



Master Program In Life Science Informatics

MASTER THESIS

Large Scale Virtual High Throughput Screening (vHTS) on an Optical High Speed Testbed

Mohammad Shahid

December, 2007

1. Supervisor: Wolfgang Ziegler
2. Supervisor: Prof. Dr. Martin Hofmann-Apitius

Submitted to:

Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn,
and
Fraunhofer Institute for Algorithms & Scientific Computing (SCAI), Sankt Augustin, Germany.

Contents

1. ABSTRACT	1
2. MOTIVATION	2
2.1. CONTRIBUTION OF THESIS WORK	3
3. OVERVIEW	4
3.1. VIRTUAL SCREENING PIPELINE	4
3.2. WHY GRID.....	5
3.3. WHY MALARIA	5
3.4. DRUG RESISTANCE IN MALARIA:	6
3.5. PREVENTION OF DRUG RESISTANCE	7
3.6. LIFE CYCLE OF PLASMODIUM SPECIES, THE MALARIA-INDUCING AGENT	8
3.7. EXTENSION OF WISDOM	9
4. GRID COMPUTING	10
4.1. WHAT IS GRID	10
4.2. CHALLENGES IN GRID COMPUTING	11
4.3. THE GRID MIDDLEWARE.....	12
4.4. UNICORE.....	13
5. TARGET SELECTION AND PREPARATION.....	17
5.1. TRIOSEPHOSPHATE ISOMERASE (PFTIM)	18
5.2. ENOYL-ACYL CARRIER PROTEIN REDUCTASE (PFENR)	22
5.3. DOCKING SOFTWARE	25
5.4. PREPARATION OF TARGETS.....	26
6. RE-DOCKINGS, CROSS-DOCKINGS AND TEST RUNS.....	29
6.1. RE-DOCKING WITH FLEXX	29
6.2. CROSS DOCKINGS WITH FLEXX.....	33
6.3. TEST RUNS WITH FLEXX.....	34
6.4. RE-DOCKING WITH AUTODOCK.....	38
6.5. CROSS DOCKINGS WITH AUTODOCK	41
6.6. TEST RUNS WITH AUTODOCK	41
7. COMPOUND DATABASE DESIGN & PROJECT DEPLOYMENT	46
7.1. FILTERING THE DATABASE:	46
7.2. SIMILARITY SEARCHING.....	47
7.3. FINAL COMPOUND DATABASE	50
7.4. GRID DEPLOYMENT	52
7.5. OPTICAL TESTBED ENVIRONMENT	52
7.6. SOFTWARE APPLICATIONS USED IN THE PROJECT:.....	54
7.7. TEST RUNS.....	56
7.8. THE LARGE SCALE SCREENING.....	56
7.8.1. JOBS MANAGEMENT.....	57

7.9. PROBLEMS AND OPERATIONAL ISSUES.....	57
8. RESULTS AND DISCUSSION	59
8.1. RESULT ANALYSIS OF TARGETS OF PfTIM	60
8.2. RESULT ANALYSIS OF TARGETS OF PfENR.....	70
8.3. SCREENING RESULTS AND MOLECULAR SIMILARITY ANALYSIS:	77
8.4. DISCUSSION	79
8.5. SUMMARY AND CONCLUSIONS.....	80
9. REFERENCES.....	82

Approved:

Prof. Dr. Martin Hofmann-Apitius
1st Reviewer (Head of Examination Board, B-IT)

Prof. Dr. Jürgen Bajorath
2nd Reviewer

I herewith certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

Mohammad Shahid,
(Author)

Acknowledgments

I thank Prof. Dr. Martin Hofmann-Apitius for allowing me to do my thesis at Fraunhofer Institute, SCAI, and providing me continuous support and guidance as my thesis advisor. I would like to express my deepest gratitude to Wolfgang Ziegler, my thesis supervisor, for his guidance, technical support and help during the project. I am also very grateful to Marc Zimmermann for his valuable guidance and helpful suggestions during the entire work.

I would like to extend my sincere thanks to Antje Wolf, Vinod Kasam and Oliver Waeldrich, for their helpful discussions and co-operation during the course of this work.

M.Shahid.

1. Abstract

The over growing burden of antimalarial drug resistance places much emphasis on the immediate development of new antimalarial drugs. Malaria is a dreadful disease and the risk continues to the global health as the resistance develops more quickly than the development of new drugs. To combat Malaria, a large scale virtual high throughput screening activity was conducted using new potential drug targets. Triosphosphate isomerase and Enoyl-acyl carrier protein reductase were selected as novel drug targets. These enzymes play a very significant role in the life cycle of the malarial parasite. Compound libraries containing hundreds of thousands of small molecules were designed using similarity measures for the *in silico* virtual screening by molecular docking using two different docking algorithms; FlexX and AutoDock. To meet the computational demands and to speed up the discovery process, large scale virtual screening was deployed on a Grid infrastructure using the UNICORE middleware. The relevance of the Grid services in the drug discovery program is demonstrated in this project. New candidate compounds for the selected malarial targets were identified and recommended for further optimization and biological analysis.

2. Motivation

Malaria is a global health concern which kills over a million people each year and causes significant human sufferings in more than 100 countries. The control of this disease has been a major focus of attention of the World Health Organization and United Nations Children's Fund. The traditional antimalarial drugs available so far has lost their clinical effectiveness because of the over growing drug resistance of the parasite causing malaria. Resistance to Chloroquine, Mefloquine and Sulfadoxine has widespread over many parts of the world [1, 2]. This also produces a risk of rapid drug resistance for the new generation of currently effective chemotherapies and many other antimalarials.

This growing risk of drug resistance motivates the urgent development of additional new antimalarial drugs. The drug development is a time consuming and expensive process. Large chemical libraries are physically screened against a biological target (experimental high throughput screening HTS). To meet the challenges in combating against a disease, cost and time are two important factors. Since the early 1990's, several new technologies have been aided to the drug development process such as computational chemistry, advances in molecular biology and informatics tools in the life sciences. An approach known as Virtual Screening is now widely used in the modern drug discovery programs because of better hit rate of new ligand discovery than experimental high throughput screening [3] and to speed up this discovery process along with reducing its cost. Computational methodologies have become a crucial component of medicinal chemistry and many drug discovery programs [4, 5]. Use of such new computational strategies and informatics tools has gained vital importance as the underlying research has become increasingly data intensive [6]. Recently, there has been a continuous focus on the fast and cost-effective strategies to meet these demands.

In virtual screening, millions of chemical compounds are screened *in silico* against a biological target and selection and ranking for the best candidate drug is performed on the basis of their ability to interact with the target. In this way, the large compound library is filtered and reduced from a few millions to a few hundreds for the *in vitro* testing. But there are certain challenges faced by the virtual screening workflow, which includes the data challenge and the required computing power. Thus there is a need for an IT environment which facilitates intensive computing deployment, data storage and resource sharing. Grid computing is a technology that offers such an environment and which can meet these critical

needs. Utilizing the potential of the Grid has become indispensable when dealing with both the complexity of model and the huge quantity of data produced as a result [7]. The Grid is not only a computational paradigm that just provides computational resources, but an infrastructure that also supports resource sharing and collaboration to speed up the computations. The motivation of this project is the importance of drug discovery for this disease [8] in a significantly reduced computation time by distributing data to process over the Grid computers [9]. The major goals of this research work include:

- to implement a target-focused virtual screening workflow by utilizing the potential of the available Grid technology,
- to find new potential anti-malarial drug candidates by evaluating a set of ligand libraries against the advanced malarial targets,
- to extend the WISDOM [10] project by focusing on the problems noticed during the first WISDOM challenge against malaria and,
- to deploy a large scale *in silico* virtual high throughput screening (vHTS) workflow and test the newly established, high bandwidth optical Grid environment VIOLA [11] for advanced bioinformatics applications using UNICORE [29].

2.1. Contribution of Thesis Work

The contribution of the thesis work in this large scale virtual high throughput screening project includes the selection of new and validated malarial drug targets, efficient designing of a compound library for the screening workflow by performing a ligand based similarity searching using different methods, deployment of the software applications and the compound database on to the Grid environment, execution and monitoring of the software using the Grid middleware and analysing the results of this large scale virtual screening calculations.

3. Overview

Research in the field of drug discovery has been aided to a great extent by the use of computations in the last two decades. The reason behind this is the massive increase in information that has become available after analysing the biochemical and physiological systems of the living organisms. The “-omics” termed fields of science have put a deep impact on the research in drug discovery which has led to a progress of medicines. The computational tools combined with these sciences allow us to reach the genetic basis of diseases and explore the new points of attacks for future medicines [12]. New strategies in bioinformatics have been adopted which enable the researchers to predict the *in vitro* activity of candidate drug molecules. Rational drug design using computational modelling to identify these small molecules has a big complementary role in the drug discovery process. Research on academic level based on the *in silico* chemical toolkit is much related to that carried in a pharmaceutical industry. But there exist certain key challenging issues for instance, time and cost among many others, lying in the way of this research. The current ongoing project is about implementation of a large-scale virtual screening workflow using new and cost-effective strategies.

3.1. Virtual Screening Pipeline

The goal of virtual screening is to find possible clinical drug candidates by evaluating large molecular libraries. These libraries are either from known compounds or constructed in a combinatorial way, and are computationally screened for the compounds, which bind to a known drug target. Although, there are chances of false-positive and false-negative predictions in the process of virtual screening, but it offers a practical way with an increased hit finding rate than the experimental screening [3, 13 and ref. therein]. Screening by molecular docking is a routine practice when the structure of the drug target is known. When the structure is not known, the structure can be predicted by homology modelling. The resulting top ranking compounds are evaluated by various docking programs based on their scoring functions and best fitting in the receptors binding position, and submitted further for experimental testing. However, there are some limitations of the scoring schemes used in these applications. It is suggested that efforts should be invested in testing alternative applications and produce a consensus scoring and improve the probability of identifying the true hits [4]. There is another important factor regarding screening of large compound libraries, which is one of the major goals of this research project. The compound libraries

exist in very large databases containing millions of compounds. Screening hundreds of thousands of compounds in these libraries requires significant computational resources. A variety of application software used in the screening programs has variable memory and computing power requirements. One can easily account for the processing time that screening one million molecules will take. Thus, there is need for a powerful environment, which facilitates these complexities by providing collaborative sharing of computing, data and storage. Grid computing is such a new technology that can provide such resources, which will be discussed in more detail under the section “Grid Computing”.

3.2. Why Grid

The central theme of the Grid computing is sharing resources. In high-throughput computing, the Grid is mainly used for scheduling large numbers of independent tasks or sub-tasks of a main task in a distributed computing manner, with the goal of utilizing the unused processors [14], thus focusing of available resources on a single problem. Today the biological applications as like other many scientific applications have become increasingly data-intensive, with the focus of extracting and aggregating new information, the data is maintained and distributed in geographically widespread repositories and databases. This geographically distributed data across heterogeneous networks put a great demand for high-speed broadband networks, which should allow the data, applications and hardware access from everywhere. A huge quantity of data stored in a variety of databases, is dealt with in the process of vHTS. Similarly, there are many software applications available today in the field of drug discovery, which demand for more computational power due to the limitations of the algorithms. It has therefore become necessary to switch to the Grid environment when dealing with such high demanding issues.

3.3. Why Malaria

The focus of the current large-scale virtual screen project moves around combating malaria. This disease has put about 40% of the world’s population into the risk of its infection. More than 500 million people worldwide become severely infected with malaria every year. Malaria is widely distributed around the world including Asia, Latin America, Middle East and parts of Europe, where as in sub-Saharan Africa it is more dreadful [15]. Over one million people die each year due to this disease, of which most of them are young children or pregnant women [16].

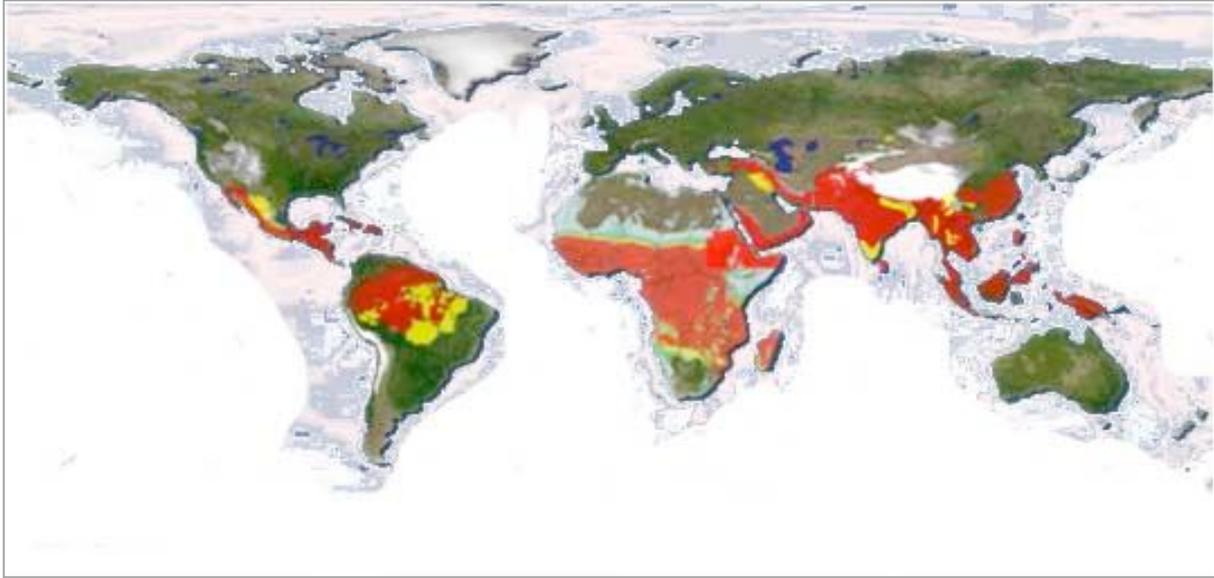


Figure 1. World wide Geographical distribution of Malaria highlighting the affected areas [taken from 17].

3.4. Drug resistance in malaria:

The rapid spread of resistance against antimalarial drugs over the past few decades has necessitated the increased monitoring for further drug resistance and malaria control efforts. Chloroquine, for example, which has been a major antimalarial treatment for decades, the resistant strains has acquired mutations in Pfct gene and thus has become less effective against the resistant strains. Similarly, sulfadoxine-pyrimethamine has also lost its effectiveness in many parts of the world because of several point mutations in the DHFR and DHPS genes which has also become the main cause of resistance to antifolates. This is leading to a much higher degree of pressure on the use of currently effective artemisinin-based combination therapies (ACT's), which, as a result, will also increase the likelihood of parasite resistance genotypes. The ACTs could lose their potency because they are the only effective antimalarial drugs available today [2, 18].

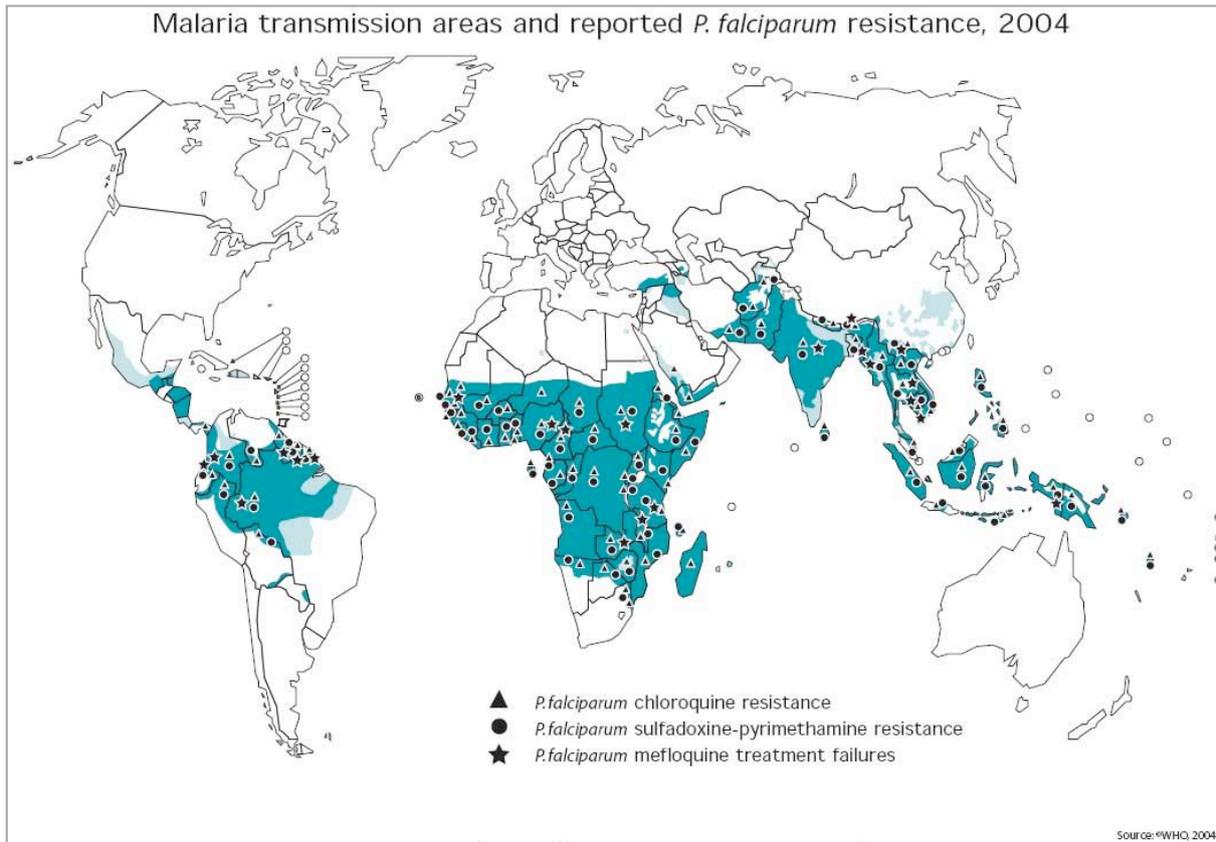


Figure 2. Reported *P. falciparum* resistance to various antimalarials worldwide [taken from 15].

3.5. Prevention of drug resistance

New antimalarial drugs will continuously be needed as a consequence of drug resistance. As long as the drugs are used for the eradication of malaria, the chances of resistance to those drugs will be present. *Plasmodium falciparum* has developed resistance to nearly all available antimalarial drugs and there is a high risk that the parasite will become resistant to any drug that is used widely. The other species may also develop resistance to the drugs in use today and may result in, for instance, resurgence of vivax malaria [19]. It has been observed that the resistance develops more quickly than the development of new antimalarial drugs. Therefore, there is a great need to speed up the development process for new antimalarial drugs along with strict monitoring on the use of current antimalarials. Information on the mechanism of drug resistance comes through the way of the ongoing genetic and biochemical research. The initiatives of drug discovery are used to find new leads that have good efficiency against the resistant malaria [20].

3.7. Extension of WISDOM

WISDOM, an international initiative for Grid-enabled drug discovery against neglected and emerging diseases, started in 2005. The initial initiative has been focussing its effort on virtual screening by molecular docking against targets in malaria. Its first large-scale deployment allowed 42 million dockings in about 6 weeks, using thousands of CPUs through the Grid infrastructure worldwide [10]. WISDOM has its own production environment using the Grid middleware on the EGEE Grid [21]. The software environment deployed on the EGEE Grid allows the submission and monitoring of jobs. Our current large-scale virtual screening program was deployed on VIOLA [11], a high-speed optical testbed, using the UNICORE middleware. This will allow us to compare both execution environments and cover the issues relating to deployment, network, security and improvement in the job submission and monitoring services.

4. Grid Computing

Grid computing is still considered a relatively new concept for distributed computing and the research under this area is still growing. The research communities are required to agree on standards and need to have efficient resource management to assess the feasibility and its usage. A computing Grid is a collection of computing resources that are available for an application to perform the tasks for that application. In fact, the term “Grid” is analogous to the electric power Grid, which provides a seamless and easy access to electrical power by sockets regardless of their source and generation [14]. The computing Grid provides a way to access geographically widespread computing centres and their resources, data and software. These computing resources are generally heterogeneous and being geographically dispersed, usually are parts of different administrative domains. Such heterogeneous resources are integrated into the Grid environment, which allows equal accessibility to them through a standard interface. The Grid is not only a computing paradigm for just providing computational resources; it is an infrastructure that can unify globally remote and diverse resources to achieve organizational and scientific objectives in a timely manner.

4.1. What is Grid

Different authors have proposed various definitions in some directions. Ian Foster and Carl Kesselman defined the Grid in 1998 as “*A computational Grid is a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities*” [22]. In another definition of the Grid, “*The Grid is the computing and data management infrastructure that will provide the electronic underpinning for a global society in business, government, research, science and entertainment*” [23] the importance of Grid computing in the real world scenarios is highlighted. As already mentioned, the Grid consists of resources which are held in dynamic, geographically dispersed heterogeneous environment, which is clearly stated in this definition: “*Grid computing is concerned with coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organisations*” [24]. Later on, Ian Foster further suggested a three point checklist for the Grid system in [22] that a Grid is a system that “*coordinates resources that are not subject to centralized control*”, “*using standard, open, general purpose protocols and interfaces*” and “*delivers nontrivial quality of service*”.

Today, powerful computers are easily available and the internet has become more popular with the availability of low-cost high-speed network components. This has resulted in exploiting the networks of computers and the underutilized resources as a single computing resource. It is now fairly possible to cluster or couple together computing resources including supercomputers by using devices distributed geographically, thus forming a “computational Grid”. Such a Grid not only provides a unified computing resource but also provides storage systems to facilitate storage of large data sets. The Grid infrastructure is becoming a flexible, secure coordination of resource sharing among dynamic collections of individuals or institutions that are referred to as *virtual organisations*. Anyone in the virtual organization can potentially get the power equivalent to that of a supercomputer from the pooling of computing resources.

There are a number of Grid categories or types and there exist no hard boundaries between these categories which are not discussed in this document. They are organized into different systems to meet specific needs. So, our focus will remain on using the computational Grid system with emphasis on virtual screening in the process of drug discovery.

4.2. Challenges in Grid Computing

The concept of Grid continues to evolve since its transition from the previous approaches such as MetaComputing, Scalable Computing, Global Computing or Internet Computing about a decade ago [25]. The paradigm of such MetaComputing was the cooperative use of geographically distributed resources unified to act as a single powerful platform for the execution of a range of parallel and distributed applications. The increase in demand for computation, data storage and processing makes the Grid concept more attractive for scientists and users. But there exist some operational challenges in the field of Grid computing which the experts are still struggling to make it more reliable and easy to use. The different network architecture, firewall and security systems of numerous institutions (virtual organizations) create complexities hindering them from opening their computing resources to the outside world [26].

Grid computing will soon reach a tipping point when the benefits will outweigh the difficulties. Everyone will gain when the technical hurdles coming in the way of Grid system are overcome. Developers of Grid systems are working their way hard to abstract from the heterogeneity of the local Grid infrastructures, the underlying networks, and the different

security requirements and standards. Communities are required to agree on standards to assess the feasibility and usage. The Global Grid Forum, now Open Grid Forum [27] continues the efforts to facilitate standards development and Grid usage.

Despite these troubling issues, the Grid system adoption is becoming more and more widespread. Some of the issues regarding the Grid user interface and the middleware which should reduce the need of an IT enthusiast, in the course of this virtual screening program will be discussed in the section “Project Deployment on the Grid”.

4.3. The Grid Middleware

The layer of software used to create common interfaces both towards the local (computing) infrastructure and towards the user to utilize the Grid system is known as the Grid middleware. These interfaces are a prerequisite to make ease the access to and utilization of the resources available in the heterogeneous Grid system. A Grid middleware system is a set of components that can be used as part of a Grid environment. These components are the building blocks on which it is possible to create the environment where the software applications can run and utilize all the resources available over the Grid. Without the middleware system it would not be possible to utilize all the heterogeneous resources of the Grid environment. For example, every computing element comprise of a different interface would require a login on all machines for a Grid user. It is the middleware, which makes the recourses connected in to the Grid equally accessible through its user interface.

The middleware layer in Grid architecture offers services to reduce the complexities and the heterogeneity of a Grid arising from network systems and resource management systems.

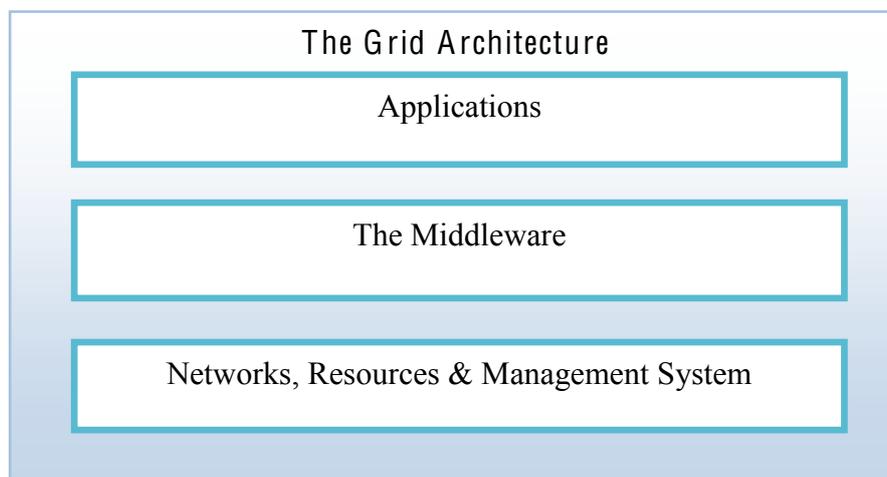


Figure 4. General Architecture of a Computer Grid.

The Globus Toolkit [28] is a set of software tools and components designed to create a Grid environment, founded by the Globus Project, in an effort to develop a software system to enable worldwide scientific cooperation. Globus, and similar systems such as UNICORE [29], gLite [30], CSAR [31], Condor [32] and MOAB Grid Suit [133], all work in a similar way.

The WISDOM project [10], deployed on the production environment of the EGEE, was based on gLite/LCG Grid middleware services. Our large-scale virtual high throughput screening was carried out using the UNICORE system, which is described in the following section.

4.4. UNICORE

The UNICORE (Uniform Interface to Computing Resources) is an open source Grid middleware started back in 1997 initially funded by the BMBF (Federal Ministry of Education and Research, Germany) and later by the European Union through projects like EUROGRID [34] and UniGrids [35]. UNICORE provides a seamless, secure and intuitive access to distributed resources and data by providing components on all level of the Grid architecture. The ideas behind the UNICORE development are to support the users by hiding the system complexities by helping to develop distributed applications [36]. Major considerations in the objectives of the UNICORE development include:

- allowing end users to transparent and seamless access the Grid resources without the need of knowledge of the complex Grid environments, applications, hardware and security issues.
- implementation of X.509 security specifications to handle security issues like authenticating users and servers by promising secure communication over the networks.

4.4.1. UNICORE Functionalities

A rich set of functionalities are provided by the UNICORE interface to the end-users which enable them to manage and execute their jobs on different sites (virtual organization) in the Grid, thus preventing them from recurring security checks. Ideally, a researcher sits down at his/her computer and logs into the virtual organization. The researcher can also access databases, software applications and the scientific instruments shared around the globe.

Making use of the large combined computing power of the collaboration, the user requests a computing job through the easy to use user interface provided. UNICORE takes care of the job on the Grid to find the data necessary for that job and its memory and processing requirements, and present the results back to the user after its completion. Some of the key functionalities provided by UNICORE are given below:

- UNICORE provides seamless execution and management for the end users jobs. A user job or task contains a sequence of steps defined by that job in the form of an Abstract Job Object (AJO) in the UNICORE environment. User can create and manage complex batch jobs that can be executed on different systems at different sites. Jobs can be interdependent on other jobs, hence a user can define complex job objects.
- Job submission interface allow users an easy way to import data to the computing sites required for a job. Therefore, UNICORE provides a transparent exchange of data between different computing sites.
- UNICORE provides support for Message Passing Interface (MPI), through which a user's application can make use of two or more computing nodes simultaneously.
- Security in the Grid system is handled in UNICORE by using standard X.509v3 certificates for authentication of both users and software components. The certificates are issued by a trusted Certificate Authority (CA). With this mechanism, a user just need to login once, the further authentication is taken care by the UNICORE system providing single sign on functionality. All the communication across the networks is SSL based.

4.4.2. UNICORE Architecture

The UNICORE architecture is based on three layer client/server model. It consists of a client that executes on the user's workstation, a gateway and multiple instances of Network Job Supervisors (NJS) that execute on dedicated securely configured servers, and multiple instances of Target System Interfaces (TSI) that run on the target system and interface with their operating system and the batch subsystems (BSS) [37] as illustrated in the Figure 5. Three layered Architecture of the UNICORE below.

Client: The UNICORE client is a java based client with advanced features. Users' security certificates should be provided to the client with the help of which the computing sites can be accessed. The AJO (Abstract Job Object) is constructed from the user job in the Job Preparation Agent (JPA) part of the client. After the user job is prepared and submitted to a site, users are able to view the status of their jobs with the help of the Job Monitor Controller (JMC) part.

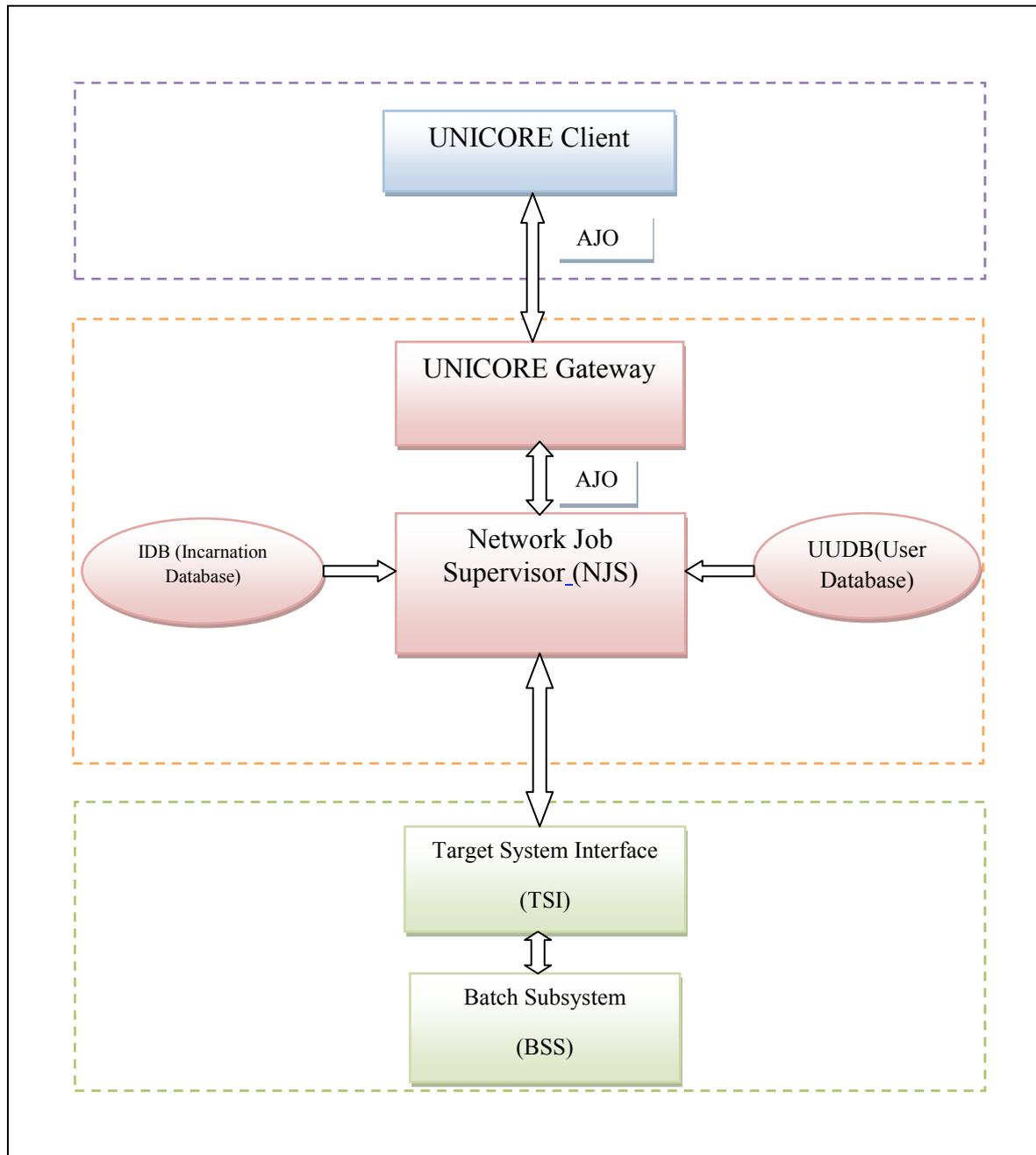


Figure 5. Three layered Architecture of the UNICORE middleware system.

Gateway: The gateway provides a single entry point for the second layer i.e a USite (UNICORE Site) and is the point of user authentication that identifies a client connection coming from a user. A UNICORE VSite is established from two components: the Network Job Supervisor (NJS) and a Target System Interface (TSI). The UNICORE NJS Server is responsible for managing all submitted jobs. It performs the user authorization by looking for a mapping of the user certificate to a valid login in the UNICORE User Data Base (UUDB). The NJS incarnates jobs from the abstract AJO definition and hands the incarnated tasks and jobs over to the TSI. The incarnation is based on the system specific information in the Incarnation Data Base (IDB). It is the NJS' task to consider the dependencies between job components and to schedule the components accordingly. The NJS server stores all job status and result information, and replies to status and result requests from the client.

Target System Interface: The Target System Interface accepts incarnated job components from the NJS, and passes them to the local batch systems for execution. In addition, file import and export tasks are handled by the TSI, and it also implements low-level status reporting and control of batch jobs.

5. Target Selection and Preparation

In humans, malaria is caused by four species of plasmodia of which *Plasmodium falciparum* is the most virulent species. Search for new antimalarials continues with full strength because of the spread of the disease with *Plasmodium falciparum* infection and the increase in drug resistance. Identification of the drug target is one of the most important steps in antimalarial drug development. To understand the mechanism of host parasite interactions for the identification of drug targets, 3D modelling of *P. falciparum* annotated proteins has been carried out, in which 476 *Plasmodium* proteins with one or more known structural templates ($\geq 40\%$ identity) have been identified [40]. There are about 200 proteins 3D structural data available from various *Plasmodium* species in the Protein Data Bank (PDB) [41].

To make new drugs available as quickly as possible, focus should be made on the most promising drug targets. After identification of a protein as a possible drug target, additional information is needed for its verification, i.e a good drug target must have an essential role in the parasite's life cycle or replication that is called a "validated drug target" [42].

Proteins involved in the energy metabolism play a very significant role in the life cycle of the parasite. Glucose metabolism is an essential pathway in erythrocytic and intraerythrocytic stages of the plasmodia. In the intraerythrocytic phase of the parasite's life cycle, glycolysis is estimated to be the predominant pathway for ATP production, making glycolytic enzymes an attractive target for inhibitory intervention. *Plasmodium falciparum* Triosephosphate isomerase (PFTIM) is a significant enzyme in the glycolytic pathway. Over the past three decades, this enzyme has been the subject of extensive structural and mechanistic investigations [44]. The focus on glycolytic enzymes in the malarial parasite results from the observation that in the asexual stage of the parasite in the human erythrocytes, the energy requirements of the organism are almost exclusively met by glycolysis [45]. Proteins involved in the lipids biosynthesis also play an essential role to meet the metabolic demands of the parasite and make this pathway an attractive drug target. *Plasmodium falciparum* Enoyl-acyl-carrier Protein Reductase (PfENR) is the key protein for the completion of fatty acid biosynthesis. Following is a brief description of these two targets involved in the energy metabolism and fatty acid biosynthesis.

5.1. Triosephosphate isomerase (PftIM)

PftIM catalyzes the interconversion of dihydroxyacetone phosphate and glyceraldehydes-3-phosphate. There are eight α -helices alternating with eight β -strands along the polypeptide chain in the structure of PftIM.

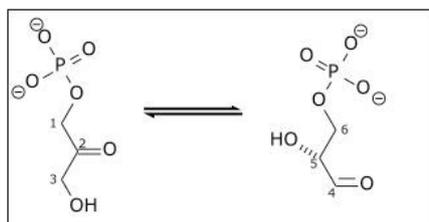


Figure 6. Interconversion of dihydroxyacetone phosphate and glyceraldehydes-3-phosphate.

The parallel β -strands are hydrogen-bonded to each other and form a central, solvent-excluded, eight-stranded β -barrel. The eight helices wrap around the central β -barrel. The strands and helices are arranged approximately antiparallel to each other [44]. The conserved residues like glutamate (E165), histidine (H95) and lysine (K12) have specific roles and many interesting differences from those of the other reported TIMs have been revealed by cloning and sequencing of the TIM gene from the malarial parasite *P. falciparum*. Among these differences, serine 96 (S96) is a notable one: The completely conserved residue in all TIM sequences reported so far is replaced by a phenylalanine (F) residue [45, 53]. The structures of several PftIM inhibitor complexes have been determined. The high-resolution structure of the PftIM-2-phosphoglycerate complex has been reported at 1.1 Å. A noticeable feature of the PftIM inhibitor complexes is the tendency of loop 6 (residues 166–176) to remain in the open conformation, even when the active site is occupied by a ligand. This is in contrast to the structures of TIMs from other sources, where the loop open forms are observed for unligated enzymes and loop closed forms are observed upon ligand binding [44].

Triosephosphate isomerase is a dimeric enzyme which contains disordered, probably mobile loops of nine amino acid residues near the active sites. However, these loops fold down and cover the sites once the substrate is bound, which suggests a kinetic and mechanistic importance of the loop flexibility.

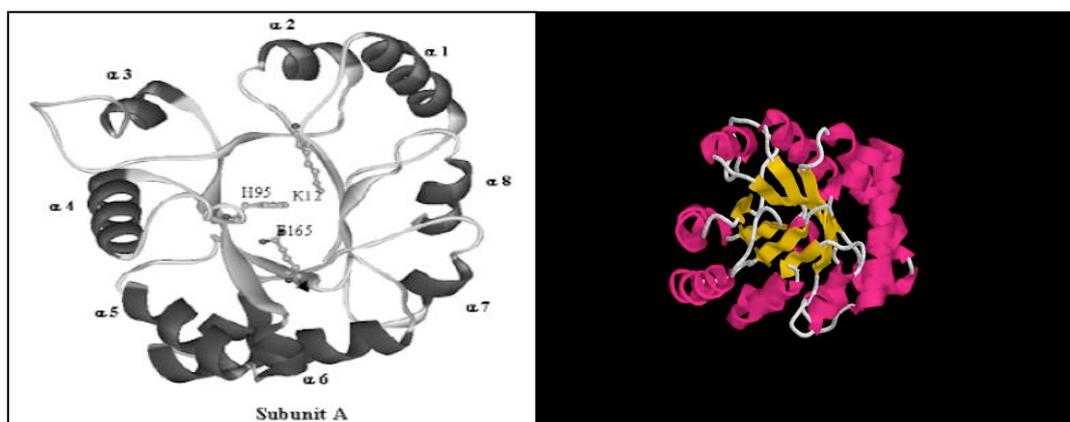


Figure 7. Structure of PfTIM subunit A with active site residues shown (left) and the eight helices wrap around the central β -barrel (right).

To consider an enzyme as a potential drug target, it is essential to investigate the key structural differences between the host and parasite enzymes, which can then be used to develop selective inhibitors. Comparing the structure of PfTIM in particular with that of human TIM has revealed several differences. In PfTIM, the residue at position 183 (a residue close to the active site) is leucine whereas in most other TIMs (also in human TIM) this residue is a glutamate. This leucine residue is completely exposed and together with the surrounding positively charged patch, may be responsible for binding TIM to the erythrocyte membrane. The occurrence of a cysteine residue (Cys13) at the dimer interface of PfTIM is another interesting feature, in contrast to human TIM where this residue is a methionine. Finally, residue 96 of human TIM (Ser96), which occurs near the active site, has been replaced by phenylalanine in PfTIM. Although the human and Plasmodium enzymes share 42% amino acid sequence identity, several key differences exist which can be exploited very well while designing specific inhibitors for PfTIM. Most significantly a strongly conserved surface residue (Glu183) is replaced by a hydrophobic residue (Leu183) in the parasite enzyme, which suggests that this hydrophobic residue may play a role in membrane attachment within the erythrocyte. This can be used as an alternative strategy for interference with the normal physiology of the parasite by specific inhibition of the enzyme attachment with the erythrocyte membrane. [47].

For the reason of comparison, a crystal structure of human TIM was also selected and included in the experiment.

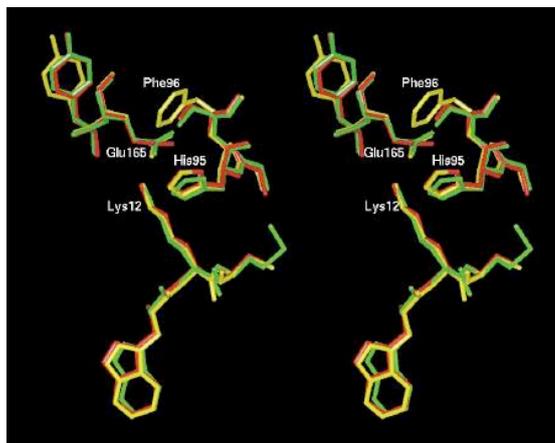


Figure 8. Superposition of the active site residues from human TIM, trypanosomal TIM and PfTIM shown in red, green and yellow respectively. The active-site residues, Glu165, His95 and Lys12, are labeled. Phe96 of PfTIM, which is a serine residue in the human and trypanosomal TIMs, is close to the active-site residues Glu165 and His95 and is also labelled [taken from 47].

5.1.1. Selection of Crystal Structures from PDB

The following PfTIM crystal structures were selected for the docking experiment:

The target structures PDB code 1LYX, at 1.90 Å resolution, having the enzyme catalytic loop in “closed” conformation, PDB code 1LZO, at 2.80 Å resolution which has the “open” conformation of the catalytic loop and PDB code 1O5X at 1.10 Å resolution in which the intact beta-subunit bound to the ligand possess the catalytic loop in both open and closed conformations, although the open conformation of the loop observed is reported to be dominant [46]. The human TIM; PDB code 1HTI at 2.80 Å resolution [48] was also selected for comparison.

3D structural super-positioning of the four TIM structures was performed using Molecular Operating Environment (MOE) package [49] which shows the catalytic loop closed in 1LYX and opened in the rest of the structures. Table 1 shows STAMP (multiple protein sequence alignment from tertiary structure comparison) results [50]. The visualization of the 3D structural superposition is similar to that presented in Figure 9.

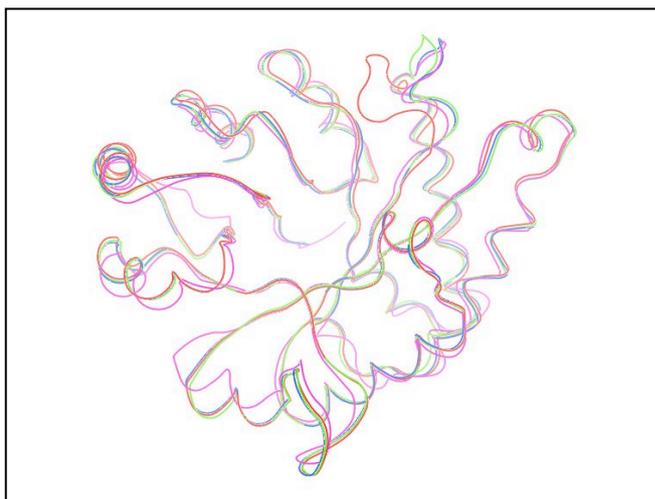


Figure 9. Super-positioning of the three PfTIM structures 1LYX (red); loop closed, 1LZO (green); loop opened and 1O5X (blue) with that of human TIM (pink).

Table 1. STAMP Results for the target structures of PfTIM.

PDB ID	Chain ID	Superimposes	Sequence Identity (%)	Stamp Score (Max 10)	RMSD (Å)
1LYX	A	Fixed Mol.	100	10.00	-
1LZO	A,B	Rotatable	100	9.44	0.658
1O5X	A,B	Rotatable	100	9.44	0.564
1HTI	A,B	Rotatable	41.53	8.76	1.147

The individual targets comparisons of the catalytic loop 6 dynamics of the single subunit of the structure 1O5X with 1LYX and 1LZO, respectively, are shown in Figure 10.

5.1.2. Characteristics of the Active Site of PFTIM:

The catalytic reaction carried on in the active site of this enzyme is achieved by simple intermolecular protonation and deprotonation. This proton transfer occurs with the help of certain elements as catalytic tools, the 10-residues loop described above and a catalytic acid (His-95) and base (Glu-165) [46]. Among the interactions involved between the ligand and protein, the most important contact is the hydrogen bonding of the flexible loop residue Gly-171 with O2P of the substrate. Certain van der Waals connections of the ligand are also made

with the residues of the catalytic flexible loop. Many other hydrogen bonds of the active site residues have been reported for Gly210, Ser211, Val212 and Leu230, Val231, Gly232, Asn233, Ala234 and which play important role in positioning and holding the ligand in the active site. Lys12, Ser211, Gly232, and Asn233 make direct interactions with the phosphate oxygens of the ligand [45].

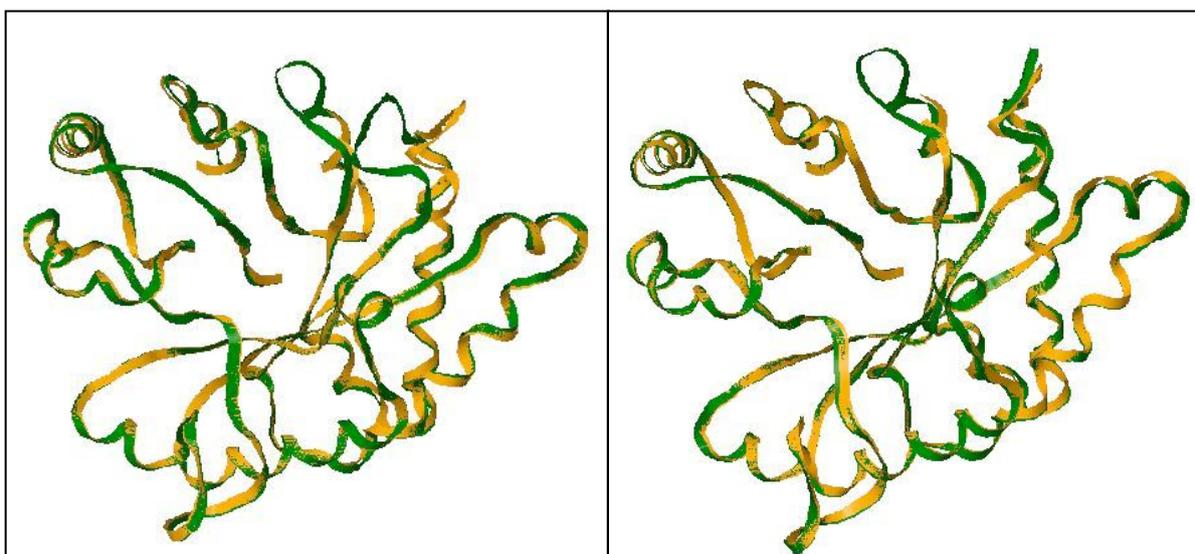


Figure 10. (Left): 3D structural superpositioning of 1O5X (yellow) with 1LYX (green). (Right): 3D structural super-positioning of 1O5X (yellow) with 1LZO (green).

5.2. Enoyl-acyl Carrier Protein Reductase (PfENR)

This enzyme has important role in the fatty acid biosynthesis pathway of the parasite. This pathway is a dissociative type of pathway known as type II FAS (Fatty acid synthesis) where each reaction is carried out by a separate enzyme. PfENR has a key regulatory function in the FAS II pathway. It catalyzes the final step in the fatty acid chain elongation and also responsible for limiting the rate of elongation. The reduction process is NADH/NADPH dependent and converts trans2-enoyl-ACP to acyl-ACP [55].

The lipids requirement is very high in the malaria parasite and in addition to growth regulation lipids are the integral part of membranes and play important role in differentiation and apoptosis [43].

PfENR is one the most promising structure-based antimalarial targets. PfENR is localized to the apicoplast, an organelle bounded by four membranes.

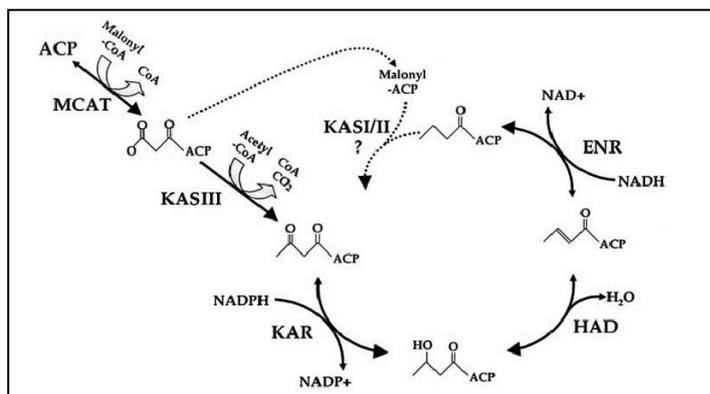


Figure 11. Role of ENR in the Fatty acid biosynthesis pathway of *Plasmodium falciparum* [taken from 42].

The apicoplast is unique to the apicomplexan parasites and has been suggested to be a rich source of potential drug targets. The difficulty inherent in delivering drugs to a target bounded by a total of seven membranes (including the parasite outer membrane, the parasitophorous vacuole membrane and the erythrocyte membrane) may well be shown by the studies with triclosan analogues. Although triclosan is more effective in inhibiting PfENR in vitro by an order of magnitude or more, these analogues are roughly equivalent to triclosan in killing parasites, indicating that they may have better penetration into the apicoplast [56].

The enzyme involved in type I and type II fatty acid synthesis pathways have major structural differences. The Type II synthesis pathway is lacking in human as it is common to plants and majority of prokaryotes, which makes this enzyme an excellent drug target. The closest analogue to this enzyme is 2,4-Dienoyl CoA reductase (DECR), PDB code (1W6U) in humans which has 19% sequence identity. Running a protein-protein BLAST [76] for the protein sequence of PfENR against the human database only, resulted in 2,4-Dienoyl CoA reductase (DECR) 1 precursor which has a very low alignment score (score < 40). The PfENR (EC no. 1.3.1.9) and the human mitochondrial DECR (EC no. 1.3.1.34) belong to the same protein family and fold (Rossmann fold) but constitute different domains. The 3D structural alignment is shown in Figure 12.

5.2.1. Selection of Crystal Structures from PDB:

The X-ray crystal structures of Triclosan bound to PfENR in the presence of NAD⁺ provides a structural basis for PfENR inhibition [57]. The following crystal structures were selected from the PDB for the docking experiment:

The crystal structure of PfENR PDB code 1NHG at 2.43 Å resolution has a complex bound to the inhibitor Triclosan (5-chloro-2-(2,4-dichlorophenoxy)phenol) and the cofactor NAD. And the crystal structure 1NNU at 2.50 Å resolution bound to a Triclosan analog (6-(4-Chloro-2-Hydroxy-Phenoxy)-Naphthalen - 2-OL) and the cofactor NAD.

Table 2. STAMP Results for the target structures of PfENR.

PDB ID	Chain ID	Superimposes	Sequence Identity(%)	Stamp Score (Max 10)	RMSD (Å)
1NHG	A	Fixed	100	10.00	-
1NNU	A	Rotatable	100	9.77	0.211
1W6U	A	Rotatable	19.48	5.87	1.783

The results of the multiple protein sequence alignment from the tertiary structures of PfENR with human DECR are presented in table 2 and the 3D structural super-positioning is shown in Figure 12.

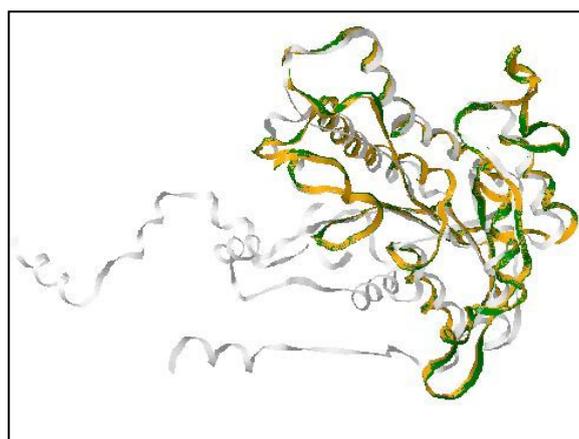


Figure 12. 3D structural super-positioning of the PfENR crystal structures 1NHG (green), 1NNU (yellow) and human DECR 1W6U (silver).

5.2.2. Characteristics of the Active Site of PfENR

The active site interactions with the bound inhibitors in the presence of co-factors have been studied in the case of PfENR compared with the ENR from other sources. Thiolactomycin and Triclosan are known as specific inhibitors of the pathway in which this enzyme is involved. The anti-microbial agent Triclosan has been used in the study and has revealed major structural differences in the substrate/inhibitor binding loop [55]. The enzyme active site pocket involves greatly van der Waals interactions and hydrogen bonding interactions as well. The residues Tyr-267, Tyr-277, Gly-313, Pro-314, Ile-323, Phe-368, Ile0369 and Ala-372 form a hydrophobic pocket of the enzyme. One more important and common interaction is the hydrogen bonding of the OH atom of Tyr-277 with the O17 atom of Triclosan. The co-factor NAD⁺ also binds to the nicotinamide binding pocket which lies a bit at the surface of the protein where it makes interactions with the side chains of the residues of the active site pocket and surrounds the ring-B (2,4-dichlorophenoxy ring) of the Triclosan at the other side [54].

5.3. Docking Software

Two docking programs are used in the current virtual screening work namely FlexX and AutoDock. A brief description about these programs is given below. More detailed information can be found in the documentations and can also be obtained for FlexX from [75] and for AutoDock from [73].

5.3.1. FlexX

FlexX is commercial software available from BioSolveIT [75], which accurately predicts the geometry as well free energy of binding of the protein-ligand complex and interaction between the protein and ligand. FlexX is very fast and is best suited for virtual high-throughput screening. FlexX employs an incremental construction algorithm for molecular docking, where the ligand is build up fragment by fragment before the complex is completely built up in the active site of the receptor. The scoring function implemented in FlexX is empirical scoring derived from the interaction types of the protein-ligand complex.

5.3.2. AutoDock:

AutoDock is a suit of automated docking tools, which allows flexible ligand docking [73] and freely available under the GNU general public license [74]. AutoDock predicts how small molecules, such as substrates or drug candidates bind to a receptor of known 3D structure.

AutoDock suit includes two main programs: the AutoGrid, which pre-calculates the grids describing the target protein and the AutoDock, which performs the docking of the ligand to the target protein.

The principal algorithm and ligand sampling method used in AutoDock is the Genetic Algorithm (GA), local search and global-local search method based on Lamarckian genetics referred to as Lamarckian Genetic Algorithm (LGA). The scoring function used is empirically derived, for empirical binding free energy force field that allows the prediction of binding free energies for docked ligands. AutoDock is based on the United Atom force-field of AMBER, which uses only polar hydrogens. This helps to reduce the number of atoms that must be modelled explicitly during the docking, thus speeding up the calculations.

5.4. Preparation of Targets

The protein targets need to be prepared and modelled according to the format requirements of the docking algorithms used. The crystal structures selected for the docking experiment were prepared for use with both FlexX and AutoDock software. The process of target preparation is described for both methods as below.

5.4.1. Target Preparation with FlexX

The protein crystal structures contain the co-crystallized ligands, cofactors if any, (the hetero atoms) and water molecules; they are removed in the very beginning. The co-crystallized ligand is separated and used as a reference, and the protein active site is defined which is comprised of the atoms within 6.5 Å radius of the reference ligand. The active site residues containing the atoms coordinate records of these atoms are kept in a separate file. A receptor description file is generated which includes the receptor surface description, reference to that active site pocket, the reference ligand and cofactors if any. The input file formats are described in detail in the FlexX introduction section. The PfTIM targets do not contain a cofactor while the PfENR targets contain NAD as cofactors. Therefore the cofactor for the PfENR targets were provided as separate files and described in the RDF file.

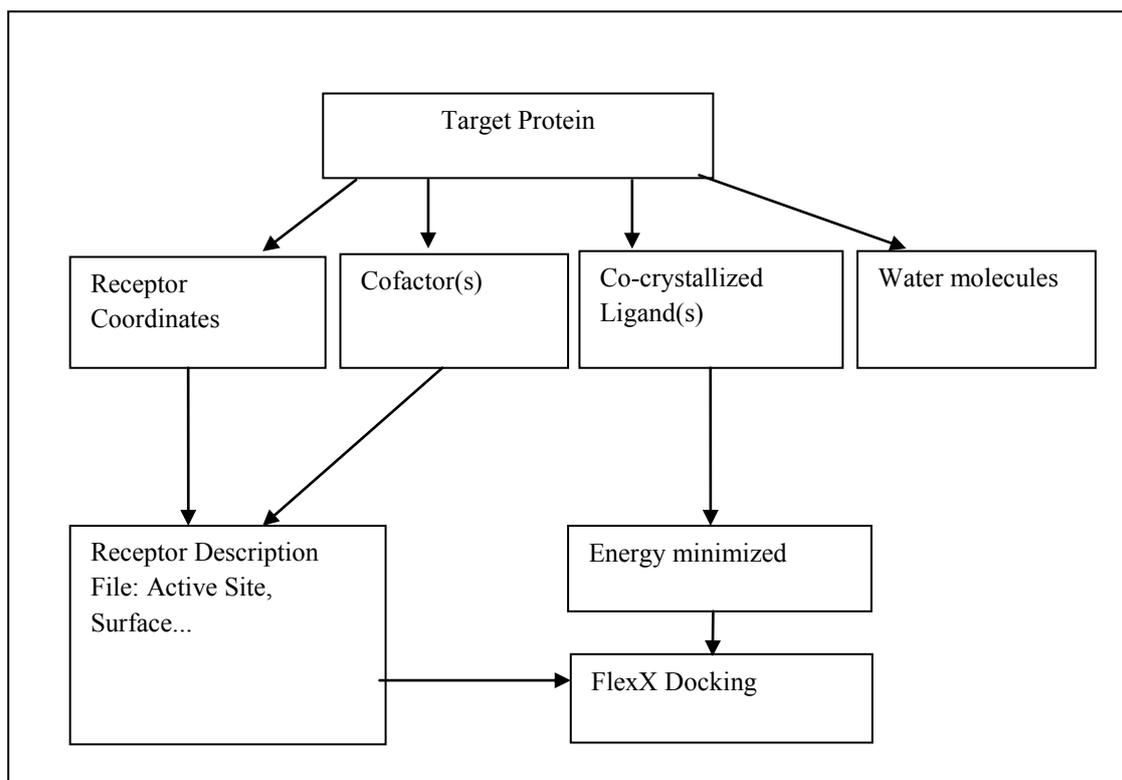


Figure 13. Target preparation steps with FlexX.

5.4.2. Target Preparation with AutoDock

Preparation of the targets for AutoDock is different than that of FlexX. The water molecules and co-crystallized ligands are removed. The cofactor (if any) is not removed and kept in the same target file. AutodockTools (ADT/MGLTools [56]) provides a nice graphical user interface for working with AutoDock. The proteins for docking with AutoDock are prepared using ADT, which includes adding polar hydrogens to the protein atoms and assigning Kollman charges afterwards. For the ligand, all hydrogen atoms must be present on it to calculate partial atomic charges on the ligand.

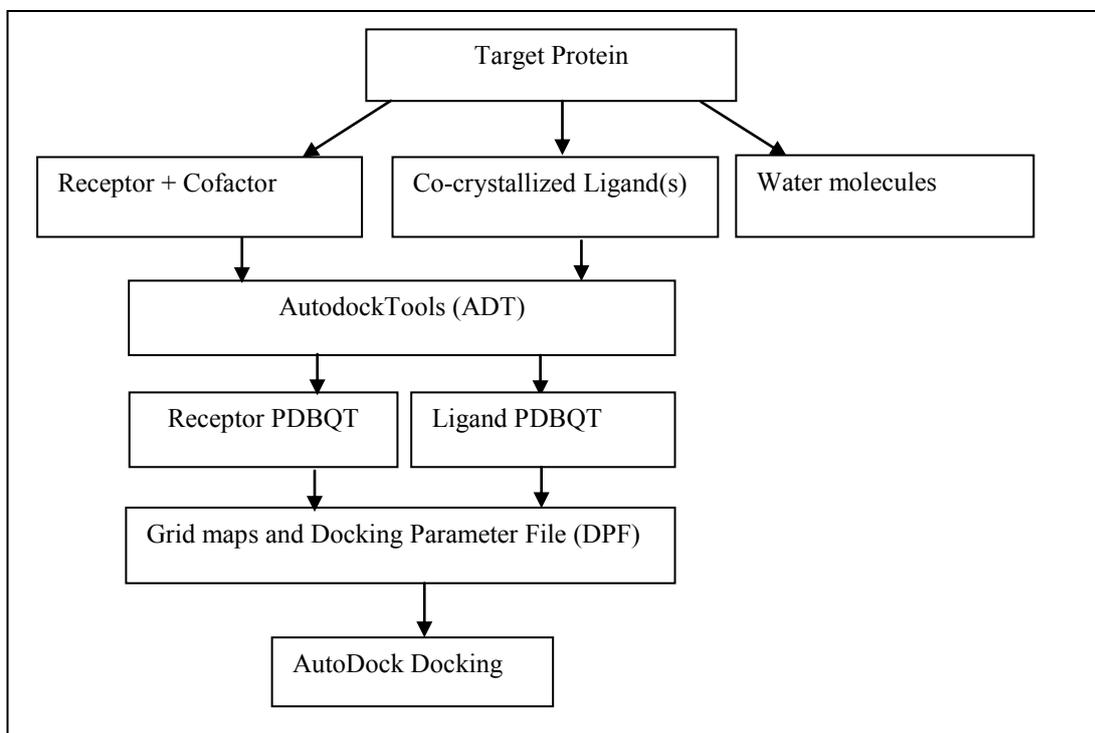


Figure 14. Target preparation steps with AutoDock.

The protein active site is defined by placing a grid over the centre of co-crystallized ligand. Hydrogens are also added to the ligand and Gasteiger charges are assigned. Before a protein is ready for docking simulations, all the necessary grid maps are calculated prior to docking. There is one grid map per ligand atom type needed along with two other maps; electrostatics and desolvation maps. The grid maps are generated with the help of AutoGrid, which is a program of the AutoDock suite.

6. Re-dockings, Cross-Dockings and Test Runs

Re-dockings, cross-dockings and tests for parameters validations were performed using both docking algorithms. In the process of re-docking, the co-crystallized ligands were taken out from the target structure complexes and re-docked into the respective receptors. The purpose of a re-docking experiment is to verify and validate the docking parameters for their input files, recover a known complex structure by the induced fit and reproduce the interactions. In the process of cross-docking, all the co-crystallized ligands are extracted from the ligand-receptor pairs and then docking every ligand to every receptor, for the purpose of finding out the effects of induced fit, binding modes and specificity of a ligand to a receptor structure. Several test runs were performed with FlexX and AutoDock with different parameters sets for the purpose of finding out suitable docking parameters of the docking algorithms, validation of the dockings and testing the execution environment of the testbed. The results of all individual tests performed with FlexX and AutoDock are described in the following sections respectively.

6.1. Re-docking with FlexX

The FlexX re-docking results for all the 6 target structures of PfTIM and PfENR are presented in the tables below. Re-docking with FlexX was performed with setting up the parameters “place particle” to 0 and 1 respectively. In the following tables, the co-crystallized ligands of the PfTIM targets are Phosphoglycolate (PGA) and 2-Phosphoglycerate (2PG). For the target structures of PfENR, Triclosan (TCL) and the TCL analog: 6-(4-chloro-2-hydroxyphenoxy) naphthalen-2-ol (TCT) are the respective co-crystallized ligands used in the re-docking experiments.

Table 3. FlexX best scores and RMSD values for re-docking with “place particles” parameter set to 0.

Target	Ligand	FlexX Score	RMSD	FlexX Sore (lowest RMSD)	Lowest RMSD	Solution RANK (lowest RMSD)
1LYX	PGA	-30.023	4.287	-24.170	1.611	6
1LZO	PGA	-21.725	2.715	-18.803	2.368	10
1O5X	2PG	-18.979	3.195	-18.979	3.195	1
1HTI	PGA	-21.691	4.179	-18.927	1.615	6
1NHG	TCL	-13.445	7.794	-11.485	0.742	12
1NNU	TCT	-19.856	1.819	-17.984	1.725	5

Table 4. FlexX best scores and RMSD values for re-docking with “place particles” set to 1.

Target	Ligand	FlexX Score	RMSD	FlexX Score (lowest RMSD)	Lowest RMSD	Solution RANK (lowest RMSD)
1LYX	PGA	-35.873	4.287	-29.685	1.611	4
1LZO	PGA	-22.430	2.412	-21.186	2.368	8
1O5X	2PG	-18.040	3.377	-16.581	3.037	7
1HTI	PGA	-22.487	4.548	-21.238	2.262	5
1NHG	TCL	-13.445	7.794	-11.485	0.742	12
1NNU	TCT	-20.105	1.819	-17.984	1.725	5

Results of FlexX re-docking experiments are presented in terms of FlexX docking scores and reference RMSD values in angstrom (Å). An RMSD value above 3Å is considered not authentic. Reference RMSD values in the 3rd column of the above tables are given for the top ranking solution based on FlexX docking score. FlexX score for lowest RMSD are given in the 5th column and the RMSD values and the solution ranks for the lowest RMSD are given in column 6th and 7th, respectively. FlexX scores and the reference RMSD values obtained for both targets structures of PfTIM with two parameters showed that the docking scores were higher for the re-docking experiments with water particles than those without water particles. Moreover, an improvement in the RMSD values was observed and the significant interactions of the ligands with the active site residues of the receptors 1LYX and 1O5X are predicted well in setting the place particle parameter to 1 as compared to setting the same parameter to 0. For the target structures of PfENR, no significant improvement was observed in the docking scores, reference RMSD values and the important ligand-receptor interactions. 2D and 3D LigPlot representations of the PfTIM structure 1LXY-PGA complex and the FlexX predicted conformation of the re-docked complex is shown in Figure 15. A similar diagram with 2D and 3D LigPlot representations and the FlexX predicted conformations of the co-crystallized ligand in the PfENR structure 1NNU-TCT complex is shown in Figure 16.

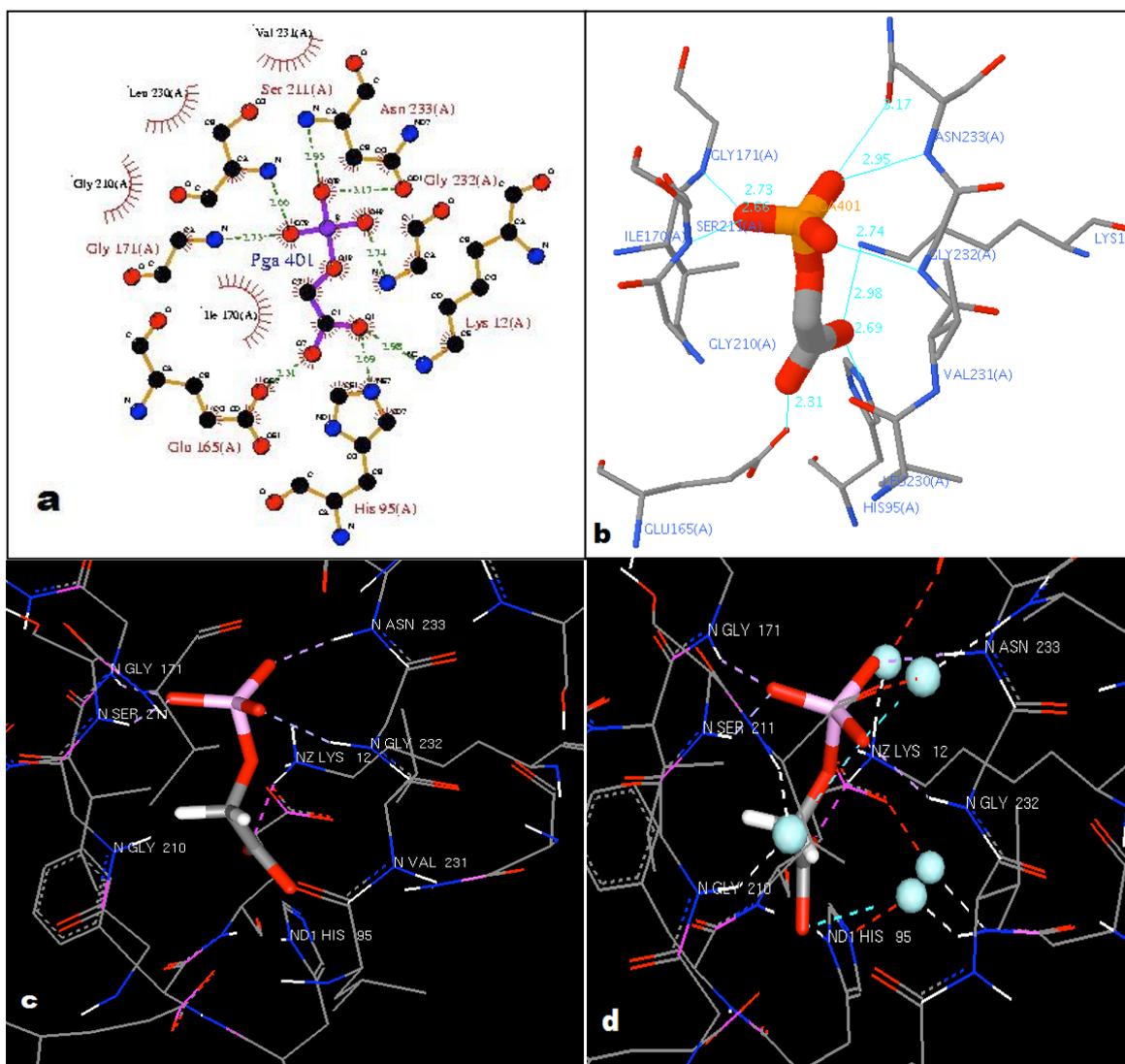


Figure 15. (a). LigPlot of target structure 1LYX of PfTIM. (b) 3D representation of the LigPlot view. (c) Geometry of the co-crystallized ligand PGA and the interaction with active site residues after re-docking with FlexX (with out water particles). Hydrogen bonds are shown with the dotted lines. (d). Predicted conformation of the co-crystallized ligand PGA and the interactions with the active site residues of target 1LYX obtained after re-docking with FlexX (with water particles presented as small balls in light blue color).

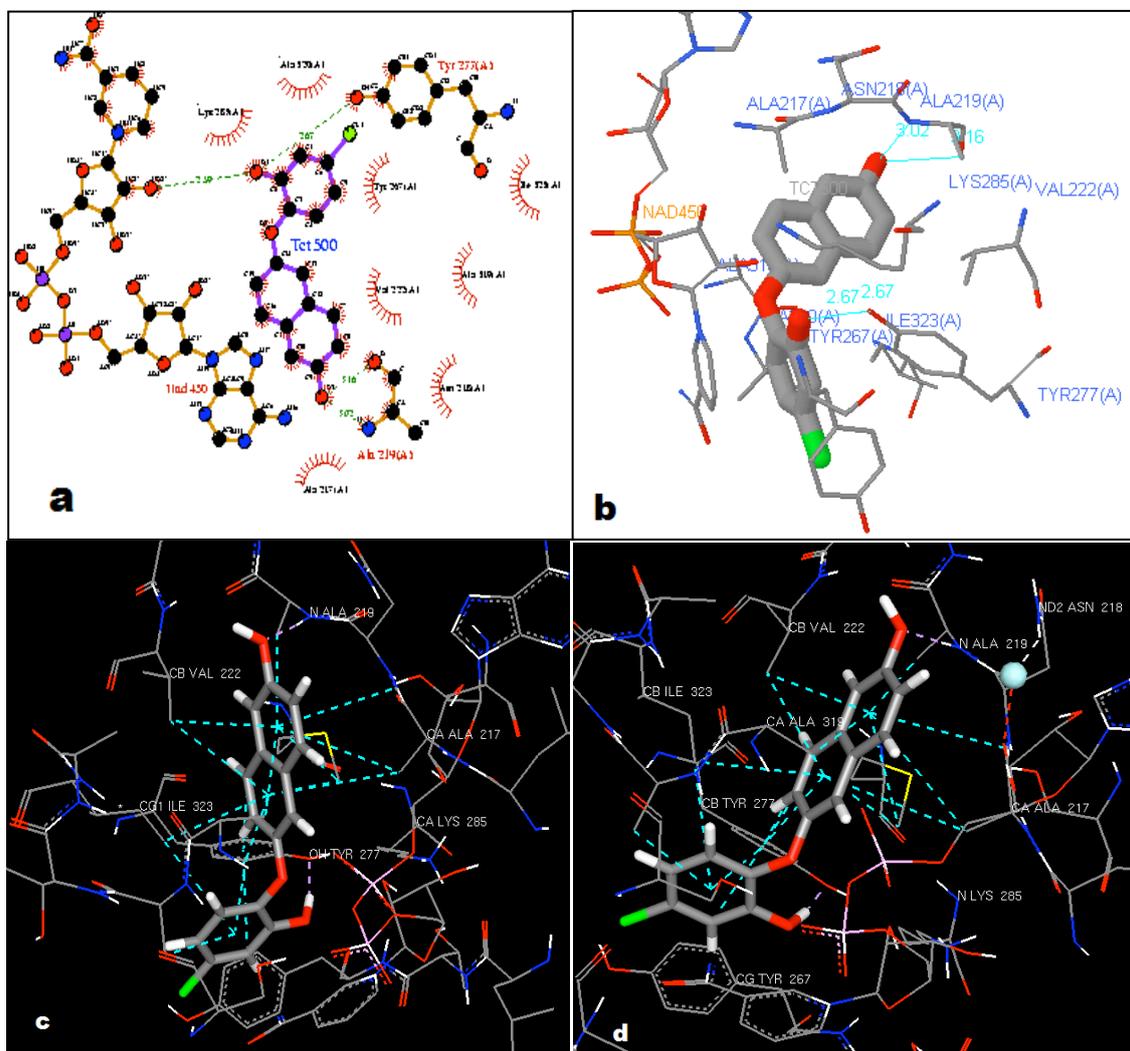


Figure 16. (a) LigPlot of target structure 1NNU of PfENR. (b) 3D representation of the LigPlot view. (c) The predicted conformation obtained after re-docking of the co-crystallized ligand TCT and the interactions with the active site residues of 1NNU (without water particles). (d) The predicted conformation obtained after re-docking of the co-crystallized ligand TCT and the interactions with the active site residues (with water particles shown as small ball in light blue color).

6.2. Cross Dockings with FlexX

Cross-docking tests were also performed for the target structures of PfENR, with the same set of parameters used in re-docking experiments i.e setting the place particle parameter to 0 and 1. The co-crystallized ligands were re-docked in each receptor structure. The RMSD values calculated for the four target structures after superimposed onto each other are presented in table 5. FlexX docking scores for the top ranking solutions and the reference RMSD values obtained after cross dockings with both parameters sets are presented in table 6 and 7, respectively.

Table 5. RMSD (Å) calculated after superposing of the four structures.

Targets	1LYX	1LZO	1O5X	1HTI
1LYX	0.00	0.54	0.46	0.95
1LZO	0.54	0.00	0.44	1.07
1O5X	0.46	0.44	0.00	0.93
1HTI	0.95	1.07	0.93	0.00

Table 6. FlexX scores and reference RMSD values for top ranking solutions obtained in cross dockings with place particles parameter set to 0 for targets of PfTIM.

Targets	1lyx_min		1lzo_min		1o5x_min		1hti_min	
	Score	RMSD	Score	RMSD	Score	RMSD	Score	RMSD
1LYX	-30.02	4.28	-28.66	3.05	-28.03	4.70	-27.22	4.34
1LZO	-21.68	4.47	-21.30	2.87	-23.33	4.92	-22.75	4.31
1O5X	-17.05	4.69	-17.82	3.89	-20.41	3.20	-16.90	4.09
1HTI	-18.74	3.63	-21.36	2.23	-25.48	3.52	-19.09	3.60

Table 7. FlexX scores and reference RMSD values for top ranking solutions obtained in cross dockings with place particles parameter set to 1 for targets of PfTIM.

Targets	1lyx_min		1LZO_min		1O5X_min		1HTI_min	
	Score	RMSD	Score	RMSD	Score	RMSD	Score	RMSD
1LYX	-35.87	4.28	-32.49	3.05	-31.74	4.65	-32.01	4.34
1LZO	-22.14	4.32	-23.69	2.48	-25.90	4.32	-24.72	4.00
1O5X	-16.29	4.52	-19.16	1.74	-22.23	2.41	-17.68	3.84
1HTI	-22.27	3.58	-23.04	2.23	-27.04	4.66	-22.10	3.39

Results of cross dockings has shown that in terms of RMSD values comparisons between the crossed ligand-receptor dockings for the top ranked solutions, there exists no significant differences, which shows no clear specificities of a ligand to another receptor. The docking scores and the important ligand-receptor interactions are also improved in the predicted conformations of the top ranking solutions for the receptor structures, with setting up the place particle parameter to 1.

6.3. Test Runs with FlexX

A set of known bioactive molecules for both targets of PfTIM and PfENR, collected from DrugBank [57] and PubChem [58] were used further in the tests. These test runs were performed with FlexX with different parameters sets for the purpose of setting up appropriate docking parameters of the docking algorithms and further validate the dockings. Furthermore, the other objectives of these tests are to test the execution environment of the Grid, test the functionalities of the middleware and formulate a strategy in creating and submitting the docking jobs to the Grid. The compounds were docked into each receptor in the presence of water molecules and without water molecules. The FlexX total score results for the targets of PfTIM and PfENR, respectively, are presented in the following tables and charts. “Total Score 0” means the FlexX score obtained with setting the place particle parameter to 0 and “Total Score 1” means the score obtained with setting the place particle parameter to 1.

Table 8. FlexX scores with and without place particles for the PfTIM targets 1LYX and 1LZO

Target 1LYX			Target 1LZO		
Compounds	Total Score 0	Total Score 1	Compounds	Total Score 0	Total Score 1
13P	-24.412	-26.782	13P	-24.175	-24.058
2PG	-29.401	-30.916	2PG	-22.204	-23.769
3PG	-25.984	-33.952	3PG	-22.846	-25.371
3PP	-24.519	-28.558	3PP	-19.467	-21.452
3PY	-25.585	-26.273	3PY	-19.652	-21.223
4PB	-25.971	-28.294	4PB	-28.248	-29.668
ACT	-14.068	-18.191	ACT	-12.985	-13.323
BTS	-18.487	-23.598	BTS	-15.48	-16.915
DMS	-8.418	-9.359	DMS	-9.288	-9.137
FTR	-27.841	-31.14	FTR	-20.308	-19.97
G3P	-26.733	-32.265	G3P	-21.047	-24.544
M129	-23.911	-29.877	M129	-20.082	-23.163
PGA549	-29.754	-35.591	PGA549	-22.02	-21.637
PGA7251	-27.103	-30.969	PGA7251	-21.725	-22.43

PGA	-30.023	-35.873	PGA	-19.956	-21.308
PGH	-26.847	-31.307	PGH	-28.764	-27.478
TBU	-8.543	-8.668	TBU	-3.503	-3.503

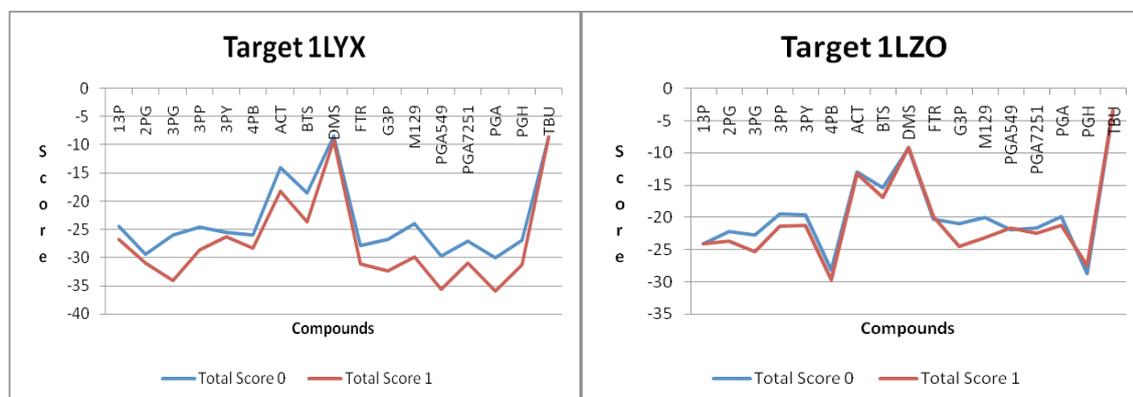


Figure 17. Comparison of FlexX total scores. Total score 0 = score without water particles. Total score 1 = score with water particles.

Table 9. FlexX scores with and without place particles for PfTIM 105X and human TIM 1HTI.

Target 1HTI			Target 105X		
Compounds	Total Score 0	Total Score 1	Compounds	Total Score 0	Total Score 1
13P	-23.675	-25.597	13P	-14.677	-18.723
2PG	-23.855	-25.682	2PG	-18.979	-18.04
3PG	-16.184	-21.338	3PG	-17.807	-15.022
3PP	-18.854	-20.723	3PP	-16.471	-20.449
3PY	-20.397	-21.408	3PY	-16.808	-17.593
4PB	-18.566	-21.523	4PB	-20.775	-23.162
ACT	-10.321	-11.857	ACT	-10.067	-11.241
BTS	-12.905	-15.543	BTS	-16.066	-14.797
DMS	-5.432	-6.628	DMS	-6.753	-6.753
FTR	-24.413	-22.629	FTR	-19.955	-20.531
G3P	-18.265	-22.76	G3P	-18.944	-19.155
M129	-18.544	-19.92	M129	-16.021	-17.74
PGA549	-21.691	-22.487	PGA549	-17.609	-16.791
PGA7251	-20.909	-22.955	PGA7251	-19.3	-20.775
PGA	-21.742	-23.274	PGA	-18.426	-18.106
PGH	-25.085	-26.511	PGH	-19.058	-21.221
TBU	--	--	TBU	-1.473	-1.473

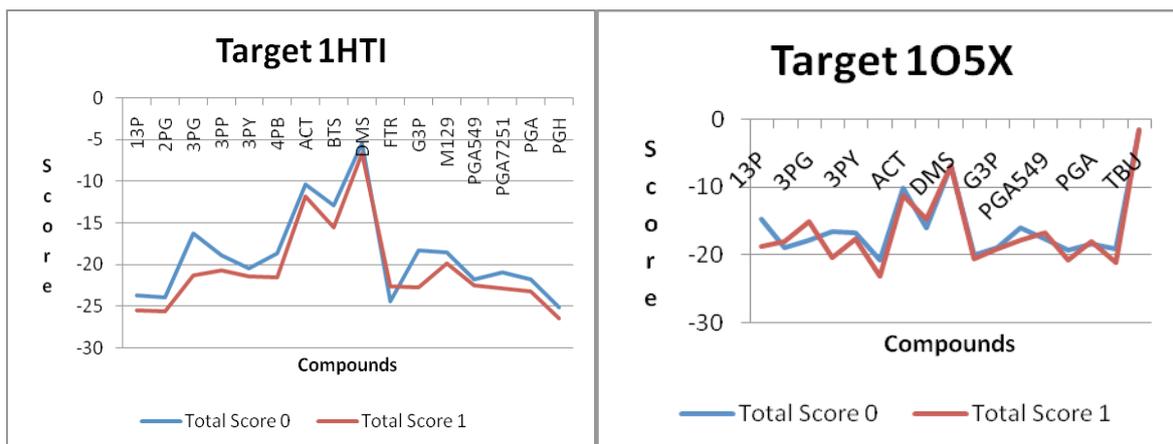


Figure 18. Comparison of FlexX total scores. Total score 0 = score without water particles. Total score 1 = score with water particles.

The figures above show that the FlexX score obtained for target Structure 1LYX (Figure 17) and 1HTI (Figure 18) were higher with including water particles than the docking scores obtained without water particles. Again, the significant interactions of most of the known compounds with the target structures and the binding modes were improved for the target structures 1LYX and 1HTI as compared as the target structures 1O5X and 1LZO, where no significant differences were observed. Based on the improvements in FlexX docking scores and the good predictions of the important interactions, the place particle parameter was selected to be set to 1 and considered as appropriate for the FlexX data challenge containing approximately 200,000 dockings. As FlexX program is very fast, therefore, the average number of dockings included in a single Grid job was set to 100.

Table 10. FlexX scores with and without place particles for the PfENR targets 1NHG and 1NNU.

Target 1NHG			Targe 1NNU		
Compounds	Total score 0	Total score 1	Compounds	Total Score 0	Total Score 1
3767	-11.056	-11.056	3767	-16.436	-16.436
654	-13.037	-13.037	654	-11.288	-11.805
826	-17.492	-17.492	826	-18.045	-17.32
AYM	-16.317	-16.317	AYM	-16.604	-17.664
DCN	-12.131	-12.319	DCN	-12.959	-12.959
GEQ	-19.952	-19.952	GEQ	-21.052	-20.269
IDN	-22.891	-23.366	IDN	-18.334	-18.411
JP1	-19.146	-19.008	JP1	-17.383	-17.383
JPA	-15.175	-15.508	JPA	-15.533	-15.533
TCC	-22.051	-22.051	TCC	-17.673	-17.612

TCL	-12.086	-12.086	TCL	-12.383	-12.383
TCT	-22.305	-23.647	TCT	-19.657	-20.088
THT	3.031	3.578	THT	3.926	3.942
TN2	-14.723	-14.723	TN2	-14.978	-14.978
TN5	-15.818	-15.818	TN5	-16.262	-16.262
ZAM	-17.39	-15.328	ZAM	-18.757	-18.757

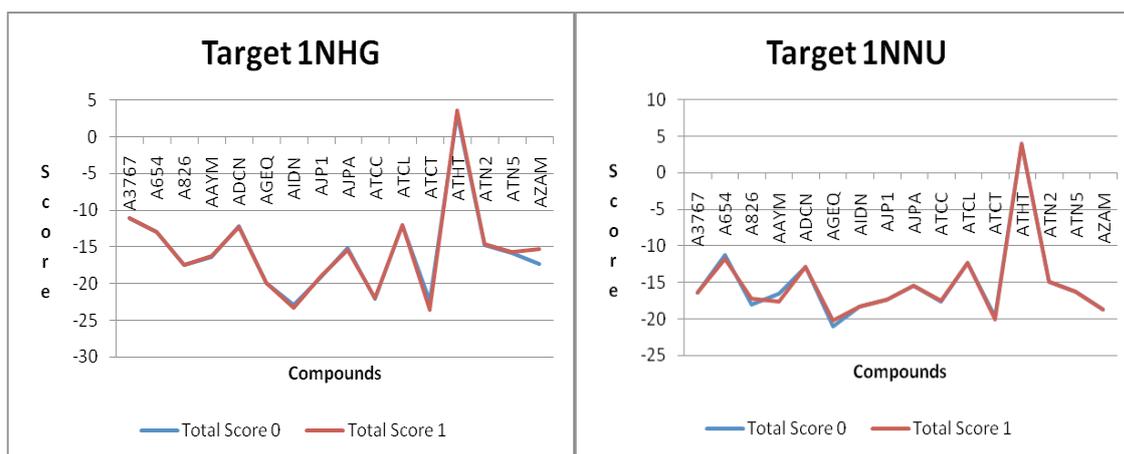


Figure 19. Comparison of FlexX total scores for target structures of PfENR. Total score 0 = score without water particles. Total score 1 = score with water particles.

Although, no differences in FlexX total scores were observed for both target structures of PfENR by testing the two parameters sets, some of the important interactions were predicted well in the top ranking solutions. Most of the active compounds were having interactions with the important residues including Tyr277, Tyr267, Ala219 and Ala319. Therefore, the place particle parameter was considered also for the targets of PfENR for the whole screening process.

6.4. Re-docking with AutoDock

The co-crystallized ligands taken out from their respective receptors were also re-docked using AutoDock. The re-docking was performed using the standard defined parameters set for AutoDock in the AutoDockTools and the Lamarckian Genetic Algorithm. The re-docking tests were performed using both versions (version-3.0.5 and version-4.0.1) of AutoDock, and the results were analysed with AutoDockTools as presented in the following table and figures.

Table 11. AutoDock re-docking tests using version 3 and 4.

Target	Ligand	AutoDock4		Autodock3	
		Binding Energy	Ref. RMSD Å	Binding Energy	Ref. RMSD Å
1LYX	1LYX_min	-5.68	2.53	-9.48	0.94
1LZO	1LZO_min	-6.93	3.99	+2.53	6.95
1O5X	1O5X_min	-5.38	3.62	-1.59	3.06
1HTI	1HTI_min	-4.11	1.95	+2.07	1.82
1NHG	1NHG_min	-8.69	0.48	-6.95	1.00
1NNU	1NNU_min	-9.51	0.92	-5.83	1.63

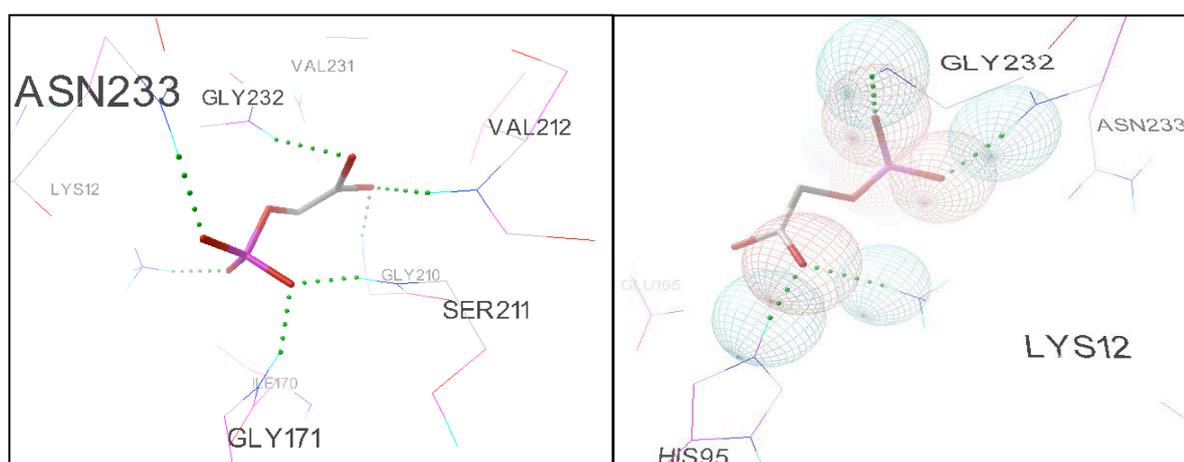


Figure 20. Re-docking experiment performed with both versions of AutoDock (3.0.5 and 4.0.1). (Left): The predicted geometry of the co-crystallized ligand PGA into the active site pocket of the structure 1LYX with AutoDock-4. Hydrogen bonding interactions are shown with green dots. (Right): Conformation of the same ligand obtained with AutoDock-3. Not all interactions were predicted with version 3 while they were mostly predicted by version 4.

In the case of PfTIM targets, most of the critical interactions were predicted and confirmed by AutoDock4. The top ranked generated conformations were also quite acceptable in terms of reference RMSD (less than 2 Å) to the co-crystallized ligand conformations for all the structures except the targets 1LZO and 1O5X where the RMSD values produced were greater than 3 angstrom. The Figure 20 above (left) shows that for the target structure 1LYX, almost all hydrogen bonding activities except for His-95 and the close contacts with the active site residues were predicted very good by AutoDock4, while in the case of AutoDock3 (right), a few of the hydrogen bonding interactions with the critical active site residues were not predicted.

For the target structure 1LZO, which has the catalytic loop crystallized in open conformation, both AutoDock4 and AutoDock3 predictions were not satisfactory in terms of reference RMSD, binding energies and in terms of critical interactions with the active site residues. The reason for the higher RMSD values and positive binding energy values obtained with AutoDock version-3 could be possibly due to conformation of the loop near the active site. For the remaining target structures of PfTIM, 1O5X and 1HTI, AutoDock3 predictions were not good while AutoDock4 predictions (as shown in Figure 21 below) for these targets were near to that described in the literature.

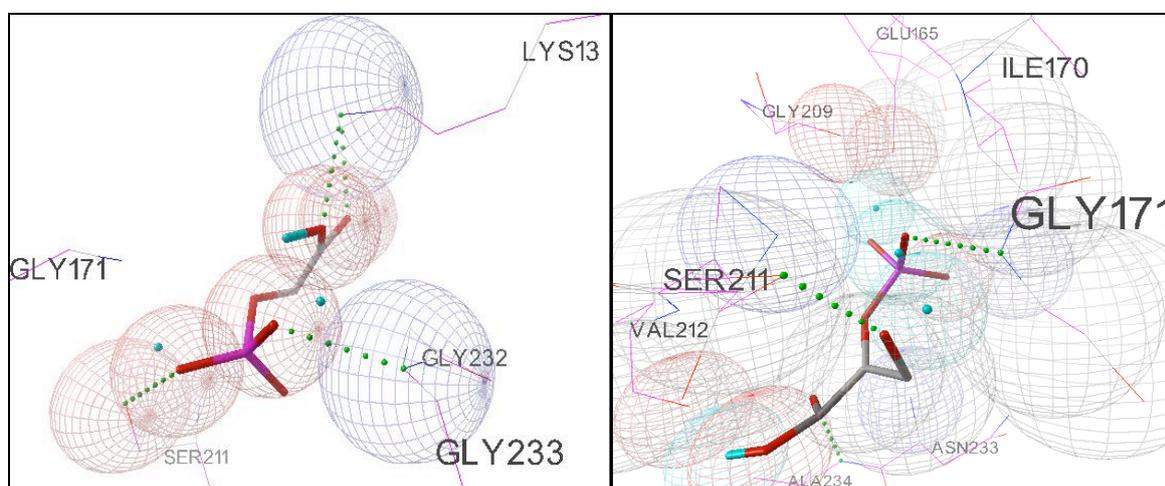


Figure 21. Re-docking experiment performed with AutoDock for target structures of PfTIM. (Left): The predicted geometry of the co-crystallized ligand PGA into the active site pocket of the structure 1HTI. Hydrogen bonding interactions are shown with green dots. (Right): Conformation of the co-crystallized ligand 2PG within the active site pocket of the structure 1O5X.

For the target structure of PfENR, 1NHG and 1NNU, the re-docking results were quite good and predicted very well by both AutoDock3 and AutoDock4 in terms of binding energies,

reference RMSD values and the interaction with the active site residues as shown in Figure 22 and Figure 23, respectively. Re-docking results in the case of target structure 1NNU are shown in the Figure 23. Here, the second conformation produced by AutoDock-4 has predicted the critical interactions with the active site residues and the cofactor NAD.

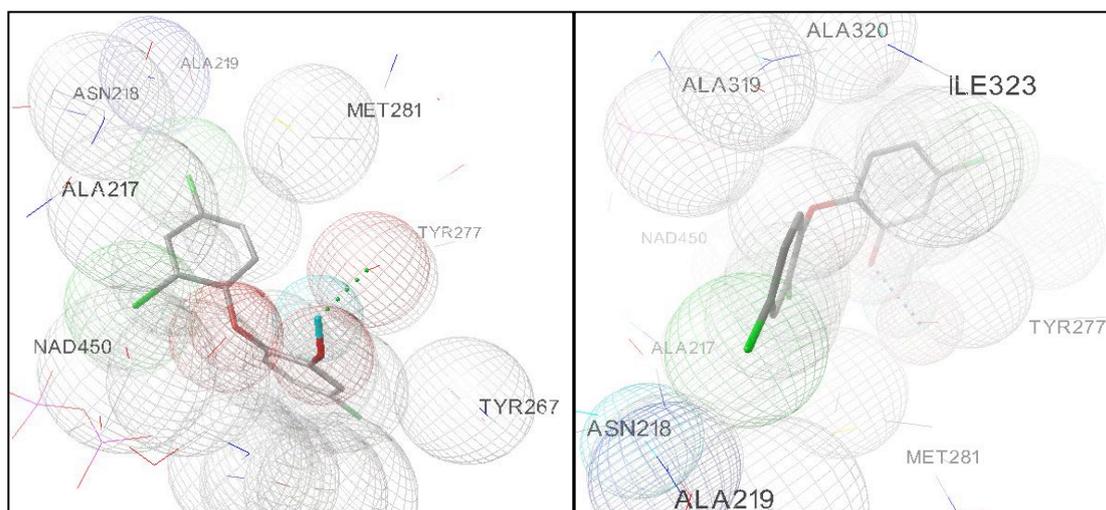


Figure 22. Re-docking experiment performed with both versions of AutoDock (3.0.5 and 4.0.1). (Left): The predicted geometry of the co-crystallized ligand TCL into the active site pocket of the structure 1NHG with AutoDock-4. Hydrogen bonding interactions (Tyr-277) are shown with green dots and spheres represents the residues in close contact with the ligand. (Right): Conformation of the same ligand & same receptor obtained with AutoDock-3.

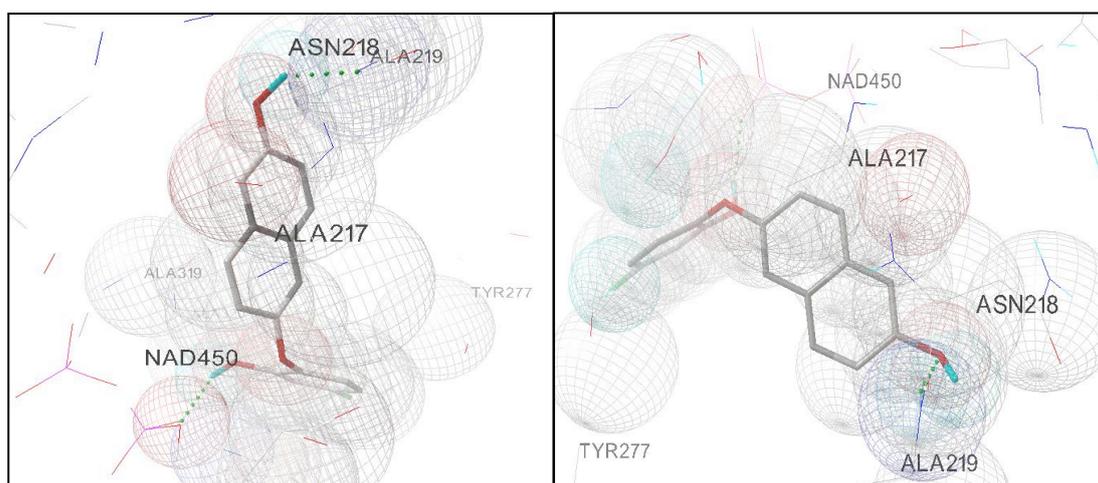


Figure 23. Re-docking experiment performed with both versions of AutoDock (3.0.5 and 4.0.1). (Left): The predicted geometry of the co-crystallized ligand TCT into the active site pocket of the structure 1NNU with AutoDock-4. Hydrogen bonding interactions are shown with green dots and spheres represents the residues in close contact with the ligand. (Right): Conformation of the same ligand & same receptor obtained with AutoDock-3.

6.5. Cross Dockings with AutoDock

Cross dockings for the target structures of PFTIM were carried out using AutoDock with optimal parameters set as given below.

- a) Total Number of energy evaluations: 5,000,000
- b) Number of Genetic Algorithm (GA) Runs: 10
- c) Number of Individuals in GA population: 200
- d) Maximum Number of Generations: 27,000

AutoDock free energies of binding and the reference RMSD values are presented in table-12 below.

Table 12. Cross Docking tests for the target structures PFTIM with AutoDock.

Targ.	1lyx_min		1lzo_min		1o5x_min		1hti_min	
	B.E	RMSD	B.E	RMSD	B.E	RMSD	B.E	RMSD
1LYX	-3.12	2.63	-3.3	2.41	-3.87	2.68	-3.71	2.4
1LZO	-1.81	6.48	-3.0	5.02	-2.73	2.61	-3.05	10.6
1O5X	-2.59	2.62	-3.81	2.92	-4.19	2.94	-4.26	2.56
1HTI	-2.29	2.37	-2.58	1.82	-2.60	3.91	-3.43	2.57

Results of cross docking experiment show that the RMSD values are greater than 3 Å for the target structure 1LZO as compared to the rest of the target structures, where the RMSD values are nearly similar and also below 3 Å.

6.6. Test Runs with AutoDock

The same set of active compounds against the targets of PFTIM and PfENR were used in the test runs for AutoDock. These tests were performed using different parameter sets for AutoDock's algorithm. The algorithm used is Lamarckian Genetic Algorithm (LGA) with Solis and Wets Local Search. The parameters that were set for each test set are as follow:

- 1. a). Total Number of energy evaluations: 250,000
 - b). Number of Genetic Algorithm (GA) Runs: 10
 - c). Number of GA population size: 150
 - d). Maximum number of generations: 27,000
- 2. a). Total Number of energy evaluations: 2,500,000
 - b). Number of Genetic Algorithm Runs: 10

- c). Number of GA population size: 150
- d). Maximum number of generations: 27,000
3. a). Total Number of energy evaluations: 2,500,000
- b). Number of Genetic Algorithm Runs: 30
- c). Number of GA population size: 150
- d). Maximum number of generations: 27,000

The above parameter sets were tested against all the selected targets of PfTIM and PfENR with their respective set of bioactive compounds. The results for these sets of tests are presented in the following tables.

Table 13. AutoDock test runs with 3 parameters sets for PfTIM targets.

Target 1LYX				Target 1LZO			
Compound	Binding Energy-1	Binding Energy-2	Binding Energy-3	Compound	Binding Energy-1	Binding Energy-2	Binding Energy-3
13P	-2.98	-3.89	-3.55	13P	1.04	1.11	1.04
2PG	-4.82	-5.12	-5.12	2PG	1.17	1.17	1.17
3PG	-4.56	-4.7	-4.79	3PG	1.17	1.17	1.17
3PP	-4.12	-4.11	-4.38	3PP	0.87	0.86	0.85
3PY	-3.81	-3.82	-3.82	3PY	0.58	0.58	0.58
4PB	-4.49	-5.18	-5.22	4PB	0.19	0.19	0.19
ACT	-3.5	-3.51	-3.51	ACT	0.03	0.03	0.03
BTS	-5.04	-5.13	-5.74	BTS	0.65	0.65	0.65
DMS	-2.34	-3.01	-3.05	DMS	-0.01	-0.01	-0.01
FTR	-4.91	-4.95	-4.99	FTR	0.14	0.13	0.13
G2H	-4.03	-4.13	-4.22	G2H	1.13	1.12	1.12
G3P	-4.02	-4.42	-4.41	G3P	0.95	0.95	0.95
m129	-2.7	-3.78	-3.27	m129	1.06	1.11	1
PGA	-4.98	-5.07	-5.05	PGA	0.9	0.9	0.9
PGH	-3.15	-2.94	-3.77	PGH	1.11	1.11	1.11

Table 14. AutoDock test runs with 3 parameters sets for PfTIM targets.

Target 1O5X				Target 1HTI			
Compound	Binding Energy-1	Binding Energy-2	Binding Energy-3	Compound	Binding Energy-1	Binding Energy-2	Binding Energy-3
13P	-3.19	-4.46	-3.77	13P	-3.54	-3.85	-3.93
2PG	-2.84	-3.05	-3.13	2PG	-2.7	-2.76	-2.75
3PG	-3.16	-3.31	-3.21	3PG	-2.7	-2.74	-2.72
3PP	-3.3	-3.39	-3.5	3PP	-3.68	-2.96	-3.92

3PY	-2.89	-2.91	-2.91	3PY	-2.82	-2.93	-2.98
4PB	-3.9	-4.53	-4.85	4PB	-3.98	-4.53	-4.48
ACT	-2.03	-2.03	-2.04	ACT	-2.23	-2.24	-2.24
BTS	-6.01	-6.78	-6.01	BTS	-6.21	-6.93	-6.71
DMS	-3.28	-3.29	-3.29	DMS	-3.24	-3.25	-3.25
FTR	-5.68	-5.91	-5.9	FTR	-6.74	-7.8	-7.83
G2H	-3.46	-3.55	-3.58	G2H	-3.02	-3.3	-3.18
G3P	-3.66	-3.82	-3.86	G3P	-3.25	-3.36	-3.37
m129	-3.43	-3.56	-4.38	m129	-3.97	-3.78	-4.2
PGA	-2.17	-2.29	-2.3	PGA	-1.94	-2.13	-2.12
PGH	-2.91	-4.07	-4.4	PGH	-3.44	-4.28	-3.67

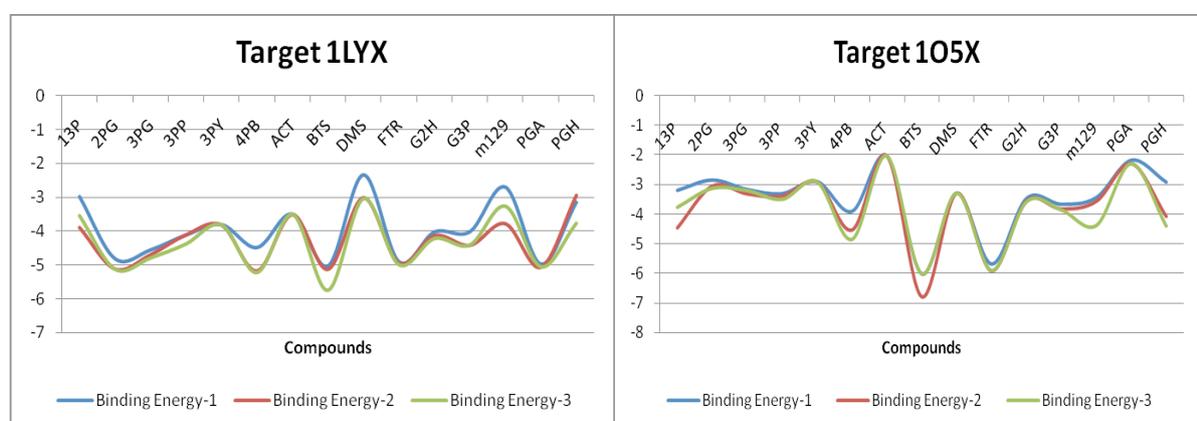


Figure 24. Graphical representations of the AutoDock binding energies for all the three parameters sets used in the test runs with the set of known active compounds.

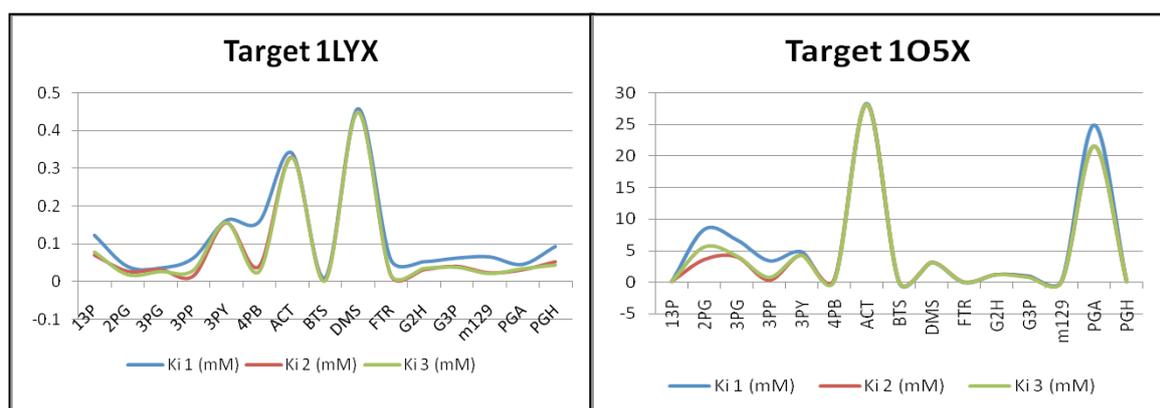


Figure 25. Line chart showing the affinity values (mM) of the known compounds predicted by AutoDock for the three parameters sets for target structures 1LYX and 105X used in test runs.

Table 15. AutoDock test runs with 3 parameters sets for PfENR targets.

Target 1NHG				Target 1NNU			
Compound	Binding Energy-1	Binding Energy-2	Binding Energy-3	Compound	Binding Energy-1	Binding Energy-2	Binding Energy-3
3767	-4.37	-4.4	-4.39	3767	-4.35	-4.4	-4.72
654	-7.4	-7.43	-7.43	654	-7.3	-7.32	-7.32
826	-8.35	-8.92	-8.97	826	-8.03	-8.22	-8.33
AYM	-8.09	-8.7	-8.67	AYM	-7.98	-8.06	-8.6
DCN	-7.71	-7.37	-7.29	DCN	-7.34	-7.6	-7.49
GEQ	-9	-8.94	-9.4	GEQ	-8.66	-8.55	-8.99
IDN	-9.09	-9.45	-9.25	IDN	-9.05	-9.47	-9.67
JP1	-9.09	-9.24	-9.25	JP1	-8.72	-9.16	-9.24
JPA	-10.43	-10.77	-10.77	JPA	-9.89	-10.55	-10.68
TCC	-7.68	-7.71	-7.93	TCC	-7.69	-7.7	-7.83
TCL	-7.85	-8.5	-8.4	TCL	-7.76	-8.09	-7.98
TCT	-9.04	-9.33	-9.36	TCT	-9.34	-9.53	-9.52
THT	-5.36	-6.53	-6.61	THT	-5.82	-6.49	-6.61
TN5	-8.34	-8.37	-8.34	TN5	-7.95	-8.63	-8.64
ZAM	-7.72	-8.79	-8.8	ZAM	-8.1	-8.77	-8.67

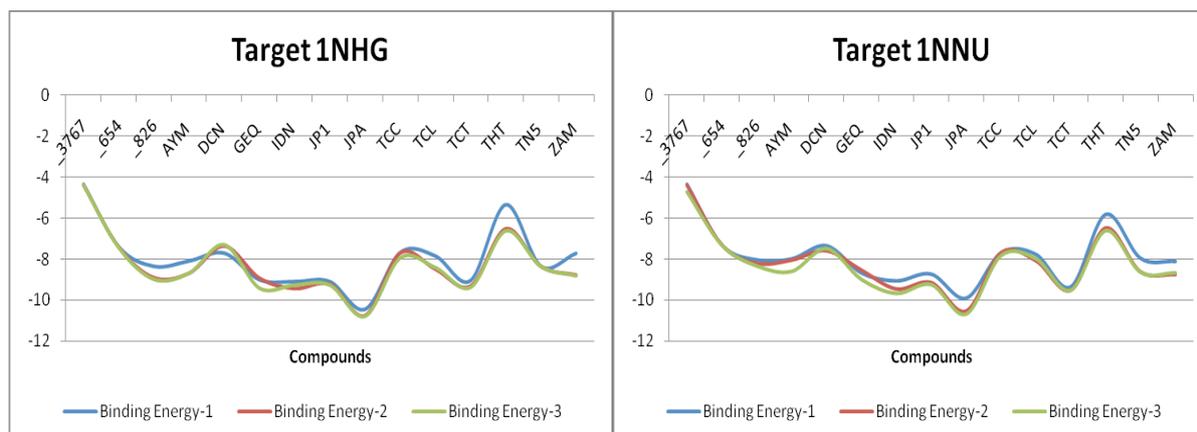


Figure 26. Graphical representations of the AutoDock binding energies for all the three parameters sets used in the test runs with the set of known active compounds.

The above data shown in the tables and both Figure 24 and Figure 26, that the second parameter set used in the test runs for AutoDock produced the optimum results in terms of AutoDock binding energies and predicted affinity values (Figure 25) compared for all the three parameters sets. The AutoDock binding energies in this parameter set were comparatively higher than the one obtained by using the first set of AutoDock parameters. For a few compounds, the binding energies noted in the parameters set-3 were higher than the first and second parameter set but no significant improvement in the prediction of ligand-

receptor interactions. Moreover, the processing time used by AutoDock for a single docking was three times higher in parameters set-3. The experimental affinity values for the representative ligand Phosphoglycolate (PGA) for the targets 1LYX and 1LZO is 0.029 mM [30] and 7.4 uM [33] for the target 1HTI, respectively. The affinity values predicted by AutoDock for the three parameters sets are presented in Figure 25 for the target structures 1LYX and 1O5X. The predicted affinity values for 1LYX-PGA complex obtained is 0.044 mM, 0.031 mM and 0.033 mM with 3 parameters sets, respectively, which is near to the experimental value. For the other targets structures the range of predicted affinity values were higher than the experimental reported values. Although, the order of predicted values remain the same as the order of values for free energies of binding with three parameters sets. The binding modes of the compounds inside the active site of the receptors were generally improved and remained similar in the parameter set-2 and 3, while there were very few compounds whose binding modes and interactions were better predicted in parameter set-1 as compared to the other sets of parameters. Overall, the results of test runs significantly improved using parameters set-2. Based on these observations and processing time of a docking run, the parameter set-2 was chosen for the large-scale virtual screening on the Grid.

The parameter set 2 was used with both AutoDock versions 4 and 3. Although AutoDock-4 has a new free-energy scoring function than used in AutoDock-3, the order of calculated binding energies for the target structures with the respective set of compounds were the same. Moreover, the new version of AutoDock is much faster than the old version. The Figure 27 below shows the binding energies calculated by both versions for the target structure 1LYX.

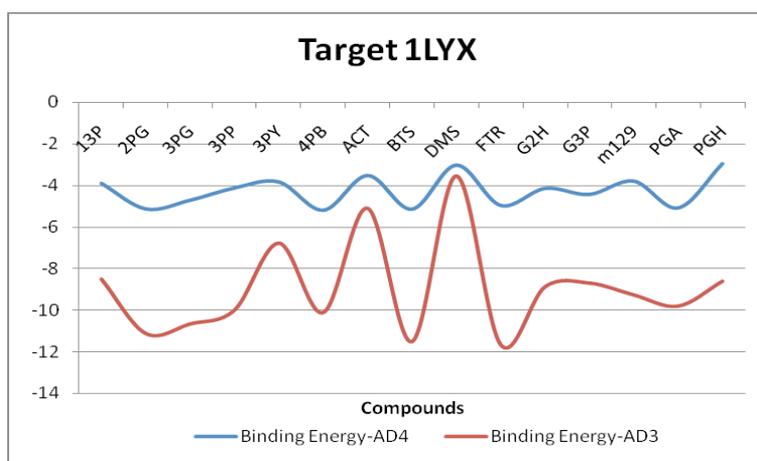


Figure 27. Comparison of the binding energies obtained with both versions of AutoDock for the second set of parameters.

7. Compound Database Design & Project Deployment

In virtual screening, large compound libraries of molecules are docked into the target structure. Such compound libraries contain millions of small molecules and are actually larger than the libraries used in experimental screenings. To screen each and every molecule from these databases is not relevant, that is, these databases of organic molecules also contain unwanted and unnecessary compounds, which have undesirable or some toxic effects. There is a great need to design such a focused library, which is more interesting biologically, thus reducing the “haystack” and increasing the chances of finding new drugs. The unnecessary and uninteresting molecules should be separated by applying certain filters such as drug-likeness (e.g. Lipinski rules). Such filtering strategy will help reducing the burden of scientific computations and save the computing resource from wasting.

7.1. Filtering the database:

As discussed earlier, to get rid of the unwanted compounds from the database, and get the drug-like compounds, the database to be designed is filtered in the first step. This filtering means the optimization of the physiochemical properties for the molecules inside the database. Such properties of the small molecules are validated by applying simple criteria like Lipinski’s “rule of five” [61] and some other filters which meets the demands of a targeted and focused library.

The compound database is filtered according the rules mentioned below:

1. Molecular weight less than 500.
2. logP less than 5.
3. Not more than 10 hydrogen bond acceptors.
4. Not more than 5 hydrogen bond donors.
5. Less than 10 rotatable bonds and less than 10 chain bonds (non-ring bonds excluding hydrogen bonds).

ZINC [62] database was used in our screening program. The database was filtered according to the filtering criteria mentioned above.

Designing the compound database or library carefully in an efficient way is the key step in a virtual screening campaign. A focused compound library is the one, which is designed for a focused screening program, where focus lies around specific targets of interest. The library is designed in such a way, which is based on the known active compounds against those specific targets of interest. The known active compounds against a class of targets can be collected from literature, PubChem, or Drug Databank.

After selection of the known active compounds, those compounds from the database are further selected which are similar to one or more active compounds. Our compound database for the current virtual high throughput screening was designed based on the molecular similarity searching. The concept of molecular similarity or similarity searching is explained in brief in the following section. The methods used for the similarity searching are also described thereafter.

7.2. Similarity searching

The concept of similarity searching is very simple. Molecular similarity searching is used for searching similar compounds in the chemical structure databases. The chemical structure databases, which contain two-dimensional (2D) or three-dimensional (3D) molecular structures, are stored in flat file formats in these databases.

The term similarity searching means comparing a set of molecular features or descriptors of a molecule to the set of descriptors of the structures in the molecular database. Molecules are characterized by such sets of descriptors, called as representation, is the principle component of similarity searching. These representations are assigned some degree of importance called a weighting scheme on the basis of which similarity coefficients are calculated which quantifies the degree of relatedness in the pair wise comparisons of the molecules. Identification of structurally similar molecules can lead to identify molecules with similar biological activity [63]. This concept of similarity is not just limited to structural similarity but also extends to identification of the bioactive molecules. Molecules, which have similar structures will possess similar properties and therefore will have similar activity. This is referred to as *similar property principle* stated by Johnson and Maggiora [64].

Similarity searching differs from the substructure searching which involves the retrieval of only those molecules from the database, which contain a user defined substructure or a short

structural fragment. Instead the molecules are compared as a whole and the query specified for searching is the entire molecule, the result of which is a list of molecules retrieved from the database and ranked in order of decreasing similarity.

To constrain the number of chemical compounds to only those, which are biologically relevant, different similarity measures were used for our compound library design. The similarity methods and their results are described in the following sections.

7.2.1. Feature Trees

Molecular similarity search was performed using Feature Trees generated for the compound database used in our screening program. FTrees is a computer program available from BioSolveIT [65], which calculates the feature trees descriptors for the molecules in the database on the basis of which similarity measures are performed. The result is a list, which contains similarity indices ranging from 0 to 1. FTrees is very fast and is well suited for searching in large compound databases as well as large combinatorial libraries [67]. It can efficiently calculate and compare the feature trees descriptors for a half million molecules at once.

In feature trees a molecule is represented by a tree called as *feature tree* [66] as shown in Figure 28, which allows a time-efficient computation of molecular similarity. The small fragments of the molecules are represented by the nodes of the tree. The nodes are labelled according to a set of features and represent the chemical properties of the small molecule and the way the nodes are linked together in a feature tree show the way the properties of the molecules linked. As a whole, a feature tree is a non-linear representation of the molecule in which a feature shows a hydrophobic fragment or a functional group of the molecule.

The FTrees program generates feature trees for each molecule in the database and creates tree index files for fast processing. Once a feature tree is generated, it can then be compared with another feature tree of another molecule. Two algorithms; match-search and split-search are used for comparisons of the molecules, which are described in detail in [66].

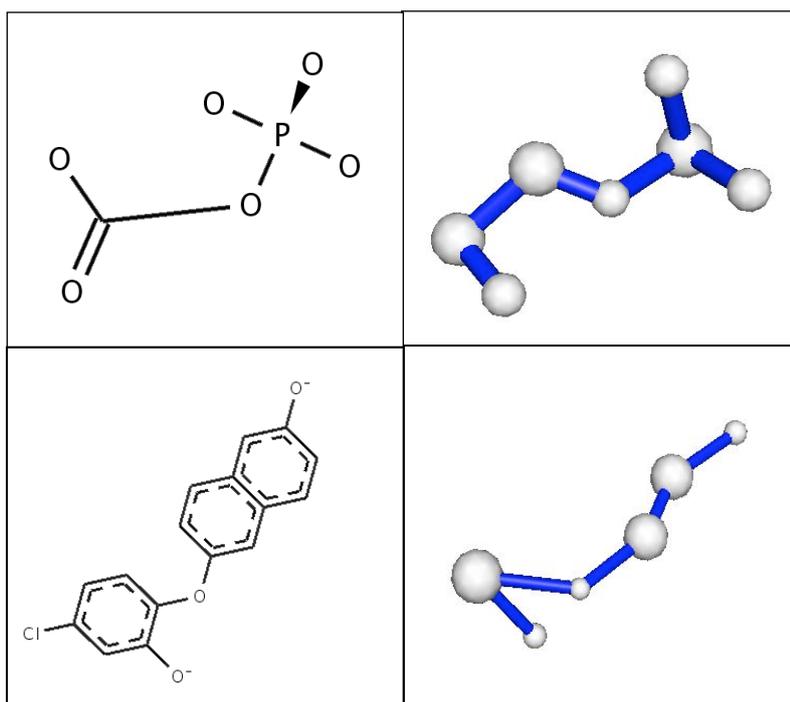


Figure 28. Representation of feature trees generated by FTrees. 2D structures of PGA (top) and TCT (bottom) are shown at left. The corresponding features trees are shown at the right. The nodes (white balls) represent fragments of the molecules, connected by edges (blue rods) which correspond to fragments that are connected in the 2D structure of the molecule.

For similarity searching in our filtered database of small molecules downloaded from ZINC, feature trees were generated by FTrees. On the other hand, feature trees for the 18 known compounds of our target (collected from DrugBank and NCBI-PubChem) were also generated. The compare function of FTrees generated lists for each compound having similarity indices ranging from 0 to 1. The result list for each compound was imported into MySQL database tables. The top similar compounds in the ranks having an Ftree similarity value of 0.98 or above from the database were joined together in one list eliminating the common hits for one or more query compounds. As a result, about 20,000 compounds were selected which has similarities in the range of 0.98 and 1.0 for each of the 18 known query compounds.

7.2.2. Chemical Fingerprints

Similarity comparisons of various molecular descriptors were performed using chemical fingerprints. Chemical fingerprint for a molecule represents a binary valued vector, also

called a bit-vector or a molecular fingerprint which encodes the topological structure of molecular structures into a bit string [68]. These binary valued vectors represent the presence or absence of a molecular feature, i.e 0 means a feature is absent and 1 means a feature is present. These vectors of bits are compared by a distance function or a similarity metric such as Euclidean distance or Tanimoto coefficient. Tanimoto coefficient is most widely used similarity metric [69]. The similarity metrics assess the bit string overlaps for the compounds in the database and are used as similarity measures for molecular similarity [70].

Chemical fingerprints were generated for the compounds in our database using the tool *GenerateMD* from ChemAxon suite of tools [71]. Several parameters of *GenerateMD* can be optimized for generating chemical fingerprints e.g length of the fingerprint bit string for storage of the information. The *ScreenMD* tool was used to screen the database for similar molecules using the Tanimoto Coefficient (Tc) as a similarity metric. The 18 known active compounds were used a query set and screened against the database. The output list of hits contains the compounds, whose 'Tc' value is calculated for each individual query compound. For our desired targeted compound library, the compounds having a 'Tc' value equal to or greater than 0.8 were selected.

7.2.3. MOE Similarity Search

Similarity searching for designing our compound library was performed using fingerprint system in MOE package [49]. There are several fingerprint calculation systems supporting a number of similarity metrics available in MOE package. MDL [72] MACCS structural keys (bit packed) were calculated for the compounds in the database. Structural keys indicate features of a molecule, which indicate the presence of one of the 166 public MDL structural keys. The Tanimoto coefficient threshold was set to 0.6. The fingerprints were compared for each of the query compound and a composite hits list of compounds was selected eliminating the hits common for one or more of the query compounds, as similarly as done with FTrees similarity search.

7.3. Final Compound Database

The results of similarity searches using the above mentioned three methods were different. There were hits ranked higher by one method, which on the other hand ranked lower by the other search method or even was not shown as a hit. Of course, there were hits, which were

picked by all the three search methods. The final compound database was designed in such a way by taking the higher in the ordered ranked list and taking unique from those common to all search methods.

The Tanimoto coefficient is used as a similarity threshold in the fingerprint searching methods, while in FTrees, the similarity value is computed by comparing features of molecules. A similarity threshold of 0.6 or above is generally used in molecular similarity searching. The threshold was set to different levels allowing a compound database of a certain size that should contain highly similar molecules. The compounds having a similarity cut-off of 0.8 and above in the case of ChemAxon fingerprint search were selected. Similarly, the compounds for which the similarity values assigned by FTrees were equal to or greater than 0.98 were chosen and in the case of MOE similarity search all the compound hits were selected with the threshold of 0.6. The compound database contained about 35,000 compounds for use with screening against four PfTIM targets and about 25,000 compounds for screening against two targets of PfENR.

7.4. Grid deployment

Access to a Grid infrastructure and its Grid services typically requires a certain level of authentication and authorisation. The Grid resources such as Computing Elements and Storage Elements are owned by different institutions and managed by independent administrative domains. Before one can actually use a Grid infrastructure, one first needs to request an X.509 user certificate from a Certification Authority (CA). Digital security certificates were obtained which is required for the security features of the UNICORE middleware as access to the Grid resources is generally restricted to a certain user community.

Deployment of the current large-scale Grid-enabled screening project is an important phase. The software applications were installed and configured on each virtual site to enable them run smoothly on the heterogeneous architecture of the Grid infrastructure and execution environment of the computing elements. The software licensing can become a great issue if the software program doesn't allow or support floating licence scheme. The FlexX program is available only with a commercial license. The FlexX license was tested and managed without problems on the test-bed. The license server is contacted from each computing site. The input files were prepared for each target structure for use with the docking applications. The molecules from the final compound database were prepared individually for FlexX and AutoDock and the same database were distributed on the main Storage Elements of the computing sites.

7.5. Optical Testbed Environment

The large-scale vHTS project was deployed on the "Optical Testbed" VIOLA (Vertically Integrated Optical Testbed for Large Applications) [11]. The VIOLA optical testbed environment, as shown in Figure 29, consists of optical network devices from different vendors providing high-speed connectivity between the partner research organisations and universities. These "virtual organisations (VO)" in the Grid infrastructure are connected with Gigabit Ethernet technology. The VOs are also known as UNICORE Sites (USites). A USite is a computer center offering UNICORE server and execution hosts. The USites in our cluster Grid infrastructure include Fraunhofer Institutes at Sankt Augustin, University of Applied Sciences at Sankt Augustin, Center of Advanced European Studies and Research (CAESAR) at Bonn and the Research Center at Juelich, Germany. The USite may further contain one or

more “Virtual Sites” (VSites), which consist of systems at that USite sharing the same data space.

WISDOM (World-wide In Silico Docking On Malaria) was the first large-scale life science application deployed on the Grid infrastructure of the EGEE. The current project is also screening by molecular docking on Malaria, which is an extension of the WISDOM project, and deployed at large scale on the optical testbed.

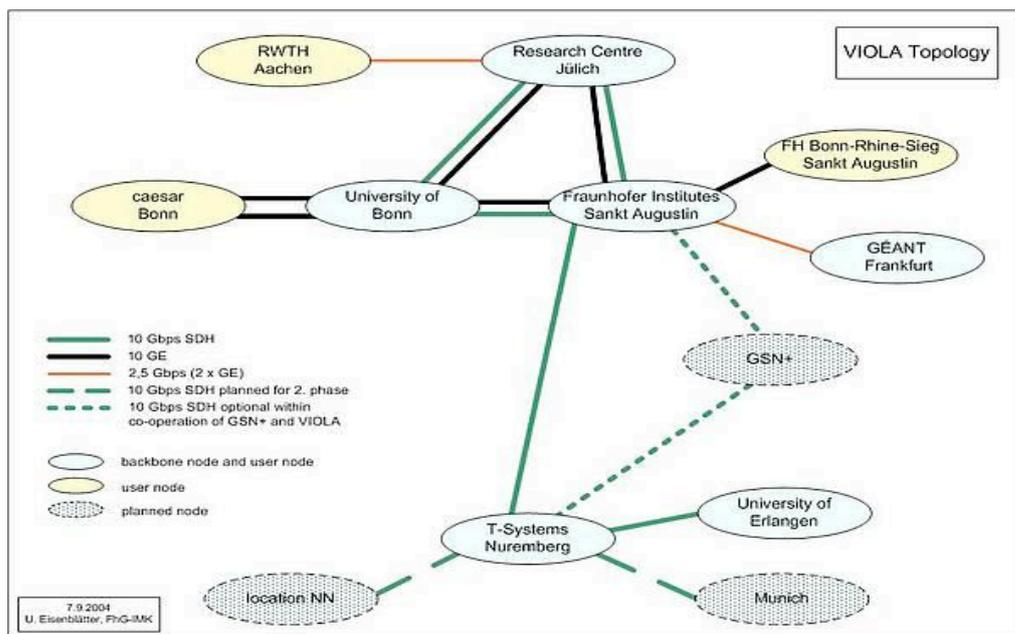


Figure 29. The VIOLA project testbed [11].

As mentioned above, there are four USites which are part of this testbed. The USite at CAESAR was not available because of maintenance reasons. The other USites, which were used during the data challenge, are shown in Figure 30. The red line in the figure shows the 10-Gbit Ethernet of the VIOLA research network providing the high-speed testbed. The execution environment includes the Cray-XD1 Cluster at the Research Center Juelich, which contains 60 dual SMP nodes, each node with 2 processors and 4GB memory per node. The WR-Cluster at the University of Applied Sciences consists of 6 nodes each with 4 processors and shared memory system. The Pack-Cluster at the Fraunhofer Institutes includes 14 dual processor nodes with the 2GB shared memory per node.

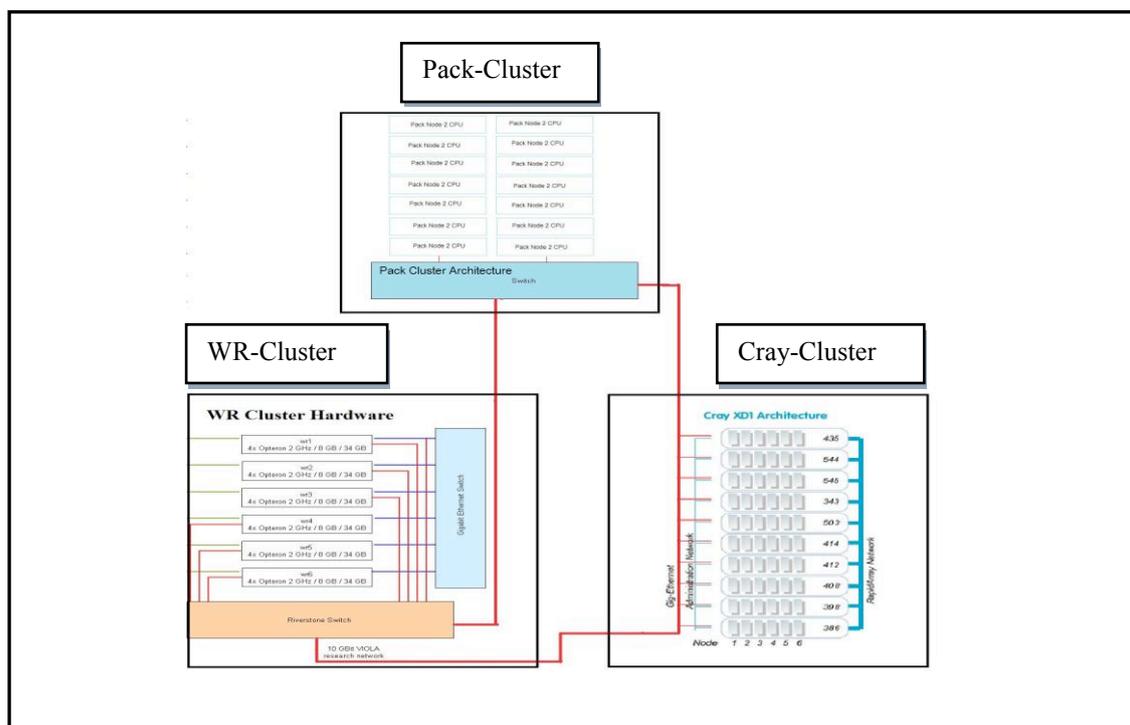


Figure 30. Overview of the Cluster Grid infrastructure of the VIOLA testbed.

7.6. Software Applications Used in the Project:

The main software applications extensively used in our current large scale virtual screening program are molecular docking programs over the Grid middleware, the UNICORE which has been described in detail in the previous section. Several other applications were used during the designing of a compound library for the virtual screening. There are several applications available related to molecular dockings and generally it is difficult to do comparisons on one's test data set. Every docking application has its own algorithm and scoring function. Improved performance can be obtained by using various applications and producing consensus result for the same input. An alternative approach to consensus scoring-AutoX had been used in a Diploma study of Antje Wolf [77], which unifies the interaction models of AutoDock and FlexX, and the results showed that AutoX can improve the docking performance instead of using AutoDock and FlexX alone. Due to the time scope of this thesis, unfortunately, this approach was not used in the current virtual screening

workflow. AutoDock, which is freely available for academic institutions and FlexX, for which we have got a licence at our institute, were used in the screening program and are described below.

Table 16. Molecular docking applications used in the current vHTS project.

Method	Algorithm	Scoring Function	Solvation Scoring
AutoDock	Genetic Algorithm	Empirically Derived	Desolvation Term
FlexX	Incremental Construction	Empirically Derived	---

Table 17. Molecular similarity searching methods used in the current vHTS project.

Method	Algorithm used	Similarity Ranking
FTrees	Feature Trees; split search and match search	Yes
MOE	MDL Structural Keys	No
ChemAxon	Chemical Fingerprints	Yes

Several other applications like Rasmol, OpenBabel and InstantJChem and modules from ChemAxon suit were used. Python and Perl scripts were routinely used during the entire project work. The main software used in this screening by molecular docking are given below.

7.6.1. FlexX:

FlexX Release 2 (version 2.0.2) was used for the screening by molecular docking. A brief description about the FlexX program has been given in section 5.3.1. FlexX is licensed software and we had enough number of licenses to use the program in the testbed environment.

7.6.2. AutoDock:

AutoDock version-4 (release 4.0.1) was used in our virtual screening. AutoDock was obtained from [74] under the GNU general public license. It is well suited for execution in a distributed environment due to no restrictions of the license. AutoDock version 3 (release 3.0.5) was also used in the test runs.

7.6.3. AutoDockTools

AutoDockTools (MGLTools) [56] is a nice graphical front-end for setting up and running AutoDock docking. It is a free graphical user interface developed by the same laboratory that develops AutoDock. Visualisation and analysis of molecular structures and results of the AutoDock dockings are performed with the built-in modules like PMV used in AutoDockTools.

7.7. Test Runs

Several test runs were performed on the Grid testbed in order to test the execution environment of the Grid infrastructure, proper functioning of the UNICORE middleware as well as to test and select the appropriate docking algorithm parameters of the docking applications in use. Deployment of the software applications and the target protein structures with the sets of known active compounds was carried out in the first place before performing the test runs. In the beginning, small UNICORE jobs of FlexX and AutoDock were constructed with the objective to test the UNICORE system such as security certificates setup and license accessibility etc. and the execution environment at the Target System Interfaces (TSI) on each computing site.

Test runs with the docking applications were performed after getting the execution environment ready for the large-scale deployment. This included the parameter setting up for both FlexX and AutoDock docking applications. The results of the test runs were analysed which are presented for FlexX in section 6.3 “Test Runs with FlexX” and for AutoDock in section 6.5 “Test Runs with AutoDock”. As far as the resource requirement is concerned, the test runs also helped in estimating the variable resource requirements at different computing sites.

7.8. The Large Scale Screening

The large-scale deployment of the current virtual high-throughput screening was started soon after the development and setting up the execution environment. As mentioned earlier, the screening is carried out using only two molecular docking applications, FlexX and AutoDock. Different strategies were employed for designing the UNICORE jobs construction and submission systems for the two applications based on the time, workload management and resource requirements. The input compound database and receptor data were split on the computing elements of the testbed for distributed mode of execution.

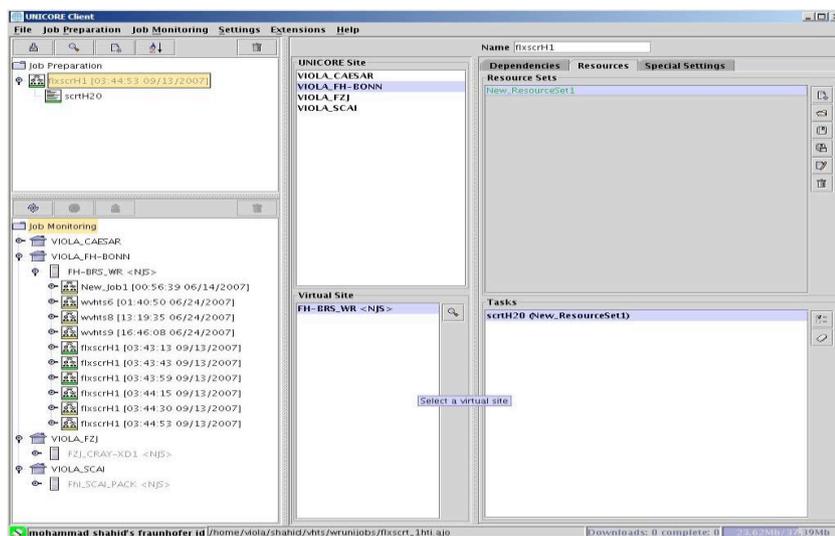


Figure 31. User jobs management with the UNICORE Client.

7.8.1. Jobs Management

The UNICORE client (ver. 5.6) was used for jobs preparation and submission to the clusters. Figure 31 represents a snapshot of the UNICORE client interface showing the users job preparation and monitoring panels, the distribution of jobs to different computing sites and allocation of resources for the jobs. A total of 3,80,000 docking runs were completed using both FlexX and AutoDock in about five weeks. Several thousands of independent UNICORE jobs were prepared and submitted to available computing sites. The duration of a single job varied from one hour to ten hours long containing 30 to 100 docking runs per job. The average time taken by a single docking run was about one minute for docking with FlexX to 30 minutes with AutoDock.

7.9. Problems and Operational Issues

The virtual screening experiment carried out in the execution environment of the VIOLA testbed faced some problems from time to time. These issues are specifically related to Grid services operations. They can generally occur temporarily without leaving great impact on the overall deployment but also cannot be neglected. The hardware problems if occur during the execution of the jobs, lead to the shutdown of the clusters, which results in the loss of jobs

or leaves incomplete jobs. The user needs to manually investigate and resubmit the jobs, which do not start automatically by the resource management system once the system is up and running again. The UNICORE client provides a much easy to operate interface for job submissions and monitoring, although if some problems occur in the UNICORE gateway due to the heavy work load or other reasons, will affect the job monitoring and create instability in the job status update at the client side. Lack of coordination between the UNICORE server, the Resource Broker System and workload management system was noticed. Scheduling policy at a computing site or maintenance shutdown can slow down the overall operation. Every single worker node in the Target System Interface is essential. A node failure can result in the limitation of the available resources. Some times an application can crash e.g AutoDock application crashed for some kind of input data. Post processing, management and analysis of the huge amount of data produced is a time consuming and laborious task in the virtual screening workflow and therefore increase the need for an automated data management approach.

8. Results and Discussion

The results analysis is the most important and demanding task of this large-scale screening program. A huge amount of data were generated and stored in millions of output files produced by the docking applications. Post-processing of the result data was carried out individually for FlexX and AutoDock as their out file formats are different from each other and therefore they were analysed and inspected separately. Different strategies were employed in the analysis of screening results. AutoDock and FlexX implement functions in their own way of scoring and ranking. In general, the strategy used in selection of the compounds has been based on the lowest docking scores, lowest binding free energies, the binding modes of the compounds inside the active site pocket of the receptor proteins and the interactions of compounds functional groups to the key residues of the receptors. The FlexX results filtering strategy has been adopted after the first WISDOM data challenge [78].

Results based on preliminary statistics performed on the virtual screening data are shown in table-18 and table-19 below. The mean, minimum and maximum for FlexX docking scores and AutoDock binding energies for all the target structures of PfTIM and PfENR are presented in separate tables. Results from similarity searching are described in a separate section at the end of the chapter. How molecular similarity searching has improved and aided in the processing of this screening by docking are also presented.

Table 18. Mean, minimum and maximum FlexX scores and AutoDock binding energies for PfTIM targets.

Score/Ene.	FlexX				AutoDock			
	1LYX	1LZO	1O5X	1HTI	1LYX	1LZO	1O5X	1HTI
Min.	-43.792	-42.04	-37.953	-42.831	-7.68	-0.700	-8.30	-8.38
Max.	8.036	6.47	5.764	6.269	-1.11	2.350	-1.78	-1.49
Mean	-14.888	-14.66	-12.346	-13.304	-3.92	0.652	-5.25	-5.08

Table 19. Mean, minimum and maximum FlexX scores and AutoDock binding energies for PfENR targets.

Score/Ene.	FlexX		AutoDock	
	1NHG	1NNU	1NHG	1NNU
Min. Score	-28.528	-28.651	-8.39	-8.330
Max. Score	3.070	3.430	-1.89	-2.260
Mean	-10.879	-11.046	-5.67	-5.714

The overall screening process consisted of docking small molecules libraries against two classes of targets, i.e. PFTIM and PfENR. The result analysis and comparisons are presented for each class with both AutoDock and FlexX.

8.1. Result Analysis of Targets of PFTIM

The docking results for the three target structures of PFTIM as well as one human TIM crystal structure are described below. The mean FlexX scores and AutoDock binding energies has shown variation for the three structures of PFTIM as compared to the structures of PfENR. Such a situation makes the selection process for candidate compounds rather very difficult and increases the chances of false positives hits. The docking results for the target structure 1LZO of the PFTIM enzyme with AutoDock produced results in terms of binding energies that were unsatisfactory. The structure 1LZO, which has the catalytic loop crystallized in open conformation and also with a high resolution could be the possible reason of getting different results. Therefore, the analysis of the screening results was mainly focused on the rest of the target structures. Figure 32 below shows the correlation plots for the FlexX scores and AutoDock binding energies for the structure 1LYX and 1O5X.

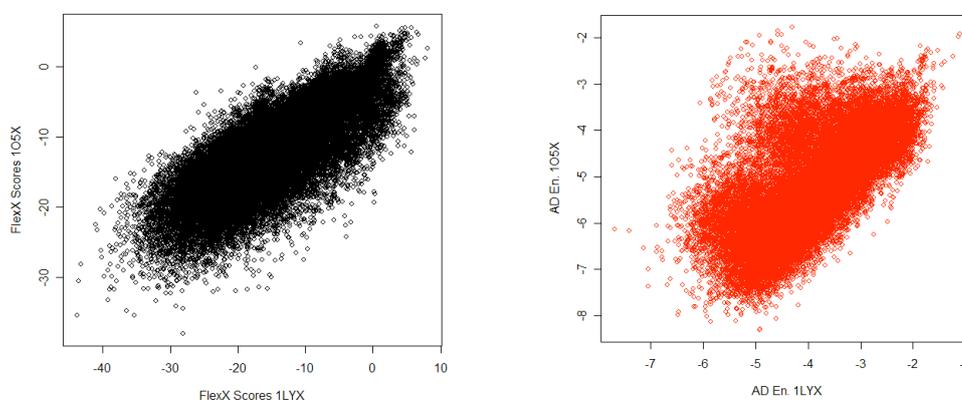


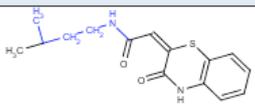
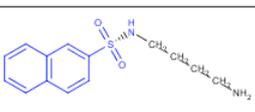
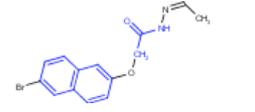
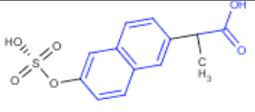
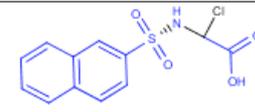
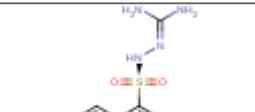
Figure 32. Correlation plots for FlexX scores and AutoDock binding energies comparisons for target structures 1LYX and 1O5X.

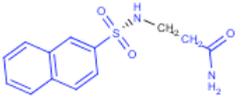
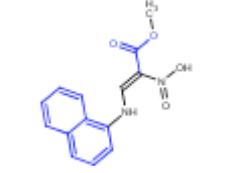
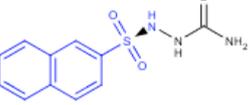
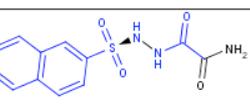
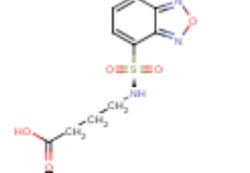
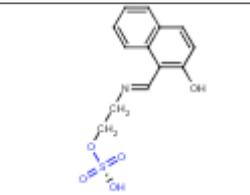
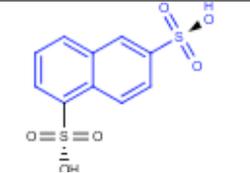
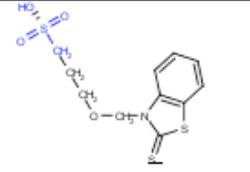
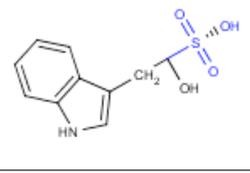
8.1.1. AutoDock Screening Results and Selected Compounds for PFTIM

AutoDock generates a docking log file, which is a flat text file for each docking run. The main strategy employed in analysing docking results with AutoDock was the visual inspection and analysis of individual log files. AutoDockTools was used to analyse the results of each docking. The top ranking compounds based on free energy of binding were first

sorted. Selection for the top ranked compounds was then performed in a consensus way based on good binding modes and fitting of predicted confirmation inside the receptors' active site pockets for the three target structures of PfTIM in comparison with the structure of human TIM. The predicted compounds confirmations were checked for the critical interactions with active site residues of the receptors. The most important interacting residues consist of ASN233, SER211, GLY232, GLY171, LYS12, HIS95, GLU165, ILE70, GLY210, LEU230 and VAL231. The predicted K_i values were also extracted from the docking results. The experimental K_i values for the representative ligand Phosphoglycolate (PGA) for the targets 1LYX and 1LZO is 0.029 mM [30] and 7.4 μ M [33] for the target 1HTI, respectively. Table-20 below presents 2D molecular structures, AutoDock binding energies (B.E) and corresponding predicted K_i values for the selected compounds.

Table 20. AutoDock binding energies (B.E) and predicted K_i values for targets of PfTIM.

ID	2D Structure	1LYX		105X		1HTI	
		B.E	K_i	B.E	K_i	B.E	K_i
1		-8.09	1.170 μ M	-8.63	472.1nM	-7.84	1.78 μ M
2		-7.02	7.150 μ M	-8.37	727.1nM	-7.16	5.68 μ M
5		-7.87	1.690 μ M	-8.79	363.37nM	-7.45	3.49 μ M
7		-9.38	132.8nM	-9.34	141.73nM	-7.43	3.6 μ M
9		-8.39	709.2nM	-9.11	210.8nM	-7.71	2.25 μ M
10		-8.10	1.150 μ M	-9.58	94.64nM	-7.56	2.87 μ M

11		-8.69	429.0nM	-8.7	422.96nM	-8.27	870.1nM
17		-7.81	1.88uM	-8.31	805.13nM	-7.30	4.43uM
18		-8.39	709.22nM	-9.79	66.48nM	-7.51	3.12uM
20		-8.69	430.09nM	-8.74	389.33nM	-7.3	4.46uM
24		-9.13	202.11nM	-9.01	248.7nM	-6.57	15.29uM
25		-7.83	1.84uM	-9.26	163.7nM	-6.14	31.33uM
29		-8.23	926.04nM	-6.75	11.25uM	-5.98	41.41uM
32		-7.49	3.25uM	-8.57	522.8nM	-6.38	21.2uM
36		-8.43	658.18nM	-8.02	1.32uM	-6.35	22.07uM

The selected compounds presented in Table 20 above are displayed in 2D format with the common structural scaffold contained in the compounds highlighted in blue color. The

identified structural moieties extracted in the selected compounds are shown in Figure 33 below. It is evident from these structural moieties that the molecules contain one or more of these functional moieties, which are also part of the active compounds against various TIM targets. Compounds such as Hexane or compounds containing Acetate ion, Sulphate or Phosphate ions, for example, are the important groups contained in the known active compounds (as shown in Figure 34) and are the parts of the selected compounds as well. Most of the selected compounds also contain a two-ring purine like substructures, which mostly interact with Glycine residues in the hydrophobic part of the active site pocket or the catalytic loop-6 of the receptors. Similar active compounds; BTS and FTR are shown the Figure 34 below.

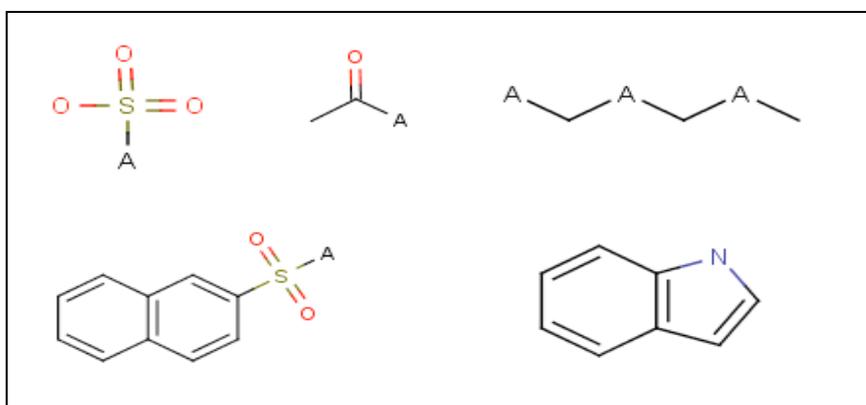


Figure 33. Structural scaffolds contained in the selected candidate compounds.

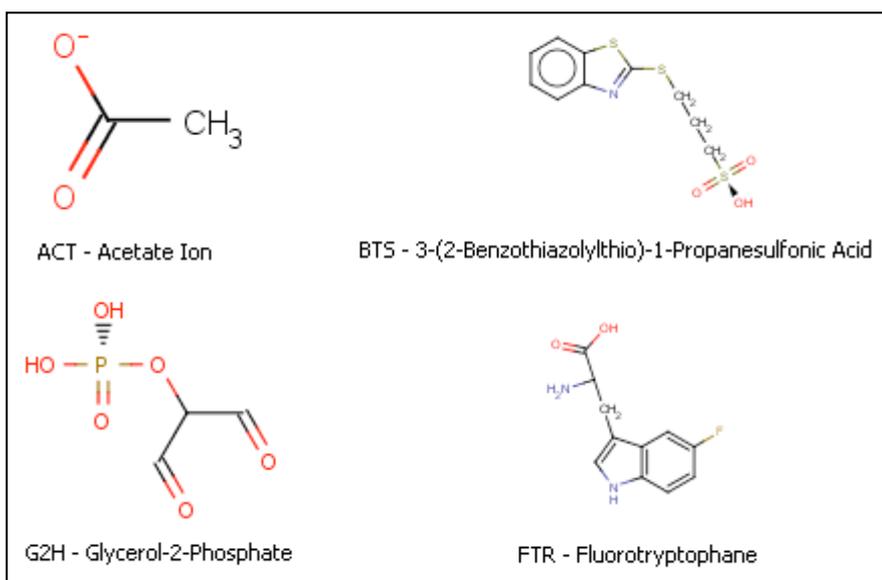


Figure 34. 2D representation of some known active compounds against TIM.

8.1.2. FlexX Screening Results and Selected Compounds for PFTIM

Post-processing of the FlexX results for PFTIM targets was carried out in a different way. The FlexX program produces output in three formats and each format stores different kind of information for a docking run. The docking output log file is also a flat text file, which contains the ranking based on the docking scores of the 10 best generated conformations of a compound. Separate output files are produced for the generated compounds geometries and the interactions of the compounds with active site residues of the receptor. The interaction information, which is also known as “match information”, is the interaction of the molecule atoms with that of the residues of the receptor. The log files were converted to CSV files and the required information was extracted. The strategy in selection of the candidate compounds in the case of FlexX is based on lowest docking score as well as the best binding interactions of the molecules with the critical residues of the receptor. The top ranked molecules were analysed with the help of FlexV, a molecular graphical application which is a part of the FlexX program.

A good docking score produced for a compound may not be the may not indicate the good binding modes for a compound. Sorting and ranking of the FlexX docking results was carried out based on the FlexX scores and based on the match information separately. Table-23 below shows that top ranked compounds for the target structure 1LYX based on lowest docking scores, are not covering the maximum match information to become the best binders. The top 25 compounds having the best FlexX scores with the corresponding ranking based on the interaction information are presented in this table. On the other hand, those compounds having good interaction with the important catalytic residues of the receptor protein do not get the good docking scores.

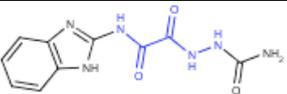
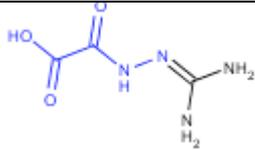
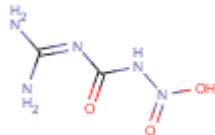
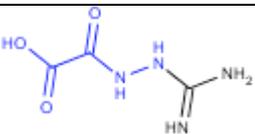
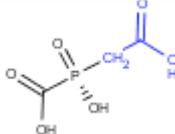
Table 21. Comparison of ranks based on FlexX scores and FlexX interaction information for the target 1LYX.

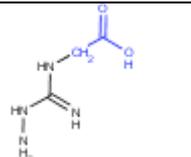
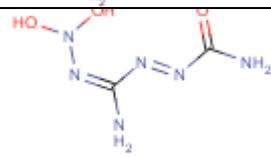
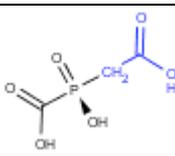
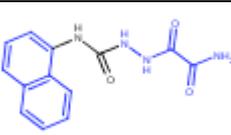
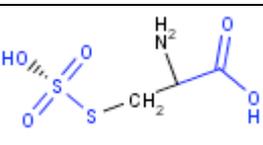
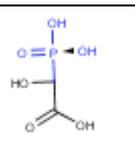
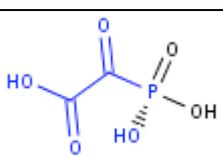
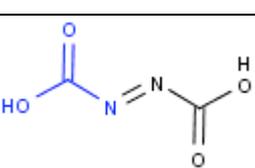
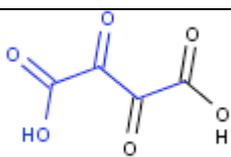
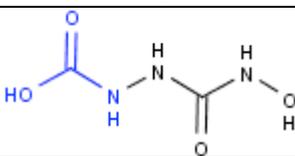
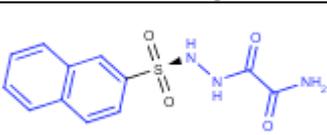
No.	Mol. ID	FlexX Score	S-Rank	Conf#	M-Rank	Mol-ID	FlexX Score	M-Rank	S-Rank
1	ZINC05399850	-43.792	1	6	35	ZINC015297	-28.625	1	1434
2	ZINC05409999	-43.621	2	1	3852	ZINC021120	-29.404	2	1083
3	ZINC05112634	-43.274	3	2	3695	ZINC056476	-29.074	3	1215
4	ZINC00084475	-41.105	4	6	60	ZINC058904	-28.224	4	1650
5	ZINC03954150	-40.835	5	2	3361	ZINC059423	-28.892	5	1311
6	ZINC04091440	-40.435	6	1	1501	ZINC060416	-25.255	6	3775
7	ZINC01485648	-40.281	7	1	2443	ZINC014929	-25.534	7	3555
8	ZINC05828162	-40.148	8	8	1851	ZINC015338	-28.148	8	1691
9	ZINC01561686	-39.8	9	7	1062	ZINC015375	-16.818	9	14274

10	ZINC04091443	-39.775	10	1	1502	ZINC015524	-26.924	10	2475
11	ZINC01605036	-39.233	11	1	340	ZINC015595	-28.501	11	1500
12	ZINC06886088	-39.228	12	5	46	ZINC016126	-34.438	12	134
13	ZINC04096920	-39.072	13	7	3407	ZINC016630	-23.895	13	5073
14	ZINC00125901	-39.052	14	4	862	ZINC017691	-31.055	14	605
15	ZINC01561449	-38.943	15	1	332	ZINC018690	-28.448	15	1532
16	ZINC03869650	-38.807	16	9	1447	ZINC021120	-28.441	16	1537
17	ZINC04403894	-38.758	17	1	3504	ZINC021639	-23.953	17	5003
18	ZINC05783646	-38.726	18	1	705	ZINC025698	-30.001	18	887
19	ZINC05220232	-38.672	19	1	3780	ZINC025846	-27.416	19	2156
20	ZINC05684039	-38.583	20	1	695	ZINC026524	-36.537	20	51
21	ZINC06576470	-38.288	21	1	237	ZINC030794	-24.962	21	4042
22	ZINC01670852	-38.185	22	1	372	ZINC033291	-31.026	22	611
23	ZINC05828173	-38.15	23	10	4066	ZINC038732	-27.226	23	2291
24	ZINC02560511	-38.07	24	1	3078	ZINC038733	-30.351	24	776
25	ZINC00125899	-38.063	25	1	861	ZINC040052	-29.313	25	1117

Therefore, selection and filtering of the candidate compounds were carried out separately based on the above mentioned strategies. The sorted list of compounds obtained from the ranking done based on the FlexX docking scores for all the targets of PFTIM were analyzed at first as given in table-24 below.

Table 22. FlexX scores and ranks for the selected candidate compounds for the targets 1LYX, 105X and 1HTI.

ID	2D Structure	1LYX		105X		1HTI	
		Score	Rank	Score	Rank	Score	Rank
1		-43.79	1	-35.376	3	-39.18	3
2		-43.62	2	-30.46	37	-34.83	16
3		-43.27	3	-28.07	150	-33.39	31
4		-40.84	5	-23.18	1341	-40.29	2
5		-40.44	6	-20.31	3496	-26.08	903

6		-40.28	7	-26.14	376	-32.78	39
7		-40.15	8	-27.17	233	-27.56	508
8		-39.78	10	-20.81	3010	-26.29	831
10		-39.23	12	-33.23	8	-36.78	8
11		-39.07	13	-25.48	522	-25.82	999
14		-38.73	18	-19.87	3948	-23.92	1881
17		-38.15	23	-22.855	1510	-26.41	787
19		-37.53	33	-20.67	3135	-23.14	2407
20		-37.46	34	-24.18	906	-27.69	482
25		-36.67	50	-26.65	293	-31.86	73
43		-36.46	56	-34.74	4	-32.39	52

The selected candidate compounds are represented in 2D display format in the table above with common structural moieties highlighted. After identifying the common structural scaffolds in the top scoring compounds, it is evident that most of the selected candidate compounds are composed of such one or more substructural moieties (as shown in Figure 35) which are also found as substructures present in the known active compounds against the targets of TIM enzyme. Interactions of these functional groups with the critical active site residues have resulted in the prediction of the best docking scores with FlexX. The Sulphate and Acetate ions have also been identified earlier in the selected compounds with AutoDock. Here in the case of screening with FlexX, Hydroxypyruvic Acid, Acetate ion and Sulphate ion containing compounds have been ranked higher in the list. 3-Hydroxypyruvic Acid (3PY) is a known active compound as shown in Figure 35, and similarly 2-Phosphoglyceric Acid (2PG) and 3-Phosphoglyceric Acid (3PG) which are the analogs of the Phosphoglycolic Acid (PGA) among the set of known active compounds.

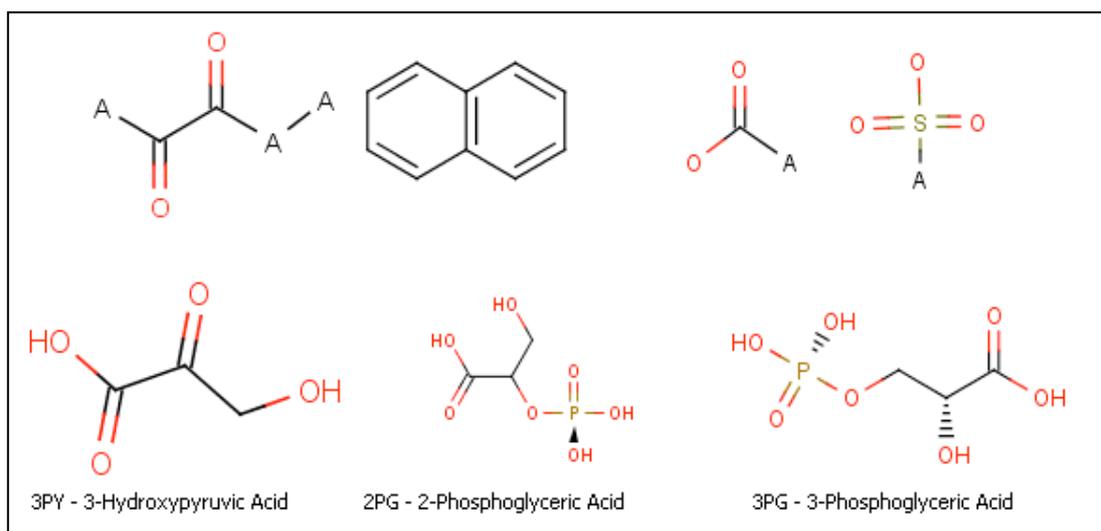


Figure 35. Common structural moieties contained in the selected candidate compounds for targets of PfTIM using FlexX and some of the known active compounds of TIM.

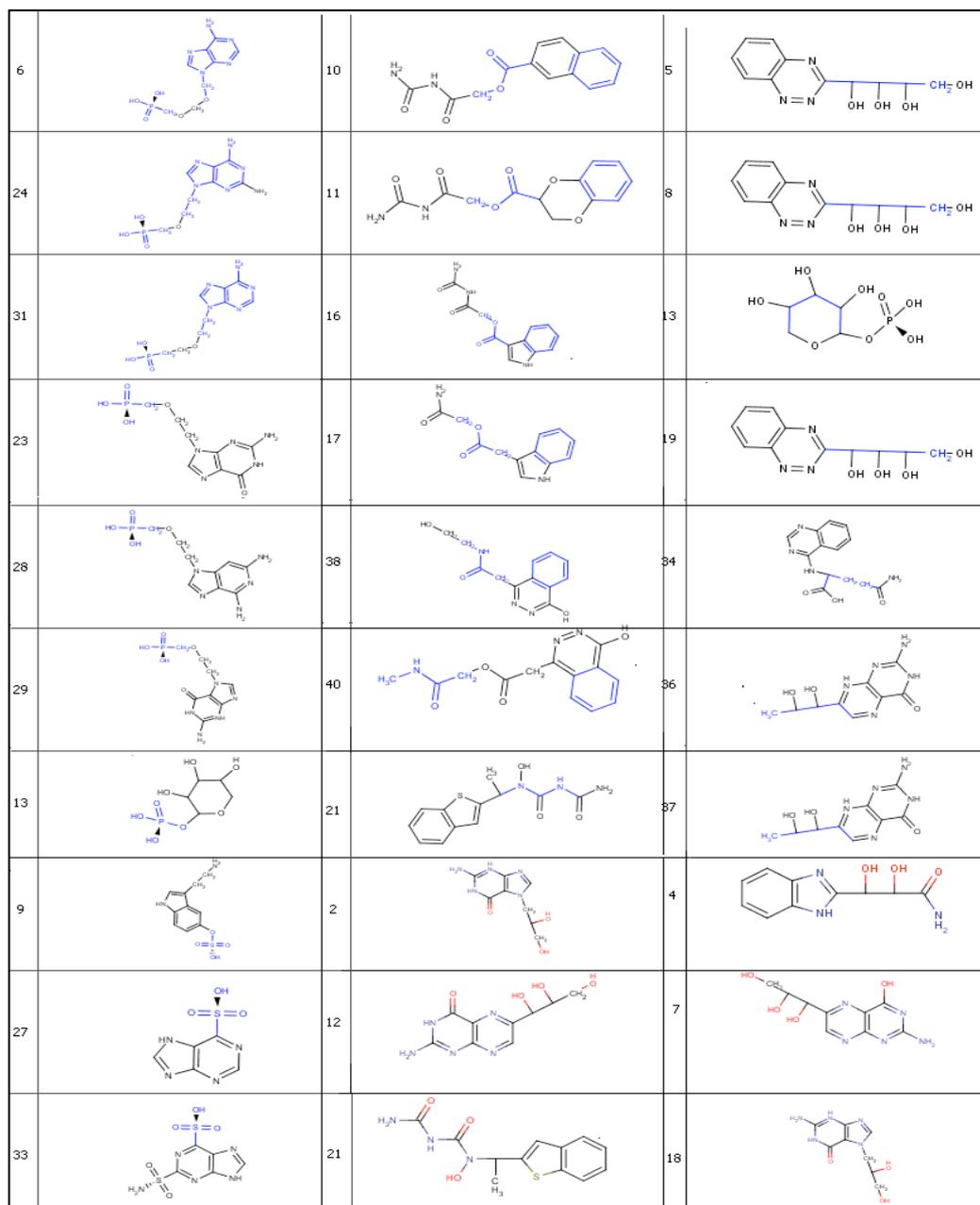


Figure 36. 2D Structural representation of selected compounds for PfTIM based on FlexX Match and Interaction information.

The compounds having good interactions with the important receptor residues which are ranked higher with AutoDock screening (as shown in Table 22) and those ranked higher based on FlexX interaction information (as shown in Figure 36) are found having similar molecular representations and interaction behaviour in the active site pocket of the receptors. The common double ring substructure of these molecules stay at the opening of the pocket forming hydrophobic interactions and interaction with Glycine residues in the opening of the active site pocket while the remaining substructure in the molecules take part in the hydrogen bonding interactions with the rest of the residues of the active site.

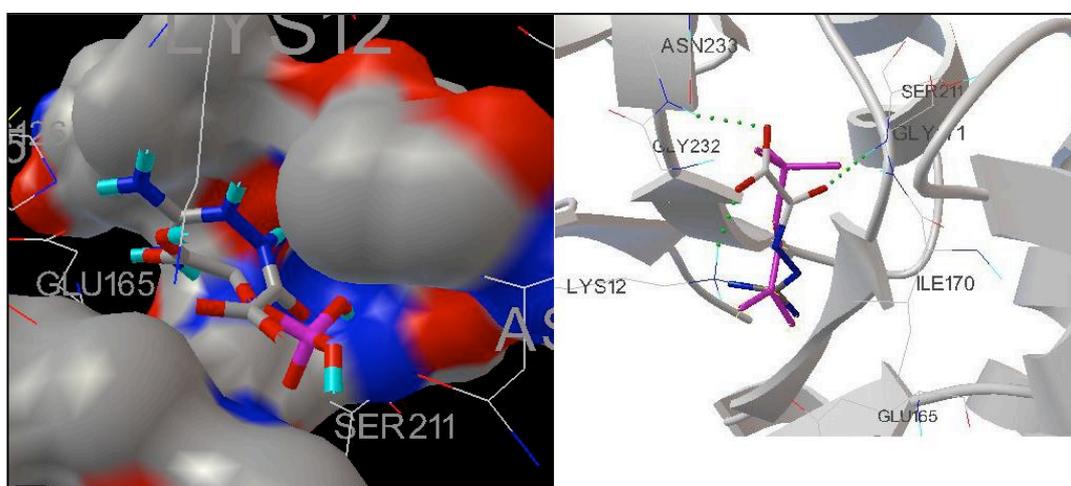


Figure 37. A compound ranked higher based on FlexX scoring in the active site pocket of 1LYX together with the cocrystallized ligand PGA (shown in magenta).

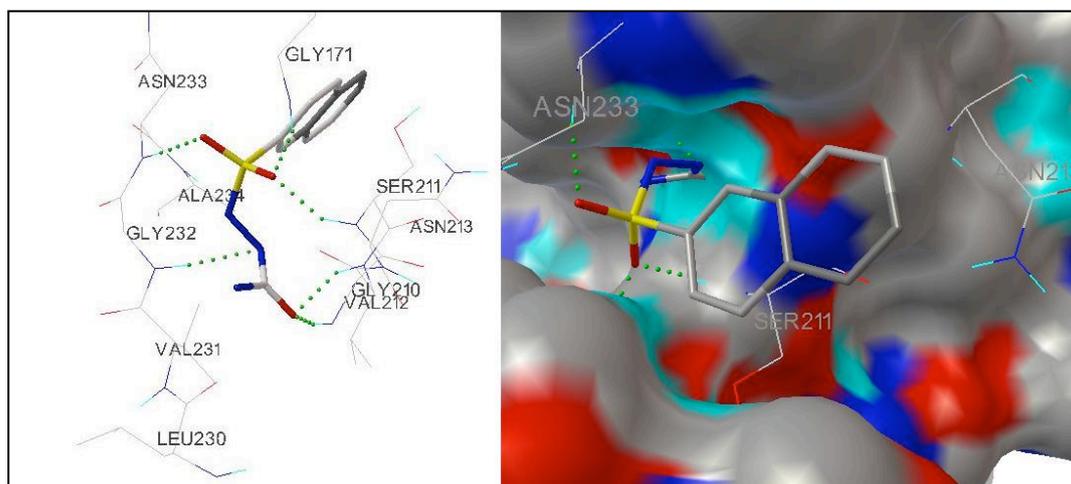


Figure 38. AutoDock predicted conformation and interactions of a candidate compound in the active site pocket of 1LYX.

8.2. Result Analysis of Targets of PfENR

The screening results for PfENR targets are presented separately in this section. The small compound database was screened against two target structures 1NHG and 1NNU. In comparison to the target structures of PfTIM, the selection and filtering for the candidate compounds was relatively straightforward. The FlexX scores and AutoDock binding energies correlation plots between the two target structures 1NHG and 1NNU are given in the Figure 39 below. The reason for the good correlation in docking scores for the compound library could be due to the nearly same resolutions of the two targets' crystal structures, a good structural alignment with lower RMSD and no active site loop dynamics. Therefore, the compounds having specific interactions with the key residues and the cofactor NAD in one structure had the similar interactions in the other structure as well. Analysis of the result data was performed in a similar way like that of the targets of PfTIM and the same strategies were employed for selection and filtering of the screened database of the molecules. The results for the targets of PfENR are presented below with both AutoDock and FlexX.

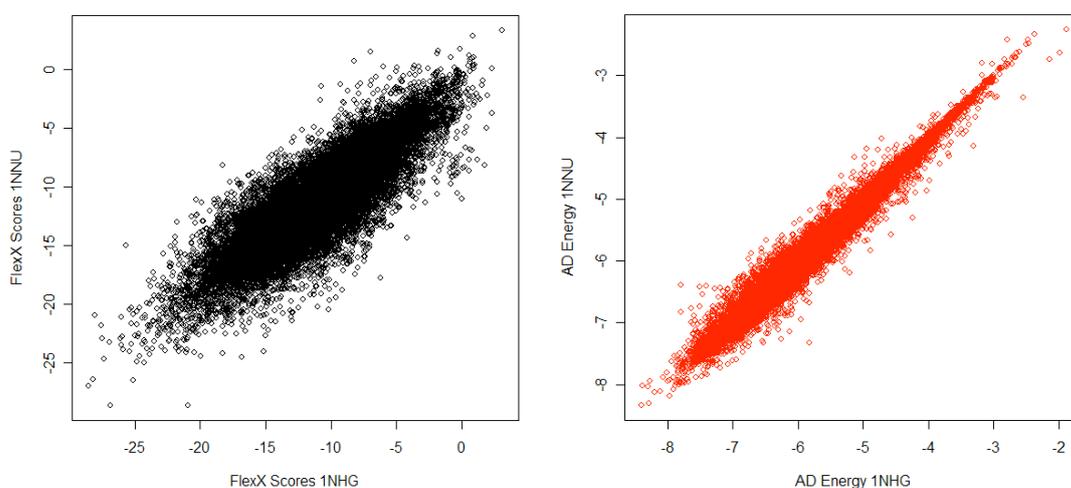


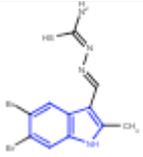
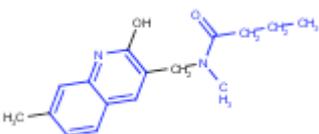
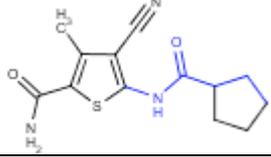
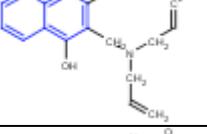
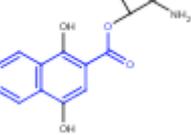
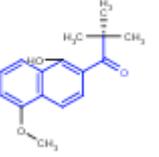
Figure 39. FlexX scores and AutoDock binding energies correlation plots between targets structures 1NHG and 1NNU.

8.2.1. AutoDock Selected Compounds for Targets of PfENR

The table- below contains the list of selected candidate compounds based on lower binding energies predicted by AutoDock for the two target structures. AutoDockTools application was used to visually inspect the top ranked docked conformations of the molecules for each target structure. The predicted geometries of the molecules were analysed for the important

interactions in the active site pocket and with the cofactor. The important residues of the active site pocket include TYR277, TYR267, LYS285, ILE323, ALA320, ALA319 and ALA217 in which the hydrogen bonding interaction of TYR277 with the ligand atom is conserved in most of the ENR enzymes from various species. AutoDock binding energies and the predicted K_i values for the selected candidate hits are presented in table-25 below. Overall, the selected compounds have almost similar binding energies and predicted K_i values for both target structures of the PfENR as predicted by AutoDock, but the affinity values are predicted higher than the experimental values. The experimental affinity values for TCL, the representative ligand of the target structure 1NHG, has been reported at 0.05 μM [54] and for some TCL analogs as 0.15 μM .

Table 23. AutoDock binding energies and predicted K_i values for targets of PfENR.

No.	Mol-ID	1NHG			1NNU		
		B.E	K_i (μM)	Rank	B.E	K_i (μM)	Rank
2		-8.19	1.04	6	-8.12	2.76	4
4		-7.97	1.76	12	-7.93	2.66	22
5		-7.94	1.6	13	-7.62	2.84	94
7		-7.84	47.7	25	-7.59	8.69	120
8		-7.83	3.79	28	-7.93	5.00	21
9		-7.83	1.88	29	-7.81	3.89	37

10		-7.8	5.22	34	-7.96	2.18	16
11		-7.79	3.98	36	-6.82	28.34	2264
12		-7.78	3.26	39	-7.97	5.08	14
15		-7.75	2.6	48	-7.83	1.82	33

The candidate compounds listed in the above table are examined along with the known active compounds against the ENR enzymes for the binding modes and critical interactions. This lead to the identification of some common pharmacophore features. Several substructural scaffolds which are present in these selected candidate compounds (as shown in 2D representations in Table 25) possess similarities to some of the known actives as a whole (as shown in stick diagram representations in Figure 43). For example, Ethionamide (2-ethylpyridine-4-carbothioamide) or Isoniazid (pyridine-4-carbohydrazide) like compounds are present as substructural moieties in the ranked hits. Some of the candidate compounds share the similarity in substructural compositions to the known active compounds (Figure 43).

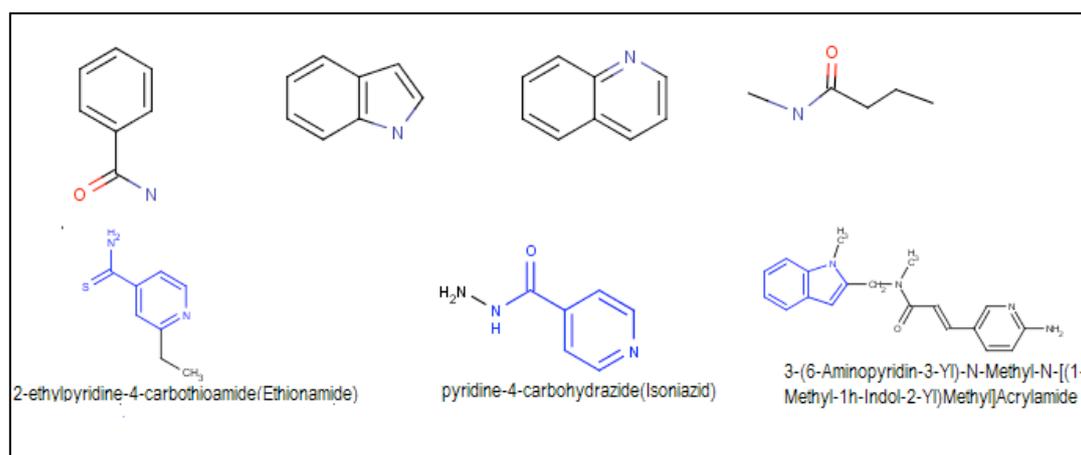
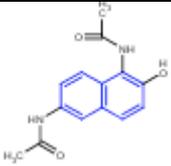
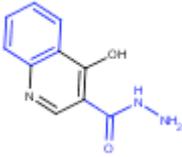
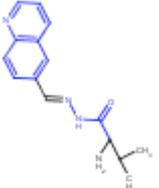
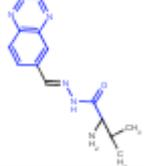
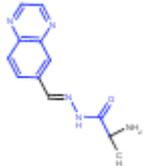
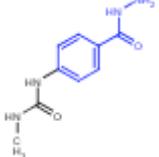


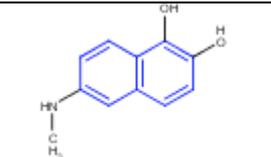
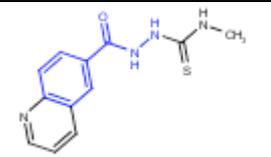
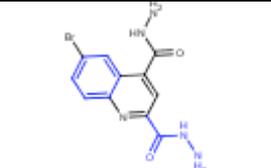
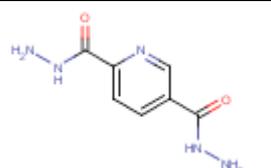
Figure 40. Identified common scaffolds in the selected candidate compounds and diagrams of some of the known active compounds.

8.2.2. FlexX Selected Compounds for Targets of PfENR

Post processing of the virtual screening with FlexX for the targets of PfENR was performed in the same way as described earlier. The top ranking compounds were analysed by looking at the best 10 predicted solutions and conformations for each hit. Analysis for the good binding modes and interactions in the active site of the receptors was done using both target structures. The table-26 below includes the selected compounds in the top ranked compounds based on good FlexX scores along with the rankings based on lowest FlexX score for the two targets.

Table 24. . FlexX scores and ranks for the selected candidate compounds for the PfENR targets 1NHG and 1NNU.

ID	2D Structure	1NHG		1NNU	
		Score	Rank	Score	Rank
1		-27.572	4	-21.864	92
2		-27.496	5	-22.972	44
3		-27.352	6	-24.716	8
4		-26.084	9	-22.931	45
5		-26.065	10	-23.528	28
7		-25.239	19	-21.708	105

8		-24.825	24	-20.669	204
10		-24.301	32	-23.098	41
11		-23.501	55	-23.175	35
15		-23.84	46	-19.979	305

The interactions between the ligands and receptor were also investigated from the match information produced for a docking experiment. Some compounds listed in the above table-26 which are ranked higher based on FlexX scoring were also among the top ranked list based on the FlexX match information between the molecule and the key residues of the receptor. Again, no noticeable differences were observed between the FlexX scores for the two target structures which helps in the identification of the true positive hits. 2D structural diagrams are shown in table-26 and Figure 45 for the selected compounds based on FlexX scoring and FlexX match information, respectively.

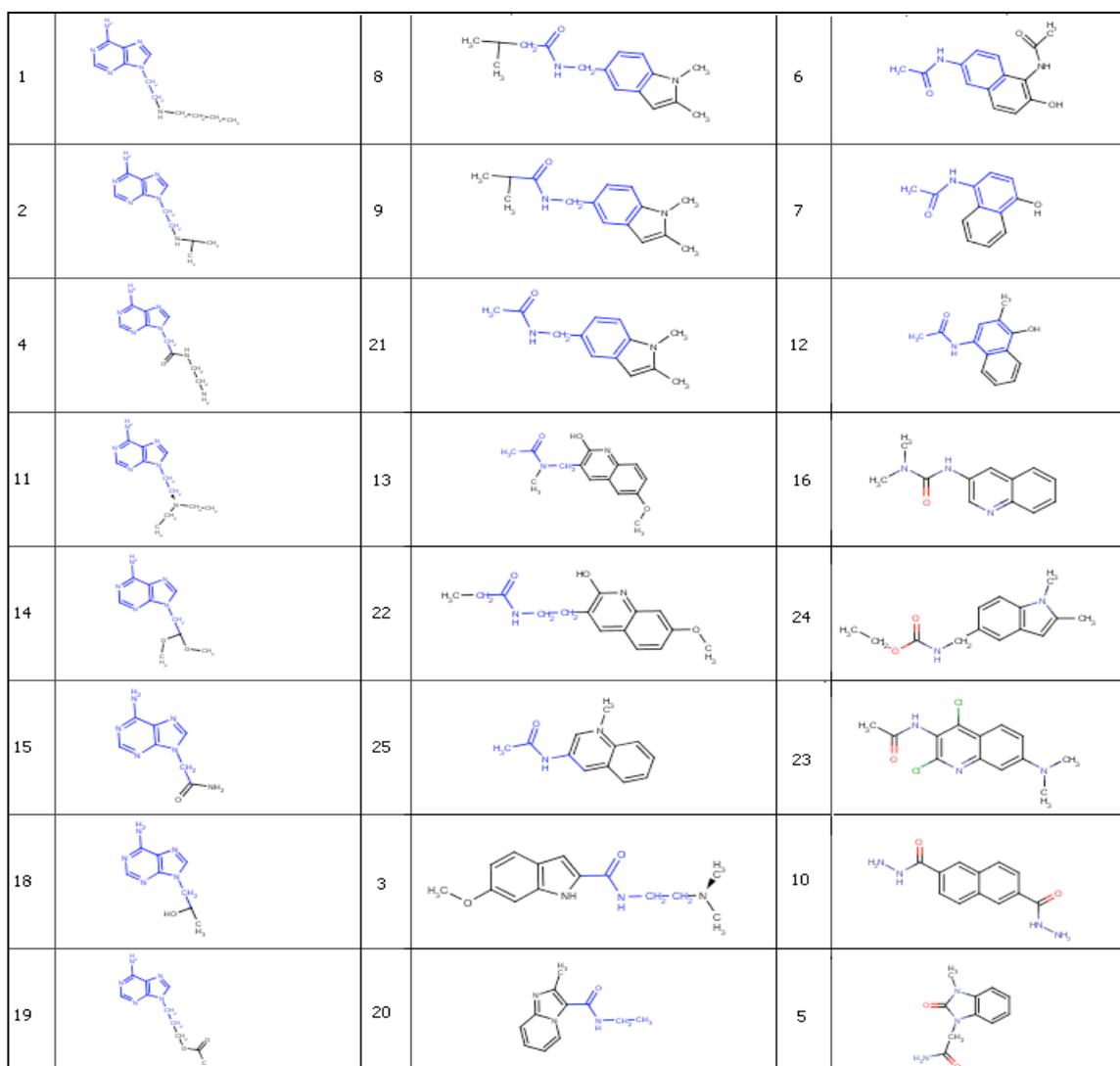


Figure 41. 2D structural representation of the selected compounds for the two targets of PfENR based on FlexX interaction information.

The substrate binding pocket of the PfENR is built up from hydrophobic residues such as TYR277, TYR267, ILE323, GLY313 and PRO314, where the chains of fatty acids accommodate for further elongation. The selected candidate molecules using FlexX as displayed in the above table also form mainly hydrophobic interactions and they are constituted from similar substructural moieties. Moreover, the pharmacophoric nature of these molecules also exhibits similarities to the candidate hits predicted by virtual screening

with AutoDock. The identified substructural moieties are shown in Figure 46 below with some of the known actives, which provide a rule of validation of these compounds as candidate hits and can more likely result in the form of competitive inhibitors of the PfENR enzyme.

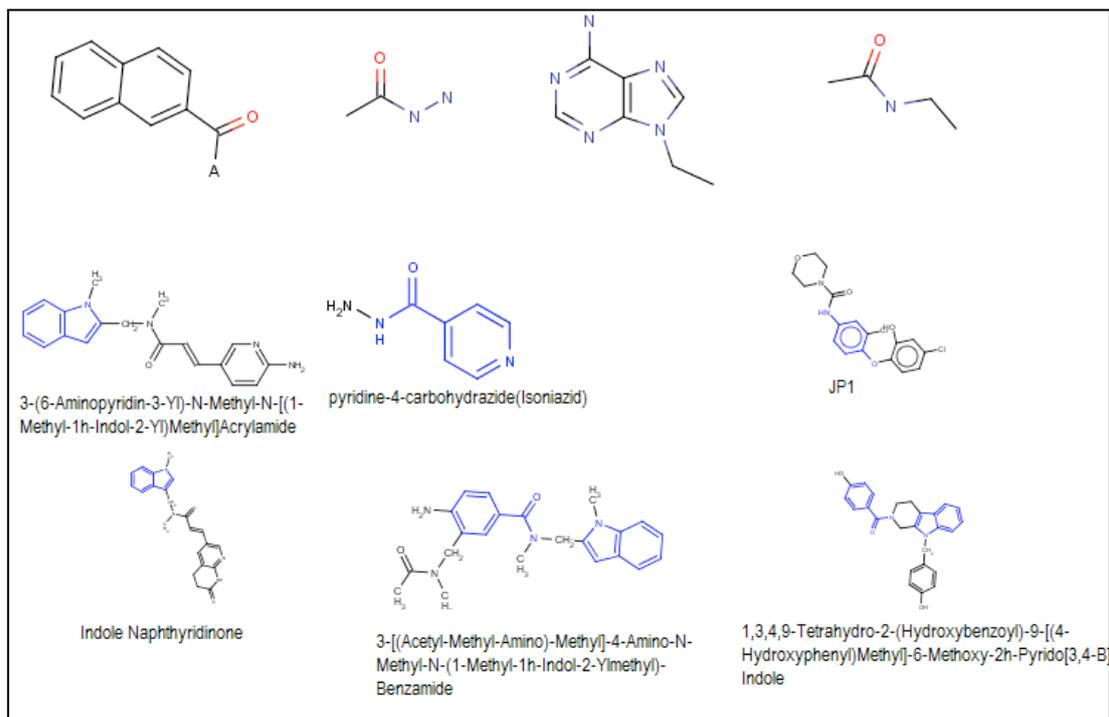


Figure 42. Identified common structural moieties in selected candidate compounds of FlexX screen and some example structures of known active compounds.

8.3. Screening Results and Molecular Similarity Analysis:

An overview of the molecular similarity search results is presented in the tables below. The number of similarity hits covered by different similarity search methods in the top 100 ranked compounds for the targets of PfTIM and PfENR with AutoDock and FlexX. This shows that for our test set of targets and the input database, the top ranked compounds are similar and the similarities are calculated more by the methods MOE Fingerprints and Feature Trees (FTrees) than the similarity screening with ChemAxon.

Table 25. FlexX top ranking solutions based on scoring & match information for PfTIM and percentage of hits from similarity search methods.

FlexX Sim. Search	Score based			Matching based		
	1LYX	1O5X	1HTI	1LYX	1O5X	1HTI
MOE	24	30	32	31	20	20
FTrees	78	73	76	73	79	79
ChemAxon	13	11	11	2	2	2

Table 26: FlexX top ranking solutions based on scoring & match information for PfENR and percentage of hits from similarity search methods:

FlexX Sim. Search	Score based		Matching based	
	1NHG	1NNU	1NHG	1NNU
MOE	81	85	96	96
FTrees	22	16	0	2
ChemAxon	0	0	4	3

Table 27: AutoDock top ranking solutions based on lowest binding energy for PfTIM and percentage of hits from similarity search methods.

AutoDock Sim. Search	Binding Energy based				
	1LYX	1O5X	1HTI	1NHG	1NNU
MOE	68	1	1	63	65
FTrees	50	63	76	37	37
ChemAxon	3	36	24	0	0

The above tables show that in FlexX screening, the ranked compounds based on FlexX scoring and interaction information for all the targets of PfTIM were picked as more similar by FTrees than MOE similarity search. For the targets of PfENR this was not the same and the reason could be that the similarity threshold was set to a higher level in FTrees. Similar is

the case in AutoDock screening, which shows that the results of similarity measures can be different, based on the input test set.

As mentioned before, the selected compounds with AutoDock screening (as shown in Table 22) and those ranked higher based on FlexX interaction information (as shown in Figure 36) are more similar than the compounds selected on the basis of FlexX docking scores. FTrees similarity searching was performed for a set of 4 known compounds of PFTIM against the above three sets of selected candidate compounds. It was shown that the compounds resulted in AutoDock screening were more similar to those resulted in filtering the compounds based on FlexX match score (Figure 43). The known compounds used in this similarity searching were Phosphoglycolate (PGA), 3-Phosphoglyceric Acid (3PG), Fluorotryptophane (FTR) and 3-(2-Benzothiazolylthio)-1-Propanesulfonic Acid (BTS).

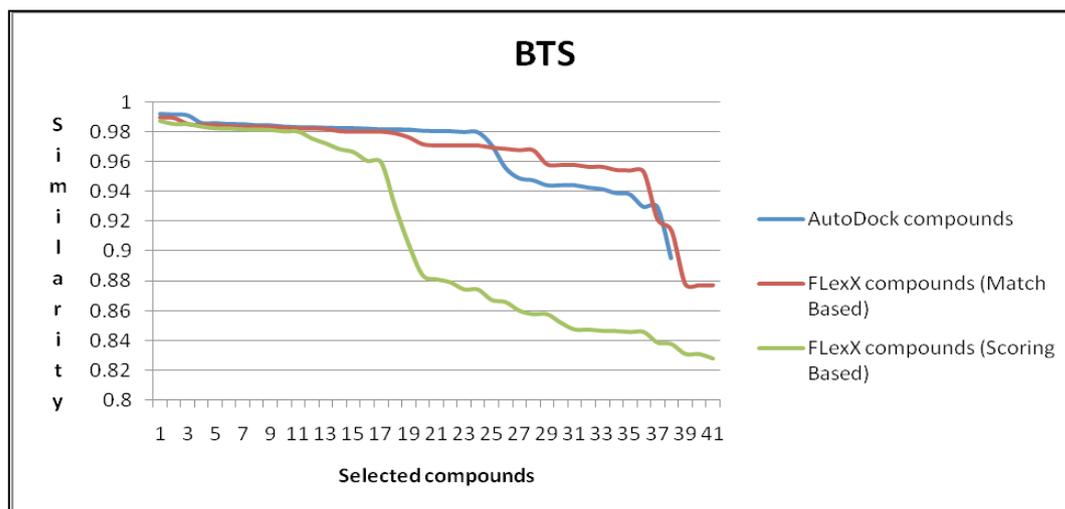


Figure 43. Similarity comparison of BTS compound with three sets of selected candidate compounds.

8.4. Discussion

The large scale screening by molecular docking was carried out against novel malarial drug targets, resulting in a set of new candidate compounds. The work undergone was another attempt after the WISDOM project on different sets of drug targets, execution environment and software applications. Similarity searching, that plays a major role in a high throughput screening, was performed in designing a compound database for screening against the two targets of *Plasmodium falcifarum*. If the whole available compound databases are screened in a high throughput screening, again the filtering is done at the end. Although, such approaches are useful in a diversity analysis, but has certain limitations such as require a lot of resources and tools. Therefore, efficient designing of a compound database has a pivotal role in a virtual screening. Three different similarity measures were used in developing a compound database for our large scale screening. Results of similarity searching showed differences (as presented in section 8.3), therefore, it was quite helpful to use various similarity methods instead of a single one and avoid false negative hits of compounds.

Two different docking software, FlexX and AutoDock, were used in the process of this screening by molecular docking. Both software produced different results. There is a need to improve performance by using various applications and combine the results in a consensus way. An alternative approach, AutoxX (mentioned in section 7.6) can be used to increase the performance of results. Employing different approaches in result analysis (as mentioned in section 8) can be much helpful in formulating filtering criteria, on the basis of which a substructure searching in the available compound databases can result in compounds having similar activities. The compounds can be studied further using molecular modelling techniques in the compound optimization stage. Interestingly, it was also observed that the top ranking compounds predicted by AutoDock have similarities to those filtered on the basis of FlexX match scores (as shown in section 8.3). Another observation regarding the screening results of FlexX is that, for the targets of PfTIM, FlexX predicted compounds (as presented in Table-22 in section 8.1.2) similar to most of the known active compounds (as shown in Figure 35 in section 8.1.2) or sharing common functional groups, which can result more likely in the form of competitive inhibitors.

Successful completion of this large scale virtual screening was made possible using computational Grid. As the WISDOM project was deployed on the world wide EGEE Grid with its own execution environment and gLite Grid middleware, this virtual screening was

deployed using UNICORE in an environment providing high speed connectivity allowing fast data transfer between the computing sites. UNICORE provides a more secure and easy-to-use interface for submitting and monitoring Grid jobs. This helps reduce the complexities faced by a normal user operating in the Grid environment. As compared to the issue of jobs success rate, failure and re-submission issue in the WISDOM environment, the rate of jobs failure was relatively less except failures that occurred due to hardware problems. Due to such failures, the manual jobs re-submission was a similar problem in this environment.

8.5. Summary and Conclusions

The essence of Grid computing and its cognizant relatedness to the long going process of drug discovery have been addressed in the current project. Grid computing provides an infrastructure for such a large-scale deployment of data and processing intensive tasks. The use of computational tools and methods deployed on the newly adopted Grid technology will result in new drug candidates more quickly and at lower cost in the preclinical period of the complex drug discovery program.

The current large-scale virtual high throughput screening was deployed on the VIOLA testbed providing high speed connectivity. This screening by molecular docking program was another attempt to combat the Malaria disease after the WISDOM deployment on a public Grid infrastructure. The emerging resistance to the currently available and effective antimalarials necessitates the selection for new potential and validated biological drug targets which can be used in a screening program and produce new drug candidates. In this study a small molecule database was designed and screened against a set of new drug targets involved in energy metabolism and fatty acid biosynthesis of the malarial parasite.

Two different docking software tools FlexX and AutoDock were used on the same input data in the screening campaign. The input compound database was designed using three different similarity search measures. The reason for this is to invest additional efforts and test alternatives in predicting the true positive hits. The screening project was deployed on a Grid infrastructure, which availed the significant advantage of Grid computing without which this large-scale screening project would have been completed in years instead of weeks. Several novel candidate compounds were identified and recommended for further optimization and testing in the experimental screening.

Grid services and operations were tested and found successful in the deployment of such a data intensive project. Although some difficulties and minor operational issues were noticed, which are reported in this experiment. Moreover, a suitable alternative approach is needed to efficiently manage and analyse quickly the time consuming and huge amount of output data.

9. References

- [1]. Nwaka S., Ridley R. Virtual Drug discovery and development for neglected diseases through public-private partnership. *Nat. Rev. Drug Discov*, 2003, 2: 919-928.
- [2]. Yeh I., Altman R. B. Drug Targets for *Plasmodium falciparum*: A post genomic Review/Survey. *Mini-reviews in medicinal chemistry*, 2006, 6: 177-202.
- [3]. Shoichet B. K. Virtual Screening of Chemical Libraries. *Nature*, 2004, 432: 862.
- [4]. Kitchen D. B., Decornez H., Furr J. R. and Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 2004. 3: 935-949.
- [5]. Lee H. C., Salzemann J., Jacq N., Chen H.Y., Ho L.Y., Merelli I., Milanesi L., Breton V., Lin S.C. and Wu Y.T. Grid Enabled High-Throughput *In Silico* Screening Against Influenza A Neuraminidase. *IEEE*, 2006, 5: 4, 288-295.
- [6]. Brooijmans N., Kuntz, I. D. Molecular Recognition And Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.* 2003, 32:335–73.
- [7]. BioinfoGRID: Bioinformatics Grid Application for life science.
(<http://www.bioinfoGrid.eu/>).
- [8]. Mehlin C. Structure-Based Drug Discovery for *Plasmodium falciparum*. *Combinatorial Chemistry & High Throughput Screening*, 2005, 8: 5-14.
- [9]. Buyya R., Branson K., Giddy J., Abramson D. The Virtual Laboratory. A Toolset to Enable Distributed Molecular Modeling for Drug Design on the World-Wide Grid. *Concurrency Computat: Pract. Exper.* 2003, 15: 1-25.
- [10]. WISDOM: “Initiative for Grid-enabled drug discovery against neglected and emerging diseases”. (<http://wisdom.healthGrid.org/>).
- [11]. VIOLA (Vertically Integrated Optical Testbed for Large Applications in DFN), (2004-2006). <http://www.viola-testbed.de/>.
- [12]. Drews J. Drug Discovery: A Historical Perspective. *Science*. 2000, 287: 5460. 1960-1964.

- [13]. Jorgensen W. L. The Many Roles of Computation in Drug Discovery. *Science*. 2004, 303: 5665. 1813-1818.
- [14]. Foster I., Kesselman C. Computational Grids. *"The Grid: Blueprint for a New Computing Infrastructure"*, Morgan-Kaufman Publishers, 1999, ISBN: 1558604758.
- [15]. World Health Organization (WHO). Fact sheet on malaria, 2007.
(<http://www.who.int/topics/malaria/en/>).
- [16]. World Malaria Report 2005. World Health Organization. (<http://rbm.who.int/wmr2005/>).
- [17]. National Center for Infectious Diseases, Division of Parasitic Diseases, 2006.
(<http://www.cdc.gov/malaria/>).
- [18]. Report on global monitoring, 2004. Susceptibility of Plasmodium Falciparum to Antimalarial Drugs. World Health Organization (WHO/HTM/MAL/2005.1103).
- [19]. Bloland, P. B. Drug Resistance in Malaria. World Health Organization, 2001.
(WHO/CDS/CSR/DRS/2001.4).
- [20]. Wellems, T. E. Plasmodium Chloroquine Resistance and the Search for a Replacement Antimalarial Drug. *Science*, 2002. 298: 5591. 124-126.
- [21]. EGEE (Enabling Grids for E-science). (<http://www.eu-egee.org/>).
- [22]. Foster I. What is the Grid? A Three Point Checklist. *GRIDToday*, July 20, 2002.
- [23]. Berman F., Fox G, Hey A. J. G. The Grid: Past, Present, Future, in *Grid Computing: Making The Global Infrastructure a Reality*, Chichester (UK): John Wiley & Sons Inc., 2003.
- [24]. Foster I., Kesselman C., Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International J. Supercomputer Applications*, 2001. 15 (3).
- [25]. Berman F., Anthony J.G. Hey, Fox G. *Grid Computing: Making The Global Infrastructure a Reality*". John Wiley & Sons; (April 8, 2003), ISBN: 0470853190. 65-100.
- [26]. Clery D. Can Grid Computing Help Us Work Together? *Science*, 2006. 313: 5786. 433.
- [27]. Open Grid Forum (OGF) [Online]. (<http://www.ogf.org>).

- [28]. Globus Toolkit. [Online]. (<http://www.globus.org/toolkit/>).
- [29]. Romberg, M. The UNICORE Grid Infrastructure. *Sci, Program*, 2002, 10. 149-157.
- [30]. gLite [Online]. (<http://glite.web.cern.ch/glite/>).
- [31]. Computer Services for Academic Research (CSAR). (<http://www.csar.cfs.ac.uk/>).
- [32]. Litzkow, M. J. et al. Condor – A Hunter of Idle Workstations. *Proceedings of the 8th International Conference of Distributed Computing Systems (1998)*. 104-111.
- [33]. Cluster Resources, Inc. (<http://www.clusterresources.com/>).
- [34]. EUROGRID. (<http://www.eurogrid.org/>).
- [35]. UniGridS. (<http://www.unigrids.org/>).
- [36]. Romberg M. The UNICORE Architecture: Seamless Access to Distributed Resources. *IEEE*, 2002, 22:43: 287-293.
- [37]. Erwin D. UNICORE Plus Final Report - Uniform Interface to Computing Resources. Joint Project Report for the BMBF Project UNICORE Plus, Grant Number: 01 IR 001 A-D, 2002, ISBN 3-00-011592-7.
- [38]. *Proceedings of the 8th IEEE International Symposium on High Performance Distributed Computing (HPDC-1999)*, Redondo Beach, USA, IEEE Computer Society Press, 1999, 287-293.
- [39]. Birkholtz L.M., Bastien O., Wells G., Grando D., Joubert F., Kasam V., Zimmermann M., Ortet P., Jacq N., Saïdan N., Roy S., Hofmann-Apitius M., Breton V., Louw A.I., Maréchal E. *Malaria Journal*, 2006. 5:110.
- [40]. Gowthaman R., Sekhar D., Kalita M.K., Gupta D. A database for Plasmodium falciparum protein models. *Bioinformatics*, 2005. 1(2): 50-51.
- [42]. Berman, H.M., et al.. The Protein Data Bank, *Nucleic Acid Res.* 2000. 28, 235-242.
- [43]. Yeh, I., Altman R.B. Drug Targets for Plasmodium falciparum: A Post-Genomic Review/Survey. *Mini-Reviews in Medicinal Chemistry*, 2006. 6: 177-202.
- [44]. Ravindra G., Balaram P. Plasmodium Falciparum Triosephosphate Isomerase: New insights into an old enzyme. *Pure Appl. Chem.* 2005. 77, 1:281–289.

- [45]. Parthasarathy S., Ravindra G., Balaram H., Balaram P., Murthy M. R. N. Structure of the Plasmodium Falciparum Triosephosphate Isomerase-Phosphoglycolate Complex in Two Crystal Forms: Characterization of Catalytic Loop Open and Closed Conformations in the Ligand-Bound State". *Biochemistry*, 2002, 41, 13178-13188.
- [46]. Parthasarathy S, Eaazhisai K, Balaram H, Balaram P, Murthy M.R. Structure of Plasmodium falciparum triose-phosphate isomerase-2-phosphoglycerate complex at 1.1-Å resolution. *J. Biol. Chem.* 2003, 278(52):52461-70.
- [47]. Velanker S.S., Ray S.S., Gokhale R.S., Suma S. , Balaram H., Balaram P., Murthy M.R. Triosephosphate Isomerase from Plasmodium Falciparum: the crystal structure provides insights into antimalarial drug design. *Structure*, 1997, 5, 6:751-761.
- [48]. Mande S.C., Mainfroid V., Kalk K.H., Goraj K., Martial J.A., Hol W.G. Crystal structure of recombinant human triosephosphate isomerase at 2.8 Å resolution. Triosephosphate isomerase-related human genetic disorders and comparison with the trypanosomal enzyme. *Protein Sci.* 1994, 3(5):810-21.
- [49]. Molecular Operating Environment; Chemical Computing Group Inc.: Montreal, Quebec, Canada (2005). <http://www.chemcomp.com>.
- [50]. Russell R.B., Barton G.J. STAMP: multiple protein sequence alignment from tertiary structure comparison. *Proteins*, 1992, 14:309–323.
- [51]. Wade R.C., Davis M. E., Lu B. A., McCammon J. A. A commentary on gating of the active site of triose phosphate isomerase: Brownian dynamics simulations of flexible peptide loops in the enzyme. *Biophys. J. Biophysical Society*, 1993, 64: 1-2
- [52]. Knowles, J. R. Enzyme catalysis: not different, just better. *Nature*, 1991, 350:121–124.
- [53]. Ranie J., Kumar V. P., Balaram H. Cloning of the triosephosphate isomerase gene of Plasmodium falciparum and expression in Escherichia coli. *Mol. Biochem. Parasitol.*, 1993, 61:159–169.
- [54]. Perozzo R., Kuo M., Sidhu A.S., Jacob T., Valiyaveetil, Bittman R., William R., Jacobs, Jr., David A., Fidock, Sacchettini J.C. Structural Elucidation of the Specificity of the Antibacterial Agent Triclosan for Malarial Enoyl Acyl Carrier Protein Reductase. *J. Biol. Chem.*, 2002. 277, 15:13106–13114.

- [55]. Pidugu L.S., Kapoor M., Surolia N., Surolia A., Suguna K. Structural Basis for the Variation in Triclosan Affinity to Enoyl Reductases. *J. Mol. Biol.* 2004, 343:147–155.
- [56]. MGLTools from Molecular Graphics Laboratory, The Scripps Research Institute. (<http://mgltools.scripps.edu/>).
- [57]. Wishart D.S. et al., DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 2006, 1:34.
- [58]. PubChem. (<http://pubchem.ncbi.nlm.nih.gov/>).
- [59]. Mehlin C. Structure-Based Drug Discovery for Plasmodium falciparum. *Combinatorial Chemistry & High Throughput Screening*, 2005, 8, 5:14 5.
- [60]. Lu J.Z., Lee P.J, Waters N.C., Prigge S.T. Fatty Acid Synthesis as a Target for Antimalarial Drug Discovery. *Combinatorial Chemistry & High Throughput Screening*, 2005, 8:15-26.
- [61]. Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* 1997, 23:3–25.
- [62]. ZINC (<http://zinc.docking.org/>).
- [63]. Willet P., Bernard J.M., Downs G.M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* 1998, 38: 983-996.
- [64]. Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
- [65]. BiosolveIT GmbH. (<http://www.biosolveit.de/FTrees>).
- [66]. Rarey, M and Dixon J.S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput. Aided Mol. Des.* 1998, 12(5): 471–490.
- [67]. Rarey, M and Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aided Mol. Des.* 2001, 15(6): 497–520.

- [68]. Schneider G., Clement-Chomienne O., Hilfiger L., Schneider P., Kirsch S., Bohm H.J., Neihart W. Virtual Screening for Bioactive Molecules by Evolutionary De Novo Design *Angew. Chem. Int. Ed.* 2000, 39: 4130-4133.
- [69]. Raymond, J. W., Willett, P. Effectiveness of graph-based and fingerprint based similarity measures for virtual screening of 2D chemical structure databases. *J. Comput. Aid. Mol. Design*, 2002, 16: 59–71.
- [70]. Godden, J.W., Stahura, F.L., Bajorath, J. Anatomy of Fingerprint Search Calculations on Structurally Diverse Sets of Active Compounds. *J. Chem. Inf. Model.* 2005, 45(6): 1812-1819.
- [71]. ChemAxon. (<http://www.chemaxon.com>).
- [72]. MDL Information System Inc. (<http://www.mdl.com>).
- [73]. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. Automated docking using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* 1998, 19: 1639–1662.
- [74]. The Scripps Research Institute. (<http://autodock.scripps.edu/>).
- [75]. FlexX. (<http://www.biosolveit.de/FlexX/>).
- [76]. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J Mol Biol.* 1990, 215(3): 403-410.
- [77]. Wolf A., Zimmermann M., Hofmann-Apitius M. Alternative to Consensus Scorings: A New Approach Toward the Qualitative Combination of Docking Algorithms. *J. Chem. Inf. Model.* 2007, 47:1036-1044.
- [78]. Kasam V., Zimmermann M., Maass A., Schwichtenberg H., Wolf A., Jacq N., Breton V., Hofmann-Apitius M. Design of New Plasmepsin Inhibitors: A Virtual High Throughput Screening Approach on the EGEE Grid. *J. Chem. Inf. Model.* 2007, 47:1818-1828