Use Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Idaho State University, I agree that the Library shall make it freely available for inspection. I further state that permission to download and/or print my thesis for scholarly purposes may be granted by the Dean of the Graduate School, Dean of my academic division, or by the University Librarian. It is understood that any copying or publication of this dissertation for financial gain shall not be allowed without my written permission.

Signature _____

Date 07/28/2017

MOLECULAR DOCKING AND SIMULATION STUDIES: IN SILICO DESIGN

By

Dipesh Budhathoki

A thesis

Submitted in partial fulfillment

Of the requirements for the degree of

Master of Science in the Department of Biomedical and Pharmaceutical Science

Idaho State University

Summer 2017

Copyright

Committee approval form

To the Graduate Faculty:

The members of the committee appointed to examine the thesis of Dipesh Budhathoki find it satisfactory and recommend that it be accepted.

_____

Dr. Kirk E. Hevener,
Major Advisor

_____

Dr. Dong Xu,
Committee Member

_____

Dr. Jared Papa
Graduate Faculty Representative

Dedication

I would like to dedicate this thesis work to my family members: to my parents Dan Bahadur Budhathoki and Indu Devi Budhathoki, my sister Sabita Budhathoki and my brother in law Shashi Sigdel for always supporting me.

Acknowledgements

My sincere appreciation to my supervisor and research advisor Dr. Kirk E. Hevener for his guidance, supervision, encouragement and vital suggestions during my graduate studies and guiding through my research work.

I would express my sincere thanks to Dr. Dong Xu for his helpful suggestions, advice and organizing the graduate committee members. He also provided me with the homology model structure of n-acetyl choline receptor. I acknowledge my other committee member Dr. Jared Papa too.

I thank Dr. Haydie LeCorbeiller in the Department of Writing Center at Idaho State University for her assistance with the proofreading service. I would also thank my friends who provided warm environment during my college days and stay at Idaho State University.

Lastly, I am eternally indebted to my parents, sister and brother in law for encouraging and supporting me throughout my graduate studies.

Table of Contents

List of Tables

List of Figures

List of Equations

List of Abbreviations

AMBER	Assisted Model Building With Refinement
BLAST	Basic Local Alignment Search Tool
CADD	Computer aided drug design
CAPRI	Critical Assessment of Prediction Of Interaction
FAS	Fatty Acid Synthesis
FDA	Food and Drug Administration
FF	Force Field
FFT	Fast Fourier Transform
FMN	Flavin Monooxygenase
GLIDE	Grid Based Ligand Docking With Energetics
GPU	Graphical Processing Unit
HADDOCK	High Ambiguity Driven Biomolecular Docking
HMM	Hidden Markov Model
HTS	High Throughput Screening
LBDD	Ligand-Based Drug Design
MD	Molecular Dynamics
NCBI	National Center of Biotechnology Information
NCGC	National Institute Of Health Chemical Genomics Center
NMR	Nuclear Magnetic Resonance
PBC	Periodic Boundary Conditions
QSAR	Quantitative Structure Activity Relationship
RMSD	Root Mean Square Deviation
SBDD	Structure-Based Drug Design

SP	Standard Precision
VMD	Visual Molecular Dynamics
VS	Virtual Screening
1D	One Dimensional
2D	Two Dimensional
3D	Three Dimensional

Abstract

The first project of the thesis included an approach to identify inhibitors of the enzyme enoyl-acyl carrier protein reductase (FabK) of *Clostridium difficile* using homology modeling. The second project involved molecular dynamics studies to obtain the lowest energy structure of *Salmonella typhimurium* DT 104 using homology modeling.

FabK is an important enzyme in the bacterial fatty acid synthesis process. It has become a potential drug target. This study used molecular docking to identify small molecule inhibitors of the enzyme in *Clostridium difficile*. We developed a homology model of this enzyme using the Modeller package with multiple structure templates and analyzed the models using various programs. The best model was chosen for the molecular docking of more than 3 million compounds from seven different libraries. We analyzed the top ten percent compounds from each library and found key interactions of top ranked compounds with the residues of the binding site. These studies identified potential hits with good scores for further experimental study.

Salmonella enterica serovar typhimurium DT 104 is one of the major reasons of Salmonellosis and is resistant to many antibiotics. It secretes AB5 toxin possessing ArtA and ArtB subunits. The study used homology modeling approaches from Schrödinger to create model structures of both the subunits. Protein-protein docking was used to obtain the complete ArtAB structure. Molecular dynamics simulations were conducted. The simulations helped us determine the energy of the system and conformational changes in the structure. The identification of the lowest energy conformation based on potential energy obtained may assist in future crystallography studies.

Chapter 1. Introduction

1.1 Computer-Aided Drug Design

It takes a long duration of time and hundreds of millions of dollars to introduce a novel medication.[1] There are significant chances of failure during the development process, and nearly 90 % of compounds which enter clinical studies do not get consent from the FDA, and fail to reach the patients.[1] High throughput screening (HTS) has been used primarily for screening compounds.[1-2] Despite the fact that the HTS technique can be successful in obtaining several hits which are prone to lead selection, the number of hits obtained from this method are usually very few.[2] To reduce the cost and time involved in the drug discovery and design process, CADD methods have been used extensively in drug design studies. Many pharmaceutical companies and other scientific groups have been using computational approaches as an important component in the initial phase of drug design to accelerate the procedure of developing new drugs, decrease the costs associated, and reduce chances of failure in the latter stages.[3] CADD is a rational approach towards drug discovery and development.

The methods used in CADD are grouped into two categories.[1-3]

a) Structure-Based Drug Discovery (SBDD)

b) Ligand-Based Drug Discovery (LBDD)

1.1.1 Structure-Based Drug Discovery (SBDD)

In this approach the three dimensional structure of the biological target enzyme or receptor is used for the purpose of designing or screening of the ligands.[3] Information about the binding site of the target protein also facilitates the SBDD process. Different

experimental and computational methods can be used for determining the structure of the target. Experimental techniques include X-ray Crystallography, Nuclear Magnetic Resonance (NMR) Spectroscopy and Cryo-Electron Microscopy whereas computational techniques include homology modelling, and molecular simulations.[4-5] Obtaining the three dimensional structure of the target provides a means to examine the potential topology of the binding site and the occurrence of cleft and binding pockets.[6] Knowledge of the valid binding site assists us to use several *in silico* methods like virtual screening and molecular docking to identify potential compounds.[1, 6]

The SBDD method uses different approaches as shown in Figure 1.1.

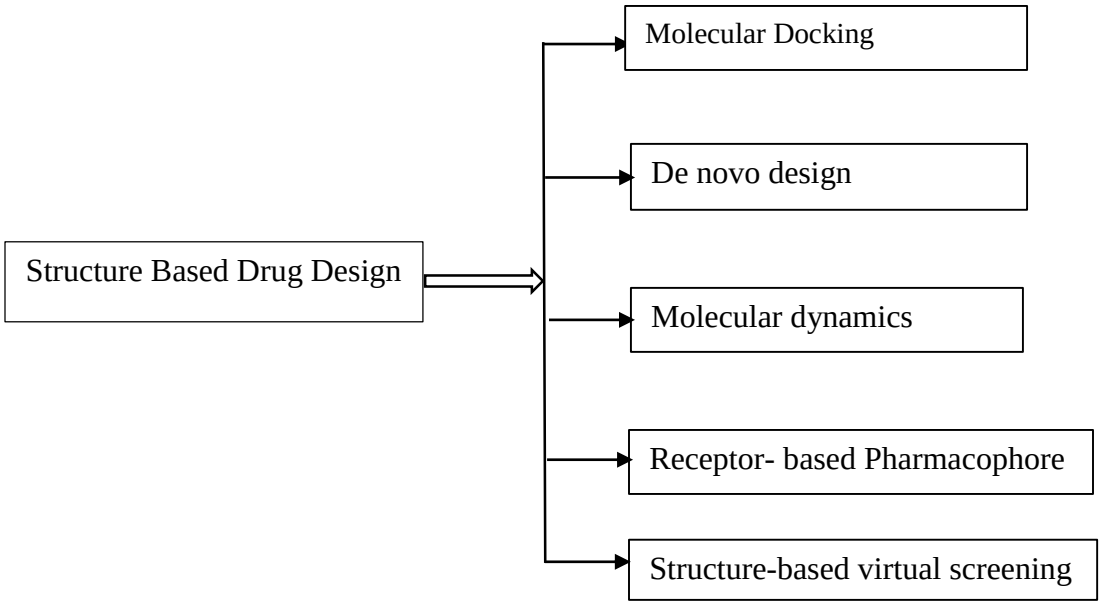


Figure 1.1. Different approaches used in Structure Based Drug Discovery

1.1.2 Ligand-Based Drug Discovery

There are some cases where the actual three-dimensional structure of the target protein is unrevealed and determining the structure using computational approaches like homology

modelling or ab initio methods becomes demanding, so the alternate approach in use in this case is ligand-based drug design.[8-9] In this method, the facts obtained from some compounds, which have measured activity against a given target, are used to determine the important structural properties accountable for showing the given activity.[3] Knowing the structural properties from the known ligands help us to design our ligand-based model.

Figure 1.2 shows some of the strategic approaches that are used in LBDD.[10]

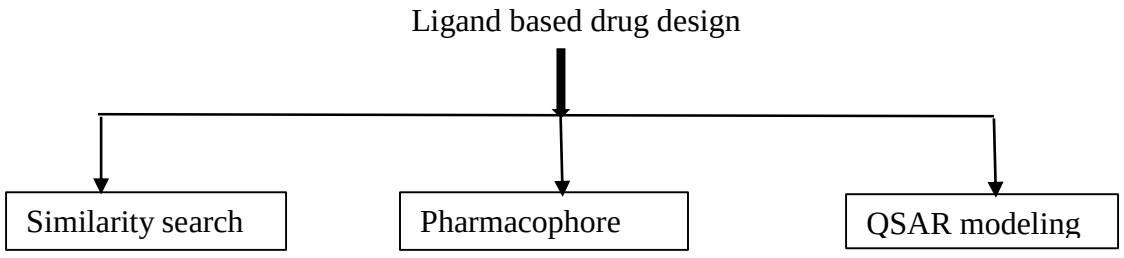


Figure 1.2 Approaches used in Ligand-based Drug Discovery

All the research projects carried out in this thesis utilized structure-based drug design. In our research projects, we did not use approaches of ligand-based drug design. Hence, we are explaining some of the approaches used in LBDD in brief.

a) Similarity search

This is one of the simplest methods used in LBDD. In this method, fingerprints of the known compounds are first generated ,and then these fingerprints are screened for larger compound containing databases to find ligands with identical fingerprints.[11] Both 2D and 3D fingerprints can be used for this purpose but a 2D fingerprint is the most preferred one. To determine the similarities between the query and the target fingerprint, different

types of coefficient of similarity are available.[12] Some of the frequently used coefficients of similarity in 2D fingerprint are as follows:[12] Tanimoto coefficient (Tc), Cosine, Forbes, Euclidean distance, dice, Russel-Rao and Soergel distance.

b) Pharmacophore modeling

In simple words, Pharmacophore refers to the molecular skeleton that shows important properties responsible for the pharmacological activity of the given ligand.[1] Pharmacophore modeling based on ligand design is one of the essential approaches in drug design when the target structure is not available.[13] Generally, this approach is carried out by obtaining common structural and chemical attributes from the three dimensional structures of a known group of ligands showing interaction with the target of interest.[13] At present various kinds of pharmacophore developing software are used in the field of drug discovery namely HipHop[14], HypoGen[15], DISCO[16], GASP[17], Schrödinger, MOE etc. A successful example of ligand-based pharmacophore technique is Indole-3-carbinol, a naturally existing anticancer agent that had poor metabolic property was optimized using this technique to develop a new compound analog SR13668 which has high potency against various types of cancer cells.[18]

c) Quantitative Structure Activity Relationship (QSAR)

The method used in QSAR is based on the reality that compounds that have identical structures prone to show similar biological activity.[19] In simple words, the QSAR technique is used to design a mathematical method that aims to obtain a statistical correlation between the physiochemical properties of the compounds and their biological actions, which is then used to determine activities of other novel molecules.[20] Some of the important points for generating a QSAR model are as follows:[1]

- Determining biologically active molecules and the actions that bind to the target of interest using either the large computational databases or high throughput screening.
- Obtaining the required physicochemical and structural properties of the compounds like bond angle, atom counts, important functional groups, surface area etc.
- Generating the QSAR model and finding correlation between the studied properties and the biological actions of the compounds.
- Authentication of the generated QSAR model.
- Application of the model to optimize the newly recognized active molecules for its biological actions.

Experimental study of the biological activities of the QSAR methods can be classified from 1D to 6D QSAR based on the structural depiction and the manner in which the descriptors are obtained.[20] Norfloxacin, a drug for the treatment of urinary tract infection, was designed using the method of QSAR.[21] Frequently, using QSAR techniques can get good results.[20] However, numerous times we fail to get favorable results for our purpose in spite of good correlation determined from the data used in the study.[20]

1.2 Homology modelling

Homology modeling is one of the useful computational techniques used for three dimensional structure prediction of biological targets. Identification of the structure of a protein with experimental procedures like x-ray crystallography and NMR techniques is a tedious task and not always successful in determining the structure of all proteins, particularly membrane proteins.[22] The overall steps to carry out the homology modeling are depicted in Figure 1.3.

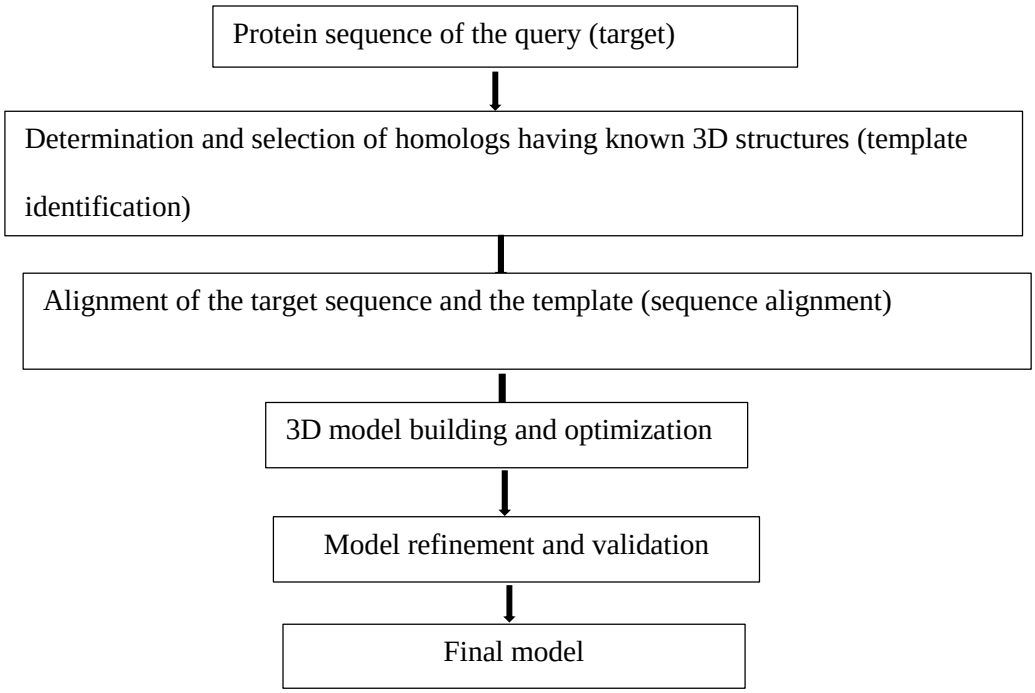


Figure 1.3. Flowchart for overall steps in generation of homology model

In this situation, homology modeling becomes useful provided the amino acid sequence of the query protein is known. This method is based on the foundation that evolutionary-related proteins have some kind of structural similarity.[1] Identifying protein structures that have identical sequences of amino acid to the query sequence helps in predicting the query's three dimensional structure and probable function as well as information about the ligand binding site.[1]

For generation of a homology model, the beginning step is to determine the homologous structure/protein to the target sequence.[1] The query sequence is compared with the sequence of the structure for which the three dimensional structure is available using several program or server.[22-23] One of the most commonly used servers for sequence alignment is Basic Local Alignment Search Tool (BLAST).[24-25] Applying BLAST for

local alignment purpose and searching the database with the query sequence provides a record of protein structures that matches up with the target sequence.[22] BLAST uses a pairwise sequence alignment method that predicts homology by aligning pair of sequences. The percentage of sequence identity between the query and the template sequences can determine the appropriate template for further homology modeling steps. BLAST provides good result if the sequence identity is over 30 percent. BLAST may not provide a good template if the sequence identity is less than 30 %, which means the templates obtained may not be accurate.[22]

A mistake due to alignment is one of the main reasons for divergence in comparative modeling even when we select an accurate template.[22] Profile-sequence and Hidden Markov Model (HMM) sequence alignments use an aligned group of related sequences shown by profile or HMM respectively. Position specific iterated BLAST (PSI-BLAST)[26] is a profile-sequence alignment. Similarly, the HMM includes Sequence alignment and Modelling(SAM)[27] and HMMER[28] which are also called profile HMMs. SAM generates an HMM with the use of single query sequence by repetitively searching templates in the database and then uses multiple alignment to produce an HMM .[27]

Another sequence-based method is a profile-profile alignment method that uses comparison of sequence families to discover evolutionary identity. This alignment has improved sensitivity and accuracy of the alignment.[29-30] Every profile-profile alignment includes steps like development of multiple sequence alignment, assessment of profile that can be used, profile alignment with sequence profile using online database like PDB and evaluation of the final outcome of the alignment with its statistical significance.[31] HH search[32], FFASO3[31] and profilescan[33] are commonly used profile-profile alignments.

The multiple sequence alignment method concurrently aligns a set of sequences already identified by other means to determine conserved section, estimate the functional area and help in evolutionary analysis.[34] These tools are helpful to improve quality of profiles and HMMs for the search of homologs. The majority of the multiple sequence alignments use heuristics called progressive alignment.[22, 35] Some of the multiple sequence alignment tools are ClustalW[36], ClustalX[36], T-Coffee[37], 3D Coffee/Expresso[38], M coffee[39], MAFFT[40] etc.

The structure alignment step follows the sequence alignment. Supplementary information about the secondary and tertiary structure of the protein is used to assist in alignment of the sequence. The structural alignment aligns areas of the protein sequence that are structurally similar instead of depending entirely on the sequence information. Some of the methods for structural alignment include distance matrix alignment (DALI)[41], sequential structure alignment program (SSAP)[42], combinatorial extension(CE).[43]

The DALI method uses the 3D structure of every protein and computes the distance matrices.[44] Distance matrices are calculated for each contact pattern of a hexapeptide sequence, and identical contact patterns are coupled as well as combined into greater consistent pairs.[44] Later these distance matrices are used for alignment purposes. The SSAP is a dynamic programming method which uses atom to atom vectors for structure alignment.[42, 45] Similarly, the CE method creates a pairwise structural alignment by applying local geometry to align short fragments. The algorithm used by CE evaluates a group of proteins and gives a record of proteins that are identical structurally .[43]

Model building

The next stage after sequence and structural alignment is model building. Different approaches used for model building are rigid-body assembly/fragment assembly[46-48], segment matching, satisfaction of spatial restraint[49] and artificial evolution[50]. The fragment assembly approach to the model building depends on different parts of the protein such as preserved center regions of the protein, a loop region that links them and the side chain regions.[22]

The foundation of the segment matching method is building of the query model structure using the sequence of the amino acids and the positions of the atoms.[51] The query structure is divided into a group of smaller fragments. [51] Later the database is explored for the similar fragments that are fitted onto the structure of the query/target.[51] Important points while selecting similar fragments from the database are homogeneity of the amino acid sequences, similarity with the atomic positions and closeness with the query structure.[51]

The method of 3D model building by satisfaction of spatial restraints contain three steps: aligning the query sequence with the template protein sequence, obtaining spatial restraints from the previous alignment step and acquiring complacency of the obtained restraints.[52]

Predicting loop region with good correctness is one of the important factors for using the obtained homology model for further structural study.[53] Several loop prediction methods are available for optimizing the loop regions. These are either in-built in some homology-model building software or available as separate loop modelling software. Some of the methods for loop modeling are i) template/database methods such as Superlooper and

FREAD that align model with the templates, measure the backbone segment spreading on the specific region and seek the database for a section of identical length that stretches on a section of same size and alignment.[22] ii) Other non-template based methods such as Molecular dynamics/Monte Carlo simulations that are useful in building larger loops. [22]

Model refinement

The model obtained after the model building needs refinement as it may contain structural problems like steric clashes, bond angle deviations and dihedral angles. Many homology-modeling programs already include a refinement stage in their algorithm. The first step in refinement is minimization of the model structure to reduce the bad contacts using a molecular mechanics force field. Further steps in refinement include all atom molecular dynamics simulation or Monte Carlo simulations for a longer duration of time using accurate force fields.[54] Commonly available force fields for both minimization and simulation include AMBER[55-56], CHARMM[57-58], and OPLS[59]. Once the structure is refined, it needs quality assessment as explained in the section below.

Model validation/assessment

The obtained model structure needs validation and assessment as well as suitable stereochemistry.[60] Generally, the model structures are compared with the template structures used for model generation. The Root Mean Square Deviation (RMSD) is used for measuring the average distance between the C alpha atoms in the backbone chain by superimposing the template and model protein structures. The RMSD indicates how greatly the model structure relates to and deviates from the template structure.

Next, the model structure can be assessed based on the Ramachandran plot between the template and model structures.[22] The Ramachandran plot provides information about the favored regions, generously allowed regions and disallowed regions of the amino acids. The greater the number of amino acids in the favored region, the better is the quality of the model structure. Various programs are available that use statistical scoring functions based on the observed properties of the amino acids of the structure. Correct protein stereochemistry including symmetry check, chirality, torsion angles, bond angle and packing volume can be assessed using programs like WHATCHECK[61] and PROCHECK[62]. Other programs such as VERIFY3D[63], PROSAII[64] and ANOLEA[65] are used to evaluate the fitting of the sequence to the predicted model and then later score for each residue which fits to the current environment. Besides using several programs for evaluation, manual inspection of the model structure is also an essential part of the assessment process.

1.3 Protein-Protein docking

There are various kinds of protein-protein interactions taking place in living organisms ranging from bacteria to human tissue. These interactions play vital roles in different biological activities that take place inside the body. To understand the mechanism of how these individual proteins combine to form protein complexes, several experimental techniques like x- ray crystallography, NMR spectroscopy and new method Cryo-Electron Micsoscopy are available.[66] Nevertheless, shortcomings of the above mentioned methods are related to the large protein area in the complex, flexible residues, and intensity of the interaction.[66] Hence, computational protein-protein docking plays an important role in molecular modeling of protein-protein interaction.

The main goal of the protein-protein docking is to determine the shape of the protein complex when two separate proteins are available.[67] While carrying out the docking, one protein, generally the smaller one, is considered as a ligand and the larger one as a receptor. Although structure of the proteins can be obtained from Protein Data Bank (PDB) that contains structures from X-ray crystallography, we can also use homology model structures which have smaller resolution.[67] There are two types of situation in protein-protein docking. The simplest situation is the case of bound docking. This docking tries to regenerate a known protein complex from the bound co-crystal structure of ligand -receptor.[67] The aim is to reconstruct the original complex after artificial isolation of the proteins. The molecule from the co-crystal structure acts as a beginning phase that contains more than one molecule. There are no conformational changes in bound docking.[67]

The unbound docking situation is more difficult than the bound. In this case, the proteins are present in their unbound native conformation, which are docked to predict the association between the molecules. There are significant conformational changes in the three dimensional structure of the proteins while interacting with each other to form the complex.[68]

There are three important components in protein-protein docking: i) Presentation of the molecules in the system ii) the search algorithm iii) scoring of the potential solutions.[69]

Presentation of molecules in the system

Molecules participating in the protein-protein docking should be represented computationally. Mathematical models like geometrical surface descriptors, grid or dynamic/static treatment of protein frame in flexible and rigid docking are frequently

used to portray these kind of surfaces.[69] A regularly used geometric surface is a Connolly that applies portion of the Vander Waals surface, which is attainable to the contact surface, called as probe surface joined by a group of concave, saddle and convex surfaces gets smooth over the crevices and pits over the atoms.[68-69] Other physiochemical properties can be used to complement the geometrical surface.[69] Another way that can represent surface is by utilizing the three dimensional grid in association with Fourier correlation search algorithm for distinguishing inner portion, surface and outer region of the proteins.[68]

Search algorithm

Searching algorithm searches for the suitable docking solutions of the complex. At present, ab initio methods are being employed in several programs where one of the proteins is definite and the other is either translated or rotated on the first one.[70] The major limitation of ab inito method in searching through the complete conformational space is that calculations are very costly and this seldom leads to specific solution.[70] Hence searching algorithms need to be descriptive, efficient, fast and accurate.[71] "*The docking method is generally based on the idea of complementarity between the interacting molecules, which may be geometric, electrostatic, hydrophobic, or all three* ".[68] The algorithms can be classified according to the flexibility reported into i) Rigid body docking, the primary model, in which both the proteins are treated as rigid bodies. There are no conformational changes in the proteins. ii) Semi-flexible docking, in which one of the proteins, mostly the smaller one, is regarded as flexible and the larger one is rigid in nature. There might be conformational changes in the smaller protein structure. iii) Flexible docking, in which both the protein structures have flexibility. These proteins

might have conformational changes. [69] These algorithms tend to find the solution with the minimum energy and stability. Two approaches preferred for searching of docking solutions include i) an entire space solution search, in which the entire conformational space is examined for the solutions.[69] ii) gradual guided progression through search space, in which only a part of the conformational spaces is examined for the solutions .[69] Methods like molecular dynamics simulation, Monte Carlo, simulated annealing and other evolutionary approaches like the genetic algorithm use search algorithm to search part of the space either in a partly random and criteria-guided fashion or by using ones that fit the solution.[69] Fourier Correlation technique like Fast Fourier transform (FFT) algorithm are also widely used.[68-69] It applies the correlation function and evaluates the overlap between the molecular surfaces of the two proteins and penetration during their relative shifts.[72] Later correlation values are calculated that explains the extent of overlap.[72]

Scoring function

The essential part of protein-protein docking is to find the pose with the global local minima after assessment of energies of all the available protein-protein docking poses. [73] Therefore, a proper scoring function differentiates between accurate native structures that have small RMSD and other protein complexes within an appropriate period of time.[69] There is no uniform opinion about the criteria to be incorporated for good scoring functions. Previously, many docking algorithms used geometric complementarity of molecular surfaces as the individual criterion for scoring function.[73] Criteria like steric complementarity in the site of interface, hydrogen bond and electrostatic interaction have been incorporated in the scoring functions.[68] A few scoring functions also apply

solvation potential, contact between the atoms and the residues, free energies, evolutionary usefulness of the interacting areas and clustering size.[68-69] Despite development in scoring functions, a lot of improvement in this area is necessary in the future to get a good predicted solution.

Evaluation of the result

After generating a protein-protein docked structure and scoring it, the next stage is the evaluation of the results. One of the evaluation steps is to compare the RMSD, generally between the C-alpha atoms of the native and the docked structures that we created. The lower the RMSD between the structures, the better is the result. Critical Assessment of Prediction of Interaction (CAPRI) is a benchmark that helps in assessment of protein-protein docking.[74]

Refinement of the docked structures

The docked structure needs refinement in further stage. Firstly, minimization of the protein-protein docked structure can be performed using a molecular mechanics force field. This helps to prevent bad contacts between the residues and minimize the energy of the system. Secondly, we can conduct molecular dynamics simulations of the structure.

Software used in protein-protein docking

Some of the software programs, their methods and sources available for the protein-protein docking process are mentioned below.

GRAMMX Server:

This is a widely used protein-protein docking software. It uses the rigid body FFT method for searching algorithm and Lennard-Jones potential for addressing the

conformational variation.[68] This is accompanied by refinement or minimization of the structure and rescoring.[68] The web interface for this open access server is http://vakser.compbio.ku.edu/resources/gramm/grammx[75].

ClusPro

This server applies the rigid body searching method, evaluates the poses based on knowledge-based scoring potentials such as atomic touch potential and utilizes the electrostatic potential for refining.[68] Ranking is based on the clustering features of the low energy composite.[68] The web interface for this server is https://cluspro.bu.edu/login.php[76].

Rosetta Dock

One of the most commonly used kinds of docking software, the Rosetta Dock uses minimization by Monte- Carlo process for both rigid- body as well as side-chain conformation to determine complexes containing low free energy.[77] The website for this software is http://rosettadock.graylab.jhu.edu/[78] .

HADDOCK server

HADDOCK stands for High Ambiguity Driven biomolecular Docking. The docking takes place in three steps: i) randomization of the orientations and rigid-body minimization of the energy in the protein complexes, ii) partial rigid simulated annealing in the torsional angle space and iii) lastly optimization in Cartesian space with explicit solvation.[68, 70] The address for the webserver: http://haddock.science.uu.nl/services/HADDOCK2.2 [70, 79]

Patchdock

Patchdock[80] is a docking program based on a geometry-dependent algorithm.[68, 81] It focuses on finding out the molecular docking transfiguration that provides better molecular shape compatibility.[81] The algorithm divides the Connolly dot representation of the molecules into various concave, convex and flat patches, and these patches are matched to generate reasonable transformations.[68] A web interface to Patchdock is http://bioinfo3d.cs.tau.ac.il/PatchDock/[81].

ZDOCK

ZDOCK uses a rigid body Fast Fourier Transformation (FFT) dependent algorithmic property that incorporates complementarity of the shape, desolvation energy and electrostatistics.[68] ZDOCK could be applied in coalition with RDOCK to help in minimization, further refinement and ranking of the complexes.[82-85] The web interface of ZDOCK server is http://zdock.umassmed.edu/[86-87].

HexServer

HexServer is the first to use graphic processors for this kind of docking .[88] It provides an easy way to operate the GPU- driven FFT dependent rigid body docking of protein complexes.[88] An easy to use web interface for HexServer is http://hexserver.loria.fr/ [88].

Schrodinger bioluminate:

Bioluminate is a commercial program provided inside the Schrödinger software package that performs the protein-protein docking to study the protein interactions. It performs the rigid body docking of protein complexes.

1.4 Molecular dynamics Simulation

Study of the protein dynamics/flexibility is important to know the protein functions. Moreover, protein dynamics study helps us to know about different possible conformations of the protein structure. Experimental methods like x-ray crystallography are available to study the flexibility of protein in ligand binding but they are expensive and require great hard work.[89] This problem led to the development of various computational methods to study protein dynamics.[89] Molecular dynamics (MD) studies the motion of the molecules using computer simulation. These simulations have been used in the research involving protein and biomolecules.

Conventional MD is a process to investigate the interaction and movement of atoms or molecules using Newton's laws of motion.[90] We use force fields to evaluate the force present on the interacting molecules and compute the total energy that exists on the system.[90] Later on during the simulation process, the representative configurations of the evolving system are generated that produce trajectories, give velocities and positions of the atoms throughout the time period.[90] Several features such as free energy calculation, potential energy and kinetic properties can be calculated from the generated trajectories that can be compared with the experimental values if available.[90]

The initial step in the MD simulation is representing the molecular system in terms of a computer model applying data from x-ray crystallography, NMR or a model structure. [89] Then the forces affecting each of the atoms in the system is calculated using Newton's law of motion. This can be achieved by resolving the differential equations of Newton's second law of dynamics as shown in Equation 1.1 below.[90]

$$Fi\ (t)\ =\ mi\ ai\ (t)\ =\ -\ dV\ (x\ (t))\ /\ dxi\ (t) \qquad \textbf{Equation 1.1}$$

Fi is the force at a particular point of time acting on the atom i, mi is the mass of the atom i and the acceleration is represented by ai. Similarly, the configuration of the system is shown by x (t).

To simplify the force acting on the system, potential energy function is introduced in Equation 1.1 and a simplified depiction of the model is generated called molecular mechanics (MM) or, in other terms, the Force field (FF).[90] In this force field, two types of forces, namely forces resulting from interaction between bonded atoms and non-bonded atoms, are taken into consideration.[89] The MM with the several component forces are represented in the Equation 1.2 below.[89]

$$\text{Etotal}\ =\ \sum_{bonds} K_r\left(r-\ r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta-\ \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{U_n}{2}[1 + \cos(n\emptyset\ -\ \gamma)]\ +\ \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}}-\frac{B_{ij}}{R_{ij}^{6}}+\frac{q_i\ q_j}{cR_{ij}}\right] \qquad \textbf{Equation 1.2}$$

The foremost three energy terms, namely bonds, angles and dihedrals, represent the bond stretching, bending and torsions between the atoms that interact with each other. [90] These represent interactions between the bonded atoms. Van deer Waals and Coulombic interactions respectively portray the non-bonded interactions. Van deer Waals interactions are considered using the 6-12 Lennard-Jones potential.[90] Likewise, the electrostatic interactions between the two atoms are depicted using the Coulombic potential where partial charges between the pair of atoms are considered.[90] To propagate

the true behavior of the molecules , the force field terms mentioned above may undergo parameterization with the quantum-mechanical and experimental values.[89] The benefit of using molecular mechanics is that it boosts up the calculation process in comparison to the quantum mechanics method.[90] Some of the commonly used molecular mechanics force fields for simulation of biological molecule are AMBER[55-56], CHARMM[91], GROMOS[92], and OPLS[59]. Besides amino acids, the guardian force fields have also accommodated the parameters for carbohydrates, nucleic acids, lipids and some ionic molecules in the recent years.[90] Nevertheless, the parameters for ligand and other non-standard residues have problems dealing with these force fields. Hence, the user needs to provide parameters for ligand and some non-standard residues.[90] Force fields like GAFF[56] and CGenFF[57] are available for parameterization of organic molecules/ligands that are used for AMBER and CHARMM respectively.[90] Programs such as ANTECHAMBER[93], paramfit[94] , Hopkins-Roitberg's approach used for amber[95] and general automated atomic model parameterization (GAAMP)[96] are frequently used for the purpose of parameterization of small molecules.

After the force field has been specified and the forces affecting every atom on the system has been calculated, the particles are moved based on Newton's law of motion.[89] Since the bio molecular systems are larger in size and difficult to solve, the integration of Newton's law of motion needs to be divided into distinct intervals of time ie. time steps, dt.[90] As forces acting upon the atom depend upon positions of the atom that fluctuate over time, a slight dt introduced helps to portray valid forces over the time.[90] One of the simplest and best algorithms used for integration of Newton's equation of motion is the Verlet algorithm[97]. This algorithm utilizes positions, velocities and accelerations at a time

t to calculate the new positions at time t+dt which is shown by the Equation 1.3 given below.[90]

$$X(t + dt) = x(t) + v(t)dt + 1/2a\,(t)dt2 \qquad \textbf{Equation 1.3}$$

Likewise, accelerations are computed by measuring the force affecting per atom that includes the first order derivative of potential energy in regard to its positions.[90] When these tiny steps are integrated, they form the trajectory of the whole system for a given time span. The time steps need to be advanced by one or two femtoseconds and the method is repeated multimillion of times.[89] We can accelerate the simulation procedure by affixing the length of the covalent interaction, which helps to accelerate the time steps from 2 to 6 femtoseconds.[98] Algorithms like SHAKE[99] and RATTLE[100] are used to constrain bond lengths. SHAKE is one of the most commonly used algorithms for this purpose.

Another way of running long time scale simulation is Hydrogen Mass Repartitioning. Studies have shown that this method is a convenient way of increasing the speed of molecular dynamics simulations.[101] Repartitioning the weight of heavy atoms into the bonded hydrogen atoms, the motions of biomolecule under study can be reduced.[101] This helps to increase the time steps of MD simulation by factor of 2.[101] With this particular technique, the simulations can be increased to 4-femtosecond time step.

Ensembles used in molecular dynamics simulation

During the MD simulation, we use different kinds of ensemble approaches. Straight implementation of most of the integrators will result in simulations that use micro canonical ensemble i.e. constant number of particles (N), volume and energy (NVE

ensemble).[90, 98] Nevertheless, the simulation of the system can be carried out under a constant temperature condition called canonical ensemble where N, V and T are constants (NVT ensemble).[98] Several coupling algorithms are used to maintain the constant temperature during the simulation process[98], and are called thermostats[102]. Commonly used thermostats in the MD simulation are Berendsen[103], Anderson[104] and Noose-hoover thermostats[105-106]. Similarly, simulations are also carried under the conditions of constant pressure where N, P and T are constants (Isothermic isobaric ensemble). Several barostat algorithms[97, 107] are available to maintain the constant pressure during simulation. The pressure is checked by fluctuating the volume of the system.[90]

Periodic Boundary Conditions (PBC)

In a molecular dynamics simulation, periodic boundary conditions are used to illustrate the large characteristics of the system of a finite size .[90] In PBC, the system is kept inside a unit cell and then reproduced by translation in every direction to make system more like an infinite geometry to fill the space.[90] Practically cut off distances are appointed to judge non-bonded interactions and make sure that only one group of closest neighbors are present around a simulation cell.[108] Usually a cut-off distance of 8-10 Angstroms is preferred. An interesting point is that if a particle/atom moves out of the unit cell in one or other cell coordinates by any means, its corresponding image enters the cell from the spot where the atom has left by lattice symmetry.[108] Some important features of PBCS are they help to conserve particle number, mass and liner momentum in the system.[108] The most important behavior of PBC is to confirm that no particle inside the unit cell notices a vacuum within the cut-off area and hence surface artifacts are prevented.[108]

Solvation models in MD simulation

Solvation has an effect on the simulation process. There are two types of solvation model used in MD simulation. They are explicit and implicit solvation models. Water is the most widely used solvent for simulation of biomolecules. In explicit solvation, water molecules are added explicitly to the system. Commonly used water models in the simulation process include SPC[109], TIP3P[110], TIP4P[110] and TIP5P[111]. Although all the water models mentioned above have been improved to match one or other physical properties of water including radial distribution, diffusion and density, none of the models represent all the properties. [112] Since the solvent molecules are also present in the simulation, application of explicit solvent in the simulation makes the process computationally extensive.

Implicit solvation is less rigorous computationally. The implicit model treats the solvent molecule as a consistent high-dielectric medium surrounding the target and the solute is treated as a small dielectric area with spatial distribution of charge.[114] Commonly used implicit models are the Poisson Boltzmann (PB) model and the Generalized Born (GB) model.[115]

Figure 1.4 represents an explicit water solvation model of the transmembrane bound human nicotinic acetyl choline receptor (nAchr) alpha 7 homology model[113]. We created the explicit water solvation model utilizing the Visual Molecular Dynamics (VMD). The figure shows the graphical representation of the transmembrane n-acetyl choline receptor (nAchr) created using VMD. Yellow CPK represents chloride ions. Purple CPK represents sodium ions. New ribbon depicts protein structure. The lowermost surface represents lipid bilayer. The outermost transparent surface depicts water box.

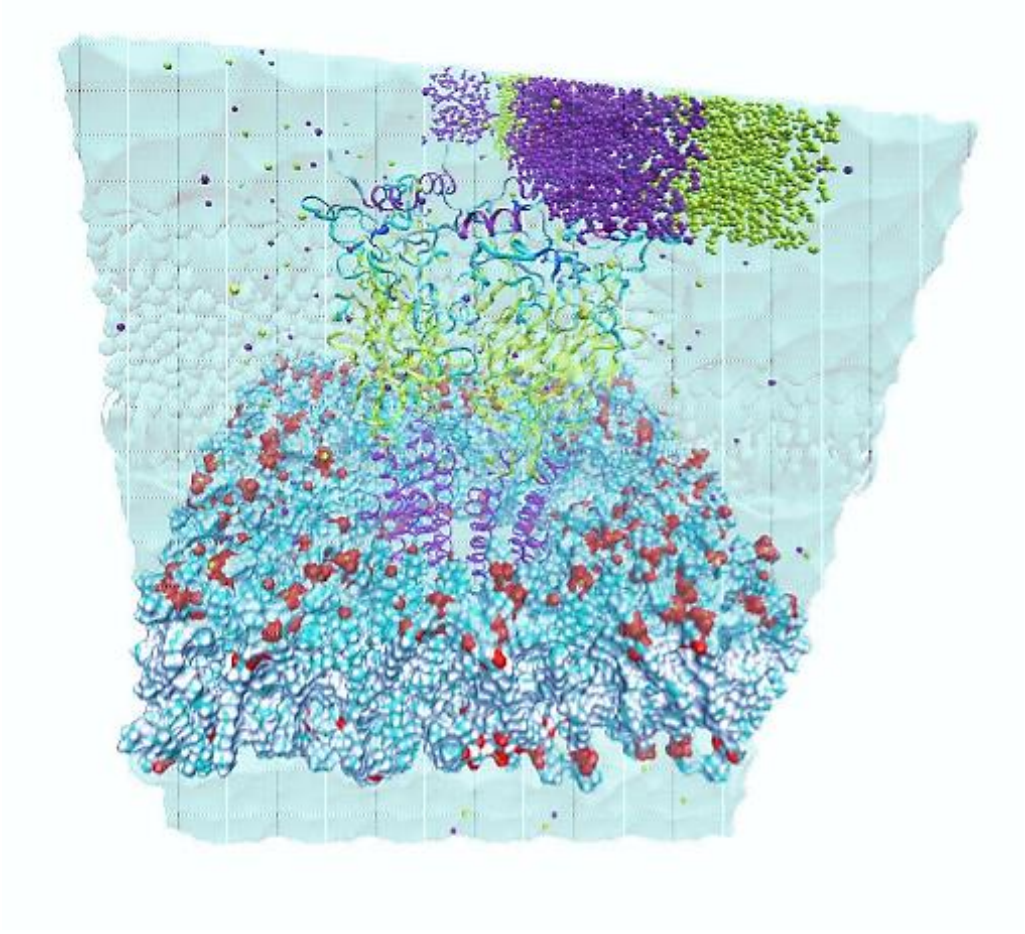


Figure 1.4. Explicit water solvated protein with ions included.

The most commonly used MD simulation packages are AMBER[116], CHARMM[91], GROMOS[92, 117], GROMACS[118-120], NAMD[121-122] and DESMOND[123]. We have used AMBER for molecular dynamics simulation in our Salmonella typhimurium ArtAB project.

1.5 Molecular docking

During the drug discovery stage, it is a commonly used technique to utilize big and diverse libraries of small molecules for the purpose of *in silico* and experimental screening. [124] Virtual high throughput screening is a widely used computational screening

of huge libraries of compounds from the database against a desirable macromolecular target to find the active molecules. Later the biological features such as affinity, potency and efficacy can be identified for the selected compounds using various experimental methods.[125] Virtual screening has advantages over high throughput screening in that it is cheaper, faster and if carried out cautiously gives good hit rates. Virtual screening is classified into structure-based and ligand-based virtual screening approaches. We have already discussed ligand-based approach in the previous section. Since we have used structure-based virtual screening in our project, we are mainly focusing on it.

Molecular docking is the most commonly used structure-based virtual screening technique. Molecular docking is a widely used computational approach in virtual screening that helps to estimate the potential binding mode of a molecule in a specific binding site of a target.[126] The three-dimensional structure of the target/receptor helps to dock the compound in the binding site. With the application of molecular docking, study of the compound binding mode and the intermolecular interaction that takes place when the ligand binds with the receptor can be performed.[127] Moreover, this kind of study provides estimation of the binding energy and ranks the docked ligands depending on the binding association of ligand and receptor.[127-128] Molecular docking of ligand and receptor complex involves two main stages[127, 129] i) searching of the conformational space or sampling to find the docking poses and ii) scoring and ranking of the generated docked poses.

Searching conformational space or sampling

Sampling in molecular docking is mentioned as the production of genuine compound binding conformation in the active site of the target.[127] Two commonly used algorithms

for conformational search are i) systematic search and ii) stochastic search.[130-131] The systematic conformation search causes small changes in the structural parameters and slowly alters the conformation of the compounds.[132] These search algorithms are applied for flexible ligand docking and produce all feasible conformations of the compound docked.[127] Table 1.1 shows different docking programs and their search algorithms.[6]

Table 1.1. Docking programs under stochastic and systematic algorithms

Stochastic algorithmic search	**Systematic algorithm search**
GOLD[133]	DOCK[134]
ICM[135]	LUDI[136]
QXP[137]	FRED[138]
AUTODOCK[139]	eHiTS[140]
Prodock[141]	GLIDE[142]
PSI-DOCK[143]	FleXX[144]
Molegro Virtual Docker[145]	Surflex-Dock[146]
MolDock[147]	ADAM[148]
PRO_LEADS[149]	SLIDE[150]
MOE[151]	FLOG[152]

The stochastic algorithm performs sampling of ligand conformations/orientations by causing random changes in the structural parameters of the compound. [6, 127] The stochastic algorithm produces a combination of ligand conformations, examines the energy space and expands the chances of obtaining the final ligand-binding mode at the global minima.[6]

Scoring functions

A scoring function is an important aspect of molecular docking programs. Scoring functions evaluate the binding affinity of the ligand and target complex.[6] Scoring functions also help in ranking the docked poses according to the binding affinity. Scoring functions can be classified into three groups[127, 153] i) force field ii) empirical iii) knowledge based scoring function and iv) consensus scoring function.

The force field scoring function, as the name suggests, takes into account various bonded and non-bonded interactions that we mentioned earlier and hence calculates the binding affinity of the compound into the receptor.[6] One drawback about this scoring function is it does not correctly predict the entropic effect. [6]

The empirical scoring function utilizes the statistics obtained from experimental structures and tries to match them with the parameters.[1] It calculates the binding affinity by taking into account the properties like hydrophobic impact and total hydrogen bond information and then matches them with the experimental details.[1, 3]

The knowledge based scoring function uses statistical or structural data from the experimental compound-receptor complexes.[1, 64, 154-155] While calculating the energy, it uses a dataset of structures and considers the number of occurrence of atom pairs present within a known distance.[6] After that, several interactions that are present between the atom pairs in the dataset are categorized and the ultimate score in the docking process is the aggregate of these interactions.[6]

Many of the above scoring functions may have some shortcomings.[6] Therefore, a consensus scoring function increases the chances of obtaining a good ultimate solution

from the docking by combining several scoring functions and neutralizes the shortcomings from the independent one.[127] Table 1.2 shows several scoring functions used in molecular docking and their examples.

Table 1.2. Different scoring functions and their examples

Scoring function	Examples
Force-field based	Autodock,[139] DOCK,[134] Goldscore,[133], ICM[135]
Empirical based	Ligscore,[156] LUDI,[157] Glidescore,[142] X-score,[158] Chemscore,[159] Fresno,[160] SCORE[161]
Knowledge based	Drugscore,[162] SMoG,[163] PMF-Score,[164] Posescore[165]
Consensus	Multiscore,[166] X-Cscore,[158] GFscore[167]

In our research project we used GLIDE (Grid-Based Ligand Docking with Energetics) for molecular docking and virtual screening purposes. This is a program available in the Schrödinger suite. During the glide docking process, the grid, generated by the grid generation process represents the receptor. [168] The appearance and features of the grid in the grid generation stage are computed by using several groups of force fields that facilitate the correct scoring of the ligand in the receptor.[142] Initially the glide program produces poses of a ligand in the docking procedure. [169] Later the glide program applies a hierarchical group of filtering criteria to explore a feasible location of the compound in the binding site of the receptor grid.[142, 169] The glide program produces a group of compound conformations.[142] With these compound conformations present, preliminary screening of complete conformational space takes place, and generates a set of favorable

ligand poses during the procedure.[142] Minimization of the poses of the compound occurs with the application of force field OPLS-AA[170] in the grid.[142, 169] A few of the small energy ranking poses also undergo Monte Carlo calculation to find out the minima.[142] After the minimization steps eventual scoring takes place with the Glidescore.[169] Glidescore is employed to estimate the binding affinity of the compound and also aids in ranking of the compounds in the virtual database.[142] However, we have used a different matrix called glide ligand efficiency score in ranking some of our compound libraries.

The glide ligand efficiency score in our calculation determines the efficiency of the ligand. The larger the docked ligand molecule in the receptor, the greater the chances of more interaction with the amino acid residues and increased score. To solve this problem, the glide score has been normalized taking into account the number of heavy atoms present in the ligand.

Chapter 2. Modelling and docking studies on the *Clostridium difficile* FabK enzyme

2.1 Introduction

Clostridium difficile is one of the major causes of nosocomial diarrhea, which affects people causing mild diarrhea to severe cases of colitis.[171] *C. difficile* leads to more than 450000 infections and approximately more than 29000 deaths every year in the United States.[172] In regards to expenditure on health services, *C. difficile* infection has caused more than $4 billion of surplus health care expenses considering acute health service only.[173-175] Infections from *C. difficile* may occur after an interruption of the gut protective normal flora increase growth rate of *C. difficile* in the intestine.[176-177]

Studies have shown that the two toxins A and B produced by the bacteria play major roles as virulence elements in the pathogenesis of the disease.[178-179] Besides, these toxins also act as an important indicator in the diagnosis of the disease caused by the *C. difficile* infections.[179] Hence, the present diagnosis and treatment of infections mainly focuses on these toxins. Drug therapies like Metronidazole and Vancomycin are common in the treatment of infections caused by *C. difficile*.[180-182] These infections can be treated with antibiotics but relapse of the disease even after the treatment has been a major issue for a long time.[181] Moreover, these antibiotics also harm normal flora present in the intestine. There is an immediate necessity for the new drug targets in these bacteria for the treatment of the infections.

2.1.1 *Clostridium difficile* FabK enzyme

One of the possible targets in the *Clostridium difficile* is the reductase enzyme called enoyl-acyl carrier protein reductase II (FabK). Different types of enzymes having various features catalyzing the bacterial fatty acid synthesis process (FAS II)[183-184] are shown in Figure 2.1[185] below.

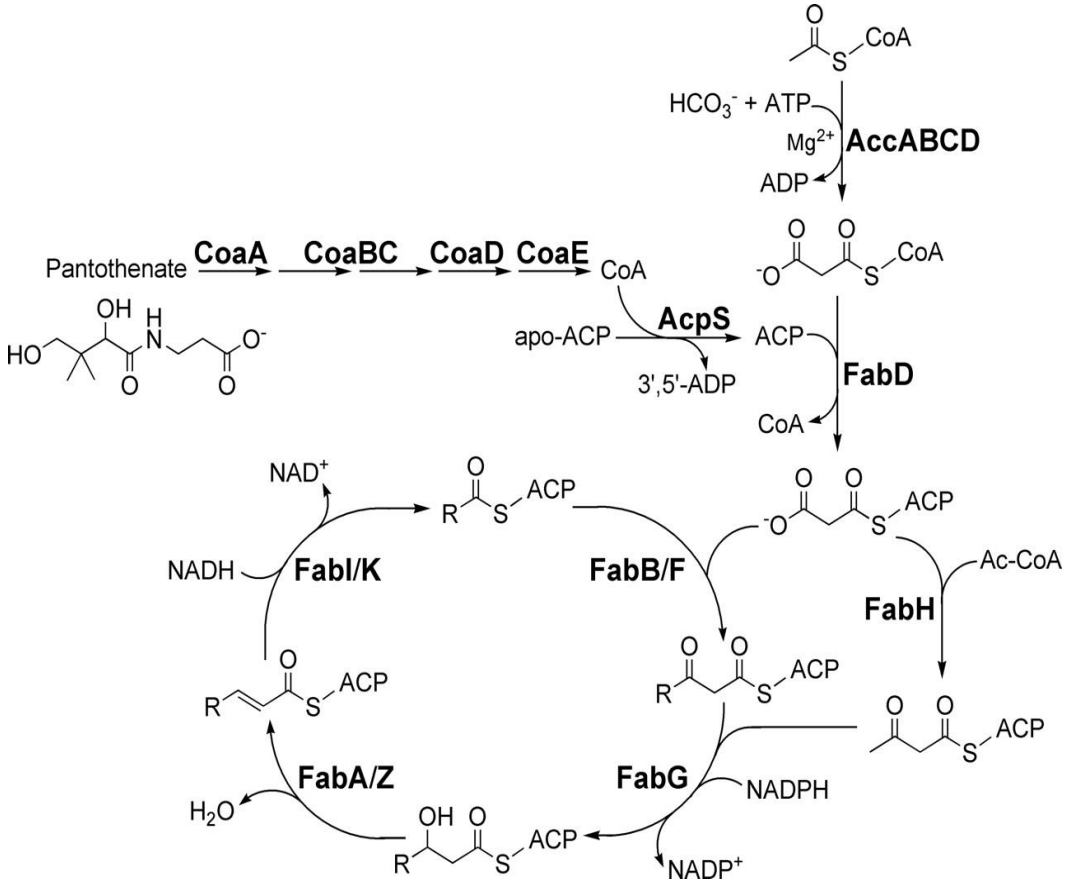


Figure 2.1. Bacterial fatty acid synthesis process

This figure was reproduced with the permission of the journal[185].

FabK is a reductase enzyme, a flavoprotein that needs NADPH for its biological action[186] and is resistant to the drug Triclosan[187]. Many of the enzymes that catalyze the bacterial fatty acid synthesis can be potential target of antibacterial agents. For example, Triclosan

inhibits FabI, an important enoyl-acyl ACP reductase, in the fatty acid synthesis process of bacteria[187]. Likewise, FabK is the only enoyl reductase enzyme that plays a crucial role in the fatty acid synthesis process of *C. difficile*. It is distinct from FabI enzyme. Currently, no study has investigated the crystal structure of *C. difficile* FabK and targeting this enzyme with the inhibitors.

Our hypothesis for this study is that the small molecule inhibitors of FabK enzyme show selective antidifficile action against *C. difficile* with good binding affinity. The basis of our hypothesis relies on the fact that FabK is the only reductase in *C. difficile* that leads to selectivity. Some studies have shown the inhibitors of FabK in other bacteria.

2.1.2 Aim of the study

The aim of this project is to identify small molecule inhibitors of the *Clostridium difficile* FabK enzyme. We believe that inhibition of this bacterial fatty acid synthesis process in *C. difficile* will provide us compounds for further experimental study.

To fulfil our aims, we used a computational approach to develop a validated homology model of *C. difficile* FabK, as it does not have a crystal structure. After that, the refined model was used for molecular docking using compounds from seven different libraries.

2.2 Materials and Research methodology

To fulfil our aims and carry out our research activities we used different software programs for various purposes. The several kinds of software used in this project are summarized in Table 2.1 below. In the project, Maestro from Schrodinger[188], Chimera[189-190] and VMD[191-192] were used for the purpose of visualization; other programs and software are itemized in detail in the research methodology.

Table 2.1. Software used in *C. difficile* FabK modelling and docking

Program	Supplier
Chimera-Modeller interface	Chimera
Swiss Model	Swiss model expasy server
Verify3D	UCLA-MOE lab
Protein Preparation wizard	Schrödinger LLC
Ligand preparation program	Schrödinger LLC
Maestro	Schrödinger LLC
Glide docking	Schrödinger LLC
Chimera	University of California San Francisco (UCSF) Biocomputing team
VMD	University of Illinois at Urbana-Champaign (UIUC) Biophysics group

2.2.1 Creating the homology model of *C. difficile* FabK

We obtained the FASTA amino acid sequence of enoyl- (acyl-carrier-protein) reductase II enzyme of *Clostridium difficile* strain 630[193-194] from the National Center of Biotechnology Information (NCBI) Protein database[195]. To generate the homology model of *C. difficile* FabK, we used the Modeler interface available in Chimera. Using the query sequence, we selected templates from the blast query considering whether the template contained the FabK enzyme or not. Two templates were chosen evaluating

sequence identity, E value, score and resolution of the structure. The first template was 2Z6J_A i.e. chain A of the crystal structure of *Streptococcus Pneumoniae* FabK in complex with the phenyimidazole inhibitor, TUI (contains 58 % sequence identity, Evalue = 1.79694e-114, score=334 and resolution 2.3). The second template was 4IQL_A i.e. chain A of the crystal structure of *Porphyromonas gingivalis* FabK with sequence identity 46.1 percent, Evalue = 5.09482e-71, score=223 and resolution 1.94.

Initially blast alignment was used for aligning the target and the templates. Later we realigned the sequences using Clustal Omega, which is useful for multiple sequence alignment. We viewed the sequence alignment in the Multialign viewer.

We ran a structure prediction from the Chimera-Modeller graphical user interface locally and generated 10 different models. One Na+ ion from the center and FMN from 4IQL were retained on models. Likewise, we retained the inhibitor TUI from 2Z6J template. FMN and the Na+ ion lie in the active site of the receptor. The TUI retained in the homology model could provide us with the coordinates of the centroid of the ligand during the later docking steps. We deleted NADPH as it acted as the competitive inhibitor of FMN. Similarly, we removed two Na ions in the template 4IQL that were far from the active site. Among ten different models generated from the Modeller-Chimera interface, the selected model was the one with majority of best scores in terms of the lowest RMSD, lowest zDOPE or normalized discrete optimized protein energy (more negative value of the structure) score, highest GA341 score and the best-estimated overlap of the model and template.

The model was subjected to 2500 iterations of energy minimization from the Schrödinger software[188] (Tasks – minimization – forcefield) to minimize the bad contacts, steric clashes and other structural problems.

We performed further evaluation of the model. PROCHECK[62] was used to evaluate the stereo chemical properties utilizing Swiss model expasy server[196]. VERIFY 3D[197-198] helped to evaluate the compatibility of the three dimensional model structure with its own amino acid sequence (one dimensional structure). This analysis was performed using the structural analysis and verification server (SAVES)[199]. QMEAN[200] was used from the Swiss model expasy server[196] to evaluate the quality of the model taking into consideration the six different energy terms that are mentioned in the results later.

2.2.2 Sequence identity of the active sites of the homology model

After we evaluated the final selected homology model, we calculated the percentage sequence identity of the residues of the active site utilizing the sequence alignment. First, we selected 5-Angstrom region around the TUI inhibitor as the active site. The residues were chosen within this region. We opened the two templates (2Z6J and 4IQL) in the Chimera. The model structure was superimposed with both the templates (Tools -- Structure comparison -- Matchmaker). The model structure was considered as a reference and the two templates as matching structures. The matching was restricted to the selected areas of the active residues. After the superposition, structure-based multiple sequence alignment was calculated for the model and the two templates. Structure alignment of the residues of the active site of the homology model with the templates helped us to calculate the percentage sequence identity of the active site residues with reference to the templates.

2.2.3 Redocking validation using inhibitor of the model structure

The following step included the docking revalidation of the inhibitor TUI in the final refined homology model. Validation is performed to check how far docking of the native ligand reproduces the result.[169] Moreover, this also provides us with an idea of the parameters that can be used for molecular docking of compound libraries. We extracted the TUI inhibitor from the model structure. This was followed by ligand preparation stage. For this, we used the ligprep program that prepares the ligand for docking process. Glide suggests processing of the ligands by ligprep before the docking procedure.

The next step was preparation of the model structure using the protein preparation wizard of the Schrödinger software. The protein preparation wizard[201] has two stages: i) Preprocess that assigns bond order, adds hydrogen to the protein, produces zero-order bond to metals and fills missing side chains that may occur in the protein; and ii) Optimization that optimizes the intra and intermolecular hydrogen bonds present between the amino acid residues.

The next step was generation of the receptor grid from the task menu of the Schrödinger. We did not include the original TUI in the grid. An enclosing outer box was prepared (size 29*29*29). This is the outermost region that the docked compounds can occupy. Similarly, an inner box of 12*12*12 size centered on the native TUI was created. This defined the size of the inner/ligand diameter midpoint box. This was adjusted from the advanced settings option. No constraints were used in the procedure.

After preparation of the ligands from the ligprep and generation of the receptor grid, we performed the glide docking. Flexible ligand sampling and Glide standard precision modes were selected. We analyzed the docked compounds based on the RMSD between

the heavy atoms of the docked pose and the native TUI inhibitor present in the homology model. The three docked poses generated from the glide docking were saved in mol2 format. The RMSD between the docked poses and the native TUI ligand in the homology model were calculated in Chimera using rmsd command from the command line. Hydrogen molecules were removed from the docked poses to make an equal number of heavy atoms between the docked ligand and the native TUI.

2.2.4 Redocking validation of the ligands of the crystal structures

We conducted the redocking validation of TUI to 2Z6J and FMN to 4IQL. We used Auto dock vina from Chimera for these docking jobs. We removed TUI from 2Z6J and FMN from 4IQL PDB structures. This was followed by the protonation of both the protein structures utilizing PDB2PQR in Chimera (Tools – Structural editing – PDB2PQR). The cartesian centers of TUI and FMN were calculated using the linux command line. We protonated TUI and FMN at pH 7.4 using Openbabel. These structures were retrieved in mol2 format.

Next, we conducted the docking using Auto dock vina from Chimera (Tools – Surface/Binding analysis). For 2Z6J, the pqr file of 2Z6J was considered as the receptor and the mol2 file of TUI was considered as the ligand. Likewise, for 4IQL, the pqr file of 4IQL was considered as the receptor and the mol2 file of FMN was considered as the ligand. The cartesian coordinates of 2Z6J and 4IQL were x=9.688, y=1.095 and z=6.557 and x = -19.26, y = 22.3899 and z = -21.9837 respectively. We defined the box size of 23*23*23 for both the redocking validations. All the other options were default. We analyzed the docked poses comparing the RMSD between the heavy atoms of the docked pose and the native ligand available in the crystal structures. RMSD between the docked

poses and the original ligands in both the crystal structures were calculated in Chimera using rmsd command from the command line. Hydrogen molecules were removed from the docked poses to make an equal number of heavy atoms between the docked ligand and the native ligands in the crystal structures.

2.2.5 Ligand preparation of multiple libraries for glide docking.

Seven different libraries were obtained from the database. They were Human Approved Drugs collection, NIH Chemical Genomics Center (NCGC) Pharmaceutical collection, Maybridge, Chembridge express, Chembridge core, Lifechemicals and Specs. We carried out the ligand preparation from the ligprep program of the Schrödinger. Ligprep helps to generate 3D ligands from their 2D structure, potential ionization states at pH 7 and +/- 2, tautomers, chiralities and other features. During the ligand preparation stage, filtering criteria were used to include the appropriate molecules in the libraries. Since Human Approved and NCGC pharmaceutical libraries are approved for use in humans, we only applied a filter for a molecular range of 150 to 800 Dalton. For the other five larger libraries, a filtering file was prepared that included criteria for molecular weight and definition of different functional groups as shown in B.1 (appendices). The final libraries contained compounds ready for docking.

Table 2.2 shows the number of compounds in each library before ligand preparation from the ligprep.

Table 2.2. Number of compounds in seven different libraries

Name of the library	Number of compounds in the library
Human Approved Drugs	7794
NCGC Pharmaceutical collection	3273
Maybridge	52160
Chembridge express	More than 480,000
Chembridge core	More than 640,000
Lifechemicals	More than 450,000
Specs compound	More than 200,000

2.2.6 Molecular docking of compound databases.

Protein preparation of the homology model was conducted in a similar way as the redocking validation of the inhibitor. This preparation was followed by the receptor grid generation for molecular docking. We assigned formal charges to the Na metal available in the active site. The force field used for grid generation was OPLS_2005. We did not include the TUI inhibitor of the model in the grid generation. It was excluded from the receptor grid by picking up the ligand. In the site tab, an outer enclosing box of 29 * 29 * 29 was considered for the receptor grid. It represents the grid box within which all docked compounds are contained. The center of the box was specified on the centroid of the workspace ligand. Size of the inner/ligand diameter midpoint box was adjusted to 13 * 13 * 13 from the advanced settings option. We considered Na+ ion and FMN in the

active site during the generation of the receptor grid. These were included in the grid box. No constraints were used in the grid generation.

We performed docking of compound libraries that were prepared from the ligprep. The receptor grid was specified, which was generated in the grid generation. Ligands were screened with FMN and Na ion available in the active site. In the beginning, the Human Approved Drugs library and NCGC pharmaceutical collection were used for glide docking. We chose options like "no docking and scoring of ligands more than 300*"*, and "number of rotatable bonds not more than 100". No constraints were used in the procedure. We chose the flexible ligand sampling to consider flexibility of the docked ligands. Other default settings remained as they were. Since Human Approved Drugs and NCGC Pharmaceutical libraries were smaller in size compared to other libraries, standard precision (SP) mode was used for docking

We selected the top ten compounds from Human Approved and NCGC libraries based on glide score (from the highest negative glide score to the lowest). For the other five larger libraries, first we conducted high-throughput virtual screening of these compounds. We sorted these compounds according to their glide ligand efficiency ln values. This normalizes the standard glide score with the number of heavy atoms present in the given ligand. We proceeded with SP docking of the top ten percent of these sorted compounds. The same procedure was repeated for Maybridge, Chembridge core, Chembridge express, Lifechemicals and Specs libraries. Default parameters were selected for docking. Moreover, the compounds from standard precision docking were also sorted according to their glide ligand efficiency ln values. The ligands were sorted from lower (higher

negative) to higher (lower negative and positive) glide ligand efficiency. The top ten compounds were chosen for analysis.

2.3 Results

2.3.1 Homology model of *C. difficile* FabK enzyme

We generated ten different homology models of *C. difficile* FabK using the Chimera-Modeller interface. These models have different individual GA341 scores, zDOPE scores, estimated root mean square deviations (RMSDs) and estimated overlap between the templates and models. The best model, 2.9, was chosen based on highest GA341 score, lowest zDOPE score, smallest estimated RMSD and the highest estimated overlap. Table 2.3 shows ten different models with their different individual scores.

Table 2.3. Top ten models of *C. difficile* FabK from Chimera-Modeller.

Model	GA341	zDOPE	Estimated RMSD	Estimated Overlap (3.5 A)
2.1	1.00	-0.72	2.628	0.862
2.2	1.00	-0.77	2.541	0.873
2.3	1.00	-0.59	2.993	0.833
2.4	1.00	-0.73	2.501	0.875
2.5	1.00	-0.74	2.582	0.865
2.6	1.00	-0.77	2.443	0.883
2.7	1.00	-0.79	2.346	0.881
2.8	1.00	-0.79	2.347	0.880
2.9	1.00	-0.83	2.298	0.885
2.10	1.00	-0.64	2.764	0.854

We subjected the model to 2500 iterations of minimization from the Schrödinger suite to refine the structure and minimize the bad contacts. The minimized structure was further analyzed. The quality of the selected model was assessed using the Ramachandran plot, PROCHECK scores, verify 3D and QMEAN6 score.

The Ramachandran plot generated from the PROCHECK result of the selected final model is shown in Figure B.2 (in Appendix). It gives information about the residues in different regions that include the favored, allowed, generously allowed and disallowed regions. The Ramachandran plot/statistics shows that 0.4% of the residues fall in the disallowed, 91.3 % in the most favored, 7.2% in the additionally allowed and 1.1% in the generously allowed regions.

Similarly, the PROCHECK program generated several stererochemical properties of the final model. Approximately 97.2 % of the residues had appropriate bond length, and 90 % of them had bond angle within the selected limit. All the planar groups of the side chains present in the protein were within the limit.

Detailed plots of bond length, bond angle and major distances from planarity are shown in Figures B.3, B.4 and B.5 (in Appendices). Major stererochemical properties from the Procheck can be summarized in Figure 2.2.

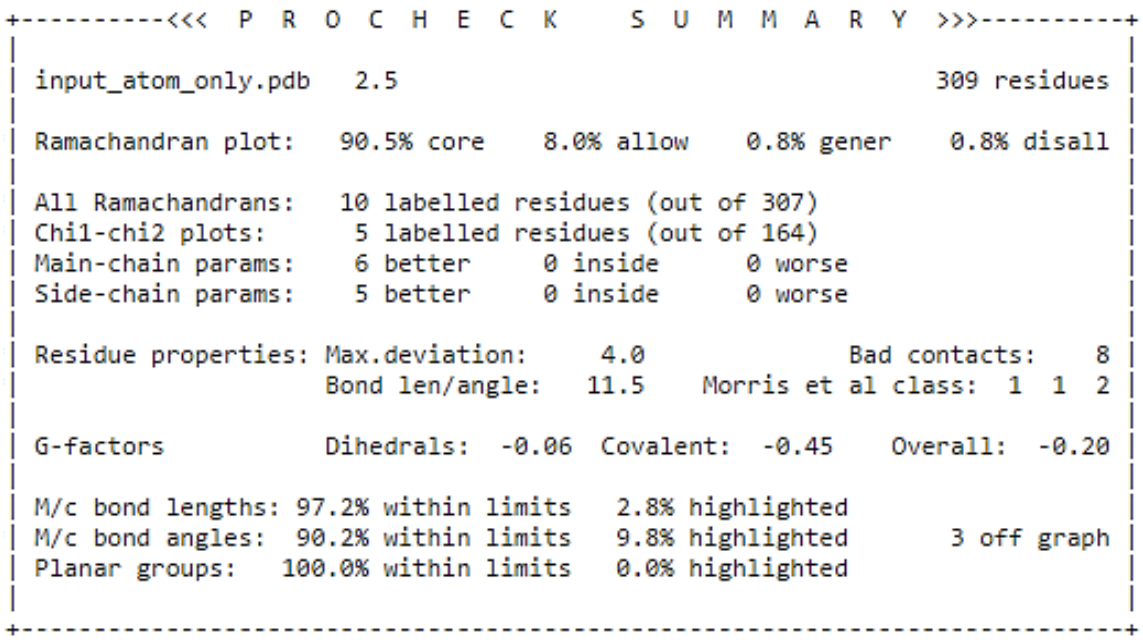


Figure 2.2. The PROCHECK summary of the *C. difficile* FabK homology model.

Verify 3D results

The verify 3D result of the final selected model showed that 94.50 % of the residues had a mean 3D-1D score greater than or equal to 0.2. The pass criteria for verify 3D is that at least 80 % of the amino acid residues should have scored greater than or equal to 0.2 in the 3D/1D profile.

QMEAN 6 result

QMEAN6 evaluation of the model was performed using the Swiss expasy web server. The various components of the QMEAN6 score with their raw and Z scores are shown in Table 2.4 below.

Table 2.4. QMEAN6 score of *C. difficile* FabK homology model

Scoring function term	Raw score	Z-score
C_beta interaction energy	-84.39	-1.05
All-atom pairwise energy	-5342.17	-1.48
Solvation energy	-39.54	0.80
Torsion angle energy	-73.74	-0.67
Secondary structure agreement	82.8%	0.16
Solvent accessibility agreement	83.8%	0.61
QMEAN6 score	0.775	0.01

The average QMEAN6 score of the model was 0.775. The values of a QMEAN6 score ranges from 0 to 1. This score consists of combinations of scores of several components that have an important role to play in the structure. The higher the score of the QMEAN, the better is the structure

2.3.2 Sequence identity of the active sites of the homology model

We matched the selected residues of the active site of the homology model with the two templates. After the superposition, structure-based multiple sequence alignment was conducted for the model and the two templates. We calculated the percentage sequence identity of the active site residues with the templates. Figures 2.3 and 2.4 show the sequence alignments of the active site residues of the model with the templates 2Z6J and 4IQL separately.

Figure 2.3. Sequence alignment of the active site residues of the model with 2Z6J

Pink region inside the box shows the sequence alignment of the residues of the active site with the residues of the template 2Z6J. There are 19 residues in the active site. 18 of the 19 residues of the active site matched with the structural template. The sequence identity of the active site is 95%.

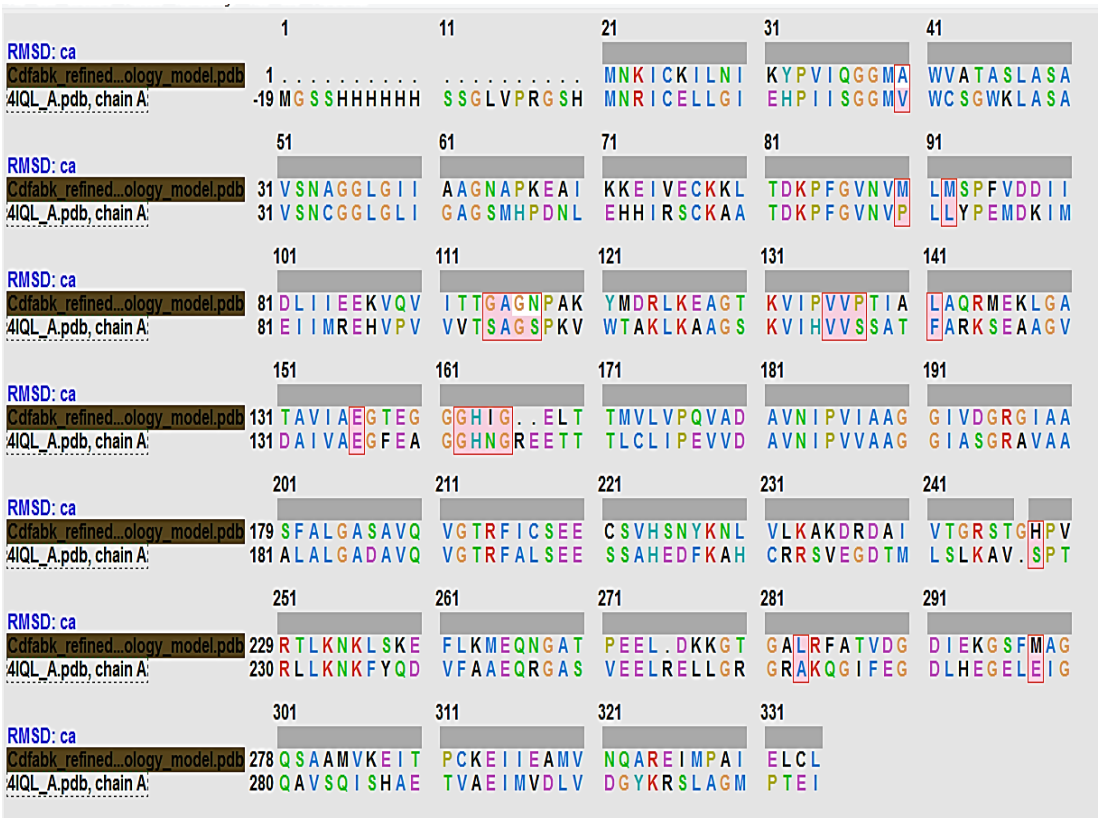


Figure 2.4. Sequence alignment of the active site residues of the model with 4IQL

In the figure, pink region inside the box shows the sequence alignment of the residues of the active site with the residues of the template 4IQL. 8 of the 19 residues of the active site matched with the residues of the template 4IQL.The sequence identity is 42%.

The sequence alignment of the active site residues of the model with the template 2Z6J showed high sequence identity (95%) compared to the overall sequence identity of 58%. Similarly, the sequence alignment of the active site residues of the model with the template 4IQL showed sequence identity of 42% compared to the overall sequence identity of 46%. Utilizing multiple templates improved the sequence identity of the active site residues.

2.3.3 Redocking validation of the native TUI inhibitor of the modeled structure

Redocking was performed by extracting the TUI inhibitor presented in the homology model to validate the process. The docking poses were evaluated comparing the calculated RMSD between the coordinates of the heavy atoms of the docked ligand to that of the TUI inhibitor of the model. The RMSD was calculated in Chimera using the command rmsd in the command line. Table 2.5 represents RMSD between the three docked poses and the original TUI inhibitor presented in the homology model.

Table 2.5. RMSD measurement of top docked poses of redocking validation

Compound	RMSD between the pose and native ligand (in Angstrom)
Pose 1	1.359
Pose 2	11.207
Pose 3	1.551

From Table 2.5, it is seen that the best RMSD is 1.359 Angstrom between pose 1 and the native TUI of the homology model (lies in the 2 A^{o} cutoff).

The three dimensional visualization of the best-docked pose (pose 1) with the native ligand available in the homology model is shown in Figure 2.5 below. The top docked pose 1 is colored by their element properties. The original TUI of the homology model is shown in green. The FMN cofactor is shown in grey. Sodium ion is shown in yellow.

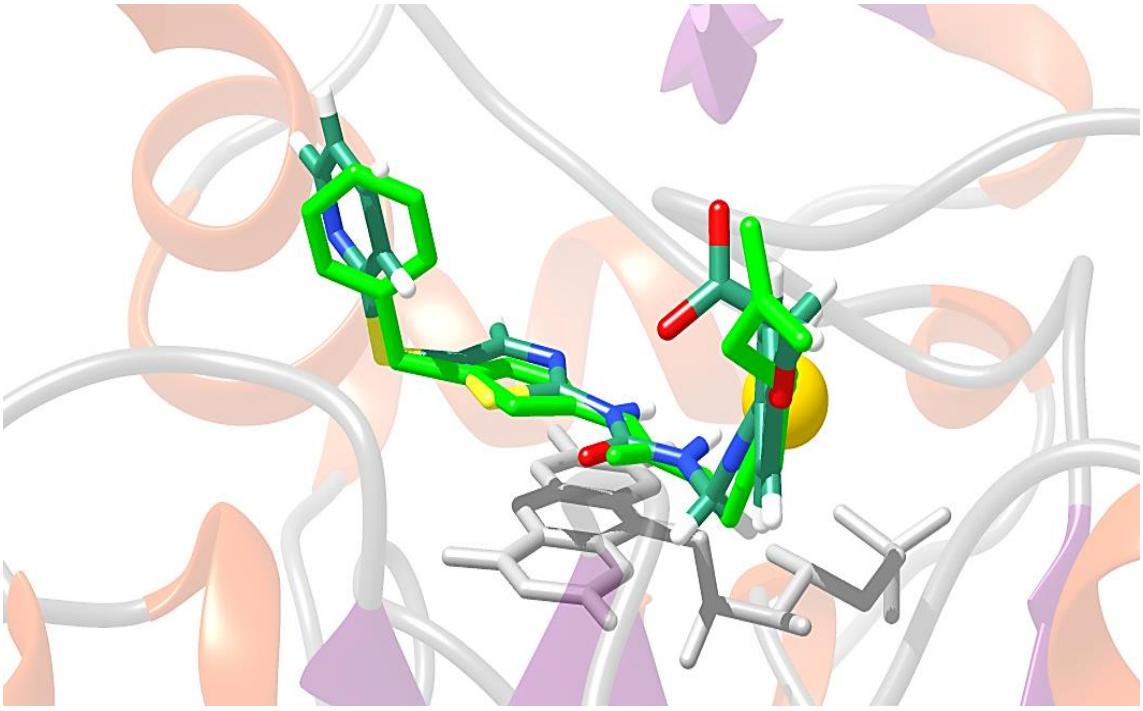


Figure 2.5. Top docked pose with the native ligand in the receptor

2.3.4 Redocking validation of the ligands of the crystal structures

We conducted the redocking validation of both TUI and FMN of 2Z6J and 4IQL respectively. The poses obtained from redocking validation were evaluated comparing the RMSD between the heavy atoms of the ligand of the docked pose to that of the ligand of the template. Table 2.6 shows the RMSD between the docked poses and the original ligand FMN of 4IQL. Table 2.6 shows that the best RMSD is 1.004 Angstrom between Pose 2 and FMN of 4IQL

Table 2.6. RMSD between the docked poses and FMN of 4IQL

Compound	RMSD between the docked pose and FMN
Pose 1	4.607
Pose 2	1.004
Pose 3	2.584
Pose 4	6.706
Pose 5	7.083
Pose 6	6.630
Pose 7	6.665
Pose 8	6.717
Pose 9	4.855

The 3D visualization of the best-docked pose (Pose 2) as well as FMN of 4IQL is shown in Figure 2.6 below.

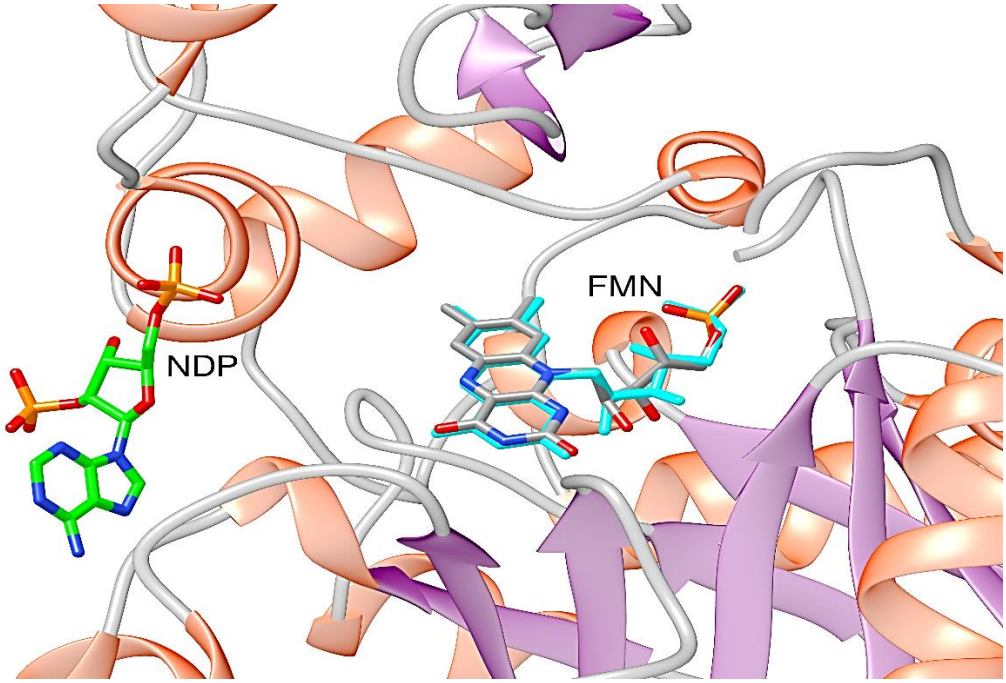


Figure 2.6. Top redocked pose with FMN of 4IQL

The top docked pose (Pose 2) is colored by their element properties. The FMN of 4IQL is shown in cyan and NDP (left) is shown in heteroatoms.

Similarly, we also conducted the redocking validation of the TUI on 2Z6J (FMN present on the active site). Table 2.7 shows the RMSD between the docked poses and the original TUI of 4IQL.

Table 2.7. RMSD between the docked poses and TUI of 2Z6J

Compound	**RMSD between the docked pose and original TUI of 2Z6J**
Pose 1	6.44
Pose 2	6.747
Pose 3	6.11
Pose 4	6.33
Pose 5	10.183
Pose 6	10.611
Pose 7	12.407
Pose 8	4.24
Pose 9	9.577

The RMSDs between the docked poses and the TUI of 4IQL are higher than 2 Angstroms. Table 2.7 shows that the best RMSD is 4.24 Angstroms between Pose 8 and the TUI of 4IQL. The best-docked pose (Pose 8) and native TUI of 2Z6J are shown in Figure 2.7 below

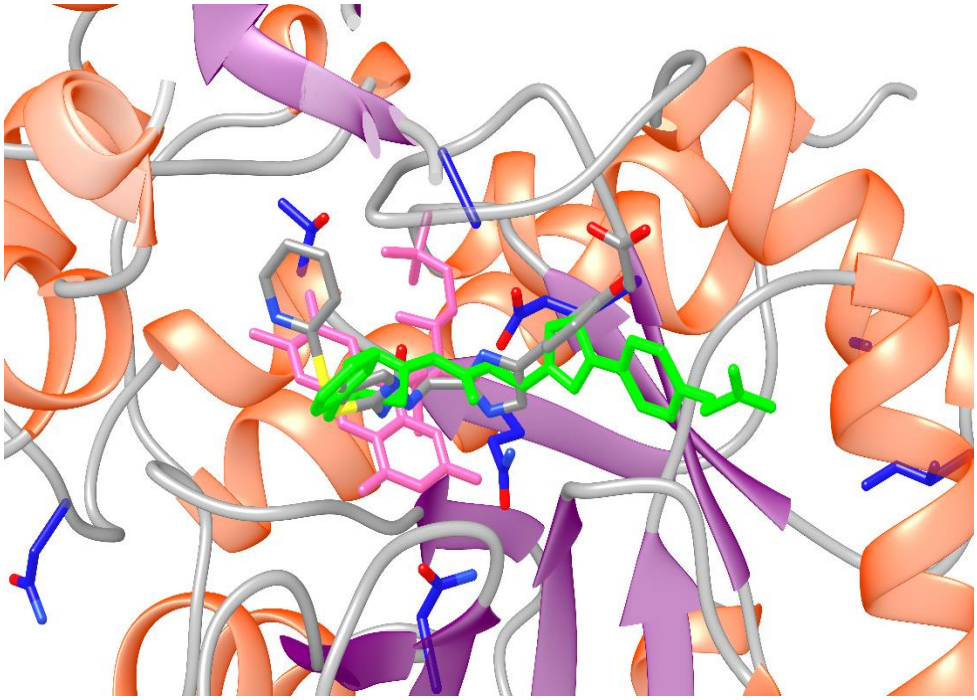


Figure 2.7. Top docked pose with TUI of 2Z6J

The top docked pose (Pose 8) is shown by their element properties. The original TUI in 2Z6J is represented in green. FMN is depicted in pink.

2.3.5 Docking of compound libraries

We performed the glide docking of seven different libraries after ligand preparation from the ligprep program of the Schrödinger software. For Human Approved Drugs and NCGC pharmaceutical collection libraries, glide SP docking was performed (no HTVS docking for these libraries). We selected the compounds from these libraries based on the glide score. For five other libraries, initial HTVS docking was performed. The compounds obtained from these HTVS docking were sorted according to the glide ligand efficiency ln score and the top ten percent compounds were extracted. This was followed

by the glide SP docking of the selected top ten percent compounds. The number of compounds obtained in HTVS and SP docking are shown in Table 2.8 below.

Table 2.8. Number of compounds obtained for HTVS and SP docking

Name of the library	No. of compounds from HTVS docking	No. of compounds from SP docking
Human approved drugs	None	6372
NCGC Pharmaceutical collection	None	7982
Maybridge	43781	4378
Chembridge express	550856	55085
Chembridge core	1766180	176618
LifeChemicals	572848	57286
Specs compound	226131	22613
Total	3,159,796	330,334

After glide SP docking of these compounds, we again sorted the compounds according to the glide ligand efficiency. The top ten compounds from each class were represented in table format.

Human Approved Drugs library

This library contains FDA approved drugs for the treatment of diseases in human beings. The ligands from the SP docking were sorted based on their glide score (from highest negative to the lowest negative/positive). Table 2.9 shows the top ten compounds obtained from Human Approved Drugs molecular docking. Salmeterol had the best glide score of -9.829. All top ten compounds had scores greater than -8.6.

Table 2.9. Top ten compounds from Human Approved Drugs Library

Compounds	Molecular weight	Pubchem_cid	Glide gscore
Salmeterol	415.578	5152 6604001 56801	-9.829
Thiamphenicol	356.227	27200 146678 5433	-9.308
Pranlukast	481.515	115100	-9.29
Domperidone	425.922	3151 145924 6604595 24870822	-9.106
Glibenclamide	494.014	3488	-8.792
Tetragastrin	596.711	446569	-8.788
Cephaloglycin	405.433	19150	-8.751
Latamoxef	520.481	24870866 6604567 16757704 24871007 47499	-8.697
Ketoconazole	531.443	456201 47576 5702077 16757695	-8.683
Betiatide	367.383	185457	-8.654

Figures 2.8 and 2.9 show Salmeterol docked into the active site of the receptor and the ligand interaction diagram respectively. Salmeterol is colored based on their element properties (shown in ball and stick representation). The FMN is shown in tube (pink). The receptor is shown in ribbon. The Sodium ion is represented in CPK (purple).

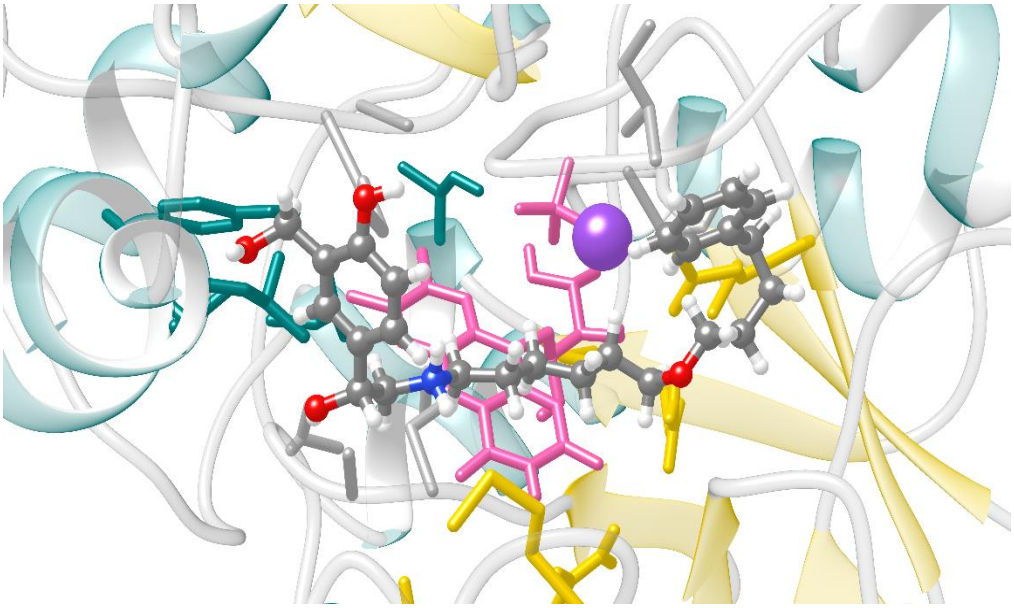


Figure 2.8. Salmeterol docked in the binding pocket of receptor

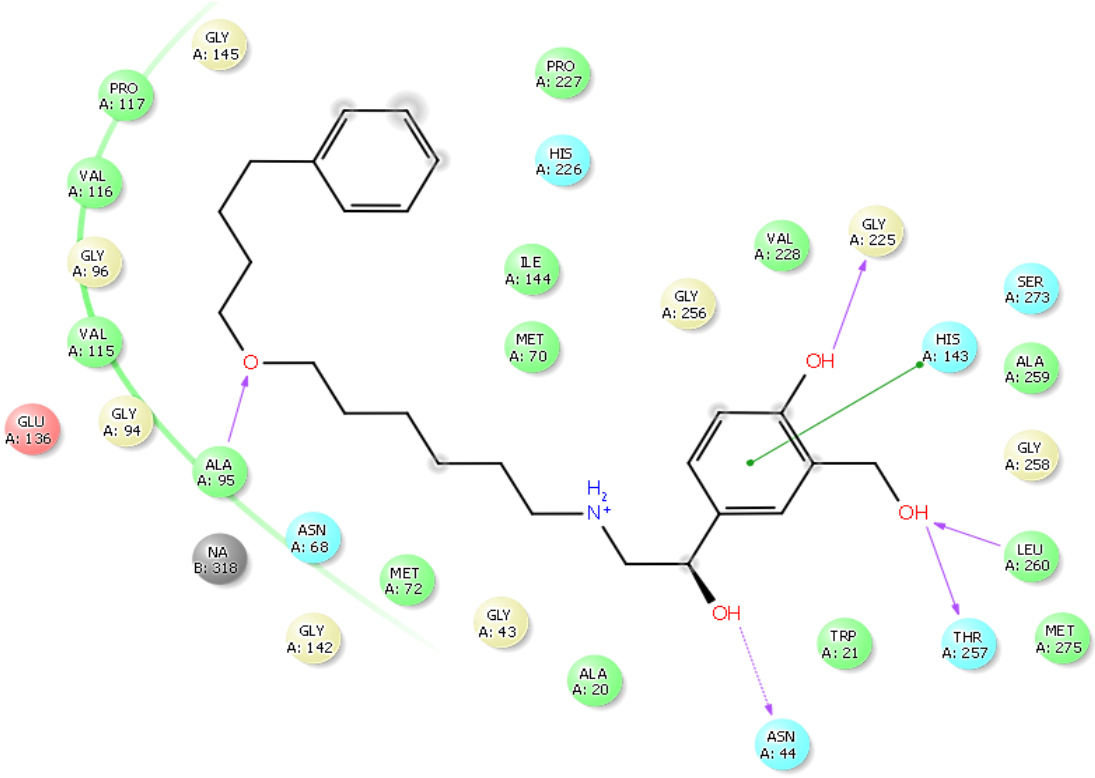


Figure 2.9. 2D interaction of Salmeterol with key residues

Sodium (NA) is shown in grey. Purple lines show hydrogen bond, and green lines represent Pi-Pi stacking between the molecules.

NCGC Pharmaceutical collection library

The top ten compounds in the NCGC library are shown in Table 2.10 below.

Table 2.10. Top ten compounds from NCGC pharmaceutical collection

Compounds	Molecular weight	Pubchem_cid	Glide gscore
Salmeterol	416.586	5152 6604001 56801	-9.829
Thiamphenicol	356.227	27200 146678 5433	-9.308
Pranlukast	480.507	115100	-9.29
Domperidone	426.93	3151 145924 6604595 24870822	-9.106
Glibenclamide	494.014	3488	-8.792
Latamoxef	518.465	24870866 6604567 16757704 24871007 47499	-8.697
Ketoconazole	531.443	456201 47576 5702077 16757695	-8.683
Betiatide	366.375	185457	-8.654
Entecavir	277.285	170343 153941	-8.648
Talampicillin	482.539	71446 24870952 6604402	-8.64

As in the Human Approved Drugs library, Salmeterol was the top-ranked compound with the highest glide score -9.829 (the highest negative value). All top ten compounds had scores above -8.6.

The top docked pose (Salmeterol) of the NCGC pharmaceutical collection is as same as Human Approved Drugs collection library. We have already shown the 3D visualization and key interactions of Salmeterol from the Human Approved Drugs library.

Maybridge library docking

The top ten compounds obtained from glide SP docking of Maybridge library are shown in Table 2.11 below.

Table 2.11. Top ten compounds from Maybridge library

Compound code	Molecular weight	logP value	Glide ligand efficiency ln score
S14092	329.738	1.75	-2.485
HTS01959	383.358	1.97	-2.480
HTS01978	350.156	2.11	-2.404
NRB04602	190.225	1.84	-2.367
FM00043	168.219	0.89	-2.367
HTS01851	339.302	0.37	-2.347
HTS01858	317.32	0.02	-2.335
SCR00961	286.329	2.05	-2.329
CD04894	412.271	3.86	-2.296
SCR00955	383.451	1.26	-2.294

HTVS docking of the Maybridge library resulted in 43781 compounds. The results were sorted according to the glide ligand efficiency score. We used the top ten percent compounds (4378) from HTVS for glide SP docking.

The compound with code S14092 has the highest ligand efficiency score of -2.485. Figures 2.10 and 2.11 show 3D visualization and the ligand interaction diagram of the top docked pose (S14092) respectively.

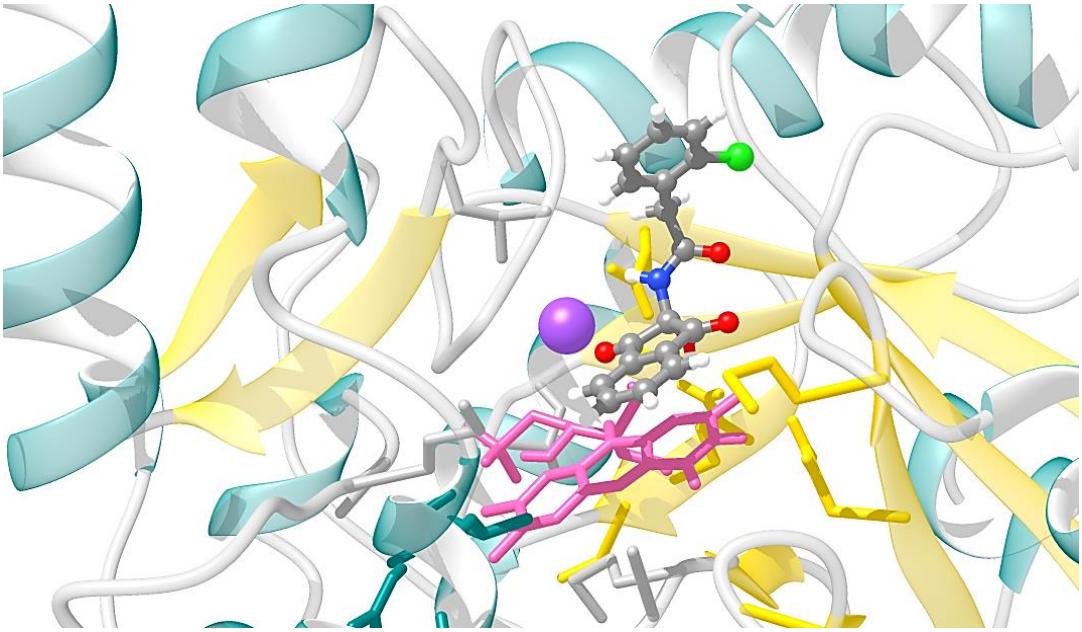


Figure 2.10. Compound S14092 of Maybridge docked in the receptor

Compound S14092 is colored based on their element properties (shown in ball and stick representation). The pink ligand represents the cofactor FMN. The receptor is shown in the ribbon. The sodium ion is represented in CPK (purple).

Compound S14092 shows key interactions with residues of the binding site. These key interactions are illustrated in the ligand interaction diagram in Figure 2.11 below.

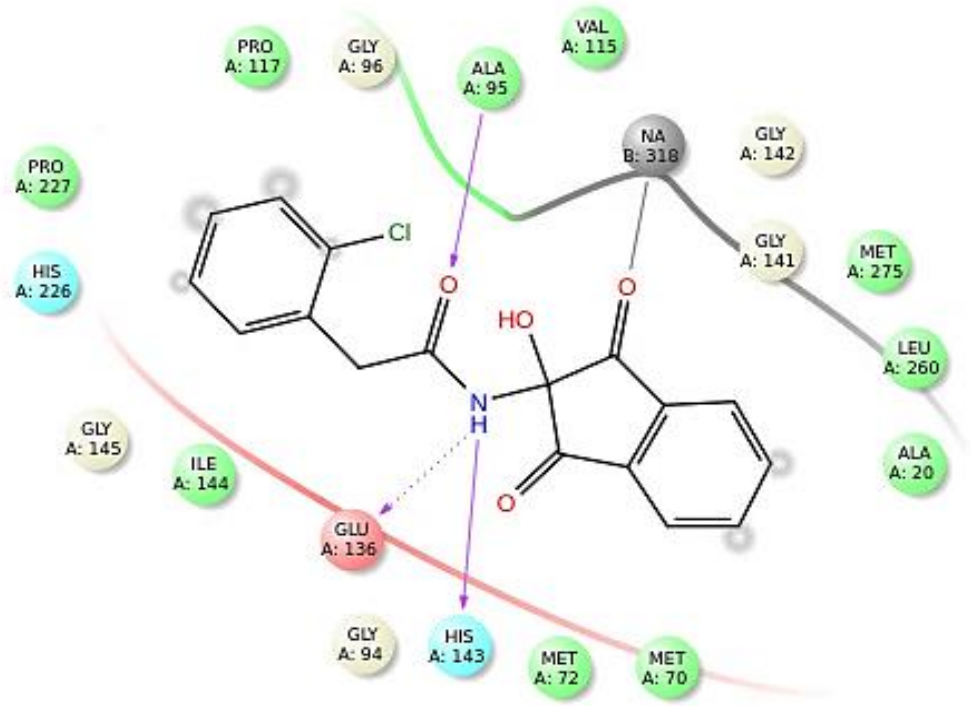


Figure 2.11. 2D interaction of compound S14092 with residues of binding site

Red amino acid residues have negative charges; blue are polar residues; and green are hydrophobic residues. Sodium (NA) is shown in grey. Purple lines show hydrogen bond, and green lines represent Pi-Pi stacking between the molecules.

Chembridge express library

550856 compounds were obtained from HTVS glide docking of this library. Compounds were sorted based on the glide ligand efficiency ln score. We used the selected top ten percent of the compounds (55085) for more advanced glide SP docking. Again, the compounds were categorized according to the ligand efficiency. The top ten compounds from the final SP docking result are shown in Table 2.12.

Table 2.12. Top ten compounds from Chembridge express

Compound ID	Molecular weight	cLogP	Glide ligand efficiency ln score
7567382	594.472	4.412	-2.475
7113385	261.26	0.74	-2.473
9214223	230.228	1.944	-2.468
9202609	308.383	3.258	-2.434
5624606	315.375	4.03	-2.431
9033325	307.355	2.857	-2.427
7998731	393.424	3.63	-2.409
6046846	298.276	2.07	-2.406
5728840	337.315	2.15	-2.404
7849307	275.375	3.66	-2.401

The compound with ID 7567382 had the best ligand efficiency score of -2.475 and clogP value of 4.412. All the compounds had ligand efficiency scores of more than -2.0 and molecular weight from 230 to 595 Dalton

Figures 2.12 and 2.13 depict binding of the compound 7567382 in the receptor and the ligand interaction diagram respectively. Compound S14092 is colored based on their element properties (shown in ball and stick representation). FMN is shown in pink. The receptor is shown in ribbon. The sodium ion is represented in CPK (purple).

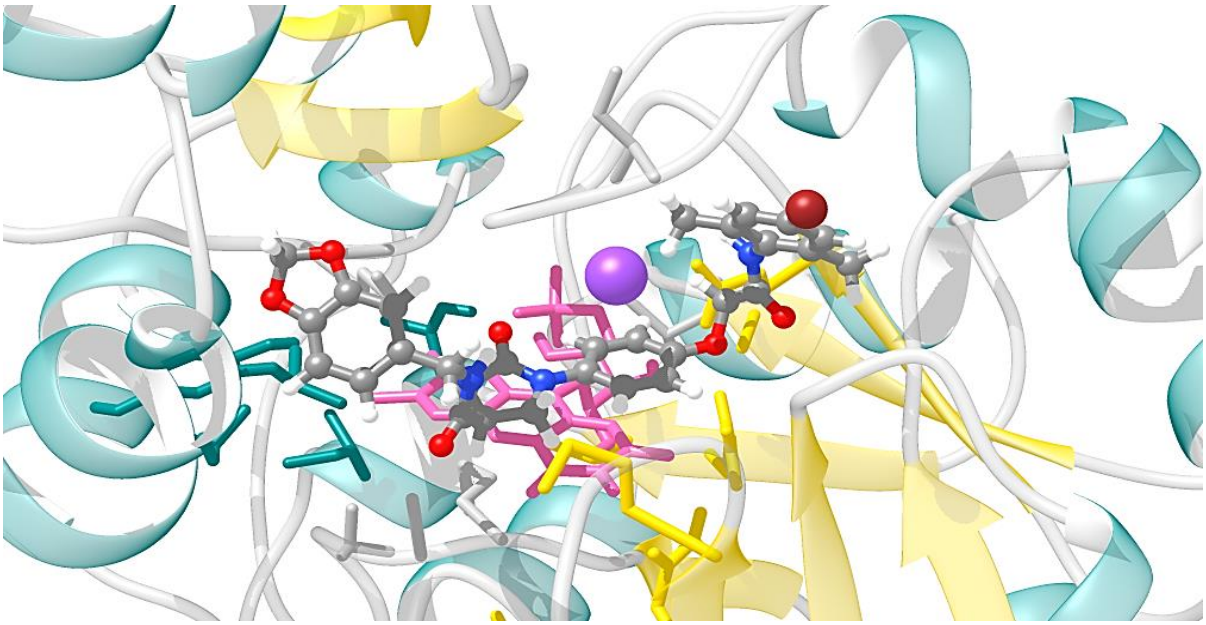


Figure 2.12. Compound 7567382 of Chembridge express docked into receptor

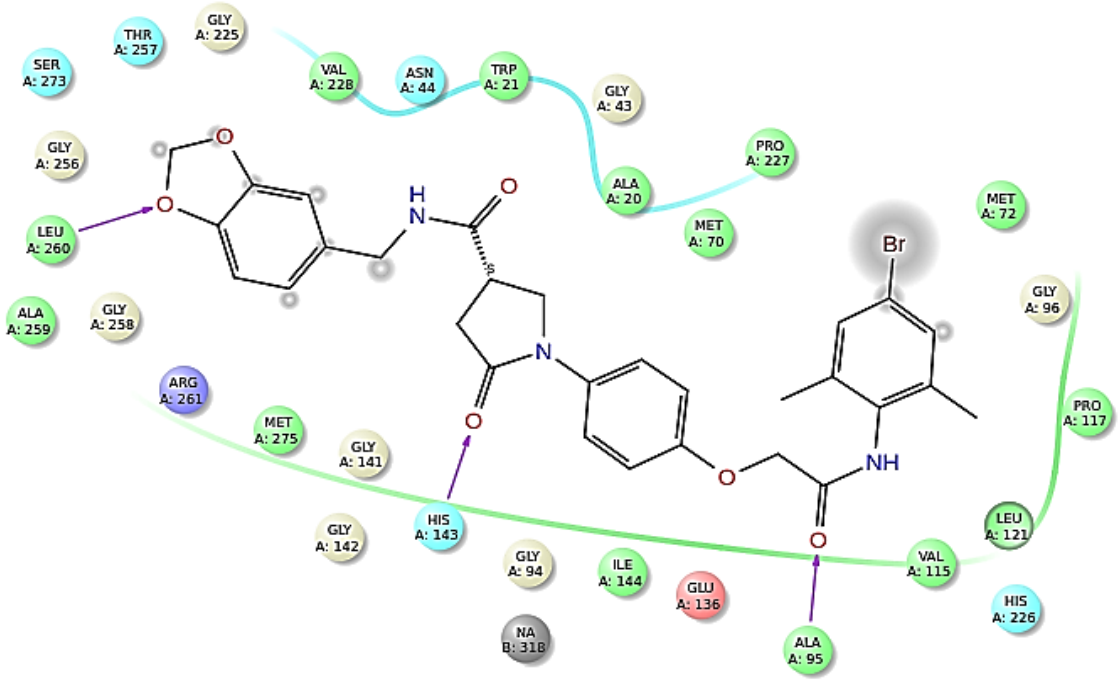


Figure 2.13. 2D interaction of Compound 7567382 in the receptor

Red amino acid residues have negative charges; blue are polar residues; and green are hydrophobic residues. Sodium (NA) is shown in grey. Purple lines show hydrogen bond, and green lines represent Pi-Pi stacking between the molecules.

Chembridge Core library

Among all the libraries that we docked, this is the one containing the largest number of compounds. HTVS glide docking provided 1766180 compounds. After sorting these compounds based on the ligand efficiency score, we chose the top ten percent of these compounds for glide SP docking. Glide SP docking provided nearly 176618 compounds. Lastly, we categorized the SP docked compounds based on the glide ligand efficiency. The top ten compounds from SP docking results are depicted in Table 2.13 below.

Table 2.13. Top ten compounds from Chembridge core

Compound ID	Molecular weight	clogP	Glide ligand efficiency ln score
77077035	279.278	-0.6	-2.532
58649928	284.32	0.57	-2.530
66161357	357.416	-0.54	-2.487
7567382	594.472	4.412	-2.475
7113385	261.26	0.74	-2.473
9214223	230.228	1.944	-2.468
97483872	346.408	1.92	-2.463
10835968	396.428	1.53	-2.454
74288240	291.377	0.02	-2.441
34788033	321.379	1.03	-2.438

The compound with ID 77077035 had the highest score of -2.53. It had clogP value of -0.6. Some of the top compounds from express also scored well in core library. Figures

2.14 and 2.15 represent the binding mode and the ligand interaction of Compound 770770335 in the receptor respectively.

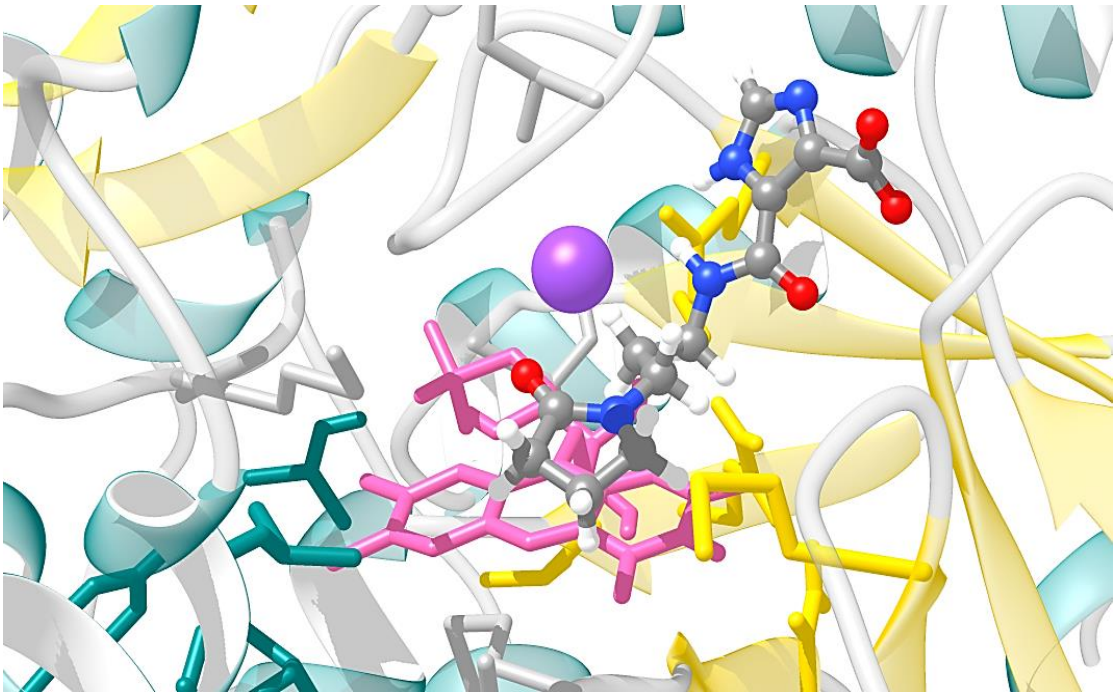


Figure 2.14. Compound 77077035 docked in the receptor-binding site

The compound 77077035 is colored based on their element properties (shown in ball and stick representation). FMN is shown in pink. The receptor is shown in ribbon. The sodium ion is represented in CPK (purple).

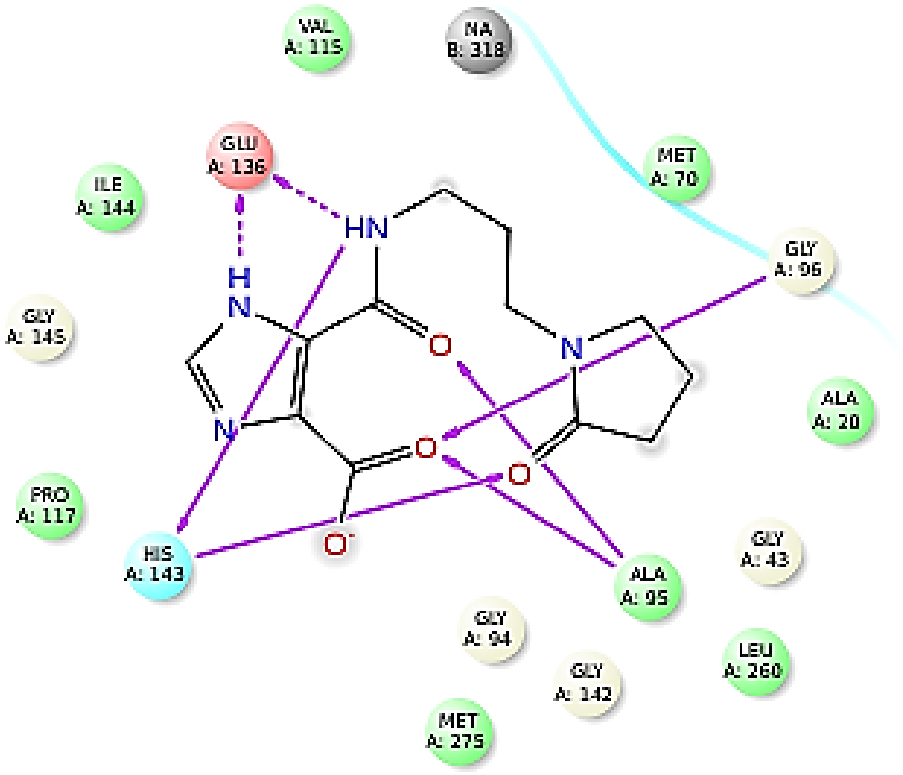


Figure 2.15. 2D interaction of Compound 77077035 with important residues

Red amino acid residues have negative charges; blue are polar residues; and green are hydrophobic residues. Sodium (NA) is shown in grey. Purple lines show hydrogen bond, and green lines represent Pi-Pi stacking between the molecules.

LifeChemicals compound library

HTVS glide docking of the Lifechemicals library yielded 572848 compounds that we sorted based on their ligand efficiency scores. The top ten percent of the compounds were used for glide SP docking and provided 57286 compounds. The selected top ten percent compounds were sorted according to their ligand efficiency scores (highest negative to lowest negative/positive). The top ten compounds obtained from the sorted SP docking result are shown in Table 2.14 below.

Table 2.14. Top ten compounds from Lifechemicals library

Compound ID	Molecular weight	clogP	Glide ligand efficiency ln score
F6413-0485	406.529	4.56	-2.523
F5060-0158	316.297	-0.05	-2.511
F2783-0073	344.417	3.83	-2.506
F6413-1972	406.529	4.56	-2.5
F6413-1957	378.422	3.17	-2.488
F6418-1972	408.545	3.66	-2.48
F6413-2949	372.396	3.48	-2.477
F2024-1610	362.407	2.21	-2.475
F5097-2910	417.512	2.06	-2.474
F6413-0186	428.508	4.05	-2.473

Compound F6413-0485 has the best ligand efficiency score (-2.523) with clogP 4.56. Figure 2.16 shows compound F6413-0485 bound to the receptor. Figure 2.17 represents the ligand interaction diagram of the compound F6413-0485.

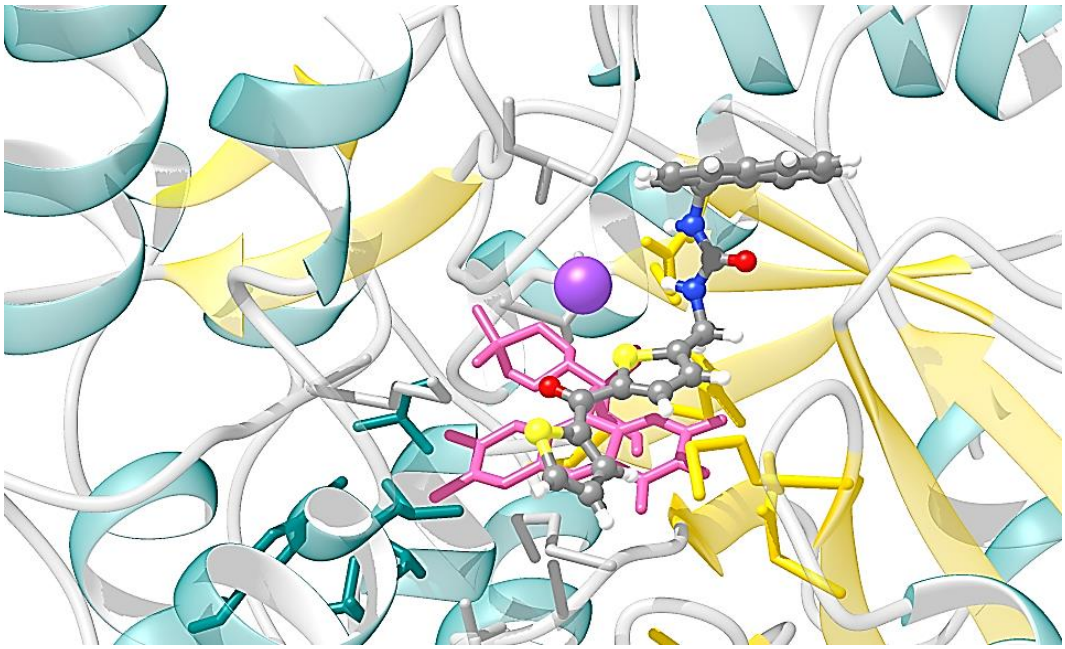


Figure 2.16. Compound F6413-0485 docked in the binding site of receptor

The compound F6413-0485 is colored based on their element properties (shown in ball and stick representation). FMN is shown in pink. The receptor is shown in ribbon. The sodium ion is represented in CPK (purple).

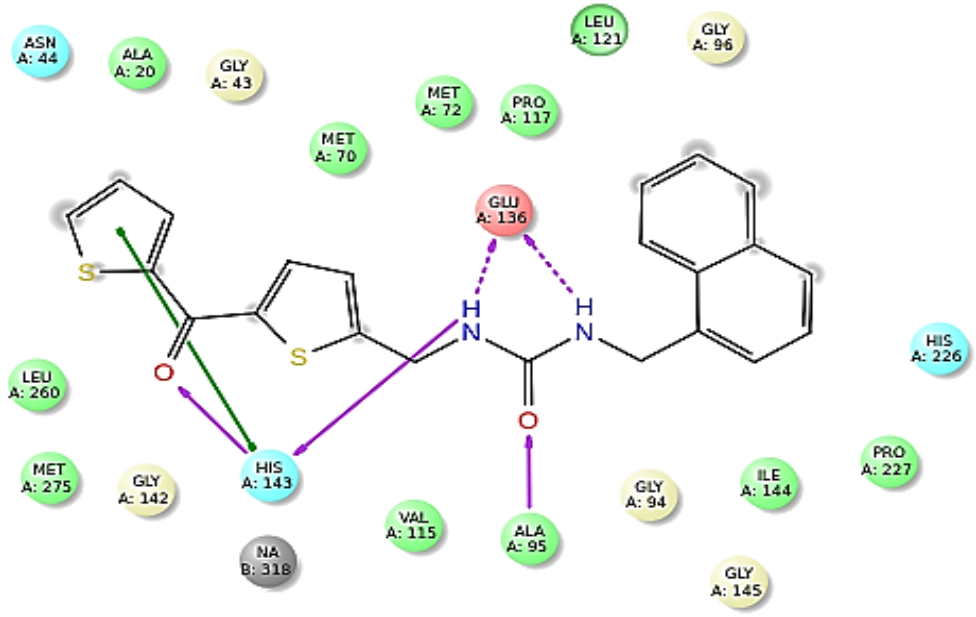


Figure 2.17. 2D interaction of Compound F6413-0485 in the receptor

Red amino acid residues have negative charges; blue are polar residues; and green are hydrophobic residues. Sodium (NA) is shown in grey. Purple lines show hydrogen bond, and green lines represent Pi-Pi stacking between the molecules.

Specs Compound library

Initially we performed HTVS glide docking. We got 226131 compounds from this result. These compounds were sorted based on glide ligand efficiency. The top ten percent were utilized for further glide SP docking. The compounds were categorized according to ligand efficiency scores from the highest to the lowest. Table 2.15 shows the top ten compounds of SP docking results.

Table 2.15. Top ten compounds from Specs chemical library

Compound ID	Molecular weight	clogP	Glide ligand efficiency ln score
AF-399/42487793	305.383	3.28	-2.395
AK-968/41018290	278.7	1.16	-2.383
AH-357/03489045	190.225	1.45	-2.367
AP-185/15474006	232.197	-0.27	-2.361
AT-057/43348336	310.356	2.78	-2.341
AN-329/41328917	326.399	2.86	-2.335
AN-329/43450308	372.855	3.18	-2.33
AN-988/15131258	584.743	3.06	-2.323
AO-476/43407140	298.307	-1.5	-2.321
AO-081/40847547	358.828	2.72	-2.307

The compound AF-399/42487793 displayed best ligand efficiency score (-2.395) and had clogP of 3.28. Figures 2.18 and 2.19 depict the binding mode of the docked pose and the ligand interaction diagram of AF-399/42487793 respectively.

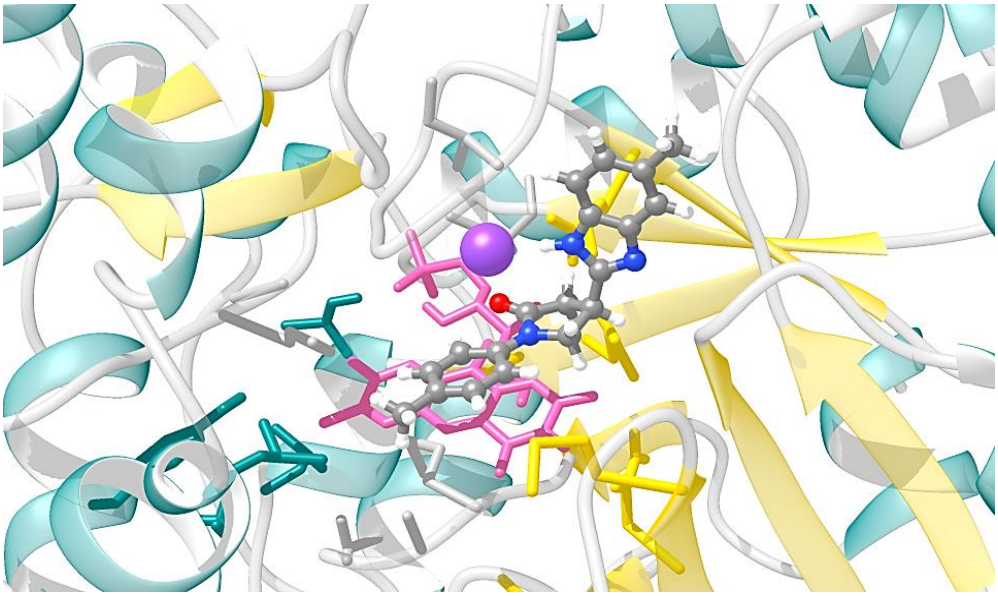


Figure 2.18. Compound AF-399/42487793 docked in the binding site

The compound AF-399/42487793 is colored based on their element properties (shown in ball and stick representation). FMN is shown in pink. The receptor is shown in ribbon. The sodium ion is represented in CPK (purple).

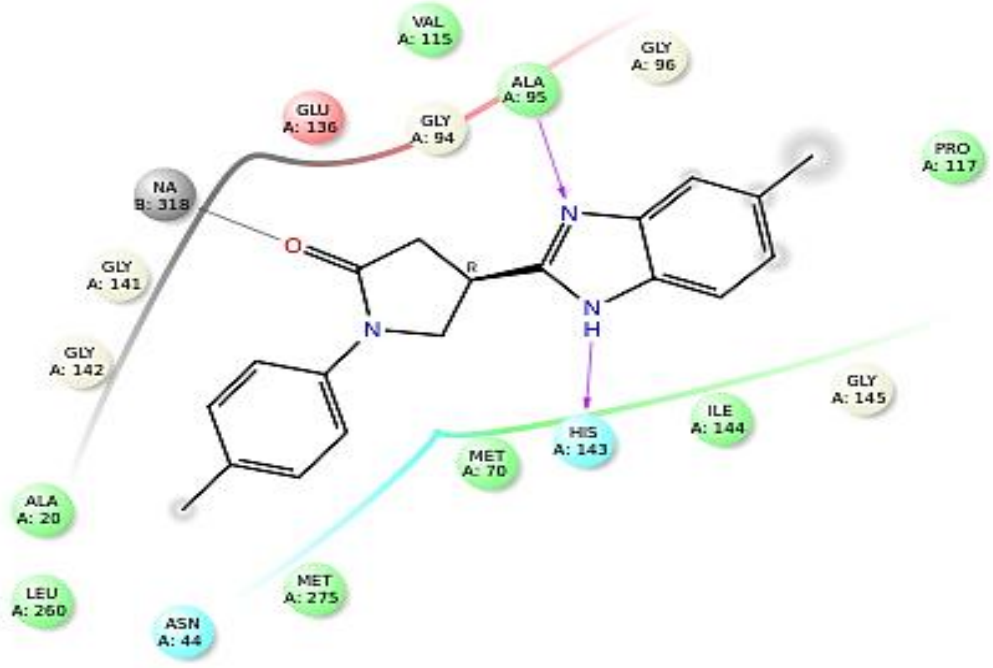


Figure 2.19. 2D interaction of AF-399/42487793

Red amino acid residues have negative charges; blue are polar residues; and green are hydrophobic residues. Sodium (NA) is shown in grey. Purple lines show hydrogen bond, and green lines represent Pi-Pi stacking between the molecules.

2.4 Discussion

The goal of this project was to create a *C. difficile* FabK model and perform molecular docking of seven different compound libraries. From this result, we obtained inhibitors that can be tested in experimental screening. Virtual screening is a useful approach in CADD that helps to screen larger libraries and obtain potential hits.

2.4.1 Homology model of *C. difficile* FabK

We created the homology model of *C. difficile* FabK using two different templates: 2Z6J_A and 4IQL_A. Both the templates had sequence identities over 40 % making them good templates for our purpose. We realigned the sequences using Clustal Omega, which is a useful alignment tool for multiple sequence alignment. All ten different generated models had GA341 scores of 1, indicating the folds of all the models are of good quality. A value greater than 0.7 indicates that the model is good and that there is a 95 % probability of having accurate folds in the model structure.[202] The estimated backbone Ca RMSD values of all the models are less than 3 Angstroms. From the results, it is seen than four of the ten different models have RMSD less than 2.5 A, which shows that these models are in acceptable agreement with the template structures. As seen from the table, model 2.9 is considered the best model as it has the best RMSD and zDOPE score and possesses good overlap with the template. The estimated overlap of approximately 89 percent depicts that nearly this percentage of C-alpha atoms of this model structure lie within 3.5 A^0 of the respective atoms in the template structures after superposition. The

chosen model structure could provide us with the starting structure of *C. difficile* Fabk enzyme for further evaluation of stererochemical and other properties.

After the required steps of minimization, the selected model was subjected to further evaluation. The Ramachandran plot showed that the majority of residues are in the favored region. A good quality model is assumed to have more than 90 % of the residues in the most favored region of the model. Only a single residue in the model is in the disallowed regions. Visual inspection of the model showed that Serine 223 lies in the loop regions and does not fall in the active site of the model. Loop regions in the protein are considered highly flexible and difficult to model. This would suggest that the active site containing FMN and sodium ion has no problematic amino acid residues around it.

A good overall compatibility of the three-dimensional structure of the model with the amino acid sequence is supported by verify3d results. The plot of verify 3D shows that a few of the residues in some loop regions (those having residue numbers 116, 117 and 298 to 309) have less compatibility of 3D structure with the amino acid sequence. This could be because 3D structure of loop regions are hard to model from the amino acid sequence. We decided to accept the selected model for further steps after evaluating the overall properties from various programs.

2.4.2 Redocking validation of the inhibitor of the model structure.

Redocking was used to determine whether the docking parameters we used effectively predict the pose of the ligand to that of the original homology model. Docking validation is necessary to obtain best outcome for future docking of compound libraries. Flexible ligand docking was incorporated, resulting in three poses. Out of three different docked poses of ligand generated from glide docking of the native TUI, two gave acceptable

results with a backbone Ca RMSD less than 1.5 Angstroms. The visual inspection of these two poses also showed good orientation of the ligand in compatible with the original ligand of the binding pocket. One of the poses gave a relatively higher RMSD value. Viewing the particular pose in chimera revealed that it had a slight different orientation of the docked ligand in this pose compared to the native ligand. Since the majority of poses were able to produce good results, we concluded that this docking validation method provided us with starting parameters for docking compound libraries.

2.4.3 Molecular docking of compound libraries

We continued the glide docking procedure for seven different libraries, using the inner boundary box of 13 * 13 * 13 in grid generation. Compounds that we docked from several libraries may have compounds of different sizes (larger than the native inhibitor present in the model).

The first round of HTVS docking yielded more than 3 million compounds. The top ten percent were sorted according to the ligand efficiency scores for glide SP docking. The top ten compounds obtained from all these libraries have ligand efficiency scores of more than -2. An interesting thing is that the highest scoring compound of each library has key interaction with histidine 143 residue (forms hydrogen bond with most of the residues). Furthermore, this particular residue also displayed Pi-Pi interaction with some top compounds like Salmeterol and F6413-0485 as seen from the ligand interaction diagrams in Figures 2.8 and 2.16 respectively**.**

FabK is considered a flavoprotein enzyme and the reaction is determined by the complex formation of NADPH and a FMN enzyme and follows a Ping-Pong reaction mechanism to reduce the substrate.[203] Similarly, the crystal structure of FabK enzyme of *S.*

pneumoniae has shown that histidine 144 acts as a catalytic residue, the conformation of which changes upon complex formation of NADPH and FMN.[186] This helps us assume that histidine 143 also plays some role in the catalytic process of NADPH - FMN reaction in FabK of C. difficile.

Similarly, compounds Salmeterol, ID S14092, ID 7567382, ID 77077035, F6413-0485 and AF-399/42487793 form a hydrogen bond with Alanine 95. A study conducted on the FabK enzyme of *S. pneumoniae* has shown that conformational alteration takes place in the major chains of Glycine 95 and Alanine 96 in the loop region, and forms hydrophobic interactions with phenyimidazole inhibitors.[204] Some of the top compounds also showed interactions with Sodium ion (metal ion) present in the active site. This study showed that docked compounds have some key interactions in the binding site as seen in the previous study of FabK enzymes of other organisms. Binding of drugs with these key interactions might help to identify novel inhibitors of the FabK enzyme.

2.5 Conclusion

By combining homology modeling and molecular docking, we determined several compounds that can act as potential inhibitors of the FabK enzyme of *C. difficile*. Homology modeling was used to predict the structure of the enzyme and the selected model was analyzed for its properties using various programs. Redocking validation prior to molecular docking was utilized to check whether the docking process reproduces the pose of the original inhibitor of the binding site. The docking validation also provided us some parameters for future library docking. Molecular docking of seven different libraries provided us with several high ranked compounds. These compounds have

interactions with the key residues of the binding site and provides a suitable starting point for future experimental study.

Chapter 3. Modeling of the *Salmonella typhimurium* ArtAB toxin

3.1 Introduction

It has been reported that *Salmonella enterica* causes more than one million cases of infection each year in USA and is the major reason of hospital stay and mortality due to food related diseases[205]. As Salmonella infections play a major role in public health, the United States Department of Health and Human Services has aim to reduce the incidence of these infections by one-fourth by 2020.[206] In various nations, one of the major reasons of Salmonellosis in human beings and other living animals is *Salmonella enterica serotype Typhimurium* aka *Salmonella typhimurium*.[207] In the past many years, these bacterial infections have rose in various countries of the world.[208]

Recently *Salmonella serovar typhimurium definitive phage type (DT) 104*, which is resistant to many drugs, has been recorded in various countries.[205, 209-211] This particular strain of *Salmonella typhimurium* shows resistance to multiple antibiotics like chloramphenicol, ampicillin, streptomycin, tetracycline and sulfonamide.[212] These pathogens can have genetic factors that are transferable from one bacteria to another like plasmid DNA, prophages and genomic islands, and these factors can act as virulence elements.[212] New cases of *Salmonella typhimurium DT 104* has been rising in both human beings and animal populations, but an increased virulence-related phenotype for this microbe has not been noticed.[213] It has been proposed that a new virulence pathway might have developed in these organisms that has led to the incidence growth.[209]

3.1.1 *Salmonella typhimurium* DT 104 ArtA and ArtB

Numerous microorganisms contain code for specific ADP-ribosyltransferase toxins.[214] *S. typhimurium* DT 104 has been reported to have ADP-ribosyltransferase toxins known as ArtA and ArtB.[214] These ArtA and ArtB are the subunits A and B of the new toxin found in these organisms called AB toxin.[207] The AB5 toxin in these pathogens has been called ArtAB.[214] The AB5 toxin consists of two subunits, namely catalytic A and pentamer B, and both of them are linked non-covalently to each other.[215] This toxin works in two steps[215] : first, the B subunit of the toxin binds to the particular glycan receptors in the host body that provokes intake of more toxin in the host cells, which is then accompanied by the release of the A subunit that prevents the important cell functions in the host cells. ArtA of *S. typhimurium* DT 104 is homologous to subunit A of Pertussis toxin found in the species of *Salmonella typhi* and *Salmonella paratyphi.*[214] Similarly, ArtB is homologous to Subtilase cytotoxin subunit B secreted by *Escherichia coli*,[215] periplasmic protein of *Salmonella typhi* and *Salmonella paratyphi*,[207] subunit B of pertusiss toxin (ptx) of Salmonella typhi and Salmonella paratyphi[207]. This ArtAB position is situated on the prophage in *S. typhimurium* DT104 of the pathogen.[214] Although AB5 toxin acts as a virulence element in *S. typhimurium* DT 104, no phenotypic structures of ArtA and ArtB of *S. typhimurium* DT104 are available.

Our hypothesis for the study is that predicting the structures of the ArtA and ArtB subunits of *S. typhimurium* DT104 will provide us details about the binding mechanism of ArtA with ArtB that could be useful for extracting the crystal structure of ArtAB toxin. Our basis for this hypothesis is that most of the AB5 toxins produced by various

pathogens have homologous similarity with one or more subunits of the structures of the family of bacteria possessing this toxin.

3.1.2 Aim of the study

The aim of this study was to create the final homology model of the structure of AB5 exotoxin of *S. typhimurium* DT104. We believe that the successful completion of the project will provide us with a phenotypic structure of AB5 toxin that can be used as an assisting tool for crystallographic studies.

To fulfil our aims, first we created the homology models for both ArtA and ArtB, as there are no available structures for them. Next, we conducted the protein-protein docking to obtain a full structure of AB5 toxin. Finally, a molecular dynamics simulation was performed to see the time dependent behavior of the molecular system of the combined ArtAB structure.

3.2 Materials and Methodology

Kinds of software used for this study are listed below.

- Schrödinger Prime structure prediction tool for creating the homology models.
- PROCHECK, VERIFY 3D, QMEAN for evaluating the homology model.
- Bioluminate program of Schrödinger software for protein-protein docking program
- AMBER program for minimization and molecular dynamics simulation.

3.2.1 Structure prediction through Homology modelling

We obtained the FASTA amino acid sequences of both ArtA and ArtB of *Salmonella enterica serovar typhimurium Structure DT104* using the National Center of Biotechnology Information (NCBI) protein database. We utilized the FASTA amino acid sequences to get preliminary ideas about the possible templates for both the structures using the blast homology search. After that, we created homology models of ArtA and ArtB using Schrödinger suite 15-2. First, we created the homology model of ArtB (Tasks – biologics – homology modeling – advanced homology modeling). Since ArtB is a pentamer (B5 units), homo -5-mer was created for this structure. In the input sequence step, the FASTA amino acid sequences of ArtB were entered in the box. It has 141 amino acids. Next in the find homolog step, we searched for the potential templates of the query sequence using the blast homology search. NCBI PDB (all) option was chosen in the search option. We chose template 3DWP with sequence identity 35% (highest) looking at the percentage sequence identity, sequence coverage between the query-template, E-score and other properties. Mainly the total percentage sequence identity was considered.

As we wanted to create a pentamer (homomultimer), we selected 3DWP_A, 3DWP_B, 3DWP_C, 3DWP_D and 3DWP_E as multiple templates at the same time by clicking the shift click button. This helped us create a homo pentamer in a single step in the build homology step. Even though the sequences were identical for each template chosen, each template chain was chosen to create the homomultimer. If we are building a multimer of any sort, we should not align the template chains, because they must already be in the correct spatial relation and alignment with each other to construct the multimer from them. The template chains were arranged spatially in the workspace. We verified the

symmetry between the chains of the template with the pentamer 3DWP structure. The third step is the edit alignment step. The edit alignment process helps to enhance the alignment between the templates and the query. To generate the alignment, we selected Prime STA, which is an alignment method based on the secondary structure prediction. Prime STA is a good alignment method for average sequence identity (20-50%). In our case, the identity is 35 %.

We ran the secondary structure prediction by selecting the required option. SSpro is the secondary structure prediction program available in the Prime. After that we generated the alignment from the align option. Because a few gaps in the template residues were at the 66 to 69 positions and one residue at position 66 was at strand, we unlocked the gap at the strand manually using the edit menu option available in the menu bar.

After visual inspection of the alignment and secondary structure prediction, we moved forward to the build structure step. To build the pentamer (homomultimer), first the chains of the templates were verified for their accurate location (position and orientation with respect to each other). In the multi-template model, we chose the Homo-multimer option (as we are building the pentamer B5). All templates were aligned to the target sequence. We did not include any ligands, cofactors or water available in the templates. No any constraints were used in the procedure. In the build options, all default settings like retain the rotamers for conserved residues, optimize side chains and omit structural discontinuities for insertion in template gaps of more than 20 residues were chosen. The knowledge-based approach was preferred as the model building method. We created a single homomultimer as the default option. Each template chain was used to build the model structure for the homomultimer. The template chains were in accurate position and

orientation with respect to each other. An assembled homo-5-mer model was obtained, which had all the chains in correct location and symmetry as in the template chains.

Similarly, we predicted the homology model of ArtA following the same steps as above. For this structure, we created an "A" subunit, a monomeric unit that binds with the ArtB subunit. The FASTA amino acid sequence of *Salmonella typhimurium* ArtA contains 241 amino acids. The blast homology search in the find homologs step showed 4K6L_G, the structure of *Salmonella enterica serovar typhi* (*S. typhi*) with sequence identity 61% as the best template.

The template was viewed in the workspace to see the alignment in the sequence viewer. As we are building a monomer and the sequence identity of the template is high, we only selected single template. In the following edit alignment step, ClustalW was selected as the alignment method. ClustalW generates the alignment based on sequence information only. It can be used to create the alignment when the sequence identity between the query and template is higher (more than 50%). In our case, it was 61%. We ran the secondary structure prediction from the option and finally alignment was obtained. No major gaps and other structural problems were seen during the visual inspection of the alignment. Therefore, no any manual alignment was performed in the process. Only a single template was used in the build homology step. Other parameters selected were the same as the process followed during the build step of ArtB model.

After building the homology models of both ArtA and ArtB, the models were subjected to 2500 iterations of energy minimization from the Schrödinger software (Tasks – minimization – forcefield) to minimize the bad contacts, steric clashes and other structural problems. Next, we examined the obtained homology models. We calculated

the RMSD between the alpha carbons of the homology model and the chain/chains of the template used for building the corresponding model from matchalign menu of Chimera. PROCHECK[62] was used to evaluate the stererochemical properties using the Swiss model expasy server[196, 216]. VERIFY 3D[197-198] was used to assess the compatibility of the 3D structure of the model with the amino acid sequence. Verify 3D was performed using the structural analysis and verification server (SAVES)[199]. We used QMEAN[200] from the Swiss model expasy server[196] to evaluate the quality of the model taking into consideration the six different energy terms.

3.2.2 Protein-Protein docking of *Salmonella* ArtA and *Salmonella* ArtB

Both the homology models of ArtA and ArtB were once again prepared through the protein preparation wizard. After inspection in the workspace, these models were taken further for protein-protein docking (Application – bioluminate – protein-protein docking). We chose the standard mode for the docking purpose. The ArtB Pentamer (all the chains A, B, C, D, and E) was selected as the receptor. The smaller subunit, ArtA, was chosen as the ligand. The number of ligand rotations to probe was 70000. We selected the maximum number of poses as 30. Bioluminate protein-protein docking generated the poses from top to bottom based on ranking. The final model was selected mainly based on the visual inspection from the top models.

We obtained the complete ArtAB structures of 4L63[217] and 4L6T[217] from the Protein Data Bank and observed the binding of subunits ArtA and ArtB. This provided us some insights about appropriate binding of ArtA and ArtB in our AB5 complex predicted from the protein-protein docking. The pose obtained with the ArtA correctly docked in the

interface of ArtB Pentamer was considered as the best pose of all the other poses. We selected the second best pose based on visualization and accuracy of the pose.

After creating the complete protein structure of ArtAB, we used the structure for MD simulation purposes to study the stability of the AB5 exotoxin. Moreover, we studied the conformational changes and analyzed them as described below.

3.2.3 Molecular Dynamic studies of protein-protein docked *Salmonella* ArtAB

The AMBER program was used to study the molecular dynamics simulation of the ArtAB structure obtained from protein-protein docking. First, topology and coordinate files were prepared from the leap program available in the AMBER software. The AMBER ff14SB force field was used for protein residues. We solvated the system with explicit solvation. A 12-Angstrom buffer of TIP3P was used. Two Na+ ions were added to neutralize the system. Generated topology and coordinate files were saved for further use. The next step was minimization of the ArtAB system. Initial minimization of water was performed using 10000 steps of maximum cycle with protein fixed using restraint of 100 on the protein. After relaxing and removing the bad contacts present in water, we conducted minimization of the whole system. Since the ArtAB system was larger, the minimization of the entire system was performed in two steps:

a) Minimization of 3000 maximum cycles with decreased restraint of 50.0 on the protein (since larger systems have maximum chances of blowing up the structure, the system was slowly relaxed using lower restraint)

b) Second minimization of 5000 maximum cycles removing the restraint on protein and relaxing the entire system.

After completion of the minimization, we viewed the structure in VMD to observe the differences in structure before and after the minimization.

Equilibration of the ArtAB system

Two-stage equilibrations were performed for the ArtAB structure after the minimization steps. In the initial stage of equilibration, we increased the temperature of the system from 0 to 300K for 500000 steps with time steps of 2 femtosecond. A restraint of 20.0 was used on the protein to prevent the system from blowing out instantly. A periodic boundary condition with constant volume (ntb=1) was used in this stage. NTT=3 was chosen as the temperature coupling algorithm. Our aim for the first equilibration was to safely raise the temperature from 0 to 300K. After the equilibration, we checked the temperature of the system using the cpptraj program available in AMBER. We plotted the temperature data using grace software to evaluate temperature changes over the period.

In the second stage of equilibration, we equilibrated the system at constant pressure (ntb=2) and constant temperature periodic boundary conditions for 1 ns with 2 femtosecond time steps. No restraints were used in the second equilibration. We analyzed the second equilibration for properties like density, potential energy, kinetic energy and total energy of the system as a function of time using the cpptraj program. The obtained data were plotted using the grace software.

Production run

Evaluating the output files from equilibration and studying different properties of the system like temperature, density, potential energy and kinetic energy from the equilibration stages provided us information whether the equilibrium of the system has

been reached. After the equilibration was reached, we ran production simulation for 8 ns with 2-femtosecond time steps. Constant pressure (ntb=2) and constant temperature periodic boundary conditions were used. A cutoff of 10 Angstrom was used for periodic boundary conditions.

Once the production run was completed, we analyzed the results. First, we measured the RMSD of backbone atoms of each frame (total 4000 frames) taking first frame of the production run as the reference frame using the cpptraj program. Second, we studied the Root Mean Square Fluctuation (RMSF) of each amino acid residue over the trajectory. These atomic fluctuations were studied on the backbone Carbon (Ca) atoms. We plotted the data obtained from these analyses using grace data plotting software. After this, we studied the energies of the system (especially the potential energy) during the production simulation. The potential energy helped us to locate conformations with the lowest energies. The Process_mdout_perl command helped us to interpret the energy results from the output file of the MD simulation. We extracted the top three lowest energy conformations from the last one nanosecond of production simulation in PDB format using the potential energy file generated from the above process. Later we visualized these conformations in Chimera. Finally, the lowest energy conformation was selected.

3.3 Results

3.3.1 Modelled structure of *Salmonella* ArtAB

We obtained homology models of ArtA and ArtB from the Schrödinger. The protein preparation wizard was used to prepare the protein. Minimization of 2500 steps was conducted from Schrödinger. This helps to refine the structure and reduce the structural problems. We analyzed both the structures using various programs. We calculated the

RMSD between the backbone alpha carbon of the homology model and the chain/chains of the template. From the Chimera match align, each chain of the model was superimposed with the corresponding template chain. Then RMSD was calculated. The RMSD results between the individual template and the respective model chain are shown in Table 3.1 below.

Table 3.1. RMSD between generated models and chains of the templates

ArtB homology model	
Chains	**RMSD (Angstrom)**
Model_chain_A- Template_chain_A	0.76
Model_chain_B – Template_chain_B	0.79
Model_chain_C – Template_chain_C	0.75
Model_chain_D – Template_chain_D	0.89
Model_chain_E – Template_chain_E	0.76
ArtA homology model	
Model – Template_chain_G	0.32

Next, we studied the Ramachandran plots of both ArtA and ArtB models to evaluate the amino acid residues.

Figures 3.1 and 3.2 shows the Ramachandran plots of ArtA and ArtB generated models. The Ramachandran plot of ArtA exhibits 86 % of the residues in the most favored regions, slightly above 11% in the additionally favored regions and 1.8% in the generously allowed regions. No residues fall in the disallowed regions of the graph.

The Ramachandran plot of ArtB reveals slightly more than 78 % residues in the most favored regions, nearly 19 % in the additionally favored regions and around 2% in generously allowed regions. 0.6% of them fall in the disallowed regions.

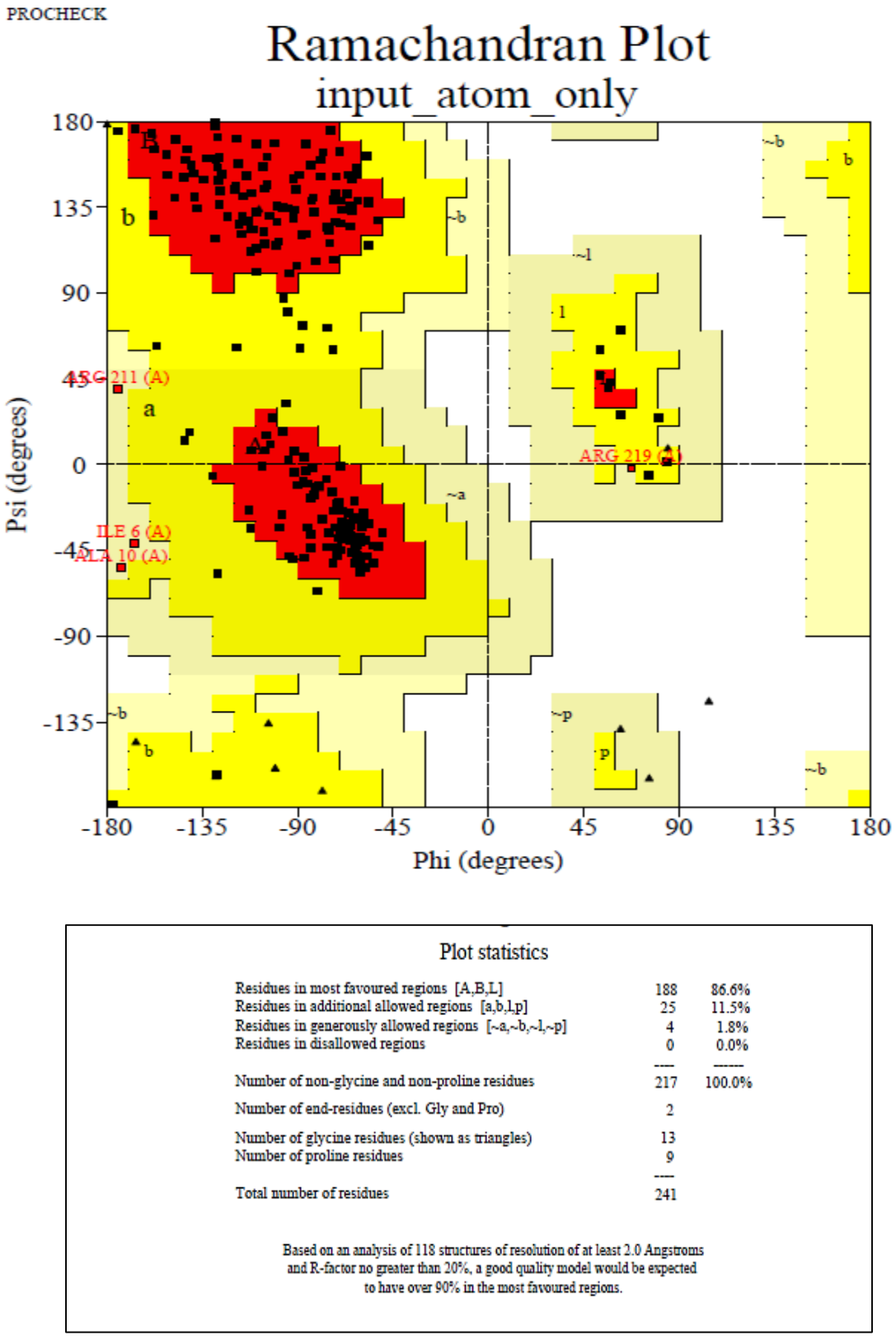


Figure 3.1. Ramachandran plot of the ArtA homology model

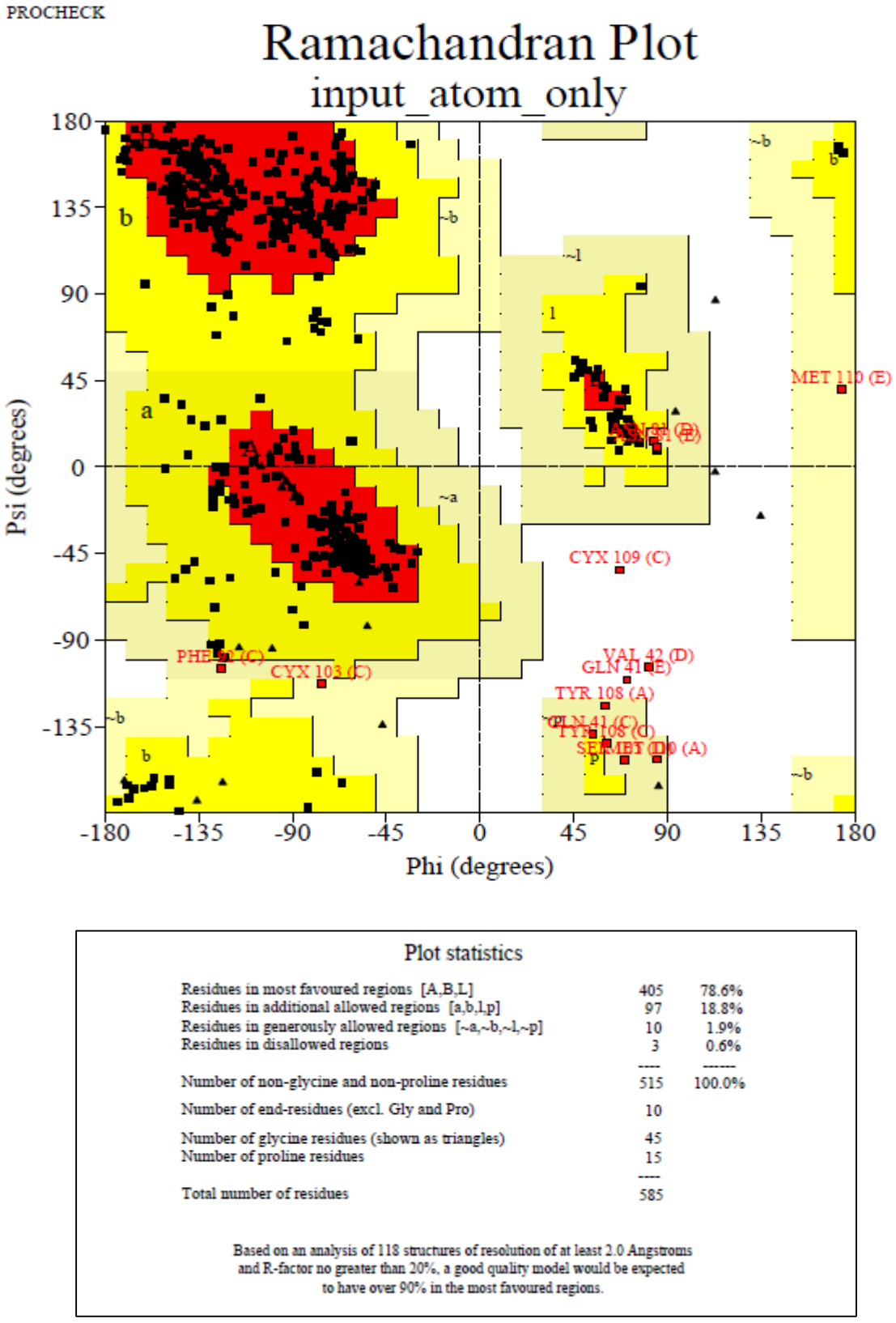


Figure 3.2. Ramachandran plot of the ArtB homology model

We studied several stererochemical properties of both ArtA and ArtB models applying PROCHECK analysis. The detailed plots of bond length, bond angle and major distances from planarity from the PROCHECK results of ArtA are shown in Figures C.1, C.2 and C.3 respectively (in Appendices). More than 99% of the residues of ArtA have main chain bond lengths within the specified limit. Likewise, main chain bond angles of 95% residues are within the limit and planar groups are 98% of the acceptable limit. Figure 3.3 below shows the PROCHECK summary of ArtA model.

```
 +----------<<<  P  R  O  C  H  E  C  K     S  U  M  M  A  R  Y  >>>----------+
  |                                                                            |
  | input_atom_only.pdb   2.5                                    241 residues  |
  |                                                                            |
 +| Ramachandran plot:   86.6% core   11.5% allow    1.8% gener    0.0% disall |
  |                                                                            |
 *| All Ramachandrans:   11 labelled residues (out of 239)                     |
 *| Chi1-chi2 plots:      5 labelled residues (out of 146)                     |
  | Main-chain params:    6 better     0 inside      0 worse                   |
  | Side-chain params:    5 better     0 inside      0 worse                   |
  |                                                                            |
 *| Residue properties: Max.deviation:     9.2             Bad contacts:    1  |
 *|                     Bond len/angle:   10.0    Morris et al class:  1  2  2  |
 +|     1 cis-peptides                                                         |
  | G-factors           Dihedrals:  -0.31  Covalent:   0.12    Overall:  -0.13  |
  |                                                                            |
  | M/c bond lengths: 99.5% within limits   0.5% highlighted                   |
 *| M/c bond angles:  95.2% within limits   4.8% highlighted       9 off graph |
 +| Planar groups:    98.1% within limits   1.9% highlighted       1 off graph |
  |                                                                            |
  +----------------------------------------------------------------------------+
```

Figure 3.3. PROCHECK summary of residues of the ArtA homology model

Similarly, PROCHECK analysis was performed for ArtB homology model. The detailed results of PROCHECK of amino acids about the bond length, bond angle and the main distance from planarity are depicted in Figures C.4, C.5 and C.6 respectively (in Appendices). The main properties from the PROCHECK results are summarized in Figure 3.4 below.

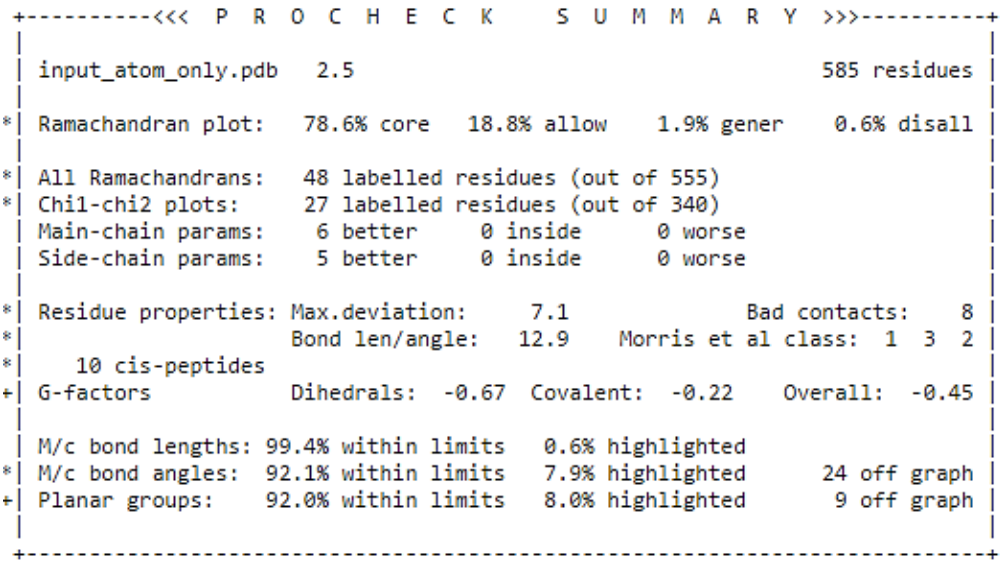


Figure 3.4. PROCHECK summary of residues of the ArtB homology model

More than 99% of the residues have main chain bond length within the favorable limit. Furthermore, 92% of the residues possess main chain bond lengths within the specified limit and planar groups are 92% within the acceptable limit.

Verify3D results of the models.

Verify3D program was utilized to predict the compatibility between the three dimensional structure of the model and the amino acid sequence. 84.65 % of the residues have a mean 3D -1D verify 3D score of more than 0.2 for the ArtA homology model whereas for the ArtB model 100.00% residues have such a score.

QMEAN6 score

The QMEAN6 score was calculated using the Swiss model expasy server. We obtained Scores of 0.471 for the ArtA model and 0.31 for the ArtB model.

3.3.2 Protein-Protein docked structure

We used the Bioluminate program of the Schrödinger package to dock the ArtA and ArtB structures. The protein-protein docked structures were ranked from top to bottom according to the poses generated. There were altogether 30 ArtAB docked poses obtained from the results. We selected the second pose based on ranking and visualization. The second pose had ArtA correctly docked in the interface of ArtB pentamer. The 3D structure of the selected ArtAB docked pose with the interface is shown in Figure 3.5 below.

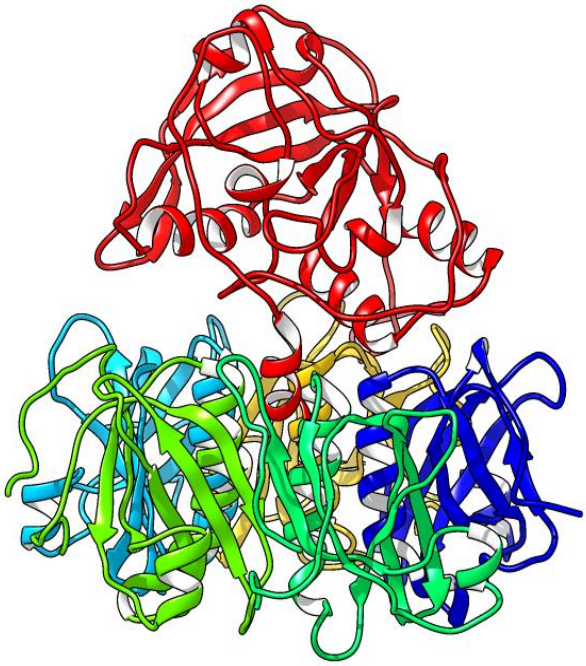


Figure 3.5. Protein-protein docked complex of ArtAB with interface

ArtA in the top of the structure is shown in red. Pentamer chains of ArtB are depicted by five different colors. Chain A: blue, Chain B: cyan, Chain C: spring green, Chain D: lime green and Chain E: yellow.

3.3.3 Molecular Dynamics Studies and analysis

The selected ArtAB model was subjected to molecular dynamics simulations. First, we carried out minimization of the system. The unrestrained 5000 iterations of final minimization of the whole system decreased the energy of the system from -4.0654E+05 to -4.1056E+05. After minimization of the whole system reduced the bad contacts, the system was heated in the first equilibration stage from 0 to 300 K using small restraint on the protein. The ArtAB system is larger and it may blow out when increasing the temperature. The restraint was used to prevent rapid fluctuation of our structure and prevent it from falling apart. The output file was viewed for temperature changes. The fluctuation in temperature during two nanoseconds of the equilibration stages is shown in Figure 3.6 below.

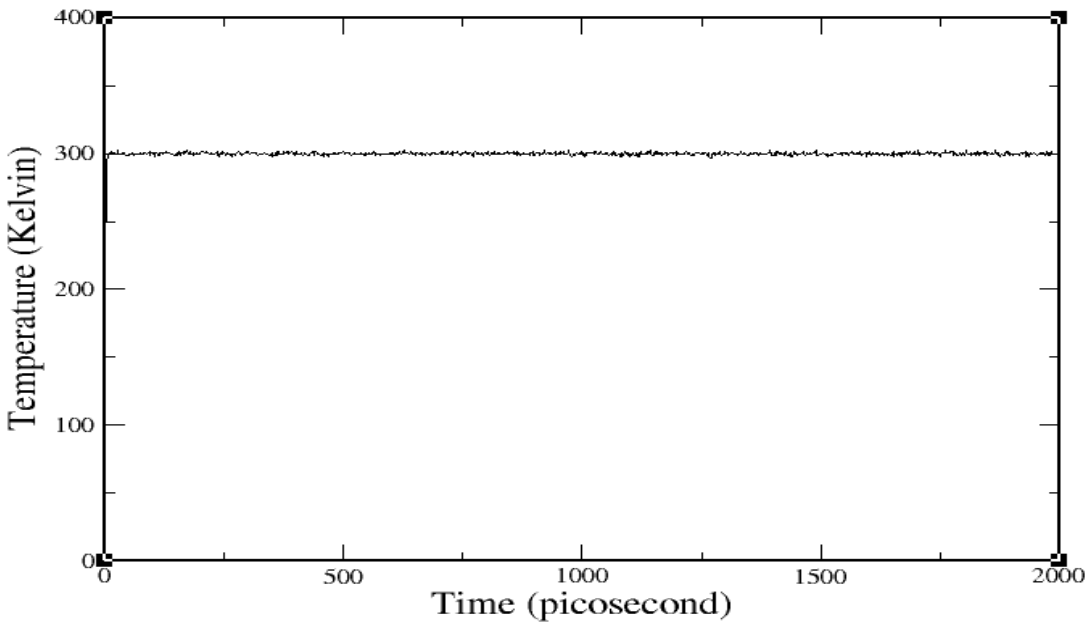


Figure 3.6. Temperature of the system over one nanosecond of equilibration at constant volume.

From the Figure 3.6, we can see that the temperature of the ArtAB system remained stable throughout the time of simulation after initial 5 picoseconds. After the temperature of the system was stabilized in the first equilibration stage, the second equilibration was conducted in the constant pressure condition to relax the density of the water and stabilize it before running the production. Restraint was removed from the system. The density of the ArtAB system during the second equilibration stage at constant pressure is depicted in Figure 3.7 below.

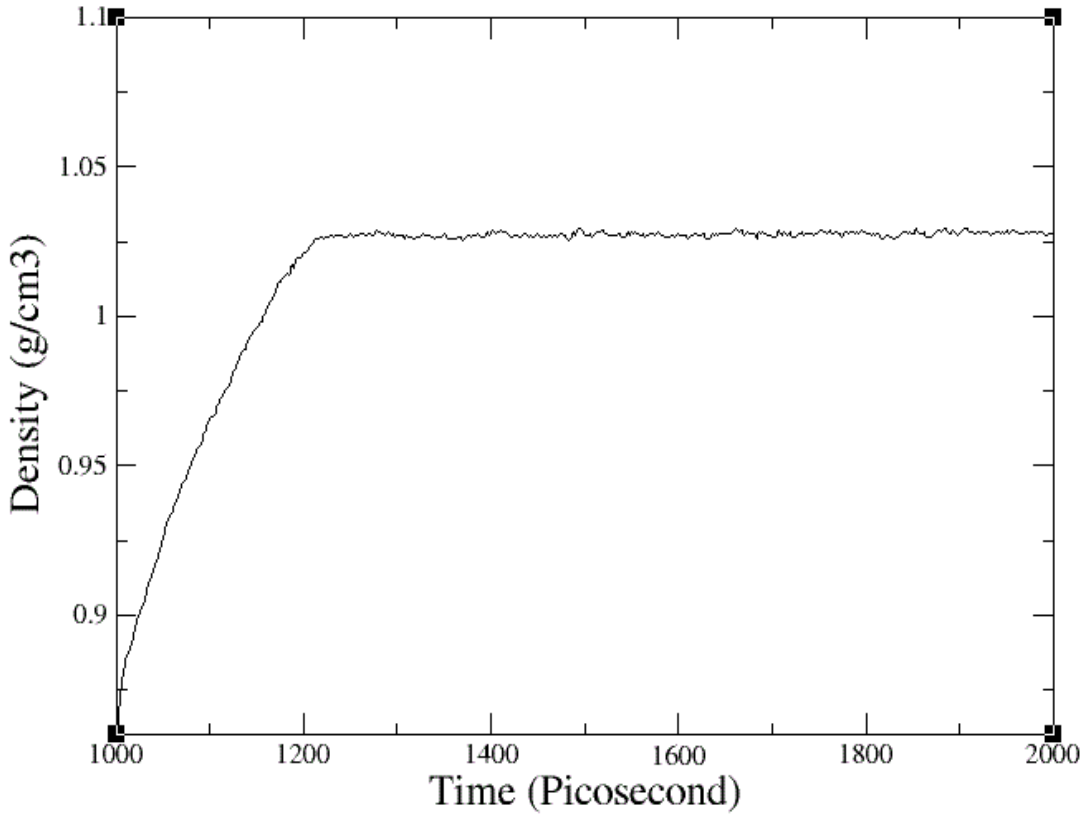


Figure 3.7. Density of the system during one nanosecond constant pressure equilibration

The density of the system has equilibrated at around 1.025 g/cm$^{3.}$ This seems to be reasonable and as per expectation. We did not calculate density in the first one

nanosecond equilibration at constant volume (as volume remains unchanged). Hence, density data represents only the constant pressure equilibration stage. We also studied the volume of the system during the simulation. The second equilibration at constant pressure provided information about volume. The volume of the system remained stable after the first 200 picoseconds of simulation at constant pressure as shown in Figure C.7 (in Appendices). For first 200 picoseconds, it decreased rapidly and after that, it remained almost constant throughout the rest of the simulation. Later, different energy terms of the system were studied for the equilibration both at constant volume and at constant pressure periodic boundary conditions. This enlightened us about the changes in energy over time for both the simulations. Figure 3.8 shows the kinetic, potential and total energy of the system during both the equilibration stages.

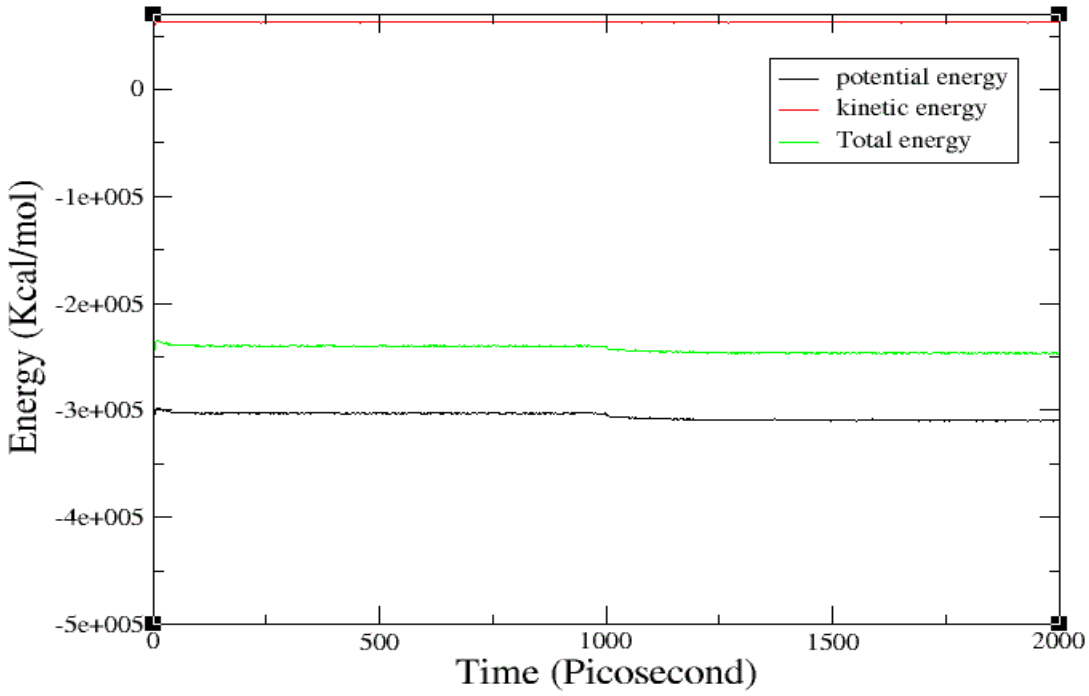


Figure 3.8. Potential, Kinetic and Total energy of the system during equilibration.

The black line represents potential energy, the red line represents kinetic energy and the green line represents total energy of the system. Kinetic energy (red line) after increasing for a few picoseconds remained steady throughout the simulation. Kinetic energy was positive (over 20000 Kcal/mole). Potential and total energy (kinetic + potential) of the system increased for a few picoseconds in the beginning and remained constant for rest of 1 ns simulation at constant volume. This was followed by slightly decreases in both the energy terms for a few picoseconds and again remained steady for rest of the 1 ns simulation at constant pressure.

We also evaluated the pressure of the system to confirm whether it had equilibrated or not, shown in Figure C.8 (in Appendices). The pressure of the system remained zero during the constant volume equilibration stage. During the constant pressure equilibration run, the pressure was initially negative for a few picoseconds. Later the pressure of the system became positive. Broadly, the pressure of the system fluctuated swiftly throughout the simulations whereas the average pressure remained stabilized around -3.7 g/cm^3.

Analysis of the equilibration stages provided us information about various properties of the system. Equilibration stages were accompanied by production run. Due to our availability of resource and time constraint, we ran 8ns of production run using the same parameters used for the constant pressure equilibration run. From the analysis of equilibration stages, the system was nearly equilibrated in terms of temperature, density, pressure and energy terms. We analyzed the structure from the production run. To see whether the structure remained sensible or not, we analyzed the trajectory of the production simulation by calculating the RMSD of the backbone atoms of the trajectory

as the function of time. The RMSD was calculated between each successive structure and the first structure of our production run (last structure of the second equilibration).The RMSD measures how identical the internal coordinates of the given structure is to the reference structure. We measured the RMSD using the cpptraj program and plotted the data using the grace software as represented in Figure 3.9 below.

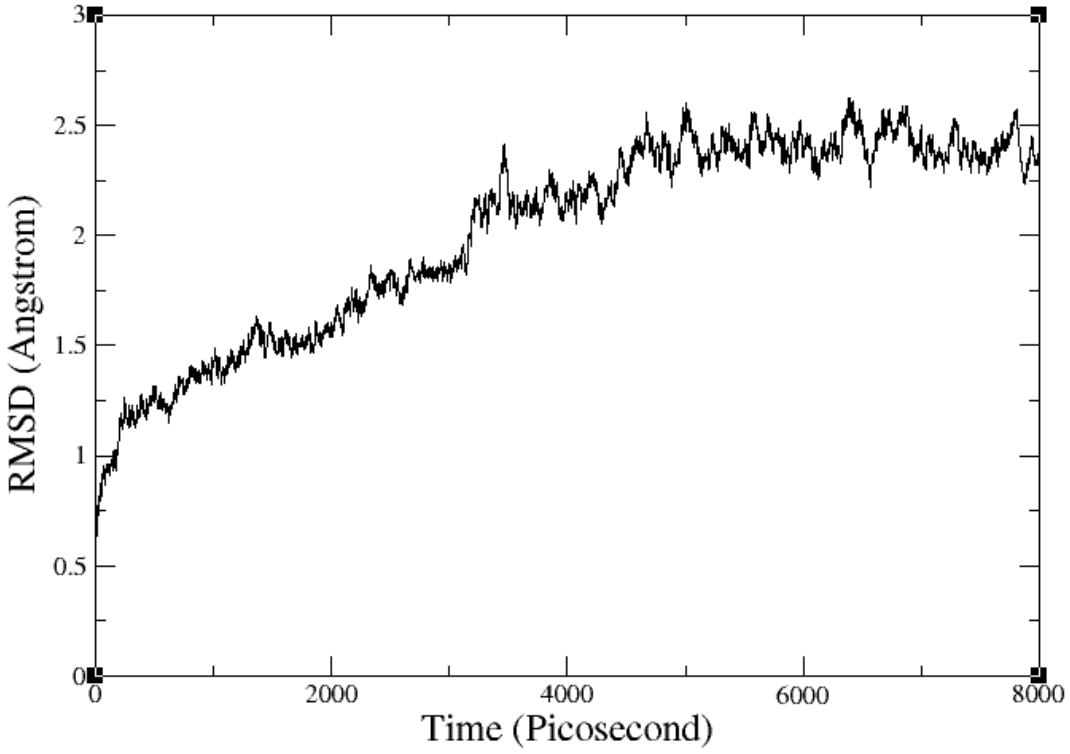


Figure 3.9. Backbone RMSD vs. Time of the production run trajectory

In the production run, there were total of 4000 frames (4000000 steps written at every 1000 steps, ntwx =1000). From the figure, we can see that initially the RMSD of the system increased rapidly until 4 ns and then it remained stable throughout rest of the simulation. Similarly, we also studied the flexibility of each residue from the trajectory. We calculated the atomic position fluctuations of each atom in terms of backbone alpha

carbon. Figure 3.10 represents the Root Mean Square Fluctuation (RMSF) per residue from the production simulation.

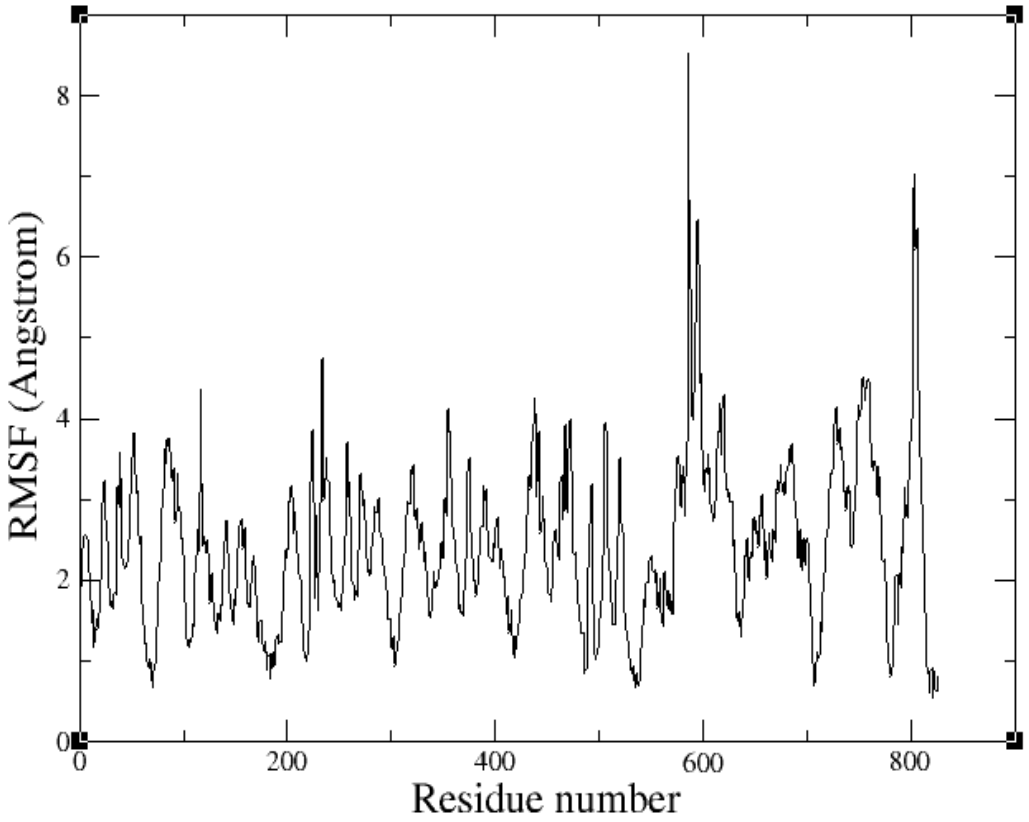


Figure 3.10. RMSF of each residue in terms of backbone carbon atoms during production simulation

For 826 residues, the majority of amino acids showed atomic fluctuations less than 4 Angstroms throughout the simulation. Some of the residues had high atomic fluctuations as shown in the figure above. When the trajectory was viewed in VMD, larger movements were obtained in amino acids whose residue numbers ranged from 582 to 598 of the ArtB structure and from 801 to 807 of the ArtA structure.

We inspected the energy plots of the production run. The potential energy of the system for 8 ns of production run was evaluated to take a few snapshots of the low energy states.

The potential energy of the system remained stable throughout the simulation as shown in Figure C.9 (in Appendices). We can identify the lowest energy states by viewing the summary.EPTOT file generated by the process_mdout.perl just as in the equilibration stages. We took three snapshots of the top three low energy states of the last 1 nanosecond of the production run. Frame 3800 at 9.60 ns had the lowest energy with potential Energy =-310796 Kcal/mole. The second and third low energy states were at 9.67 ns and 9.39 ns respectively. They represent frames 3839 (Energy = -310640 Kcal/mol) and 3695(Energy = -310560 Kcal/mol) respectively. Later we extracted the top three frames in PDB format using the cpptraj program. We viewed the structures in Chimera to see whether they looked reasonable or not. The final structure chosen was the one with the lowest energy. The three dimensional visualization of the lowest energy conformation of the ArtAB structure is represented in Figure 3.11 below.

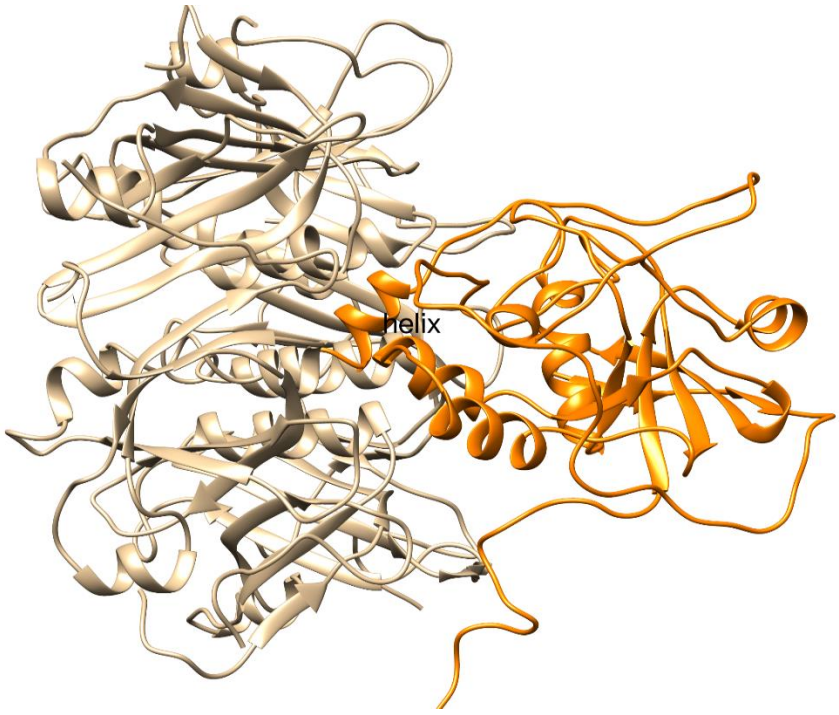


Figure 3.11. Lowest energy snapshot of ArtAB with interface of ArtA and ArtB

ArtA is represented in orange. ArtB is shown in tan. The figure shows that a part of the alpha helix of ArtA enters the junction/interface of the ArtB pentamer.

3.4 Discussion

Our aims for this project were a) to create ArtA and ArtB structures of *Salmonella typhimurium DT 104* using the homology modeling, b) to develop a protein-protein docked structure of complete ArtAB and c) to perform molecular dynamics simulations of the docked structure to generate the lowest energy conformer of the structure for future studies.

3.4.1 Homology models of ArtA and ArtB

Homology models of ArtA and ArtB were constructed from the templates 3DWP (all five chains) and chain A of 4K6L respectively. Both of these templates are from the family of AB5 toxin secreted by various bacteria and are related evolutionarily. Both final models of ArtA and ArtB compared with their template structures have backbone carbon (Ca) RMSD values less than 1 Angstrom. We also studied the Ramachandran plots of both the ArtA and ArtB structures. ArtA has no residues in the disallowed region and ArtB has nearly 0.6% residues (Cysteine 106, Glutamine 41and Valine 42) in this region. The majority of the residues in ArtA and ArtB have bond length, bond angle and planar groups within the specified limit (more than 90 % residues). Both the structures of ArtA and ArtB have good compatibility between the 3D structure and their own amino acid sequences as shown by verfiy3d scores. The only problem is with the QMEAN6 score. Both ArtA and ArtB structures have scores below 0.5 and ArtB, especially, has lower score. One of the reasons for having low QMEAN scores may be the use of low-resolution templates as crystal structures for modeling of both the homology models.

3.4.2 Protein-protein docked structure of ArtAB

After creating the homology models of both ArtA and ArtB, we conducted the protein-protein docking to create the complete structure of ArtAB. 14 of 30 (36.6%) protein-protein docked poses generated from the Bioluminate program of the Schrödinger were able to predict the correct ArtAB structure. "A" subunit is a toxic component possessing an active site and "B" component of ArtAB binds with the target in the host cells and translocates ArtA inside the host cell.[207] The catalytic subunit A binds noncovalently with the pentavalent subunit B in the interface.[215] So we chose the top ranked compound (rank 2) that had ArtA correctly bound in the interface of ArtB. The first pose did not have the correctly docked ArtAB.

3.4.3 Molecular dynamics simulations of ArtAB structure

The initial minimization of the system resulted in decreasing the energy of the system. Our goal in this stage was to remove bad contacts and prepare the ArtAB structure for further MD simulation. The first equilibration of 1ns at constant volume resulted in maintaining the temperature at around 300 K after a few picoseconds. This was what we expected. This also indicated that the Langevin dynamics for maintaining the temperature (NTT=3) that we used in our simulation worked well. Langevin temperature for equilibration (NTT=3) is considered better for maintaining and equilibrating the temperature.[218] Temperature remained around 300K throughout the simulation. We used limited restraint force on the protein in this stage to help the system from falling apart during the heating process. The second equilibration was at constant pressure for one nanosecond to equilibrate density and other energy terms. Density remained at around 1.025 g/cm^3 as shown in Figure 3.7. Density of water is 1 g/cm$^{3.}$ Adding protein and

some positive charged Na ions in the solvent might have slightly increased the density of the system.[218] While analyzing the potential, kinetic and total energy of the system, we found these energy terms equilibrated, as we wanted for our system as illustrated in Figure 3.8. Due to increase in temperature from 0 to 300k, the kinetic energy increased for a few picoseconds as expected and then remained stable throughout the simulation. After we removed the restraint and equilibrated the system at constant pressure, the potential and total energy of the system remained stable.

We ran 8ns of production simulation to study the time dependent behavior of our ArtAB system. Comparing the RMSDs between the total frames of the trajectory to that of the first reference structure, it remained below 2.5 Angstroms during the production run. Interestingly, the RMSD remained stable after 4ns of production simulation. This stability indicates that there is no huge conformational changes taking place in the structure. Although larger time scale molecular dynamics simulations of microseconds are required to study accurate conformational changes in the structure, small and stable RMSD during the simulation of ArtAB provided us the acceptable standard for studying the properties of the system.

The atomic fluctuation of backbone alpha carbon during the simulation as indicated by RMSF shows that atomic fluctuations in these residues differ. Most of the residues have RMSF from 2 to 4 Angstroms. A few residues have greater fluctuations as indicated by the residue number mentioned in the result and seen in Figure 3.10. While visualizing in VMD, these residues were mainly present in the loop regions of the ArtB structure (chain C) and some loop regions of the ArtA monomer structure.

We also visualized the interface of ArtA and ArtB where the alpha helix part of ArtA interacts at the junction of the pentamer of ArtB and found that these regions do not possess large conformational changes. Finally, we extracted the lowest energy conformer of the ArtAB based on potential energy. We proposed this structure to aid in the study of crystal structure of ArtAB.

3.5 Conclusion

In this project, we proposed a model of the complete ArtAB structure of *Salmonella enterica serotype typhimurium DT104.* We created ArtA and ArtB homology models using templates 4K6L and 3DWP respectively using the homology model program from the Schrödinger software. ArtA contains monomer and ArtB contains pentamer structure. We analyzed these structures using RMSD, Ramachandran plot, PROCHECK, Verfiy3D and QMEAN scores. We noted that these programs generated acceptable and better results for both structures except the QMEAN scores. QMEAN score values for both structures were lower than 0.5. We speculate that this may have happened because of the low-resolution x-ray crystal structure templates used for generating these homology models.

There were 30 protein-protein docked poses of ArtAB generated from Schrödinger's Bioluminate program. We visualized the docked poses in the workspace and found that approximately half of the poses were docked appropriately. We selected the top ranked accurately docked pose.

After 5000 steps of the whole system minimization, we conducted two steps of equilibration. We observed that the system attained equilibrium in terms of temperature, density, volume, pressure, kinetic, potential and total energy in these steps. Though

pressure fluctuated during the constant pressure equilibration stage, it remained at an average of -3.7 g/cm^3 throughout the process. Additionally we performed 8ns of production run to study the behavior of the system. The RMSD analysis of the backbone atoms was performed on 4000 frames of the trajectory, and comparing the first structure of production run showed that the RMSD remained within 2.5 Angstroms. After half of the production simulation time, it remained stable. However, flexibility of loop residues of both the ArtB and ArtA structures was shown by RMSF analysis. These loop regions did not fall on the interface of ArtA and ArtB.

Finally, we studied the potential energy of production run. It remained stable during whole 8ns time. We extracted the lowest energy conformer from the last one nanosecond time. Frame 3800 at time period 9.60 ns had the lowest energy with minimum potential energy (Emin) of -310796 Kcal/mole. The resulting structure of ArtAB can act as a starting point for further three-dimensional study of crystal structure.

Chapter 4. Discussion and Conclusion

4.1 General review of the thesis

The thesis work includes two different projects. The first was predicting the structure of the *C. difficile* FabK enzyme using homology modeling and finding potential inhibitors that bind in the active site of the enzyme. The second was predicting the structure of Salmonella ArtAB, and obtaining the lowest energy conformer of the structure from molecular dynamics simulations. This can be used as a starting structure for crystallography study of ArtAB. With these two projects, I was able to study and utilize several computational techniques.

The Chapter 2 project contained the study of FabK enzyme (an important enzyme in fatty acid synthesis) of *C. difficile.* Utilizing homology-modeling techniques, I created the final model structure that was analyzed using various software programs. Similarly, seven different libraries were prepared from the ligand preparation program using various filtering criteria included in the filtering file. Additionally these compounds were docked in the active site of the receptor. Some of the top compounds showed key interactions with the residues of the receptor.

The Chapter 3 project included discussions about the *Salmoenalla typhimurium* ArtAB. Using the homology-modeling approach, I created models for the ArtA monomer and the ArtB pentamer. These structures were analyzed using several programs in a similar way to that of Chapter 2. Moreover, the protein-protein docking tool was applied to create the complete ArtAB structure and the top ranked reasonable structure was selected. Molecular dynamics simulations were utilized to study the time dependent behavior of the system. Furthermore, the thermodynamic properties of the system were analyzed

along with the atomic fluctuations of the residues. Final selection of the lowest energy conformer was conducted from the potential energy data of the system.

4.2 Discussion of computational tools

The projects in this thesis employed several computational tools. These tools are desribed in the introduction (Chapter 1) of the thesis. Computational tools/methods have become an integral part of the drug discovery and development process. They reduce the time required for these processes and save a lot of money. A good computational approach requires knowledge of structural biology, molecular biology, pharmacology and various other disciplines. However, these tools have limitations. In this segment, I discuss the benefits and limitations of the computational tools used in these thesis projects.

4.2.1 Homology modeling

Homology modeling is a useful computational tool that helps to predict the 3D structure of a protein when there is no available experimental structure of the protein. If the sequence/sequences of the query structure is known, it can be used to identify the 3D structure by alignment with available template structures. However, homology modeling has some limitations because it is a theoretical model and visual inspection is required

The first consideration is that homology model only provides acceptable results if the sequence identity between the template and query is more than 30%.[34] Sequence identity below this level may not provide reliable models. Understanding this limitation is very important. In the case of these thesis projects, all homology models have more than 40% sequence identity except ArtB in one project, which has a low sequence identity of 35%.

The second consideration is a good sequence alignment between the template and query. This is required for producing accurate models for the target sequence.[34] Even though there are several alignment tools available that can generate alignment between the templates and query as discussed in chapter 1.2, there are still chances of errors in this process. The multiple chains single homology model, as in one of the projects, needs proper alignment and spatial arrangement of the chains before model building. It is always necessary to visualize the alignment before being sure about it. Sometimes manual alignment might be required in the process.

The third consideration is prediction of loop regions in the model. Loops have a reputation of being highly flexible where insertions and deletions occur frequently.[22] These regions are tough to predict with accuracy in comparison to other secondary structures like the alpha helix and beta strand.[22] Although several loop prediction methods are available (as discussed in chapter 1.2), predicting loop regions are still considered a challenge.

The homology modeling methods used in these thesis projects helped us to create acceptable models for various purposes of the studies. There are still requirements for advance methods for minimizing the errors in predicting the models particularly at the active site and if there is a low sequence identity between the query and template.[34] New method such as ligand-steered homology model tools have evolved where data from a familiar ligand are used to enhance configuration and optimize the binding site.[219-223] This technique is especially helpful in minimizing the variability in modeling the active site region.[34] Despite having some limitations, homology modeling is an essential approach in structure based drug discovery.

4.2.2 Protein-protein docking

To study the interaction of two protein partners, a technique such as protein-protein docking is an important tool. This technique has various applications in drug discovery. Although it is emerging as a fast and cost-effective tool to predict the complete structures of proteins, it has some limitations. A careful understanding of these limitations is required to use this tool appropriately in drug discovery process.

A benefit of using the protein-protein docking is that it is an easy and fast technique to obtain a docked structure. Although development of different search algorithms or scoring functions (as discussed in chapter 1.3) has facilitated the process, the protein-protein docking tool does not indicate the correct or incorrect docked poses. Predicting perfect binding condition is a tough job considering the larger nature of the protein structure.[68] A wise decision is required to predict whether the obtained docked pose is desirable or not. The poses generated from various software give different configurations. A little knowledge of the binding interface (how it binds), as in the ArtAB project, is helpful to differentiate between the accurate and inaccurate poses. Moreover, visualization of the docked poses is necessary.

Another important consideration is docking of model structures, as done in the studies of ArtAB docking. The trouble is that protein-protein docking gets complicated if the structures are homology models instead of crystal structures.[67] Thus, proper visualization of the docked pose and refinement of structures using minimization and MD simulations might help to minimize the errors/bad contacts that can occur in these structures.

A more comprehensive searching algorithm and scoring functions are required in the future to overcome the challenges in protein-protein docking. Though there are some

limitations and challenges in the process, a wise use of the protein-protein docking tool helps to understand the interaction mechanism of two protein partners.

4.2.3 Molecular docking

Another important computational method is molecular docking. Molecular docking helps in understanding the binding mode of the ligand in the receptor. In addition, it has become an essential part of virtual screening. Though it is a time saving and cost-effective way of obtaining virtual hits against a target, it has some limitations. A good understanding of the docking process and its weaknesses is necessary.

The first consideration in molecular docking is preparation of the receptor and ligands for docking. The receptor needs proper preparation with the addition of hydrogen, proper charges, bond orders and the receptor grid box for docking purpose. Similarly, preparation of ligands in terms of protonation states, partial charges, bond orders and tautomers is necessary for docking process. One of the common errors is lack of assigning partial charges to the ligand. A careful visualization of the receptor and prepared ligands is required.

The second consideration is validation. Scoring functions act differently to various types of receptors according to the nature of polar or lipophilic sites present in the binding site.[224] To overcome this limitation, there is a need to evaluate whether the sampling algorithm and scoring function present in the docking program are favorable for the binding site/target. A validation of the docking process with the original ligand of the receptor helps to provide parameters for future library screening.

Another important limitation in the docking process is scoring function. Due to the development of several search algorithms, conformational search space for docking pose has advanced properly but scoring functions are still not satisfactory. Binding affinity of the ligand in the pose is determined by the scoring function.[1] Nevertheless, contemporary algorithms for scoring (as discussed in chapter 1.5) available in several docking programs do not correctly predict the interaction energy of ligand-target composite with sufficient accuracy.[6] These scoring functions do not accurately predict the binding energy of the docked ligand. The algorithm used in scoring functions still faces some problems of desolvation and entropic impact.[225] A good visual interpretation of docking score results is necessary to identify hit compounds.

Despite some limitations of molecular docking indicated in the above sections, it is an essential tool in screening millions of compounds in a cost-effective way. A successful docking application also depends on the user and familiarity with the docking programs.

4.2.4 Molecular dynamics simulation

Molecular dynamics simulations are a convenient method in studying the molecular motion of atoms and residues in biological system. They can be used to study the time-dependent behavior of both crystal and model structures. Time scale motion of loops, alpha helix, beta strand and ligand-receptor binding can be studied utilizing MD simulations. However, MD simulations have some limits.

The first consideration is that MD simulations are time-consuming and computationally expensive. In larger systems just like the one in this thesis's second project, to study the conformational behavior of the protein structure requires a simulation of at least a microsecond. To correctly and precisely model nanosecond level protein motions, it

require a much longer simulation of a few microseconds.[226] Despite development of GPUs and increase in computational speed recently, this scale of simulation requires several weeks or months to complete the process. Increased computational requirements limit simulations of more than a few microseconds in length.[89]

The second important consideration is the force field used in molecular dynamics simulations. The molecular mechanics force fields used are approximations of quantum mechanics.[89] Molecular mechanics simulations can predict the motion of several biological molecules but quantum mechanics are required in some cases like transition metal coordination or electronic distribution.[89] Another important limitation of mechanics force fields is that they cannot be utilized for understanding the chemical reactivity as they do not take into account the bond breaking and making phenomenon.[90] These problems have been addressed by incorporating quantum mechanics force into molecular mechanics (QM/MM) and has been used successfully in many systems.[89] However, in QM/MM, only a part of the system can be treated quantum mechanically and the majority of it is treated using a classical force field. This technique is still computationally intensive, time consuming and difficult to use in larger systems.

Another important point is preparation of the system for simulations. Force fields are used to prepare the coordinate and parameter files for the residues of the system. Currently, several mechanics force fields are available that are used to prepare parameters for protein, nucleic acids, carbohydrates, lipids and some other biomolecules (as discussed in chapter 1.4). However, for non-standard residues such as ligand and cofactors, other force fields (as discussed in chapter 1.4) that are validated should be used

for preparing the parameters for these residues. The theme of this section is that proper evaluation of residues of the system is needed to conduct a successful simulation.

Lastly, although a few nanoseconds simulation was used in the second project of the thesis to obtain the lowest energy conformer of the structure, a few milliseconds simulation is required to accurately study the behavior of large protein motion and protein folding. Various enhanced sampling algorithms such as umbrella sampling[227], replica exchange[228], metadynamics[229], steered molecular dynamics[230-231], accelerated molecular dynamics[232], milestoning[233], transition-path sampling[234] and free energy perturbations(FEP)[235-236] have evolved that have increased sampling of the conformational space with accuracy. Although limitations exist for MD simulation, it is an important tool in drug discovery.

4.3 Future perspective

We studied two projects, namely *C. difficile* FabK and ArtAB, in this thesis work. We have presented some work in *C. difficile* FabK with our available resources and time. There are some avenues for future exploration. Some of the probable future paths that require exploration in *C. difficile* FabK and ArtAB projects are summarized below.

4.3.1 *C. difficile* FabK project

Future work on this project includes application of other docking programs for molecular docking. In this project, we have used glide docking from the Schrödinger to dock millions of compounds. It is always better to validate docking results from one or more docking algorithms. Although it might be time consuming to dock large number of compounds using other programs, it will be exciting to compare the results

Another future direction is experimental screening of virtually docked compounds. Ordering of some of these compounds from vendors is underway. Results from experimental assays of top hit compounds will provide crucial information about potency of these compounds in inhibiting the enzyme. Furthermore, they may also provide crucial information for the lead optimization process and help to refine further virtual screening.

4.3.2 *Salmonella* ArtAB project

We have obtained the ArtAB structure of *S typhimurium* from homology modeling and MD simulations. Future work in this project includes using a good resolution and higher sequence identity template especially for the ArtB structure. The template obtained for ArtB during the time of the project was crystal structure of subunit B of AB5 toxin of *Escherichia coli* (PDB:3DWP) possessing sequence identity of approximately 35% and resolution 2.2 Angstroms. Our aim in the project was to obtain lowest energy conformer structure of ArtAB that could aid in the analysis of crystal structure. We considered ArtB structure acceptable in this purpose. Detailed study including protein folding, loop motion, binding of ArtA and ArtB definitely requires a template that possesses higher sequence identity and a good alignment with the query sequence. Recently, a higher sequence identity template from same *Escherichia coli* (PDB: 4Z9C) has been noticed in the Protein Data Bank. I would suggest this as a good template for ArtB modeling. Comparison of both the results would provide differences between the two structures. This could improve the molecular modeling of the ArtAB structure.

Another future perspective for ArtAB is performing longer time scale MD simulation. We performed 8 ns production simulation in our project. Despite some loop motions in both the structures, no larger conformational changes took place as suggested by

thermodynamics properties, RMSD and RMSF analysis. As discussed earlier in the chapter 4.2.3, at least a few microseconds simulation is required for analyzing the system accurately. I would suggest a multiple-step microsecond simulations in the future. This can be utilized to get the lowest energy conformer of the structure and might represent the structure more accurately than the present one.

4.4 Conclusion

Homology models of both *C. difficile* FabK enzyme and Salmonella ArtAB were created and evaluated for their structural properties using multiple programs. After predicting the model, millions of compounds prepared from ligand preparation using filtering criteria were docked in the *C. difficile* FabK receptor. From this particular study, conclusions were drawn about the top compounds from different libraries and the key interactions shown by these compounds with the residues of the binding site. Similarly, ArtA and ArtB were used for protein-protein docking purposes. After the entire ArtAB structure was developed, it was modelled using molecular dynamics simulations and evaluated for their thermodynamics properties, atomic fluctuation and lowest energy conformation structure. From molecular dynamics, it was identified that the thermodynamic properties remained stable during the simulation. Flexibility was seen in the loop regions of the structure. The lowest energy conformer had the minimum potential energy and has the probability of assisting in the study of crystal structure.

I conclude this thesis by trying to give a satisfactory answer to the question: Are computational methods appropriate for drug design and biomolecular study?

We have used various computational methods in both of our projects. These include the homology model, protein-protein docking, molecular dynamics simulation and molecular

docking. The mentioned methods have some limitations as discussed in chapter 4.2. There might be several questions about the results of computational methods. However, computational methods have been used successfully in drug design and biomolecular study. Even in our docking study, we obtained some top compounds that had important key interactions with the residues, and had previously been believed to play a role in catalytic process of the FabK enzyme. Computational methods applied with caution and proper validation aid in the drug design as well as biomolecular study.

Finally, I trust that computational methods have a good future in the field of science. With the development of newer algorithms and techniques, several limitations as mentioned in chapter 4.2 will be solved. The computational field has a bright future.

List of References

1. Leelananda, S. P.; Lindert, S., Computational methods in drug discovery. *Beilstein journal of organic chemistry* **2016,** *12*, 2694-2718.

2. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr., Computational methods in drug discovery. *Pharmacological reviews* **2014,** *66* (1), 334-95.

3. Macalino, S. J.; Gosu, V.; Hong, S.; Choi, S., Role of computer-aided drug design in modern drug discovery. *Archives of pharmacal research* **2015,** *38* (9), 1686-701.

4. Anderson, A. C., The process of structure-based drug design. *Chemistry & biology* **2003,** *10* (9), 787-97.

5. Kalyaanamoorthy, S.; Chen, Y. P., Structure-based drug design to augment hit discovery. *Drug discovery today* **2011,** *16* (17-18), 831-9.

6. Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D., Molecular docking and structure-based drug design strategies. *Molecules* **2015,** *20* (7), 13384-421.

7. Jorgensen, W. L., Drug discovery: Pulled from a protein's embrace. *Nature* **2010,** *466* (7302), 42-3.

8. Loew, G. H.; Villar, H. O.; Alkorta, I., Strategies for indirect computer-aided drug design. *Pharmaceutical research* **1993,** *10* (4), 475-86.

9. Mason, J. S.; Good, A. C.; Martin, E. J., 3-D pharmacophores in drug discovery. *Current pharmaceutical design* **2001,** *7* (7), 567-97.

10. Alexander D. MacKerell, J. E. P., Andrew Coop, Chayan Acharya, Recent Advances In Ligand-Based Drug Design: Relevance and Utility of the Conformationally sampled Pharmacophore Approach *Current Computer Aided-Drug Design* **2011,** *7* (1), 13.

11. Vogt, M.; Bajorath, J., Predicting the performance of fingerprint similarity searching. *Methods in molecular biology* **2011,** *672*, 159-73.

12. Mukhsin Syuib, S. M. A., Nurul Malim, Comparison of Similarity Coefficients for Chemical Database Retrieval. *First International Conference on Artificial Intelligence, Modelling & Simulation* **2013**.

13. Yang, S. Y., Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug discovery today* **2010,** *15* (11-12), 444-50.

14. Doug Barnum, J. G., Andrew Smellie ,Peter Sprague, Identification of common functional configurations among molecules. *Journal of Chemical Information and Modelling* **1996,** *36* (3), 563-572.

15. Li, H.; Sutter, J.; Hoffmann, R., HypoGen: an automated system for generating 3D predictive pharmacophore models. *Pharmacophore perception, development, and use in drug design* **2000,** *2*, 171.

16. Martin, Y., DISCO: what we did right and what we missed. *Pharmacophore perception, development, and use in drug design* **2000,** *2*, 49-68.

17. Jones, G.; Willet, P.; Glen, R., GASP: genetic algorithm superimposition program. *Pharmacophore perception, development, and use in drug design* **2000**, 85-106.

18. Chao, W.-R.; Yean, D.; Amin, K.; Green, C.; Jong, L., Computer-aided rational drug design: a novel agent (SR13668) designed to mimic the unique anticancer mechanisms of dietary indole-3-carbinol to block Akt signaling. *Journal of medicinal chemistry* **2007,** *50* (15), 3412-3415.

19. Verma, R. P.; Hansch, C., Camptothecins: a SAR/QSAR study. *Chemical reviews* **2008,** *109* (1), 213-235.

20. Verma, J.; Khedkar, V. M.; Coutinho, E. C., 3D-QSAR in drug design-a review. *Current topics in medicinal chemistry* **2010,** *10* (1), 95-115.

21. Koga, H.; Itoh, A.; Murayama, S.; Suzue, S.; Irikura, T., Structure-activity relationships of antibacterial 6, 7-and 7, 8-disubstituted 1-alkyl-1, 4-dihydro-4-oxoquinoline-3-carboxylic acids. *Journal of medicinal chemistry* **1980,** *23* (12), 1358-1363.

22. Vyas, V. K.; Ukawala, R. D.; Ghate, M.; Chintha, C., Homology modeling a fast tool for drug discovery: current perspectives. *Indian journal of pharmaceutical sciences* **2012,** *74* (1), 1-17.

23. Bowie, J. U.; Luthy, R.; Eisenberg, D., A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **1991,** *253* (5016), 164-170.

24. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L., NCBI BLAST: a better web interface. *Nucleic acids research* **2008,** *36* (suppl 2), W5-W9.

25. Wong, W.-C.; Maurer-Stroh, S.; Eisenhaber, F., Not all transmembrane helices are born equal: Towards the extension of the sequence homology concept to membrane proteins. *Biology direct* **2011,** *6* (1), 57.

26. Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997,** *25* (17), 3389-3402.

27. Karplus, K.; Barrett, C.; Hughey, R., Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **1998,** *14* (10), 846-856.

28. Eddy, S. R., Profile hidden Markov models. *Bioinformatics* **1998,** *14* (9), 755-763.

29. Rychlewski, L.; Li, W.; Jaroszewski, L.; Godzik, A., Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science* **2000,** *9* (2), 232-241.

30. Jaroszewski, L.; Godzik, A.; Rychlewski, L., Improving the quality of twilight-zone alignments. *Protein Science* **2000,** *9* (8), 1487-1496.

31. Jaroszewski, L.; Rychlewski, L.; Li, Z.; Li, W.; Godzik, A., FFAS03: a server for profile–profile sequence alignments. *Nucleic acids research* **2005,** *33* (suppl 2), W284-W288.

32. Söding, J., Protein homology detection by HMM–HMM comparison. *Bioinformatics* **2005,** *21* (7), 951-960.

33. Marti-Renom, M. A.; Madhusudhan, M.; Sali, A., Alignment of protein sequences by their profiles. *Protein Science* **2004,** *13* (4), 1071-1087.

34. Cavasotto, C. N.; Phatak, S. S., Homology modeling in drug discovery: current trends and applications. *Drug discovery today* **2009,** *14* (13), 676-683.

35. Feng, D.-F.; Doolittle, R. F., Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution* **1987,** *25* (4), 351-360.

36. Thompson, J. D.; Gibson, T.; Higgins, D. G., Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics* **2002**, 2.3. 1-2.3. 22.

37. Notredame, C.; Higgins, D. G.; Heringa, J., T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* **2000,** *302* (1), 205-217.

38. Armougom, F.; Moretti, S.; Poirot, O.; Audic, S.; Dumas, P.; Schaeli, B.; Keduas, V.; Notredame, C., Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic acids research* **2006,** *34* (suppl 2), W604-W608.

39. Wallace, I. M.; O'Sullivan, O.; Higgins, D. G.; Notredame, C., M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic acids research* **2006,** *34* (6), 1692-1699.

40. Katoh, K.; Misawa, K.; Kuma, K. i.; Miyata, T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **2002,** *30* (14), 3059-3066.

41. Holm, L.; Kääriäinen, S.; Rosenström, P.; Schenkel, A., Searching protein structure databases with DaliLite v. 3. *Bioinformatics* **2008,** *24* (23), 2780-2781.

42. Orengo, C. A.; Taylor, W. R., [36] SSAP: sequential structure alignment program for protein structure comparison. *Methods in enzymology* **1996,** *266*, 617-635.

43. Shindyalov, I. N.; Bourne, P. E., A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic acids research* **2001,** *29* (1), 228-229.

44. Holm, L.; Sander, C., Protein structure comparison by alignment of distance matrices. *Journal of molecular biology* **1993,** *233* (1), 123-138.

45. Taylor, W. R.; Orengo, C. A., Protein structure alignment. *Journal of molecular biology* **1989,** *208* (1), 1-22.

46. Greer, J., Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins: Structure, Function, and Bioinformatics* **1990,** *7* (4), 317-334.

47. Browne, W. J.; North, A.; Phillips, D.; Brew, K.; Vanaman, T. C.; Hill, R. L., A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *Journal of molecular biology* **1969,** *42* (1), 65IN1371-7086.

48. Collura, V.; Higo, J.; Garnier, J., Modeling of protein loops by simulated annealing. *Protein Science* **1993,** *2* (9), 1502-1510.

49. Rost, B., Twilight zone of protein sequence alignments. *Protein engineering* **1999,** *12* (2), 85-94.

50. Xiang, Z., Advances in homology protein structure modeling. *Current Protein and Peptide Science* **2006,** *7* (3), 217-227.

51. Levitt, M., Accurate modeling of protein conformation by automatic segment matching. *Journal of molecular biology* **1992,** *226* (2), 507-533.

52. Sali, A.; Blundell, T., Comparative protein modelling by satisfaction of spatial restraints. *Protein structure by distance analysis* **1994,** *64*, C86.

53. Lee, J.; Lee, D.; Park, H.; Coutsias, E. A.; Seok, C., Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins: Structure, Function, and Bioinformatics* **2010,** *78* (16), 3428-3436.

54. Zhu, J.; Fan, H.; Periole, X.; Honig, B.; Mark, A. E., Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins: Structure, Function, and Bioinformatics* **2008,** *72* (4), 1171-1188.

55. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **1995,** *117* (19), 5179-5197.

56. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-1174.

57. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010,** *31* (4), 671-690.

58. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. a.; Karplus, M., CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* **1983,** *4* (2), 187-217.

59. Jorgensen, W. L.; Tirado-Rives, J., The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **1988,** *110* (6), 1657-1666.

60. Hillisch, A.; Pineda, L. F.; Hilgenfeld, R., Utility of homology models in the drug discovery process. *Drug discovery today* **2004,** *9* (15), 659-669.

61. Rodriguez, R.; Chinea, G.; Lopez, N.; Pons, T.; Vriend, G., Homology modeling, model and software evaluation: three related resources. *Bioinformatics* **1998,** *14* (6), 523-528.

62. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M., PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography* **1993,** *26* (2), 283-291.

63. Eisenberg, D.; Lüthy, R.; Bowie, J. U., [20] VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods in enzymology* **1997,** *277*, 396-404.

64. Sippl, M. J., Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. *Journal of molecular biology* **1990,** *213* (4), 859-883.

65. Melo, F.; Devos, D.; Depiereux, E.; Feytmans, E. In *ANOLEA: a www server to assess protein structures*, ISMB, 1997; pp 187-190.

66. de Vries, S. J.; Schindler, C. E.; de Beauchêne, I. C.; Zacharias, M., A web interface for easy flexible protein-protein docking with ATTRACT. *Biophysical journal* **2015,** *108* (3), 462-465.

67. Vakser, I. A., Protein-protein docking: from interaction to interactome. *Biophysical journal* **2014,** *107* (8), 1785-93.

68. Moreira, I. S.; Fernandes, P. A.; Ramos, M. J., Protein–protein docking dealing with the unknown. *Journal of computational chemistry* **2010,** *31* (2), 317-342.

69. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics* **2002,** *47* (4), 409-443.

70. Dominguez, C.; Boelens, R.; Bonvin, A. M., HADDOCK: A protein− protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **2003,** *125* (7), 1731-1737.

71. Li, C. H.; Ma, X. H.; Zu Chen, W.; Wang, C. X., A protein–protein docking algorithm dependent on the type of complexes. *Protein engineering* **2003,** *16* (4), 265-269.

72. Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A., Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences* **1992,** *89* (6), 2195-2199.

73. Yu, Y.; Lu, B.; Han, J.; Zhang, P., Scoring protein–protein docked structures based on the balance and tightness of binding. *Journal of computer-aided molecular design* **2004,** *18* (4), 251-260.

74. Camacho, C. J.; Gatchell, D. W., Successful discrimination of protein interactions. *Proteins: Structure, Function, and Bioinformatics* **2003,** *52* (1), 92-97.

75. Tovchigrechko, A.; Vakser, I. A., GRAMM-X public web server for protein–protein docking. *Nucleic Acids Research* **2006,** *34* (suppl_2), W310-W314.

76. Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J., ClusPro: a fully automated algorithm for protein–protein docking. *Nucleic Acids Research* **2004,** *32* (suppl_2), W96-W99.

77. Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D., Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology* **2003,** *331* (1), 281-299.

78. Lyskov, S.; Gray, J. J., The RosettaDock server for local protein–protein docking. *Nucleic acids research* **2008,** *36* (suppl 2), W233-W238.

79. Van Zundert, G.; Rodrigues, J.; Trellet, M.; Schmitz, C.; Kastritis, P.; Karaca, E.; Melquiond, A.; van Dijk, M.; De Vries, S.; Bonvin, A., The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology* **2016,** *428* (4), 720-725.

80. Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J., Geometry-based flexible and symmetric protein docking. *Proteins: Structure, Function, and Bioinformatics* **2005,** *60* (2), 224-231.

81. Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J., PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research* **2005,** *33* (suppl 2), W363-W367.

82. Li, L.; Chen, R.; Weng, Z., RDOCK: Refinement of rigid-body protein docking predictions. *Proteins: Structure, Function, and Bioinformatics* **2003,** *53* (3), 693-707.

83. Wiehe, K.; Pierce, B.; Mintseris, J.; Tong, W. W.; Anderson, R.; Chen, R.; Weng, Z., ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. *Proteins: Structure, Function, and Bioinformatics* **2005,** *60* (2), 207-213.

84. Chen, R.; Weng, Z., A novel shape complementarity scoring function for protein-protein docking. *Proteins: Structure, Function, and Bioinformatics* **2003,** *51* (3), 397-408.

85. Wiehe, K.; Pierce, B.; Tong, W. W.; Hwang, H.; Mintseris, J.; Weng, Z., The performance of ZDOCK and ZRANK in rounds 6–11 of CAPRI. *Proteins: Structure, Function, and Bioinformatics* **2007,** *69* (4), 719-725.

86. Chen, R.; Li, L.; Weng, Z., ZDOCK: an initial-stage protein-docking algorithm. *Proteins: Structure, Function, and Bioinformatics* **2003,** *52* (1), 80-87.

87. Pierce, B. G.; Wiehe, K.; Hwang, H.; Kim, B.-H.; Vreven, T.; Weng, Z., ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* **2014,** *30* (12), 1771-1773.

88. Macindoe, G.; Mavridis, L.; Venkatraman, V.; Devignes, M.-D.; Ritchie, D. W., HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Research* **2010,** *38* (suppl_2), W445-W449.

89. Durrant, J. D.; McCammon, J. A., Molecular dynamics simulations and drug discovery. *BMC biology* **2011,** *9* (1), 71.

90. De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A., Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem* **2016,** *59* (9), 4035-4061.

91. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* **1983,** *4* (2), 187-217.

92. Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C., The GROMOS software for biomolecular simulation: GROMOS05. *Journal of computational chemistry* **2005,** *26* (16), 1719-1751.

93. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling* **2006,** *25* (2), 247-260.

94. Betz, R. M.; Walker, R. C., Paramfit: automated optimization of force field parameters for molecular dynamics simulations. *Journal of computational chemistry* **2015,** *36* (2), 79-87.

95. Hopkins, C. W.; Roitberg, A. E., Fitting of dihedral terms in classical force fields as an analytic linear least-squares problem. *Journal of chemical information and modeling* **2014,** *54* (7), 1978-1986.

96. Huang, L.; Roux, B., Automated force field parameterization for nonpolarizable and polarizable atomic models based on ab initio target data. *Journal of chemical theory and computation* **2013,** *9* (8), 3543-3556.

97. Frenkel, D.; Smit, B., *Understanding molecular simulation: from algorithms to applications*. Academic press: 2001; Vol. 1.

98. Zhou, R., *Molecular Modeling at the Atomic Scale: Methods and Applications in Quantitative Biology*. CRC Press: 2014.

99. Best, R. B.; Buchete, N.-V.; Hummer, G., Are current molecular dynamics force fields too helical? *Biophysical journal* **2008,** *95* (1), L07-L09.

100. Andersen, H. C., Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics* **1983,** *52* (1), 24-34.

101. Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E., Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of chemical theory and computation* **2015,** *11* (4), 1864-1874.

102. Hünenberger, P. H., Thermostat algorithms for molecular dynamics simulations. *Advanced computer simulation* **2005**, 130-130.

103. MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B* **1998,** *102* (18), 3586-3616.

104. Li, D., The Andersen thermostat in molecular dynamics. *Communications on Pure and Applied Mathematics* **2008,** *61* (1), 96-136.

105. MacKerell, A. D.; Feig, M.; Brooks, C. L., Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of computational chemistry* **2004,** *25* (11), 1400-1415.

106. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **2010,** *78* (8), 1950-1958.

107. Allen, M. P.; Tildesley, D. J., *Computer simulation of liquids*. Oxford university press: 1989.

108. Cramer, C. J.; Bickelhaupt, F., Essentials of computational chemistry. *ANGEWANDTE CHEMIE-INTERNATIONAL EDITION IN ENGLISH*- **2003,** *42* (4), 381-381.

109. Berendsen, H. J.; Postma, J. P.; van Gunsteren, W. F.; Hermans, J., Interaction models for water in relation to protein hydration. In *Intermolecular forces*, Springer: 1981; pp 331-342.

110. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983,** *79* (2), 926-935.

111. Mahoney, M. W.; Jorgensen, W. L., A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of chemical physics* **2000,** *112* (20), 8910-8922.

112. Mashayak, S. Y.; Tanner, D. E., Comparing solvent models for molecular dynamics of protein. *University of Illinois at Urbana-Champaign, Champaign, IL* **2011**.

113. Wang, H.; Yu, M.; Ochani, M.; Amella, C. A.; Tanovic, M.; Susarla, S.; Li, J. H.; Wang, H.; Yang, H.; Ulloa, L., Nicotinic acetylcholine receptor α7 subunit is an essential regulator of inflammation. *Nature* **2003,** *421* (6921), 384-388.

114. Knight, J. L.; Brooks, C. L., Surveying implicit solvent models for estimating small molecule absolute hydration free energies. *Journal of computational chemistry* **2011,** *32* (13), 2909-2923.

115. Cuendet, M., Molecular dynamics simulation. EMBL: 2008.

116. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular

simulation programs. *Journal of computational chemistry* **2005,** *26* (16), 1668-1688.

117. Scott, W. R.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F., The GROMOS biomolecular simulation program package. *The Journal of Physical Chemistry A* **1999,** *103* (19), 3596-3607.

118. Berendsen, H. J.; van der Spoel, D.; van Drunen, R., GROMACS: a message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **1995,** *91* (1-3), 43-56.

119. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J., GROMACS: fast, flexible, and free. *Journal of computational chemistry* **2005,** *26* (16), 1701-1718.

120. Lindahl, E.; Hess, B.; van der Spoel, D., GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual* **2001,** *7* (8), 306-317.

121. Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K., NAMD2: greater scalability for parallel molecular dynamics. *Journal of Computational Physics* **1999,** *151* (1), 283-312.

122. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **2005,** *26* (16), 1781-1802.

123. Release, S., 1: Desmond Molecular Dynamics System, version 3.7. *DE Shaw Research, New York, NY, Maestro-Desmond Interoperability Tools, version* **2014,** *3*.

124. Kodadek, T., The rise, fall and reinvention of combinatorial chemistry. *Chemical communications* **2011,** *47* (35), 9757-9763.

125. Fang, Y., Ligand–receptor interaction platforms and their applications for drug discovery. *Expert opinion on drug discovery* **2012,** *7* (10), 969-988.

126. Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H., Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS journal* **2012,** *14* (1), 133-141.

127. Huang, S.-Y.; Zou, X., Advances and challenges in protein-ligand docking. *International journal of molecular sciences* **2010,** *11* (8), 3016-3034.

128. López-Vallejo, F.; Caulfield, T.; Martínez-Mayorga, K.; A Giulianotti, M.; Nefzi, A.; A Houghten, R.; L Medina-Franco, J., Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. *Combinatorial chemistry & high throughput screening* **2011,** *14* (6), 475-487.

129. Kapetanovic, I., Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. *Chemico-biological interactions* **2008,** *171* (2), 165-176.

130. Yuriev, E.; Agostino, M.; Ramsland, P. A., Challenges and advances in computational docking: 2009 in review. *Journal of Molecular Recognition* **2011,** *24* (2), 149-164.

131. Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E., Conformational sampling of bioactive molecules: a comparative study. *Journal of chemical information and modeling* **2007,** *47* (3), 1067-1086.

132. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J., Protein–ligand docking: current status and future challenges. *Proteins: Structure, Function, and Bioinformatics* **2006,** *65* (1), 15-26.

133. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology* **1997,** *267* (3), 727-748.

134. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of computer-aided molecular design* **2001,** *15* (5), 411-428.

135. Abagyan, R.; Totrov, M.; Kuznetsov, D., ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *Journal of computational chemistry* **1994,** *15* (5), 488-506.

136. Böhm, H.-J., The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *Journal of computer-aided molecular design* **1992,** *6* (1), 61-78.

137. McMartin, C.; Bohacek, R. S., QXP: powerful, rapid computer algorithms for structure-based drug design. *Journal of computer-aided molecular design* **1997,** *11* (4), 333-344.

138. McGann, M., FRED and HYBRID docking performance on standardized datasets. *Journal of computer-aided molecular design* **2012**, 1-10.

139. Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J., Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *Journal of computer-aided molecular design* **1996,** *10* (4), 293-304.

140. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P., eHiTS: a new fast, exhaustive flexible ligand docking system. *Journal of Molecular Graphics and Modelling* **2007,** *26* (1), 198-212.

141. Trosset, J.-Y.; Scheraga, H. A., PRODOCK: software package for protein modeling and docking. *Journal of computational chemistry* **1999,** *20* (4), 412-427.

142. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry* **2004,** *47* (7), 1739-1749.

143. Pei, J.; Wang, Q.; Liu, Z.; Li, Q.; Yang, K.; Lai, L., PSI-DOCK: Towards highly efficient and accurate flexible ligand docking. *Proteins: Structure, Function, and Bioinformatics* **2006,** *62* (4), 934-946.

144. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology* **1996,** *261* (3), 470-489.

145. Sochacka, J., Docking of thiopurine derivatives to human serum albumin and binding site analysis with Molegro Virtual Docker. *Acta Pol. Pharm* **2014,** *71*, 343-349.

146. Jain, A. N., Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of medicinal chemistry* **2003,** *46* (4), 499-511.

147. Thomsen, R.; Christensen, M. H., MolDock: a new technique for high-accuracy molecular docking. *Journal of medicinal chemistry* **2006,** *49* (11), 3315-3321.

148. Mizutani, M. Y.; Tomioka, N.; Itai, A., Rational automatic search method for stable docking models of protein and ligand. *Journal of molecular biology* **1994,** *243* (2), 310-326.

149. Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D., Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins: Structure, Function, and Bioinformatics* **1998,** *33* (3), 367-382.

150. Schnecke, V.; Kuhn, L. A. In *Database screening for HIV protease ligands: the influence of binding-site conformation and representation on ligand selectivity*, ISMB, 1999; pp 242-251.

151. Corbeil, C. R.; Williams, C. I.; Labute, P., Variability in docking success rates due to dataset preparation. *Journal of computer-aided molecular design* **2012**, 1-12.

152. Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P., FLOG: a system to select ‘quasi-flexible’ligands complementary to a receptor of known

three-dimensional structure. *Journal of computer-aided molecular design* **1994,** *8* (2), 153-174.

153. Huang, S.-Y.; Grinter, S. Z.; Zou, X., Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Physical Chemistry Chemical Physics* **2010,** *12* (40), 12899-12908.

154. Tanaka, S.; Scheraga, H. A., Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **1976,** *9* (6), 945-950.

155. Miyazawa, S.; Jernigan, R. L., Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **1985,** *18* (3), 534-552.

156. Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C.; Waldman, M., LigScore: a novel scoring function for predicting binding affinities. *Journal of Molecular Graphics and Modelling* **2005,** *23* (5), 395-407.

157. Böhm, H.-J., The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of computer-aided molecular design* **1994,** *8* (3), 243-256.

158. Wang, R.; Lai, L.; Wang, S., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design* **2002,** *16* (1), 11-26.

159. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring

function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design* **1997,** *11* (5), 425-445.

160. Rognan, D.; Lauemøller, S. L.; Holm, A.; Buus, S.; Tschinke, V., Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *Journal of medicinal chemistry* **1999,** *42* (22), 4650-4658.

161. Wang, R.; Liu, L.; Lai, L.; Tang, Y., SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *Journal of molecular modeling* **1998,** *4* (12), 379-394.

162. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *Journal of molecular biology* **2000,** *295* (2), 337-356.

163. DeWitte, R. S.; Shakhnovich, E. I., SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *Journal of the American Chemical Society* **1996,** *118* (47), 11733-11744.

164. Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W., Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *Journal of medicinal chemistry* **1999,** *42* (14), 2498-2503.

165. Fan, H.; Schneidman-Duhovny, D.; Irwin, J. J.; Dong, G.; Shoichet, B. K.; Sali, A., Statistical potential for modeling and ranking of protein–ligand interactions. *Journal of chemical information and modeling* **2011,** *51* (12), 3078-3092.

166. Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jørgensen, F. S., A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein− ligand binding affinities. *Journal of medicinal chemistry* **2001,** *44* (14), 2333-2343.

167. Betzi, S.; Suhre, K.; Chétrit, B.; Guerlesquin, F.; Morelli, X., GFscore: a general nonlinear consensus scoring function for high-throughput docking. *Journal of chemical information and modeling* **2006,** *46* (4), 1704-1712.

168. Jacobsson, M. Structure-Based Virtual Screening: New Methods and Applications in Infectious Diseases. Acta Universitatis Upsaliensis, 2008.

169. Berg, R. L. Identification of inhibitors of tryptophan hydroxylase 1. The University of Bergen, 2014.

170. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996,** *118* (45), 11225-11236.

171. Girinathan, B. P.; Braun, S. E.; Govind, R., Clostridium difficile glutamate dehydrogenase is a secreted enzyme that confers resistance to H2O2. *Microbiology* **2014,** *160* (1), 47-55.

172. Lessa, F. C.; Mu, Y.; Bamberg, W. M.; Beldavs, Z. G.; Dumyati, G. K.; Dunn, J. R.; Farley, M. M.; Holzbauer, S. M.; Meek, J. I.; Phipps, E. C., Burden of Clostridium difficile infection in the United States. *New England Journal of Medicine* **2015,** *372* (9), 825-834.

173. Miller, B. A.; Chen, L. F.; Sexton, D. J.; Anderson, D. J., Comparison of the burdens of hospital-onset, healthcare facility-associated Clostridium difficile infection and of healthcare-associated infection due to methicillin-resistant Staphylococcus aureus in community hospitals. *Infection Control & Hospital Epidemiology* **2011,** *32* (04), 387-390.

174. Magill, S. S.; Edwards, J. R.; Bamberg, W.; Beldavs, Z. G.; Dumyati, G.; Kainer, M. A.; Lynfield, R.; Maloney, M.; McAllister-Hollod, L.; Nadle, J., Multistate point-prevalence survey of health care–associated infections. *New England Journal of Medicine* **2014,** *370* (13), 1198-1208.

175. Dubberke, E. R.; Olsen, M. A., Burden of Clostridium difficile on the healthcare system. *Clinical infectious diseases* **2012,** *55* (suppl 2), S88-S92.

176. Rupnik, M.; Wilcox, M. H.; Gerding, D. N., Clostridium difficile infection: new developments in epidemiology and pathogenesis. *Nat Rev Micro* **2009,** *7* (7), 526-536.

177. Adams, H. M.; Li, X.; Mascio, C.; Chesnel, L.; Palmer, K. L., Mutations associated with reduced surotomycin susceptibility in Clostridium difficile and Enterococcus species. *Antimicrobial agents and chemotherapy* **2015,** *59* (7), 4139-4147.

178. Kuehne, S. A.; Cartman, S. T.; Heap, J. T.; Kelly, M. L.; Cockayne, A.; Minton, N. P., The role of toxin A and toxin B in Clostridium difficile infection. *Nature* **2010,** *467* (7316), 711-713.

179. Voth, D. E.; Ballard, J. D., Clostridium difficile toxins: mechanism of action and role in disease. *Clinical microbiology reviews* **2005,** *18* (2), 247-263.

180. Rupnik, M.; Wilcox, M. H.; Gerding, D. N., Clostridium difficile infection: new developments in epidemiology and pathogenesis. *Nature Reviews Microbiology* **2009,** *7* (7), 526-536.

181. Kelly , C. P.; LaMont , J. T., Clostridium difficile — More Difficult Than Ever. *New England Journal of Medicine* **2008,** *359* (18), 1932-1940.

182. Ghose, C., Clostridium difficile infection in the twenty-first century. *Emerg Microbes Infect* **2013,** *2*, e62.

183. Cronan, J. E., Bacterial membrane lipids: where do we stand? *Annual reviews in microbiology* **2003,** *57* (1), 203-224.

184. White, S. W.; Zheng, J.; Zhang, Y.-M.; Rock, C. O., The structural biology of type II fatty acid biosynthesis. *Annu. Rev. Biochem.* **2005,** *74*, 791-831.

185. Zhang, Y.-M.; White, S. W.; Rock, C. O., Inhibiting bacterial fatty acid synthesis. *Journal of Biological Chemistry* **2006,** *281* (26), 17541-17544.

186. Marrakchi, H.; DeWOLF, W. E.; Quinn, C.; Joshua, W.; Polizzi, B. J.; HOLMES, D. J.; HEATH, R. J.; PAYNE, D. J.; WALLIS, N. G., Characterization of Streptococcus pneumoniae enoyl-(acyl-carrier protein) reductase (FabK). *Biochemical Journal* **2003,** *370* (3), 1055-1062.

187. Heath, R. J.; Rock, C. O., Microbiology: A triclosan-resistant bacterial enzyme. *Nature* **2000,** *406* (6792), 145-146.

188. Release, S., 1: Maestro, version 10.1. *Schrödinger, LLC, New York, NY* **2015**.

189. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004,** *25* (13), 1605-1612.

190. Huang, C. C.; Meng, E. C.; Morris, J. H.; Pettersen, E. F.; Ferrin, T. E., Enhancing UCSF Chimera through web services. *Nucleic acids research* **2014,** *42* (W1), W478-W484.

191. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *Journal of molecular graphics* **1996,** *14* (1), 33-38.

192. Developers, M.; Chan, K.-Y.; McGreevy, R.; Trabuco, L. G.; Villa, E., Molecular Dynamics Flexible Fitting. **2016**.

193. Sebaihia, M.; Wren, B. W.; Mullany, P.; Fairweather, N. F.; Minton, N.; Stabler, R.; Thomson, N. R.; Roberts, A. P.; Cerdeño-Tárraga, A. M.; Wang, H., The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nature genetics* **2006,** *38* (7), 779-786.

194. Monot, M.; Boursaux-Eude, C.; Thibonnier, M.; Vallenet, D.; Moszer, I.; Medigue, C.; Martin-Verstraete, I.; Dupuy, B., Reannotation of the genome sequence of Clostridium difficile strain 630. *Journal of medical microbiology* **2011,** *60* (8), 1193-1199.

195. National Center of Biotechnology information (NCBI). Protein Home Page. https://www.ncbi.nlm.nih.gov/protein/ (accessed 26 November, 2016).

196. Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T., The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **2006,** *22* (2), 195-201.

197. Bowie, J. U.; Lüthy, R.; Eisenberg, D., A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **1991**, 164-170.

198. Luthy, R.; Bowie, J. U.; Eisenberg, D., Assessment of protein models with three-dimensional profiles. *Nature* **1992,** *356* (6364), 83.

199. UCLA-DOE lab homepage. The Structual Analysis and Verification Server http://services.mbi.ucla.edu/SAVES/ (accessed 25 Novemebr, 2016).

200. Benkert, P.; Biasini, M.; Schwede, T., Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **2011,** *27* (3), 343-350.

201. Release, S., 3: Schrödinger Suite 2015-3 Protein Preparation Wizard; Epik version 3.3. Schrödinger, LLC, New York, NY: 2015.

202. Sgro, J.-Y., MODELLER-II-Chimera GUI interface. **2017**.

203. Hevener, K. E.; Mehboob, S.; Boci, T.; Truong, K.; Santarsiero, B. D.; Johnson, M. E., Expression, purification and characterization of enoyl-ACP reductase II, FabK, from Porphyromonas gingivalis. *Protein expression and purification* **2012,** *85* (1), 100-108.

204. Saito, J.; Yamada, M.; Watanabe, T.; Iida, M.; Kitagawa, H.; Takahata, S.; Ozawa, T.; Takeuchi, Y.; Ohsawa, F., Crystal structure of enoyl–acyl carrier

protein reductase (FabK) from Streptococcus pneumoniae reveals the binding mode of an inhibitor. *Protein Science* **2008,** *17* (4), 691-699.

205. Villar, R. G.; Macek, M. D.; Simons, S.; Hayes, P. S.; Goldoft, M. J.; Lewis, J. H.; Rowan, L. L.; Hursh, D.; Patnode, M.; Mead, P. S., Investigation of multidrug-resistant Salmonella serotype Typhimurium DT104 infections linked to raw-milk cheese in Washington State. *Jama* **1999,** *281* (19), 1811-1816.

206. Vital signs: incidence and trends of infection with pathogens transmitted commonly through food--foodborne diseases active surveillance network, 10 U.S. sites, 1996-2010. *MMWR. Morbidity and mortality weekly report* **2011,** *60* (22), 749-55.

207. Uchida, I.; Ishihara, R.; Tanaka, K.; Hata, E.; Makino, S.-i.; Kanno, T.; Hatama, S.; Kishima, M.; Akiba, M.; Watanabe, A., Salmonella enterica serotype Typhimurium DT104 ArtA-dependent modification of pertussis toxin-sensitive G proteins in the presence of [32P] NAD. *Microbiology* **2009,** *155* (11), 3710-3718.

208. Hermans, A. P.; Abee, T.; Zwietering, M. H.; Aarts, H. J., Identification of novel Salmonella enterica serovar Typhimurium DT104-specific prophage and nonprophage chromosomal sequences among serovar Typhimurium isolates by genomic subtractive hybridization. *Applied and environmental microbiology* **2005,** *71* (9), 4979-4985.

209. Glynn, M. K.; Bopp, C.; Dewitt, W.; Dabney, P.; Mokhtar, M.; Angulo, F. J., Emergence of Multidrug-Resistant Salmonella enterica SerotypeTyphimurium

DT104 Infections in the United States. *New England Journal of Medicine* **1998,** *338* (19), 1333-1339.

210. Sameshima, T.; Akiba, M.; Izumiya, H.; Terajima, J.; Tamura, K.; Watanabe, H.; Nakazawa, M., Salmonella typhimurium DT104 from livestock in Japan. *Japanese journal of infectious diseases* **2000,** *53* (1), 15-16.

211. Threlfall, E.; Frost, J.; Ward, L.; Rowe, B., Epidemic in cattle and humans of Salmonella typhimurium DT 104 with chromosomally integrated multiple drug resistance. *Veterinary Record* **1994,** *134* (22), 577-577.

212. Briggs, C. E.; Fratamico, P. M., Molecular characterization of an antibiotic resistance gene cluster of Salmonella typhimuriumDT104. *Antimicrobial agents and chemotherapy* **1999,** *43* (4), 846-849.

213. Allen, C. A.; Fedorka-Cray, P. J.; Vazquez-Torres, A.; Suyemoto, M.; Altier, C.; Ryder, L. R.; Fang, F. C.; Libby, S. J., In vitro and in vivo assessment of Salmonella enterica serovar Typhimurium DT104 virulence. *Infection and immunity* **2001,** *69* (7), 4673-4677.

214. Saitoh, M.; Tanaka, K.; Nishimori, K.; Makino, S.-i.; Kanno, T.; Ishihara, R.; Hatama, S.; Kitano, R.; Kishima, M.; Sameshima, T., The artAB genes encode a putative ADP-ribosyltransferase toxin homologue associated with Salmonella enterica serovar Typhimurium DT104. *Microbiology* **2005,** *151* (9), 3089-3096.

215. Wang, H.; Paton, J. C.; Herdman, B. P.; Rogers, T. J.; Beddoe, T.; Paton, A. W., The B subunit of an AB5 toxin produced by Salmonella enterica serovar Typhi up-regulates chemokines, cytokines, and adhesion molecules in human

macrophage, colonic epithelial, and brain microvascular endothelial cell lines. *Infection and immunity* **2013,** *81* (3), 673-683.

216. SWISS-MODEL. http://swissmodel.expasy.org/workspace/index.php?func=tools_structureassessment1 (accessed 23 Novemeber, 2016).

217. Ng, N. M.; Littler, D. R.; Paton, A. W.; Le Nours, J.; Rossjohn, J.; Paton, J. C.; Beddoe, T., EcxAB is a founding member of a new family of metalloprotease AB 5 toxins with a hybrid cholera-like B subunit. *Structure* **2013,** *21* (11), 2003-2013.

218. 5, A. t. B. s. Running minimization and MD in explicit solvent. http://ambermd.org/tutorials/basic/tutorial1/section5.htm (accessed 6/26/2017).

219. Cavasotto, C. N.; Orry, A. J. W.; Murgolo, N. J.; Czarniecki, M. F.; Kocsi, S. A.; Hawes, B. E.; O'Neill, K. A.; Hine, H.; Burton, M. S.; Voigt, J. H.; Abagyan, R. A.; Bayne, M. L.; Monsma, F. J., Discovery of Novel Chemotypes to a G-Protein-Coupled Receptor through Ligand-Steered Homology Modeling and Structure-Based Virtual Screening. *Journal of medicinal chemistry* **2008,** *51* (3), 581-588.

220. Cavasotto, C. N.; Abagyan, R. A., Protein flexibility in ligand docking and virtual screening to protein kinases. *Journal of molecular biology* **2004,** *337* (1), 209-225.

221. Kovacs, J. A.; Cavasotto, C. N.; Abagyan, R., Conformational sampling of protein flexibility in generalized coordinates: application to ligand docking. *Journal of Computational and Theoretical Nanoscience* **2005,** *2* (3), 354-361.

222. Monti, M. C.; Casapullo, A.; Cavasotto, C. N.; Napolitano, A.; Riccio, R., Scalaradial, a Dialdehyde-Containing Marine Metabolite That Causes an Unexpected Noncovalent PLA2 Inactivation. *ChemBioChem* **2007,** *8* (13), 1585-1591.

223. Monti, M. C.; Casapullo, A.; Cavasotto, C. N.; Tosco, A.; Dal Piaz, F.; Ziemys, A.; Margarucci, L.; Riccio, R., The binding mode of petrosaspongiolide M to the human group IIA phospholipase A2: exploring the role of covalent and noncovalent interactions in the inhibition process. *Chemistry-A European Journal* **2009,** *15* (5), 1155-1163.

224. Hevener, K. E., *Structure-and ligand-based design of novel antimicrobial agents*. The University of Tennessee Health Science Center: 2008.

225. Ruvinsky, A. M., Role of binding entropy in the refinement of protein–ligand docking predictions: analysis based on the use of 11 scoring functions. *Journal of computational chemistry* **2007,** *28* (8), 1364-1372.

226. Bowman, G. R., Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *Journal of computational chemistry* **2016,** *37* (6), 558-566.

227. Torrie, G. M.; Valleau, J. P., Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977,** *23* (2), 187-199.

228. Zhou, R., Replica exchange molecular dynamics method for protein folding simulation. *Protein Folding Protocols* **2006**, 205-223.

229. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002,** *99* (20), 12562-12566.

230. Isralewitz, B.; Gao, M.; Schulten, K., Steered molecular dynamics and mechanical functions of proteins. *Current opinion in structural biology* **2001,** *11* (2), 224-230.

231. Grubmüller, H.; Heymann, B.; Tavan, P., Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **1996,** (5251), 997-999.

232. Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics* **2004,** *120* (24), 11919-11929.

233. Faradjian, A. K.; Elber, R., Computing time scales from reaction coordinates by milestoning. *The Journal of chemical physics* **2004,** *120* (23), 10880-10889.

234. Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L., Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry* **2002,** *53* (1), 291-318.

235. Jorgensen, W. L.; Ravimohan, C., Monte Carlo simulation of differences in free energies of hydration. *The Journal of chemical physics* **1985,** *83* (6), 3050-3054.

236. Jorgensen, W. L.; Thomas, L. L., Perspective on free-energy perturbation calculations for chemical equilibria. *Journal of chemical theory and computation* **2008,** *4* (6), 869-876.

Appendices

B.1. Filtering file containing filtering criteria for ligand preparation of compound libraries.

```
# Custom patterns
DEFINE Aliphatic_thioester [S;X2]([!#1])[C;X3](=O)[C;X4;H2,H3]
DEFINE Hydrazines [NX3][NX3]
DEFINE Acyl_halides [F,Cl,Br,I][C;X3]=O
DEFINE Sulfonyl_halides [F,Cl,Br,I][S;X4](=O)=O
DEFINE Sulfinyl_halides [F,Cl,Br,I][S;X3]=O
DEFINE Sulfenyl_halides [F,Cl,Br,I][S;X2]
DEFINE Alkyl_halides_wo_fluorine [Cl,Br,I][C;X4;H2,H3]
DEFINE Anhydrides [O;X2]([C;X3]=O)([C;X3]=O)
DEFINE Perhalomethylketones
[#6][C;X3](=O)[C;X4]([F,Cl,Br,I])([F,Cl,Br,I])[F,Cl,Br,I]
DEFINE Aldehydes [#6][C;H1]=[O;X1]
DEFINE Formates [O;X2][C;H1]=O
DEFINE Peroxides [O;X2]~[O;X2]
DEFINE Isothiocyanates [#6][N;X2]=C=[S;X1]
DEFINE Isocyanates [#6][N;X2]=C=O
DEFINE Phosphinyl_halides [P;X3][Cl,Br,I]
DEFINE Phosphonyl_halides [P;X4](=O)[Cl,Br,I]
DEFINE Carbodiimides [#6][N;X2]=[C;X2]=[N;X2][#6]
DEFINE Silyl_enol_ethers C=CO[Si;X4]
DEFINE Nitroalkanes [#6][C;H2][N](~[O;X1])[O;X1]
DEFINE Phosphines [#6][#15]([#6])~[#6]
DEFINE Alkyl_sulfonates [#6]O[S;X4](=O)=O
DEFINE Epoxides [O;X2;r3](C)C
DEFINE Azides [#6][N;X2]=[N;X2]=[N;X1]
DEFINE Diazoniums [#6][N;X2]#[N;X1]
DEFINE Isonitriles [#6][N;X2]#[C;X1]
DEFINE Halopyrimidines [F,Cl,Br,I]c(nc)nc
```

DEFINE 1,2-Dicarbonyls [C;X3](=O)([C;X3](=O))

DEFINE Michael_acceptors [O;X1]=C[C;H1]=[C;H1]

DEFINE beta-Heterosubstituted_carbonyls [O;X1]=C[C;H2]C[F,Cl,Br,I]

DEFINE Diazos [N;X2]~[N;X2]

DEFINE Disulfides [S;X2]~[S;X2]

DEFINE Imines [N;X2]([!#1])=[C;X3][C;H2,H3]

DEFINE Aziridines [N;X3;r3](C)C

DEFINE Thiols [S;X2;H1]

DEFINE Aliphatic_ester [O;X2]([!#1])[C;X3](=O)[C;X4;H2,H3]

DEFINE Aliphatic_ketone [C;X4;H3][C;X3](=O)[C;X4;H2,H3]

DEFINE Thiourea [#1][N-0X3][C-0X3](=[S-0X1])[N-0X3][#1]

DEFINE Cyclohexanone [O-0X1]=[C-0X3]1[C-0X4][C-0X4][C-0X4][C-0X4][C-0X4]1

DEFINE Halogenated_hydrocarbons [F,Cl,Br,I][C;X4,H1][F,Cl,Br,I]

DEFINE Hydroxylamines [#1][N-0X3][O-0X2][#1]

DEFINE Iminoquinones [N-0X2]=[C-0X3]1[C-0X3]=[C-0X3][C-0X3](=[O-0X1])[C-0X3]=[C-0X3]1

DEFINE Halocarbonyl_Cl [Cl-0X1][C-0X4][C-0X3]=[O-0X1]

DEFINE Halocarbonyl_Br [Br-0X1][C-0X4][C-0X3]=[O-0X1]

DEFINE Halocarbonyl_F [F-0X1][C-0X4][C-0X3]=[O-0X1]

DEFINE Halocarbonyl_I [I-0X1][C-0X4][C-0X3]=[O-0X1]

DEFINE Heteroatom_Heteroatom_1 [N-0X3][O-0X2]

DEFINE Heteroatom_Heteroatom_2 [N,#8][N,#8]

DEFINE Heteroatom_Heteroatom_3 [N-0X3][S-0X2]

DEFINE Heteroatom_Heteroatom_4 [n-0X2][s-0X2]

DEFINE Heteroatom_Heteroatom_5 [O-0X2][S-0X2]

DEFINE Nitro_aromatic [a]-[$([NX3](=O)=O),$([NX3+](=O)[O-])][!#8]

#

Filter criteria

#

Aliphatic_thioester >= 1

Hydrazines >= 1

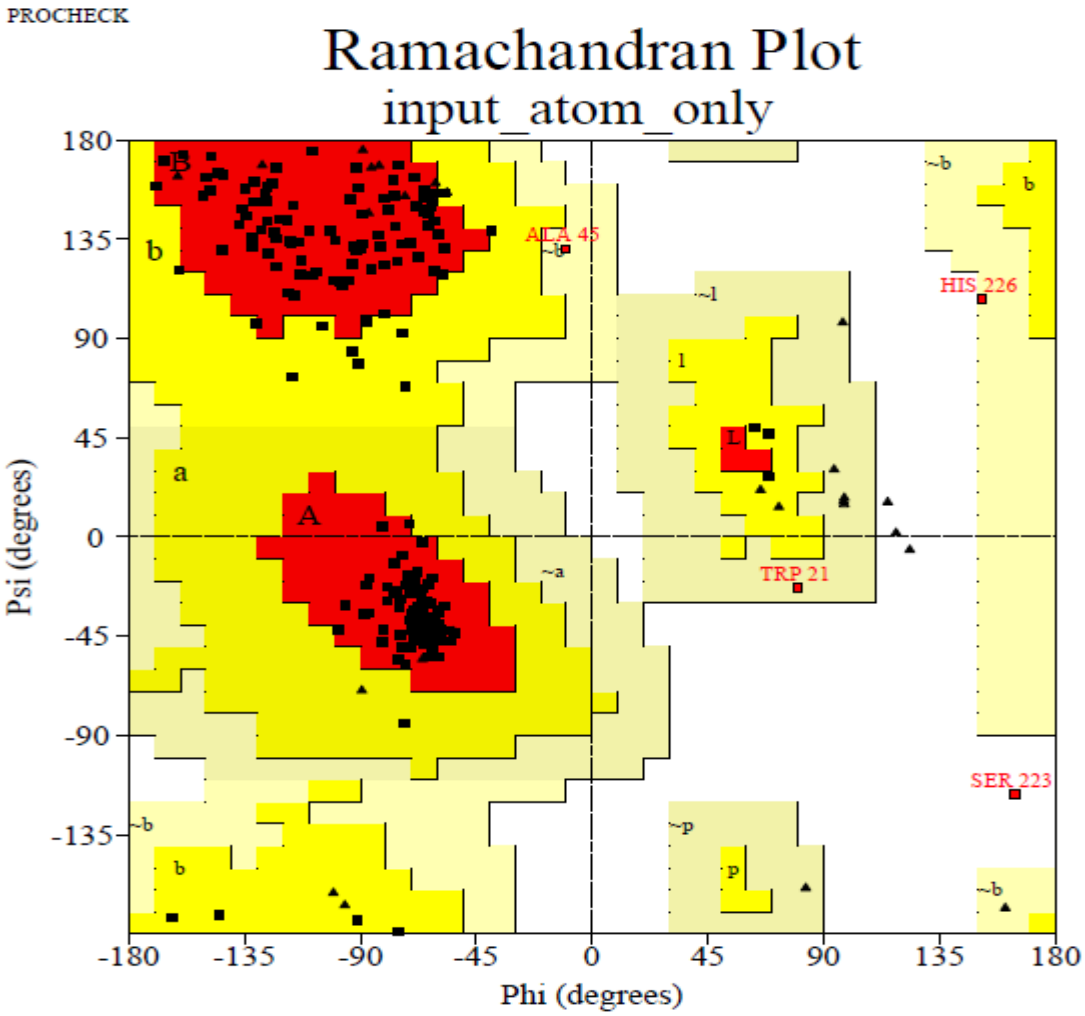
Acyl_halides >= 1

Sulfonyl_halides >= 1

Sulfinyl_halides	>= 1
Sulfenyl_halides	>= 1
Alkyl_halides_wo_fluorine	>= 1
Anhydrides	>= 1
Perhalomethylketones	>= 1
Aldehydes	>= 1
Formates	>= 1
Peroxides	>= 1
Isothiocyanates	>= 1
Isocyanates	>= 1
Phosphinyl_halides	>= 1
Phosphonyl_halides	>= 1
Carbodiimides	>= 1
Silyl_enol_ethers	>= 1
Nitroalkanes	>= 1
Phosphines	>= 1
Alkyl_sulfonates	>= 1
Epoxides	>= 1
Azides	>= 1
Diazoniums	>= 1
Isonitriles	>= 1
Halopyrimidines	>= 1
1,2-Dicarbonyls	>= 1
Michael_acceptors	>= 1
beta-Heterosubstituted_carbonyls	>= 1
Diazos	>= 1
Disulfides	>= 1
Imines	>= 1
Aziridines	>= 1
Thiols	>= 1
Aliphatic_ester	>= 1
Aliphatic_ketone	>= 1
Thiourea	>= 1

Cyclohexanone	>= 1
Halogenated_hydrocarbons	>= 1
Hydroxylamines	>= 1
Iminoquinones	>= 1
Halocarbonyl_F	>= 1
Halocarbonyl_Br	>= 1
Halocarbonyl_Cl	>= 1
Halocarbonyl_I	>= 1
Heteroatom_Heteroatom_1	>= 1
Heteroatom_Heteroatom_2	>= 1
Heteroatom_Heteroatom_3	>= 1
Heteroatom_Heteroatom_4	>= 1
Heteroatom_Heteroatom_5	>= 1
Nitro_aromatic	>= 1
Num_chiral_centers	>= 3
Molecular_weight	<= 150 OR >= 750

Figure B.2. Ramachandran plot of selected CdFabK homology model



Plot statistics

Residues in most favoured regions [A,B,L]	241	91.3%
Residues in additional allowed regions [a,b,l,p]	19	7.2%
Residues in generously allowed regions [~a,~b,~l,~p]	3	1.1%
Residues in disallowed regions	1	0.4%
Number of non-glycine and non-proline residues	264	100.0%
Number of end-residues (excl. Gly and Pro)	3	
Number of glycine residues (shown as triangles)	30	
Number of proline residues	13	
Total number of residues	310	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figure B.3 Main Chain bond lengths of the amino acid residues of CdFabK model

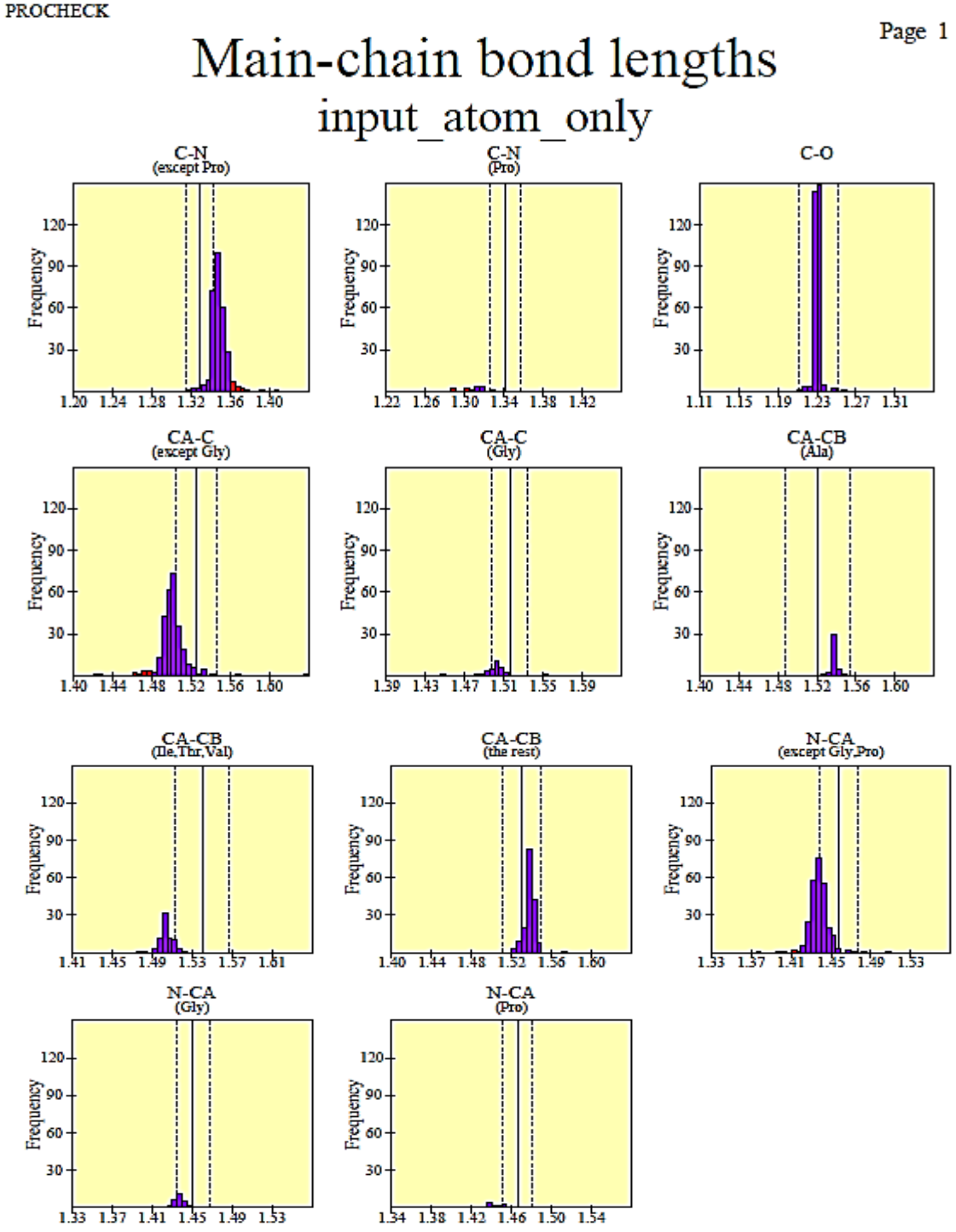
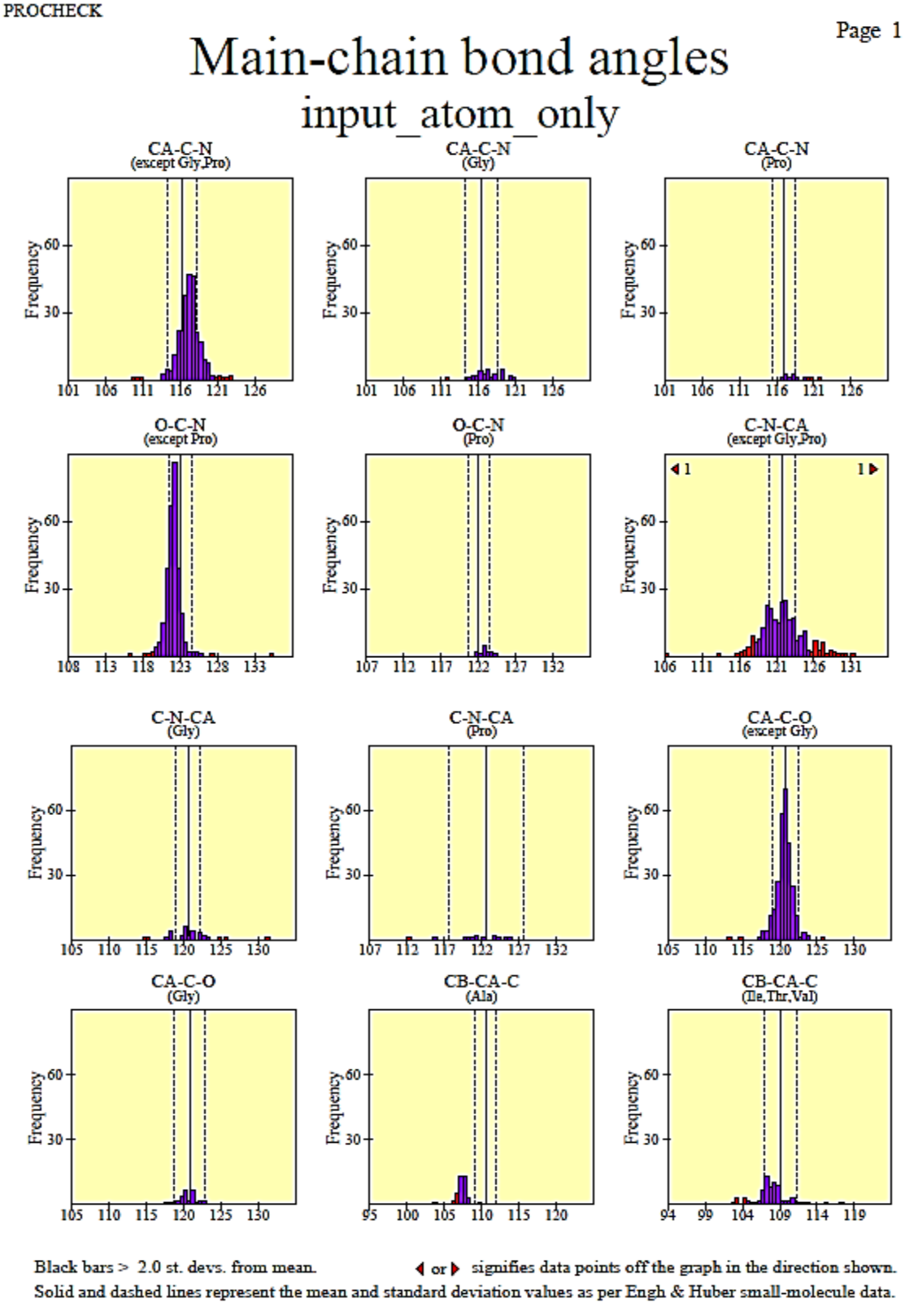


Figure B.4. Main chain bond angles of amino acid residues of CdFabK model



Main-chain bond angles

input_atom_only

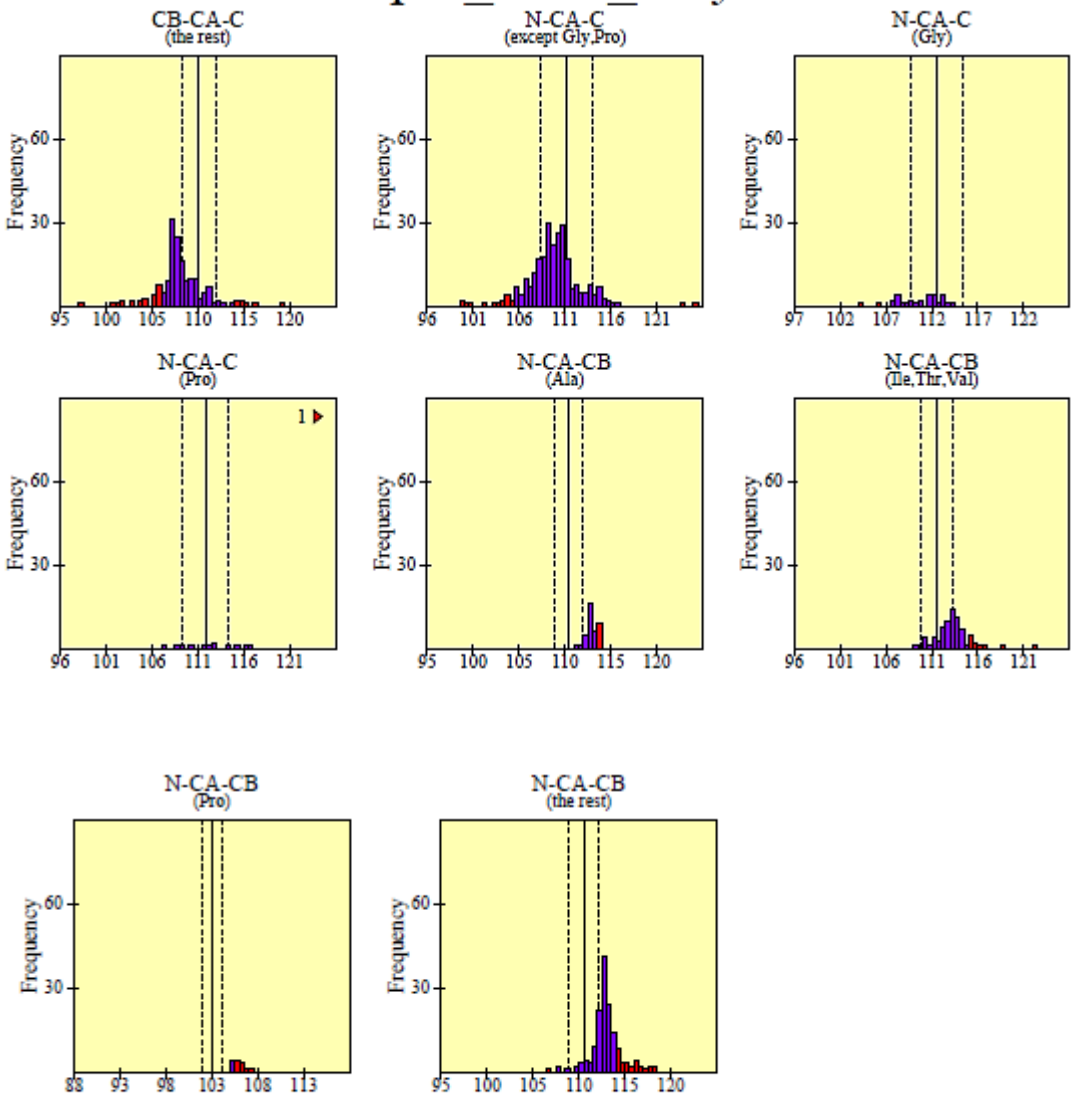


Figure B.5. RMS distances of planar atoms from the best-fit plane of CdFabK model

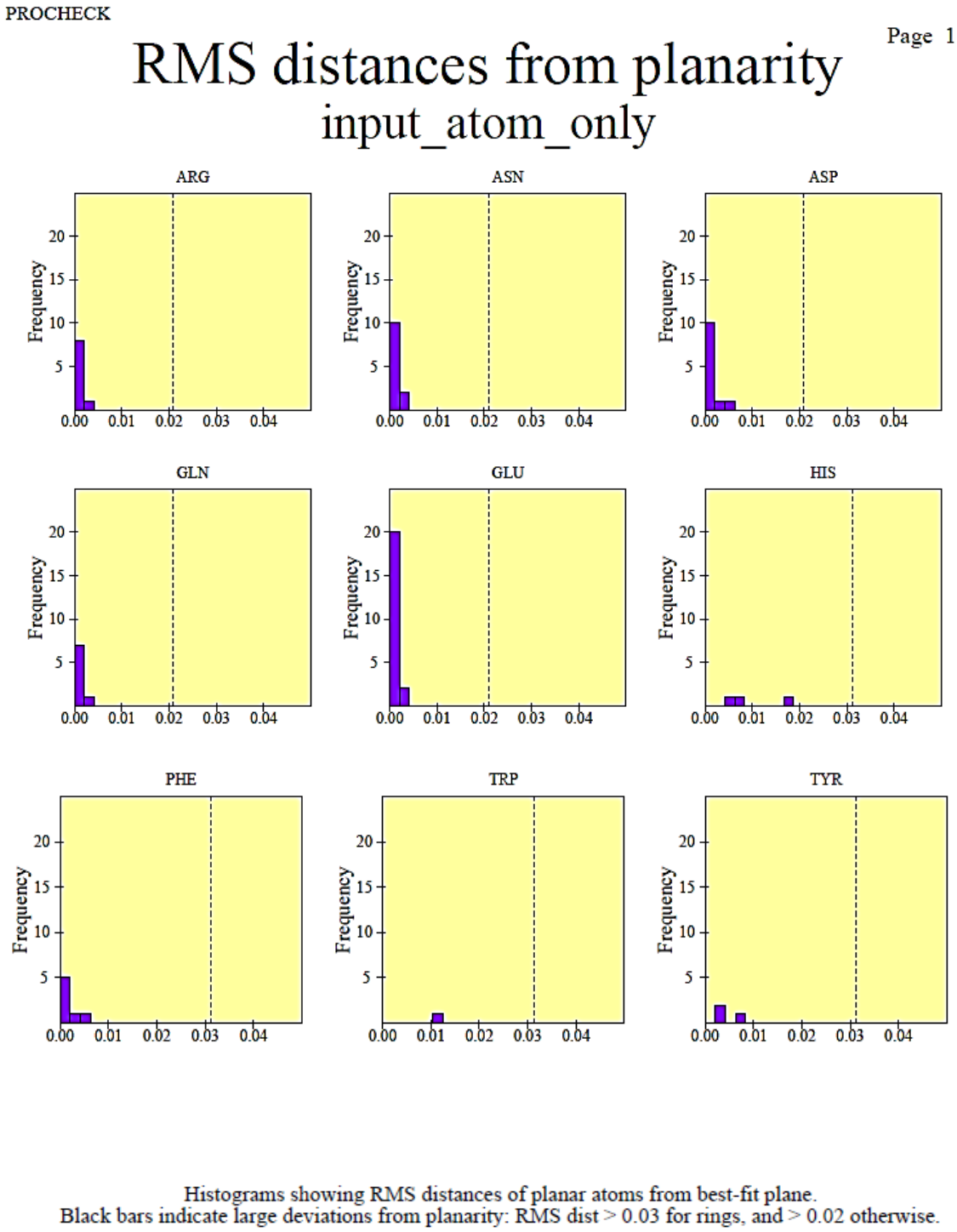


Figure C.1. Main Chain bond lengths of ArtA homology model

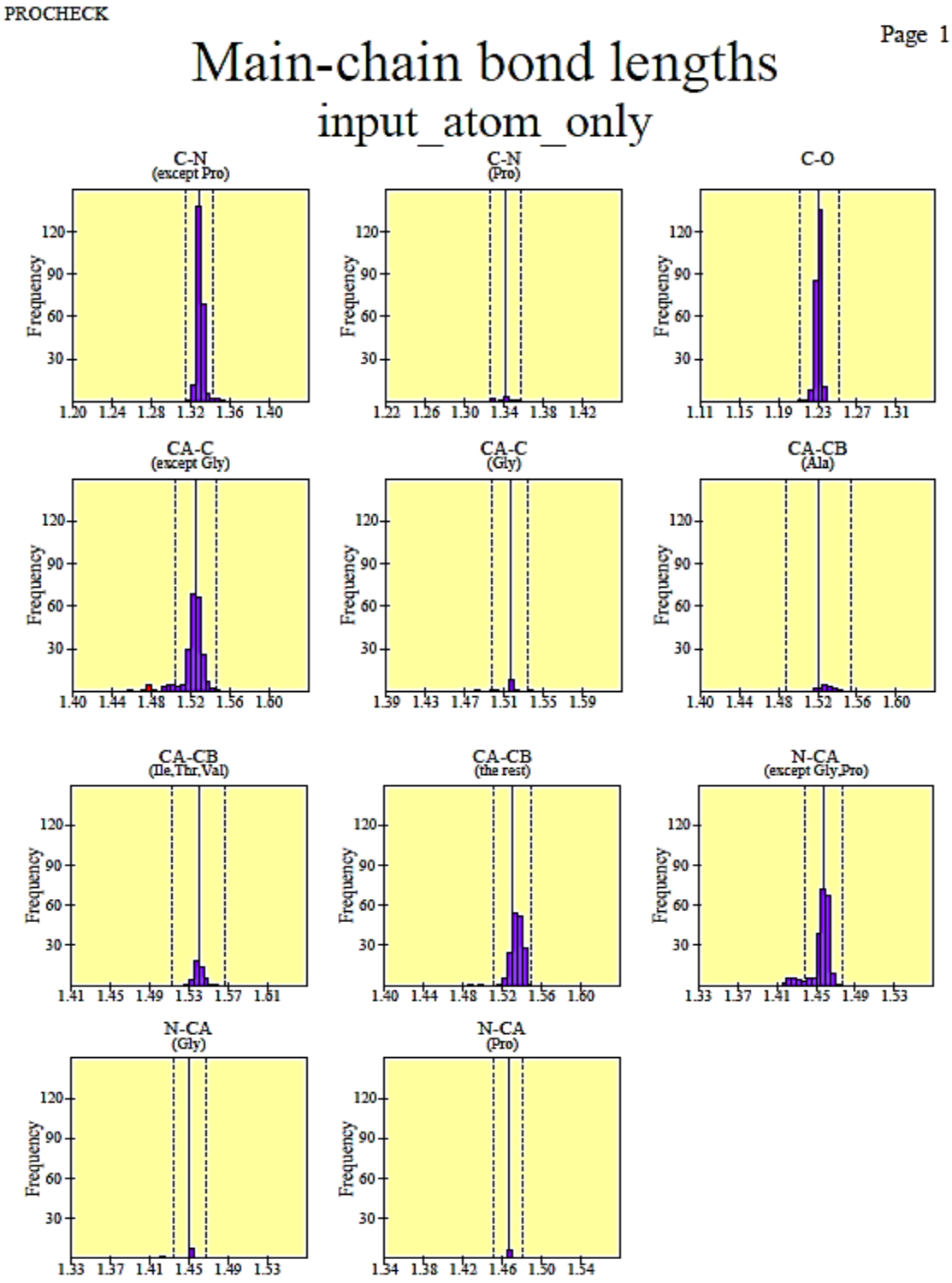
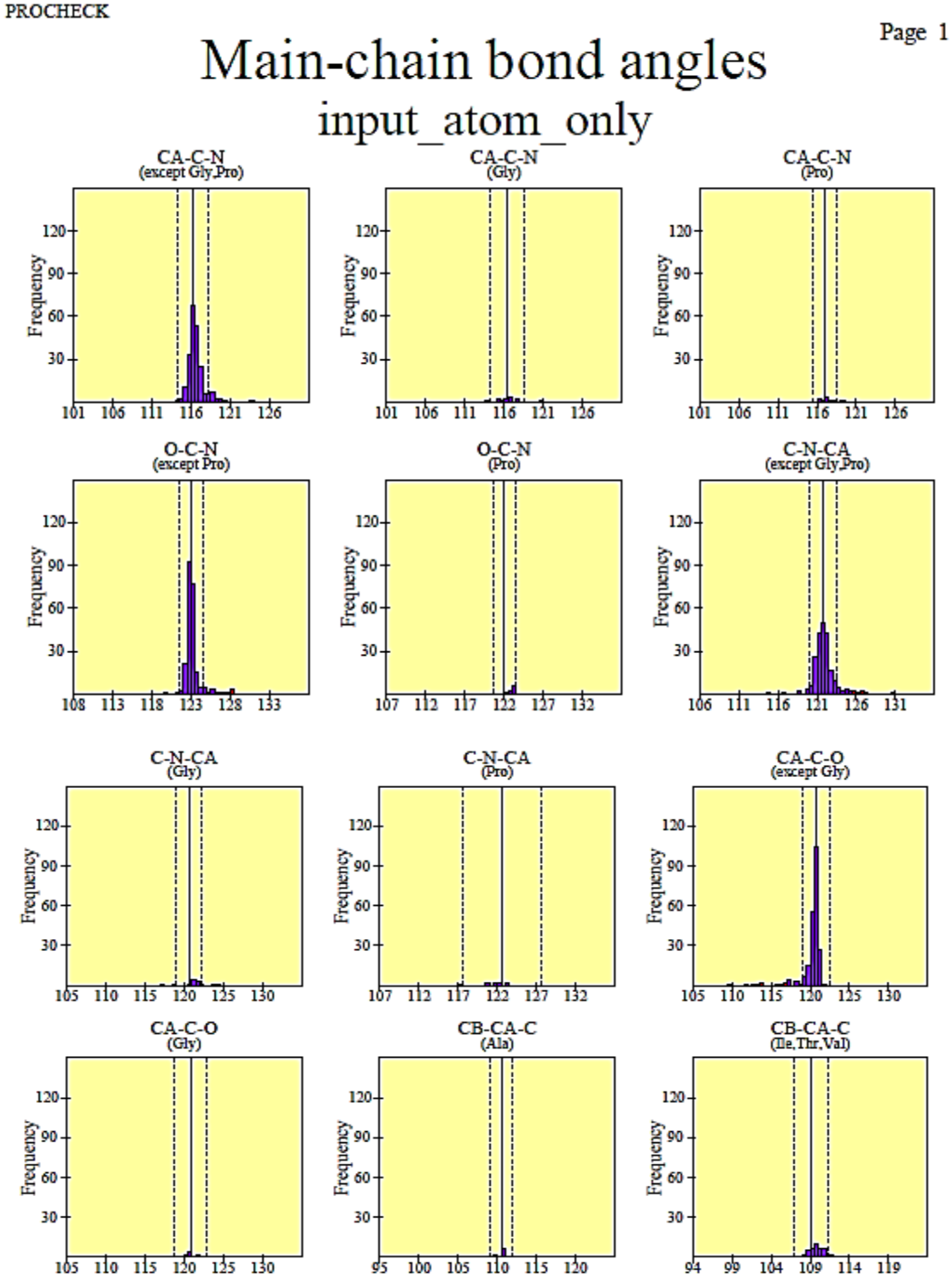
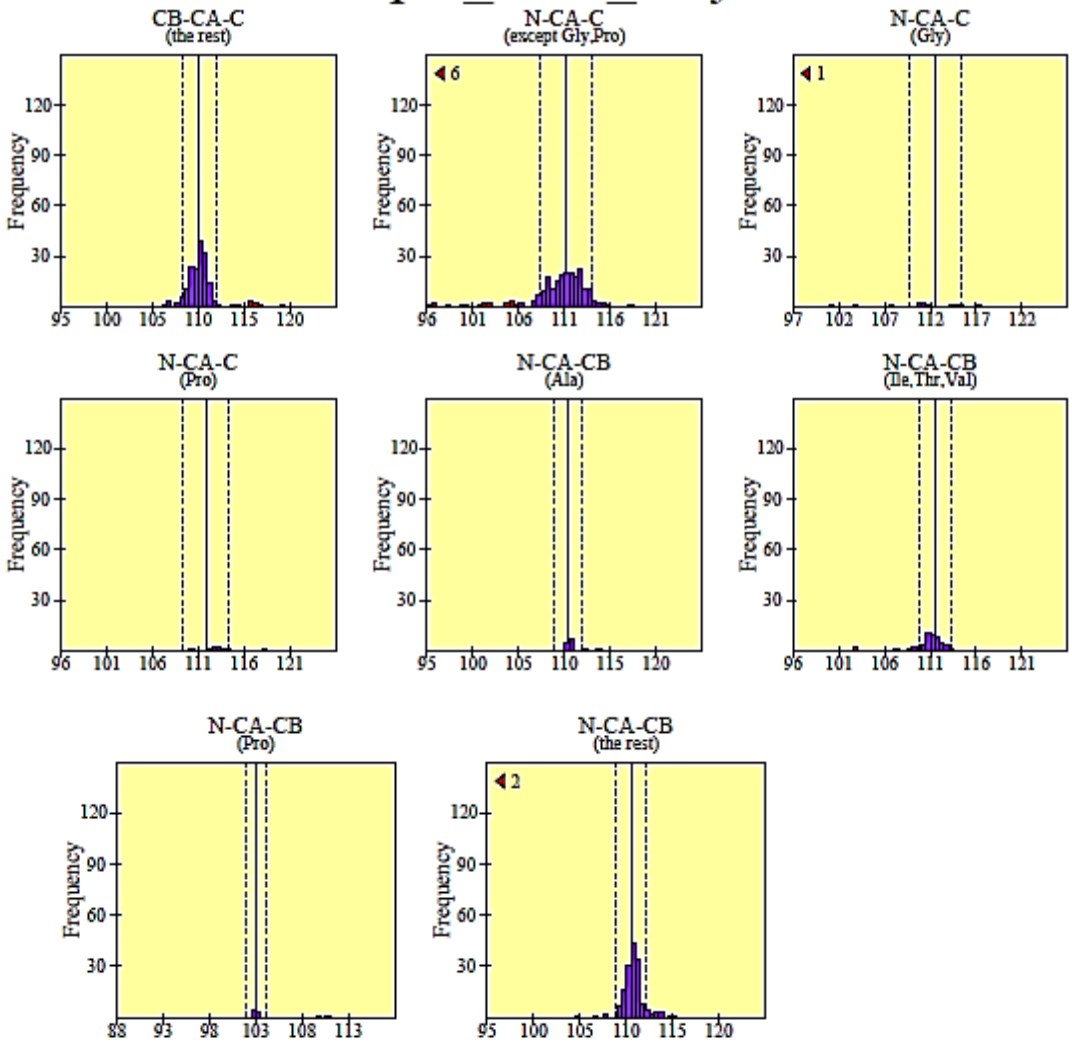


Figure C.2. Main chain bond angles of ArtA homology model



Main-chain bond angles

input_atom_only



Black bars > 2.0 st. devs. from mean. ◀ or ▶ signifies data points off the graph in the direction shown.
Solid and dashed lines represent the mean and standard deviation values as per Engh & Huber small-molecule data.

Figure C.3. Root mean square distances of planar atoms from the best-fit plane of ArtA model

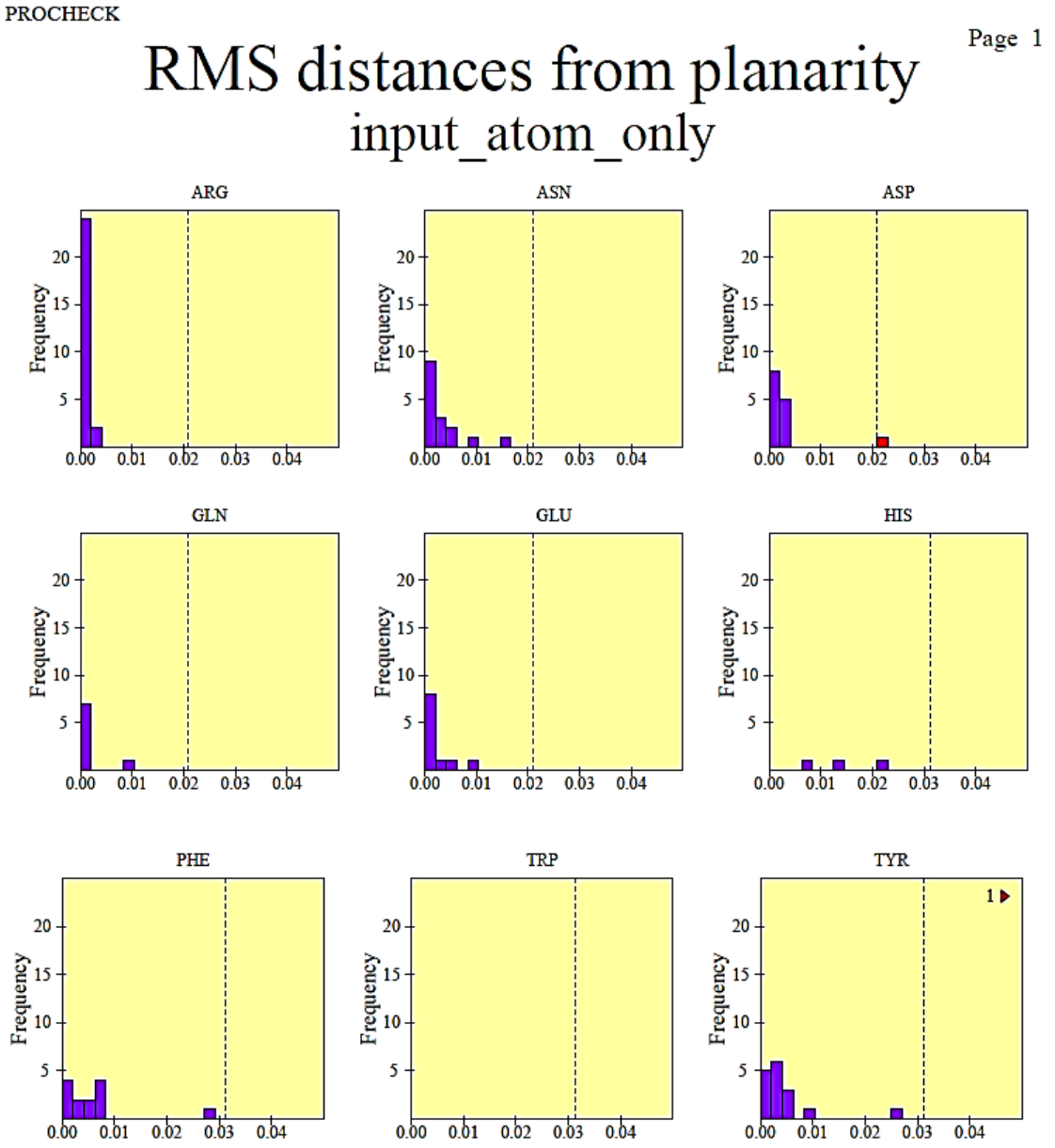


Figure C.4. Main chain bond lengths of ArtB homology model

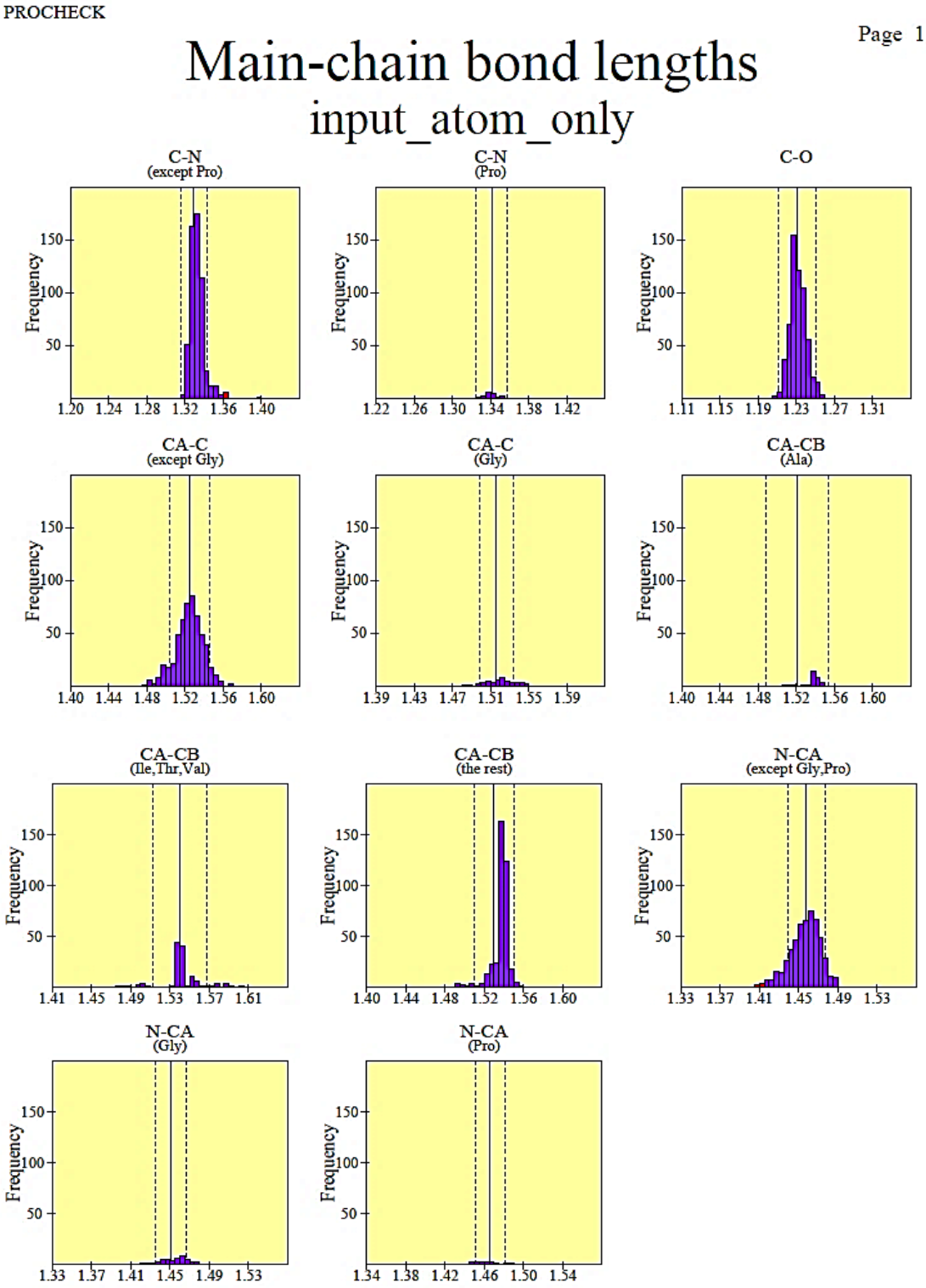
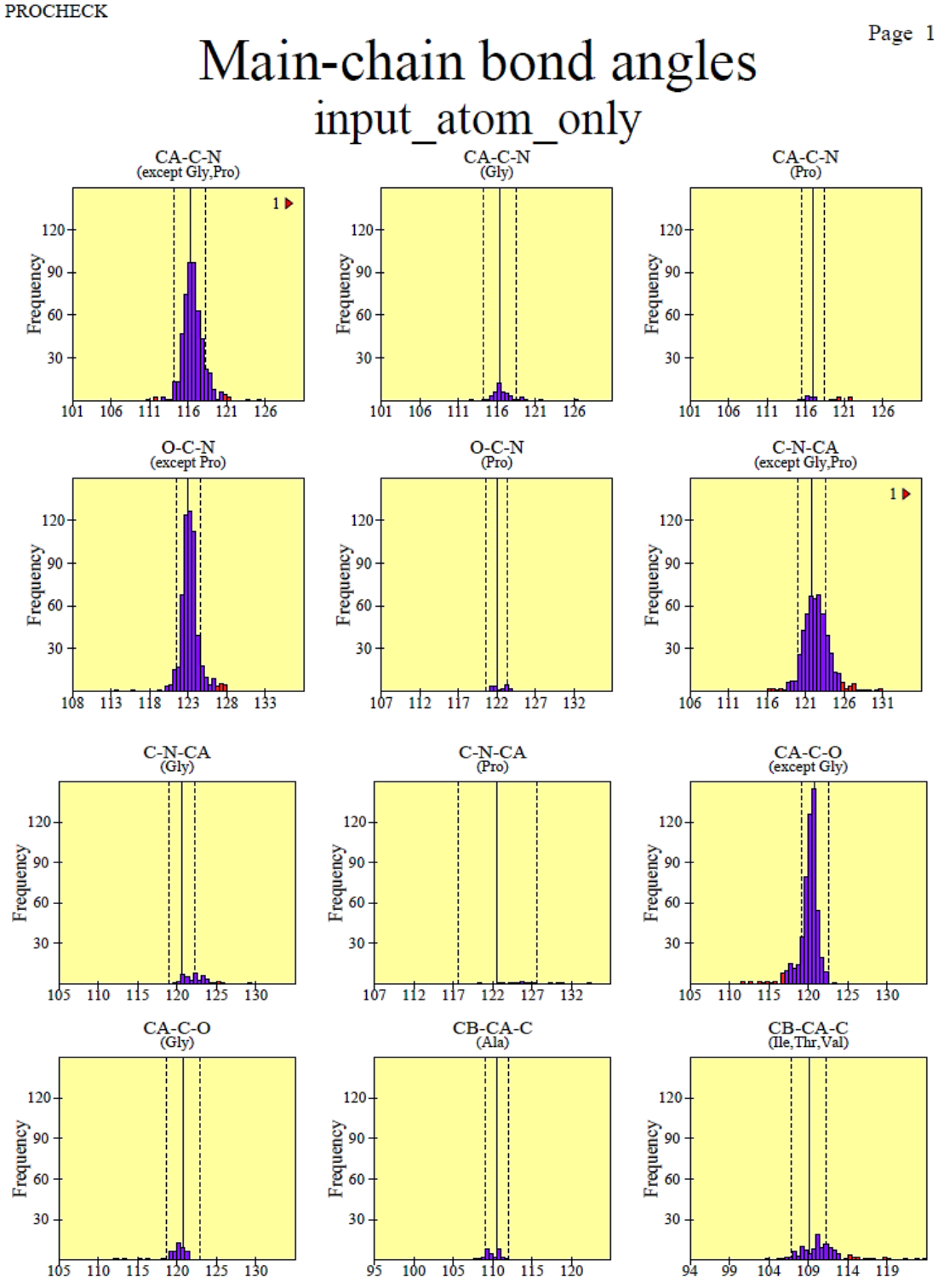
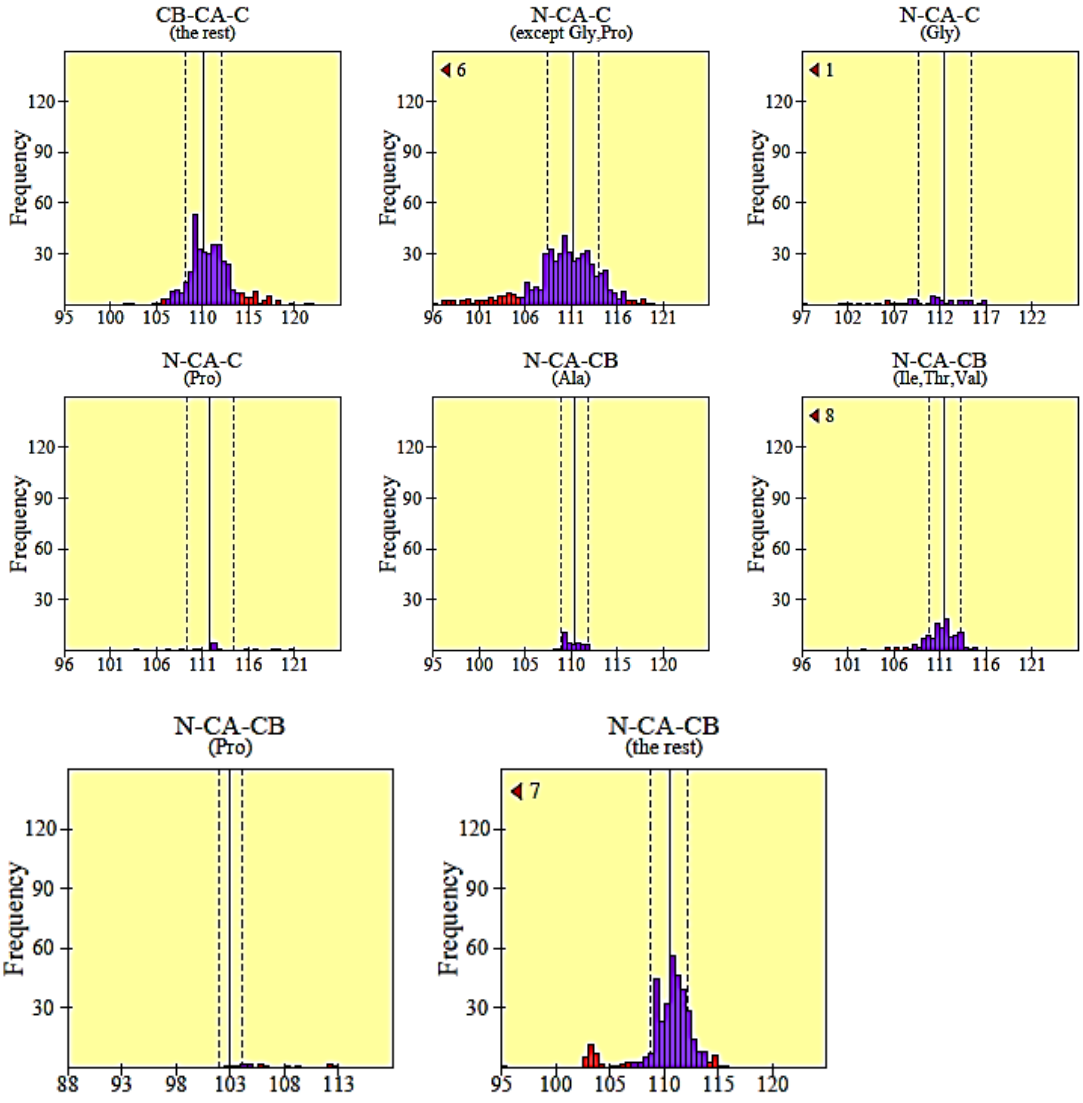


Figure C.5. Main chain bond angles of ArtB homology model



Main-chain bond angles

input_atom_only



Black bars > 2.0 st. devs. from mean. ◀ or ▶ signifies data points off the graph in the direction shown.
Solid and dashed lines represent the mean and standard deviation values as per Engh & Huber small-molecule data.

Figure C.6. Root mean square distances of planar atoms from the best-fit plane of ArtB model

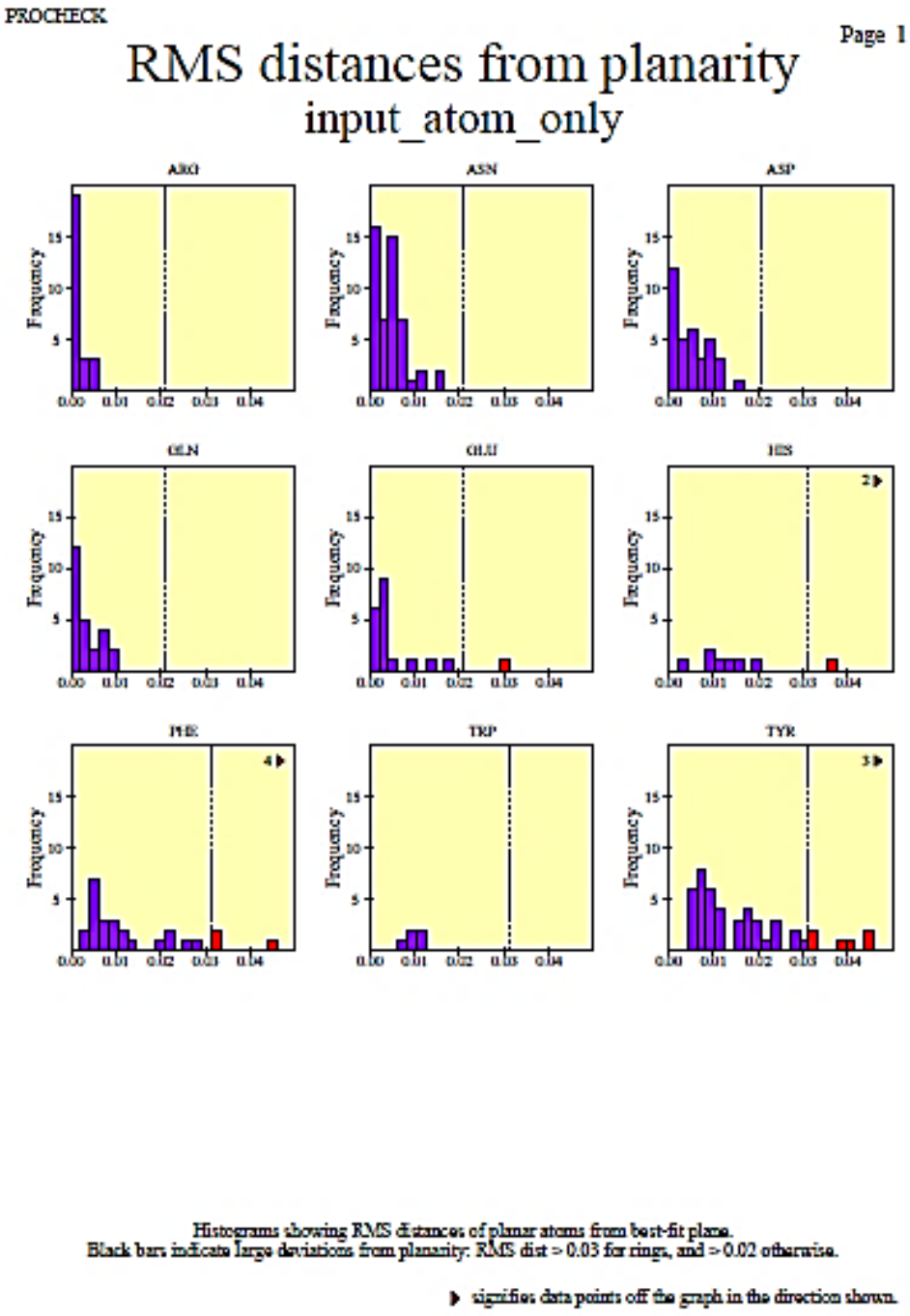


Figure C.7. Volume of the system ArtAB at constant pressure equilibration stage.

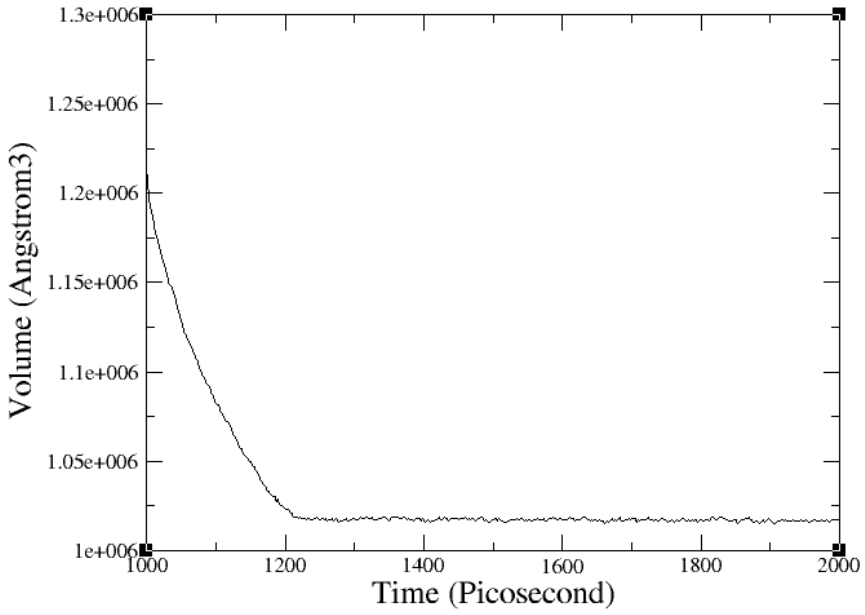


Figure C.8. Pressure of the system ArtAB at constant pressure equilibration stage.

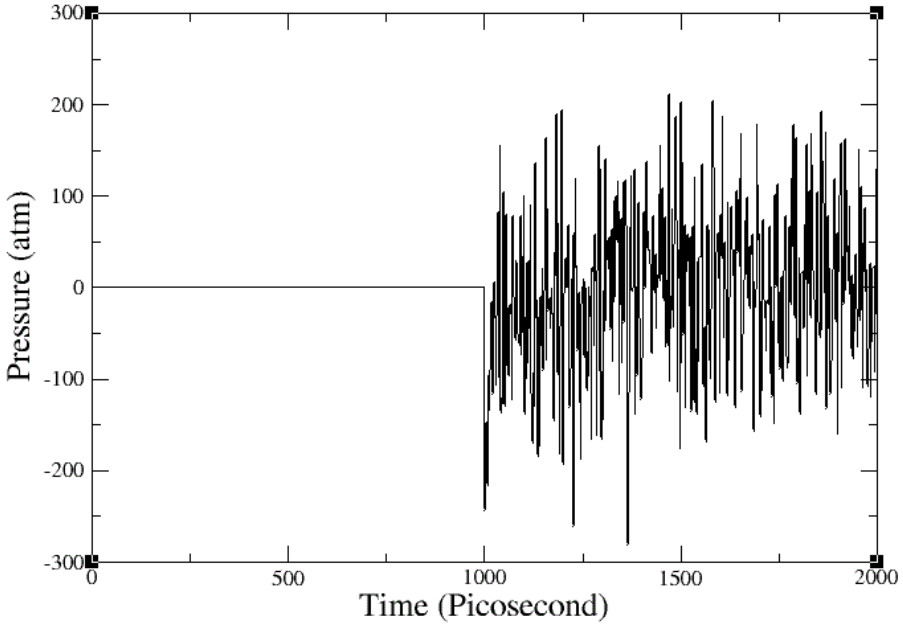
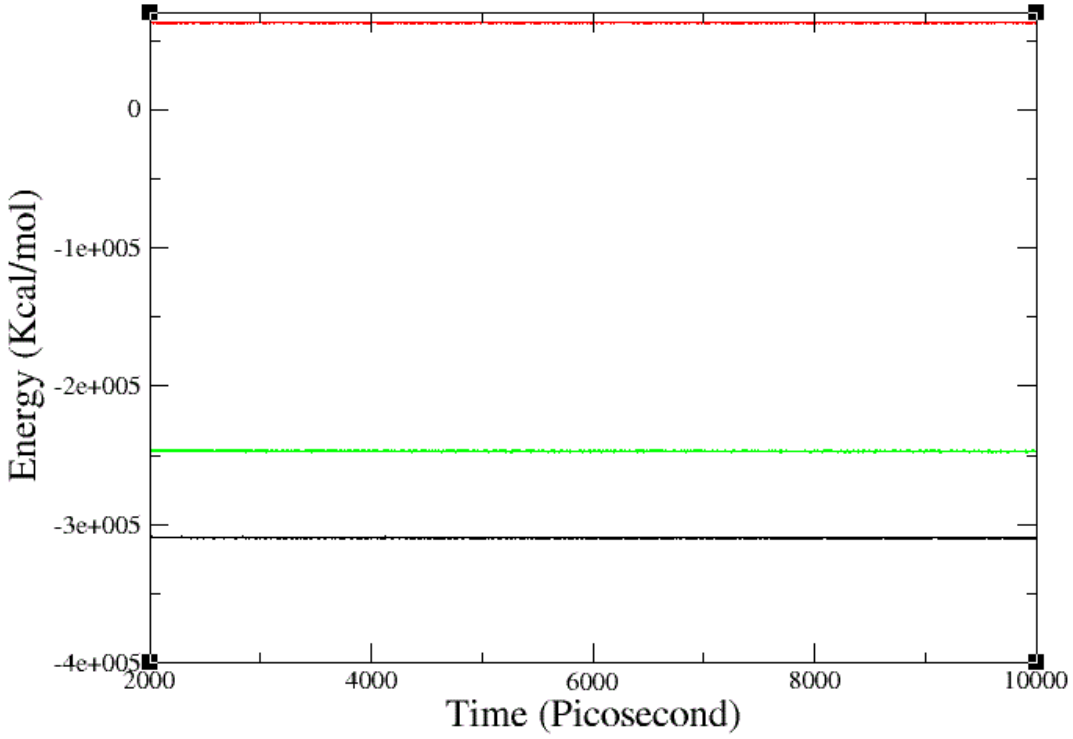


Figure C.9. Potential, Kinetic and Total energy of the system during 8ns production stage.



Black line represents potential energy, red line represents kinetic energy and green line represents total energy of the system.

Vita

Dipesh Budhathoki was born in Biratnagar, Nepal on September 30, 1988, to Dan Bahadur Budhathoki and Indu Devi Budhathoki. He completed his secondary education from St. Joseph Higher secondary School and high school from Birat Science Campus, Nepal. After graduating from high school, he joined Institute of Medicine, Nepal to continue his Bachelor of Pharmacy. He completed his Pharmacy degree in February 2014. He worked as a trainee in industry. He also has experience in academic sector as a lecturer of Pharmacy technician awarding college. In August 2015, he came to US to pursue his higher education and enrolled in Master of Pharmaceutical science at Idaho State University.