23, except for PDB 4TZK (60%), 4TZK-c7a3 (64%), and 2X23-b9 (60%). The low values of helical motif content for PDB 4TZT (33%) and 2NSD-b6 (36%) justify the earlier observations pertaining to the induced loop disordering upon ligand binding. Comparatively, 2NSD-b7 (averaged helical motif content: 55%) fares much better than pc-d8 (PDB 4TZT) and pc-b6 (2NSD). This contradicts the assumption that the bulky meta-trifluoro substituent on the A2 ring of pc-b7 should destabilise the loop more than the meta-cyano group of pc-b6.

Table 10.5 Averaged secondary structure propensities (in %) for the SBL (residues 195 to 217) of the simulated InhA protein-ligand complexes. The data were generated from 150 frames per protein-ligand complex corresponding to an overall sampling duration of 150 ns.

Motif	4TZK	4TZT	c6a3	c7a3	b3	b6	b7	b9	b12
parallel sheet	0	0	0	0	0	1	0	0	1
anti-parallel β sheet	0	3	1	0	0	2	1	1	1
3-10 helix	7	11	21	2	4	7	18	14	22
α-helix	53	22	34	62	50	30	37	46	25
π-helix	0	0	2	0	1	0	0	1	8
turn	24	28	21	20	15	18	15	17	26

Nevertheless, the low helical content in the case of pc-d8 and pc-b6 implies a binding similar in its characteristics to that of triclosan (TCL), i.e., rapid-reversible binding. The comparatively higher helical content of pc-d11 (PDB 4TZK), pc-c7a3, and pc-b9, suggests a similar ability to bring about loop ordering as that of 6PP. Furthermore, the high percentage of helical content for pc-c7a3 is in line with its activity in the high nanomolar range, much like 6PP. It is expected to be closely followed by pc-b9 and pc-b3 whose averaged helical content is in a similar range, and thereby these molecules are expected to exhibit a similar effect on the SBL.

10.3.4 Comparison of the conformational families with experimental structures

In order to further ascertain the significance of the above clustering results, and keeping in mind the two-step binding as well as the conformational changes between the EI and EI* states, a structural comparison with experimentally available structures is critical. The complexes most relevant as reference for this discussion involve the diphenyl ether series (PDB 2X23). In addition to PDB 2X23, the structural comparison with various ternary diaryl ether complexes with InhA-NAD adduct prove helpful in revealing conformational changes in key residues that happen along the transition from EI* to EI. The most recent ternary diaryl ether crystal structures have been solved by Li et al. [74] and Pan et al. [67], with a majority of them being slow-tight binders (PDB 4OXY - **PT10**, PDB 4OYR - **PT91**, PDB 4OHU - **PT92**, and 4OIM - **PT119**), along with a rapid reversible 4-hydroxy-2-pyridone deriative **PT155** (PDB 4OXN and 4OXK).

All of the slow-tight binding diaryl ethers differ from **PT70** (2X23 ligand) by substitution at the 2' position of the B ring as follows: 2'-nitro (PT10), 2'-chloro (P92), and 2'-cyano (PT119) as opposed to 2'-methyl group of PT70. Upon comparison of the conformational families from clustering of active site and extended active site with the aforementioned crystal structures, the following findings can be put forward:

1. All of the crystal structures of slow-tight binders exhibit similar conformations of key residues in the binding site in addition to a specific orientation of the SBL. This merely supports the hypotheses that the said conformation of these crystal structures corresponds to the EI* state.
2. PDB 4OIM (PT119) and PDB 4OXK/4OXN closely resemble the dominant binding conformation predicted by the hierarchical agglomerative clustering of active site and extended active site. PDB 4OIM displays a slight displacement of Ile202 away from the ligand. Consequently, Ile202 adopts the position of Val203, while Val203 is further displaced to the back (cf. Figure 10.13). This arrangement of Ile202 and Val203 is **exactly identical** to that of the Family 1 from both clusterings as well as PDB 4TZK. However, the SBL of 4TZK and in all conformational families are noticeably shifted with respect to the SBL of both 2X23 and 4OIM.
3. In the case of the rapid reversible inhibitor **PT155** (PDB 4OXK), the SBL conformation corresponds to a wide open state that is identical to PDB 4TZK and the dominant conformation from clustering of the active site and extended active site. However, in both cases, the SBL was more open as compared to the one in 4OXK. Additionally, the SBL of conformational family 3 from active site clustering comes closest to the SBL conformation in 4OXK and 4OIM.

The aforesaid observations lend credibility to the hypothesis that the conformation of the SBL as seen in 4TZK and the majority of the conformational families may correspond to the EI state. This can be seen from Figures 10.14 and 10.15, which depict the distances in between Phe97 (strand 4) and Ile202 (α6-helix) as well as Leu207 (hinge region) and Ile215 (α7-helix). These figures strongly support the hypothesis stated earlier in this chapter.

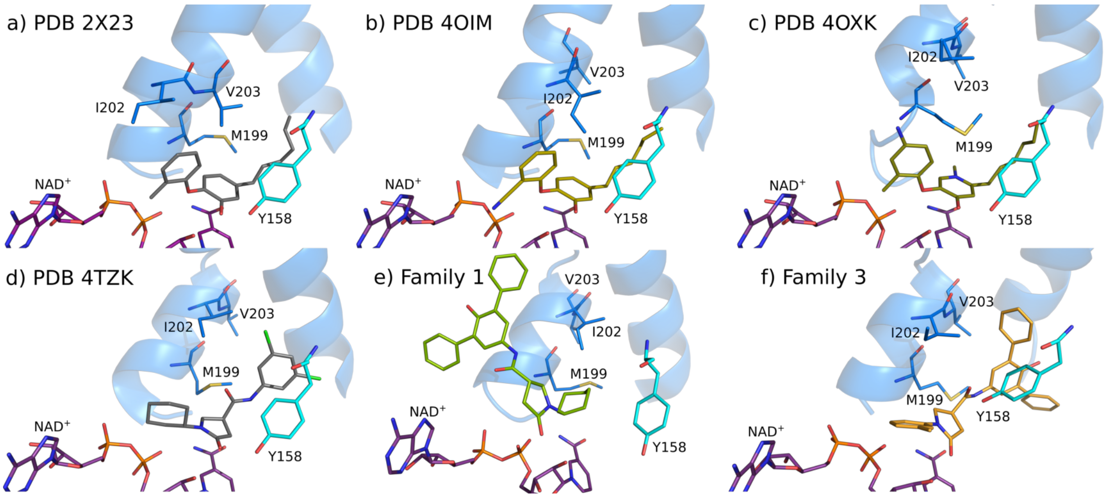


Figure 10.13 Illustration of dominant conformational families 1 and 3 of pyrrolidine carboxamides: The 6 clusters from hierarchical clustering analysis were subsumed to 3 conformational families. Subsequently, a Partitioning Around Medoids (PAM) was performed in R around each conformational family that yielded medoids or cluster representatives. The figures from the first row depict the crystal structures 2X23, 4OIM, and 4OXK, with the first two complexes having diphenyl ethers (slow-tight binders) PT70 and PT119 as bound ligands. The PDB 4OXK has PT155, a rapid-reversible 2-pyridone as a bound ligand. The lower row depicts PDB 4TZK and the medoids of conformational families 1 and 3. The family 1 is represented by 4TZK-c6a3 and corresponds to a conformation at t = 35 ns from a MD simulation of the respective protein-ligand complex. The family 3 is represented by 2NSD-b3 and corresponds to a conformation at t = 6 ns from a MD simulation of the respective protein-ligand complex. Family 1 shows a slight drift of Ile202 and Val203 as compared to PDB 4TZK, 4OXK, and family 3. On the contrary, this shift is huge in comparison with the conformation of the said residues in PDB 2X23 and 4OIM. In all figures, the SBL has been depicted as a marine coloured helix, the cofactor as purple sticks, the key residues as marine coloured sticks, while the bound ligands have been depicted as sticks coloured differently.

10.4 Determinants of rapid-reversible binding

From the clustering of the active and extended active site, the effects of the crystal structure ligands as well as of the newly designed ligands on the SBL become evident. A comparison of the backbone RMSD of the reference system (2X23 and 4TZK) to the simulated complexes in terms of SBL and the entire protein clearly revealed that maximal changes take place in the SBL upon ligand binding. Studying the reordering of the SBL in terms of the EI and EI* states and associating the conformational families with either of these two states provides a possibility of interpreting the long simulations. In the literature, it has already been known that the energy required for a rearrangement of Ile202 and Val203 contributes directly to the energy barrier that separates EI and EI* [74]. In contrast to the biased simulations of Li et al., the current simulations were entirely classical, i.e., without a biasing potential. However, all the ligands from pyrrolidine carboxamides are purportedly rapid-reversible inhibitors. Thus, the SBL conformations in all systems must correspond to the EI state. Comparing the conformational clustering

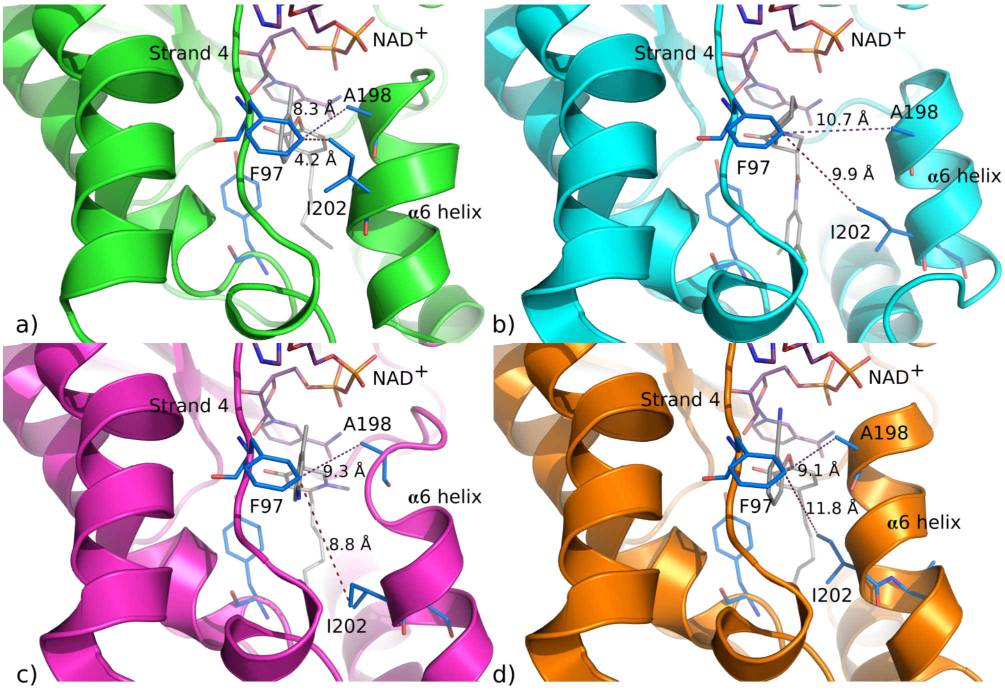


Figure 10.14 Distances in between Phe97 (atom CE1) and Ile202 (atom CD1) and Phe97 (atom CE1) and Ala198 (atom CB) for a) PDB 2X23, b) PDB 4TZK, c) PDB 4OXK, and d) PDB 4OIM.

of pyrrolidine carboxamides with that of 2X23 from the work of Merget et al. [261] thereby provides a way for ascertaining the conformational changes from EI* to EI.

The comparison of the clustering results for pyrrolidine carboxamides and diphenyl ethers (2X23, 6PP, and Triclosan) provides support to the hypothesis already stated earlier in this chapter. A major portion of the active site conformations from the simulated complexes corresponds to an open state of the SBL and thereby their apparent *rapid-reversible binding*. A major feature of all pyrrolidine carboxamides is their inability to prevent Ile202 and Val203 moving away from the binding pocket. This can be clearly seen from Figure 10.14. Firstly, the cyclohexyl ring of ligands pc-d11 (4TZK) and pc-d8 (4TZT) is unable to prevent Ile202 and Val203 turning away from the binding pocket. The case worsens in pc-c6a3 (inverted binding mode) which actually pushes Ile202 and Val203 away towards the α7-helix. In contrast, the pc-c7a3 with a cycloheptyl group (as ring C) does not push Ile202 and Val203 towards the α7-helix. This can be seen from Figure 10.12, where pc-c7a3 shows a very narrow distribution of SBL backbone RMSD. A similar observation holds true for the ligands pc-b9 and b12. Moreover, the pyrrolidine carboxamides with cyclo-hexyl/heptyl rings cannot fill the space between the B ring and the cofactor in an optimal manner. Additionally, the extreme motions of the C ring (cyclo-hexyl/heptyl) pushes the α6 helix towards the α7 helix.

Second, the loop ordering is quite variable in case of the simulated complexes. A comparison between Table 10.5 and results from clustering of diphenyl ether simulations

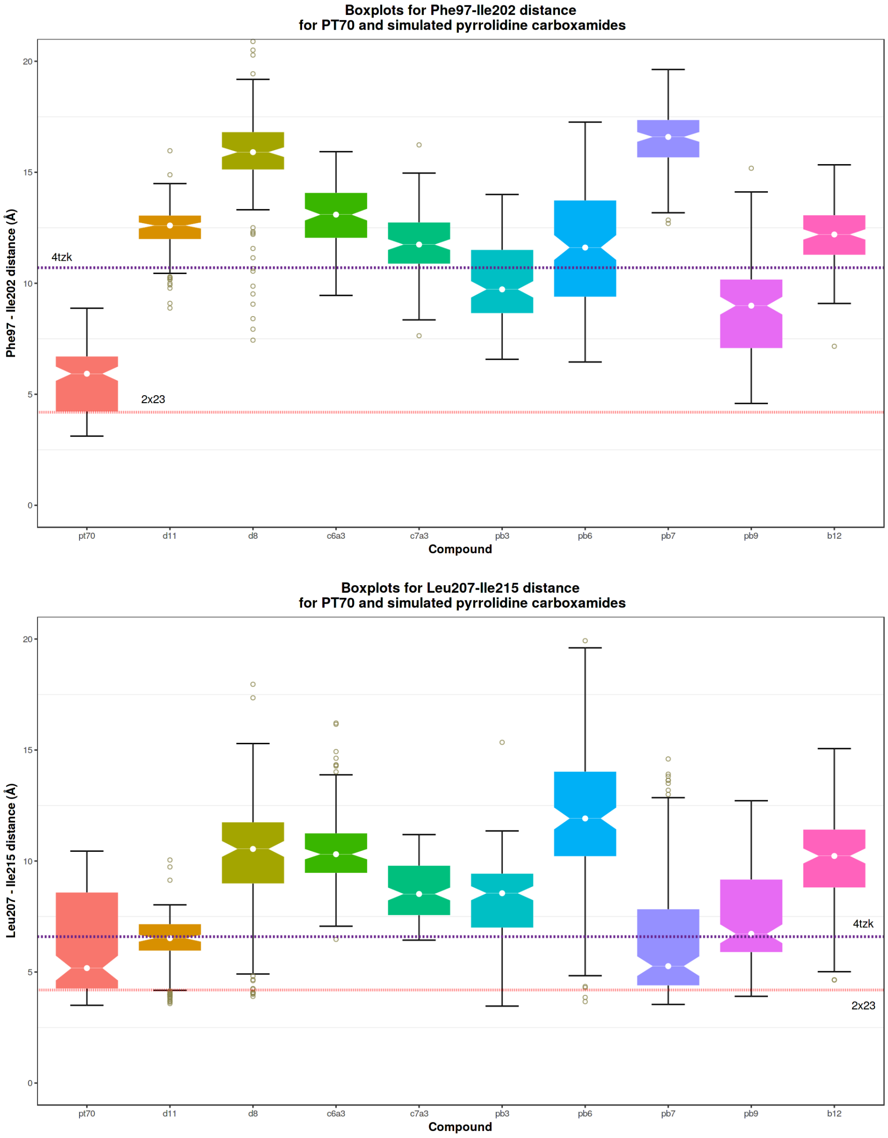


Figure 10.15 Distances in between (a) Phe97 (atom CE1) and Ile202 (atom CD1) (top) and (b) Leu207 and Ile215 (atom CD1) (bottom). Each monomer (simulated for 150 ns) was fitted individually onto PDB 2X23 chain A (reference structure) before exporting the raw distance data. The boxes represent the interquartile range for each monomer. The white circles denote the median value of the RMSD's, while outliers are shown as khaki coloured circles. In both plots, the horizontal lines indicate the distances between the said residues as observed from the crystal structures 4TZK (violet) and 2X23 (red).

reveals a low to moderate percentage of α-helical content in case of the simulated complexes as compared to ∼70% for PDB 2X23 (**PT70**). Of the simulated complexes, only pc-c7a3 and 4TZK come close, while the helical content of all other complexes suggests features of rapid reversible inhibitors. Furthermore, the low helical content for pc-c6a3 and 4TZT stresses their inability to bring about loop ordering. The causative reasons for the same can be traced to the artefactual binding mode (pc-c6a3) and inadequate filling of the hydrophobic binding pocket (pc-d8 of 4TZT), similar to triclosan (**TCL**, cf. Merget et al. [261]). The most surprising observation is made for the ligand pc-b6. Its structure-based design was guided with an aim to increase H-bonding with the backbone of the SBL residues near Val203 and prevent it from turning over towards the residues 210-218 (α7-helix). However, as seen from Table 10.5 and Figure 10.6, the rigidity of the planar cyano-group results in loop destabilisation as opposed to the intended purpose, i.e., loop ordering. A similar observation holds true for pc-b7. Compound pc-b12 also represents an interesting case with a quite low α helical content ($\sim$ 25%). It appears that pc-b12 slightly destabilises the SBL as opposed to pc-b6 and pc-b7, which is evident in the 3-10 helical content for these ligands.

Finally, proper occupation (volume filling) of the binding pocket is another determinant for binding of InhA inhibitors. The high values for the SBL as well as backbone RMSD's for the majority of the simulated complexes indicate a general inadequacy of the ligands in occupying the binding pocket. For example, pc-d8 (PDB 4TZT) has a well defined binding mode as seen from the crystal structure. However, the lack of stabilising interactions highlight its unstable binding. As a result, the ligand migrates out of the binding pocket after 150 ns. On the contrary, pc-c7a3 retains its binding mode even at the end of the 150 ns simulation. It binds like pc-d11 and consequently is able to establish stabilising interactions with the binding site residues. This can be partially attributed to the reasonable occupation of the binding pocket by pc-c7a3. Thus, a general lack of stabilising interactions also might play a dominant role in determining the type of binding for a particular ligand. Summing up, the following factors can be put forward as determinants characterising rapid-reversible binding:

1. Inability to prevent the shift of Ile202 and Val203 towards the α7-helix.
2. Non-optimal volume filling of the binding pocket and thereby lack of stabilising interactions.

10.4.1 Expected role of weak intermolecular interactions in binding stabilisation

The observations in the previous sections underscore the lack of substantial energetical barriers to the movement of Ile202 and Val203 towards the α7 helix. Insufficient barrier heights might be directly attributed to the absence of energy-lowering van der Waals interactions between the bound ligand and these residues. However, in addition to these two residues, there are several other residues that also play an important role in the stabilisation of the ligand binding. Given the hydrophobic nature of the InhA binding pocket, non-polar interactions are expected to play a crucial role in determining the type of binding ("slow-tight" vs. "rapid reversible"). This can clearly be seen by comparing the ligands of PDB 2X23 (**PT70**) and PDB 4TZK (**pc-d11**). The former exhibits a particularly stable binding which results from optimal interactions with Tyr158, Met198, Ile202, and Val203. This translates to a residence time (t_R) of 24 minutes for **PT70**. On the contrary, **pc-d11** and its related congeneric series cannot engage Ile202 and Val203 optimally. Consequently, the entire pyrrolidine carboxamide series can be seen to exhibit characteristics similar to rapid-reversible binders like 6PP and TCL.

In order to shed more light on the different types of interactions in between the bound ligand and InhA, a second type of analysis for protein-ligand interactions has been undertaken. The interaction analysis made use of 5 ns MD simulations of the same complexes used for the conformational clustering. Additionally, the 5 ns MD simulations of PT70, pc-p28 (alternate binding mode), pc-b12, and pc-b20 were also used. The sole purpose behind this addition was to probe the effect of ligand size and its placement within the InhA active site on the interactions. This analysis is based on distance and angle criteria in the flavour of structural interaction fingerprints (SiFt) [358] and pairwise atomic interactions according to CREDO [359, 360]. Over the course of 150 ns simulations, visual inspection revealed that the side chains of the following methionine residues interact weakly with pc-d11: Met103, Met155, and Met161. For the sake of completeness, our closer analysis should also contain Met198, since it has already been known to interact with the bound ligand. The interactions of the former three methionine residues, which we consider as especially important in the InhA active site, have been neglected in the literature so far and will be discussed here for the first time. A closer look at the strength of the sulphur-π interactions can reveal their importance in the ligand binding to InhA. Sherill et al. have shown by very reliable, highest-level quantum mechanical (QM) calculations (CCSD(T), augmented quadruple zeta basis set), that interaction energies of -2.5 kcal/mol are associated with sulphur-arene interactions in the gas phase [361, 362]. Diederich and colleagues still expect a net stabilisation energy of -0.5 to -1.0 kcal/mol in biological systems after accounting for the desolvation penalties [349, 361, 362]. In addition to this, the work of Beno et al. investigated the angle dependence of sulphur-π interactions: The S atom interacting with the face of an aromatic ring and its π-cloud is more favourable than one with the edge of the aromatic ring. [363].

The simulated InhA complexes have not only been assessed for sulphur-arene interactions involving the various methionine residues but also for halogen bonds, CH-π, and OH or lone pair-π interactions using distance and angle criteria in accordance with the currently state of knowledge in the literature [358–360, 364]. Of course, traditional force fields do not have special terms to describe the aforementioned interactions. Moreover, the force fields capture the strength of such interactions only partly, if at all, and are subsumed in the van der Waals or electrostatic contributions. However, the conformational sampling in unbiased MD simulations yields conformations that might be close to the physical reality in the majority of times and therefore such interactions can be expected to be present if certain angle and distance criteria are met.

Table 10.6 summarises the key interactions found in our analysis between the bound ligand and residues of the active site along 5 ns simulations of several pyrrolidine carboxamides and the diphenyl ether **PT70** (PDB 2X23). On purpose, quite strict geometrical criteria were chosen in order to obtain reliable and significant indications for the possibility of the presence or absence of these interactions and to get a clear impression of their abundance with respect to time. For example, in the case of the sulphur-π interaction, the often chosen cutoff distance of 6 Å in between the sulphur atom of the methionine and the centroid of the aromatic ring was reduced to 4.25 Å. This ensured that the sulphur atom was always in the proximity of the aromatic ring alongwith an angle that was not too far from perpendicular (90°). The same distance onset criterion was applied in analysing the CH and OH $\cdots$ π interactions. The analysis was performed utilising Arpeggio [364] and customised python code written by Dr. Thomas B. Adler, University of Würzburg.

From the inspection of Table 10.6, several observations come to the fore:

1. Sulphur-π interactions (S-π): Represents the interaction of methionine sulphur atom with any aromatic ring of the bound ligand. It was seen that in addition to **PT70** and pyrrolidine carboxamides taken from literature, the methionine S atoms were not found very often within the cutoff distance, and hence, sulphur-π interactions might contribute only weakly to the stabilisation of literature pyrrolidine carboxamides. These interactions, however, might play a noticeable role in stabilising the binding of the pyrrolidine carboxamides pc-p28, pc-b4, pc-b7, and pc-b20 to InhA. In the case of the former two compounds, i.e., pc-p28 and pc-b4, it is the sizeable occurrence probability of Met155 within the sulphur-π distance cutoff that lets expect the possibility of a weak to moderate interaction. With pc-b7 and pc-b20, it is the methionine 161 that generally shows interactions at a convincing, moderate occurrence probability. Both methionine residues are situated in the immediate proximity of Tyr158 and are expected to play a role in restricting its phenyl ring rotation so that the conserved H-bonding with the ligand can reliably be established. In other words, both Met155 and Met161 can be seen to play an important role in facilitating the binding of these compounds to InhA. As compared to **PT70** and **pc-d11**, the increase in the occurrence probability of S-π geometries is significant for pc-b7 and pc-b20. On the contrary, pc-d8 and pc-c6a3 do not often satisfy the criteria for S-π geometries with any of the surrounding methionine residues. This can partially explain the lack of stable binding for pc-d8. Furthermore, the lack to meet the S-π criteria for pc-c6a3 merely stems from its predicted binding mode, i.e., an inverted one, which lends further support to disprove of its viability.

2. Possible Met-S-CH_3-π interactions present another interesting scenario in the case of Met155 and Met161. It has been found that PT70 comes within an interacting distance mainly with Met161: over 86% of the 5 ns MD simulation. In marked contrast, all published pyrrolidine carboxamides are not found to exhibit Met-S-CH_3-π interactions with Met161, while only pc-c7a3 is expected to weakly interact with Met155. A similar behaviour can be seen for the pyrrolidine carboxamides pc-p28, pc-b6, pc-b7, and pc-b12. A notable exception to this observation are the compounds pc-b3, pc-b4, and pc-b20, of which pc-b4 is expected to be able to interact strongly with both Met155 and Met161. The compounds pc-b3 and pc-b20 interacted rather moderately with aforementioned methionine residues.

3. Almost all of the analysed complexes exhibit distances which only allow weak or no interactions at all with Met198. The exception to this are the compounds pc-p28 and pc-b20 which can be found exhibiting distances that allow moderate interactions.

4. For a sizeable number of ligands (N=4: i.e., pc-d11, pc-b3, pc-b6, and pc-b9), CH moieties of the ligands might interact with any aromatic ring in their vicinity. Such CH-π interactions might play a role not to be underestimated at all in stabilising the binding [349]. A very high occurrence probability of favourable CH-π interaction distances has been found (cf. Table 10.6). Moreover, pc-b12 (C ring) is the only ligand which might be able to additionally establish a moderate OH and lone pair-π interaction occurrence involving the ribose hydroxyl group. This kind of interaction might also be observed for pc-b20 (Figure 9.11), though it is much more subdued as compared to pc-b12 or even pc-b7.

The above findings reveal considerable differences in the geometrical arrangements in pyrrolidine carboxamides versus diphenyl ethers which would allow for specific interactions with the two methionine residues (Met155 and Met161) that flank the catalytic Tyr158. While PT70 might establish **exclusive** interactions with Met161 via MetS-$CH_3 \cdots \pi$ interaction, the same clearly lacking in case of pyrrolidine carboxamides from the literature. The preference of PT70 in interacting with Met161 alone over Met155 becomes clear by studying Figure 10.16: since a Met-S $\cdots$ π interaction must involve a considerable change of the CB-CG-S-CH_3 torsion in order to point the sulphur atom towards the π-cloud of the aromatic ring of the ligand. In contrast, the MetS-$CH_3 \cdots \pi$ interactions can readily be established without larger conformational changes (e.g., rotation of an entire group). Having three hydrogens (from S-CH_3) at free disposal together with their fast rotation allows for a high occurrence probability of the MetS-$CH_3 \cdots \pi$ interactions. Overall, this large number of admittedly minor energetic contributions with respect to the individual CH-π interactions might finally render the MetS-$CH_3 \cdots \pi$ interactions an important player in ligand binding to InhA. This merely underscores the importance of engaging Met155 and Met161 as additional stabilising factors during the binding process to InhA.

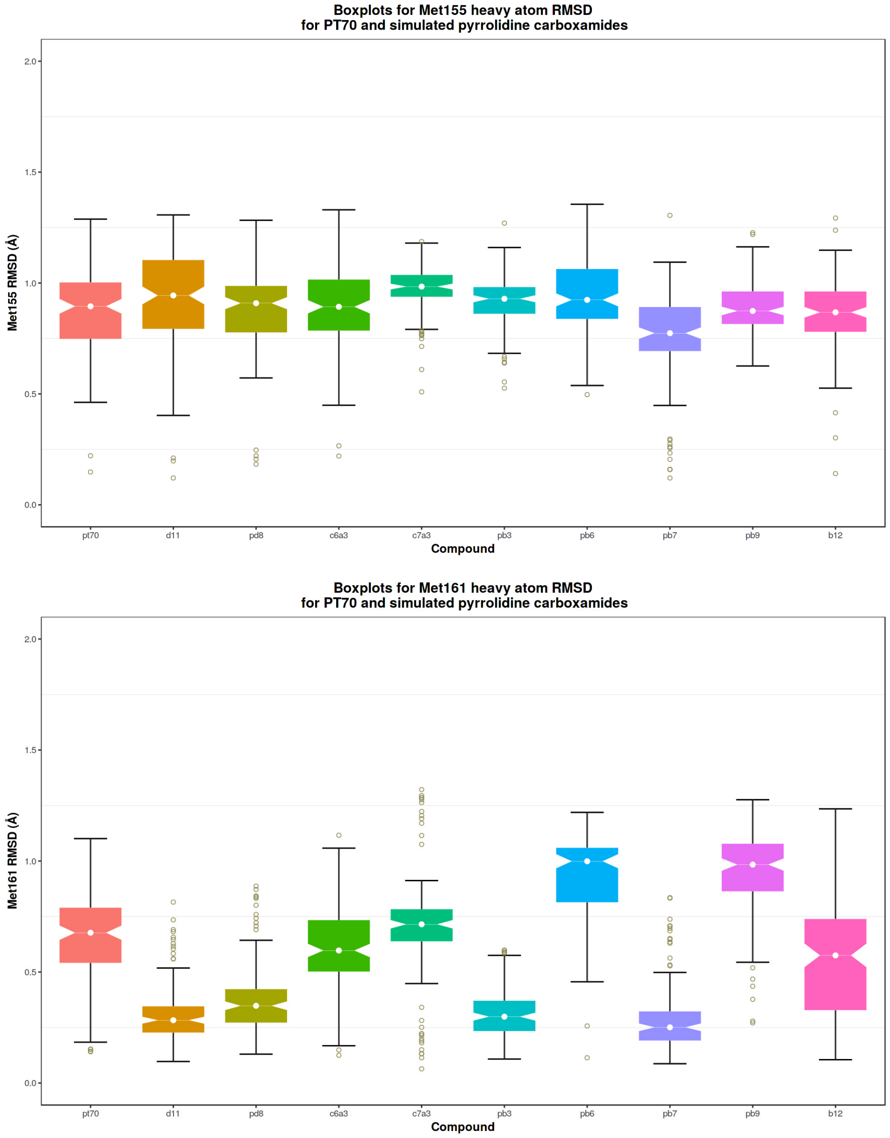


Figure 10.16 Distribution of heavy atom (all except hydrogens) RMSD for (a) methionine 155, and (b) methionine 161 of the simulated protein ligand complexes. Each monomer has been simulated for 150 ns and fitted individually onto PDB 2X23 chain A (reference structure). The boxes signify the interquartile range for each monomer. The white circles denote the median value of the RMSD's, while outliers are shown as khaki coloured circles.

Table 10.6 Summary of sulphur-π and σ - π interactions: Occurrence frequency (in %) of geometrical arrangements along a 5 ns MD trajectory indicative of various sulphur arene, MetS-CH$_3$ $\cdots$ π, CH-π, and OH-π interactions between ligand and the cofactor, several methionines or any aromatic systems. These types of interactions must be active in reality if one assumes that the macromolecular system visits conformational states that are closely related to those from MD simulation.

Interaction type	**PT70**	**D11**	**PD8**	**C63**	**C73**	**P28**	**PB3**	**PB4**	**PB6**	**PB7**	**PB9**	**B12**	**B20**
							Sulphur - π interactions: methionines-ligand						
Met103-S $\cdots$ π	0.30	0	0.02	0	0.06	0.02	0.06	0.30	0.02	0.06	0	0.30	54.62
Met155-S $\cdots$ π	0.04	0	0	0	4.30	14.70	1.70	26.70	1.88	0.32	0.26	0.04	2.04
Met161-S $\cdots$ π	6.90	0	0	0	0	0	0.14	1.54	0	26.72	0	0.08	29.42
Met198-S $\cdots$ π	0	0.14	0.16	0	0.32	10.20	0.38	0.48	0	20.08	0.04	0.16	23.08
All sulphur-π	7.24	0.14	0.18	0	4.68	24.92	2.28	29.02	1.90	47.18	0.30	0.58	109.16*
							Met-S-CH$_3$ - π interactions: methionines-ligand						
Met103-S-CH$_3$ $\cdots$ π	2.70	0	0.04	2.20	0	1.20	0	0.56	0.08	18.80	0	0.20	0
Met155-S-CH$_3$ $\cdots$ π	0	0	0	0	19.00	0.92	40.58	66.08	1.28	0	0	0.18	1.40
Met161-S-CH$_3$ $\cdots$ π	83.70	0	0	0	0	0	16.80	56.38	1.10	11.52	0	1.70	30.68
Met198-S-CH$_3$ $\cdots$ π	0	3.72	0.28	0	0.06	38.62	3.96	1.44	0	10.40	1.30	0	40.36
All Met-S-CH$_3$-π	86.40	3.72	0.32	2.20	19.06	40.74	61.34	124.46	2.46	40.72	1.30	2.08	72.44
							CH and OH-π interactions						
Ligand-CH - any-π-systems[‡]	-	55.44	0.62	0	13.66	0	91.98	18.92	97.40	10.30	80.92	24.46	21.30
NAD-OH - ligand-π-systems[∓] *	0.02	0	0	0	0	0	0	10.84	0	17.80	0	53.62	6.08

[‡] CH is a part of the methylene group adjacent to the pyrrolidine nitrogen of the B ring, while NAD$^+$ corresponds to the aromatic ring of nicotinamide.

[∓] NAD-OH corresponds to the hydroxyl group of the ribose ring of the cofactor. Ligand π system corresponds to any aromatic ring of the bound ligand.

[*] Percentages higher than 100% can occur because several S-π interactions are possible in a single frame.

The conservatively chosen onset criterion for the different interaction types is: sulphur, CH, or OH-π: 4.25 Å (ring centroid to S, C, O)

From the above, it might be deduced that almost solely Met-S-$CH_3 \cdots \pi$ interactions are decisive in the case of PT70 and by extension diphenyl ethers. In the case of pyrrolidine carboxamides, however, especially CH $\cdots$ π interactions arising from the methylene group adjacent to the pyrrolidine nitrogen of the B ring are proposed to play a role that clearly distinguishes them from the diphenyl ethers and yields in sum considerable contributions to ligand binding. With respect to the Met-sulphur-π interactions, pc-b4, pc-b7, pc-p28 and pc-b20 might be markedly involved in this (geometrically/sterically demanding) type of interaction. Especially, pc-b20 fares noticeably better than the other compounds, implying its ability to engage all of the methionine residues around itself. The noticeable energetic contribution of Met-S-π interactions is expected to aid in the stabilisation of their binding modes.

Summing up, increased CH $\cdots$ π interactions in addition to engaging the methionine residues flanking Tyr158, especially Met161 can be considered as an additional factor that distinguishes the binding of diphenyl ethers from pyrrolidine carboxamides to InhA.

10.4.1.1 Implications for structure-based optimisation in relation to determinants of rapid reversible binding

The conformational families of pyrrolidine carboxamides when compared to those of diphenyl ethers revealed the changes necessary in going from EI to EI* (or vice versa). Apparently, the pyrrolidine carboxamides are unable to prevent Ile202 and Val203 turning away from the ligand. In order to prevent this, the local interactions with these residues warrant improvement. In order to better accomplish this, a comparison of the best pyrrolidine carboxamide from the interaction analysis (pc-b20) with PDB 2X23 proves informative. A C-α atom alignment of PDB 2X23 and the protein-ligand complex revealed the near-optimal placement of the scaffold as well as of the C ring (Figure 10.17). It can be clearly seen that the meta standing halogen group of pc-b20 overlaps with the ortho methyl group of PT70. Furthermore, Merget et al. have suggested placing a methyl group para to the ortho methyl group as a barrier to restrict the motions of Val203 [261].

In the context of this work, considering the limited pocket space available for further exploitation, an additional substitution of the C ring at the other **meta** (5') position is expected to be able to act as a barricade for Val203. Suitable groups at this position might be a methyl or a cyano group. The latter group just like the proposed methyl group dissects the space between Ile202 and Val203, thereby acting as a steric barrier. In addition to this, it further lowers the logP of the resultant compound (Figure 10.18). Considering the hydrophobic space in the immediate vicinity of ring C, the methyl group is expected to fare better than the cyano group [365].

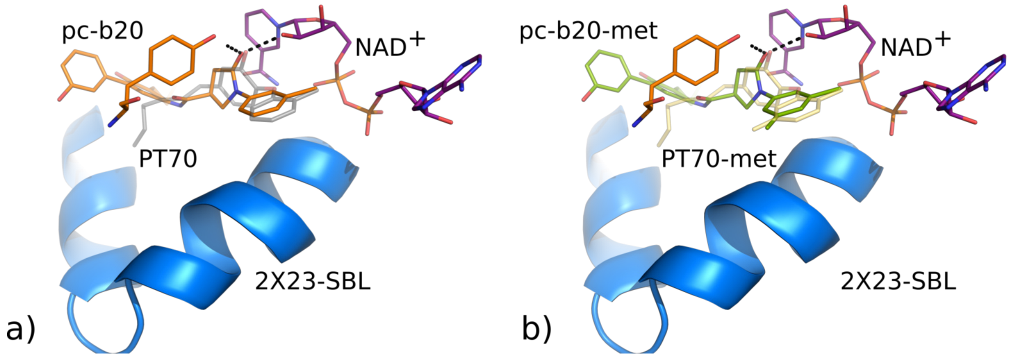
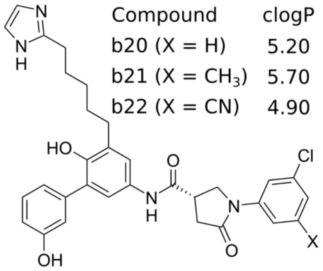


Figure 10.17 **Aligned structures of:** a) pc-b20 (orange sticks) and PT70 (transparent grey sticks), and b) proposed methyl substitution on ring B of PT70 (yellow sticks) by Merget et al. [261], and pc-b20 (green sticks). The H-bonds are depicted as black dashes, while the SBL has been depicted as a blue cartoon.



Compound	clogP
b20 (X = H)	5.20
b21 (X = CH_3)	5.70
b22 (X = CN)	4.90

Figure 10.18 Proposed methyl and cyano group substitutions of pc-b20 as inferred from the clustering and subsequent comparison of conformations representing EI (4TZK) and EI* (2X23). logP values have been calculated using Moka [327].

10.5 Conclusion

Long MD simulations combined with a cluster analysis enabled the unveiling of the dominant conformational state of pyrrolidine carboxamides bound to InhA. Comparing the conformational families of pyrrolidine carboxamides with diphenyl ethers (representing the EI* state) shows that the pyrrolidine carboxamide conformational families correspond to the EI state. The associated changes in the EI* and EI macrostates can be visualised by comparing the MD simulations of PT70 and pc-d11. They represent the reference ligands from diphenyl ether and pyrrolidine carboxamide compound families. Analysis of the conformations, secondary structure propensities, inter-helical distances and backbone RMSD underscores the differences between EI* and EI macrostates, alongwith the determinants that characterise the binding of diphenyl ethers and pyrrolidine carboxamides: firstly and most critical, proper occupation of the binding site. The second determinant pertains to engaging Ile202 and Val203 in order to prevent them from turning further away towards the $\alpha 7$ helix. This can be achieved by introducing an ortho methyl or halogen substituent on the C ring in order to engage the cofactor. Furthermore, meta-meta disubstituted C ring pyrrolidine carboxamides might also be worthwhile to be tested, since the second meta substituent (e.g., methyl group) can engage Ile202 and Val203, thereby averting them from moving away from the binding pocket. Notably, a second ortho/meta substituent has size limitations primarily due to the presence of the cofactor on one end and Met103 on the other which has to be taken into account. The final determinant involves the methionine residues flanking Tyr158. While PT70 was found to engage Met161, pyrrolidine carboxamides apparently did not and were mostly found to exhibit CH-π interactions with aromatic residues around them. This determinant however is closely linked to the first one, i.e., proper placement and occupation of the InhA binding site. A simultaneous satisfaction of the first determinant would result in the ligand satisfying the distance criteria necessary to form stabilsing interactions with the active site residues of InhA. Accruing together, all of these factors reveal the differences crucial for ”slow-tight” versus ”rapid-reversible” inhibitors as well as a strategy for further optimisation of pyrrolidine carboxamides.

Chapter 11

Summary - Part II

The molecular docking of pyrrolidine carboxamides yielded three distinct binding modes, especially for bulkier members of the compound series. The dihedral angle analysis together with RMSD and H-bond evaluation suggested that the crystal structure orientation represents the dominant binding conformation for pyrrolidine carboxamides. The dihedral angle analysis clearly showed that the in place motions of the bound ligand were mostly phenyl ring flips that contributed maximally to the bound ligand RMSD. In regards of the binding stability, the crystal structure conformation was comparatively more stable than the alternate binding modes. The pyrrolidine carboxamides were generally observed to exhibit H-bonding that was weak and transient in nature. This weak H-bonding trend was found to coincide with the potency, i.e., potent compounds binding in a crystal structure like conformation exhibited stronger H-bonding (and hence stable binding) as compared to those with lower potency. Thus, the dihedral angle analysis together with the H-bonding analysis shed light on the binding of pyrrolidine carboxamides.

In order to further gain information about the binding and in place motions of pyrrolidine carboxamides, essential dynamics and dynamic cross correlation analyses were carried out for the atom selections representing the bound ligands active site and the substrate binding loop. Essential dynamics was quite helpful in ascertaining the maximal variance (movement) and its direction while dynamic cross correlation revealed whether there was any sort of correlation amongst the movements of the atoms. The essential dynamics for the bound ligands suggested that in addition to the phenyl ring flips, the ring C and the 2° carbonyl group contributed maximally to the overall in-place motions of the bound ligand. The bulky pyrrolidine carboxamides in particular were found to exhibit a motion that resembled a pincer open and close cycle. In regards of the active site and the substrate binding loop residues, the light and bulky pyrrolidine carboxamides markedly differed in their ability to bring about motions of key residues involved in ligand binding. While the light pyrrolidine carboxamides mainly affected Met155 and Met199, the bulky pyrrolidine carboxamides engaged Ile202 and Val203. This might explain the noticeable difference in the overall potencies for the two classes. A comparison of the dynamic cross correlation maps for the representative ligands pc-d11 and pc-c7a3 as well as the residues of active site and substrate binding loop revealed that light pyrrolidine carboxamides were able to induce modest to strongly correlated motions of residues situated in the minor (near Met155 and Tyr158) as well as major exit portal (near Ile202). On the contrary, bulky pyrrolidine carboxamides only affected the residues of the major exit portal. These analyses also underscored the motions of two leucine residues (Leu197 and Leu207) which exhibited conserved motions with opposite direction. The collective motions of these two residues depicted the direction of motion of the substrate binding loop.

The structural information garnered from Chapter 8 was used to drive the structure-based optimisation of the bulky pyrrolidine carboxamide scaffold, with a main aim of maximising interactions and stabilising the binding mode. A total of 20 molecules were designed using a iterative procedure that consisted of molecular docking, rescoring, MD simulations, activity classification and mycobacterial cell wall permeability prediction. On an overall basis, all compounds exhibited modest improvements over their parent compounds as well as the reference ligand pc-d11. Nevertheless, the extensive *in-silico* analyses deemed 6 compounds with optimal interactions and clogP values worthy of further testing and evaluation.

The last chapter was clearly focussed on revealing the molecular determinants that drive the binding of pyrrolidine carboxamides to InhA. For this purpose long MD simulations (150 ns) of 9 selected protein-ligand complexes were extensively used. These complexes ensured adequate coverage of the light, bulky, and the designed pyrrolidine carboxamides. A hierarchical agglomerative clustering of the active site and extended active site residues clearly suggested that the dominant binding conformation for pyrrolidine carboxamides corresponded to a wide open state of the α6-helix and thereby the EI state. The subsequent comparison of the clusterings of pyrrolidine carboxamides(representing EI) and diphenyl ethers (representing EI*) underscored the importance of engaging Ile202 and Val203 in order to prolong the ligand-InhA interaction time. The following were shown to be important determinants driving the ligand binding; a) proper placement and occupation (volume filling) of the binding pocket; b) ability of the ligand to engage Ile202 and Val203 and prevent them from turning away from the bound ligand. In pursuit of engaging Ile202 and Val203, an introduction of a small substituent at 5' position of the C ring was proposed that would act as a steric barrier to prevent Val203 from turning over. As a result, two more ligands with methyl and cyano groups were designed. Additionally, the role of weak interactions in stabilising the ligand binding was probed. Using distance-based criteria laid down in literature, the MD simulations of 2X23-PT70 complex and selected pyrrolidine carboxamides were thoroughly analysed. The interaction analyses suggested the decisive role of a pair of methionine residues flanking Tyr158 in ligand binding. While PT70 exclusively engaged Met161 (Met-S-$CH_3 \cdots \pi$), a majority of pyrrolidine carboxamides were found to be lacking in the same. Moreover, the CH $\cdots$ π interactions arising from the methylene group adjacent to the pyrrolidine nitrogen of the B ring are proposed to play a role in their binding. Thus, increased CH $\cdots$ π interactions along with engaging Met161 can be put forward as an additional factor that distinguishes the binding of diphenyl ethers from pyrrolidine carboxamides to InhA.

In conclusion, MD simulations and associated techniques were used to achieve the aim of *in-silico* structure-based optimisation of pyrrolidine carboxamides. Additionally, they were also used to reveal the dominant binding conformation as well as determinants of the supposedly *rapid reversible* binding of pyrrolidine carboxamides. The designed pyrrolidine carboxamides arising from this work are expected to engage InhA in a more prominent fashion as compared to their parent pyrrolidine carboxamides.

Summary

Prediction of the binding mode and affinity constitute important stages in a structure-based drug design endeavour. The rational optimisation of compounds encompasses these steps after due consideration of the target receptor structure. The mycobacterial trans-enoyl-ACP-reductase FabI, or InhA, is an important and well validated target from the mycobacterial fatty acid synthesis II pathway. InhA has been shown to be inhibited by numerous compounds with varying structures. The slow-onset inhibition which involves the process of induced fit is a key phenomenon that distinguishes potent, long duration inhibitors from moderately potent and short acting ones. The ordering of a crucial and flexible region of InhA, the substrate binding loop, has been shown to be directly tied to the ability of the inhibitor to fall in either of the aforementioned class. Pyrrolidine carboxamides represent a compound series with moderate InhA inhibitory potential. And although the binding modes for the smaller members of this series to InhA are well known, the effect of their binding on the substrate binding loop is currently unknown. Moreover, the molecular determinants behind the apparent rapid reversible binding of pyrrolidine carboxamides is yet to be revealed. In such a scenario, molecules with well defined binding modes and an ability to bring about the ordering of the substrate binding loop are desirable.

With an aim of structure-based optimisation of pyrrolidine carboxamides, the binding modes for the entire pyrrolidine carboxamide series were predicted using molecular docking employing induced fit. The ensuing poses formed an input for molecular dynamics (MD) simulations. Using a small dataset of 23 compounds deemed stable through the MD simulations, affinity prediction models were generated using the Linear Interaction Energy method. Using docking and rescoring values for the same 23 compounds, an activity-based classification model employing logistic regression was generated. While the affinity prediction models were unable to achieve statistical significance, the activity-based classification models performed satisfactorily in correctly distinguishing least active compounds from the moderately and highly active pyrrolidine carboxamides. Additionally, the activity-based classification model could be used with ease as an additional filter in the preliminary stages of a virtual screening endeavour, primarily to ascertain the molecules worth pursuing further.

The binding mode prediction yielded three distinct binding modes for the bulky members of the pyrrolidine carboxamide series. The MD simulations of the entire series were subsequently subjected to extensive analysis to reveal the dominant binding mode as well as the molecular determinants for the rapid reversible binding. Summing up, molecular docking together with MD simulations provided for reasonable starting conformations for

the ligand-InhA complex that could be used in structure-based optimisation of pyrrolidine carboxamides.

The MD simulations of the pyrrolidine carboxamide series were extensively analysed for the perturbations in the bound ligand as well as the protein. An attempt to correlate the quality and strength of the H-bonds with the measured InhA inhibitory potential was also made. The dihedral angle analysis for the bound ligand clearly revealed the type of in-place motions of the bound ligand. It also aided in shedding light on the most mobile portions of the ligand that ultimately contribute to its RMSD. Together, the RMSD and the H-bond analysis suggested that bulky pyrrolidine carboxamides could bind in an orientation quite identical to that of the crystal structure ligands. These analyses also revealed that the rings A and C of the bound ligands contributed maximally to their in-place motions. Moreover, pyrrolidine carboxamides were found to exhibit weak and transient H-bonding with Tyr158 and the cofactor that might underscore their moderate InhA inhibitory potential.

A further analysis was entirely focussed on revealing the maximal variance and direction of movements of the bound ligand as well as key residues involved in ligand binding. Additionally, the correlated/anti-correlated motions of the key residues and bound ligand were analysed. The essential dynamics of the bound ligand merely corroborated the findings from the dihedral angle analysis. They also revealed the fundamental differences in the changes that occur upon binding of light and bulky pyrrolidine carboxamides. While the smaller pyrrolidine carboxamides were found to affect residues of both minor and the major exit portal, the bulkier members only affected the key residues of the substrate binding loop, i.e., Ile202 and Val203. Supporting the findings from the essential dynamics were the dynamic cross correlation maps for the representative ligands pc-d11 and pc-c7a3. Collectively, these analyses aided in revealing the modest movements of the substrate binding loop towards the ligand.

The information garnered from the essential dynamics and dynamic cross correlation analyses was used to drive the structure-based optimisation of the bulky pyrrolidine carboxamide scaffold. A total of 20 compounds were designed and subjected to extensive *in-silico* analysis that included molecular docking, rescoring, MD simulations and even the mycobacterial cell wall permeability prediction. All of these analyses revealed modest improvements over their parent molecules (pc-c6a3 and pc-c7a3) as well as the reference molecule (pc-d11). For a promising subset of new molecules, extensive MD simulations equalling 150 ns per protein-ligand complexes were performed to ascertain the changes in their binding mode as well as the protein at extended simulation duration. Almost all of the molecules, barring a few, were found to exhibit improved binding stability, although the overall ligand RMSD was comparatively higher than that of the reference or the parent ligands. These simulations were also used to shed light on the molecular determinants behind the rapid reversible binding of pyrrolidine carboxamides.

Finally, extensive use was made of clustering techniques to ascertain the dominant binding conformation for pyrrolidine carboxamides and to determine the causative factors behind the weak binding of pyrrolidine carboxamides. The clustering and subsequent comparison of the dominant conformations with that from slow-tight binders revealed that pyrrolidine carboxamides prominently feature a wide open state of substrate binding loop. A smaller conformational family was associated with a less open form of the substrate binding loop. The clustering also aided in underscoring the structural determinants of their apparent rapid reversible binding. This information was subsequently used to suggest two more pyrrolidine carboxamide inhibitors that exhibited favourable characteristics of slow-tight binders. A comparison with the findings from literature suggests the possibility of improved binding and InhA inhibitory potential for the proposed molecules as compared to the reference ligand. In conclusion, the promising molecules from the designed pyrrolidine carboxamides exhibit favourable characteristics and are worthy of further experimental investigation.

Zusammenfassung

Die Vorhersage von Bindemodus und Affinität stellen wichtige Schritte im strukturbasierten Wirkstoffdesign dar. Die rationale Optimierung von Verbindungen umfasst diese Schritte unter Berücksichtigung der Struktur des Targets. Die mykobakterielle Trans-Enoyl-ACP-Reduktase FabI oder InhA ist ein wichtiges und gut validiertes Zielprotein im mykobakteriellen Fettsäuresyntheseweg II. Es wurde gezeigt, dass InhA von einer Vielzahl von Molekülen mit unterschiedlichster Struktur inhibiert werden kann. Das langsame Einsetzen der Inhibition, verbunden mit einem Induced Fit Vorgang, ist ein Schlüsselphänomen, welches starke und lang-wirksame Inhibitoren von mäßig starken und kurz-wirksamen unterscheidet. Ebenso wurde gezeigt, dass die Anordnung einer entscheidenden flexiblen Region von InhA, dem Substratbindeloop, direkt mit Zugehörigkeit eines Inhibitors zu einer der erwähnten Klassen verbunden ist. Pyrrolidincarboxamide stellen eine Reihe von Molekülen mit mäßigem Inhibitionspotential dar. Obwohl der Bindemodus der kleineren Mitglieder dieser Reihe weitgehend bekannt ist, ist deren Einfluss auf den Substratbindeloop noch kaum untersucht. Außerdem sind die molekularen Determinanten hinter der anscheinend schnellen und reversiblen Bindung der Pyrrolidincarboxamide noch aufzuklären. In diesem Szenario sind Moleküle mit genau bestimmten Bindemodi und der Fähigkeit, eine bestimmte Anordnung des Substratbindeloops zu induzieren, wünschenswert.

Um das Ziel einer strukturbasierten Optimierung der Pyrrolidincarboxamide zu erreichen, wurden die Bindemodi für die gesamte Reihe der Pyrrolidincarboxamide mit molekularem Docking unter Verwendung von Induced-Fit-Verfahren vorhergesagt. Die sich daraus ergebenden Posen bildeten die Grundlage für Molekulardynamik (MD) Simulationen. Unter Verwendung der *Linear Interaction Energy* Methode wurden auf der Grundlage eines kleinen Datensatzes von 23 Verbindungen, die während der gesamten MD Simulation stabil waren, Modelle zur Affinitätsvorhersage erstellt. Die Docking- und Rescoring-Ergebnisse für diese 23 Verbindungen wurde zur Entwicklung eines Klassifizierungsmodells mit Hilfe von logistischer Regression genutzt. Während die Modelle zur Affinitätsvorhersage keine statistische Signifikanz erzielten, erreichte das aktivitätsbasierte Klassifizierungsmodell zufriedenstellende Ergebnisse hinsichtlich der korrekten Unterscheidung von kaum aktiven Verbindungen von mäßig und stark aktiven Pyrrolidincarboxamiden. Darüber hinaus könnte das aktivitätsbasierte Klassifizierungsmodell in einfacher Weise als zusätzlicher Filter in den Vorstufen eines virtuellen Screenings verwendet werden, um die vielversprechensten Moleküle zu ermitteln.

Die Vorhersage des Bindemodus lieferte drei verschiedene Bindungsmodi für die räumlich anspruchsvollen Vertreter der Pyrrolidincarboxamide. Die MD-Simulationen der gesamten

Reihe wurden anschließend einer umfangreichen Analyse unterzogen, um den dominanten Bindungsmodus sowie die molekularen Determinanten für die schnelle reversible Bindung zu ermitteln. Zusammenfassend lieferte das molekulare Docking in Kombination mit MD-Simulationen plausible Startkonformationen für die Ligand-InhA-Komplexe, die zur strukturbasierten Optimierung von Pyrrolidincarboxamiden verwendet werden konnten.

Die MD-Simulationen der Pyrrolidincarboxamid-Serie wurden eingehend auf die Bewegungen im gebundenen Liganden sowie dem Protein analysiert. Ein Versuch, die Qualität und Stärke der Wasserstoffbrückenbindungen mit der gemessenen InhA-Inhibition zu korrelieren, wurde ebenfalls unternommen. Die Diederwinkelanalyse für den gebundenen Liganden zeigte deutlich die Art der direkten Bewegungen des gebundenen Liganden. Sie half auch, die mobilsten Teile des Liganden zu identifizieren, die letztlich zum RMSD beitragen. Die RMSD- und die Wasserstoffbrücken-Analyse legen beide nahe, dass sperrige Pyrrolidincarboxamide in einer Orientierung binden können, die jener der Kristallstrukturliganden entspricht. Diese Analysen zeigten auch, dass die Ringe A und C der gebundenen Liganden am meisten zu den direkten Bewegungen beitragen. Darüber hinaus wurde festgestellt, dass Pyrrolidincarboxamide eine schwache und kurzlebige Wasserstoffbrückenbindung mit Tyr158 und dem Cofaktor ausbilden, was mitverantwortlich für ihr ingesamt moderates InhA-Hemmpotential sein dürfte.

Eine weitere Analyse konzentrierte sich auf die Bestimmung der maximalen Varianz und Richtung der Bewegungen des gebundenen Liganden sowie der wichtigsten Aminosäuren, die an der Bindung des Liganden beteiligt sind. Zusätzlich wurden die korrelierten und antikorrelierten Bewegungen dieser Aminosäuren und des gebundenen Liganden untersucht. Die *Essential Dynamics*-Analyse des gebundenen Liganden bestätigte nicht nur die Erkenntnisse aus der Diederwinkelanalyse, sie zeigte auch die grundlegenden Unterschiede in den Veränderungen, die bei der Bindung von leichten und sperrigen Pyrrolidincarboxamiden auftreten. Während die kleineren Verbindungen sowohl Einfluss auf Aminosäuren des Minor als auch des Major Portal haben, haben die sperrigen Vertreter nur Einfluss auf die entscheidenden Aminosäuren des Substratbindeloops, d.h. Ile202 und Val203. Die *Dynamic Cross Correlation* Karten für die beiden representativen Liganden pc-d11 und pc-c7a3 unterstützen die *Essential Dynamics*-Ergebnisse. Beide Analysen zusammen halfen die geringen Bewegungen des Substratbindeloops zum Liganden hin aufzudecken.

Die Informationen aus den *Essential Dynamics* und *Dynamic Cross Corrrelation* Analysen wurden verwendet, um eine strukturbasierte Optimierung des sperrigen Pyrrolidincarboxamid-Gerüsts durchzuführen. Insgesamt wurden 20 Verbindungen entworfen und einer umfangreichen *in-silico* Analyse unterzogen, die molekulares Docking, Rescoring, MD-Simulationen und auch die mykobakterielle Zellwandpermeabilitätsvorhersage einschloss. Alle diese Analysen zeigten moderate Verbesserungen gegenüber den Ausgangsmolekülen (pc-c6a3 und pc-c7a3) und dem Referenzmolekül (pc-d11). Für ein vielversprechendes Subset neuer Moleküle wurden umfangreiche MD-Simulationen von jeweils 150 ns pro Protein-Ligand-Komplex durchgeführt, um die Veränderungen im Bindungsmodus sowie im Protein bei erweiterter Simulationsdauer zu ermitteln. Mit Ausnahme einiger weniger, zeigten alle Moleküle eine verbesserte Bindungsstabilität, obwohl der Gesamtligand-RMSD vergleichsweise höher war als jener des Referenz- oder der Ausgangsliganden. Diese Simulationen wurden auch verwendet, um die molekularen Determinanten der schnellen reversiblen Bindung von Pyrrolidincarboxamiden zu beleuchten.

Schließlich wurden unter umfangreicher Verwendung von Clustering-Techniken, die dominanten Bindungskonformationen für Pyrrolidincarboxamide ermittelt und versucht, die ursächlichen Faktoren hinter deren relativ schwacher Bindung zu bestimmen. Das Clustering und der anschließende Vergleich der dominanten Konformationen mit jener von *slow-tight binders* zeigte, daß Pyrrolidincarboxamide einen weit offenen Zustand des Substratbindeloops zeigen. Eine kleinere Konformationsfamilie war mit einer weniger offenen Form des Substratbindeloops verbunden. Das Clustering zeigte auch die strukturellen Determinanten der offensichtlichen schnellen und reversiblen Bindung auf. Diese Information wurde später verwendet, um zwei weitere Inhibitoren vorzuschlagen, die günstige Eigenschaften von *slow-tight binders* zeigen sollten. Ein Vergleich mit den Erkenntnissen aus der Literatur deutet auf die Möglichkeit einer verbesserten Bindung für die vorgeschlagenen Moleküle im Vergleich zum Referenzliganden hin. Zusammenfassend lässt sich sagen, dass unter den entworfenen Pyrrolidincarboxamiden vielversprechende Moleküle enthalten sind, die günstige Eigenschaften aufweisen und eine weitere experimentelle Untersuchung rechtfertigen würden.

Bibliography

[1] P. N. Fonkwo. Pricing infectious disease. *EMBO rep.*, 9(1S):S13–S17, 2008.

[2] K. Williams. The introduction of "chemotherapy" using arsphenamine - the first magic bullet. *J. R. Soc. Med.*, 102(8):343–348, 2009.

[3] R. I. Aminov. A brief history of the antibiotic era: Lessons learned and challenges for the future. *Front. Microbiol.*, 1:134, 2010.

[4] 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, 2014.

[5] J. Davies. Where have all the antibiotics gone? *Can. J. Infect. Dis. Med. Microbiol.*, 17(5):287, 2006.

[6] R. J. Fair and Y. Tor. Antibiotics and bacterial resistance in the 21^{st} century. *Perspect. Medicin. Chem.*, 6:25, 2014.

[7] P. McGann, E. Snesrud, R. Maybank, B. Corey, A. C. Ong, R. Clifford, M. Hinkle, T. Whitman, E. Lesho, and K. E. Schaecher. *Escherichia coli* Harboring *mcr-1* and bla_{CTX-M} on a Novel IncF Plasmid: First report of *mcr-1* in the USA. *Antimicrob. Agents. Chemother.*, 2016.

[8] M. C. Enright, D. A. Robinson, G. Randle, E. J. Feil, H. Grundmann, and B. G. Spratt. The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc. Natl. Acad. Sci.*, 99(11):7687–7692, 2002.

[9] I. M. Gould. VRSA - doomsday superbug or damp squib? *Lancet Infect. Dis.*, 10(12):816–818, 2010.

[10] M. Wootton, R. Howe, T. Walsh, P. Bennett, and A. MacGowan. *In vitro* activity of 21 antimicrobials against vancomycin-resistant *Staphylococcus aureus* (VRSA) and heteroVRSA (hVRSA). *J. Antimicrob. Chemother.*, 50(5):760–760, 2002.

[11] P. J. Brennan. Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. *Tuberculosis*, 83(1):91–97, 2003.

[12] V. Jarlier and H. Nikaido. Mycobacterial cell wall: Structure and role in natural resistance to antibiotics. *FEMS Microbiol. Lett.*, 123(1-2):11–18, 1994.

[13] H. Nikaido. Preventing drug access to targets: cell surface permeability barriers and active efflux in bacteria. In *Semin. Cell Dev. Biol.*, volume 12, pages 215–223. Elsevier, 2001.

[14] Y. Zhang, B. Heym, B. Allen, D. Young, and S. Cole. The catalase-peroxidase gene and isoniazid resistance of *Mycobacterium tuberculosis*. *Nature*, 358(6387):591–593, 1992.

[15] J. M. Musser, V. Kapur, D. L. Williams, B. N. Kreiswirth, D. Van Soolingen, and J. D. Van Embden. Characterization of the catalase-peroxidase gene (*katG*) and *inhA* locus in Isoniazid-Resistant and-Susceptible strains of *Mycobacterium tuberculosis* by Automated DNA Sequencing: Restricted Array of Mutations Associated with Drug Resistance. *J. Infect. Dis.*, 173(1):196–202, 1996.

[16] R. Rawat, A. Whitty, and P. J. Tonge. The isoniazid-NAD adduct is a slow, tight-binding inhibitor of InhA, the *Mycobacterium tuberculosis* enoyl reductase: Adduct affinity and drug resistance. *Proc. Natl. Acad. Sci.*, 100(24):13881–13886, 2003.

[17] R. Colangeli, D. Helb, S. Sridharan, J. Sun, M. Varma-Basil, M. H. Hazbón, R. Harbacheuski, N. J. Megjugorac, W. R. Jacobs, A. Holzenburg, et al. The *Mycobacterium tuberculosis iniA* gene is essential for activity of an efflux pump that confers drug tolerance to both isoniazid and ethambutol. *Mol. Microbiol.*, 55(6):1829–1840, 2005.

[18] G. Li, J. Zhang, Q. Guo, Y. Jiang, J. Wei, L. L. Zhao, X. Zhao, J. Lu, and K. Wan. Efflux pump gene expression in multidrug-resistant *Mycobacterium tuberculosis* clinical isolates. *PLoS One*, 10(2):e0119013, 2015.

[19] 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 Sci.*, 19(3):458–474, 2010.

[20] G. E. Louw, R. M. Warren, N. C. Gey Van Pittius, C. R. E. McEvoy, P. D. Van Helden, and T. C. Victor. A Balancing Act: Efflux/Influx in Mycobacterial Drug Resistance. *Antimicrob. Agents. Chemother.*, 53(8):3181–3189, 2009.

[21] P. D. Lister, D. J. Wolter, and N. D. Hanson. Antibacterial-resistant *Pseudomonas aeruginosa*: Clinical Impact and Complex Regulation of Chromosomally Encoded Resistance Mechanisms. *Clin. Microbiol. Rev.*, 22(4):582–610, 2009.

[22] T. Strateva and D. Yordanov. *Pseudomonas aeruginosa* - A phenomenon of bacterial resistance. *J. Med. Microbiol.*, 58(9):1133–1148, 2009.

[23] K. J. Kieser and E. J. Rubin. How sisters grow apart: mycobacterial growth and division. *Nat. Rev. Microbiol.*, 12(8):550–562, 2014.

[24] T. Wirth, F. Hildebrand, C. Allix-Béguec, F. Wölbeling, T. Kubica, K. Kremer, D. van Soolingen, S. Rüsch-Gerdes, C. Locht, S. Brisse, et al. Origin, Spread and Demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.*, 4(9):e1000160, 2008.

[25] World Health Organization. Global tuberculosis report 2015, 2015.

[26] World Health Organization. *Treatment of tuberculosis: guidelines*. World Health Organization, 2010.

[27] N. S. Shah, A. Wright, G.-H. Bai, L. Barrera, F. Boulahbal, F. Drobniewski, C. Gilpin, M. Havelkov, R. Lepe, R. Lumb, et al. Worldwide emergence of extensively drug-resistant tuberculosis. *Emerg. Infect. Dis.*, 2007.

[28] R. Prasad. Multidrug and extensively drug-resistant TB (M/XDR-TB): Problems and Solutions. *Indian J. Tuberc.*, 57(4):180–191, 2010.

[29] G. Migliori, G. De Iaco, G. Besozzi, R. Centis, and D. Cirillo. First tuberculosis cases in Italy resistant to all tested drugs. *Euro Surveill.*, 12(5):E070517, 2007.

[30] G. L. Calligaro, L. Moodley, G. Symons, and K. Dheda. The medical and surgical treatment of drug-resistant tuberculosis. *J. Thorac. Dis.*, 6(3):186, 2014.

[31] A. A. Velayati, M. R. Masjedi, P. Farnia, P. Tabarsi, J. Ghanavi, A. H. ZiaZarifi, and S. E. Hoffner. Emergence of new forms of totally drug-resistant tuberculosis bacilli: Super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran. *Chest*, 136(2):420–425, 2009.

[32] K. Rowland. Totally drug-resistant TB emerges in India. *Nature*, 1:9797, 2012.

[33] G. Lamichhane. Novel targets in *M. tuberculosis*: search for new drugs. *Trends Mol. Med.*, 17(1):25–33, 2011.

[34] T. R. Ioerger and J. C. Sacchettini. Structural genomics approach to drug discovery for *Mycobacterium tuberculosis*. *Curr. Opin. Microbiol.*, 12(3):318–325, 2009.

[35] E. Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.*, 27(3):2985–2993, 1894.

[36] D. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci.*, 44(2):98–104, 1958.

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

[38] H. Lu and P. J. Tonge. Drug - target residence time: Critical information for lead optimization. *Curr. Opin. Chem. Biol.*, 14(4):467–474, 2010.

[39] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[40] B. Ma, S. Kumar, C. J. Tsai, and R. Nussinov. Folding funnels and binding mechanisms. *Protein Eng.*, 12(9):713–720, 1999.

[41] C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov. Folding funnels, binding funnels, and protein function. *Protein Sci.*, 8(06):1181–1190, 1999.

[42] S. Kumar, B. Ma, C.-J. Tsai, N. Sinha, and R. Nussinov. Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Sci.*, 9(1):10–19, 2000.

[43] C.-J. Tsai, B. Ma, and R. Nussinov. Folding and binding cascades: Shifts in energy landscapes. *Proc. Natl. Acad. Sci.*, 96(18):9970–9972, 1999.

[44] D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, 5(11):789–796, 2009.

[45] D. E. Shaw, R. O. Dror, J. K. Salmon, J. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the conference on high performance computing networking, storage and analysis*, pages 1–11. IEEE, 2009.

[46] S. Piana, J. L. Klepeis, and D. E. Shaw. Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.*, 24:98–105, 2014.

[47] G. Tiana and C. Camilloni. Ratcheted molecular-dynamics simulations identify efficiently the transition state of protein folding. *J. Chem. Phys.*, 137(23):235101, 2012.

[48] 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. *Nat. Commun.*, 6, 2015.

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

[50] P. Pan and P. J. Tonge. Targeting InhA, the FAS II enoyl-ACP reductase: SAR studies on novel inhibitor scaffolds. *Curr. Top. Med. Chem.*, 12(7):672, 2012.

[51] U. H. Manjunatha, S. P. Rao, R. R. Kondreddi, C. G. Noble, L. R. Camacho, B. H. Tan, S. H. Ng, P. S. Ng, N. L. Ma, S. B. Lakshminarayana, et al. Direct inhibitors of InhA are active against *Mycobacterium tuberculosis*. *Sci. Transl. Med.*, 7(269):269ra3–269ra3, 2015.

[52] X. He, A. Alian, R. Stroud, and P. R. Ortiz de Montellano. Pyrrolidine carboxamides as a novel class of inhibitors of enoyl acyl carrier protein reductase from *Mycobacterium tuberculosis*. *J. Med. Chem.*, 49(21):6308–6323, 2006.

[53] 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.

[54] R. L. Hunter, M. Olsen, C. Jagannath, and J. K. Actor. Trehalose 6,6'-dimycolate and lipid in the pathogenesis of caseating granulomas of tuberculosis in mice. *Am. J. Pathol.*, 168(4):1249–1261, 2006.

[55] M. S. Glickman, J. S. Cox, and W. R. Jacobs. A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*. *Mol. Cell*, 5(4):717–727, 2000.

[56] M. S. Glickman and W. R. Jacobs. Microbial pathogenesis of *Mycobacterium tuberculosis*: Dawn of a discipline. *Cell*, 104(4):477–485, 2001.

[57] E. Dubnau, J. Chan, C. Raynaud, V. P. Mohan, M. A. Lanéelle, K. Yu, A. Quémard, I. Smith, and M. Daffé. Oxygenated mycolic acids are necessary for virulence of *Mycobacterium tuberculosis* in mice. *Mol. Microbiol.*, 36(3):630–637, 2000.

[58] A. K. Ojha, A. D. Baughn, D. Sambandan, T. Hsu, X. Trivelli, Y. Guerardel, A. Alahari, L. Kremer, W. R. Jacobs, and G. F. Hatfull. Growth of *Mycobacterium tuberculosis* biofilms containing free mycolic acids and harbouring drug-tolerant bacteria. *Mol. Microbiol.*, 69(1):164–174, 2008.

[59] Y. Yuan, Y. Zhu, D. D. Crane, and C. E. Barry III. The effect of oxygenated mycolic acid composition on cell wall function and macrophage growth in *Mycobacterium tuberculosis*. *Mol. Microbiol.*, 29(6):1449–1458, 1998.

[60] M. S. Glickman, S. M. Cahill, and W. R. Jacobs. The *Mycobacterium tuberculosis cmaA2* gene encodes a mycolic acid trans-cyclopropane synthetase. *J. Biol. Chem.*, 276(3):2228–2233, 2001.

[61] K. Takayama, C. Wang, and G. S. Besra. Pathway to synthesis and processing of mycolic acids in *Mycobacterium tuberculosis*. *Clin. Microbiol. Rev.*, 18(1):81–101, 2005.

[62] A. Bhatt, V. Molle, G. S. Besra, W. R. Jacobs, and L. Kremer. The *Mycobacterium tuberculosis* FAS-II condensing enzymes: Their role in mycolic acid biosynthesis, acid-fastness, pathogenesis and in future drug development. *Mol. Microbiol.*, 64(6):1442–1454, 2007.

[63] H. Marrakchi, M. A. Lanéelle, and M. Daffé. Mycolic acids: Structures, biosynthesis, and beyond. *Chem. Biol.*, 21(1):67–85, 2014.

[64] M. Allen, C. Bailey, I. Cahatol, L. Dodge, J. Yim, C. Kassissa, J. Luong, S. Kasko, S. Pandya, and V. Venketaraman. Mechanisms of control of *Mycobacterium tuberculosis* by NK cells: Role of glutathione. *Front. Immunol.*, 6, 2015.

[65] S. W. White, J. Zheng, Y.-M. Zhang, and C. O. Rock. The structural biology of type II fatty acid biosynthesis. *Annu. Rev. Biochem.*, 74:791–831, 2005.

[66] R. P. Massengo-Tiassé and J. E. Cronan. Diversity in enoyl-acyl carrier protein reductases. *Cell. Mol. Life Sci.*, 66(9):1507–1517, 2009.

[67] P. Pan, S. Knudson, G. R. Bommineni, H. J. Li, C. T. Lai, N. Liu, M. Garcia-Diaz, C. Simmerling, S. S. Patil, R. A. Slayden, et al. Time-dependent diaryl ether inhibitors of InhA: SAR studies of enzyme inhibition, antibacterial activity, and in vivo efficacy. *ChemMedChem*, 9(4):776–791, 2014.

[68] H. Cheng, J. Hoffman, P. Le, S. K. Nair, S. Cripps, J. Matthews, C. Smith, M. Yang, S. Kupchinsky, K. Dress, et al. The development and SAR of pyrrolidine carboxamide 11β-HSD1 inhibitors. *Bioorg. Med. Chem. Lett.*, 20(9):2897–2902, 2010.

[69] B. P. Moore, D. H. Chung, D. S. Matharu, J. E. Golden, C. Maddox, L. Rasmussen, J. W. Noah, M. I. Sosa, S. Ananthan, A. Tower, Nichole, et al. (s)-n-(2,5-dimethylphenyl)-1-(quinoline-8-ylsulfonyl) pyrrolidine-2-carboxamide as a small molecule inhibitor probe for the study of respiratory syncytial virus infection. *J. Med. Chem.*, 55(20):8582–8587, 2012.

[70] Q. Ding, N. Jiang, J. Liu, J. Zhang, and Z. Zhang. Substituted Pyrrolidine-2-Carboxamides, May 19 2011. US Patent App. 12/898,955.

[71] A. Cole, J. Letourneau, and K. Ho. Pyrrolidine carboxamide compounds, May 27 2010. WO Patent App. PCT/US2009/065,290.

[72] 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*. *J. Biol. Chem.*, 285(19):14330–14337, 2010.

[73] T. J. Sullivan, J. J. Truglio, M. E. Boyne, P. Novichenok, X. Zhang, C. F. Stratton, H.-J. Li, T. Kaur, A. Amin, F. Johnson, et al. High affinity InhA inhibitors with activity against drug-resistant strains of *Mycobacterium tuberculosis*. *ACS Chem. Biol.*, 1(1):43–53, 2006.

[74] H. J. Li, C. T. 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 Chem. Biol.*, 9(4):986–993, 2014.

[75] 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. *Biophys. J.*, 89(2):876–884, 2005.

[76] E. H. S. Sousa, L. A. Basso, D. S. Santos, I. C. N. Diógenes, E. Longhinotti, L. G. de França Lopes, and Í. de Sousa Moreira. Isoniazid metal complex reactivity and insights for a novel anti-tuberculosis drug design. *J. Biol. Inorg. Chem.*, 17(2):275–283, 2012.

[77] J. S. Oliveira, E. H. Sousa, L. A. Basso, M. Palaci, R. Dietze, D. S. Santos, and Í. S. Moreira. An inorganic iron complex that inhibits wild-type and an isoniazid-resistant mutant 2-trans-enoyl-ACP (CoA) reductase from *Mycobacterium tuberculosis*. *Chem. Commun.*, 1(3):312–313, 2004.

[78] E. M. Cohen, K. S. Machado, M. Cohen, and O. N. de Souza. Effect of the explicit flexibility of the InhA enzyme from *Mycobacterium tuberculosis* in molecular docking simulations. *BMC Genom.*, 12(4):1, 2011.

[79] K. F. 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 *Mycobacterium tuberculosis*. *Int. J. Quantum Chem.*, 106(13):2689–2699, 2006.

[80] 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. *Chem. Biol. Drug. Des.*, 72(5):444–449, 2008.

[81] 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. *Eur. J. Med. Chem.*, 45(12):5585–5593, 2010.

[82] 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 QSAR Environ. Res.*, 25(6):473–488, 2014.

[83] F. G. Parak. Proteins in action: The physics of structural fluctuations and conformational changes. *Curr. Opin. Struct. Biol.*, 13(5):552–557, 2003.

[84] L. S. Busenlehner and R. N. Armstrong. Insights into enzyme structure and dynamics elucidated by amide H/D exchange mass spectrometry. *Arch. Biochem. Biophys.*, 433(1):34–46, 2005.

[85] W. J. Greenleaf, M. T. Woodside, and S. M. Block. High-resolution, single-molecule measurements of biomolecular motion. *Annu. Rev. Biophys. Biomol. Struct.*, 36:171, 2007.

[86] R. A. Copeland. Conformational adaptation in drug-target interactions and residence time. *Futur. Med. Chem.*, 3(12):1491–1501, 2011.

[87] S. A. Mousa, J. M. Bozarth, U. P. Naik, and A. Slee. Platelet GPIIb/IIIa binding characteristics of small molecule RGD mimetic: Distinct binding profile for Roxifiban. *Br. J. Pharmacol.*, 133(3):331–336, 2001.

[88] S. Kapur and P. Seeman. Antipsychotic agents differ in how fast they come off the dopamine D2 receptors. implications for atypical antipsychotic action. *J. Psychiatry Neurosci.*, 25(2):161, 2000.

[89] S. A. Lipton. Paradigm shift in neuroprotection by NMDA receptor blockade: memantine and beyond. *Nat. Rev. Drug Discov.*, 5(2):160–170, 2006.

[90] H. Wu, D. S. Pfarr, Y. Tang, L. L. An, N. K. Patel, J. D. Watkins, W. D. Huse, P. A. Kiener, and J. F. Young. Ultra-potent antibodies against respiratory syncytial virus: Effects of binding kinetics and binding valence on viral neutralization. *J. Mol. Biol.*, 350(1):126–144, 2005.

[91] C. M. Song, S. J. Lim, and J. C. Tong. Recent advances in computer-aided drug design. *Brief. Bioinform.*, 10(5):579–591, 2009.

[92] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe. Computational methods in drug discovery. *Pharmacol. Rev.*, 66(1):334–395, 2014.

[93] J. Bajorath. Computer-aided drug discovery. *F1000Research*, 4, 2015.

[94] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 3(11):935–949, 2004.

[95] A. R. Leach. *Molecular modelling: principles and applications.* Pearson education, 2001.

[96] K. A. Sharp. Statistical thermodynamics of binding and molecular recognition models. *Prot.-Lig. Interact*, 3, 2012.

[97] B. O. Brandsdal, F. Österberg, M. Almlöf, I. Feierberg, V. B. Luzhkov, and J. Åqvist. Free energy calculations and ligand binding. *Adv. Protein Chem.*, 66:123–158, 2003.

[98] M. K. Gilson and H. X. Zhou. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36(1):21, 2007.

[99] J. Åqvist, C. Medina, and J. E. Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.*, 7(3):385–391, 1994.

[100] M. Almlöf, J. Carlsson, and J. Åqvist. Improving the accuracy of the linear interaction energy method for solvation free energies. *J. Chem. Theory Comput.*, 3(6):2162–2175, 2007.

[101] H. Gutiérrez-de-Terán and J. Åqvist. Linear interaction energy: Method and applications in drug design. In *Computational Drug Discovery and Design*, pages 305–323. Springer, 2012.

[102] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J. Am. Chem. Soc.*, 120(37):9401–9409, 1998.

[103] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, et al. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33(12):889–897, 2000.

[104] S. Genheden and U. Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.*, 10(5):449–461, 2015.

[105] F. S. Lee, Z. T. Chu, M. B. Bolger, and A. Warshel. Calculations of antibody-antigen interactions: Microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603. *Protein Eng.*, 5(3):215–228, 1992.

[106] Y. Y. Sham, Z. T. Chu, H. Tao, and A. Warshel. Examining methods for calculations of binding free energies: LRA, LIE, PDLD-LRA, and PDLD/S-LRA calculations of ligands binding to an HIV protease. *Proteins: Struct., Funct., Bioinf.*, 39(4):393–407, 2000.

[107] D. A. Pearlman. Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. *J. Med. Chem.*, 48(24):7796–7807, 2005.

[108] P. Mikulskis, S. Genheden, and U. Ryde. Effect of explicit water molecules on ligand-binding affinities calculated with the MM/GBSA approach. *J. Mol. Model.*, 20(6):1–11, 2014.

[109] P. W. Rose, C. Bi, W. F. Bluhm, C. H. Christie, D. Dimitropoulos, S. Dutta, R. K. Green, D. S. Goodsell, A. Prlić, M. Quesada, et al. The RCSB protein data bank: New resources for research and education. *Nucleic Acids Res.*, 41(D1):D475–D482, 2013.

[110] I. Muegge and M. Rarey. Small molecule docking and scoring. *Rev. Comput. Chem.*, 17:1–60, 2001.

[111] H. J. Böhm and M. Stahl. The use of scoring functions in drug discovery applications. *Rev. Comput. Chem.*, 18:41–88, 2002.

[112] N. Brooijmans and I. D. Kuntz. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.*, 32(1):335–373, 2003.

[113] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161(2):269–288, 1982.

[114] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Bioinf.*, 47(4):409–443, 2002.

[115] R. Dias, J. de Azevedo, and F. Walter. Molecular docking algorithms. *Curr. Drug Targets*, 9(12):1040–1047, 2008.

[116] W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner, and R. Farid. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.*, 49(2):534–553, 2006.

[117] W. Sherman, H. S. Beard, and R. Farid. Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug. Des.*, 67(1):83–84, 2006.

[118] Y. P. Pang, E. Perola, K. Xu, and F. G. Prendergast. EUDOC: A computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.*, 22(15):1750–1771, 2001.

[119] M. McGann. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.*, 51(3):578–596, 2011.

[120] C. A. Sotriffer. Protein–ligand docking: From basic principles to advanced applications. In *In Silico Drug Discovery and Design*, pages 155–188. Informa UK Limited, 2015.

[121] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261(3):470–489, 1996.

[122] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.*, 15(5):411–428, 2001.

[123] M. A. Neves, M. Totrov, and R. Abagyan. Docking and scoring with ICM: The benchmarking results and strategies for improvement. *J. Comput. Aided Mol. Des.*, 26(6):675–686, 2012.

[124] M. Liu and S. Wang. MCDOCK: A Monte Carlo simulation approach to the molecular docking problem. *J. Comput. Aided Mol. Des.*, 13(5):435–451, 1999.

[125] D. Rognan. Docking methods for virtual screening: Principles and recent advances. *Virtual Screening: Princ. Challenges, Pract. Guidel.*, pages 153–176, 2011.

[126] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19(14):1639–1662, 1998.

[127] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor. Improved protein - ligand docking using GOLD. *Proteins: Struct., Funct., Bioinf.*, 52(4):609–623, 2003.

[128] C. Abrams and G. Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2013.

[129] R. Mannhold, H. Kubinyi, H. Timmerman, H. D. Höltje, and G. Folkers. *Molecular Modeling: Basic Principles and Applications*, volume 5. John Wiley & Sons, 2008.

[130] D. Zilian and C. A. Sotriffer. SFCscore^{RF}: A random forest-based scoring function for improved affinity prediction of protein - ligand complexes. *J. Chem. Inf. Model.*, 53(8):1923–1933, 2013.

[131] P. J. Ballester and J. B. Mitchell. A machine learning approach to predicting protein - ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

[132] H. J. Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.*, 8(3):243–256, 1994.

[133] H. J. Böhm. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.*, 6(1):61–78, 1992.

[134] 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, et al. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–1749, 2004.

[135] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.*, 11(5):425–445, 1997.

[136] C. A. Sotriffer, P. Sanschagrin, H. Matter, and G. Klebe. SFCscore: Scoring functions for affinity prediction of protein - ligand complexes. *Proteins: Struct., Funct., Bioinf.*, 73(2):395–419, 2008.

[137] P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, and C. L. Brooks. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.*, 47(12):3032–3047, 2004.

[138] J. Åqvist and J. Marelius. The linear interaction energy method for predicting ligand binding free energies. *Comb. Chem. & High Throughput Screen.*, 4(8):613–626, 2001.

[139] D. Huang and A. Caflisch. Efficient evaluation of binding free energy using continuum electrostatics solvation. *J. Med. Chem.*, 47(23):5791–5797, 2004.

[140] I. Muegge and Y. C. Martin. A general and fast scoring function for protein - ligand interactions: A simplified potential approach. *J. Med. Chem.*, 42(5):791–804, 1999.

[141] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295(2):337–356, 2000.

[142] G. Neudert and G. Klebe. DSX: A knowledge-based scoring function for the assessment of protein - ligand complexes. *J. Chem. Inf. Model.*, 51(10):2731–2745, 2011.

[143] H. F. Velec, H. Gohlke, and G. Klebe. DrugScore^{CSD} knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.*, 48(20):6296–6303, 2005.

[144] D. M. Krüger, J. I. Garzón, P. Montes, and H. Gohlke. Predicting protein-protein interactions with DrugScore^{PPI}: fully-flexible docking, scoring, and in silico alanine-scanning. *J. Cheminform.*, 3:1–1, 2011.

[145] C. A. Sotriffer and H. Matter. The Challenge of Affinity Prediction: Scoring Functions for Structure-Based Virtual Screening. *Virtual Screening: Princ. Challenges, Pract. Guidel.*, pages 177–221, 2011.

[146] I. Reulecke, G. Lange, J. Albrecht, R. Klein, and M. Rarey. Towards an integrated description of hydrogen bonding and dehydration: Decreasing false positives in virtual screening with the HYDE scoring function. *ChemMedChem*, 3(6):885–897, 2008.

[147] T. Cheng, X. Li, Y. Li, Z. Liu, and R. Wang. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.*, 49(4):1079–1093, 2009.

[148] A. Kukol. Consensus virtual screening approaches to predict protein ligands. *Eur. J. Med. Chem.*, 46(9):4661–4664, 2011.

[149] P. S. Charifson and W. P. Walters. Filtering databases and chemical libraries. *Molec. Divers.*, 5(4):185–197, 2000.

[150] P. S. Charifson, J. J. Corkery, M. A. Murcko, and W. P. Walters. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.*, 42(25):5100–5109, 1999.

[151] S. Betzi, K. Suhre, B. Chétrit, F. Guerlesquin, and X. Morelli. GFscore: A general nonlinear consensus scoring function for high-throughput docking. *J. Chem. Inf. Model.*, 46(4):1704–1712, 2006.

[152] S. Bar-Haim, A. Aharon, T. Ben-Moshe, Y. Marantz, and H. Senderowitz. SeleX-CS: A new consensus scoring algorithm for hit discovery and lead optimization. *J. Chem. Inf. Model.*, 49(3):623–633, 2009.

[153] H. Gohlke and G. Klebe. Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. *J. Med. Chem.*, 45(19):4153–4170, 2002.

[154] H. Matter, D. W. Will, M. Nazaré, H. Schreuder, V. Laux, and V. Wehner. Structural requirements for factor Xa inhibition by 3-oxybenzamides with neutral P1 substituents: combining X-ray crystallography, 3D-QSAR, and tailored scoring functions. *J. Med. Chem.*, 48(9):3290–3312, 2005.

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

[156] H. Kubinyi. Comparative molecular field analysis (CoMFA). *Handb. Chemoinformatics: From Data to Knowl. 4 Volumes*, pages 1555–1574, 2008.

[157] A. M. Henzler and M. Rarey. Protein Flexibility in Structure-Based Virtual Screening: From Models to Algorithms. *Virtual Screening: Princ. Challenges, Pract. Guidel.*, pages 223–244, 2011.

[158] H. Zhang, T. Zhang, K. Chen, S. Shen, J. Ruan, and L. Kurgan. On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins: Struct., Funct., Bioinf.*, 76(3):617–636, 2009.

[159] P. Cozzini, G. E. Kellogg, F. Spyrakis, D. J. Abraham, G. Costantino, A. Emerson, F. Fanelli, H. Gohlke, L. A. Kuhn, G. M. Morris, et al. Target Flexibility: An Emerging Consideration in Drug Discovery and Design. *J. Med. Chem.*, 51(20):6237–6255, 2008.

[160] N. Furnham, T. L. Blundell, M. A. DePristo, and T. C. Terwilliger. Is one solution good enough? *Nat. Struct. Mol. Biol.*, 13(3):184–185, 2006.

[161] J. H. Lin, A. L. Perryman, J. R. Schames, and J. A. McCammon. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J. Am. Chem. Soc.*, 124(20):5632–5633, 2002.

[162] R. E. Amaro, R. Baron, and J. A. McCammon. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. Aided Mol. Des.*, 22(9):693–705, 2008.

[163] J. Nilmeier and M. Jacobson. Multiscale Monte Carlo sampling of protein sidechains: Application to binding pocket flexibility. *J. Chem. Theory Comput.*, 4(5):835–846, 2008.

[164] A. Shehu, L. E. Kavraki, and C. Clementi. Multiscale characterization of protein conformational ensembles. *Proteins: Struct., Funct., Bioinf.*, 76(4):837–851, 2009.

[165] M. Thorpe, M. Lei, A. Rader, D. J. Jacobs, and L. A. Kuhn. Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. Model.*, 19(1):60–69, 2001.

[166] M. Lei, M. I. Zavodszky, L. A. Kuhn, and M. Thorpe. Sampling protein conformations and pathways. *J. Comput. Chem.*, 25(9):1133–1148, 2004.

[167] A. R. Leach, A. P. Lemon, et al. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Protein Struct. Funct. Genet.*, 33(2):227–239, 1998.

[168] A. C. Anderson, R. H. O'Neil, T. S. Surti, and R. M. Stroud. Approaches to solving the rigid receptor problem by identifying a minimal set of flexible residues during ligand docking. *Chem. Biol.*, 8(5):445–457, 2001.

[169] C. Hartmann, I. Antes, and T. Lengauer. IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci.*, 16(7):1294–1307, 2007.

[170] B. Brooks and M. Karplus. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci.*, 80(21):6571–6575, 1983.

[171] S. Hayward and B. L. De Groot. Normal modes and essential dynamics. In *Molecular Modeling of Proteins*, pages 89–106. Springer, 2008.

[172] F. Österberg, G. M. Morris, M. F. Sanner, A. J. Olson, and D. S. Goodsell. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins: Struct., Funct., Bioinf.*, 46(1):34–40, 2002.

[173] G. Bottegoni, I. Kufareva, M. Totrov, and R. Abagyan. Four-dimensional docking: A fast and accurate account of discrete receptor flexibility in ligand docking. *J. Med. Chem.*, 52(2):397–406, 2008.

[174] H. Claußen, C. Buning, M. Rarey, and T. Lengauer. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, 308(2):377–395, 2001.

[175] F. Jiang and S. H. Kim. "Soft Docking": Matching of molecular surface cubes. *J. Mol. Biol.*, 219(1):79–102, 1991.

[176] A. M. Ferrari, B. Q. Wei, L. Costantino, and B. K. Shoichet. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.*, 47(21):5076–5084, 2004.

[177] V. Schnecke and L. A. Kuhn. Database screening for HIV protease ligands: The influence of binding-site conformation and representation on ligand selectivity. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 242–251, 1999.

[178] M. Mangoni, D. Roccatano, and A. Di Nola. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins: Struct., Funct., Bioinf.*, 35(2):153–162, 1999.

[179] H. Merlitz and W. Wenzel. Comparison of stochastic optimization methods for receptor - ligand docking. *Chem. Phys. Lett.*, 362(3):271–277, 2002.

[180] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997.

[181] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30(16):2785–2791, 2009.

[182] M. Zacharias. Rapid protein - ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: Binding of FK506 to FKBP. *Proteins: Struct., Funct., Bioinf.*, 54(4):759–767, 2004.

[183] R. Tatsumi, Y. Fukunishi, and H. Nakamura. A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor. *J. Comput. Chem.*, 25(16):1995–2005, 2004.

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

[185] B. Alder and T. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27(5):1208, 1957.

[186] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[187] H. Ode, M. Nakashima, S. Kitamura, W. Sugiura, and H. Sato. Molecular dynamics simulation in virus research. *Front Microbiol.*, 3:258, 2012.

[188] R. Petrenko and J. Meller. Molecular dynamics. *eLS*, 2010.

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

[190] W. F. van Gunsteren and H. J. Berendsen. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Ed.*, 29(9):992–1023, 1990.

[191] A. D. MacKerell Jr. Atomistic models and force fields. *Comput. Biochem. Biophys.*, pages 7–38, 2001.

[192] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31(4):671–690, 2010.

[193] 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. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[194] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25(13):1656–1676, 2004.

[195] W. L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988.

[196] 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. *J. Phys. Chem.*, 97(40):10269–10280, 1993.

[197] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.*, 65(3):712–725, 2006.

[198] R. Chelli, V. Schettino, and P. Procacci. Comparing polarizable force fields to *ab initio* calculations reveals nonclassical effects in condensed phases. *J. Chem. Phys.*, 122(23):234107, 2005.

[199] P. Cieplak, F. Y. Dupradeau, Y. Duan, and J. Wang. Polarization effects in molecular mechanical force fields. *J. Phys. Condens. Matter*, 21(33):333102, 2009.

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

[201] L. Verlet. Computer ”experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159(1):98, 1967.

[202] M. P. Allen and D. J. Tildesley. *Computer simulation of liquids.* Oxford university press, 1989.

[203] R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics, Desktop Edition Volume I*, volume 1. Basic books, 2013.

[204] D. Beeman. Some multistep methods for use in molecular dynamics calculations. *J. Comput. Phys.*, 20(2):130–139, 1976.

[205] J. P. Ryckaert, G. Ciccotti, and H. J. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327–341, 1977.

[206] D. J. Tobias and C. L. Brooks III. Molecular dynamics with internal coordinate constraints. *J. Chem. Phys.*, 89(8):5115–5127, 1988.

[207] H. C. Andersen. RATTLE: A "Velocity" version of the SHAKE algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52(1):24–34, 1983.

[208] S. Miyamoto and P. A. Kollman. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13(8):952–962, 1992.

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

[210] H. J. Berendsen, J. P. Postma, W. F. van Gunsteren, A. DiNola, and J. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[211] T. Schlick. *Molecular modeling and simulation: An interdisciplinary guide*, volume 21. Springer Science & Business Media, 2010.

[212] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.*, 103(11):4613–4621, 1995.

[213] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26(16):1781–1802, 2005.

[214] D. C. Rapaport. *The art of molecular dynamics simulation.* Cambridge University Press, 2004.

[215] T. Darden, D. York, and L. Pedersen. Particle Mesh Ewald: An N log (N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[216] F. S. Lee and A. Warshel. A local reaction field method for fast evaluation of long-range electrostatic interactions in molecular simulations. *J. Chem. Phys.*, 97(5):3100–3107, 1992.

[217] J. Norberg and L. Nilsson. On the truncation of long-range electrostatic interactions in DNA. *Biophys. J.*, 79(3):1537–1553, 2000.

[218] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw. Biomolecular simulation: A computational microscope for molecular biology. *Annu. Rev. Biophys.*, 41:429–452, 2012.

[219] H. Fujisaki, K. Moritsugu, Y. Matsunaga, T. Morishita, and L. Maragliano. Extended phase-space methods for enhanced sampling in molecular simulations: A review. *Front. Bioeng. Biotechnol.*, 3, 2015.

[220] B. Roux. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.*, 91(1):275–282, 1995.

[221] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, 1977.

[222] J. Kästner. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Science*, 1(6):932–942, 2011.

[223] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.*, 120(24):11919–11929, 2004.

[224] L. C. 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. *J. Chem. Theory Comput.*, 8(9):2997–3002, 2012.

[225] J. Hritz and C. Oostenbrink. Hamiltonian replica exchange molecular dynamics using soft-core interactions. *J. Chem. Phys.*, 128(14):144121, 2008.

[226] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.*, 11(2):224–230, 2001.

[227] P. E. A. da Silva, A. Von Groll, A. Martin, and J. C. Palomino. Efflux as a mechanism for drug resistance in *Mycobacterium tuberculosis*. *FEMS Immunol. Med. Microbiol.*, 63(1):1–9, 2011.

[228] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[229] S. G. Kini, A. R. Bhat, B. Bryant, J. S. Williamson, and F. E. Dayan. Synthesis, antitubercular activity and docking study of novel cyclic azole substituted diphenyl ether derivatives. *Eur. J. Med. Chem.*, 44(2):492–500, 2009.

[230] A. Kumar and M. I. Siddiqi. Receptor based 3D-QSAR to identify putative binders of *Mycobacterium tuberculosis* enoyl acyl carrier protein reductase. *J. Mol. Model.*, 16(5):877–893, 2010.

[231] A. Tripathi, N. Wadia, D. Bindal, and T. Jana. Docking studies on novel alkaloid tryptanthrin and its analogues against enoyl-acyl carrier protein reductase (InhA) of *Mycobacterium tuberculosis*. *Indian J. Biochem. Biophys.*, 49(6):435–441, 2012.

[232] S. D. Joshi, U. A. More, S. R. Dixit, D. Dubey, A. Tripathi, and H. Kulkarni Venkatrao. Discovering potent inhibitors against the enoyl-acyl carrier protein reductase (InhA) of *Mycobacterium tuberculosis*: Structure-based design, synthesis and antimicrobial activity of quinoline hydrazones. *Indo Am. J. Pharm. Res.*, 4(2):864–877, 2014.

[233] I. Pauli, R. N. dos Santos, D. C. Rostirolla, L. K. Martinelli, R. G. Ducati, L. F. Timmers, L. A. Basso, D. S. Santos, R. V. Guido, A. D. Andricopulo, et al. Discovery of new inhibitors of *Mycobacterium tuberculosis* InhA enzyme using virtual screening and a 3D-pharmacophore-based approach. *J. Chem. Inf. Model.*, 53(9):2390–2401, 2013.

[234] T. Kinjo, Y. Koseki, M. Kobayashi, A. Yamada, K. Morita, K. Yamaguchi, R. Tsurusawa, G. Gulten, H. Komatsu, H. Sakamoto, et al. Identification of compounds with potential antibacterial activity against *Mycobacterium* through structure-based drug screening. *J. Chem. Inf. Model.*, 53(5):1200–1212, 2013.

[235] G. Jose, T. S. Kumara, G. Nagendrappa, H. Sowmya, D. Sriram, P. Yogeeswari, J. P. Sridevi, T. N. G. Row, A. A. Hosamani, P. S. Ganapathy, et al. Synthesis, molecular docking and anti-mycobacterial evaluation of new imidazo [1, 2-a] pyridine-2-carboxamide derivatives. *Eur. J. Med. Chem.*, 89:616–627, 2015.

[236] A. L. Perryman, W. Yu, X. Wang, S. Ekins, S. Forli, S.-G. Li, J. S. Freundlich, P. J. Tonge, and A. J. Olson. A virtual screen discovers novel, fragment-sized inhibitors of *Mycobacterium tuberculosis* InhA. *J. Chem. Inf. Model.*, 55(3):645–659, 2015.

[237] H. Kanetaka, Y. Koseki, J. Taira, T. Umei, H. Komatsu, H. Sakamoto, G. Gulten, J. C. Sacchettini, M. Kitamura, and S. Aoki. Discovery of InhA inhibitors with anti-mycobacterial activity through a matched molecular pair approach. *Eur. J. Med. Chem.*, 94:378–385, 2015.

[238] M. Muddassar, J. W. Jang, H. S. Gon, Y. S. Cho, E. E. Kim, K. C. Keum, T. Oh, S. N. Cho, and A. N. Pae. Identification of novel antitubercular compounds through hybrid virtual screening approach. *Bioorg. Med. Chem.*, 18(18):6914–6921, 2010.

[239] M. Sgobba, F. Caporuscio, A. Anighoro, C. Portioli, and G. Rastelli. Application of a post-docking procedure based on MM-PBSA and MM-GBSA on single and multiple protein conformations. *Eur. J. Med. Chem.*, 58:431–440, 2012.

[240] J. A. Shilpi, M. T. Ali, S. Saha, S. Hasan, A. I. Gray, and V. Seidel. Molecular docking studies on InhA, MabA and PanK enzymes from *Mycobacterium tuberculosis* of ellagic acid derivatives from *Ludwigia adscendens* and *Trewia nudiflora. In Silico Pharmacol.*, 3(1):1, 2015.

[241] C. Menendez, A. Chollet, F. Rodriguez, C. Inard, M. R. Pasca, C. Lherbet, and M. Baltas. Chemical synthesis and biological evaluation of triazole derivatives as inhibitors of InhA and antituberculosis agents. *Eur. J. Med. Chem.*, 52:275–283, 2012.

[242] N. Huang, B. K. Shoichet, and J. J. Irwin. Benchmarking sets for molecular docking. *J. Med. Chem.*, 49(23):6789–6801, 2006.

[243] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.*, 55(14):6582–6594, 2012.

[244] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, and D. T. Mainz. Extra Precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.*, 49(21):6177–6196, 2006.

[245] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[246] 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. *J. Med. Chem.*, 47(7):1750–1759, 2004.

[247] T. A. Halgren. New Method for Fast and Accurate Binding-site Identification and Analysis. *Chem. Biol. Drug. Des.*, 69(2):146–148, 2007.

[248] T. A. Halgren. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.*, 49(2):377–389, 2009.

[249] Sitemap, version 3.0. *Schrödinger, LLC, New York, NY*, 2014.

[250] M. P. Jacobson, G. A. Kaminski, R. A. Friesner, and C. S. Rapp. Force field validation using protein side chain prediction. *J. Phys. Chem. B*, 106(44):11673–11680, 2002.

[251] M. P. Jacobson, R. A. Friesner, Z. Xiang, and B. Honig. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.*, 320(3):597–608, 2002.

[252] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. Day, B. Honig, D. E. Shaw, and R. A. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf.*, 55(2):351–367, 2004.

[253] Prime, version 4.5. *Schrödinger, LLC, New York, NY*, 2015.

[254] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105(28):6474–6487, 2001.

[255] E. Gallicchio, L. Y. Zhang, and R. M. Levy. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comput. Chem.*, 23(5):517–529, 2002.

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

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

[258] S. L. Parikh, G. Xiao, and P. J. Tonge. Inhibition of InhA, the enoyl reductase from *Mycobacterium tuberculosis*, by triclosan and isoniazid. *Biochemistry*, 39(26):7645–7650, 2000.

[259] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, and M. Müller. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.*, 12(1):1, 2011.

[260] H. Yabuuchi. *enrichvs: Enrichment assessment of virtual screening approaches*, 2011. R package version 0.0.5.

[261] 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.

[262] V. Kumar and M. E. Sobhia. Insights into the bonding pattern for characterizing the open and closed state of the substrate-binding loop in *Mycobacterium tuberculosis* InhA. *Futur. Med. Chem.*, 6(6):605–616, 2014.

[263] D. Case, V. Babin, J. Berryman, R. Betz, Q. Cai, D. Cerutti, T. Cheatham III, T. Darden, R. Duke, H. Gohlke, et al. Amber 14, 2014.

[264] Schrödinger. LLC. Maestro, protein preparation wizard. *New York*, 2012.

[265] M. H. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen. PROPKA 3: Consistent treatment of internal and surface residues in empirical pK_a predictions. *J. Chem. Theory Comput.*, 7(2):525–537, 2011.

[266] C. R. Søndergaard, M. H. Olsson, M. Rostkowski, and J. H. Jensen. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pK_a values. *J. Chem. Theory Comput.*, 7(7):2284–2295, 2011.

[267] Macromodel, version 9.9. *Schrödinger, LLC, New York, NY*, 2012.

[268] J. L. Banks, H. S. Beard, Y. Cao, A. E. Cho, W. Damm, R. Farid, A. K. Felts, T. A. Halgren, D. T. Mainz, J. R. Maple, et al. Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.*, 26(16):1752–1780, 2005.

[269] Tripos International, 1699 South Hanley Road, St. Louis, Missouri, 63144, USA. *SYBYL-X 1.0*, August 2009.

[270] Ligprep 2.5. *Schrödinger, LLC, New York NY*, 2012.

[271] T. Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

[272] P. Markt, D. Schuster, and T. Langer. Pharmacophore models for virtual screening. *Virtual Screening: Princ. Challenges, Pract. Guidel.*, pages 115–152, 2011.

[273] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

[274] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. ROCR: Visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005.

[275] H. Yabuuchi, S. Niijima, H. Takematsu, T. Ida, T. Hirokawa, T. Hara, T. Ogawa, Y. Minowa, G. Tsujimoto, and Y. Okuno. Analysis of multiple compound - protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.*, 7(1):472, 2011.

[276] Schrödinger LLC., The pymol molecular graphics system, version 1.8, 2015.

[277] H. Wickham. *ggplot2: elegant graphics for data analysis.* Springer Science & Business Media, 2009.

[278] D. R. Cox. The regression analysis of binary sequences. *J. R. Stat. Soc.*, 215–242, 1958.

[279] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[280] H. A. Carlson. Protein flexibility and drug design: How to hit a moving target. *Curr. Opin. Chem. Biol.*, 6(4):447–452, 2002.

[281] Y. Georgievskii, C. P. Hsu, and R. Marcus. Linear response in theory of electron transfer reactions as an alternative to the molecular harmonic oscillator model. *J. Chem. Phys.*, 110(11):5307–5317, 1999.

[282] M. Almlöf, B. O. Brandsdal, and J. Åqvist. Binding affinity prediction with different force fields: Examination of the linear interaction energy method. *J. Comput. Chem.*, 25(10):1242–1254, 2004.

[283] M. Linder, A. Ranganathan, and T. Brinck. "adapted Linear Interaction Energy": A Structure-Based LIE Parametrization for Fast Prediction of Protein–Ligand Affinities. *J. Chem. Theory Comput.*, 9(2):1230–1239, 2013.

[284] Y. Su, E. Gallicchio, K. Das, E. Arnold, and R. M. Levy. Linear interaction energy (LIE) models for ligand binding in implicit solvent: Theory and application to the binding of NNRTIs to HIV-1 reverse transcriptase. *J. Chem. Theory Comput.*, 3(1):256–277, 2007.

[285] T. Hansson, J. Marelius, and J. Åqvist. Ligand binding affinity prediction by linear interaction energy methods. *J. Comput. Aided Mol. Des.*, 12(1):27–35, 1998.

[286] G. King and A. Warshel. A surface constrained all-atom solvent model for effective simulations of polar solutions. *J. Chem. Phys.*, 91(6):3647–3661, 1989.

[287] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[288] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, 112(20):8910–8922, 2000.

[289] I. Jolliffe. *Principal component analysis.* Wiley Online Library, 2002.

[290] K. Peason. On lines and planes of closest fit to systems of point in space. *Philos. Mag.*, 2:559–572, 1901.

[291] 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. *Comput. Stat. Data Anal.*, 49(4):974–997, 2005.

[292] H. V. Henderson and P. F. Velleman. Building multiple regression models interactively. *Biometrics*, 391–411, 1981.

[293] S. Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[294] S. Menard. *Applied logistic regression analysis*, volume 106. Sage Publications Inc., 2002.

[295] D. A. Case, T. Darden, T. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. Walker, W. Zhang, K. Merz, et al. Amber 11. Technical report, University of California, 2010.

[296] M. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, J. Montgomery, T. Vreven, K. Kudin, J. Burant, et al. Gaussian 03, revision c. 02, 2008.

[297] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260, 2006.

[298] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Antechamber: An accessory software package for molecular mechanical calculations. *J. Am. Chem. Soc.*, 222:U403, 2001.

[299] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78(1):1–20, 1999.

[300] 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. *Theor. Chem. Acc.*, 101(6):426–434, 1999.

[301] M. Feig and C. L. Brooks. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.*, 14(2):217–224, 2004.

[302] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, 2005.

[303] R. Salomon-Ferrer, A. W. Goetz, D. Poole, S. Le Grand, and R. C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J. Chem. Theory Comput.*, 9(9):3878–3888, 2013.

[304] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual Molecular Dynamics. *J. Mol. Graph.*, 14(1):33–38, 1996.

[305] D. R. Roe and T. E. Cheatham III. Ptraj and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.*, 9(7):3084–3095, 2013.

[306] H. Gohlke and G. Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.*, 41(15):2644–2676, 2002.

[307] D. Sarkar. *Lattice: multivariate data visualization with R.* Springer Science & Business Media, 2008.

[308] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[309] J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar. R core team (2014) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-117, 2014.

[310] K. P. Burnham and D. R. Anderson. Multimodel inference understanding AIC and BIC in model selection. *Socio. Meth. Res.*, 33(2):261–304, 2004.

[311] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Science & Business Media, 2003.

[312] A. Perdih, G. Wolber, and T. Solmajer. Molecular dynamics simulation and linear interaction energy study of D-Glu-based inhibitors of the MurD ligase. *J. Comput. Aided Mol. Des.*, 27(8):723–738, 2013.

[313] L. Capoferri, M. C. Verkade-Vreeker, D. Buitenhuis, J. N. Commandeur, M. Pastor, N. P. Vermeulen, and D. P. Geerke. Linear interaction energy based prediction of cytochrome P450 1A2 binding affinities with reliability estimation. *PLoS One*, 10(11):e0142232, 2015.

[314] D. F. Lewis, S. Modi, and M. Dickins. Structure - activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metab. Rev.*, 34(1-2):69–82, 2002.

[315] M. L. Lamb, J. Tirado-Rives, and W. L. Jorgensen. Estimation of the binding affinities of FKBP12 inhibitors using a linear response method. *Bioorg. Med. Chem.*, 7(5):851–860, 1999.

[316] A. Amadei, A. Linssen, and H. J. Berendsen. Essential dynamics of proteins. *Proteins: Struct., Funct., Bioinf.*, 17(4):412–425, 1993.

[317] R. Levy, A. Srinivasan, W. Olson, and J. McCammon. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers*, 23(6):1099–1112, 1984.

[318] M. Karplus and J. N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.

[319] A. E. García. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, 68(17):2696, 1992.

[320] P. Hünenberger, A. Mark, and W. Van Gunsteren. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J. Mol. Biol.*, 252(4):492–503, 1995.

[321] K. Kasahara, I. Fukuda, and H. Nakamura. A Novel Approach of Dynamic Cross Correlation Analysis on Molecular Dynamics Simulations and Its Application to Ets1 Dimer–DNA Complex. *PLoS One*, 9(11):e112419, 2014.

[322] C.-T. Lai, H.-J. Li, W. Yu, S. Shah, G. R. Bommineni, V. Perrone, M. Garcia-Diaz, P. J. Tonge, and C. Simmerling. Rational Modulation of the Induced-Fit Conformational Change for Slow-Onset Inhibition in *Mycobacterium tuberculosis* InhA. *Biochemistry*, 54(30):4683–4691, 2015.

[323] SciFinder, version 2006; Chemical Abstracts Service: Columbus, OH, 2016.

[324] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1):141–151, 1999.

[325] E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, et al. Definition of the hydrogen bond (IUPAC Recommendations 2011). *Pure Appl. Chem.*, 83(8):1637–1641, 2011.

[326] I. Y. Torshin, I. T. Weber, and R. W. Harrison. Geometric criteria of hydrogen bonds in proteins and identification of bifurcated hydrogen bonds. *Protein Eng.*, 15(5):359–363, 2002.

[327] F. Milletti, L. Storchi, G. Sforna, and G. Cruciani. New and original pK$_a$ prediction method using grid molecular interaction fields. *J. Chem. Inf. Model.*, 47(6):2172–2181, 2007.

[328] R. M. Yates, A. Hermetter, G. A. Taylor, and D. G. Russell. Macrophage activation downregulates the degradative capacity of the phagosome. *Traffic*, 8(3):241–250, 2007.

[329] M. Grosso, A. Kalstein, G. Parisi, A. E. Roitberg, and S. Fernandez-Alberti. On the analysis and comparison of conformer-specific essential dynamics upon ligand binding to a protein. *J. Chem. Phys.*, 142(24):06B619_1, 2015.

[330] J. E. T. Pineda, R. Callender, and S. D. Schwartz. Ligand binding and protein dynamics in lactate dehydrogenase. *Biophys. J.*, 93(5):1474–1483, 2007.

[331] O. F. Lange and H. Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins: Struct., Funct., Bioinf.*, 62(4):1053–1061, 2006.

[332] A. Kitao and N. Go. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.*, 9(2):164–169, 1999.

[333] S. Hayward, A. Kitao, and N. Gō. Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. *Proteins: Struct., Funct., Bioinf.*, 23(2):177–186, 1995.

[334] J. McCammon. Protein dynamics. *Rep. Prog. Phys.*, 47(1):1, 1984.

[335] Q. Cui and I. Bahar. *Normal mode analysis: Theory and applications to biological and chemical systems.* CRC press, 2005.

[336] A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80(1):505–515, 2001.

[337] C. Chennubhotla, A. Rader, L. W. Yang, and I. Bahar. Elastic network models for understanding biomolecular machinery: From enzymes to supramolecular assemblies. *Phys. Biol.*, 2(4):S173, 2005.

[338] I. Daidone and A. Amadei. Essential dynamics: Foundation and applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(5):762–770, 2012.

[339] B. Sanjeev and S. Vishveshwara. Essential dynamics and sidechain hydrogen bond cluster studies on eosinophil cationic protein. *Eur. Phys. J. D*, 20(3):601–608, 2002.

[340] T. Ichiye and M. Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Struct., Funct., Bioinf.*, 11(3):205–217, 1991.

[341] T. M. Cover and J. A. Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

[342] A. Bakan, L. M. Meireles, and I. Bahar. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577, 2011.

[343] J. Ribas, E. Cubero, F. J. Luque, and M. Orozco. Theoretical study of alkyl-π and aryl-π interactions. reconciling theory and experiment. *J. Org. Chem.*, 67(20):7057–7065, 2002.

[344] M. J. Plevin, D. L. Bryce, and J. Boisbouvier. Direct detection of CH/π interactions in proteins. *Nat. Chem.*, 2(6):466–471, 2010.

[345] B. Brauer, M. K. Kesharwani, S. Kozuch, and J. M. Martin. The s66x8 benchmark for noncovalent interactions revisited: Explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.*, 2016.

[346] A. Kumar and M. I. Siddiqi. CoMFA based de novo design of pyrrolidine carboxamides as inhibitors of enoyl acyl carrier protein reductase from *Mycobacterium tuberculosis*. *J. Mol. Model.*, 14(10):923–935, 2008.

[347] W. Jorgensen. QikProp, version 3.0. *Schrodinger, LLC: New York*, 2006.

[348] C. W. Yap. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, 32(7):1466–1474, 2011.

[349] L. M. Salonen, M. Ellermann, and F. Diederich. Aromatic rings in chemical and biological recognition: Energetics and structures. *Angew. Chem. Int. Ed.*, 50(21):4808–4842, 2011.

[350] B. C. Doak, B. Over, F. Giordanetto, and J. Kihlberg. Oral druggable space beyond the rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.*, 21(9):1115–1142, 2014.

[351] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[352] A. E. Torda and W. F. van Gunsteren. Algorithms for clustering molecular dynamics configurations. *J. Comput. Chem.*, 15(12):1331–1340, 1994.

[353] A. Wolf and K. N. Kirschner. Principal component and clustering analysis on molecular dynamics data of the ribosomal L11·23S subdomain. *J. Mol. Model.*, 19(2):539–549, 2013.

[354] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, 2001.

[355] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Commun. Stat.*, 3(1):1–27, 1974.

[356] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227, 1979.

[357] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[358] Z. Deng, C. Chuaqui, and J. Singh. Structural interaction fingerprint (SiFT): A novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.*, 47(2):337–344, 2004.

[359] A. Schreyer and T. Blundell. CREDO: A Protein–Ligand Interaction Database for Drug Discovery. *Chem. Biol. Drug. Des.*, 73(2):157–167, 2009.

[360] A. M. Schreyer and T. L. Blundell. CREDO: A Structural Interactomics Database For Drug Discovery. *Database*, 2013:bat049, 2013.

[361] T. P. Tauer, M. E. Derrick, and C. D. Sherrill. Estimates of the ab Initio Limit for Sulfur- π Interactions: The H_2S - Benzene dimer. *J. Phys. Chem. A*, 109(1):191–196, 2005.

[362] A. L. Ringer, A. Senenko, and C. D. Sherrill. Models of S/π interactions in protein structures: Comparison of the H_2S - Benzene complex with PDB data. *Protein Sci.*, 16(10):2216–2223, 2007.

[363] B. R. Beno, K. S. Yeung, M. D. Bartberger, L. D. Pennington, and N. A. Meanwell. A Survey of the Role of Noncovalent Sulfur Interactions in Drug Design. *J. Med. Chem.*, 58(11):4383–4438, 2015.

[364] H. C. Jubb, A. P. Higueruelo, B. Ochoa-Montaño, W. R. Pitt, D. B. Ascher, and T. L. Blundell. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.*, 429(3):365–371, 2017.

[365] C. S. Leung, S. S. Leung, J. Tirado-Rives, and W. L. Jorgensen. Methyl effects on protein - ligand binding. *J. Med. Chem.*, 55(9):4489, 2012.

[366] D. Toummini, A. Tlili, J. Bergès, F. Ouazzani, and M. Taillefer. Copper-Catalyzed Arylation of Nitrogen Heterocycles from Anilines under Ligand-Free Conditions. *Chem. Eur. J.*, 20(45):14619–14623, 2014.

[367] N. Miyaura and A. Suzuki. Palladium-catalyzed cross-coupling reactions of organoboron compounds. *Chem. Rev.*, 95(7):2457–2483, 1995.

List of Figures

List of Tables

Abbreviations

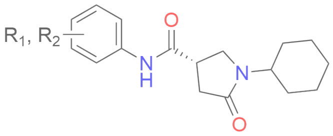
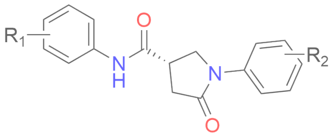
ACP	Acyl-Carrier Protein
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
amu	atomic mass unit
AUC	Area under the curve
CHC	Calinski-Harabasz criterion
CNS	Central Nervous System
DHF	Dihydrofolate
DNA	Deoxyribonucleic acid
DPE	Diphenyl ethers
ENR	Enoyl ACP reductase
FAS II	Fatty Acid Synthesis II
FEP	Free Energy Perturbation
GBIS	Generalized Born Implicit Solvent
IQR	Inter-quartile range
IFD	Induced Fit Docking
LIE	Linear Interaction Energy
MD	Molecular Dynamics
MDR-TB	Multidrug-resistant *Mycobacterium tuberculosis*
MM/GBSA	Molecular Mechanics-Generalized Born Surface Area
MM/PBSA	Molecular Mechanics-Poisson Boltzmann Surface Area
MRSA	Methicillin-resistant *Staphylococcus aureus*
Mtb	*Mycobacterium tuberculosis*
NMR	Nuclear Magnetic Resonance
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
RAMD	Random Accelerated Molecular Dynamics

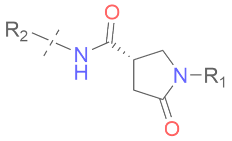
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuation
RNA	Ribonucleic acid
ROC	Receiver Operating Characteristics
SBL	Substrate Binding Loop
SBDD	Structure Based Drug Design
SMD	Steered Molecular Dynamics
TB	Tuberculosis
THF	Tetrahydrofolate
VRSA	Vancomycin-resistant *Staphylococcus aureus*
XDR-TB	Extensively drug-resistant *Mycobacterium tuberculosis*

Appendix A

Inhibitors of InhA derived from literature used for docking and affinity prediction

Table A.1 Overview of the different scaffolds investigated by molecular docking in the current work.

	**"light" pyrrolidine carboxamides**
	**C-ring replaced pyrrolidine carboxamides**


R1 = C6-C8 cycloalkanes
R2 = multiple (hetero-)cycles
	A-ring replaced/"bulky" pyrrolidine carboxamides

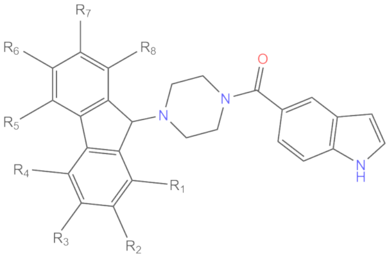
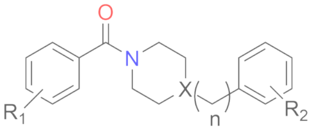
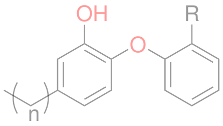
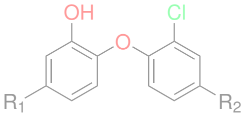
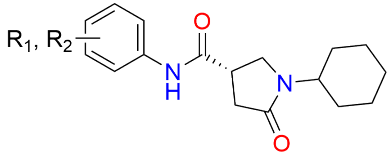
 R$_7$ R$_6$ R$_8$ R$_5$ N N O N H R$_4$ R$_1$ R$_3$ R$_2$	**genzyme series**
	**arylamide series**
	**diphenyl ether series**
	**triclosan based diphenyl ether series**

Table A.2 Protein from which top ranked pose was selected, experimental activity, XPscore, Substructure RMSD, C-α and bound ligand RMSD for top ranked poses from light pyrrolidine carboxamide subset



"Light" pyrrolidine carboxamide subset

Protein	Ligand	R_1	R_2	pIC_{50}	Poses (N) generated (with)	XPscore (kcal/mol)	RMSD (Å)			Used in LIE (Y/N)	Reason
							Substr.	C-α	Bound lig.		
4TZK	s1 (4U0J)∓	H	H	4.97	1	-9.30	0.69	1.44	0.32	N	NC*
4TZK	s2	H	2-$COOCH_3$	4.45	5 (XP)	-9.29	0.74	1.27	0.34	Y	C**
4TZK	s4 (4TRJ)∓	H	3-Br	6.05	1	-9.60	0.38	1.30	0.29	Y	C
4TZK	s5	H	4-Br	4.55	5 (XP)	-8.77	0.91	1.21	0.69	Y	C
4TZK	s6‡	H	3-Cl	5.86	1 (mut.)	-9.64	0.40	1.16	0.67	Y	C
4TZK	s10‡	H	3-CH_3	4.77	1 (mut.)	-9.48	0.39	1.09	0.34	Y	C

continued on next page

∓ Crystal structure ligands whose XPscore is actually scored **"in-place"** using XPscore after preparing the updated PDB codes.

***NC** stands for "non-converged". This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the "stability" criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).

N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.

****C** stands for "converged" which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

‡ Molecules whose orientations were derived by mutating pc-d11 (PDB 4TZK ligand) to their structure followed by restrained minimisation and "in-place" rescoring with XPscore.

Protein	**Ligand**	$\mathbf{R_1}$	$\mathbf{R_2}$	$\mathbf{pIC_{50}}$	**Poses (N) generated (with)**	**XPscore (kcal/mol)**	**RMSD (Å)**			**Used in LIE (Y/N)**	**Reason**
							Substr.	**C-α**	**Bound lig.**		
4TZK	s11[‡]	H	3-CF_3	5.45	1 (mut.)	-10.34	0.38	0.96	0.74	Y	C
4TZK	s12[‡]	H	3-NO_2	4.97	1 (mut.)	-8.92	0.40	1.16	0.55	Y	C
4TZK	s15[‡]	H	3-$CH(CH_3)_2$	5.25	1 (mut.)	-9.85	0.47	1.30	0.75	Y	C
4TZK	s17	H	4-NO_2	4.13	5 (XP)	-9.43	0.87	1.29	1.06	N	NC
4TZK	d1[‡]	2-Cl	4-Cl	4.25	1 (mut.)	-9.67	0.56	1.61	0.92	N	NC
4TZK	d2	2-Cl	5-Cl	4.24	5 (XP)	-9.87	0.49	1.13	0.35	Y	C
4TZK	d3 (4U0K)[∓]	2-CH_3	5-Cl	6.01	1	-9.91	0.42	1.46	0.54	N	NC
4TZK	d4[‡]	3-CH_3	5-Cl	4.43	1 (mut.)	-9.48	0.47	1.17	0.45	Y	C
4TZK	d6	2-CH_3	5-CH_3	4.99	5 (XP)	-9.79	0.39	1.31	0.67	N	NC
4TZK	d7	3-CH_3	5-CH_3	5.50	5 (XP)	-9.89	0.58	1.79	0.32	N	NC
4TZK	d8 (4TZT)[∓]	2-CH_3	3-Cl	4.63	1	-9.81	0.52	1.16	0.39	Y	C
4TZK	d9[‡]	2-CH_3	4-NO_2	4.50	1 (mut.)	-8.78	0.57	1.15	0.61	Y	C
4TZK	d10	3-F	5-F	5.82	5 (XP)	-10.10	0.51	1.44	0.57	N	NC
4TZK	d11 (4TZK)[∓]	3-Cl	5-Cl	6.40	1	-10.09	0.37	1.11	0.49	Y	C
4TZK	d12	3-Br	5-CF_3	6.07	5 (XP)	-10.73	0.38	1.15	0.78	Y	C
4TZK	d13	3-OCH_3	5-CF_3	5.88	5 (XP)	-9.90	0.65	1.22	0.81	Y	C
4TZK	d14[‡]	2-CF_3	5-CF_3	5.44	1 (mut.)	-10.76	0.38	1.09	0.83	Y	C

continued on next page

[∓] Crystal structure ligands whose XPscore is actually scored **”in-place”** using XPscore after preparing the updated PDB codes.

[*]**NC** stands for ”non-converged”. This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the ”stability” criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).

N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.

[**]**C** stands for ”converged” which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

[‡] Molecules whose orientations were derived by mutating pc-d11 (PDB 4TZK ligand) to their structure followed by restrained minimisation and ”in-place” rescoring with XPscore.

Protein	Ligand	R_1	R_2	pIC_{50}	Poses (N)	XPscore	RMSD (Å)			Used in LIE	Reason
					generated (with)	(kcal/mol)	Substr.	C-α	Bound lig.	(Y/N)	
4TZK	d15[‡]	2-OCH_3	5-Cl	5.79	1 (mut.)	-8.31	0.69	1.21	0.53	Y	C
4TZK	d16	3-Cl	4-F	4.82	5 (XP)	-9.96	0.81	1.21	0.47	Y	C

[∓] Crystal structure ligands whose XPscore is actually scored **”in-place”** using XPscore after preparing the updated PDB codes.

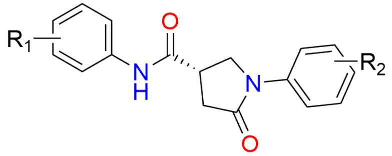
NC stands for ”non-converged”. This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the ”stability” criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).

N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.

C stands for ”converged” which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

[‡] Molecules whose orientations were derived by mutating pc-d11 (PDB 4TZK ligand) to their structure followed by restrained minimisation and ”in-place” rescoring with XPscore.

Table A.3 Protein from which pose was selected, experimental activity, XPscore, C-α RMSD, bound ligand RMSD for top ranked poses of ring C modified "light" pyrrolidine carboxamides derived from GlideXP docking



C-ring replaced pyrrolidine carboxamides

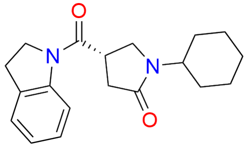
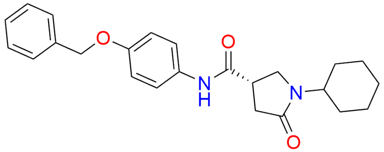
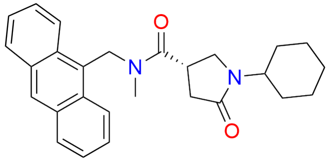
Protein	Pyrrolidine carboxamide	R_1	R_2	pIC$_{50}$	Poses (N) generated (with)	XPscore (kcal/mol)	RMSD (Å)			Used in LIE (Y/N)	Reason
							Substr.	C-alpha	Bound lig.		
4TZK	3a	3-Cl	H	5.40	5 (XP)	-9.31	0.55	1.39	0.82	N	NC
4TZK	3i	3-Br	H	4.86	5 (XP)	-9.27	0.71	1.26	0.75	Y	C
4TZK	3j	3-Br	4-F	4.53	5 (XP)	-8.87	0.89	1.36	0.88	N	NC

NC stands for "non-converged". This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the "stability" criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).

N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.

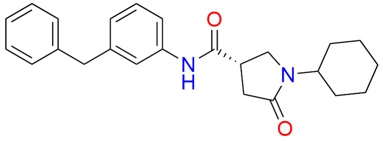
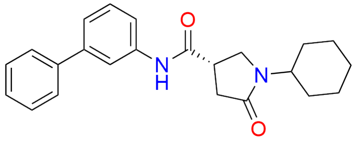
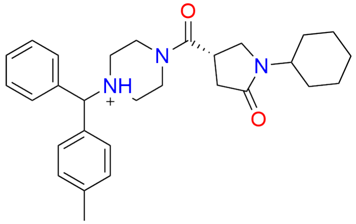
C stands for "converged" which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

Table A.4 Protein from which pose was selected, experimental activity, XPscore, C-α RMSD, bound ligand RMSD for top ranked poses of bulky pyrrolidine carboxamides (A-ring replaced) derived from Induced fit/Glide docking; IFD stands for induced fit, IFDT for induced fit with trimmed side chains, SP stands for GlideSP.

Protein	**Pyrrolidine**		**pIC$_{50}$**	**Poses (N)**	**XPscore**		**RMSD (Å)**		**Used in LIE**	**Reason**
	carboxamide			**generated (with)**	**(kcal/mol)**	**Substr.**	**C-alpha**	**Bound lig.**	**(Y/N)**	
2X23	r7		5.29	20 (IFD)	-9.34	2.21	1.22	0.71	Y	C
2NSD	p9		5.46	20 (IFD)	-10.79	1.56	1.41	0.98	N	NC
2NSD	p20		6.12	20 (IFD)	-12.37	1.40	1.59	1.41	N	NC

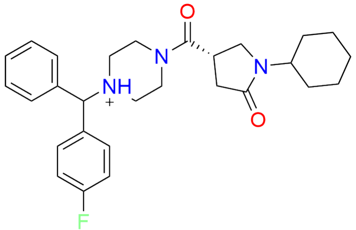
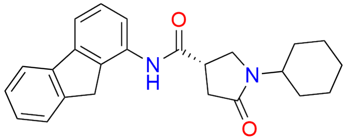
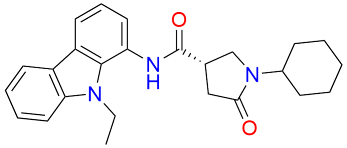
continued on the next page

NC stands for ”non-converged”. This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the ”stability” criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).
N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.
C stands for ”converged” which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

Protein		Pyrrolidine	pIC_{50}	Poses (N)	XPscore		RMSD (Å)		Used in LIE	Reason
		carboxamide		generated (with)	(kcal/mol)	Substr.	C-alpha	Bound lig.	(Y/N)	
2NSD	p21		6.39	20 (IFDT)	-5.72	3.59	1.53	1.83	N	NC
2NSD	p24		6.41	20 (IFDT)	-8.67	1.42	1.48	1.19	N	NC
2NSD	p27		5.28	20 (IFDT)	-9.93	3.16	-	-	N	NC

continued on the next page

NC stands for "non-converged". This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the "stability" criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).
N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.
C stands for "converged" which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

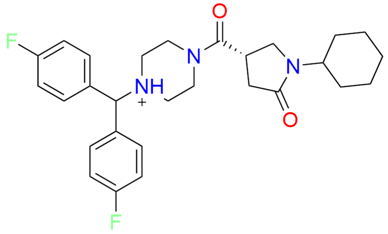
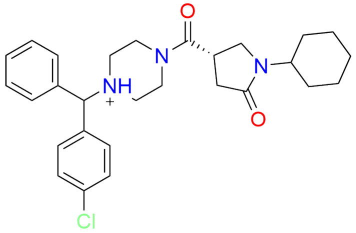
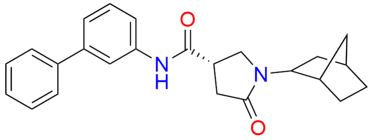
Protein	Pyrrolidine carboxamide		pIC_{50}	Poses (N) generated (with)	XPscore (kcal/mol)	RMSD (Å)			Used in LIE (Y/N)	Reason
						Substr.	C-alpha	Bound lig.		
2NSD	p28		5.13	20 (IFDT)	-7.7	2.76	1.74	1.13	N	NC
2X23	p31		5.86	20 (IFDT)	-10.79	1.92	1.12	0.75	Y	C
2NSD	p33		5.59	20 (IFDT)	-5.25	3.07	1.75	1.64	N	NC

continued on the next page

NC stands for "non-converged". This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the "stability" criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).

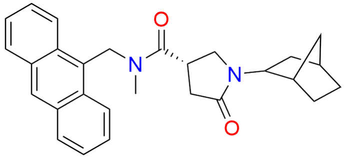
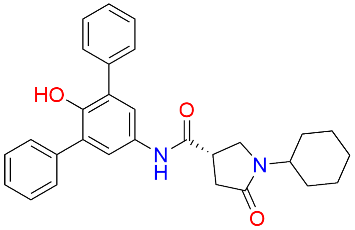
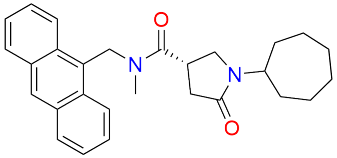
N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.

C stands for "converged" which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

Protein	Pyrrolidine		pIC_{50}	Poses (N)	XPscore		RMSD (Å)		Used in LIE	Reason
	carboxamide			generated (with)	(kcal/mol)	Substr.	C-alpha	Bound lig.	(Y/N)	
2NSD	p36		5.25	20 (IFDT)	-8.81	3.15	1.62	1.16	N	NC
2NSD	p37		5.34	20 (IFD)	-8.80	3.03	-	-	N	NC
2X23	c1a1		6.33	20 (IFDT)	-11.11	3.13	1.36	0.69	N	NC

continued on the next page

NC stands for ”non-converged”. This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the ”stability” criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).
N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.
C stands for ”converged” which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

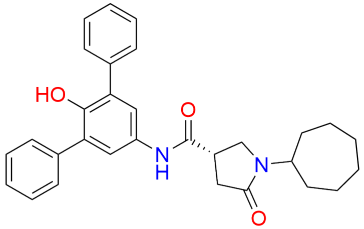
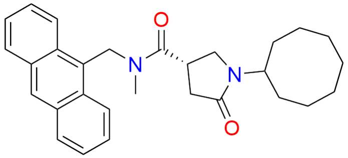
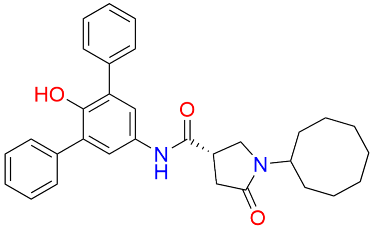
Protein	**Pyrrolidine carboxamide**		**pIC$_{50}$**	**Poses (N) generated (with)**	**XPscore (kcal/mol)**	**RMSD (Å)**			**Used in LIE (Y/N)**	**Reason**
						Substr.	**C-alpha**	**Bound lig.**		
2NSD	c1a2		6.07	20 (IFD)	-6.26	1.37	1.52	0.71	N	NC
2H7M	c6a3		6.85	20 (SP)	-5.17	6.46	1.30	1.50	N	NC
2NSD	c7a2		6.49	20 (IFDT)	-12.06	1.89	1.44	1.18	N	NC

continued on the next page

NC stands for "non-converged". This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the "stability" criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).

N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.

C stands for "converged" which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

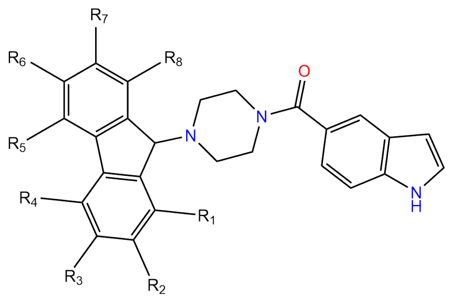
Protein		**Pyrrolidine carboxamide**	**pIC$_{50}$**	**Poses (N) generated (with)**	**XPscore (kcal/mol)**	**Substr.**	**RMSD (Å) C-alpha**	**Bound lig.**	**Used in LIE (Y/N)**	**Reason**
2H7M	c7a3		6.56	20 (IFDT)	-9.18	1.07	1.22	0.81	Y	C
2X23	c8a2		6.20	20 (IFDT)	-12.69	1.62	1.29	1.00	Y	C
2NSD	c8a3		5.88	20 (IFD)	-7.83	1.03	1.47	1.19	N	NC

NC stands for ”non-converged”. This simply means that the pose exhibited substantial mobility during MD simulations and did not satisfy the ”stability” criterion (C-α RMSD $\leq$ 1.3Å; Bound ligand RMSD $\leq$ 1Å).

N stands for not used in LIE model derivation, while Y denotes that the ligand was used in LIE model generation.

C stands for ”converged” which implies that the pose exhibited minimal mobility during MD simulations and thereby stable binding. Also used in LIE model derivation.

Table A.5 Genzyme series of InhA inhibitors.



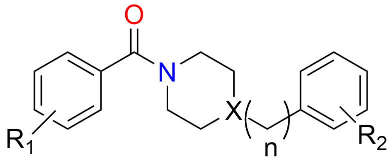
Genzyme series scaffold

Protein	Genzyme	R	pIC_{50}	Poses (N) generated (with)	XPscore (kcal/mol)	RMSD (Å)
1P44	10850	H	6.79	5 (XP)	-11.14	0.46
1P44	11918	1-CO_2CH_3	6.46	5 (XP)	-12.23	0.52
1P44	12638	2,7-I	6.88	5 (XP)	-12.79	0.61
1P44	12639	2-NHAc	6.55	5 (XP)	-12.80	0.53
1P44	12640	2-NMe_2	6.04	5 (XP)	-12.31	0.59
1P44	12641	2-NHCHO	6.74	5 (XP)	-12.01	0.51
1P44	12643	2,4,7-Cl	6.76	5 (XP)	-13.13	0.46
1P44	12644	2-NO_2	6.88	5 (XP)	-12.09	0.58
1P44	12645	3-NO_2	6.88	5 (XP)	-11.08	1.06

continued on the next page

Protein	**Genzyme**	**R**	**pIC$_{50}$**	**Poses (N)**	**generated (with)**	**XPscore (kcal/mol)**	**RMSD (Å)**
1P44	12646	2,7-Br	6.92	5	(XP)	-13.18	0.57
1P44	13100	2-NHCOBu	6.22	5	(XP)	-13.34	0.54
1P44	13108	2-NHCOPr	6.33	5	(XP)	-13.55	0.56
1P44	13348	2-OCH$_3$	6.28	5	(XP)	-12.46	0.57
1P44	13349	4-OCH$_3$	6.08	5	(XP)	-11.23	0.54

Table A.6 Arylamide series of InhA inhibitors



Arylamide scaffolds

Protein	**Ligand**	***X***	**n**	**R$_1$**	**R$_2$**	**pIC$_{50}$**	**Poses (N)**		**XPscore**	**RMSD**
							generated (with)		**(kcal/mol)**	**(Å)**
2NSD	a1	N	0	H	H	4.41	5	(XP)	-9.70	1.23
2NSD	a2	N	0	4-CH$_3$	H	4.77	5	(XP)	-9.27	1.29
2NSD	a3	N	0	4-CH$_3$	3-CF$_3$	5.20	5	(XP)	-11.84	1.48
2NSD	a4	N	0	4-CH$_3$	3-Cl	5.51	5	(XP)	-9.01	1.18
2NSD	a5	N	0	3-CH$_3$	3-Cl	5.02	5	(XP)	-10.42	1.24
2NSD	a6	N	0	3-CH$_3$	4-NO$_2$	4.81	5	(XP)	-7.88	1.19
2NSD	a7	N	0	3,4-Me$_2$	3-Cl	6.00	5	(XP)	-9.83	1.20
2NSD	a8	N	0	3,4-Me$_2$	3-CF$_3$	5.73	5	(XP)	-11.53	1.29
2NSD	a13	N	0	2-F	3-Cl	4.85	5	(XP)	-10.88	1.24
2NSD	a14	N	0	4-F	3-Cl	5.01	5	(XP)	-10.55	1.26
2NSD	a15	N	0	3-Cl	3-Cl	5.17	5	(XP)	-10.41	1.35
2NSD	a16	N	0	3,4-Cl	3-Cl	5.21	5	(XP)	-10.5	1.23
2NSD	a17	N	0	3,4-Cl	H	4.75	5	(XP)	-9.83	1.22
2NSD	a18	N	1	H	H	4.50	5	(XP)	-10.73	0.88
2NSD	b1	C	1	3-Cl	H	5.11	5	(XP)	-11.08	0.81
2NSD	b2	C	1	2-F	H	4.85	5	(XP)	-11.28	0.89
2NSD	b3	C	1	4-CH$_3$	H	5.28	5	(XP)	-9.26	0.68
2NSD	b4	C	1	3-CH$_3$	H	5.13	5	(XP)	-11.17	0.81

Table A.7 XPscore and substructure RMSD values for top ranked poses of arylamides identified by means of micro-titer synthesis and in-situ screening

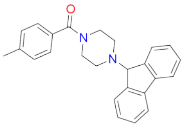
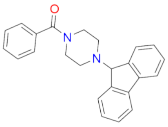
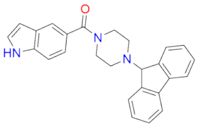
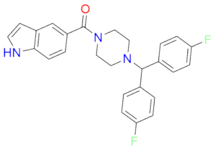
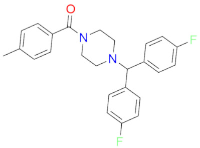
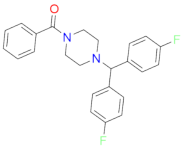
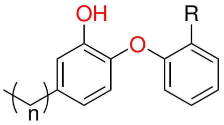
Protein	**Ligand**	**pIC$_{50}$**	**Poses (N) generated (with)**	**XPscore (kcal/mol)**	**RMSD (Å)**
2NSD	p1 	6.39	5 (XP)	-12.66	1.04
2NSD	p2 	7.04	5 (XP)	-12.74	1.09
2NSD	p3 	6.69	5 (XP)	-13.35	1.15
2NSD	p4 	5.98	5 (XP)	-14.21	0.94
2NSD	p5 	5.72	5 (XP)	-12.70	0.93
2NSD	p6 	5.69	5 (XP)	-13.78	0.96

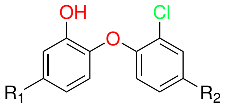
Table A.8 GlideXP Docking and substructure RMSD values for top ranked poses of n-alkyl-diphenyl ether series as Mtb-InhA inhibitors



Diphenyl ether series scaffold

Protein	**Ligand**	**N**	**R-group**	**pIC$_{50}$**	**Poses (N) generated (with)**	**XPscore (kcal/mol)**	**RMSD (Å)**
2X23	TCL	NA	NA	6.69	5 (XP)	-7.39	0.19
2X23	2PP	2	H	5.69	5 (XP)	-8.62	0.58
2X23	4PP	4	H	7.09	5 (XP)	-9.99	0.75
2X23	5PP	5	H	7.76	5 (XP)	-10.86	0.49
2X23	6PP	6	H	7.95	5 (XP)	-10.58	0.52
2X23	8PP	8	H	8.30	5 (XP)	-10.91	0.78
2X23	14PP	14	H	6.82	5 (XP)	-13.39	0.74
2X23	PT-70	6	CH$_3$	8.28	5 (XP)	-10.99	0.82

Table A.9 Triclosan based diphenyl ether series of InhA inhibitors.



Triclosan based diphenyl ether series scaffold

Protein	Ligand	R_1	R_2	pIC_{50}	Poses (N) generated (with)	XPscore (kcal/mol)	RMSD (Å)
2X23	2	CH_3	Cl	6.09	5 (XP)	-8.10	1.02
2X23	7	$CH_2(C_6H_{11})$	Cl	6.95	5 (XP)	-12.26	0.70
2X23	7-mod	$CH_2(C_6H_{10})$	Cl	6.95	5 (XP)	-12.32	0.70
2X23	8	CH_2CH_3	Cl	6.92	5 (XP)	-9.46	0.71
2X23	9	$(CH_2)_2CH_3$	Cl	7.04	5 (XP)	-9.01	1.01
2X23	10	$(CH_2)_3CH_3$	Cl	7.25	5 (XP)	-10.44	0.78
2X23	11	$CH_2CH(CH_3)_2$	Cl	7.01	5 (XP)	-10.59	0.70
2X23	12	$(CH_2)_3CH(CH_3)_2$	Cl	7.20	5 (XP)	-10.85	0.24
2X23	13	$CH_2CH(CH_3)CH_2CH_3$	Cl	6.88	5 (XP)	-11.01	0.55
2X23	17	CH_2(2-pyridyl)	Cl	7.53	5 (XP)	-11.62	0.70
2X23	18	CH_2(3-pyridyl)	Cl	7.37	5 (XP)	-10.94	0.73
4TZK	19	CH_2(4-pyridyl)	CN	7.12	5 (XP)	-11.52	0.80
2NSD	20	o-CH_3-Ph	Cl	5.88	5 (XP)	-11.08	1.31
2NSD	22	m-CH_3-Ph	Cl	6.06	5 (XP)	-11.66	1.49
2X23	24	CH_2-Ph	Cl	7.29	5 (XP)	-10.51	0.71
2X23	25	$(CH_2)_2$-Ph	Cl	7.67	5 (XP)	-11.90	0.40
2X23	26	$(CH_2)_3$-Ph	Cl	7.30	5 (XP)	-12.66	0.77

Table A.10 Mean number of poses per compound getting selected post pose selection phase.

Compound	Protein			
	1P44	2H7M	2NSD	2X23
Genzyme	2.5	3	2.5	1.5
Carboxamides	3.1	3.4	2.9	3.1
Arylamides	1.5	1.5	1.4	1.5
Diphenyl ethers	2.3	2.8	2.5	2.8

Table A.11 Number of poses docked and not docked across all four proteins

Protein	Molecules Docked (N)	Molecules not docked (N)	% Success
1P44	112	1	99.11
2H7M	109	4	96.46
2NSD	108	5	95.57
2X23	97	16	85.84

A.0.1 Logistic Regression models for entire pyrrolidine carboxamide dataset:

This section describes the logistic regression models generated using poses for the entire pyrrolidine carboxamide dataset in an activity-based separation endeavour.

A.0.1.1 "MOD" binomial *logreg* models

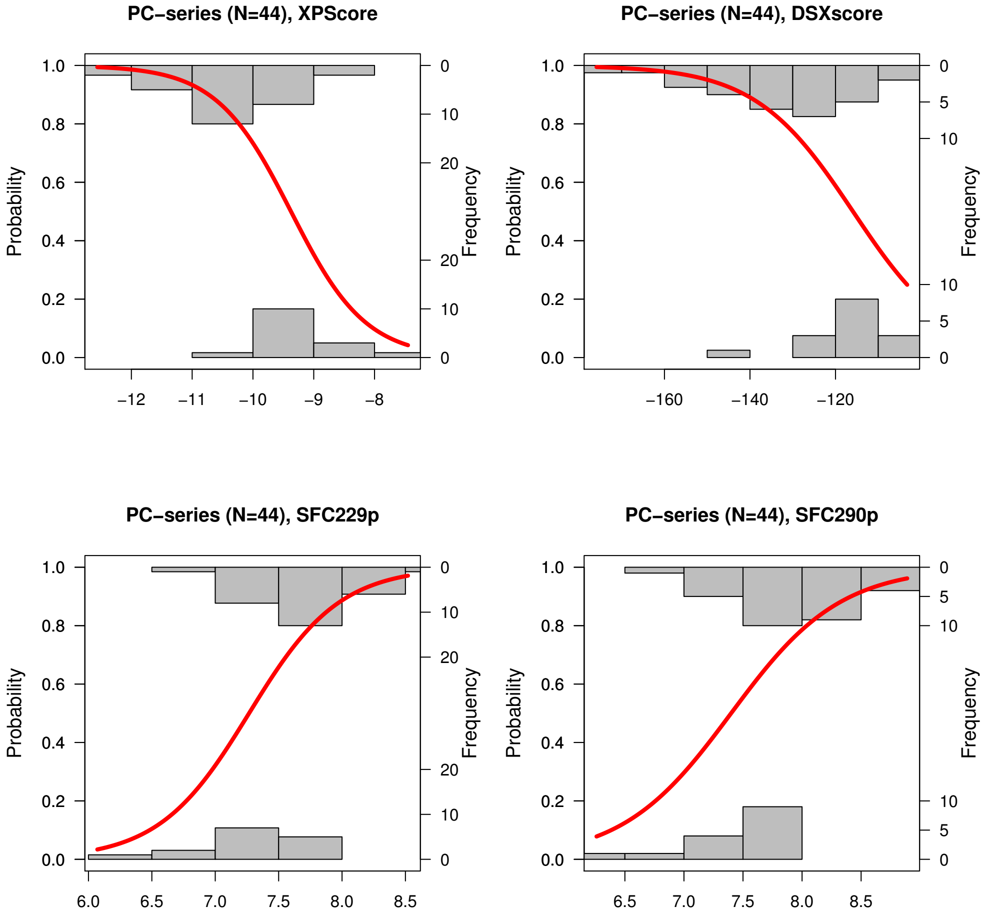


Figure A.1 Binomial logistic regression models of a) XPscore (top left), b) DrugscoreX (top right), c) SFC229p (bottom left), and d) SFC290p (bottom right) to detect moderately active compounds [for entire pyrrolidine carboxamide dataset (N=44)].

A.0.1.2 "HIGH" binomial *logreg* models

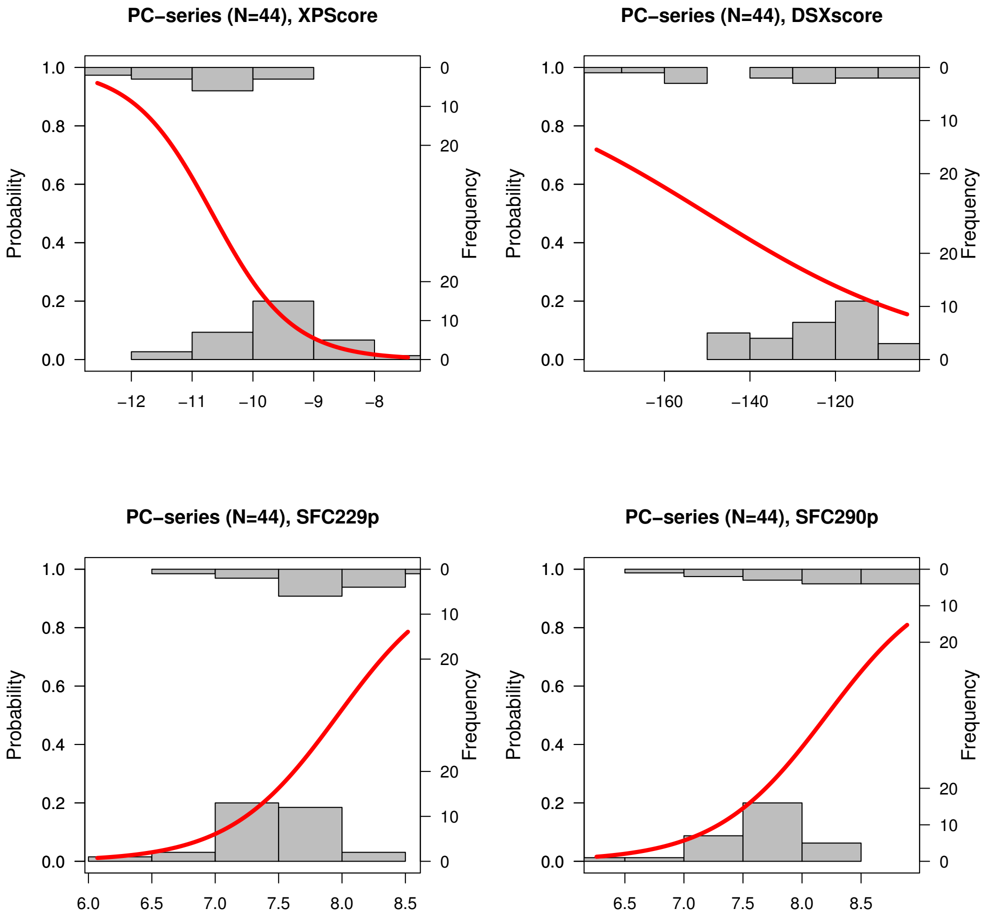


Figure A.2 Binomial logistic regression models of a) XPscore (top left), b) DrugscoreX (top right), c) SFC229p (bottom left), and d) SFC290p (bottom right) to detect highly active compounds [for entire pyrrolidine carboxamide dataset (N=44)].

Appendix B

Activity Based Separation models

B.1 Logistic regression models

The following are the miscellaneous *logreg* models derived from scoring functions as well as force field terms.

B.1.1 ”Scoring functions” based *logreg* models

B.1.1.1 ”MOD” binomial *logreg* models

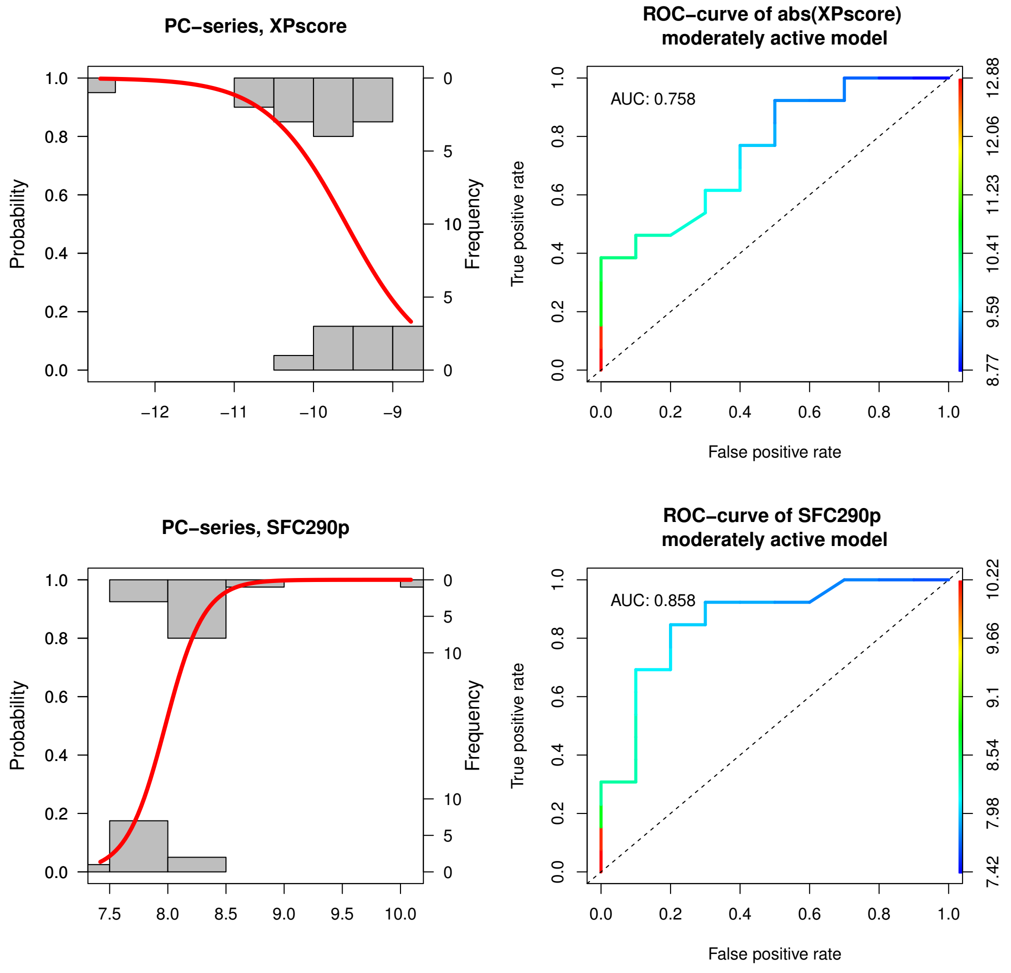


Figure B.1 Binomial logistic regression models of a) XPscore (top) and b) SFC290p (bottom) to detect moderately active compounds [for ”final” dataset (N=23)]

B.1.1.2 ”HIGH” binomial *logreg* models

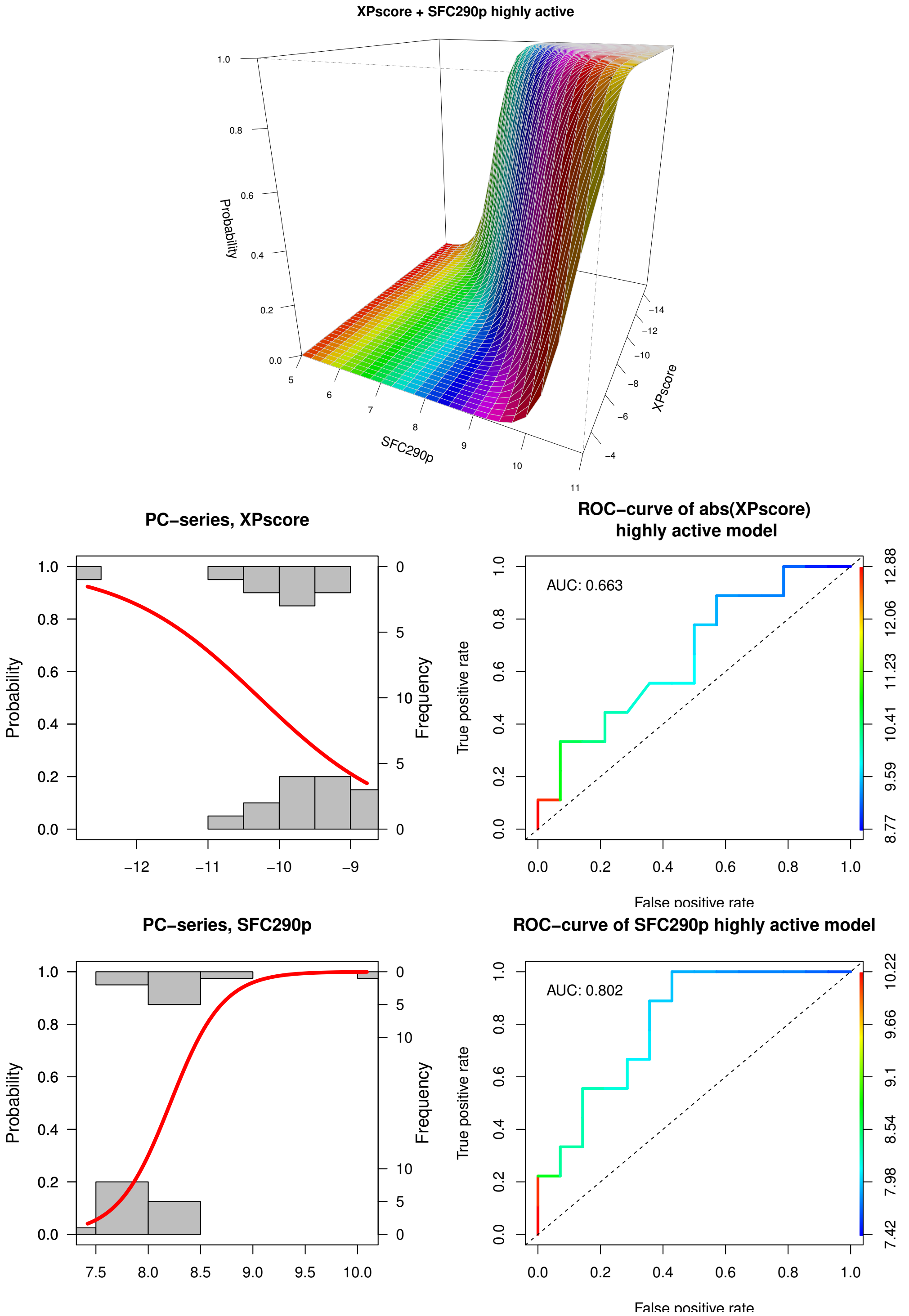


Figure B.2 Binomial logistic regression model of a) XPscore-SFC290p combination (top); b) XPscore (middle), and c) SFC290p (bottom) to detect highly active compounds [for ”final” dataset (N=23)]

B.1.1.3 Multinomial *logreg* models

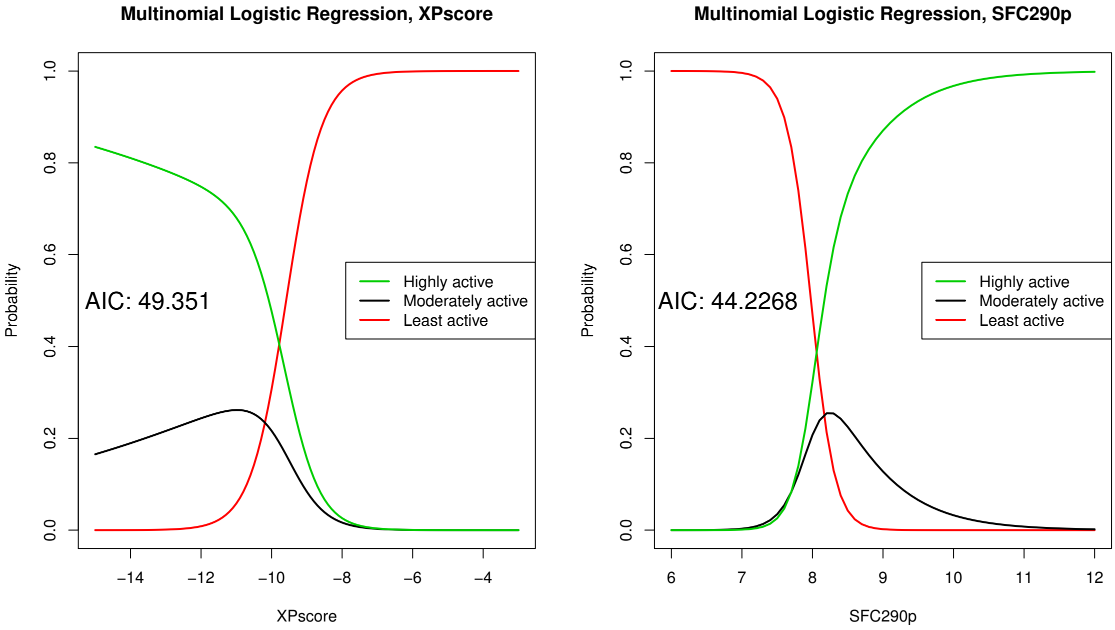


Figure B.3 Multinomial logistic regression models of XPscore and SFC290p to classify compounds as least active, moderately active, and highly active.

B.1.2 ”Force field” terms based *logreg* models

B.1.2.1 ”MOD” binomial *logreg* models

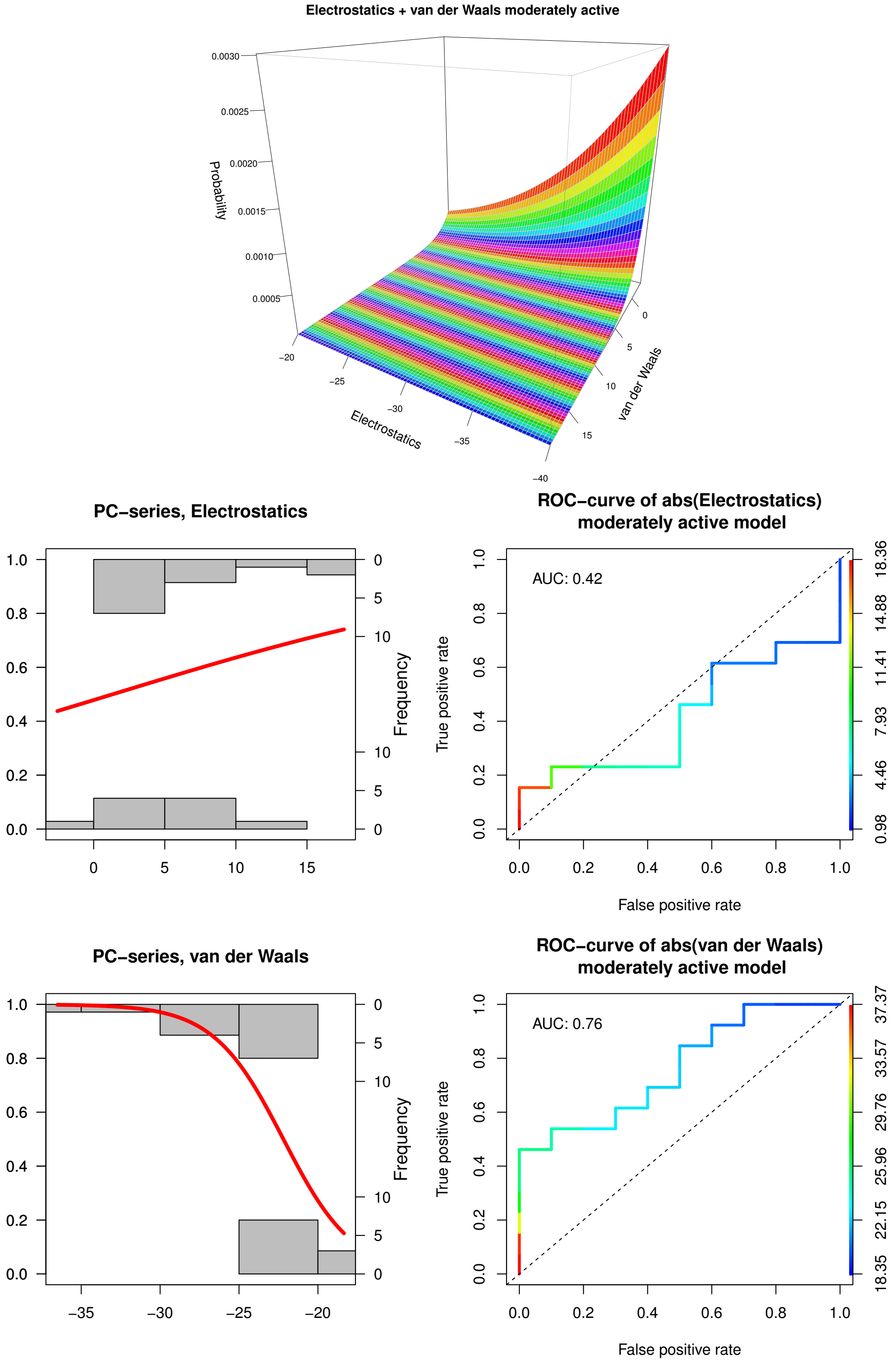


Figure B.4 Binomial logistic regression models for moderately active molecules recognition for: electrostatics and van der Waals combined; electrostatics alone; and van der Waals alone; derived using ”final” dataset (N=23)

B.1.2.2 ”HIGH” binomial *logreg* models

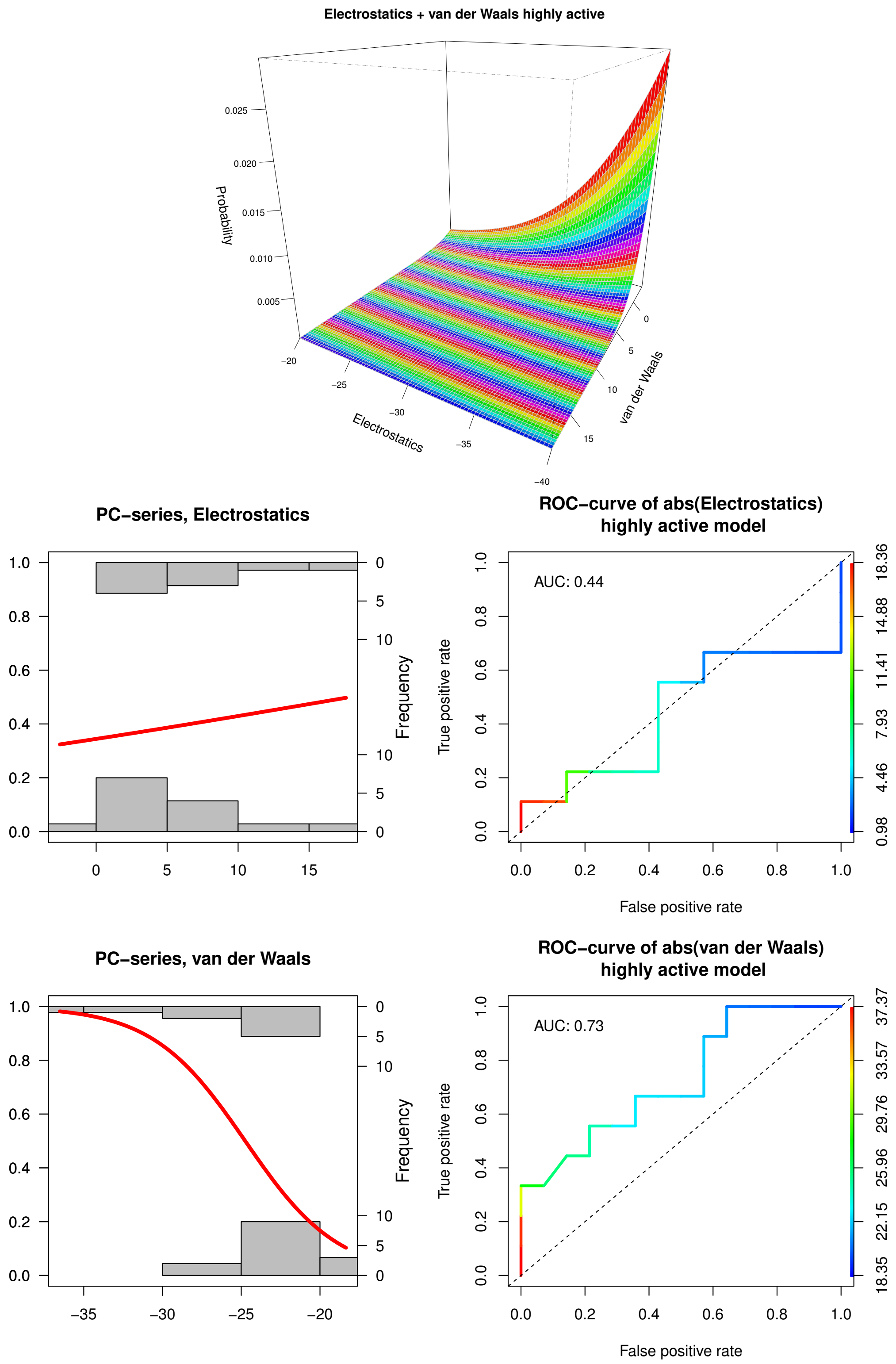


Figure B.5 Binomial logistic regression models for highly active molecules recognition for: electrostatics and van der Waals combined; electrostatics alone; and van der Waals alone; derived using ”final” dataset (N=23)

Table B.1 Pairwise interaction energies for the pyrrolidine carboxamide dataset in the bound (B) and free (F) states, respectively. The molecules of training and test sets used for the affinity prediction model are separated by a double horizontal line

	Bound (B)		**Free (F)**		**Difference (B-F)**	
	(kcal/mol)					
Compound	Electrostatics	van der Waals	Electrostatics	van der Waals	Electrostatics	van der Waals
s2	-44.93	-51.74	-46.73	-29.01	1.80	-22.73
s4	-38.69	-52.63	-41.22	-28.08	2.53	-24.55
s5	-37.65	-47.92	-40.06	-28.33	2.41	-19.59
s6	-35.75	-49.20	-41.69	-27.09	5.94	-22.11
s10	-39.44	-48.91	-44.10	-26.95	4.66	-21.96
s11	-38.68	-49.40	-41.44	-28.03	2.76	-21.37
s12	-44.60	-53.89	-47.61	-29.13	3.01	-24.76
s15	-39.25	-52.42	-41.91	-30.05	2.66	-22.37
d2	-30.33	-52.47	-38.38	-28.67	8.05	-23.80
d4	-38.61	-52.13	-42.83	-29.35	4.22	-22.78
d8	-30.65	-51.58	-41.92	-28.18	11.27	-23.40
d9	-40.69	-52.57	-51.81	-29.63	11.12	-22.94
d11	-34.05	-51.82	-34.99	-29.61	0.94	-22.21
d12	-38.53	-53.66	-40.86	-30.93	2.33	-22.73
d13	-34.14	-58.77	-44.72	-31.11	10.58	-27.66
d14	-36.72	-56.37	-39.06	-30.96	2.34	-25.41
d15	-39.32	-53.61	-40.32	-29.65	1.00	-23.96
d16	-39.21	-49.88	-42.12	-27.65	2.91	-22.23
3i	-36.64	-47.27	-42.92	-26.95	6.28	-20.32
r7	-17.10	-53.12	-34.43	-27.47	17.33	-25.65
p31	-32.82	-60.24	-50.36	-32.42	17.54	-27.82
c73	-36.24	-72.51	-53.83	-40.46	17.59	-32.05
c82	-32.40	-72.84	-38.97	-36.33	6.57	-36.51
s1	-31.59	-45.22	-36.08	-26.13	4.49	-19.09
s17	-41.51	-53.39	-48.63	-28.77	7.12	-24.62
d1	-33.38	-50.02	-38.45	-28.49	5.07	-21.53
d3	-34.20	-48.37	-35.60	-28.96	1.40	-19.41
d6	-39.91	-49.55	-44.33	-27.87	4.42	-21.68

continued on next page

	Bound (B)		**Free (F)**		**Difference (B-F)**	
	(kcal/mol)					
Compound	Electro-statics	van der Waals	Electro-statics	van der Waals	Electro-statics	van der Waals
d7	-40.91	-53.26	-42.46	-30.14	1.55	-23.12
d10	-35.30	-48.41	-38.75	-26.56	3.45	-21.85
3a	-37.05	-49.07	-43.11	-26.08	6.06	-22.99
3j	-34.42	-50.80	-41.80	-27.40	7.38	-23.40
p9	-26.60	-66.67	-51.01	-34.75	24.41	-31.92
p20	-22.13	-62.74	-36.32	-35.00	14.19	-27.74
p21	-39.13	-62.57	-49.82	-33.11	10.69	-29.46
p24	-31.93	-61.21	-50.64	-31.99	18.71	-29.22
p28	-100.00	-68.78	-95.01	-38.73	-4.99	-30.05
p33	-34.77	-65.63	-47.89	-33.01	13.12	-32.62
p36	-95.70	-67.19	-94.31	-39.77	-1.39	-27.42
c11	-30.71	-60.28	-49.33	-32.80	18.62	-27.48
c12	-29.43	-70.66	-38.20	-36.42	8.77	-34.24
c63	-43.25	-61.33	-53.61	-38.17	10.36	-23.16
c72	-30.01	-64.19	-40.81	-35.23	10.80	-28.96
c83	-44.20	-79.90	-58.52	-40.75	14.32	-39.15

Appendix C

Structure and synthesis of new pyrrolidine carboxamides

C.1 Synthesis of representative compounds

The basic synthesis scheme for the pyrrolidine carboxamides has been discussed in literature [52]. It basically consists of coupling an amine and a carboxylic acid giving the corresponding pyrrolidine carboxamide. This section describes in detail the synthesis routes for the representative compounds and the sole molecule identified from Scifinder® [323].

C.1.1 Synthesis of Scfinder hit

The routes for synthesis for the solitary hit from Scifinder® can be seen in Figure C.1. The synthesis of this compound consists of two steps as follows [52]:

1. **Synthesis of pyrrolidine carboxylic acid:**

   In the first step, itaconic acid and the amine corresponding to the C ring i.e. 5,6,7,8 tetrahydronaphthlene-1-amine were heated from room temperature to 200°C for 30 minutes. Thereafter, the molten mass was allowed to cool followed by addition of water and chilling the mixture in an ice bath. The mixture was then dissolved in aqueous sodium hydroxide and the solution filtered to remove any impurities. Subsequent to the filtration, the mixture was acidified with diluted hydrochloric acid followed by recrystallisation of the precipitiate with methanol and diethyl ether.

2. **Fusion with corresponding amine (microtiter synthesis):**

   In the second phase, a small portion of the pyrrolidine carboxylic acid was dissolved in dimethyl formamide (DMF). To this solution, HBTU (2-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate) and DIEA (N,N-diisopropylethylamine) were added in 1:2 ratio under constant shaking at 45°C for 2 hours, with the reaction being assessed by thin layer chromatographic (TLC) analysis.

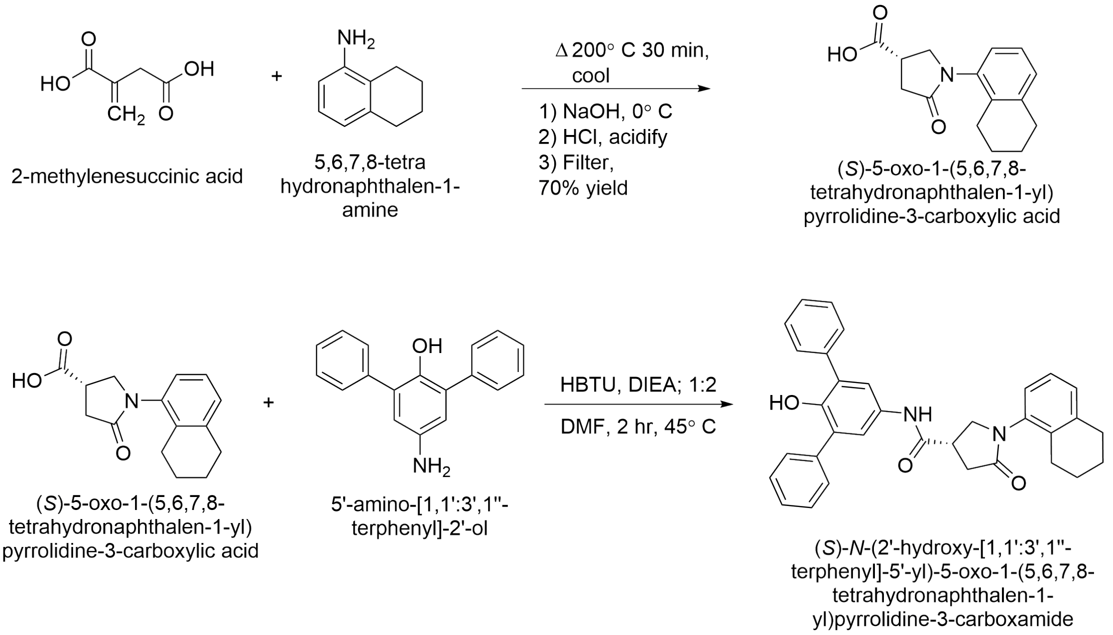

2-methylenesuccinic acid
+
5,6,7,8-tetra hydronaphthalen-1-amine
Δ 200° C 30 min, cool
1) NaOH, 0° C
2) HCl, acidify
3) Filter,
70% yield
(S)-5-oxo-1-(5,6,7,8-tetrahydronaphthalen-1-yl) pyrrolidine-3-carboxylic acid
(S)-5-oxo-1-(5,6,7,8-tetrahydronaphthalen-1-yl) pyrrolidine-3-carboxylic acid
+
5'-amino-[1,1':3',1''-terphenyl]-2'-ol
HBTU, DIEA; 1:2
DMF, 2 hr, 45° C
(S)-N-(2'-hydroxy-[1,1':3',1''-terphenyl]-5'-yl)-5-oxo-1-(5,6,7,8-tetrahydronaphthalen-1-yl)pyrrolidine-3-carboxamide

Figure C.1 Synthesis scheme for solitary hit from scaffold search in Scifinder®.

The above process can be tweaked in order to synthesize the new pyrrolidine carboxamides as shown in Figure C.2. However, in case of the new pyrrolidine carboxamides containing the alkyl chains, the process has to be expanded and modified in order to attach the alkyl ring to the ring-A1 or while replacing the C ring. The entire procedure is described in Appendix C.1.1.1.

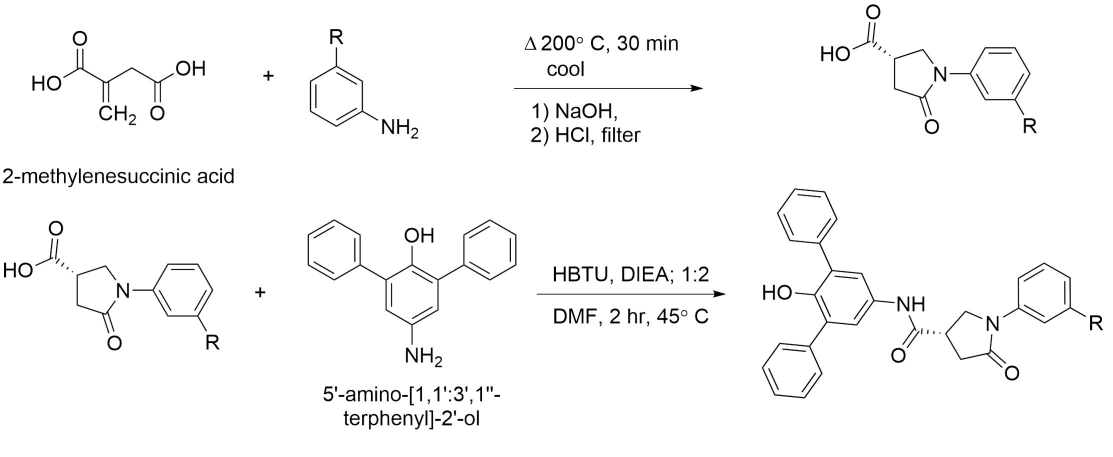


Figure C.2 Synthesis scheme for new pyrrolidine carboxamides that do not contain any alkyl chain.

C.1.1.1 Synthesis of new pyrrolidine carboxamides containing alkyl chain

In case of new pyrrolidine carboxamides with alkyl chains or with unreactive C-ring amine precursors, alternative synthesis schemes were studied. Although, the overall synthesis consisted of fusing the amine and pyrrolidine carboxylic acid to give the final product. The synthesis of pyrrolidine carboxamides with alkyl chains consists of two stages. The first stage deals with the pyrrolidine carboxylic acid synthesis. If the corresponding C-ring amine was unreactive with itaconic acid, the pyrrolidine carboxylic acid was obtained by copper catalysed arylation of the pyrrolidinone ring in presence of arenediazonium species. A saponification or catalytic reduction of the arenediazonium species yielded the pyrrolidine carboxylic acid [366]. The amine corresponding to the A ring system typically contains an alkyl chain and a reactive hydroxyl group. This makes the alkylation or arylation of the A1 ring challenging. This problem can be solved in two steps. Starting with 4-amino-2-bromophenol, a Friedel-Crafts alkylation could be performed that attaches the alkyl chain (of desired chain length) at the 5’ position. Arylation of the intermediate at the 2’ position can proceed via Suzuki coupling [367]. Finally, in the 2nd stage of the pyrrolidine carboxamide synthesis, fusion of the amine and pyrrolidine carboxylic acid can proceed according to the reaction described in Figure C.1. The whole steps are depicted in Figure C.3.

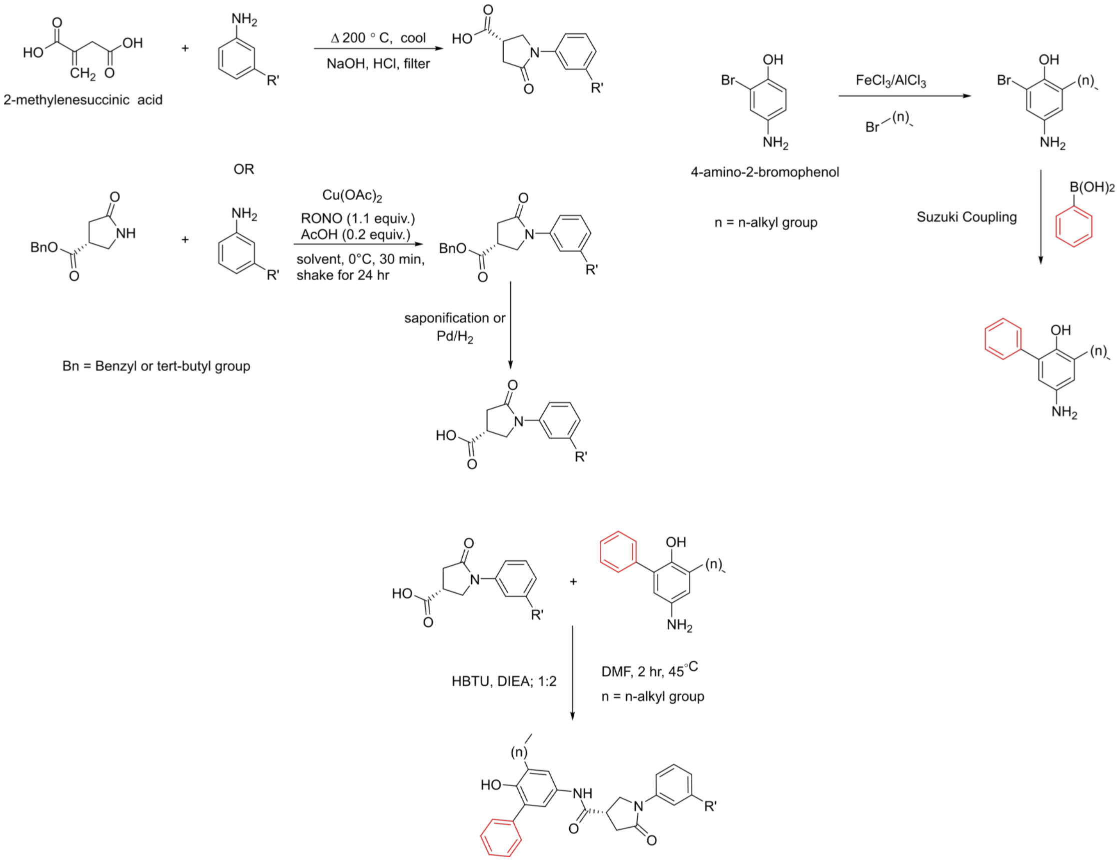


Figure C.3 Synthesis routes for designed pyrrolidine carboxamides containing alkyl chains.

Publications

Narkhede, Y., Wagner S., and Sotriffer, C.A., (in preparation) ”Activity-based classification circumvents affinity prediction problems for pyrrolidine carboxamide inhibitors of InhA”

Poster presentations

Parts of this work were presented at the following conferences as posters :

- Closing Syposium for SFB630 (2015), Würzburg, Germany
- 8. Joint Ph.D. Student Meeting of the SFBs 766 - 630 - FOR 854 (2014), Retzbach, Germany
- Annual meeting of the German Pharmaceutical Association (2014), Frankfurt, Germany
- Novel Agents against Infectious Diseases - An Interdisciplinary Approach (2013), Würzburg, Germany
- Chem-SyStM (2012), Würzburg, Germany

Affidavit

I hereby confirm that my thesis entitled, "***In-silico*** **structure-based optimization of Pyrrolidine carboxamides as *Mycobacterium tuberculosis* enoyl ACP reductase (InhA) inhibitors**" is the result of my own work . I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Signed: ______________________________

Place, Date:
Würzburg, 22.05.2017

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation "**In-silico Struktur-basierte Optimierung von Pyrrolidin-Carbonsäureamiden als *Mycobacterium tuberculosis* Enoyl-ACP-Reduktase-Inhibitoren**" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Unterschrift:

Ort, Datum:

Würzburg, 22.05.2017

Yogesh Narkhede