PREDICTION OF POTENTIAL INHIBITORS AGAINST *YERSINIA* YOPE AND *SALMONELLA* SOPE BY VIRTUAL SCREENING AND DOCKING

by

Gizem Özbüyükkaya

B.S., Chemical Engineering, Boğaziçi University, 2010

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering
Boğaziçi University
2012

ACKNOWLEDGEMENTS

It is my pleasure to thank all the people who made this thesis possible. Foremost, I would like to thank my supervisor Prof. Kutlu Ülgen and my co-supervisor Assist. Prof. Elif Özkırımlı Ölmez for the wise advice, guidance and encouragement they provided at all levels of this research. I am grateful for writing my thesis with their expertise and support.

I would like to thank the other members of my committee, Prof. Viktorya Aviyente, Prof. Türkan Haliloğlu and Assist. Prof. Barkın Berk for their revision and thoughtful comments they provided on this thesis.

I wish to thank all of my friends and colleagues with whom I have shared experiences in graduate studies. My sincere gratitude goes to Seval Aladağ for providing her knowledge and help on this work, whenever I needed. Thank you, Seval. I am tempted to individually thank all other KB 407 and KB 440 members for their friendship, entertainment and emotional support, especially Deniz Menekşedağ, Begüm Alaybeyoğlu, Ceyda Kasavi and Oya Gürsoy. It was really fun working with you guys. I also wish to thank everyone I know from childhood to this day, for sharing experiences and memories in life. Since it would be too long to list all the names, I will simply say thank you very much and I love you all. Anyway, some of you are lucky: İrem, Bahadır and Ferit, thank you for being such good companies. And dearest Cihan Kaya, despite words are unnecessary, I thank you for standing by me for all these years.

Above all, I would like to express my deep gratefulness to my parents. They bore and raised me, loved and supported me unconditionally, taught me and cared for me. I wish to make them proud every day, and I dedicate this thesis to my wonderful parents.

Financial support provided by TUBITAK through project 111M444 is gratefully acknowledged.

ABSTRACT

PREDICTION OF POTENTIAL INHIBITORS AGAINST *YERSINIA* YOPE AND *SALMONELLA* SOPE BY VIRTUAL SCREENING AND DOCKING

Outer proteins of gram negative bacteria *Yersinia* and *Salmonella* secrete virulence factors that invade eukaryotic cells via type III secretion system. *Yersinia* outer protein E, YopE, and *Salmonella* outer protein E, SopE, are cytotoxic and pathogenic for humans, and they target small GTPases of Rho family. Rho family of GTPases acts as molecular switches in many cell signaling pathways that affect actin skeleton arrangement, cell motility and apoptosis. Interaction of YopE or SopE with Rho GTPases induces cytoskeletal disruption and depolymerisation of actin stress fibers within the cell, and can lead to many diseases, including cancer. In this work, to the aim was to discover novel inhibitors that would block biological functioning of the bacterial pathogen *Salmonella* SopE and *Yersinia* YopE, with the help of computational drug design tools. To this end, small molecule database formation, 3D database screening with pharmacophore building, molecular docking and scoring were carried out to propose a set of biologically active YopE and SopE inhibitors. For target YopE, quantitative structure-activity relationship (QSAR) method was incorporated into ligand-based pharmacophore building; whereas structure-based pharmacophores were constructed for target SopE. 3D pharmacophore models were constructed, and database generated from ZINC containing 25 million conformers was pre-filtered using the selected models. Molecular docking of filtered compounds was carried out with Schrödinger software Glide, taking ligand flexibility into account. The accuracy of each docking was presented with GlideScore and ranked accordingly. Top ranking results were further analyzed according to their interactions with target, druglikeness and bioavailability. Ultimately, a total of ten molecules were proposed as potent inhibitors against YopE and SopE, among which Y1 (ZINC02736077) and S1 (ZINC00370772) showed the highest predicted binding affinity towards YopE and SopE, respectively.

ÖZET

POTANSİYEL *YERSINIA* YOPE VE *SALMONELLA* SOPE KARŞITI İNHİBİTÖRLERİN SANAL TARAMA VE DOKLAMA İLE TAHMİNİ

Gram negatif bakteri *Yersinia* ve *Salmonella*, tip III salgı sistemi üzerinden virülans faktörlerini ökaryot hücre içlerine aktarırlar. Bu bakterilerin efektör proteinlerinden *Yersinia* dış protein E, YopE ve *Salmonella* dış protein E, SopE, proteinleri efektör proteinlerini salgılamak için Rho GTPasez ailesi proteinlerini hedefler; bu yüzden insan için sitotoksik ve patojeniklerdir. Rho GTPaz ailesi proteinleri hücre sinyalleşmesini düzenleyerek aktin iskelet düzenlenmesi, hücre hareketliliği ve apoptoz gibi birçok hücresel fonksiyonu yerine getiren bir anahtar olarak görev alır. YopE ve SopE'nin salgılanması ve Rho GTPazlarla etkileşimi hücre aktin liflerinin depolimerizasyonuna, iskeletin bozulmasına ve sinyal mekanizmasının değişmesine yol açarak, kanser dahil birçok hastalığa yol açabilir. Bu çalışmada, hesapsal ilaç tasarım yöntemleriyle bakteriyel patojen proteinler *Salmonella* SopE ve *Yersinia* YopE'nin biyolojik fonksiyonunu engelleyecek yeni inhibitörlerin keşfi amaçlanmıştır. Bu doğrultuda; küçük molekül veritabanı oluşumu, farmakofor model geliştirilmesi ve 3B veritabanı taraması, moleküler doklama ve skorlama ile YopE ve SopE proteinlerinin fonksiyonlarını bloke edebilecek bir dizi biyolojik aktif molekül önerilmiştir. YopE proteini için oluşturulan ligand tabanlı farmakofor modeline kantitatif yapı-etki ilişkileri (QSAR) yöntemi dahil edilmiştir. SopE proteini içinse hedef tabanlı farmakofor modeli oluşturulmuştur. Oluşturulan 3B farmakofor modelleri ile 25 milyon konformerlı ZINC veritabanı ön filtreden geçirilmiştir. Farmakofor filtresinden geçen moleküller, Schrödinger yazılımı Glide ile hedef proteinlere, ligand esnekliğine izin vererek, iki basamakta doklanmıştır. Doklanan her molekül ve konformer GlideScore ile skorlanmış ve sıralanmıştır. Yüksek skorlu sonuçlar ilaçbenzerliği, biyoyararlanım ve hedef ile moleküler etkileşimler baz alınarak analiz edilmiştir. Sonuç olarak, hedef proteinlerine yüksek ilgi gösterebilecek toplam on molekül, başta Y1 (ZINC02736077) ve S1 (ZINC00370772) olmak üzere, potansiyel YopE ve SopE inhibitörü olarak önerilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a, A	Number of actives
Br	Bromine
C	Carbon
Cl	Chlorine
Da	Dalton
F	Fluorine
H	Hydrogen
I	Iodine
N	Nitrogen
n	Sample size
N	Total number of decoys and actives
O	Oxygen
Q^2	Pearson correlation coefficient
r	Radius
R^2	Correlation of determination
S	Sulfur
X, Y, Z	Three dimensional Cartesian coordinate axes
Å	Angstrom
α	Alpha
β	Beta
γ	Gamma
ΔG, DG	Binding free energy

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two dimensional
3D	Three dimensional
A	Hydrogen bond acceptor
Ala	Alanine
AccptHB	Number of hydrogen bond acceptors
ADME	Absorption, distribution, metabolism, and excretion
Arf	Alternative reading frame
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
AUC	Area under the receiver-operating characteristic curve
avg	Average
BTB	Bric-a-brac domain
Cdc42	Cell division control protein 42 homolog
Coul	Coulomb energy
CPU	Central processing unit
Cys	Cysteine
D	Hydrogen bond donor
DBL	Duffy-binding-like domain
DH	Dbl homology domain
DHR2	Dock homology region 2
DonorHB	Number of hydrogen bond donors
EF	Enrichment factor

EGFR	Epidermal growth factor receptor
ExoS	Exoenzyme S
F	Variance ratio
FISA	Hydrophilic component of the total surface accessible area
FOSA	Hydrophobic component of the total surface accessible area
G, Gly	Glycine
GAP	Gtpase-activating protein
GDI	Gdp dissociation inhibitor
GDP	Guanosine driphosphate
GEF	Guanine exchange factor
Gln	Glutamine
Glu	Glutamic acid
GTP	Guanosine triphosphate
H	Hydrophobic group
Hbond	Hydrogen bonding energy
HIV	Human immunodeficiency virus
His	Histidine
HOA	Human oral absorption
HP	Hewlett-Packard
HTS	High-throughput screening
HTVS	High-throughput virtual screening
IC_{50}	Half maximum inhibitory concentration
IUPAC	International Union of Pure and Applied Chemistry
Leu	Leucine
Lipo	Lipophilic energy
LogP	Partition coefficient

Lys	Lysine
Met	Methionine
MIF	Molecular interaction fields
Min	Minimized
MM-GBSA	Molecular mechanics/generalized born surface area
MW	Molecular weight
N	Negatively charged group
NCI	National cancer institute
NMR	Nuclear magnetic resonance
OPLS	Optimized potentials for liquid simulations
P	Positively charged group
PDB	Protein Data Bank
PH	Pleckstrin homology domain
Phe	Phenylalanine
PISA	Phi component of the total solvent accessible surface area
PLS	Partial least squares
Pro	Proline
PSA	Van der Waals surface area of polar nitrogen and oxygen atoms
QPlogPo/w	Predicted octanol/water partition coefficient
QSAR	Quantitative structure-activity relationship
R	Aromatic ring
Rab	Ras-related proteins in brain
Rac	Ras-related botulinum toxin substrate
Ran	Ras-related nuclear protein
Ras	Rat sarcoma

RCSB	Research Collaboratory for Structural Bioinformatics
Rho	Ras homolog gene family
RIE	Robust initial enhancement
Rif	Ras-related protein interacting factor
RMSD	Root mean square deviation
RMSE	Root mean square error
Rnd	Resistance-nodulation-cell division family
RotB	Rotatable bond penalty
S(1,2,3,4,5)	Proposed SopE molecules 1, 2, 3, 4, 5
SASA	Total solvent accessible surface area in square angstroms
SD	Standard deviation
Ser	Serine
SMILES	Simplified molecular input line entry system
SolvGB	Solvation energy of the complex
Sop	Salmonella outer protein
SP	Standard precision
SptP	Salmonella protein tyrosine phosphatase
Thr	Threonine
Trp	Tryptophan
TTF	Thyroid transcription factor
Tyr	Tyrosine
Val	Valine
vdW	Van der Waals
WPSA	Weakly polar component of the total solvent accessible surface area

XP	Extra precision
Y(1,2,3,4,5)	Proposed YopE molecules 1, 2, 3, 4, 5
Yop	Yersinia outer protein

1. INTRODUCTION

Guanine nucleotide-binding proteins, or G-proteins, are intracellular cell membrane proteins that actively participate in cell signaling [1]. They collect and distribute signals from many hormones, neurotransmitters, and other signaling factors [2, 3]. Therefore, their proper functioning is essential for the organization and survival of the cell. Rho GTPase proteins, which are small G-proteins, are the most important regulators of the actin cytoskeleton, and they are implicated in cell cycle progression, cell division, vesicular trafficking and motility [3-7] . Three types of regulators control GTPase signaling: guanine nucleotide exchange factors (GEFs), GTPase-activating proteins (GAPs) and guanine nucleotide dissociation inhibitors (GDIs) [2, 3, 8]. Several bacterial protein toxins alter the state and functioning of Rho GTPases by mimicking these GTPase regulators. Examples of such bacteria are *Pseudomonas*, *Salmonella*, *Shigella* and *Yersinia* [9, 10].

In this work, two bacterial protein toxins that target and manipulate Rho GTPases were investigated. One of these proteins is *Salmonella* outer protein E (SopE) and the other is *Yersinia* outer protein E (YopE). The pathogenic bacteria *Salmonella* and *Yersinia* are gram-negative enterobacteria that can be found in plants, animals and humans [11].

During the entry of *Yersinia* and *Salmonella* into mammalian cells, the bacterial outer proteins YopE and SopE are injected via a conserved type III secretion system into the cytosol of the host cells [9, 12, 13]. This secretion system allows toxins to be injected directly from the bacterial cytoplasm into the host cytoplasm through a needle complex, rather than simple secretion into the extracellular medium [14]. Their invasion results in a disruption of actin cytoskeleton [9]. Injected YopE proteins cause depolymerisation of the actin stress fibers in eukaryotic cells, prevents phagocytosis and interferes with host cell's immune system [15, 16]. SopE proteins induce membrane ruffling and promotes bacterial uptake of host cell [17, 18]. Additionally, by interfering with host cell signaling mechanism, YopE and SopE can lead to many disease states in the cell, including cancer, cardiovascular disease, hepatic disease and developmental disorders, depending on the host cell functions [19].

In this research, the aim to discover potent inhibitors targeting the bacterial pathogen proteins *Salmonella* SopE and *Yersinia* YopE, with the use of computer-aided drug discovery tools and principles. For this purpose, compound library formation, ligand-based virtual screening which involves pharmacophore building, quantitative structure-activity relationship (QSAR) and library filtering, structure based virtual screening which involves molecular docking and scoring was carried out to propose a set of biochemically active molecules with inhibition potential against SopE and YopE. Protein Data Bank [20] was used to obtain the necessary protein structures, and Schrödinger Suite 2011 modules (Section 3.1) were used for virtual screening approaches.

The content of this thesis is as follows: Background section includes necessary theoretical information about computer-aided drug discovery methods along with structural information and infection mechanism of the bacterial proteins YopE and SopE. The computational tools used throughout the study are explained in Methods section. Result and Discussion section covers all the computational results and implications, including tables and figures. The interpretation of results is also presented in this section. The following section Conclusions and Recommendations includes the summary of the research goals, used methods, results and major conclusions along with recommendations for future work. Several additional data and information regarding the study are represented in Appendices.

2. BACKGROUND

2.1. Family of Small RhoGTPases: Attractive Targets for Pathogens

G-proteins, short for guanine nucleotide binding proteins, are a family of proteins involved in signal transmission in the cell [1]. G-protein generally refers to the membrane-associated heterotrimeric G-proteins, which are large G-proteins. These proteins are made up of alpha (α), beta (β) and gamma (γ) subunits. G-proteins are bound to either GDP or GTP. If GDP is bound to G-protein, all subunits form a complex and the protein is inactive. Once a signal molecule arrives at the extracellular space of the cell, it binds to the receptor proteins within the transmembrane region. Signal transition to cell cytosol induces conformational change of receptor protein, which is sensed by membrane-associated G-proteins. This conformational change catalyzes the exchange of bound GDP for GTP, which activates G-proteins. Both nucleotides (GDP and GTP) bind to α-subunit of G-proteins. Activated G-proteins are detached from the receptor, and separated into two units, α-subunit and βγ-subunit. These separate units are able to transmit received signal to same or different effector proteins [1, 21]. The visual representation of G-protein cycle is shown in Figure 2.1. Some of the cellular processes that are mediated by G-proteins are represented in Table 2.1.

Table 2.1. Examples of physiological processes mediated by G-proteins [1].

Stimulus	Receptor	Effector	Physiological response
Epinephrine	β-adrenergic receptor	adenylate cyclase	glycogen breakdown
Light	rhodopsin	c-GMP phosphodiesterase	visual excitation
Acetylcholine	muscarinic receptor	potassium channel	controls the rate of the heartbeat

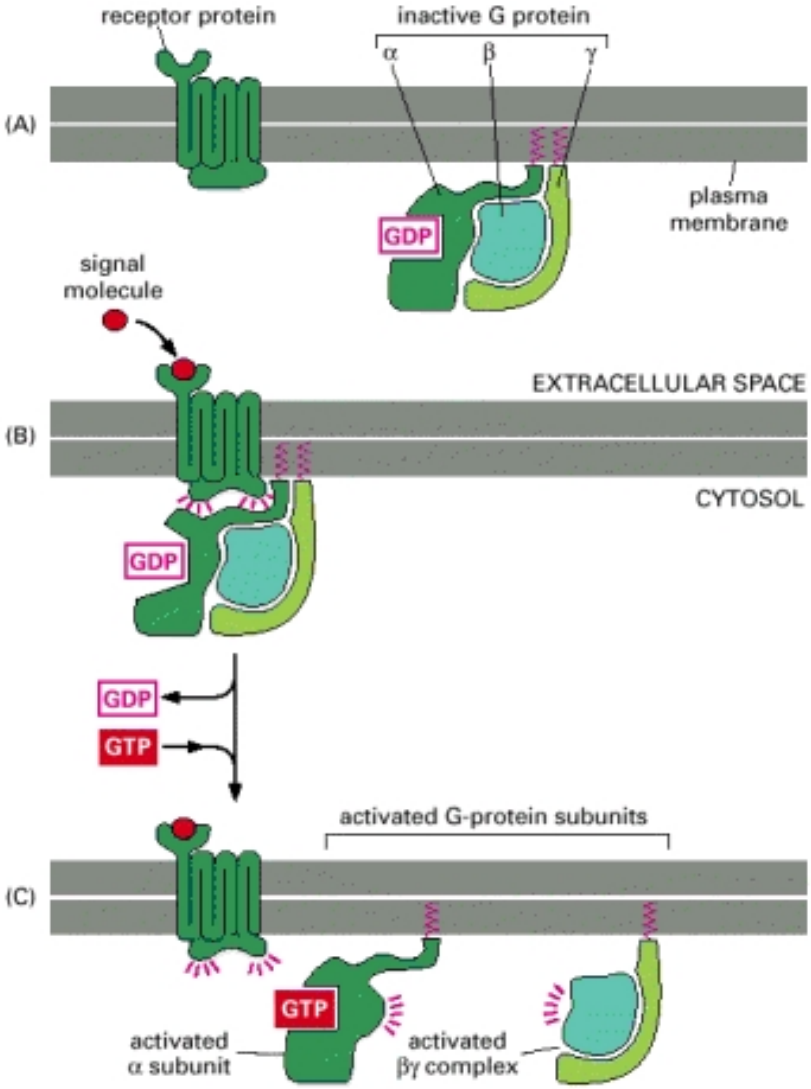


Figure 2.1. G-protein activation cycle [21].

Besides heterotrimeric G-proteins, there are also small G-proteins or small GTPases that are monomeric. These GTPases are small hydrolase enzymes which resemble the α-subunit of heterotrimeric G-proteins, therefore they are capable of nucleotide binding [21]. Small GTPases constitute a large family, which is called Ras superfamily of small GTPases, containing more than 150 proteins [22]. Ras superfamily has been divided into five subfamilies, which are Ras, Rho, Rab, Ran and Arf [3, 22]. Among them, Rho GTPase subfamily is the most crucial regulator of actin reorganization and has been studied in great detail [22]. Up to date, 22 human members of the Rho family have been identified and can be divided into 10 groups: Cdc42, Rac1, RhoA, RhoD, Rif/RhoF, Rnd3/RhoE, TTF/RhoH, RhoV, and mitochondrial Rho or Rho-related BTBdomain-containing protein [2, 3].

Rho-family proteins are regulators of the actin cytoskeleton, gene transcription and cell-cycle progression. They also play a role in several cellular processes, such as adhesion and migration, neurite extension and retraction, cellular morphogenesis and polarization, growth and cell survival [2]. Like other G-proteins, Rho-family proteins can serve as molecular switches, by binding to either GDP or GTP (Figure 2.2). Since nucleotide dissociation is normally a slow process, there are some regulators that catalyze the process of switching between GDP and GTP bound states of Rho GTPases. These regulators are guanine nucleotide exchange factors (GEFs), GTPase-activating proteins (GAPs) and guanine nucleotide dissociation inhibitors (GDIs) [2, 3, 6].

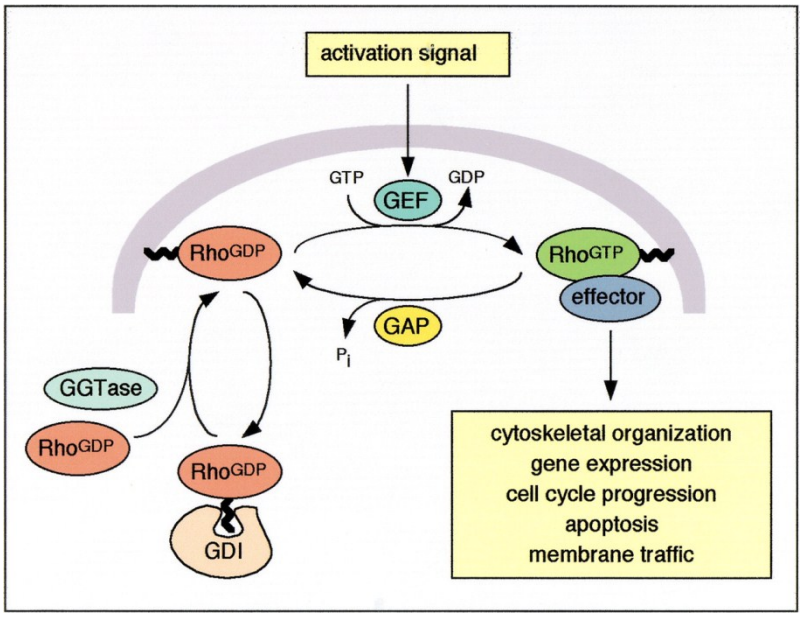


Figure 2.2. Rho GTPase switch [8].

Guanine nucleotide exchange factors (GEFs) stimulate the exchange of GDP for GTP to activate Rho GTPases [2]. Activated Rho GTPases are capable of transmitting signal to effector proteins, initiating many cell processes. When Rho GTPases are activated, their conformations change in especially two regions, called “switch 1” and “switch 2”, which form an interaction surface for downstream effector proteins [2, 8]. In contrast, GTPase activating proteins (GAPs) stimulate the exchange of GTP to GDP, to inactivate Rho GTPases [23]. Majority of the GAP proteins have a conserved arginine residue, which is strongly associated with GTP release from GTPases [24]. GAP activity stops the signaling event. The third class of GTPase regulators is guanine nucleotide

dissociation inhibitors (GDIs). They interact with the GDP-bound form of GTPases to control cycling between membranes and cytosol. GDIs prevent the dissociation of GDP from GTPases and keep them in their non-signaling state [2].

2.1.1. Type III Secretion System Mediated Infection of GTPases

There are several examples of bacterial virulence factors that specifically target members of the Rho GTPase family and their regulators. Because Rho GTPases play an important role in cell signaling events and several cellular functions, they are favored by many bacterial pathogens as targets [9]. Bacterial cytotoxins invade host cells via a specialized secretion system and regulate Rho GTPases by mimicking either GEF or GAP activity [25, 26].

Gram-negative bacteria use a special type of secretion system to deliver its virulence factors into the target host cell. Type III secretion system is a delivery system used by many bacteria, such as *Burkholderia*, *Chlamydia*, *Salmonella*, *Shigella* and *Yersinia* [9]. With this secretion system, bacteria inject their virulence proteins directly into the host cell cytosol. Besides pathogenic bacteria, type III secretion systems can be found in animals and plants [12]. Type III secretion systems are one of the most complex secretion systems found in bacteria, involving more than 20 proteins, and the secretion mechanism is not completely understood [27]. A simplified representation of type III secretion system is shown in Figure 2.3.

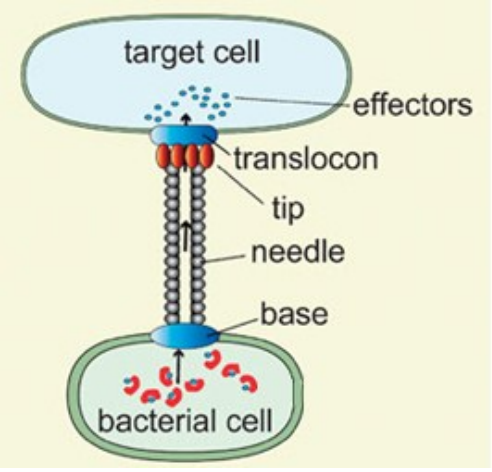


Figure 2.3. A simple illustration of type III secretion system [28].

Type III secretion apparatus consist of a base, needle, tip, and translocon. The core of the secretion apparatus is the base, which is placed between inner and outer membranes of gram-negative bacteria and composed of several membrane rings. These porous rings function as an anchored base for the needle structure. The needle structure, or injectisome, is located between the cytosols of gram-negative bacteria and host cell, and functions as an export apparatus for bacterial effector proteins. The needle passes through both inner and outer membranes of bacteria. The end of the needle consists of a tip protein, which makes a small hole in the cell membrane of the host protein. This hole enables the transfer of bacterial effector proteins to be injected into the host cell through the needle. First effectors that are secreted into the host cell cytoplasm produce a channel called translocon within the host membrane, from which upcoming effectors can be inserted [28]. Once the effectors enter the host cell cytoplasm, they can manipulate host cell GTPases by mimicking GEF and GAP activities [29].

2.2. *Yersinia* and *Salmonella* Pathogenesis

The genus of the Gram-negative bacterium *Yersinia* has 11 known members, from which three species are pathogenic to humans: *Yersinia pestis*, *Yersinia enterocolitica* and *Yersinia pseudotuberculosis*. Y. *pestis* is the causative agent of plague, which results from direct inoculation of bacteria into the bloodstream. Y. *enterocolitica* and Y. *pseudotuberculosis* are naturally present in soil, and acquired by ingestion of contaminated food or water. All three species have a common 70 kB virulence plasmid, on which type III secretion system is encoded. This plasmid functions to export virulence-associated effector proteins (Yops) from plasma membrane and deliver them to cytosol of the host cell. Survival and replication purpose of Yops seem to interfere with regulation of host immune system, since Yop secretion inhibits phagocytosis and causes macrophage cell death. *Yersinia* is known to translocate six effector proteins (YopE, YopH, YopJ, YopM, YopO and YopT) into the target cell cytosol where they all have different functions. The major anti-host effect induced by YopE is the disruption of the actin skeleton of the host cell [30-33].

The genus of the Gram-negative, rod-shaped, motile bacterium *Salmonella* has 2 species: *Salmonella enterica* which is divided into six subspecies and *Salmonella bongori*.

Salmonella genus are found widespread in cold- and warm-blooded animals and cause illnesses like typhoid fever, paratyphoid fever, and foodborne illness. Similar to *Yersinia*, *Salmonella* delivers its outer effector proteins (SopB, SopE, SopE2) into the host cell by type III secretion system. These effector proteins modulate different cellular processes and therefore are involved in various stages of bacterial infection. The SopE protein initiates bacterial internalization by membrane ruffling [34-36].

2.3. Bacterial GAP Effector Protein YopE

Yersinia outer protein E, YopE, is an effector protein of the gram-negative bacteria *Yersinia*, along with other outer effector proteins YopH, YopJ, YopM, YopO and YopT [16]. The YopE protein is found to be an important virulence factor that target small Rho GTPases by acting as a GTPase activating protein (GAP) [37]. *In vivo*, YopE is observed to target Rac1, RhoA and Cdc42 proteins of Rho GTPases [23, 37] . Bacterial uptake of *Yersinia* outer proteins are schematically shown in Figure 2.4.

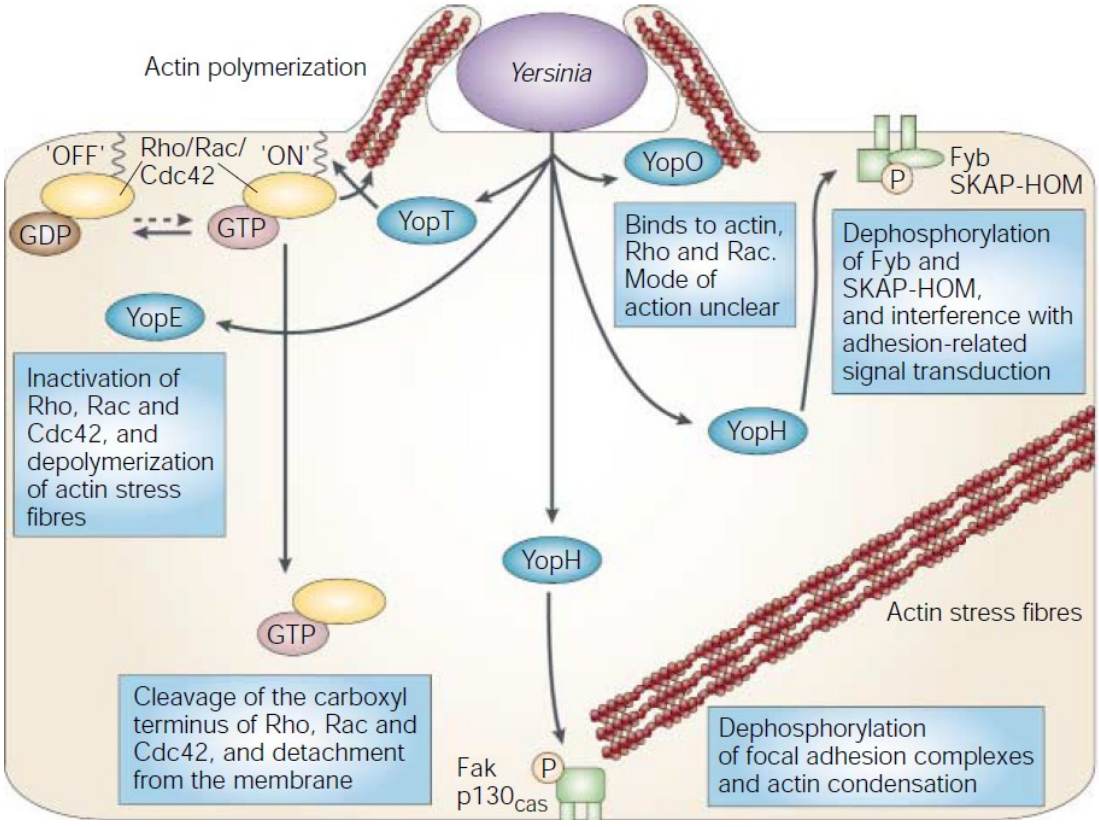


Figure 2.4. Targets and functions of *Yersinia* outer proteins [37].

YopE is injected into the host cells by type III secretion system and mimics eukaryotic GAP proteins functionally, and makes contact with the catalytic site of the target Rho GTPase. YopE protein inserts an arginine finger to this site, which stimulates the nucleotide exchange from GTP to GDP [16]. Thus, GTPase becomes inactivated, and signal translation to downstream effectors is prevented by YopE activity [23, 37].

With its GAP activity, YopE disorganizes the actin cytoskeleton by depolymerisation of actin stress fibers of the host cell. Beside its toxic effects on the cytoskeleton, YopE, together with YopH, ensures the antiphagocity of other *Yersinia* effectors by deactivating host cell's immune system. Additionally, YopE is also found to affect cytokine production [16, 38]. Overall, it is evident that YopE takes an important role in different virulence mechanisms induced by *Yersinia*. The three dimensional structure of GAP domain of YopE determined by X-ray crystallography is shown in Figure 2.5.

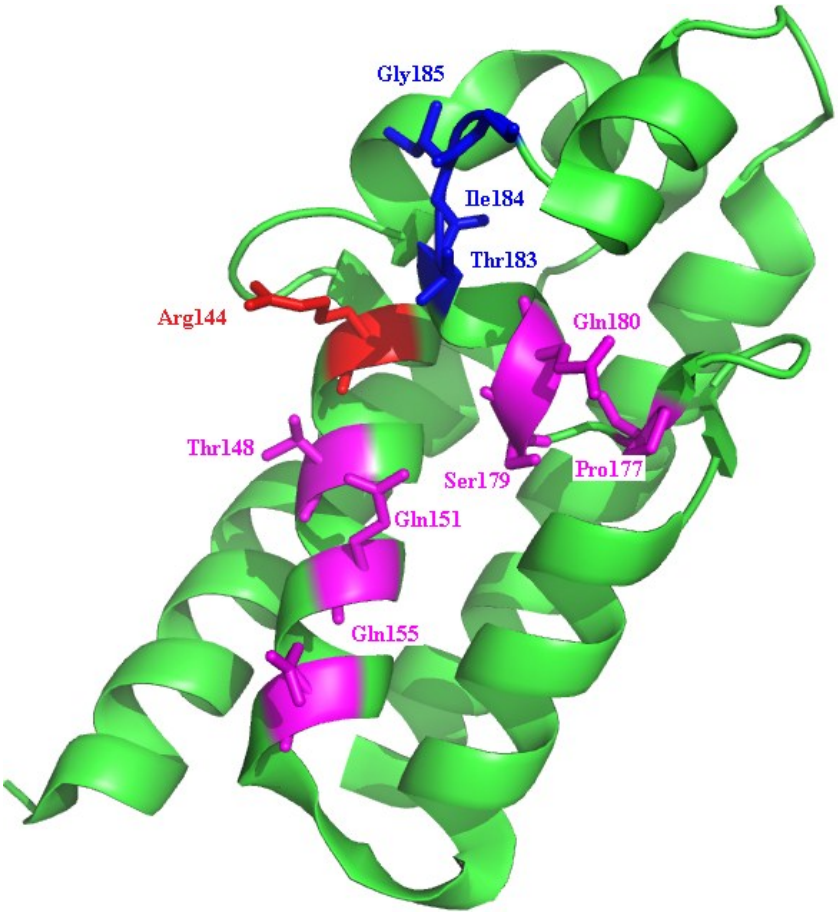


Figure 2.5. Cartoon representation of YopE (PDB id: 1hy5). Residues that contact both with nucleotide and switch regions are shown as stick.

Its N-terminal domain (residues 1-89) contains the signals that target the protein for secretion from the bacterium and translocation into eukaryotic cells by the type III secretion in *Yersinia* pestis, and its C-terminal domain (residues 90-219) is the seat of the GAP activity [23, 39, 40]. YopE interactions with G-proteins are investigated and important YopE residues that govern its activity are reported. To date, no crystals of YopE-GTPase protein complexes have been obtained. Using the similarity between $YopE_{GAP}$ and other bacterial GAPs, a model was constructed for $YopE_{GAP}$ and Rac1 (GTPase) [1, 21, 22]. Based on this complex model, residues Ile106, Leu109, Thr138, Gly139, Ser140, and Gln149 are observed to interact with Switch II region of the GTPases. The key residue Arg144 along with Thr183, Ile184, and Gly185, are reported to contact nucleotide and both of the switch regions. Additionally, residues Thr148, Gln151, Gln155, Pro177, Ser179, and Gln180 are found to interact with nucleotide and Switch I region of Rho GTPases. Nevertheless, these residues are not conserved in all three bacterial GAPs. Conserved residues are Gly139, Thr148, Gln180, Thr183, and Gly185 [41]. All these residues, particularly arginine finger and other conserved residues that contact both of the switch regions, were aimed to be observed in virtual screening approaches.

2.4. Bacterial GEF Effector Protein SopE

Salmonella outer protein E, SopE, is an effector protein of the gram-negative bacteria *Salmonella*, along with other outer effector proteins SopB and SopE2 [18]. The bacterial SopE, along with other effector proteins, injected via a conserved type III secretion system into the cytosol of the host cells [9]. SopE acts as a guanine nucleotide exchange factor (GEF) and stimulates the nucleotide exchange from GDP to GTP. SopE activity induces signal transduction within the host Rho GTPases [42]. *In vivo*, SopE is observed to target Rac1 and Cdc42 proteins of Rho GTPases. Bacterial uptake of SopE induces membrane ruffling and actin cytoskeleton rearrangement [42].

The structure of SopE is represented in Figure 2.6 as ribbons and colored according to the secondary structure type. SopE is a bacterial GEF protein with 165 amino acids on its C-terminal catalytic domain. Its structure consists of six α-helices in two three-helix bundles and a two stranded β-sheets followed by an 8 amino acid long peptide segment.

This peptide segment contains the catalytic core of SopE, a GAGA (residues 166-169) motif, which is the central green loop in Figure 2.6.

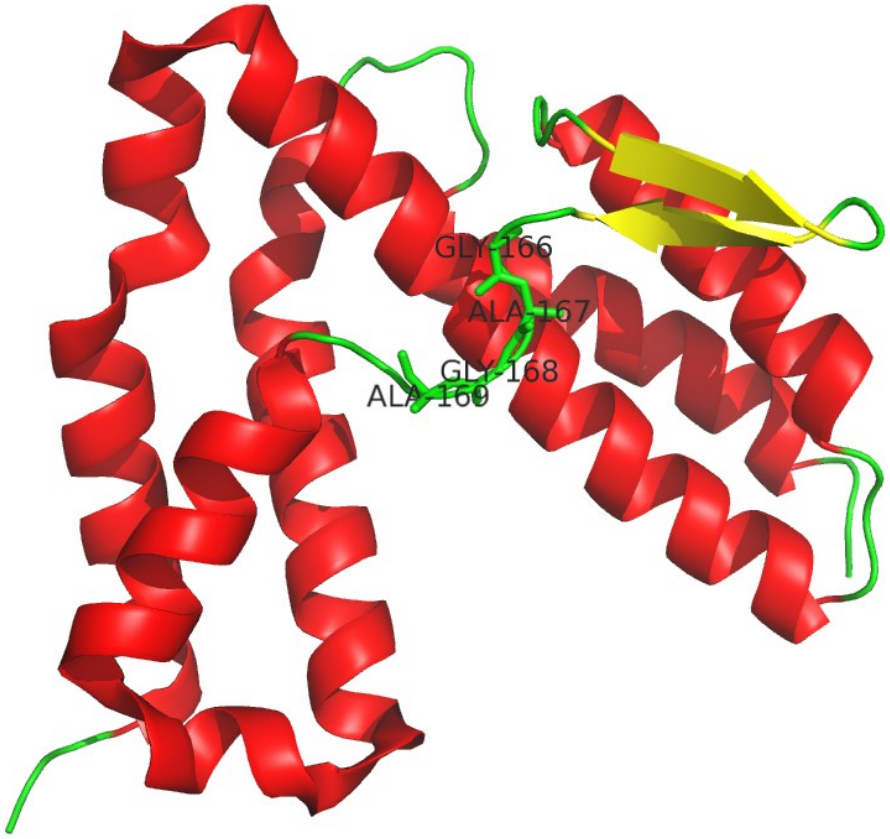


Figure 2.6. Crystal structure of the catalytic domain of SopE, colored by secondary structure (PDB id: 1gzs:B). GAGA loop residues are labeled.

Catalytic core of SopE does not resemble the typical catalytic domain of other Rho-GEFs; whose structures have either DH-PH or DHR2 domains [16]. The structural analysis of the SopE-Cdc42 (GTPase) complex reveals that, insertion of the GAGA loop between the nucleotide-binding regions of Cdc42 leads to major conformational changes in the Switch I and Switch II regions, resulting in reduced GDP affinity. Therefore, it is reported that the GAGA loop is the most important region that governs SopE's bacterial GEF activity. Similar conformational changes in these switch regions have been observed in other eukaryotic GEF proteins that regulate Rho family of GTPases [42].

SopE interaction with Rho GTPase protein Cdc42 is investigated by solving crystals of SopE-Cdc42 complex, and important SopE residues are reported [42]. Multiple contacts

are observed between Switch I region of Rho GTPase and Asp124 residue of SopE. Ile131 residue is also reported to have hydrophobic contacts with Switch I region. Other reported Switch I interactions involve Gln194 and Lys198 residues of SopE. GAGA loop residues Gly166, Gly168 and Gly169 are reported to interact with the Switch II region. Other Switch II contacts involved SopE residues Asp103, Gln109, Ser165 and Thr174. These residues, GAGA loop residues in particular, were aimed to be observed in virtual screening approaches.

2.5. Previous Research on Inhibition of YopE and SopE

There are several studies regarding *Yersinia* and *Salmonella* outer proteins, but they generally focus on the type III secretion mechanism or the adverse effects that occur within the cell after the bacterial uptake [43-48]. There are no known or reported co-crystallized ligands for YopE or SopE. Studies involving computer-aided drug discovery and virtual screening approaches targeting YopE and SopE could not also be found.

Small molecule inhibitors that the target type III secretion of Yersinia outer proteins *in vitro* have been identified by preparing gene reporter assays and screening commercially available small molecules with these assays. As a result, 23 compounds that belong to a class of acylated hydrazones of different salicylaldehydes were found. These compounds were observed to specifically block Yop effector secretion *in vitro*. Inhibitory concentrations of these compounds were reported [46].

These compounds, which are titled as compound1-compound23, were utilized for virtual screening purposes in this study. 2D structures of these inhibitors are represented in Table 2.2, and they were denoted as known inhibitors of YopE in this study.

Table 2.2. 2D structures of known inhibitors [46].

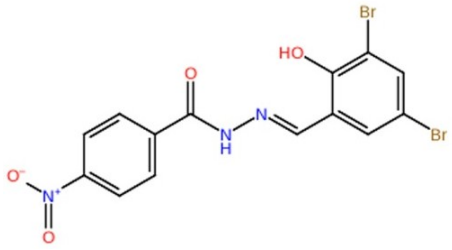
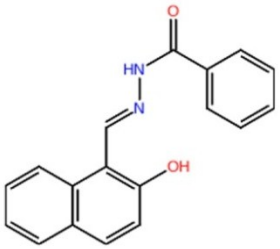
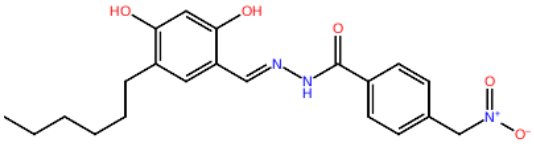
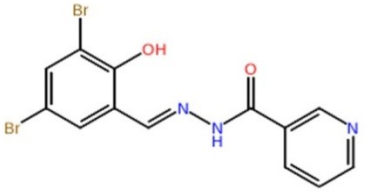
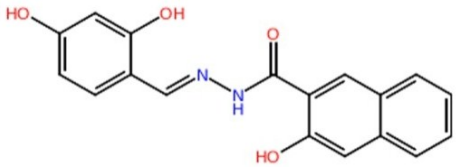
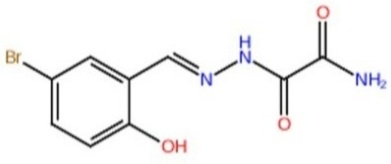
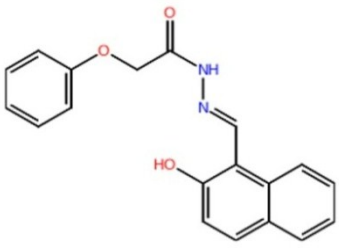
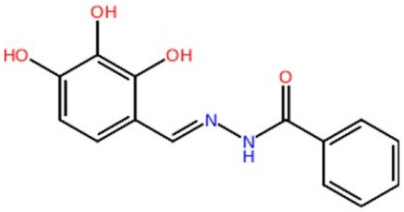
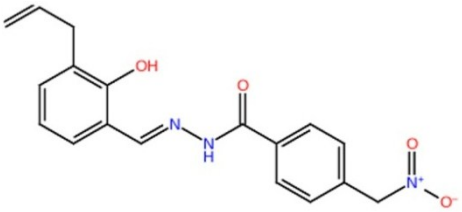
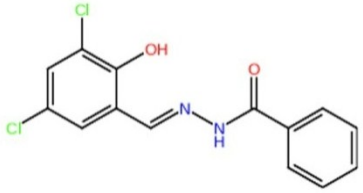
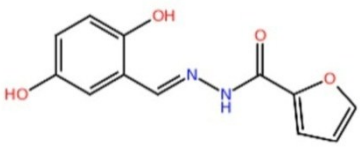
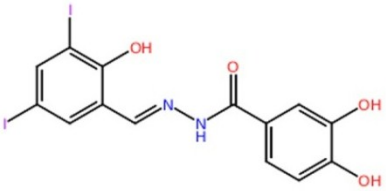
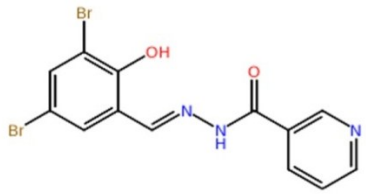
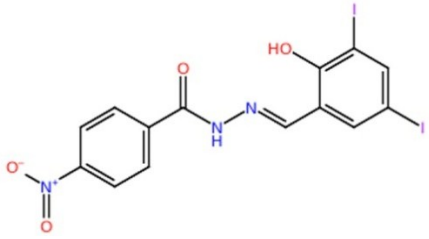
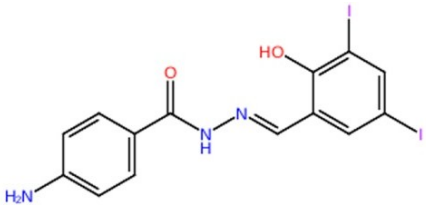
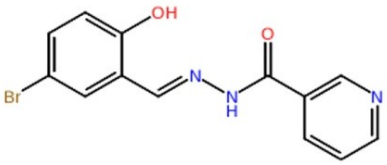
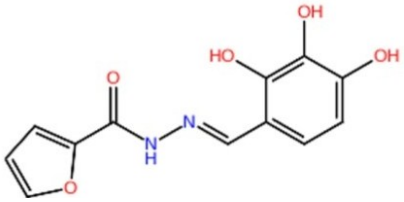
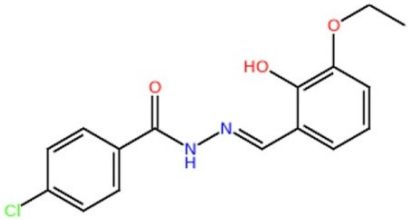
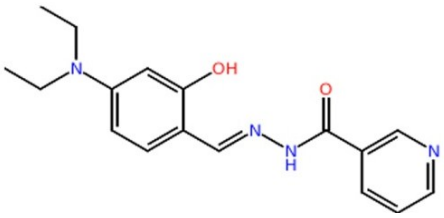
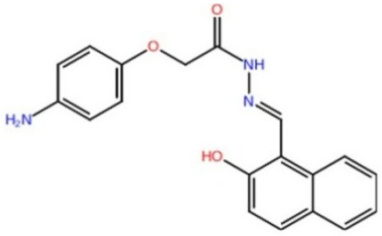
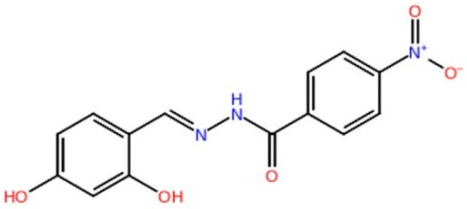
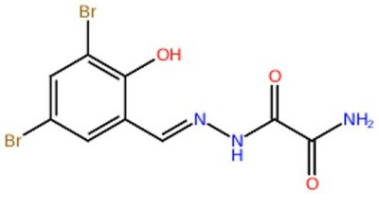
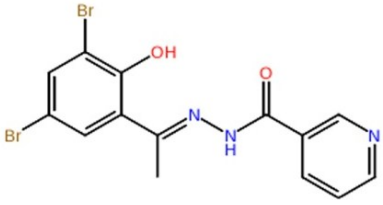
Title	Structure	Title	Structure
1	 Chemical formula: $C_{14}H_9Br_2N_3O_4$	7	 Chemical formula: $C_{18}H_{14}N_2O_2$
2	 Chemical formula: $C_{21}H_{25}N_3O_5$	8	 Chemical formula: $C_{13}H_9Br_2N_3O_2$
3	 Chemical formula: $C_{18}H_{13}N_2O_4$	9	 Chemical formula: $C_9H_8Br_1N_3O_3$
4	 Chemical formula: $C_{19}H_{15}N_2O_3$	10	 Chemical formula: $C_{14}H_{12}N_2O_4$
5	 Chemical formula: $C_{18}H_{16}N_3O_4$	11	 Chemical formula: $C_{14}H_{10}Cl_2N_2O_2$
6	 Chemical formula: $C_{12}H_{10}N_2O_4$	12	 Chemical formula: $C_{14}H_8I_2N_2O_4$

Table 2.3. 2D structures of known inhibitors [46] (cont.).

Title	Structure	Title	Structure
13	 Chemical formula: $C_{15}H_9Br_2N_3O_4$	19	 Chemical formula: $C_{14}H_9I_2N_3O_4$
14	 Chemical formula: $C_{14}H_{11}I_2N_3O_2$	20	 Chemical formula: $C_{13}H_{10}Br_1N_3O_2$
15	 Chemical formula: $C_{12}H_{10}N_2O_5$	21	 Chemical formula: $C_{16}H_{15}Cl_1N_2O_3$
16	 Chemical formula: $C_{17}H_{20}N_4O_2$	22	 Chemical formula: $C_{19}H_{16}N_3O_3$
17	 Chemical formula: $C_{14}H_{11}N_3O_5$	23	 Chemical formula: $C_9H_7Br_2N_3O_3$
18	 Chemical formula: $C_{14}H_{11}Br_2N_3O_2$		

2.6. Computer-aided Drug Discovery

Drug discovery is the design or discovery of novel therapeutic agents for a protein involved in a specific disease. Traditionally, these therapeutic agents, also known as lead compounds or ligands, are identified with high throughput screening (HTS), which involves formation of targeted protein-ligand complexes in biological assays [49]. In HTS, a large number of drug-like compounds are rapidly screened against a biological target via experiment testing, to observe if screened compounds inhibit the unfavorable function of the target protein [50, 51]. With the recent advances in pharmaceutical research and technology, computers can aid the drug development process, as a complementary tool to experimental testing [52, 53]. Using computer-aided approaches, new drug candidates can be identified in a timelier manner [54].

A method for identifying drug candidates is to virtually screen compound libraries against the target protein. These compound libraries contain structures of a high number of biological and synthetic compounds in electronic format [55, 56]. As a result of the virtual screening, compounds that show predicted binding affinity to target protein, which are called hits, are obtained. There are several companies that offer such compound libraries containing commercial, small drug-like molecules [57]. Pharmaceutical companies may also have built-in libraries, which are used by their research and development departments [58].

These promising compounds are tested *in vitro*, in order to investigate their biological activity against the target protein. Drug development process can start if a compound exhibits a notable inhibition in these tests [49]. Despite the recent developments in technology and perception of protein-ligand interactions, drug discovery is still very expensive and longstanding [59]. Thus, computer-aided drug development is an important and continuously improving area in drug discovery [60]. Computer-aided drug design can be either processed with ligand-based design methods such as library screening methods that use information about known active ligands as the basis for the design of novel lead compounds, or structure-based design methods such as molecular docking which use the information about the target protein [61-63].

2.7. Virtual Screening for Lead Discovery

Virtual screening is a computational method that enables the discovery of novel compounds showing activity against a known biological target, which is generally a protein receptor [64]. Virtual screening approach uses various tools and principles, including molecular mechanics and quantum mechanics calculations, chemoinformatics, database of molecular structures, quantitative structure-activity relationship techniques and molecular graphics for visualizing molecules [65, 66]. Virtual screening approaches are extensively used in drug development currently, for lead identification purposes [54]. Table 2.4 shows examples of clinically approved drugs which were developed by the help of computer aided-drug discovery methods [67].

Table 2.4. Examples of drugs discovered by computer-aided drug discovery tools.

Compound	Trade name	Target	Disease	Company	Year
Saquinavir	Fortovase	HIV Protease	HIV	Roche	1995
Dorzolamide	Trusopt	CarbonicAnhydrase	Glaucoma	Merck	1998
Zanamivir	Relenza	Neuraminidase	Influenza	GSK	1999
Oseltamivir	Tamiflu	Neuraminidase	Influenza	Roche	1999
Lopinavir	Kaletra	HIV Protease	HIV	Abbott	2000
Imatinib	Glivec	Tyrosine Kinase	Leukemia	Novartis	2003
Erlotinib	Tarceva	EGFR Kinase	Cancers	OSI Pharma	2004
Raltegravir	Isentress	HIV Integrase	HIV	Merck	2007
Boceprevir	Victrelis	Protease	Hepatitis C	Merck	2011

There are two main virtual screening techniques: ligand-based and structure-based. Ligand-based methods make use of structural information of a ligand against a protein. Structure-based methods do not require ligand knowledge, but rather they utilize structural information of the receptor protein [62]. These two methods can also be combined together in virtual screening studies [63, 68].

2.7.1. Ligand-based Approaches

Ligand-based virtual screening uses the structural information of a set of compounds with known activities against the target protein. Ligand-based approaches are commonly applied in the absence of knowledge about the target in question, but they can also be utilized in lead modification or library filtering [69, 70]. The basic assumption in ligand-based approaches is that molecules that exhibit structural similarity are more likely to have similar activity [70]. Ligand-based methods can be roughly categorized by whether they use 2D or 3D methods to search the compound library [71]. Former method searches the library according to the chemical structure of the known ligands, whereas the latter method takes 3D forms of known ligands into account, in addition to their structures [71].

Frequently used 2D search methods are substructure searching and similarity searching [72, 73]. Substructure search screens the compounds based on their 2D structure (e.g. SMILES) and finds commonly observed partial matches between structures of known actives and library compounds [73]. This method is very fast, but it does not take physicochemical properties or 3D structures of molecules into account. The other method is similarity searching, which is more detailed and slow than substructure search [74]. Similarity searching method calculates similarity coefficients between known ligands and compound libraries. The most popular similarity coefficients are Tanimoto coefficient and Dice coefficient [75, 76]. These coefficients are determined comparing a large number of physical and chemical properties of molecules, such as common fragment substructures, molecular weight, molecular volume, atom count, ring count, octanol/water partition coefficient, surface area, dipole moment [74-76]. Higher coefficient number indicates higher similarity according to this method.

Ligand-based screening based on 3D structures of molecules is called pharmacophore searching or building. Pharmacophore search also involves identifying similar functional group features, such as hydrogen bonding behavior, polarity, hydrophobicity and aromaticity [77]. In pharmacophore matching, set of features that are frequent in known ligands is identified, and later used as library filtering criteria [77]. A compound must include the structural features of a pharmacophore with a proper 3D alignment, in order to be identified as similar to the known ligands [78]. Besides 3D

search, constraints can also be used upon database filtering, such as druglikeness according to Lipinski rule, to reduce the screening time [79]. If the structural information about the target structure's binding site is available, it can be incorporated in the 3D search method in the form of excluded volumes. Excluded volumes ensure that screened compounds do not occupy that excluded space.

In addition to pharmacophore building, quantitative structure-activity relationship (QSAR) models can be obtained using 3D molecular descriptors, if the activities of known ligands are available [80]. Pharmacophore-based QSAR models need the alignment of known ligands according to the pharmacophore feature points [80]. Once ligands are aligned, QSAR descriptors can be calculated. With these calculations, QSAR equation can be derived using a suitable regression tool, such as partial least squares method [81]. The independent descriptors are used to derive the equation that predicts the dependent descriptor, such as activity. After the QSAR model has been built, validation must be performed using internal and external tests [82]. Internal validation consists of predicting the property of known ligands that were used in creation of the QSAR model. In contrast, external validation consists of predicting the property of known ligands that were not used upon QSAR building. The predictive ability of the QSAR equation is generally indicated by multiple correlation coefficients, which are R^2 value for internal set and the analogous Q^2 value for external set. QSAR model with higher predictive ability has higher correlation coefficients, closer to unity [82]. QSAR method can be applied to compare different pharmacophore feature sets, and to determine the set that best describes the activities of the known ligands.

Based on the similarity to the known ligands, library compounds can be selected and tested in biological assays. Alternatively, ligand-based approaches can be used to reduce the number of molecules in the compound library, which can be later used in structure-based approaches.

2.7.2. Structure-based Approaches

Structure-based virtual screening approaches require structural information about target protein and binding site [83]. Structure-based approaches are more extensively used

in virtual screening than ligand-based approaches, since knowledge about the binding site can provides valuable information about ligand binding interactions [84]. Target protein's structure is directly obtained from X-ray crystallography and NMR structure, or homology modeling can be used.

Structure-based approaches are usually performed with molecular docking method, in which library compounds are placed near the binding site using various search algorithms. Once the position and orientation of a compound near the protein's binding site is determined, the binding affinity can be estimated using different types of scoring functions [85]. Molecular docking requires a search algorithm, scoring function, 3D structure of the binding site of the target protein, and a library of compounds in a suitable electronic format [85]. Subsequent to docking, compounds are ranked according to their scores. High scoring compounds can be chosen and tested in biological assays.

Molecular docking includes two major aspects: pose prediction and scoring. Pose can be defined as the position and orientation of library compounds near binding site of the target protein [86]. Pose prediction is performed by search algorithm of the related docking program [86]. Schrödinger Suite's molecular docking module Glide [87] uses the Monte Carlo method to generate random conformations of compounds to be docked. These conformations are then docked and scored. Poses are either accepted or rejected based upon predefined criteria, which is the Boltzmann probability function for Monte Carlo method [87]. There are different search algorithms used by various docking programs, some of which are listed in Table 2.5.

Table 2.5. Common docking programs used in virtual screening approaches [88, 89].

Docking program	Search algorithm	Scoring function
Autodock	Genetic algorithm	Force field
Dock	Fragmentation	Force field
FlexX	Fragmentation	Empirical
Gold	Genetic algorithm	Empirical
Glide	Stochastic search	Empirical

Once the poses are predicted, they are ranked according to a scoring function, which predicts the binding affinity of docked compound to the target protein [87]. Majority of the available docking programs take the receptor as a static structure, whereas docked compounds can be either rigid or flexible [90]. In rigid docking, structures of compounds are kept in their input states. In flexible docking, multiple conformations of a single compound can be docked, and energy minimization can be performed on compounds to be docked. Flexible docking is much more accurate at the expense of the computational time and effort spent [29].

Frequently used docking programs use one of the following scoring functions: force field, empirical and knowledge-based [91]. Force field scoring function calculates energy of the ligand-receptor complex based on the atomic positions. Force fields include terms for bond stretching, angle bending, dihedral angles; and non-bonded terms for van der Waals and electrostatic interactions [89]. Force field scoring function extensively uses quantum mechanics principles [89]. Empirical scoring functions are derived from the experimental binding energy data of known receptor-ligand complexes with the help of regression methods [89]. Therefore, every docking program that uses an empirical scoring function has its own proprietary function. For example, Glide utilizes its empirical scoring function GlideScore to score and rank the predicted poses [87]. On the other hand, knowledge-based scoring function aims to obtain knowledge about known receptor-ligand binding that is available in PDB. With statistical analysis based on this knowledge, occurrence frequency of atomic-interaction pair potentials are determined and used in predicting new affinities [89]. The scoring functions of commonly used docking programs with their scoring functions are tabulated in Table 2.5.

2.8. Objectives and Plan of the Study

The main objective of this study is to use computer-aided drug discovery tools to identify potent inhibitors against YopE and SopE. The outline of the work presented can be broken down to two main steps, which are YopE and SopE virtual screening approaches.

For YopE approach, molecules that exert inhibition against YopE *in vitro* were utilized for pharmacophore building. 3D pharmacophores were constructed considering the

following common features: hydrogen bond acceptor, hydrogen bond donor, hydrophobe, negative ionizable, positive ionizable, aromatic ring. For pharmacophore hypotheses derived from YopE inhibitors, QSAR method was applied, in which validation was carried out to determine the hypothesis that best represents the activities of the inhibitors. Small molecule database generated from ZINC was screened and filtered with pharmacophore hypotheses of YopE. Pre-filtered database molecules were flexibly docket to target protein YopE. Standard precision (SP) and extra precision (XP) docking calculations were performed, successively. Binding modes of molecules were predicted with the search algorithm of the docking program, and scored according to their affinity to target proteins. The docking scores for each pose and molecule were calculated with empirical GlideScore function and ranked accordingly.

For SopE approach, fragment-based pharmacophore building was carried out since literature survey did not yield any identified SopE inhibitory molecules. Small fragments provided by Schrödinger [104] were docked to SopE and using energetic contributions of fragments to docking score, pharmacophore features were determined and hypotheses were generated. Small molecule database generated from ZINC was screened and filtered with selected pharmacophore hypotheses of SopE. Pre-filtered database molecules were flexibly docked to target protein SopE. Again, SP and XP docking were performed. Binding modes of molecules were predicted with the search algorithm of Glide, and scored according to their affinity to target proteins. The docking scores for each pose and molecule were calculated with empirical GlideScore function. From each approach, top ranking molecules were further investigated to propose a set of molecules that has a potential to exhibit activity against YopE and SopE in biological assays, which is the main goal of this study.

3. METHODS

3.1. General

All virtual screening applications were performed using Schrödinger Suite 2011 on Linux platform using HP xw6600 Workstation. The following Schrödinger modules were used: Protein Preparation Wizard [92] for receptor preparation, LigPrep [93] and ConfGen [94] for ligand preparation, Phase [95] for database preparation, pharmacophore modeling and database filtering with QSAR application, QikProp [96] for predicting molecular descriptors and druglikeness, Glide [97] for receptor grid generation and ligand docking, MacroModel [98] for ligand minimization, and Prime MM-GBSA [99] for binding free energy calculation. All modules were accessed via Maestro graphical interface [100].

3.2. Receptor Protein Preparation

Structures of the receptor proteins were extracted from RCSB Protein Data Bank (PDB id: 1gzs for SopE and 1hy5 for YopE). However, it was not appropriate to use co-crystallized PDB structures directly in any Schrödinger applications. Especially, docking calculations require 3D coordinates and charges for all heavy atoms including polar hydrogens in the structure. Therefore, they were pre-processed using Protein Preparation Wizard module in Maestro interface. This procedure can be outlined as follows:

- All bond orders and partial charges on atoms were assigned.
- Hydrogen atoms were added to heavy atoms of proteins since PDB structure files lack this information.
- All selenomethionines were converted to methionines, since mislabeled methionines could not be recognized. Selenomethionines are not naturally present in protein structure, but rather they are used in X-ray structure determination.
- Missing side chains and loops were predicted with Prime module, if necessary.
- Waters that are not important for ligand binding were located and deleted.

After the pre-processing step, structures were visually inspected in Maestro workspace and unwanted parts of the structures were deleted, such as dimeric chains or protein structures other than target. Multimeric structures were simplified by deleting the duplicate chain. SopE structure had also another protein present in PDB structure (Small G-protein Cdc42). This protein was also deleted and the receptor protein SopE was reduced to a single unit.

Next, hydrogen-bonding network was optimized by reorienting hydroxyl and thiol groups, amide groups of asparagines, glutamines and imidazole ring of histidines. Since orientations of these hydroxyl and thiol groups cannot be determined from X-ray structure, their optimization was needed. Correct protonation states of certain residues (histidine, aspartic acid and glutamic acid) were also predicted at neutral pH. Histidines also have multiple tautomers since their imidazole group form isomers at different protonation states. After protonation state prediction, tautomerization state of histidines was also predicted. At the last stage of protein preparation, a complete minimization of the atoms was performed using Impref utility. Impref minimization was carried out using the OPLS_2005 (Optimized Potentials for Liquid Simulations) force field. This utility allows removing initial steric clashes, and optimizing the positions of the hydrogen atoms added to protein structure in pre-processing step. This minimization was performed ensuring that the final structures do not deviate much from their input geometries. The allowed maximum RMSD of heavy atoms was selected as 0.30, which is default. Modified protein structures were exported in the Maestro file format, which became ready-to-use for docking simulations.

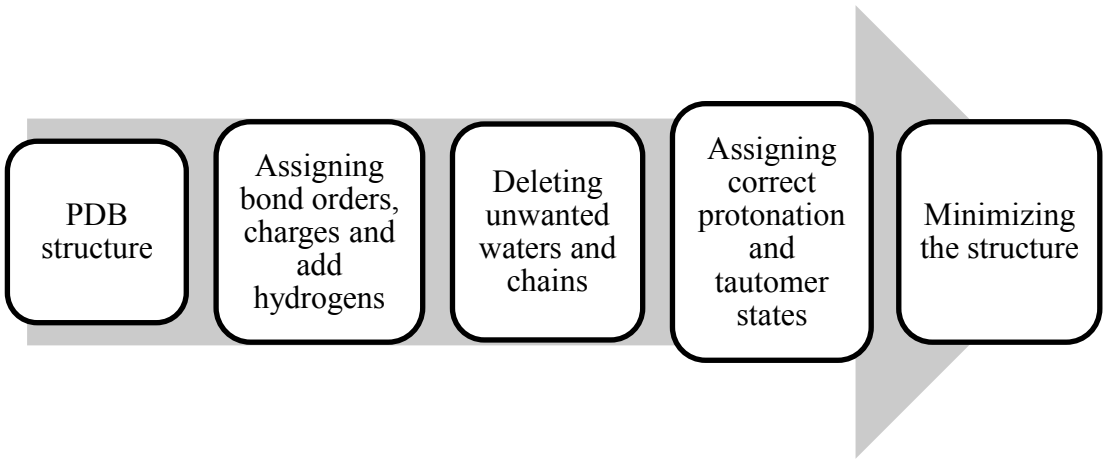


Figure 3.1. Workflow of protein preparation wizard.

3.3. Ligand Preparation and Conformer Generation

Structures of YopE inhibitors were not available in electronic format; therefore they were drawn in 2D form in Maestro workspace. Similar to proteins, ligands were also prepared prior to virtual screening simulations using LigPrep module. This module was used for preparing low-energy, all atom 3D structures with various ionization states, tautomers, stereochemistries, ring conformations, and correct chiralities for the inhibitors. This procedure can be outlined as follows:

- Inhibitor structures were checked, implicit hydrogen atoms were added if necessary.
- Presence of water molecules and ions in salts were investigated. Inhibitors did not include salt or water molecules; therefore their removal was not needed.
- Charged groups were neutralized with addition or removal of hydrogens.
- 2D structures of inhibitors were converted to 3D, generating multiple states for tautomers, stereoisomers, ring conformations and ionization at a selected pH range.
- Ionization states of ligands were generated by simply adding or removing protons. Epik module [101] was used for generating states at pH 7±2.
- Tautomers were generated up to eight conformations per ligand (default).
- Up to 32 stereoisomers were generated per ligand with determination of chiralities from 3D structure (default).
- Low-energy ring conformations were generated.
- Finally, an energy minimization of the 3D conformers with the OPLS_2005 force field was performed.

Output structures of ligand were exported in Maestro file format and became ready-to-use in docking calculations. An example of LigPrep output is represented in Figure 3.2.

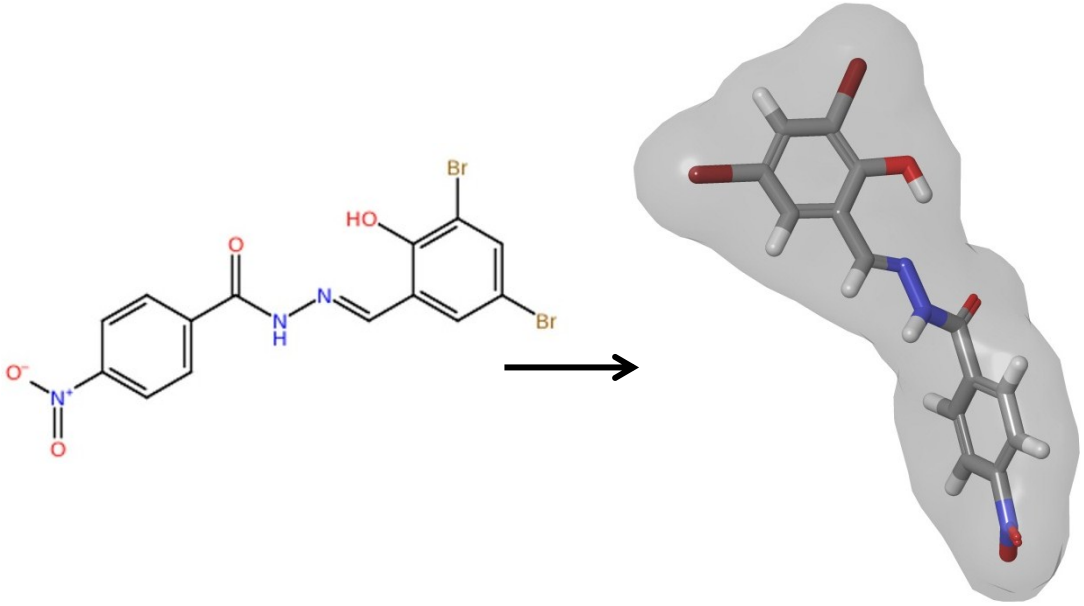


Figure 3.2. 2D to 3D conversion of compound1 in LigPrep.

3.4. Receptor Grid Generation

Glide docking requires generation of small identical grids that represents shape and the properties of receptor binding site. Electrostatic and potential energies of the receptor can be pre-calculated within these grids. Once grid energetics is calculated, they can be used in ligand docking; therefore it reduces the screening and scoring time effectively. Grid generation for YopE and SopE was performed using Receptor Grid Generation panel in Glide module. Grid generation requires a fully prepared protein structure. Hence, output files obtained from the Protein Preparation Wizard were fed to Receptor Grid Generation. The panel has five tabs: receptor, site, constraints, rotatable groups and excluded volumes.

In the Receptor tab, prepared receptor proteins were imported to the Maestro workspace. In this tab, van der Waals radius scaling should also be determined. Glide does not allow for receptor flexibility in docking, but reducing van der Waal's radii of atoms with partial charges may induce slight flexibility. Atoms are represented as Van der Waals spheres and their radius can be multiplied by a user-defined scaling factor. Default value is 1.0 scaling factor with 0.25 partial charge cutoff, where no scaling is done. Changing these parameters is essentially advised in case of a very tight and buried binding site, aiming to

soften the binding site. Since it was not the case for both YopE and SopE, the vdW radius of atoms were not scaled.

In the Site tab, a site on the receptors that could be favored by ligand binding was defined. Grids are generated and energetics was only pre-calculated within this site. This site was defined as a 3D cubic box, for which user-given dimensions are needed. The box size should be large enough that it could efficiently enclose ligands to be docked. All ligand atoms should be placed in this box. There is also an inner box, in which the center of the docked ligands must reside. Inner box dimensions were kept in their default value (10 Å). For YopE, box size was determined according to the size of the known ligands (18 Å). Since SopE does not have any native or known ligands, its box size was kept larger as a safe option (25 Å). In addition to size, center location of the box was provided. Arg144 for YopE and GAGA loop for SopE were selected as centers.

To include constraints in Glide docking, all constraints must be defined in grid generation steps under the Constraints tab. Positional, hydrogen bond, metal or hydrophobic interaction constraint can be determined, based on the protein's binding site information. Maximum allowable constraints that can be determined are 10. For YopE, one hydrogen bond constraint was defined on Arg144 residue. For SopE, three hydrogen bond constraints were defined on Gly166, Gly168 and Ala169.

In Rotatable Groups tab, the hydroxyl groups in residues such as Ser, Thr, and Tyr and the thiol group in Cys can be selected as flexible, if it is known that their different orientations improve ligand interactions. Since no prior knowledge was available, the rotations are not allowed for both YopE and SopE to limit the screening time.

In Excluded Volumes tab, ligands can be prevented from occupying some regions within the binding site. If there is a region near the binding site that should not be occupied by ligands, this option can be used. No volumes within the protein surface were excluded throughout the simulations, both for YopE and SopE.

After receptor, binding site and enclosing box size and center position were determined with constraints, grid files were generated for later use in docking. An example of grid representation with enclosing and inner box around is shown in Figure 3.3.

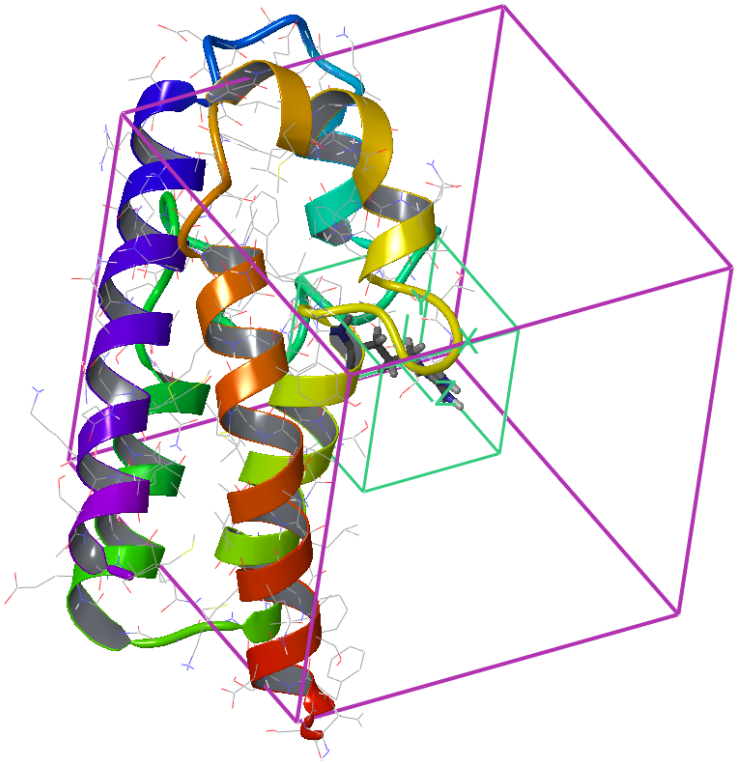


Figure 3.3. Grid box generated for YopE centering Arg144 (PDB:1hy5, chain A).

3.5. Small Molecule Database Preparation

In the initial stages of a virtual screening, it was necessary to prepare the database of compounds to be screened. For this purpose ZINC database was used, which is a collection of commercially-available molecules that can be used in virtual screening applications. ZINC database is provided by the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California San Francisco, containing 3D formats of approximately Åmillion compounds. ZINC offers subsets by a total of 219 vendors. In this project, only big vendor subsets were obtained in mol2 file format. List of all big vendors and total number of database molecules are represented in Table 3.1.

Table 3.1. Names and access dates of used vendors.

Vendor name	Access Date
Aldrich CPR	2011-09-10
Asinex	2011-09-10
AsisChem	2011-09-10
ChemBridge	2011-09-10
ChemDiv	2011-09-10
Chemical Block	2011-09-10
Enamine	2011-09-10
IBScreen	2011-09-10
Labotest	2011-09-10
Life Chemicals	2011-09-10
NCI Plated 2007	2011-09-10
Otava	2011-09-10
PBMR Labs	2011-09-10
Pharmeks	2011-09-10
Princeton BioMolecular Research	2011-09-10
Ryan Scientific	2011-09-10
Specs	2011-09-10
TimTec	2011-09-10
UORSY	2011-09-10
Vitas-M	2011-09-10
Total number of entries	498883

All database molecules were combined and prepared using Phase Database Generation module. In this module, multiple mol2 structure files were given as input. Database was prepared with the following steps, successively:

- Structures were prepared using the embedded LigPrep and ConfGen scripts. Tautomers (structural isomers) and ionizations states of molecules were generated using Epik in pH of 7±2.
- Low energy conformers were also generated. Maximum number of conformers was selected as 100, and up to 10 conformers per rotatable bond were retained. These are the default values.

- States and conformers with very high energy were automatically removed.

3.5.1. Assessment of Druglikeness

The initial database was reduced in size by applying several physical and chemical filters, in an attempt to have a database of compounds that have physical properties and chemical functionality consistent with the majority of known drugs. Common filtering protocols for 'druglikeness' were applied to the database generated from ZINC. Lipinski's rule-of-five, which is an empirical set of rules based on molecular weight, lipophilicity and hydrophobicity that provides a simple profile for orally bioavailable compounds were applied to the prepared database. Moreover, duplicates were also eliminated. Filtering was processed with running embedded QikProp script in Phase Database Generation module. Molecular properties of the molecules were calculated with this script. Molecules were pre-filtered with Lipinski's rule of five (molecular weight<500, hydrogen bond donor<5, hydrogen bond acceptor<10, partition coefficient<5). In addition, molecules with reactive functional groups were removed. Defined set of reactive groups are available in Appendix A:. At the end, after conformer generation and filtering, a total of 25.8 million conformations were generated from 498883 molecules. The collection of molecules was exported as a Phase database.

3.6. Phase Ligand-based Pharmacophore Modeling

A pharmacophore is an abstract concept, which defines the spatial arrangement of groups or atoms within a small molecule that would be important upon ligand binding to a protein. The IUPAC defines a pharmacophore to be "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response." [102]. In the case where there are experimentally or computationally verified ligands for the specific target, pharmacophore can be built to initially filter the compound database for a more effective screening. A pharmacophore model is dependent on the following features: h-bonding interactions, lipophilic areas, aromatic interactions and electrostatic interactions. Pharmacophore modeling in Phase module consists of five main steps, which are outlined in the following subsections.

3.6.1. Preparing Ligands for Pharmacophore Development

Pharmacophore model was developed using the experimentally determined 23 inhibitors of YopE (Table 2.2 and Table 2.3). All-atom 3D structures of each known inhibitor were needed for realistic representation of the structures. As explained in Section 3.3, structures of inhibitors were already converted to 3D using LigPrep with different tautomerization and ionization states at neutral pH, and energetically minimized conformers for each molecule were generated with ConfGen. Therefore, additional preparation was not needed. However, it is possible to prepare ligands directly in Phase pharmacophore development workflow.

3.6.2. Creating Pharmacophore Sites

In pharmacophore development, the second stage was to represent each inhibitor structure, along with their different states and conformers, by a set of points in space. When these set of points were aligned, some of them were found to coincide with each other, which indicates a structural feature. These common structural features are also known as pharmacophore sites. A built-in set of six pharmacophore features are available in Phase, which are:

- Hydrogen bond acceptor (A)
- Hydrogen bond donor (D)
- Hydrophobic group (H)
- Negatively charged group (N)
- Positively charged group (P)
- Aromatic ring (R)

Phase locates the hydrogen bond acceptor site on a surface-accessible atom carrying at least one lone pair. Similarly, hydrogen bond donor site is located on a hydrogen atom that is attached to a heavy atom. Mappings of donor and acceptor sites on structures are represented in Figure 3.4.

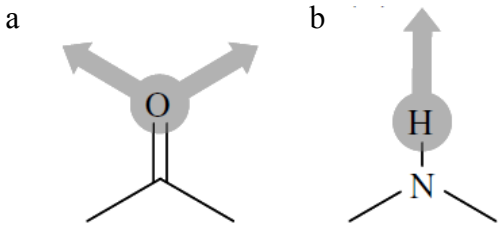


Figure 3.4. An example of (a) hydrogen bond acceptor (b) hydrogen bond donor pharmacophore site point.

Each ring, halogenic moiety, isopropyl group, t-butyl group or chain having at most four carbons is assigned as a single hydrophobic site. Likewise, an aromatic ring site is located at the center of each ring, for which a two-headed normal vector is placed. Negatively and positively charged group sites were located on the corresponding group's charged atom. Examples of hydrophobic, aromatic ring and charged group pharmacophore sites are represented in Figure 3.5.

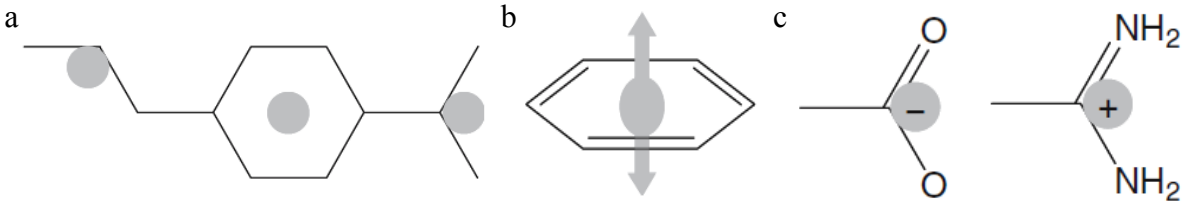


Figure 3.5. An example of (a) hydrophobic (b) aromatic ring (c) charged group pharmacophore site point.

With this procedure, all possible pharmacophore sites were defined for each inhibitor in Phase. Pharmacophore features determined on an inhibitor can be visualized in Maestro workspace.

3.6.3. Finding Common Pharmacophores

In this stage, generated pharmacophores from all conformations were combined and examined. Phase uses a very complex tree-based partitioning technique [80, 103] to identify common pharmacophores, which will be simply explained in this section.

Common set of pharmacophore features that are observed repetitively in inhibitors with similar spatial arrangements were identified and grouped. Phase allows maximum number of sites to be seven in a group, whereas the minimum value is three and default value is five. According to this limitation, all possible grouped combinations were listed, e.g. AADDR. Then, all groups were individually investigated and tree-based partitioning technique was applied. If the grouped pharmacophore points do not coincide with at least one arbitrary pharmacophore site of each inhibitor, it was eliminated. With this elimination, only common pharmacophore groups, which are then called hypotheses, were kept for further use.

3.6.4. Scoring Hypotheses

In this stage, pharmacophore hypotheses were scored according to their alignments to the input molecules. The quality of each alignment is measured in three ways:

- Alignment (site) score, which is the root-mean-squared deviation (RMSD) in the site-point positions.
- Vector score, which is the average cosine of the angles formed by corresponding pairs of vector features (acceptors, donors, and aromatic rings) in the aligned structures.
- Volume score, which is based on the overlap of van der Waals spheres of the non-hydrogen atoms in each pair.

There is also an additional scoring, which estimates the selectivity of hypotheses. Selectivity is an empirical prediction defined on log scale, which calculates the fraction of molecules that could match the hypothesis, whether it is an active or inactive ligand. For example, a selectivity value of 2 means that, 1 molecule out of 100 (log 10^2) arbitrary ligand molecules would match the hypothesis, regardless of their activity value. Therefore, higher selectivity is favored, since it induces uniqueness to the ligand set. However, it is reported that selectivity score does not affect the final scoring as site, vector and volume scores do [80]. Overall, a non-weighted combination of site, vector, volume and selectivity scores yields the final scoring function of the hypothesis, which is the survival score. Weights of individual scores can be changed and customized. Scoring functions were kept in their default forms in this work.

3.6.5. Building QSAR Model

A QSAR model can be incorporated to the hypotheses if activities of the input inhibitors are known in terms of their IC_{50} concentrations. Since this data is available for YopE inhibitors, QSAR model was built, in order to identify the best hypothesis that represents the inhibitor set. Basically in QSAR, inhibitors are aligned according to the hypothesis pharmacophore sites, and their activity values were correlated with their 3D structural information. For this purpose, total set of inhibitors were divided into two sets: training set and test set. Training set was used to fit the QSAR model, and test set was used to validate the built QSAR model, to check if it is able to predict activities of test set molecules. These sets can be randomly or manually generated.

Before constructing the model, the space occupied by the training set inhibitors were defined by a collection of equally-sized small cubes of 1 Å side. Cubes can be occupied by either a pharmacophore site or atoms of the molecules, or both. Computationally, cubes can be defined on volume bits, whose value was determined from the number of occupation coming from different molecules (i.e. zero volume bits for no occupancy). Similarly, space occupied by a single molecule can be represented by a string of zeros or ones. A sample representation of this procedure is shown in Figure 3.6. Here, volume occupied by the molecule was mapped on a grid. According to cube occupancy of atoms, molecule can be mapped onto volume bits. Each color represents a different site point category, e.g. hydrogen bond donor, acceptor or a charged group.

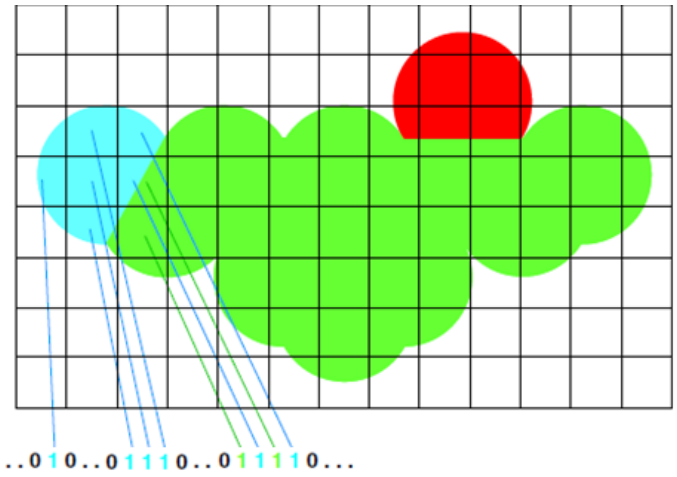


Figure 3.6. Representation of a molecule on grids and volume bits [80].

Each volume bit string provided independent variables of the training set. To these independent variables, partial least squares regression (PLS) method was applied, and a regression coefficient was obtained for every bit. According to the coefficients of the bits and atoms occupied by corresponding cubes, QSAR model determined the structural features within a molecule that increases or decreases the activity. Strength of the regression is determined by the number of PLS factors included to the fit. If there are more training set molecules, PLS factors can be increased, which results in a better fit. Number of the training set compounds multiplied by 0.20 gives the maximum PLS factors that can be selected. PLS factors were selected as three since 16 training set compounds were available, which will be further discussed in Section 4.1.3.1. Regression was carried out with training set compounds, and test set was used to externally validate the model. Quality of the QSAR model was measured with a series of statistical parameters, explained in Section 4.1.3.1.

According to the QSAR model, the best hypothesis was selected and the database generated from ZINC was pre-filtered with 3D similarity analysis. All of the hypothesis site points were used to match with the database molecules. Site matching tolerance was selected as 2 Å, which is default. In addition, molecules having vector score lower than 0.65 and having volume score lower than 0.25 were rejected, which is also default.

3.7. Structure-based Pharmacophore Modeling (E-pharm)

A novel method discovered for pharmacophores building based on receptor structure was used. This method was chosen for SopE target, since structural information of known or native inhibitors against SopE are not available, therefore ordinary pharmacophore modeling could not be done. Instead, E-pharmacophores script with Phase and Glide module was used for hypothesis generation for SopE receptor.

First, a docking was performed with the receptor SopE and the fragment library provided from Schrödinger [104]. This library was prepared from 441 unique small fragments, which are derived from molecules in the medicinal chemistry literature. The set includes a total of 667 fragments with accessible low energy ionization and tautomeric

states and metal and state penalties for each compound from Epik. Molecular weight range of fragments is 32-226 with 6-37 atoms. Ionization/tautomer states range between 1 and 7.

All fragments obtained from Schrödinger were docked to SopE, and the results of the docking calculations were recorded on Glide XP descriptor file. Grid generated with prepared SopE structure was used, and an extra precision Glide docking was performed without any constraints. The XP descriptor file, which includes energetic terms of docking along with receptor structure and fragment binding modes, was given to E-pharmacophores script as an input.

The script read the energetic terms from the Glide XP descriptor file, and mapped the energetic contributions that make up GlideScore, onto the atom centers. Energetic contributions of all 667 fragments were mapped, and summed on each atom center. Hence, each possible pharmacophore site on atoms was quantified with their summed energy value. These sites were ranked individually according to their energies, and a set of pharmacophore sites that would make a hypothesis was determined. A workflow of the procedure is represented in Figure 3.7.

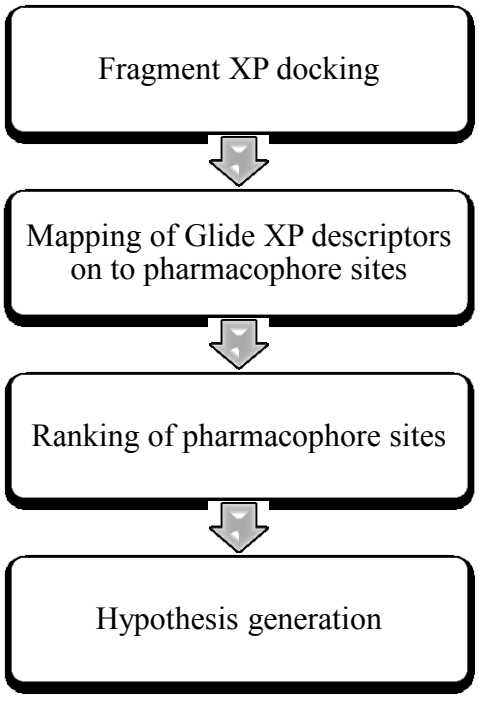


Figure 3.7. Overview of structure-based pharmacophore modeling.

After the hypothesis was determined, the database generated from ZINC was filtered according to the 3D similarity. All Phase parameters were kept in their default values for database matching, as specified in the previous section.

3.8. Glide Docking Protocol

For docking, Glide module was used, which investigates interactions between a set molecules and a target receptor, which is a protein. Molecules that passed the pharmacophore pre-filtering stage were docked to the receptor with Glide. Glide requires a grid generated around binding site prior to docking, as explained in Section 3.4. Energetic terms of the receptor were already calculated and stored in grids. These constructed grids were used for receptor structures YopE and SopE, and pre-filtered database molecules were provided as ligands for docking.

Glide basically searches for the possible locations of ligands within the binding site of the protein. Glide keeps receptor atoms frozen during docking, whereas ligands can be optionally rigid or flexible. When flexibility is allowed, for each input molecule, Glide generates conformers. Position and orientation of a molecule (or ligand) with respect to the receptor is defined as the ligand pose [87]. Glide uses a series of hierarchical filters to the generated poses, in order to eliminate unreasonable poses, which is handled by Glide's sampling algorithm. After filtering, ligand poses are scored using Schrödinger's GlideScore [87] scoring function, which is based on ChemScore [105], but includes a steric-clash term, adds rewards and penalties such as buried polar terms, amide twist penalties and hydrophobic enclosure terms. Overall, conformational sampling and scoring are two aspects that are important in docking calculations. Sampling searches for binding orientations or conformations of ligand near a binding site, whereas scoring predicts the binding tightness of a ligand to the receptor. Glide uses random search algorithm and uses an empirical scoring function, which is given in Equation 3.1. Energetic contributions to the GlideScore are represented in Table 3.2.

$$GlideScore = 0.05 \times vdW + 0.15 \times Coul + Lipo + Hbond + Metal + Rewards + RotB + Site \quad (3.1)$$

Table 3.2. GlideScore components [87].

Component	Description
vdW	Van der Waals energy. This term is calculated with reduced net ionic charges on groups with formal charges, such as metals, carboxylates, and guanidiniums.
Coul	Coulomb energy. This term is calculated with reduced net ionic charges on groups with formal charges, such as metals, carboxylates, and guanidiniums.
Lipo	Lipophilic term derived from hydrophobic grid potential. Rewards favorable hydrophobic interactions.
Hbond	Hydrogen-bonding term. This term is separated into differently weighted components that depend on whether the donor and acceptor are neutral, one is neutral and the other is charged, or both are charged.
Metal	Metal-binding term. Only the interactions with anionic or highly polar acceptor atoms are included.
Rewards	Rewards and penalties for various features, such as buried polar groups, hydrophobic enclosure, correlated hydrogen bonds, amide twists, and so on.
RotB	Penalty for freezing rotatable bonds.
Site	Polar interactions in the active site. Polar but non-hydrogen-bonding atoms in a hydrophobic region are rewarded.

In Glide module, three different docking modes are available, namely, high-throughput virtual screening (HTVS), standard precision (SP) and extra precision (XP). These three modes majorly differ in their sampling methods and scoring functions. Scoring function of HTVS and SP precision is essentially same, but HTVS mode uses a much more restricted sampling and refining, therefore it is recommended to use for rapid screening of very large number of compounds [87]. HTVS mode was not used in this work. Standard precision mode performs a more through sampling compared to the HTVS mode. XP, on the other hand, employs a much more extensive sampling, with greater conformational requirements and ligand-receptor shape complementarity, in order to eliminate false positives. Additionally, XP scoring function includes several penalties, such as a charge penalty for having a ligand charge in a region without water.

Therefore, in this work, Glide SP mode was employed to pre-filtered database molecules, and top 10% of molecules in terms of GlideScore were re-docked to YopE and SopE with Glide XP mode. Between SP and XP docking, a minimization was performed

on ligands to relieve their strains. This minimization was carried out via Premin script in MacroModel module, which employs a default 4r distance-dependent dielectric model and uses Macro-Model's efficient truncated-Newton minimizer up to 500 iterations. Flexibility of ligands was allowed for both docking modes. Epik state penalties for ligands were incorporated to the docking score. Advanced docking settings, e.g. sampling parameters and number of poses to keep for energy minimization, were kept in their default values. Docking and scoring of atoms having more than 300 atoms and 50 rotatable bonds were not allowed. To soften the potential nonpolar parts of the ligand, vdW radius of ligand atoms, having less than 0.15 partial charges were scaled by 0.8. This scaling parameter was also the default setting of Glide. Constraints that were defined during receptor grid generation were reviewed. One hydrogen bond constraint was selected for YopE, whereas at least one hydrogen bond constraint out of three was selected for SopE. Constraints were used for both SP and XP modes. At the end of the docking run, a pose viewer file was obtained, in which binding modes of ligands on the receptor and docking energetics terms were present.

3.8.1. Post-docking Evaluation

Docking results can be improved and evaluated with the help of post-processing scripts. In this work, rescoring with ligand strain correction, calculation of binding free energies of ligands and enrichment metrics were performed.

3.8.1.1. Ligand Strain Calculation. After docking was performed, the output file was fed to Strain Rescore script. Since Glide performs docking with a rigid receptor, some ligand strain is allowed to compensate the lack for receptor flexibility. However, it is difficult to tell whether ligand strain is unnatural, or an indication of a false positive. Therefore, Glide developed the strain correction as a way to identify ligands with too much strain. For each ligand pose in the input file, a tightly constrained minimization and an unconstrained minimization are performed with MacroModel module. The energy difference is used to determine the GlideScore penalty. Ligands with more than 4 kcal/mol energy difference between the docked and free conformations receive penalties by default.

3.8.1.2. Enrichment Metrics. To evaluate the performance of the docking protocol, Enrichment Calculator script was utilized. Known actives were combined with a set of ligand-like inactive decoy molecules. Glide XP docking was performed with this combination and receptor structures. All molecules, including decoys and inactives, were ranked and sorted according to the GlideScore. The ability of docking protocol to distinguish between actives and decoys were investigated. If the docking protocol is successful, ranks of the actives must be higher than the decoys. Additionally, enrichment factor was calculated, with respect to the number of total ligands (Equation 3.2):

$$EF = (a/n)/(A/N) \tag{3.2}$$

where a is the number of actives found in sample size n, A is the total number of actives, and N is the total number of ligands (decoys and actives). Enrichment factor can be defined as the concentration of the active ligands among the top scoring docking hits compared to their concentration throughout the entire database [106].

3.8.1.3. Binding Free Energy Calculation. The binding free energies of ligand-receptor complexes were estimated using Prime MM-GBSA module. The binding free energy of each ligand was calculated from the following equation:

$$\Delta G_{bind} = E_{complex\,(min)} - (E_{ligand\,(min)} + E_{receptor\,(min)}) \tag{3.3}$$

Here, sum of minimized total energy of the ligand and receptor alone were extracted from the minimized energy of the ligand-receptor complex. Prime requires pre-positioning of the ligands with respect to the receptor. Therefore, pose viewer file including binding modes of the ligands were given to Prime as an input. The obtained ligand poses were minimized using the local optimization feature in Prime, whereas the energies of complex were calculated with the OPLS_2005 force field. During the simulation process, the ligand strain energy was also considered. Flexibility of receptor atoms within a 5 Å radius from the ligand was allowed.

4. RESULTS AND DISCUSSION

Investigation of novel inhibitors targeting the bacterial pathogen *Yersinia* YopE and *Salmonella* SopE were carried out with the help of computational drug design tools. Accordingly; library formation, pharmacophore modeling, pharmacophore-based high throughput screening, molecular docking and scoring were carried out to propose a set of biochemically active molecules with inhibition potential against YopE and SopE.

4.1. YopE Results

Structure-based virtual screening (docking) approach was used for target YopE since library of small molecule compounds, target's structure knowledge and 3D representation of the binding site are available. However, docking each compound in small molecule database (approximately 25 million) to target protein required excessive CPU time. Therefore, ligand-based virtual screening (pharmacophore building) was combined with molecular docking, in order to pre-filter the small molecule database. Pharmacophore models were constructed utilizing the knowledge of a set of 23 compounds with known activity against YopE [46]. QSAR method was also applied to predict the activity of hit compounds which have not been tested *in vitro*. Compounds in small molecule database that passed pre-filtering were docked to target YopE in standard (SP) and extra precision (XP) mode, successively. With scoring, ranking and post-docking analysis, number of potential leads was reduced to a manageable size for visual inspection. Workflow of this hybrid virtual screening methodology is presented in Figure 4.1.

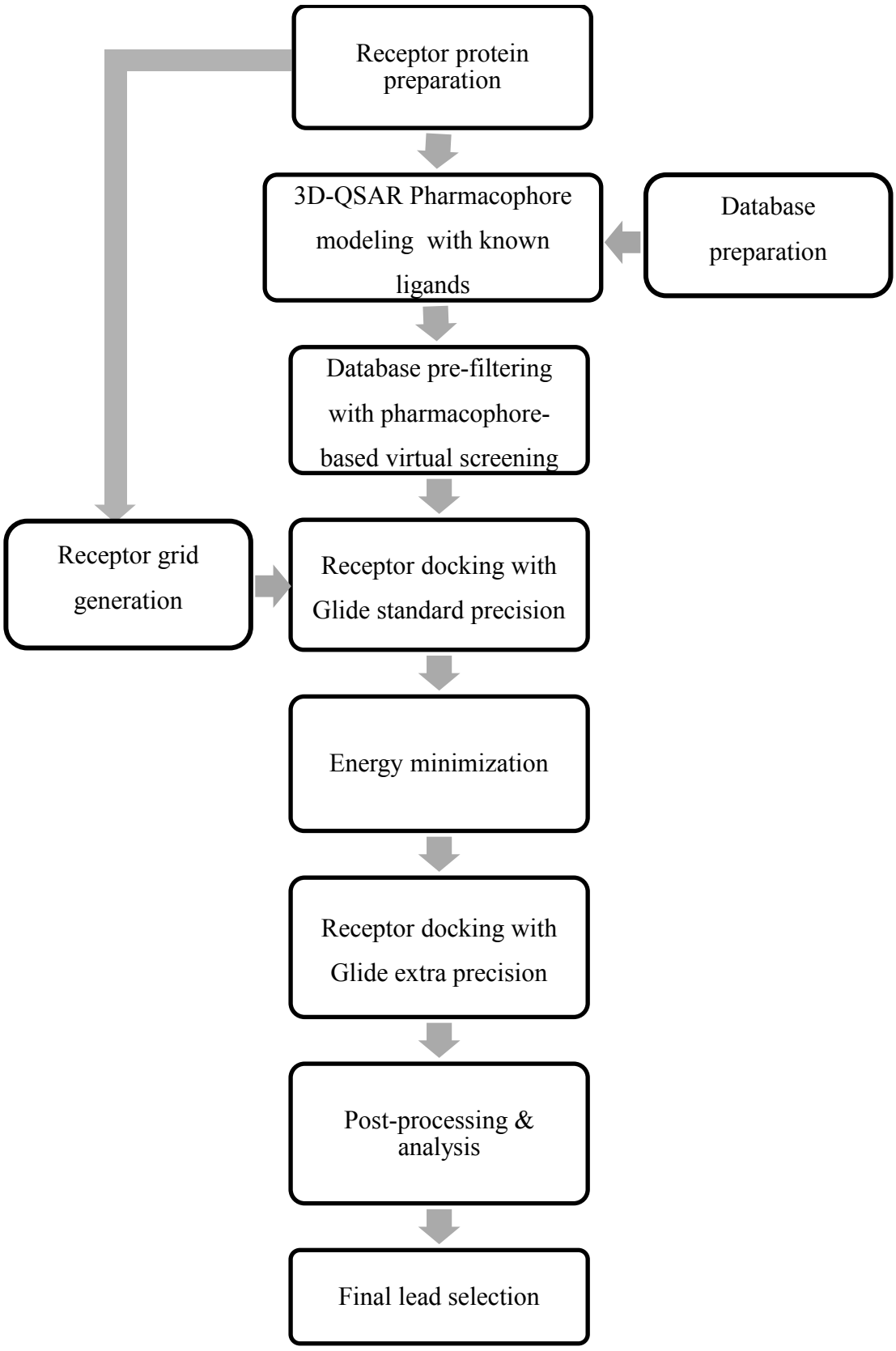


Figure 4.1. Schematic workflow of virtual screening and docking of YopE.

4.1.1. Remarks on Receptor Preparation and Grid Generation

Initially, structure of target protein YopE (PDB code: 1hy5) was prepared using Protein Preparation Wizard module in Maestro workspace. Since PDB file does not contain hydrogen atoms, these were added manually to protein's heavy atoms, and bond orders were automatically assigned.

The structure was subjected to the Prime refinement tool in Protein Preparation Wizard to predict missing side chains and loops in the receptor. No residues or loops were missing in PDB file; therefore no changes were induced by this script. Similarly, there were no metal ions or cofactors present; thus their treatment was omitted. There were a total of 67 explicit water molecules in protein structure. All water molecules were deleted since there was no information on the presence of structural waters that could mediate receptor-ligand interactions. Also, since receptor was rigid during docking, remaining waters would have been treated as a part of the receptor environment and have interfered with ligands. The structure included more than one chain (two dimers), only chain A was kept for further use.

Addition of hydrogen atoms to heavy atoms was not sufficient since docking requires predicting correct protonation states of some residues. An automated optimization of hydrogen bonding network was carried out at neutral pH, and proper orientations of aspartic acid, cysteine, glutamic acid, histidine and lysine residues were determined. Possible steric clashes were also prevented with hydrogen network optimization. After assigning charge and protonation states, final energy minimization was done using Impref module in Protein Preparation Wizard. Additionally, residue numbering of PDB was changed and adjusted to residue numbering of its publication. Figure 4.2 represents structure of YopE before and after protein preparation.

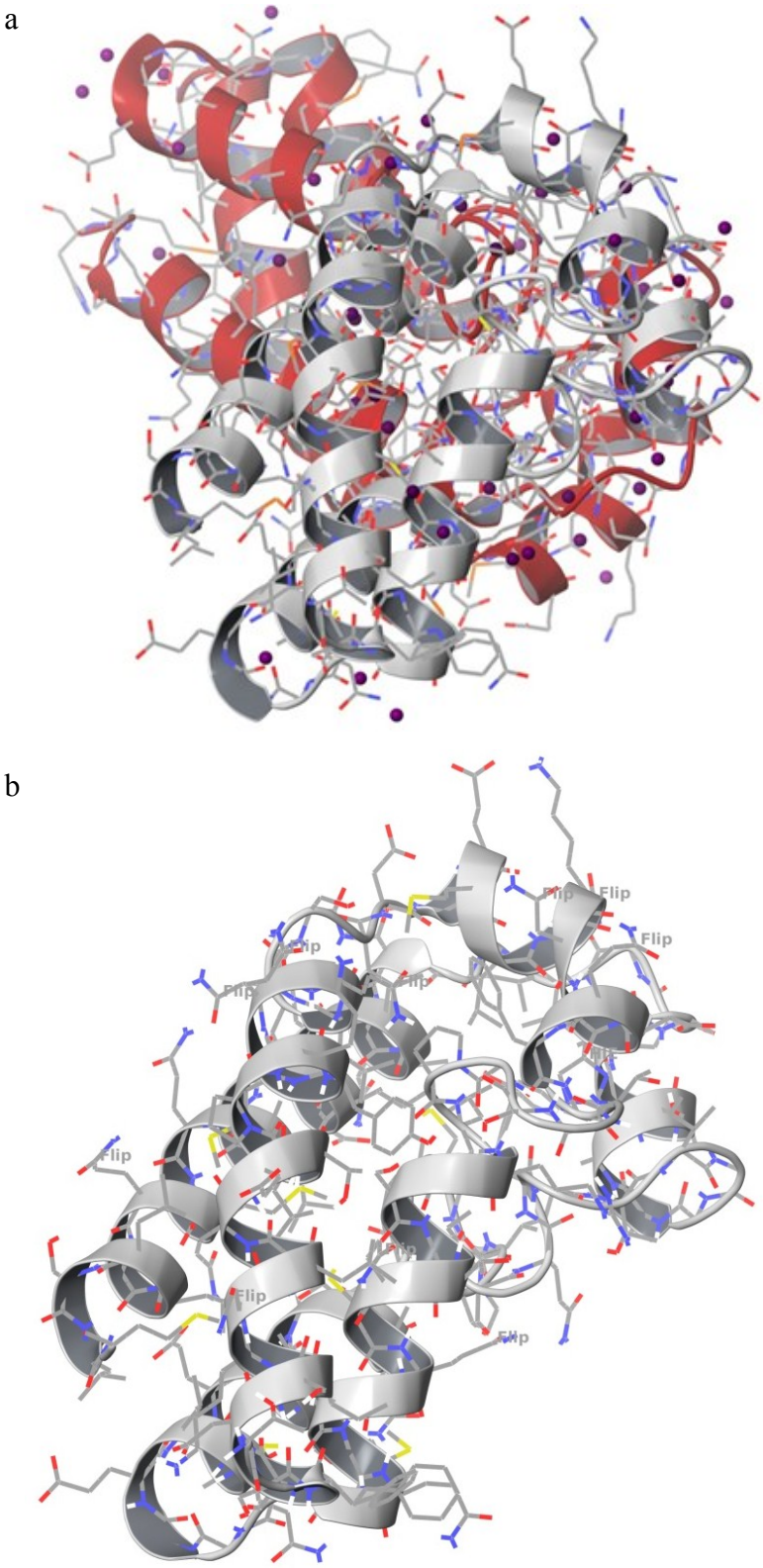


Figure 4.2. 3D structure of YopE (a) before preparation (b) after preparation. Both ribbon and tube molecular representations are displayed. Secondary structures are colored by chain name and waters are represented as ball & stick.

Prior to molecular docking, the receptor grid was generated, where an energetically favorable site of interest (possible binding site) on the receptor was identified. The grid was generated using the prepared structure file of YopE. Due to its importance on catalytic activity [41], Arg144 was selected as the center of the enclosing grid box defining the binding site. Box size was determined to make sure that it efficiently encloses the space that docked ligands are likely to occupy. The side length of this cubic box was selected as 18 Å since sizes of known inhibitors vary between 15-18 Å [46]. A hydrogen bond constraint on Arg144 was defined. Rotation of hydroxyl groups on certain residues (serines and tyrosines) was not allowed since an implication on the importance of these groups' flexibility was not found. Default vdW radius scaling settings were employed for generation of the receptor grid. Enclosing box that was used for grid generation was represented in Figure 4.3 and Figure 4.4.

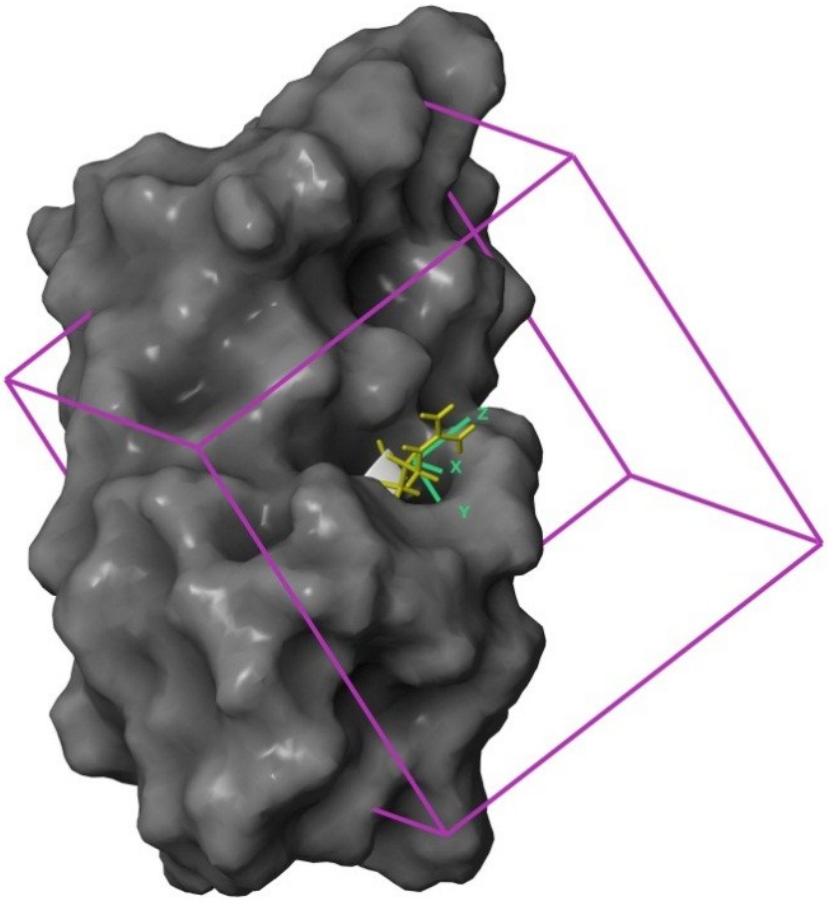


Figure 4.3. Enclosed grid box of YopE in molecule surface representation. Centered Arg144 residue is shown in wire and ball & stick representation, respectively.

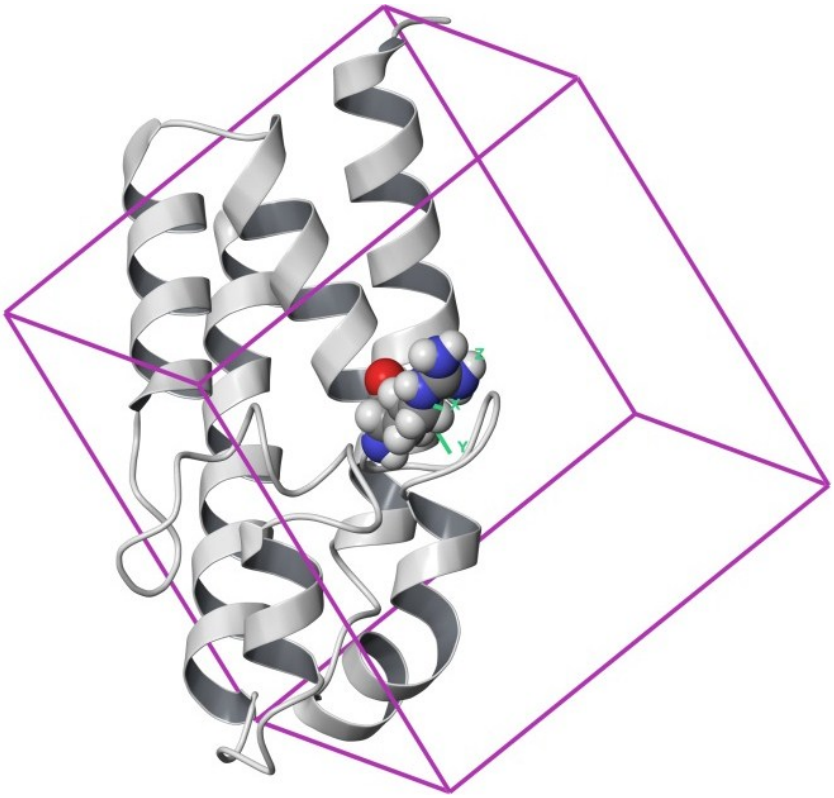


Figure 4.4. Enclosed grid box of YopE in ribbon representation. Centered Arg144 residue is shown in wire and ball & stick representation, respectively.

4.1.2. Interactions and Binding Modes of Inhibitors with YopE

23 chemically synthesized compounds belonging to a class of acylated hydrazones of salicylaldehydes that inhibit the YopE secretion *in vitro* were used as active ligands (Table 2.2 and Table 2.3). Since ligands were not available in any structure file format, they were built from scratch in 2D form in Maestro workspace. 2D structures of ligands were converted to 3D using LigPrep module in Maestro, generating all possible states at neutral pH using Epik. After ligand preparation, possible conformers of ligands were generated using ConfGen module.

Since YopE structure does not include any co-crystallized ligands, binding modes and key interactions of ligands with the receptor could not be obtained. In order to get a feel of possible important interactions and binding modes, an unconstrained extra precision docking was carried out with 23 inhibitors, as described in Section 3.8. Arg144 constrained was not applied intentionally, to see if ligands bind to this residue naturally or not. After

Glide XP docking, ligand interactions were visualized in Maestro workspace and represented in Figure B.1-Figure B.24, and docking results in terms of GlideScores are listed in Table B.1. The most common interactions between YopE and 23 inhibitors were observed in residues Arg144, Thr183 and Ile184 of YopE. These three residues are claimed to interact with both of the switch regions of G-proteins and GTP [41]. Additionally, importance of Arg144 in catalytic activity is reported by various studies by mutation analyses [23, 24]. It was observed that Arg144 made hydrogen bonds with the ligands' oxygen atom in formic hydrazide group (NNC=O); whereas Ile184 was bound to phenol group and Thr183 was bound to nitrogen monohydride group. A representative diagram including these interactions is given in Figure 4.5. Gly182 of YopE had also notable contact with inhibitors. It is observed that Gly182 interactions mostly occurred when inhibitors did not contact with Thr183. Fewer interactions were observed on residues Thr148, Gln151, Ser179 of YopE with ligands. These residues are also claimed to contact with Switch 1 regions of GTPases [41].

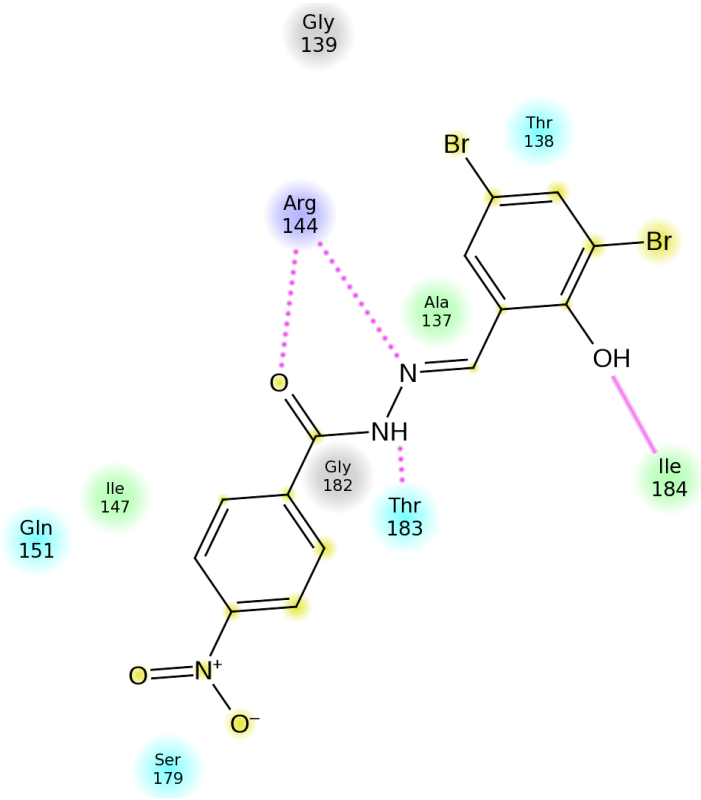


Figure 4.5. Interaction diagram of compound1 with YopE. Solid pink lines indicate backbone h-bond and dashed pink line indicate side chain h-bond. Solvent exposed areas of ligand are shown in yellow spheres.

Table 4.1. Summary of hydrogen bonds between inhibitors and YopE.

Compound	Residues of YopE							
	Ala137	Arg144	Thr148	Gln151	Ser179	Gly182	Thr183	Ile184
1		++					+	+
2		+				+	+	+
3		+				++	+	++
4	+	+					+	
5			+					
6		+					+	+
7						+		++
8						+		++
9		++		+	+		+	+
10			+					
11						+		++
12			++	+			+	
13		+					+	+
14		+					+	+
15			+	+	+			
16		+				+		++
17	+	+					+	
18		+		+			+	
19		+					+	+
20		+			+		+	+
21		+					+	+
22		+	+					
23		++		+	+		++	

Binding modes of all ligands were superimposed. Majority of ligands were located in the vicinity of Arg144 despite it was not constrained. Figure 4.6a shows binding modes of all ligands and Figure 4.6b shows binding modes of ligands that interact with Arg144.

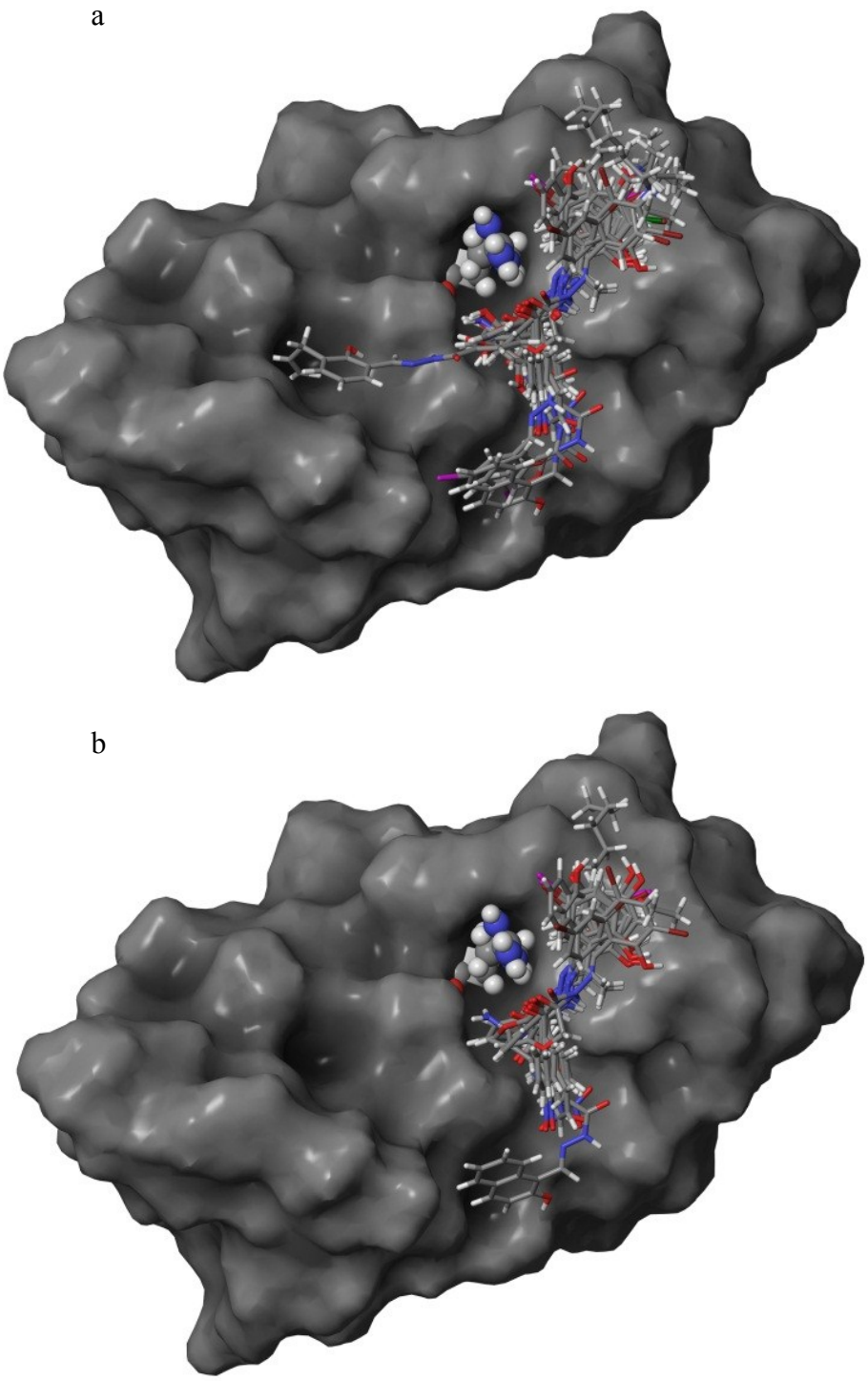


Figure 4.6. Binding modes of (a) all inhibitors (b) inhibitors interacting Arg144 residue. YopE structure is shown as molecular surface and Arg144 is shown as ball&stick.

Next, Glide XP docking was performed with Arg144 constraint. Constrained docking was performed in order to compare GlideScores of 23 compounds with database hits that will be obtained at the end of the virtual screening workflow. Docking results are given in

Table 4.2. GlideScores of 23 inhibitors varied between -3.4 kcal/mol and -6.88 kcal/mol, with an average of 4.96 kcal/mol.

Table 4.2. Glide XP scores of known inhibitors against YopE with Arg144 constraint.

Compound	GlideScore (kcal/mol)	Compound	GlideScore (kcal/mol)
1	-3.412	13	-3.992
2	-4.359	14	-5.033
3	-5.149	15	-6.748
4	-5.896	16	-5.161
5	-3.833	17	-4.527
6	-5.425	18	-4.792
7	-4.576	19	-3.542
8	-3.944	20	-5.058
9	-5.659	21	-4.316
10	-5.781	22	-6.048
11	-4.713	23	-5.140
12	-6.879	avg	-4.956

4.1.3. Pharmacophore Hypothesis Selection

23 ligands that are known to react with YopE were used for pharmacophore construction via Phase. Before starting any ligand-based approaches, it was checked that if known ligands have varying physical and chemical properties. Some of the molecular and physiochemical descriptors that characterize properties of molecules were calculated in Maestro workspace and represented in Table 4.3. Hydrogen bond donors vary between two and four whereas acceptors vary between three and six. Notable diversity was observed also in molecular weights, logP values, atom counts and bond counts of the YopE ligands. These descriptors, along with many others, were used upon modeling of quantitive-structural activity relationship. Therefore, their diversity is favored to ensure a robust modeling in QSAR step.

Table 4.3. Physiochemical descriptors of known YopE inhibitors.

Compound	MW	Chemical formula	Heavy Atom count	Bond count	H bond donor	H bond acceptor	IC50 (μM)
1	443.06	$C_{14}H_9Br_2N_3O_4$	23	34	2	4	12.5
2	399.45	$C_{21}H_{25}N_3O_5$	29	40	3	6	19.5
3	322.32	$C_{18}H_{13}N_2O_4$	24	37	3	4	17.7
4	320.35	$C_{19}H_{15}N_2O_3$	24	34	2	4	49.6
5	339.35	$C_{18}H_{16}N_3O_4$	25	35	2	5	22.3
6	246.22	$C_{12}H_{10}N_2O_4$	18	22	3	5	57.0
7	290.32	$C_{18}H_{14}N_2O_2$	22	33	2	3	28.9
8	399.05	$C_{13}H_9Br_2N_3O_2$	20	28	2	5	21.8
9	286.09	$C_9H_8Br_1N_3O_3$	16	20	2	4	14.5
10	272.26	$C_{14}H_{12}N_2O_4$	20	29	4	5	21.8
11	309.15	$C_{14}H_{10}Cl_2N_2O_2$	20	28	2	3	17.3
12	524.06	$C_{14}H_8I_2N_2O_4$	22	33	4	5	25.0
13	457.09	$C_{15}H_9Br_2N_3O_4$	24	37	2	4	45.6
14	507.07	$C_{14}H_{11}I_2N_3O_2$	21	30	4	4	23.1
15	262.22	$C_{12}H_{10}N_2O_5$	19	26	4	5	15.8
16	312.37	$C_{17}H_{20}N_4O_2$	23	33	2	6	40.0
17	301.26	$C_{14}H_{11}N_3O_5$	22	31	3	5	21.8
18	413.08	$C_{14}H_{11}Br_2N_3O_2$	21	31	2	4	70.0
19	537.05	$C_{14}H_9I_2N_3O_4$	23	34	2	5	51.0
20	320.15	$C_{13}H_{10}Br_1N_3O_2$	19	25	2	4	18.3
21	318.76	$C_{16}H_{15}Cl_1N_2O_3$	22	31	4	5	45.0
22	335.37	$C_{19}H_{16}N_3O_3$	25	36	2	4	76.0
23	364.99	$C_9H_7Br_2N_3O_3$	17	23	2	4	80.0

Basically, each ligand structure was represented by a set of points in 3D space, and the geometries of these ligands were compared in order to reveal matching features. Six key features that are likely to be important for drug-receptor interactions were automatically identified for each ligand through the use of atom types. All six built-in

pharmacophore features (or sites) were searched: hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic group (H), negatively charged group (N), positively charged group (P) and aromatic ring (R). Number of pharmacophore sites present in each ligand is represented in Table 4.4. None of the ligands had a charged group; therefore four features were used upon pharmacophore building.

Table 4.4. Pharmacophore sites of known YopE inhibitors.

Name	A	D	H	N	P	R	Total
compound1	3	2	2	0	0	2	9
compound2	4	3	2	0	0	2	11
compound3	5	4	0	0	0	3	12
compound4	4	2	0	0	0	3	9
compound5	3	2	1	0	0	2	8
compound6	5	3	0	0	0	2	10
compound7	3	2	0	0	0	3	8
compound8	4	2	2	0	0	2	10
compound9	4	4	1	0	0	1	10
compound10	5	4	0	0	0	2	11
compound11	3	2	2	0	0	2	9
compound12	5	4	2	0	0	2	13
compound13	3	2	3	0	0	2	10
compound14	3	4	2	0	0	2	11
compound15	6	4	0	0	0	2	12
compound16	4	2	2	0	0	2	10
compound17	4	3	0	0	0	2	9
compound18	4	2	3	0	0	2	11
compound19	3	2	2	0	0	2	9
compound20	4	2	1	0	0	2	9
compound21	4	2	2	0	0	2	10
compound22	4	4	0	0	0	3	11
compound23	4	4	2	0	0	1	11

An example of site point representation on compound2 can be seen in Figure 4.7. Blue spheres indicate hydrogen bond donors whereas red spheres are hydrogen bond acceptors, green spheres are hydrophobic groups and orange rings are aromatic rings.

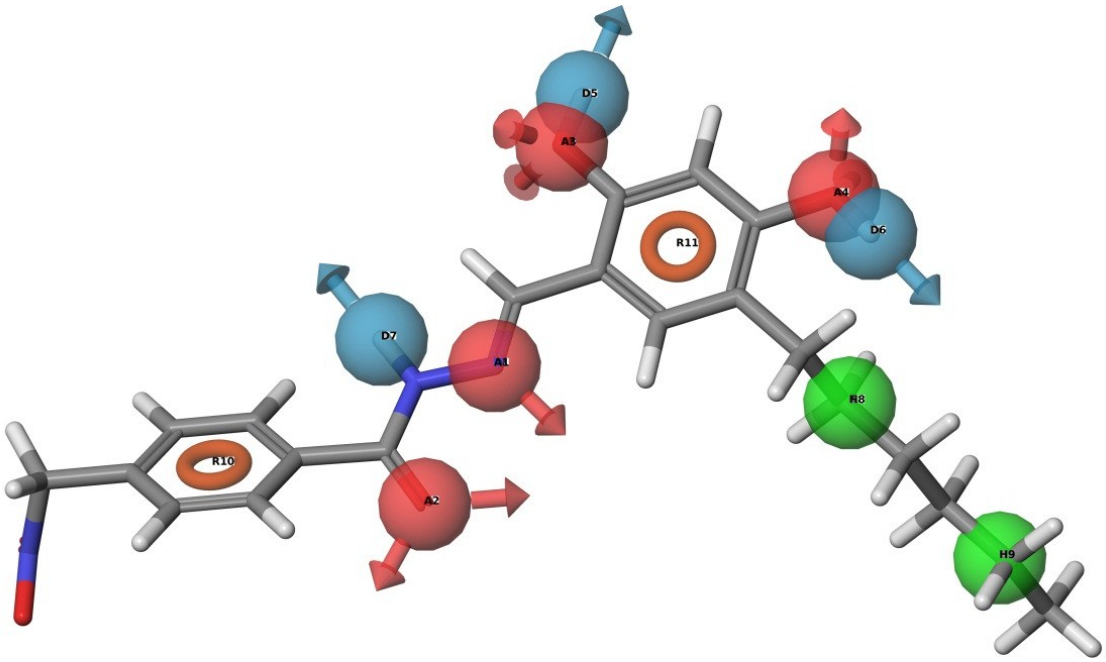


Figure 4.7. Pharmacophore site points on compound2 shown in tube presentation. Oxygen atoms are presented in red, nitrogen atoms are blue and hydrogen atoms are white.

Pharmacophores from all conformations of the ligands were aligned, and those pharmacophore sites that contain identical sets of features with very similar spatial arrangements were grouped together, e.g. AADR. All possible group combinations were generated using a tree-based partitioning algorithm [80]. These grouped pharmacophore sites, known as hypotheses, were constructed with a given number of pharmacophore sites. For a single hypothesis, Phase allows number of pharmacophore site points to be between three and seven [80]. Maximum site point was selected as five, because when more site points were included, the likelihood of finding common pharmacophores was decreased. Higher number of site points (six or seven) were not selected since some compounds had only eight pharmacophore sites, and inclusion of almost all site points to pharmacophore building was not reasonable (Table 4.4). After hypotheses were constructed, they were scored and ranked according to their coordination to the ligands, which can be seen in Table 4.5. The table shows top 10 hypotheses sorted on survival score out of 23 hypotheses produced by Phase. Complete list of hypotheses are given in Table C.1.

Table 4.5. Scores of 5-point hypotheses constructed from known YopE inhibitors.

ID	Survival	Site	Vector	Volume	Selectivity
AAADR.169	3.670	0.89	0.977	0.806	1.147
AAADR.71	3.664	0.89	0.977	0.802	1.120
AADDR.41	3.661	0.89	0.984	0.792	1.223
AADDR.49	3.647	0.87	0.982	0.791	1.193
AAADR.79	3.221	0.64	0.947	0.638	1.139
AADDR.37	3.189	0.61	0.947	0.631	1.209
AADDR.161	3.087	0.57	0.819	0.696	1.232
AADDR.43	3.002	0.67	0.882	0.453	1.235
AAADR.81	2.995	0.67	0.877	0.452	1.157
AADDR.48	2.953	0.44	0.889	0.621	1.200

The quality of each hypothesis was measured by its survival score, which is a weighted combination of site, vector, volume and selectivity scores. Site and vector scores indicate root mean squared deviation of ligand molecules from the hypothesis site positions, whereas volume score measures how ligand overlays with pharmacophore sites based on van der Waals spheres. Site, volume and vector scores range between 0-1, and higher values indicate better alignment of ligands on hypothesis. Selectivity is an empirical estimate defined at logarithmic scale that gives fraction of any random drug-like molecule to match the hypothesis [80]. Higher selectivity indicates that hypothesis is more likely to be unique to the active-set ligands, which is favorable. Selectivity score of the highest ranking hypothesis AAADR.169 is 1.147, which means the hypothesis will match $1/10^{1.147}$ of all drug-like molecules available, regardless of their activity towards YopE.

Survival score of top four hypotheses were fairly close to each other and subscores that constitute survival score, which are site, volume, vector and selectivity, did not point to a single hypothesis that represents all parameters the best. For example, 4th scoring hypothesis AADDR.49 had a higher vector score than top two hypotheses, whereas 3rd scoring hypothesis AADDR.41 has the highest selectivity score among top four hypotheses. Consequently, these hypotheses were further investigated with QSAR method. Hypothesis that yielded the most accurate ligand activity prediction was selected for small

molecule database filtering, rather than the highest survival score. Fitness of each ligand to the selected hypotheses was measured and represented in Table 4.6, which can be between -1 and 3.

Table 4.6. Fitness of all YopE inhibitors to top four hypotheses.

ID	AAADR.169	AAADR.71	AADDR.41	AADDR.49	pIC_{50}
compound1	2.89	2.94	2.88	2.94	4.902
compound2	2.71	2.74	2.70	2.74	4.711
compound3	2.78	2.81	2.78	2.81	4.752
compound4	2.12	2.13	2.08	2.09	4.305
compound5	2.79	2.82	2.79	2.81	4.651
compound6	2.89	2.85	2.89	2.85	4.244
compound7	2.23	2.15	2.24	2.11	4.540
compound8	2.87	2.85	2.87	2.85	4.662
compound9	2.83	2.78	2.82	2.78	4.839
compound10	2.94	2.89	2.95	2.83	4.662
compound11	3.00	2.96	3.00	2.96	4.761
compound12	2.94	2.87	2.81	2.84	4.602
compound13	1.83	1.85	1.92	1.79	4.341
compound14	2.85	3.00	2.84	3.00	4.636
compound15	2.87	2.80	2.87	2.75	4.802
compound16	2.80	2.78	2.71	2.78	4.398
compound17	2.83	2.78	2.73	2.78	4.662
compound18	1.91	1.90	2.00	1.85	4.155
compound19	2.88	2.95	2.88	2.95	4.292
compound20	2.96	2.94	2.96	2.94	4.737
compound21	2.90	2.91	2.90	2.90	4.347
compound22	2.08	2.11	2.07	2.07	4.119
compound23	2.85	2.80	2.85	2.80	4.097
avg combined fitness	12.21	12.18	12.17	12.11	

For each hypothesis, average of combined fitness values of 23 inhibitors was calculated. Table 4.6 shows that majority of the ligands are in a good agreement with the top four hypotheses. Average combined fitness values are also very close to each other, meaning that all hypotheses showed similar fitness values when inhibitor activities are considered. AAADR.169 and AADDR.41 have the highest fitness to compound11 while AAADR.71 and AADDR.49 have the highest fitness to compound14. Alignments of these ligands to hypotheses are illustrated in Figure 4.8 and Figure 4.9.

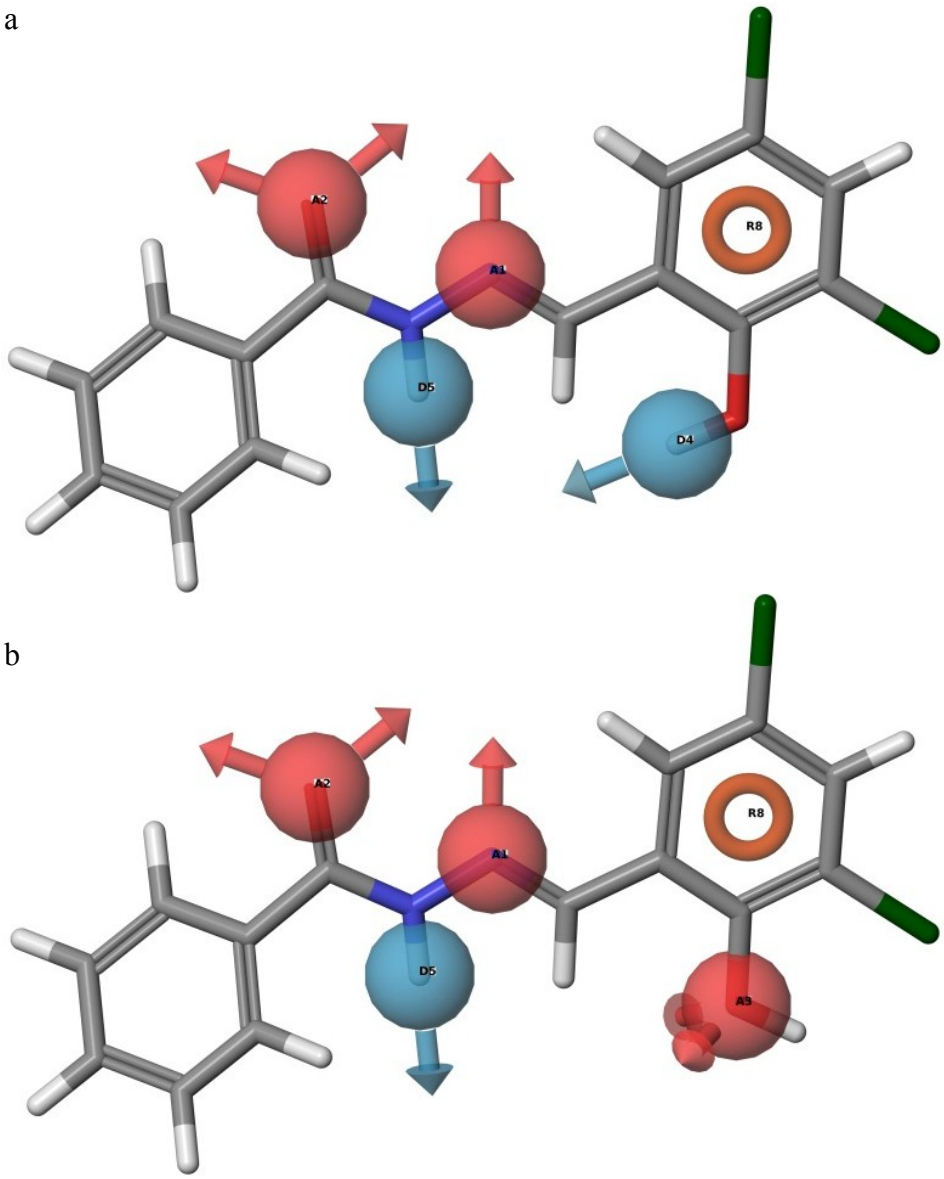


Figure 4.8. Alignment of (a) highest scoring hypothesis AAADR.169 (b) 3rd highest scoring hypothesis AADDR.41 site points on compound11. Site points were indicated as spheres.

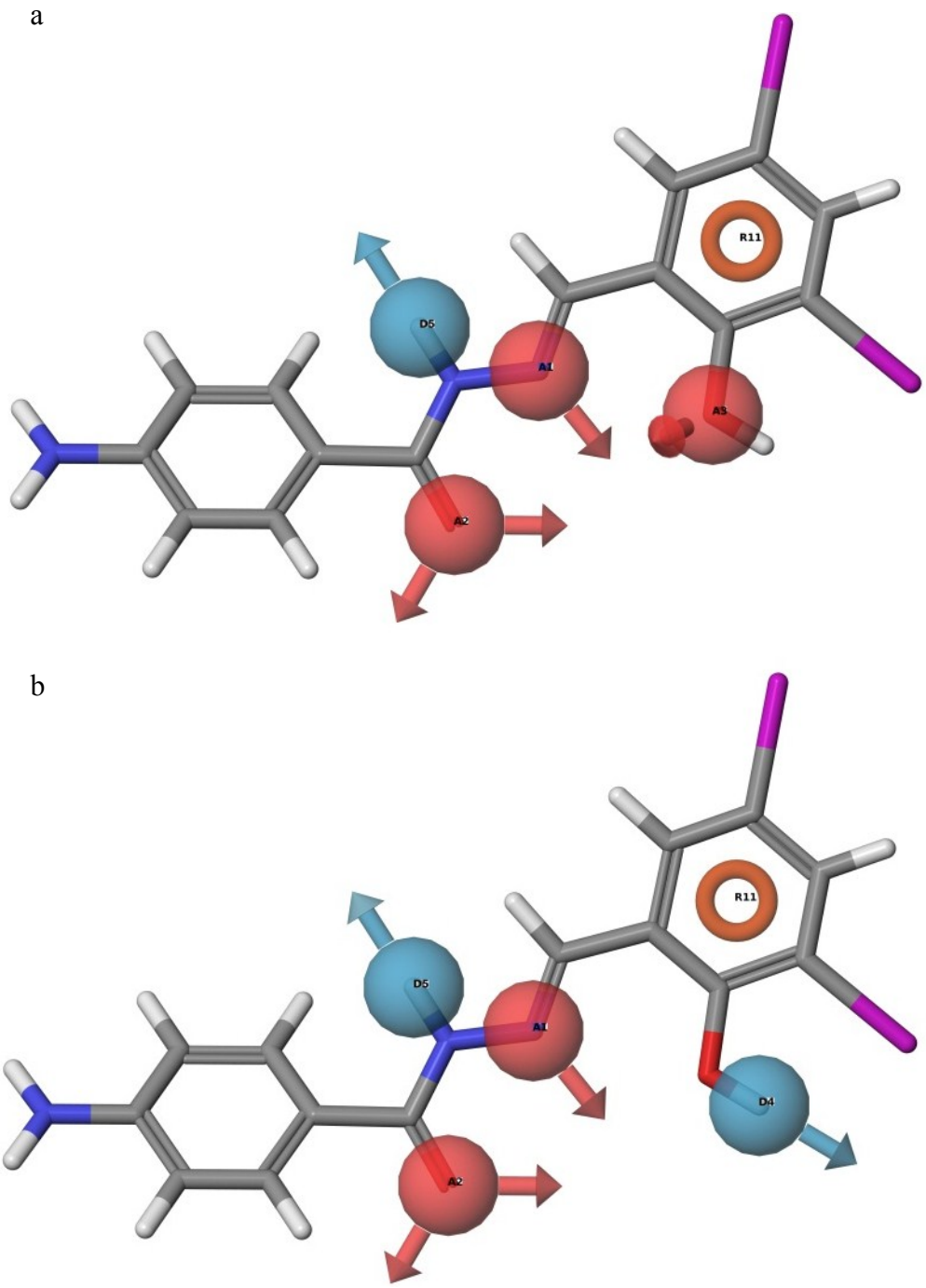


Figure 4.9. Alignment of (a) 2nd highest scoring hypothesis AAADR.71 (b) 4th highest scoring hypothesis AADDR.49 site points on compound14. Site points were indicated as spheres.

When interaction of the ligands with YopE were investigated (Table 4.1), it was observed that most common interactions involve lone O atom, -NH of formic hydrazide group and -OH of the phenol group. Although binding site information was not used upon

pharmacophore building, these three atoms were successfully aligned with the pharmacophore sites for all four hypotheses.

4.1.3.1. Quantitative Structure-Activity Relationship (QSAR) Model. After hypotheses were generated and scored, their QSAR models were developed, in order to select and validate the best hypothesis. QSAR model is a mathematical relationship between a biological activity of a molecular system and its geometric and chemical characteristics [80]. Therefore, along with pharmacophore-based alignments of ligands, their activity data was used, and aspects of their molecular structure that affect activity were investigated. The 3D-QSAR model was developed from 23 ligands that have a range of known activities using partial least squares (PLS) regression method, as described in methods section.

Active ligands were divided to training set and test set. The training set was used to generate QSAR models and test set was used to predict the activity, in order to validate the proposed models. The regression was done by constructing a series of models with an increasing number of PLS factors. It is known that the accuracy of the model increases with increasing number of PLS factors [80], therefore PLS factor was increased to maximum possible number. Since maximum PLS factor can be no larger than 1/5 number of the training set molecules [80], the training set was selected to comprise 70% of all active ligands. 23 ligands were randomly divided into a training set (16 compounds) and a test set (7 compounds). This way, regression was carried out with three PLS factor.

QSAR models were developed with the top four hypotheses, and the robustness of the models was internally validated by statistical parameters, i.e. standard deviation, root mean square error and variance ratio, which can be seen in Table 4.7. The negative logarithm of the measured IC_{50} (pIC_{50}) value of known YopE inhibitors and predicted activities are also tabulated in Table 4.8 and Table 4.9.

Table 4.7. Quantitative structure-activity relationship results for top four hypotheses.

Set	Parameters	AAADR.169	AAADR.71	AADDR.41	AADDR.49
Training	SD	0.0780	0.0745	0.078	0.0676
	R-squared	0.9317	0.9377	0.9318	0.9488
	F	54.6	60.2	54.6	74.1
Test	RMSE	0.0755	0.0801	0.1019	0.0621
	Q-squared	0.7187	0.6839	0.4887	0.8097
	Pearson-R	0.8721	0.9633	0.9086	0.9821

Statistical parameters SD, R^2, F, RMSE, Q^2 and Pearson-R were used to evaluate the quality of the QSAR model. QSAR results are shown for all number of factors in the partial least squares regression model. Here, SD is the standard deviation of the regression, F is the variance ratio, RMSE is root-mean-square error of predicted actives, Q-squared is analogous to R^2 value for the predicted activities and Pearson-R is the value for the correlation between the predicted and observed activity for the test set. SD, R-squared and F measure the predictive ability of QSAR model on training set, whereas RMSE and Q-squared measure the predictive ability of model on test set.

R-squared value shows how model successfully interprets structure-activity relationship for training set compounds. For example, R-squared value of 0.95 means that the model accounts for 95% of the variance in the observed activities for the training set. From Table 4.7, it is seen that the highest R-squared value was observed for model that was fit to hypothesis AADDR.49, although other hypotheses had also good coverage on the training set. Variance ratio, F, identifies the model that best fits to the training set, and large values of F indicate a more statistically significant regression. Along with R-squared, AADDR.49 yielded the highest variance ratio. SD value also measures the strength of the fit to the training set compounds. Lowest standard deviation, which is 0.06, was again observed in hypothesis AADDR.49. Hence using three PLS factors, satisfactory model fit to training set compounds was observed best with AADDR.49 with R^2=0.95, F=74.1 and SD=0.06 (Figure 4.10).

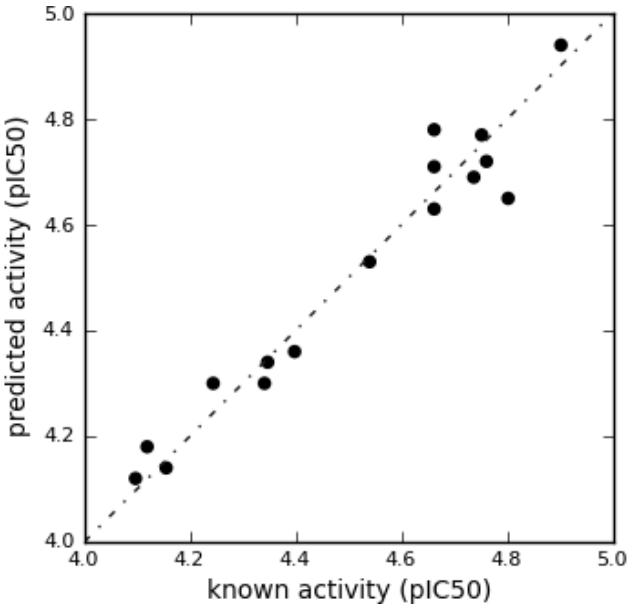


Figure 4.10. Scatter plot for the predicted and actual pIC50 values for hypothesis AADDR.49 applied to the training set compounds.

Table 4.8. Actual and predicted activities of training set ligands for AADDR.49.

Ligand Name	QSAR Set	Activity (pIC_{50})	Predicted Activity (pIC_{50})	Error (%)
compound1	training	4.902	4.94	0.78
compound3	training	4.752	4.77	0.38
compound6	training	4.244	4.30	1.32
compound7	training	4.540	4.53	0.22
compound8	training	4.662	4.63	0.69
compound10	training	4.662	4.78	2.53
compound11	training	4.761	4.72	0.86
compound13	training	4.341	4.30	0.94
compound15	training	4.802	4.65	3.17
compound16	training	4.398	4.36	0.86
compound17	training	4.662	4.71	1.03
compound18	training	4.155	4.14	0.36
compound20	training	4.737	4.69	0.99
compound21	training	4.347	4.34	0.16
compound22	training	4.119	4.18	1.48
compound23	training	4.097	4.12	0.56

High R^2 value is necessary but not sufficient condition for a good QSAR model, since predictive ability of the model is also important. Quality of model fit on test set compounds were measured by Q-squared, RMSE and Pearson-R values. Q-squared is an analogous R^2 statistic, except that it comes from applying the QSAR model to the test set. Q^2 value varies between 0.49 and 0.81 for top four hypotheses (Table 4.7). Highest Q^2=0.81 was observed in AADDR.49, which implies a satisfactory predictive ability, since it is comparable in value to R^2. Pearson-R value shows the correlation between the predicted and observed activity for the test set, for which increased numbers are favorable. Hypothesis AADDR.49 showed the best prediction on the test set compounds with 0.98 Pearson-R value and lowest root-mean-square error (0.06).

Overall, five-point pharmacophore hypothesis AADDR.49, having two hydrogen acceptors, two donors and one ring group, gave satisfactory statistical significance and predictive ability on all ligands when it was associated with a 3D-QSAR model. Figure 4.11 shows the actual versus predicted activities of test set compounds and their values were tabulated in Table 4.9 for the selected hypothesis.

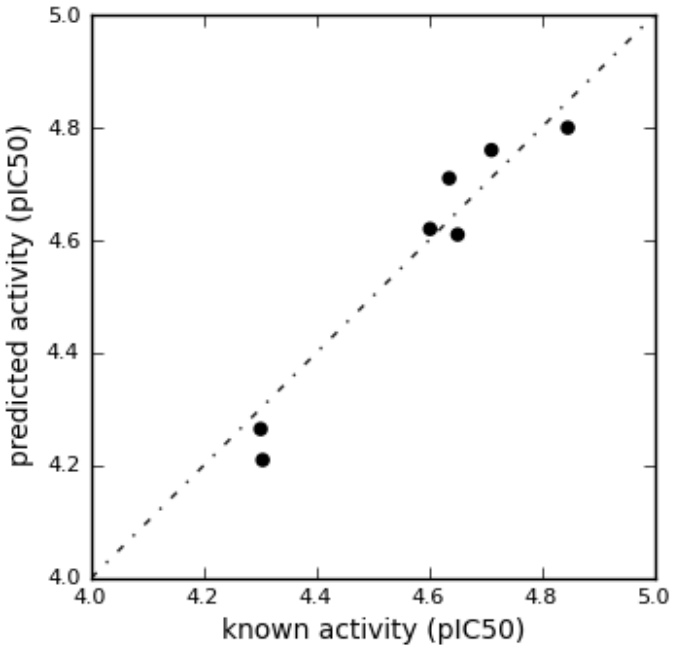


Figure 4.11. Scatter plot for the predicted and actual pIC50 values for hypothesis AADDR.49 applied to the test set compounds.

Table 4.9. Actual and predicted activities of test set ligands for hypothesis AADDR.49.

Ligand Name	QSAR Set	Activity (pIC50)	Predicted Activity (pIC50)	Error (%)
compound2	test	4.711	4.76	1.04
compound4	test	4.305	4.21	2.21
compound5	test	4.651	4.61	0.88
compound9	test	4.839	4.81	0.60
compound12	test	4.602	4.62	0.39
compound14	test	4.636	4.71	1.60
compound19	test	4.292	4.25	0.98

An example of QSAR representation was visualized in the context of the most active (compound1) and the least active (compound23) ligands, with a pre-defined regression coefficient threshold (Figure 4.12).

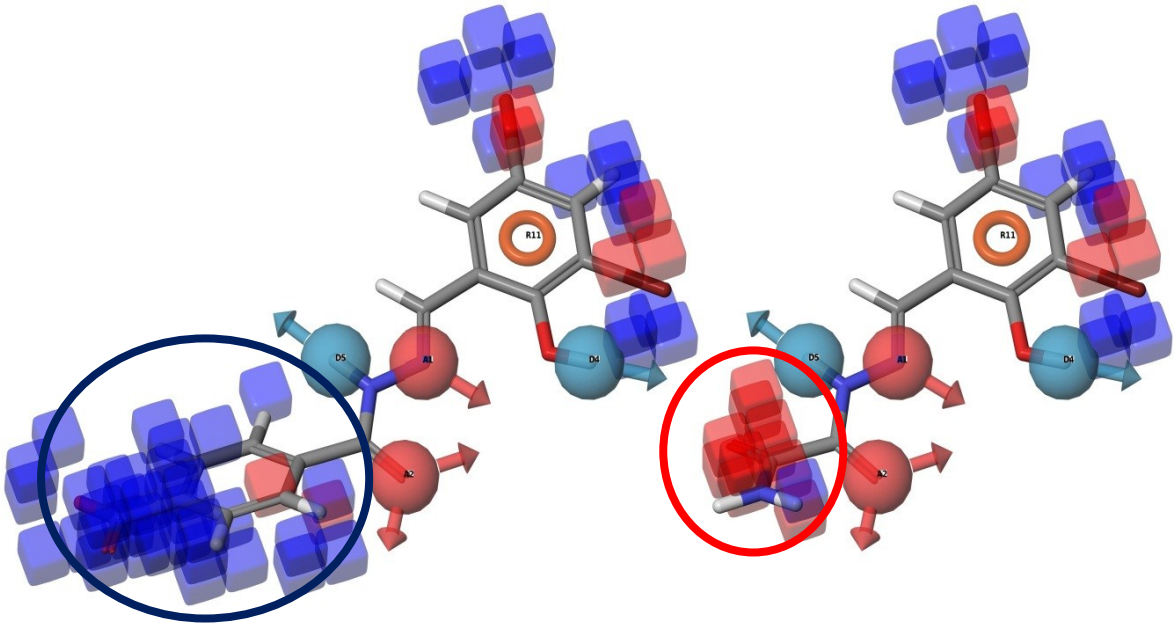


Figure 4.12. An example representation of QSAR with most active compound1 (left) and least active compound23. Regions that govern activity change were indicated in circles.

In this figure, blue cubes denote the volume occupied by the ligand that has a positive influence on calculated activity. By contrast, red cubes denote a negative influence on activity. Structure of compound1 and compound23 is quite similar, with an exception that induces an activity change. Compound1 has a formamide group (NH_2-C-O) that is

attached to the formic hydrazine group (NNC=O). Contrarily, compound23 has a nitrogen dioxide (NO_2) attached to the formic hydrazine group. Therefore, according to the model, formamide group favored activity whereas nitrogen dioxide presence resulted in a reduced activity. (Regression coefficient thresholds are 0.08 for positive and -0.08 for negative cubes).

After pharmacophore model was developed and validated by QSAR, it was used to search 3D phase small molecule database to identify the molecules that satisfy the hypothesis. With pharmacophore pre-filtering, small molecule database generated from ZINC was reduced to 84128 hits from initial 2.5 million compounds. These hits were sorted according to their fitness to hypothesis AADDR.49. Library compounds with less than 1 fitness value to hypothesis were further rejected. Figure 4.13 shows fitness of library molecules that matched the hypothesis AADDR.49 with a fitness value above 1. Fitness of top 22200 molecules ranged between 1 and 2.90. The small molecule database was reduced to 22200 compounds. For YopE, structure-based virtual screening was processed with these molecules.

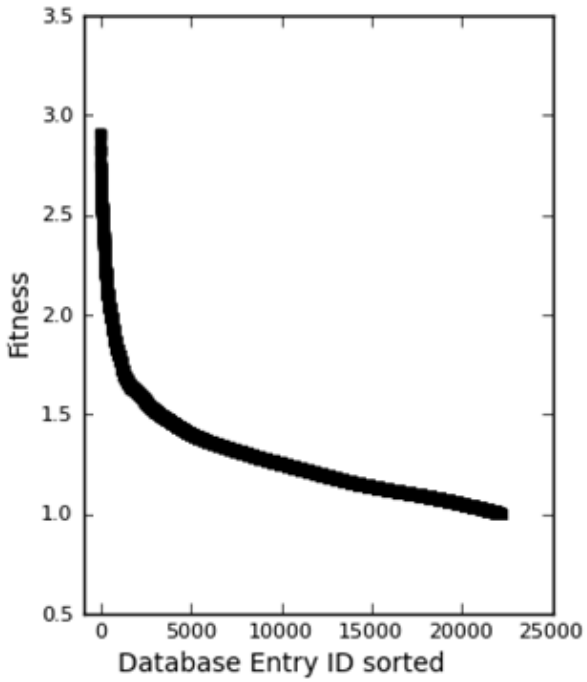


Figure 4.13. Scatter plot of fitness scores of 22200 small molecule database compounds filtered by hypothesis AADDR.49.

4.1.4. Docking Results and Post-docking Analysis

Pharmacophore pre-filtering with hypothesis AADDR.49 yielded 22200 compounds in the small molecule database generated from ZINC. These compounds were docked to receptor YopE with Glide (Grid-based Ligand Docking with Energetics) program to search for interactions between ligands and the protein. Each ligand's position and orientation relative to the receptor protein was determined and scored with Glide's internal scoring function GlideScore.

For virtual screening, two docking modules present in Glide, standard precision docking (Glide SP) and extra precision docking (Glide XP) were used, successively. Glide SP provides a more rapid screening at the cost of accuracy, therefore it is reported to be more useful in docking large number of ligands [87]. On the other hand, Glide XP provides a more detailed scoring step with additional terms evaluated at the cost of screening time, and it is recommended for docking a modest number of ligands [87]. Therefore, remaining 22200 compounds were docked with Glide SP mode, and top 10% of poses were re-docked to YopE with more expensive Glide XP precision mode.

4.2.3.1. Standard Precision Docking. Glide SP docking was carried out by providing receptor grid and .maegz file (Maestro file extension) containing structures of 22200 compounds. Receptor grid was generated using YopE structure described in Section 4.1.1 with a binding site defined by a 18 Å box around residue Arg144. One hydrogen bond docking constraint on Arg144 was included. Compounds were docked using the standard precision setting with the flexible docking option enabled for the ligands. All other Glide docking settings were used with default values. Figure 4.14 shows plot of Glide SP scores of first 2220 compounds, corresponding to top 10% of inputs.

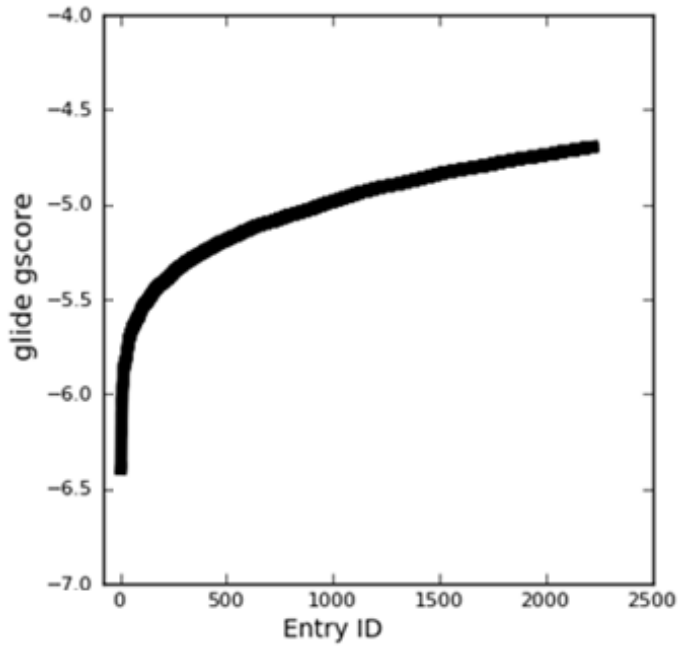


Figure 4.14. Plot of GlideScores of top 2220 small molecule database compounds docked to YopE in Glide SP mode. Docking scores are presented in kcal/mol.

GlideScores of top 2220 compounds vary between -6.4 and -4.6 kcal/mol. A more negative GlideScore corresponds to a higher predicted binding affinity. Therefore an increase in negativity indicates a higher or better score, and tighter binding. Components of GlideScore are explained in detail in Section 3.8.

Virtual screening protocol was preceded with top 10% compounds from Glide SP docking results. However, it is not recommended to use poses generated with Glide SP as an input to Glide XP docking, since Glide induces some ligand internal strain during pose generation [87]. Using Premin utility in MacroModel application, ligands were relaxed and energetically minimized before Glide XP docking, as explained in Section 3.8.

4.2.3.2. Extra Precision Docking. After energy minimization was carried out to top 2220 compounds obtained from the Glide SP docking, all compounds were re-docked to YopE structure in Glide XP mode. The same receptor grid and hydrogen bond constraint on Arg144 residue were used for this mode. Compounds were docked using the extra precision setting with the flexible docking option enabled. All other Glide docking settings were used with default values. All input compounds successfully docked to YopE,

satisfying the hydrogen bond constraint on Arg144. Figure 4.15 shows plot of Glide XP scores of all 2220 compounds.

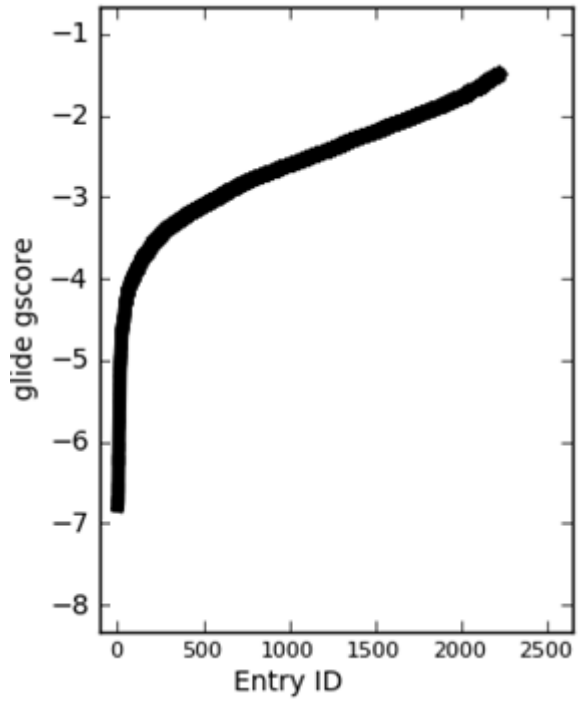


Figure 4.15. Plot of GlideScores of 2220 small molecule database compounds docked to YopE in Glide XP mode. Docking scores are presented in kcal/mol.

GlideScores of all compounds vary between -6.8 and -1.5 kcal/mol in Glide XP mode. Detailed analysis of docking parameters was carried out with hits having -5 kcal/mol or above Glide XP score. Threshold value was selected as -5 kcal/mol since average Glide XP score of known inhibitors of YopE with Arg144 constraint was found as 4.96 kcal/mol (Table 4.2). This criterion yielded 15 hits out of 2220. The quality of each docking was presented with a score and ranked accordingly. Docking results for these 15 hits were tabulated in Table 4.10. The docking score was calculated with GlideScore, which is based on ChemScore empirical scoring function and given in Section 3.8.

Hits on Table 4.10 were sorted according to their GlideScore values. First column of the table shows GlideScore values of each ligand, and the following seven columns include components of GlideScore function, all values in kcal/mol. At the last column, another scoring function is also represented, which is Emodel.

Table 4.10. Glide XP results of top 15 hits (receptor:YopE, hypothesis: AADDR.49). All values are shown in kcal/mol.

#	Title	Glide score	Glide lipo	Glide hbond	Glide rewards	Glide evdw	Glide ecoul	Glide rotb	Glide esite	Glide emodel
1	ZINC02736077	-6.784	**-1.290**	-0.304	-1.159	**-25.062**	-12.941	0.939	-5.63E-02	-45.830
2	ZINC16545543	-6.774	-0.242	-1.056	-1.635	-7.100	**-22.699**	0.523	-9.05E-02	-38.617
3	ZINC16525119	-6.737	-0.955	-0.509	**-1.863**	-14.977	-17.259	0.912	-3.84E-02	-36.007
4	ZINC16525119	-6.719	-0.420	-0.271	**-1.847**	-14.271	-18.109	0.912	-2.72E-02	-39.238
5	ZINC16525119	-6.568	-0.435	-0.280	**-1.842**	-14.998	-16.536	0.912	-2.69E-02	-36.826
6	ZINC01663005	-6.227	-0.443	-0.123	-1.696	-16.032	-16.607	0.833	-4.81E-02	-40.732
7	ZINC16545543	-6.220	-0.307	**-1.116**	-1.635	-11.771	-18.504	0.523	-1.14E-01	-38.933
8	ZINC04587520	-5.783	-0.808	-0.522	-1.371	-16.567	-15.629	0.612	-2.66E-02	-36.517
9	ZINC04830947	-5.707	-0.490	-0.247	-1.164	-13.462	-13.972	0.918	-5.60E-05	-32.419
10	ZINC05949341	-5.548	-0.858	-1.092	-0.959	-19.685	-15.161	0.869	-1.02E-01	-41.082
11	ZINC05949125	-5.471	-0.725	-1.008	-1.007	**-24.589**	-12.738	1.120	-1.26E-01	-39.203
12	ZINC46587237	-5.363	**-1.358**	-0.892	-0.493	**-29.619**	-16.385	0.859	**-1.76E-01**	**-63.423**
13	ZINC05728414	-5.237	-0.690	**-1.156**	-1.409	-16.758	-13.246	1.062	-7.98E-02	-35.553
14	ZINC17043967	-5.065	-0.261	-0.926	-1.025	-23.921	**-20.604**	0.841	-1.11E-01	-53.011
15	ZINC00067731	-5.009	**-1.068**	-0.848	-1.027	-15.464	-17.082	0.936	**-1.85E-01**	-42.433

Hits can be ranked according to either GlideScore or Emodel. For both scoring functions, lower values are better. Difference between GlideScore and Emodel scoring functions is that, Emodel is greatly dependent on internal strain energy of the generated poses. Therefore, it is reported that Emodel is not appropriate for comparing chemically distinct structures [87]. Additionally, it is reported that empirical GlideScore function should be used for ranking in virtual screening applications since it is more optimized for docking accuracy and database enrichment [87]. Therefore, GlideScore was used for comparing poses of different hits, and Emodel was used for deciding between different conformers (or poses) of a particular ligand. Highest GlideScore components within each column are indicated in bold in Table 4.10. It can be concluded that individual values of GlideScore components are not directly correlated to ranking of corresponding ligand. Their contributions to final score are quite diverse throughout the hits. For example, ligand with the highest GlideScore, ZINC02736077, has only highest lipophilic and van der Waals energy among all contributions.

Five molecules out of final 15 hits were selected as potent inhibitors of YopE Based on the GlideScore, as well as additional criteria, such as absorption, distribution, metabolism and excretion (ADME) considerations, druglikeness of ligands and strain energy differences of ligands. These post-docking processes can be seen in the following sections. Consequently, hits that violate Lipinski's rule of five, have high ligand strain or show poor ADME properties, were eliminated.

4.2.3.3. ADME and Molecular Properties of Final Hits. Another Schrödinger program QikProp was used for this step, which predicts a large number of descriptors, such as pharmacodynamic properties and oral bioavailability characteristics for candidate ligand molecules. QikProp also recommends ranges for each descriptor, according to their similarity with 95% of known drugs [96].

Table 4.11 shows some of the properties of hits determined by QikProp. Selected properties are #stars, SASA and its components, PSA, QPlogS and human oral absorption. Value of #stars indicates number of property that is not in 95% range of properties determined by known drugs. Recommended value for #stars is between 0-5, and high numbers indicate that the molecule is less drug-like. Number of stars is determined from a

number of descriptors, such as MW, SASA, FOSA, FISA, PISA, WPSA, donorHB, accptHB, QPlogPo/w [96]. All hits in Table 4.11 were in recommended range in terms of number of stars; hence the star values are not explicitly shown.

SASA is the total solvent accessible surface area in square angstroms, which has four different components. These are hydrophobic component (FOSA), hydrophilic component (FISA), π component (PISA) and weak polar component (WPSA). Recommended range of SASA value is between 300 and 1000 $Å^2$ [96]. SASA values of hits in Table 4.11 are in the acceptable range, ZINC02736077 and ZINC46587237 having the highest values. These values of SASA and HOA are indicated in bold in Table 4.11.

Table 4.11. Pharmacodynamic properties of top 15 hits (receptor:YopE). Eliminated molecules are indicated by strikethrough.

Molecule	SASA	FOSA	FISA	PISA	WPSA	PSA	QPlogS	Human Oral Absorption
ZINC02736077	**666.2**	135.6	232.6	220.7	77.3	124.3	-4.40	**71.4**
ZINC16545543	474.0	78.4	198.7	196.9	0.0	118.7	-1.79	63.1
ZINC16525119	465.8	84.2	189.7	191.8	0.0	110.6	-1.44	61.3
ZINC16525119	465.5	86.6	188.6	190.3	0.0	110.1	-1.44	61.5
ZINC16525119	465.4	84.5	190.5	190.4	0.0	110.0	-1.44	61.1
ZINC01663005	465.5	92.8	190.0	182.8	0.0	115.9	-1.22	60.9
ZINC16545543	474.1	76.4	201.3	196.4	0.0	118.1	-1.79	62.5
ZINC04587520	512.0	193.4	217.3	101.3	0.0	136.2	-1.74	58.5
ZINC04830947	640.7	366.6	203.7	70.4	0.0	138.2	-3.10	70.0
ZINC05949341	617.6	87.0	289.8	239.6	1.2	158.0	-2.37	37.6
~~ZINC05949125~~	~~644.6~~	~~89.1~~	~~297.6~~	~~257.9~~	~~0.0~~	~~186.7~~	~~-0.84~~	~~10.1~~
ZINC46587237	**793.3**	196.7	149.6	359.6	87.4	114.0	-5.16	**88.9**
ZINC05728414	556.3	165.9	171.8	186.2	32.4	109.3	-2.35	67.8
~~ZINC17043967~~	~~661.1~~	~~194.5~~	~~309.3~~	~~157.3~~	~~0.0~~	~~206.6~~	~~-1.50~~	~~8.0~~
ZINC00067731	597.6	294.2	150.0	153.4	0.0	128.7	-2.84	78.9

PSA is the van der Waals surface area of polar nitrogen and oxygen atoms and QPlogS is predicted aqueous solubility. Range for PSA is between 7 and 200 whereas QPlogS range is between -6.5 and -0.5 [96]. All hits in Table 4.11 are within these ranges. Last column represents predicted human oral absorption (HOA) on percent scale. QikProp suggest that molecules with more than 80% HOA exhibit good absorption and molecules having less than 25% HOA exhibit poor absorption [96]. ZINC05949125 and ZINC17043967 have poor predicted human oral absorptions, at 10% and 8%, respectively. These molecules are indicated with a strikethrough in Table 4.11. Therefore, these two molecules were eliminated.

Table 4.12 shows the druglikeness properties of top 15 hits. The formulated rule is denoted as Lipinski's rule of five since it sets limits involving five or multiples of five. According to this rule, in order for a compound to be drug-like and orally active, it should have a molecular weight less than 500 Da, hydrogen bond donor equal to or less than five, hydrogen bond acceptor equal to or less than 10 and partition coefficient (QPlogPo/w) less than five [96].

Since hydrogen bond acceptor and donor values of hits were determined from a number of configurations, they were calculated as non-integers. QPlogPo/w (or LogP) is the predicted partition coefficient of a compound between octanol and water solution, which gives an estimate about compound's lipophilicity. Up to certain limit, compounds with higher lipophilicity have higher ability to permeate across biological membranes, which is necessary for a drug candidate. High lipophilicity, on the other hand, can result in a poor aqueous solubility. Lipinski determined the upper favorable limit of partition coefficient (QPlogPo/w) as five [79]. No violations of hits to the rule of five have been observed; therefore no compounds were eliminated with these criteria. Within desired limits, ZINC02736077 and ZINC46587237, which are indicated in bold in Table 4.12, showed the highest lipophilicity.

Table 4.12. Druglikeness of top 15 hits according to Lipinski's rule (receptor:YopE, hypothesis: AADDR.49). Eliminated molecules are indicated by strikethrough.

Molecule	MW	donorHB	accptHB	QPlogPo/w	Rule of Five
ZINC02736077	380.39	3	7.4	**2.125**	0
ZINC16545543	260.25	4	8.1	-0.285	0
ZINC16525119	238.24	5	8.3	-0.847	0
ZINC16525119	238.24	5	8.3	-0.839	0
ZINC16525119	238.24	5	8.3	-0.854	0
ZINC01663005	251.24	4	9.8	-0.914	0
ZINC16545543	260.25	4	8.1	-0.308	0
ZINC04587520	300.27	4	9.55	-0.523	0
ZINC04830947	353.37	4	8.8	1.039	0
ZINC05949341	365.40	4.25	7.75	0.62	0
~~ZINC05949125~~	~~372.38~~	~~4.5~~	~~7.5~~	~~-0.204~~	~~0~~
ZINC46587237	458.97	4	7.75	**3.42**	0
ZINC05728414	293.34	3	7.7	0.287	0
~~ZINC17043967~~	~~405.41~~	~~4.25~~	~~8.25~~	~~0.197~~	~~0~~
ZINC00067731	349.35	3	9.4	1.017	0

4.2.3.4. Strain Energy Calculation. Glide docking enables ligand flexibility whereas the receptor remains essentially frozen. Therefore, the docking allows ligands to be strained, so that they can fit better into the binding site. However, program cannot identify if strain is an outcome of docking or it is a false positive. Thus, after docking, it is recommended to use the Strain Rescore script, which determines if ligand has too much strain [87]. Energies of free and docked conformations of each hit were calculated. If the energy difference is more than 4 kcal/mol between conformations (default value), corresponding hit received penalty, added to the GlideScore. Bound and free energy of hits, strain energy differences, penalties and adjusted GlideScores are listed in Table 4.13. Majority of the hits did not receive strain penalties. Only two hits, ZINC05949341 and ZINC17043967 received penalties. ZINC17043967 were already eliminated due to its poor ADME properties. ZINC05949341 were also eliminated after strain correction, since energy of this molecule is almost tripled in its bound form.

Table 4.13. Strain energy penalties of top 15 hits (receptor:YopE, hypothesis: AADDR.49). Eliminated molecules are indicated by strikethrough. All values are shown in kcal/mol.

	Title	Bound energy	Free energy	Strain Energy	Strain penalty	Strain GlideScore	Glide emodel
1	ZINC02736077	24.706	23.478	1.228	0.000	-6.784	-45.830
2	ZINC16545543	32.416	32.416	0.000	0.000	-6.774	**-38.617**
3	ZINC16525119	23.901	22.083	1.818	0.000	-6.737	-36.007
4	ZINC16525119	20.015	19.471	0.545	0.000	-6.719	**-39.238**
5	ZINC16525119	19.114	18.852	0.262	0.000	-6.568	-36.826
6	ZINC01663005	25.096	24.851	0.245	0.000	-6.227	-40.732
7	ZINC16545543	32.516	32.416	0.100	0.000	-6.220	-38.933
8	ZINC04587520	41.054	38.386	2.668	0.000	-5.783	-36.517
9	ZINC04830947	54.556	54.543	0.013	0.000	-5.707	-32.419
~~10~~	~~ZINC05949341~~	~~8.737~~	~~2.772~~	~~5.965~~	~~0.491~~	~~-5.057~~	~~-41.082~~
~~11~~	~~ZINC05949125~~	~~13.903~~	~~10.454~~	~~3.449~~	~~0.000~~	~~-5.471~~	~~-39.203~~
12	ZINC46587237	11.442	7.882	3.561	0.000	-5.363	-63.423
13	ZINC05728414	32.342	27.543	4.799	0.200	-5.237	-35.553
~~14~~	~~ZINC17043967~~	~~27.537~~	~~18.683~~	~~8.854~~	~~1.214~~	~~-3.851~~	~~-53.011~~
15	ZINC00067731	27.426	26.330	1.096	0.000	-5.009	-42.433

Overall, three hits were eliminated from the top 15 hits list. From the remaining hits, five molecules were selected. Since all remaining hits showed acceptable predicted ADME properties and did not violate Lipinski's rule, final hit selection was done based on GlideScores with strain corrections applied (Table 4.13).

Eventually, ZINC02736077, ZINC16545543, ZINC16525119, ZINC01663005 and ZINC04587520 were chosen for proposal. More than one poses of two different hits were observed in the table. ZINC16525119 has three poses with rankings 3, 4, 5 and ZINC16545543 has two poses with rankings 2 and 7. As stated earlier, selection of poses of a particular molecule was performed based on their Emodel values. Fourth ranking pose of ZINC16525119 have the highest model score (-39.2 kcal/mol), therefore it is selected

for visual inspection. Similarly, Emodel values of two ZINC16545543 poses were investigated, and their values were observed to be very close and indistinguishable. Therefore, pose with highest docking score, rank 2, were selected for visual inspection. Selected poses of ZINC16525119 and ZINC16545543 are indicated in bold in Table 4.13.

4.2.3.5. Binding Free Energy Calculation. After docking and post-docking processing were done, an additional module, Prime MMGB-SA was used to estimate relative binding affinity of top 15 hits. The aim of this calculation was to see if there is a correlation between estimated free energy of binding and docking score. For this purpose, pose viewer file of hits and the receptor YopE is provided to MM-GBSA module as an input. Binding free energies of each ligand along with its components are calculated as explained in Section 3.8.1.3 and represented in Table 4.14.

Table 4.14. Binding free energies of top 15 hits (receptor:YopE, hypothesis: AADDR.49). Eliminated molecules are indicated by strikethrough. All values are shown in kcal/mol.

Title	DG bind Coulomb	DG bind Covalent	DG bind vdW	DG bind SolvGB	DG bind Lipo	DG bind total
ZINC02736077	-22.44	2.33	-33.43	17.97	-28.51	-66.34
ZINC16545543	-34.47	1.75	-15.06	18.57	-12.43	-45.34
ZINC16525119	-30.21	0.25	-17.99	13.81	-16.49	-50.01
ZINC16525119	-14.87	-1.89	-25.01	9.93	-17.11	-49.82
ZINC16525119	-26.87	1.87	-23.12	14.62	-17.29	-51.43
ZINC01663005	2.08	-6.20	-24.36	11.74	-17.33	-35.91
ZINC16545543	-34.00	0.36	-14.76	18.05	-12.54	-45.88
ZINC04587520	-26.56	0.71	-27.03	17.54	-21.32	-56.67
ZINC04830947	-36.33	4.08	-14.70	22.45	-14.46	-39.71
~~ZINC05949341~~	~~-36.73~~	~~4.38~~	~~-14.93~~	~~19.29~~	~~-14.36~~	~~-47.73~~
~~ZINC05949125~~	~~-38.98~~	~~6.14~~	~~-19.93~~	~~21.71~~	~~-19.71~~	~~-60.38~~
ZINC46587237	-35.85	1.40	-34.57	22.17	-21.51	-74.07
ZINC05728414	-36.85	1.44	-18.55	14.96	-19.70	-57.99
~~ZINC17043967~~	~~-27.97~~	~~3.74~~	~~-31.12~~	~~16.56~~	~~-18.83~~	~~-58.63~~
ZINC00067731	-27.54	6.08	-26.51	18.03	-20.77	-56.08

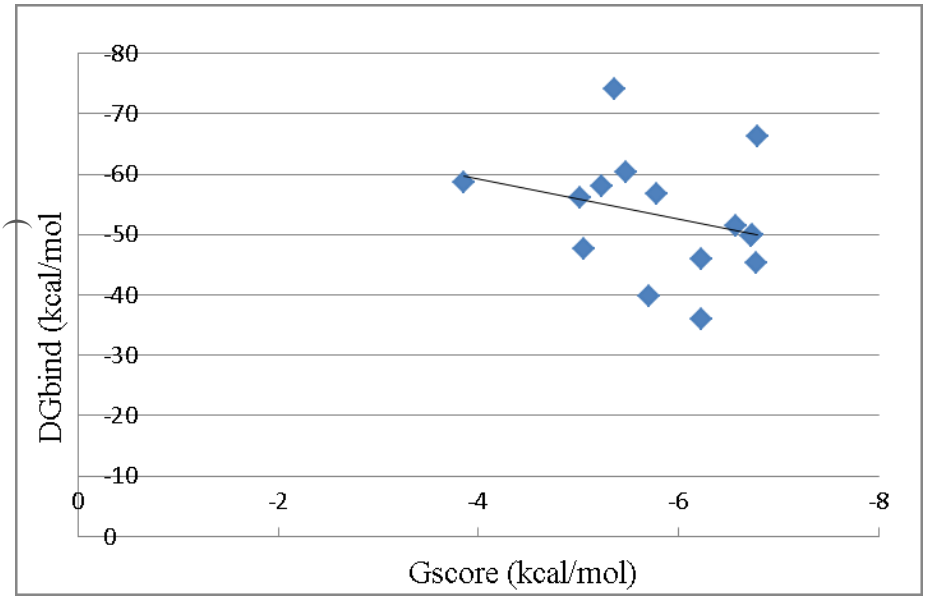


Figure 4.16. Correlation between GlideScores and MM-GBSA predicted binding affinities for top 15 hits for YopE.

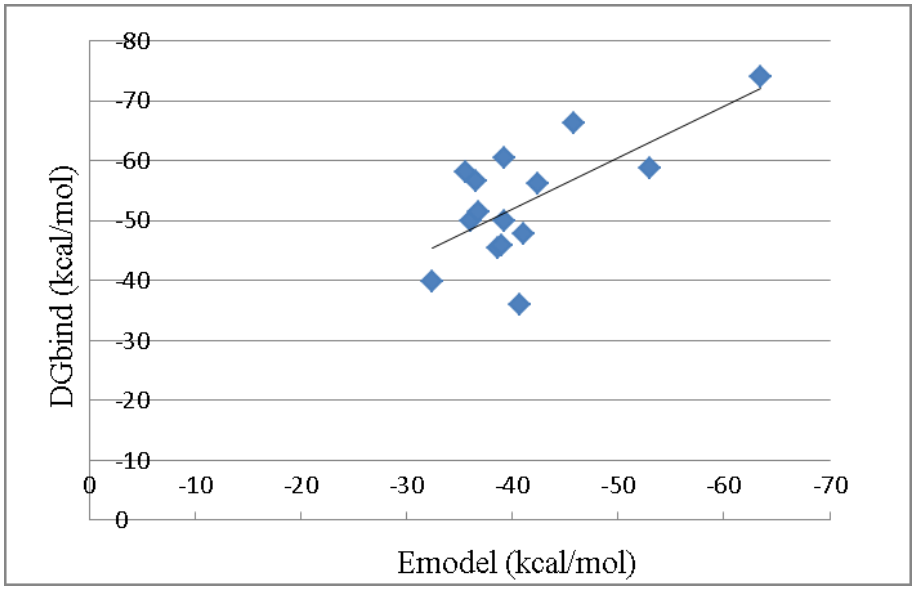


Figure 4.17. Correlation between Emodel and MM-GBSA predicted binding affinities for top 15 hits for YopE.

A scatter plot including free binding energies versus GlideScores of top 15 hits were shown in Figure 4.16. No direct correlation was observed between binding free energies and docking scores. Additionally, binding free energies were not consistent with rankings of the hits. Therefore, this approach was not considered in decision-making step of

proposed molecules. Additionally, correlation between Emodel and binding free energies were plotted (Figure 4.17). Positive correlation was observed in this case.

4.1.5. Visual Inspection of Proposed Hits

Binding modes of selected hits and their interactions with the receptor YopE were analyzed in this section. However, two proposed hits, ZINC16525119 and ZINC01663005 were almost similar in terms of their molecular properties and 2D structures (Figure 4.18).

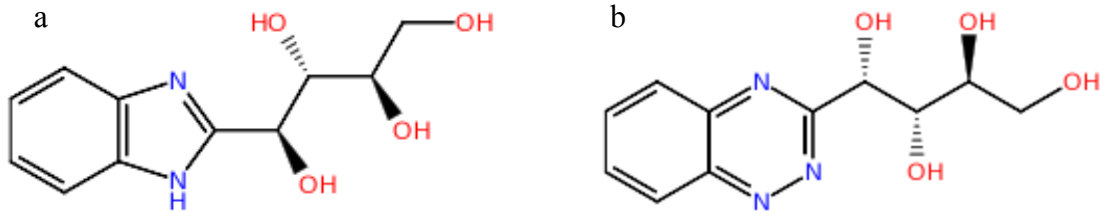


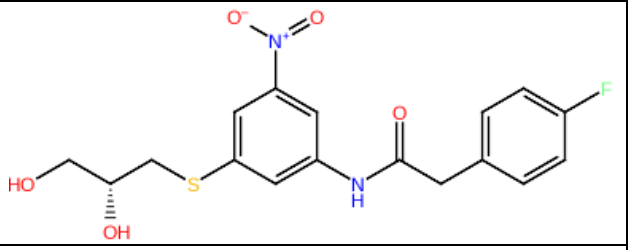
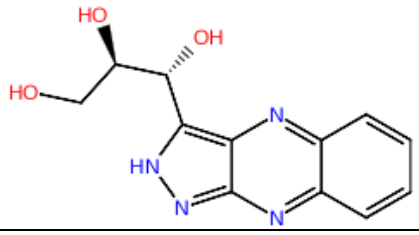
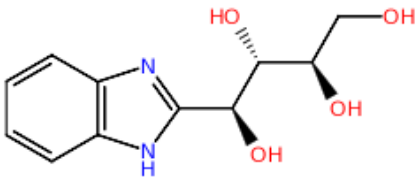
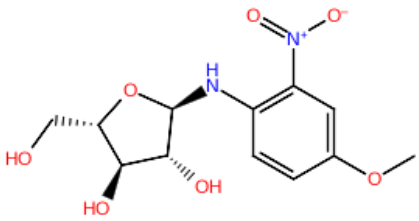
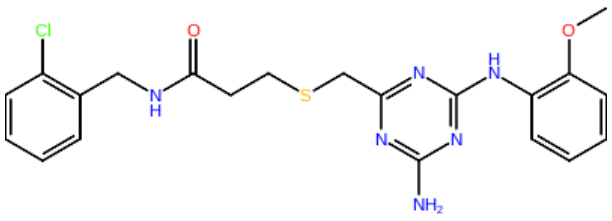
Figure 4.18. 2D structure of (a) ZINC16525119 (b) ZINC01663005.

Since chemical and structural diversity were also aimed in ligand proposal stage, one of these hits was decided to be eliminated. In terms of both docking score and Emodel score, ZINC16525119 showed better affinity (Table 4.13). Therefore, ZINC01663005 was discarded from proposed hits list and ZINC46587237 was included, since it exhibited some interesting properties. Despite its ranking, ZINC46587237 has the highest predicted human oral absorption percentage (88%), highest Emodel value (-63.423 kcal/mol), QPlogPo/w (3.24) and solvent accessible surface area (793.3 $Å^2$) among all top 15 hits. With this alteration, the list of proposed molecules for inhibition of YopE is given in Table 4.15.

Table 4.15. Summary of proposed molecules for YopE.

Initial rank	Name	Database Title	GlideScore (kcal/mol)
1	Y1	ZINC02736077	-6.784
2	Y2	ZINC16545543	-6.774
4	Y3	ZINC16525119	-6.645
8	Y4	ZINC04587520	-5.783
12	Y5	ZINC46587237	-5.363

Table 4.16. Chemical formulas and structures of proposed YopE inhibitors.

Name	Chemical formula	2D Structure
Y1	$C_{17}H_{17}F_1N_2O_5S_1$	
Y2	$C_{12}H_{12}N_4O_3$	
Y3	$C_{11}H_{14}N_2O_4$	
Y4	$C_{12}H_{16}N_2O_7$	
Y5	$C_{21}H_{23}Cl_1N_6O_2S_1$	

The binding mode and ligand interaction maps of these molecules were represented in Figure 4.19-Figure 4.23. The YopE binding site is viewed from the same perspective in these figures for easier comparison. Backbone interactions are shown with solid line whereas side chain interactions are shown with dashed lines. The residues are shown in their three-letter codes and are colored according to their amino acid types. Hydrogen bonds are represented as pink lines in interaction diagrams. Solvent exposed areas of ligands are indicated with yellow spheres. Hydrophobic contacts are not explicitly shown in figures, but they are explained within the text.

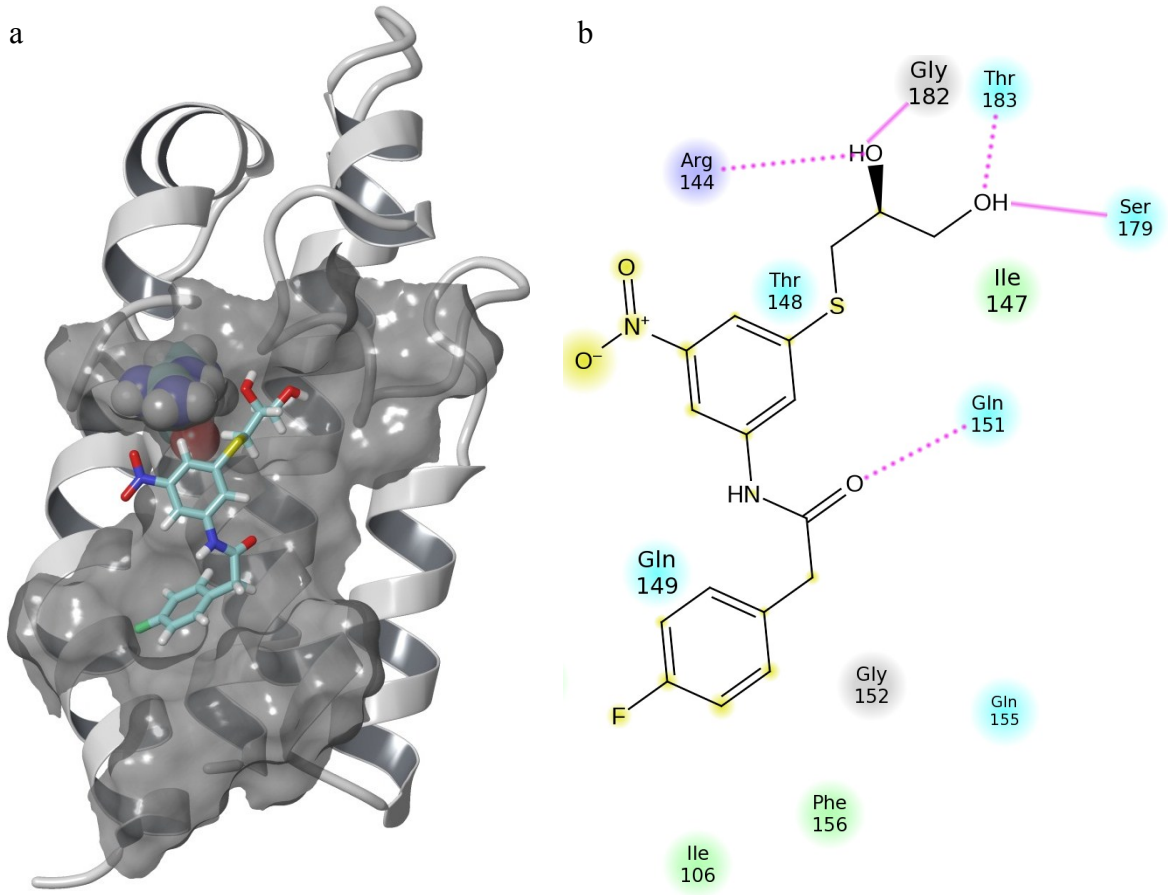


Figure 4.19. (a) Binding mode (b) Ligand interaction map of Y1.

Compound Y1, with molecular formula $C_{17}H_{17}F_1N_2O_5S_1$, has two ring systems, 26 heavy atoms and eight rotatable bonds. Y1 makes a total of five hydrogen bonds with YopE residues. Lone O atom attached to the propylamine group of Y1 (CCCN) makes a hydrogen bond with Gln151 of YopE. Both of two hydroxyl groups of the ligand make two hydrogen bonds, each with different residues. Hydroxyl group at the end makes hydrogen bond with side chain of Thr183 and backbone of Ser179. Other hydroxyl group makes hydrogen bond with the side chain of Arg144 and backbone of Gly182. This area of the ligand also has hydrophobic contacts with Ile147. Other hydrophobic contacts are between fluorobenzene ring of ligand with Ile106 and Phe156. Carbon atom attached to the fluorobenzene has also hydrophobic contact with Phe156. NO_2 group attached to the benzene ring is the surface-exposed region of the ligand. Besides satisfying Arg144 constraint, Y1 was able to interact spontaneously with four of the important residues listed in Table 4.1, which are Gln151, Ser179, Gly182, Thr183.

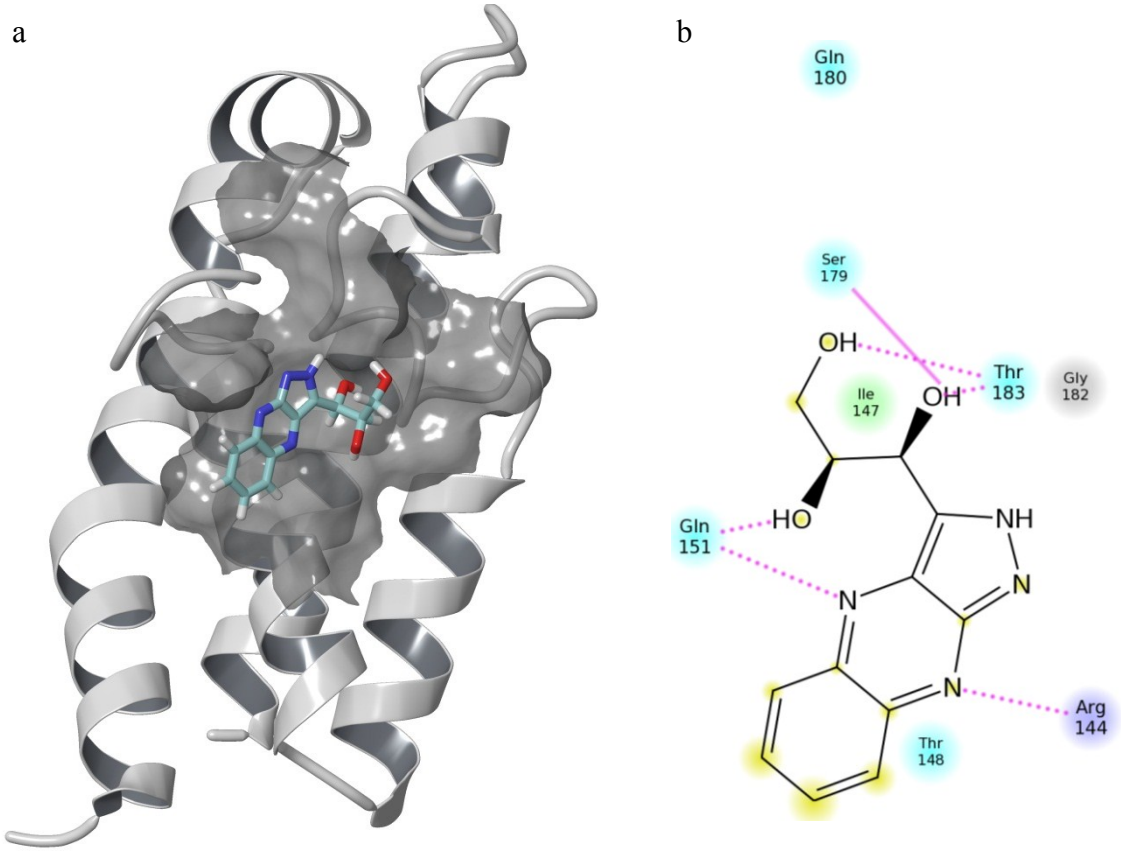


Figure 4.20. (a) Binding mode (b) Ligand interaction map of Y2.

Compound Y2, with molecular formula $C_{12}H_{12}N_4O_3$, has three ring systems, 19 heavy atoms and three rotatable bonds. Y2 makes a total of six hydrogen bonds with YopE residues. Nitrogen atom in the cyclohexane ring makes a hydrogen bond with side chain of Arg144. Other nitrogen atom of this ring makes hydrogen bond with side chain of Gln151. Gln151 also makes a hydrogen bond with the hydroxyl group that is flipped through the rings. Both of two other hydroxyl groups interact with the side chain of Thr183. Backbone of Ser179 also makes a hydrogen bond with the farther hydroxyl group. The region enclosed by three hydroxyl groups has hydrophobic contact with Ile147 of YopE. Cyclohexadiane group of Y2 is the solvent exposed region of the ligand and do not show any interactions with the receptor. Besides satisfying Arg144 constraint, Y2 was able to interact spontaneously with three of the important residues listed in Table 4.1, which are Gln151, Ser179, Thr183.

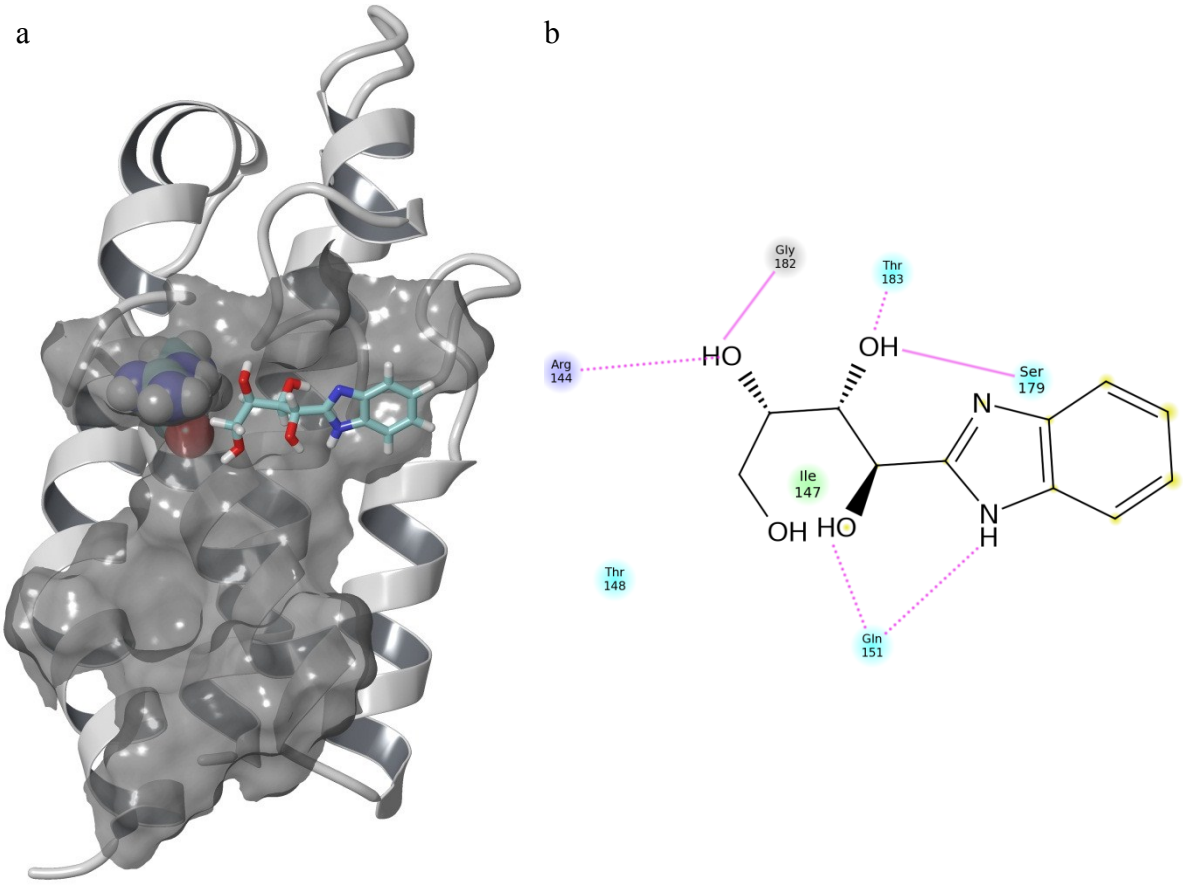


Figure 4.21. (a) Binding mode (b) Ligand interaction map of Y3.

Compound Y3, with molecular formula $C_{11}H_{14}N_2O_4$, has two ring systems, 17 heavy atoms and four rotatable bonds. Y3 makes a total of six hydrogen bonds with YopE residues. Side chain of Gln151 makes hydrogen bond with –NH group of the imidazole ring. Gln151 also make a hydrogen bond with hydroxyl group that is in the vicinity of Ile147. Remaining four hydrogen bonds are shared between other hydroxyl groups of the ligand. Hydroxyl group at the upper left end of the ligand make hydrogen bond with side chain of Arg144 and side chain of Gly182. Side chain of Thr183 and backbone of Ser179 interacts with other hydroxyl group. The region enclosed by four hydroxyl groups is in hydrophobic contact with Ile147. Cyclohexadiane group of Y3 is the solvent exposed region of the ligand and do not show any interactions with the receptor. Besides satisfying Arg144 constraint, Y3 was able to interact spontaneously with four of the important residues listed in Table 4.1, which are Gln151, Ser179, Gly182 and Thr183.

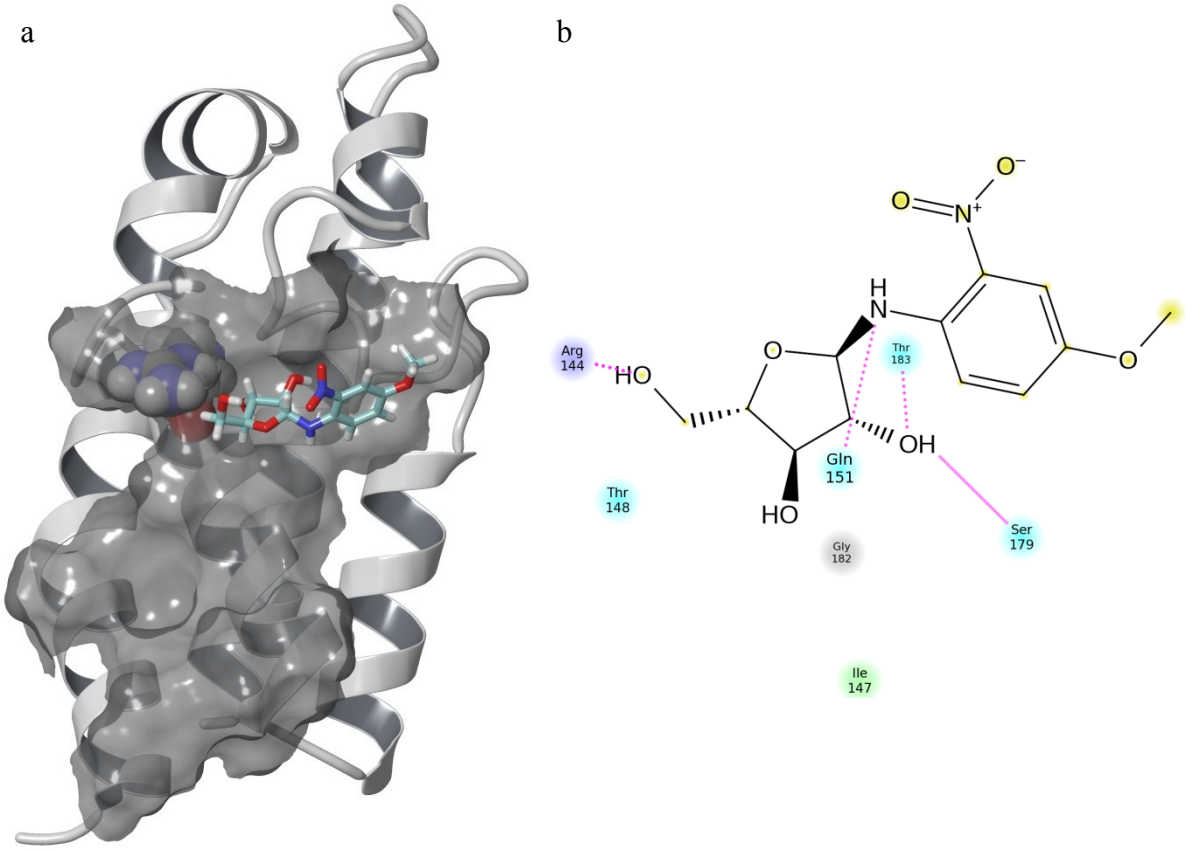


Figure 4.22. (a) Binding mode (b) Ligand interaction map of Y4.

Compound Y4, with molecular formula $C_{12}H_{16}N_2O_7$, has two ring systems, 21 heavy atoms and five rotatable bonds. Y4 makes a total of four hydrogen bonds with YopE residues. Side chain of Arg144 makes hydrogen bond with the hydroxyl group of the ligand that is flipped towards Arg144. Similarly, side chain of Gln151 makes a hydrogen bond with –NH group that is attached to the benzene ring. Remaining two hydrogen bonds are made with same hydroxyl group of the ligand. Hydroxyl group attached to the cyclopentane group make hydrogen bond with both side chain of Thr183 and backbone of Ser179. Cyclopentane ring group is in hydrophobic contact with Ile147. Nitrogen dioxide and –O-C groups attached to the benzene ring of Y4 is the solvent exposed region of the ligand and do not show any interactions with the receptor. Besides satisfying Arg144 constraint, Y4 was able to interact spontaneously with three of the important residues listed in Table 4.1, which are Gln151, Ser179 and Thr183.

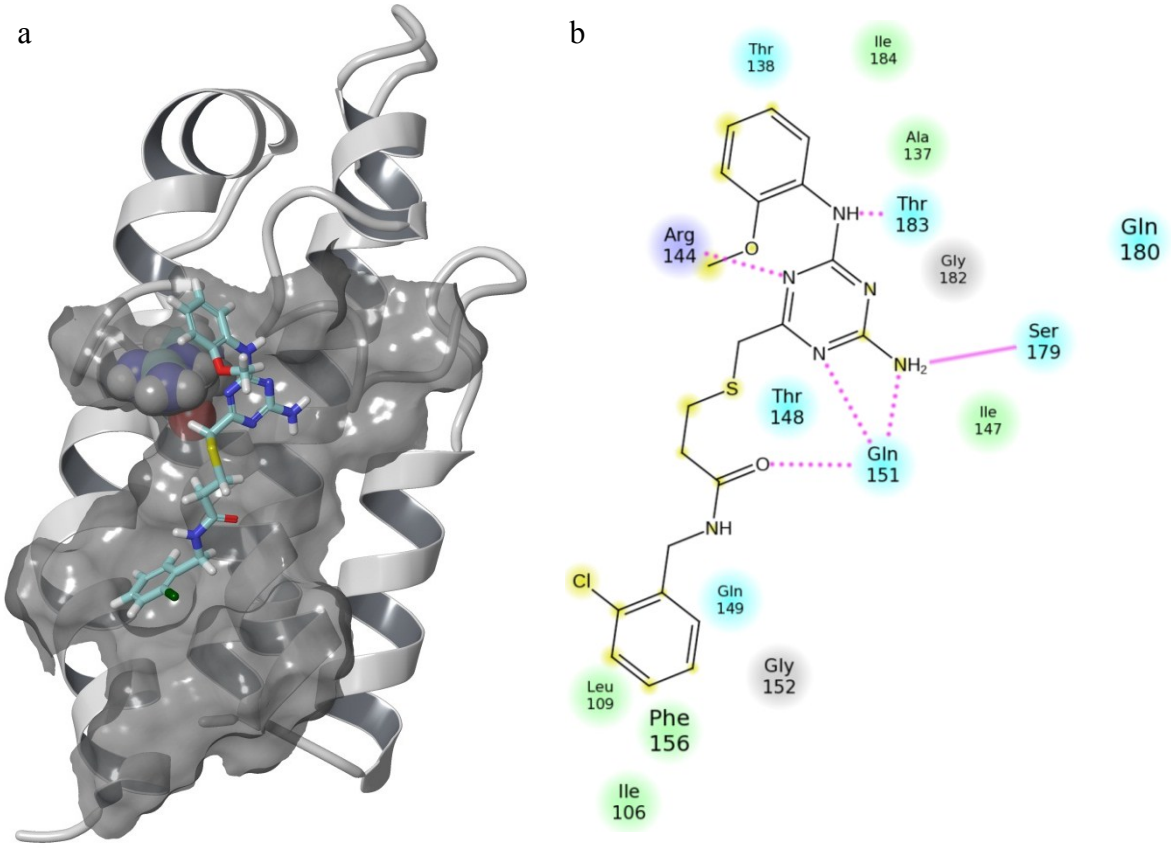


Figure 4.23. (a) Binding mode (b) Ligand interaction map of Y5.

Compound Y5, with molecular formula $C_{21}H_{23}Cl_1N_6O_2S_1$, has three ring systems, 31 heavy atoms and 10 rotatable bonds. Y5 makes a total of six hydrogen bonds with YopE residues. Side chain of Arg144 makes hydrogen bond with the nitrogen atom of the triazine ring. Similarly, side chain of Gln151 makes three hydrogen bonds with amine group that is attached to the triazien ring, nitrogen atom of the triazine ring and the lone oxygen atom in the vicinity of Thr148. Amine group attached to the triazine ring also makes hydrogen bond with the backbone of Ser179. Hydrophobic contacts occur between chlorobenzene ring and Ile106 and Leu109 of YopE. Besides satisfying Arg144 constraint, Y5 was able to interact spontaneously with three of the important residues listed in Table 4.1, which are Gln151, Ser179 and Thr183. Summary of hydrogen bond interactions between proposed molecules and YopE are represented in Table 4.17. Here, "+" sign indicates presence of a hydrogen bond and number of signs indicates the number of hydrogen bonds between ligand and a particular receptor residue.

Table 4.17. Summary of H-bond interactions between YopE and proposed molecules.

Name	Arg144	Gln151	Ser179	Gly182	Thr183
Y1	+	+	+	+	+
Y2	+	++	+		++
Y3	+	++	+	+	+
Y4	+	+	+		+
Y5	+	+++	+		+

From eight important residues for YopE activity [41], five residues made hydrogen bonds with proposed molecules. In addition to their acceptable ADME properties, visual inspection showed that proposed molecules represent favorable interactions with YopE using Glide docking protocol. Arg144 residue preferred interactions with either hydroxyl groups or nitrogen atom in ring systems. Thr183 and Ser179 interactions were mostly on hydroxyl groups of ligands as well but Gln151 did not preferred specific atom or group type upon ligand binding. Ile147 was observed to be an important residue in hydrophobic interactions. The binding modes of proposed molecules are superimposed in Figure 4.24.

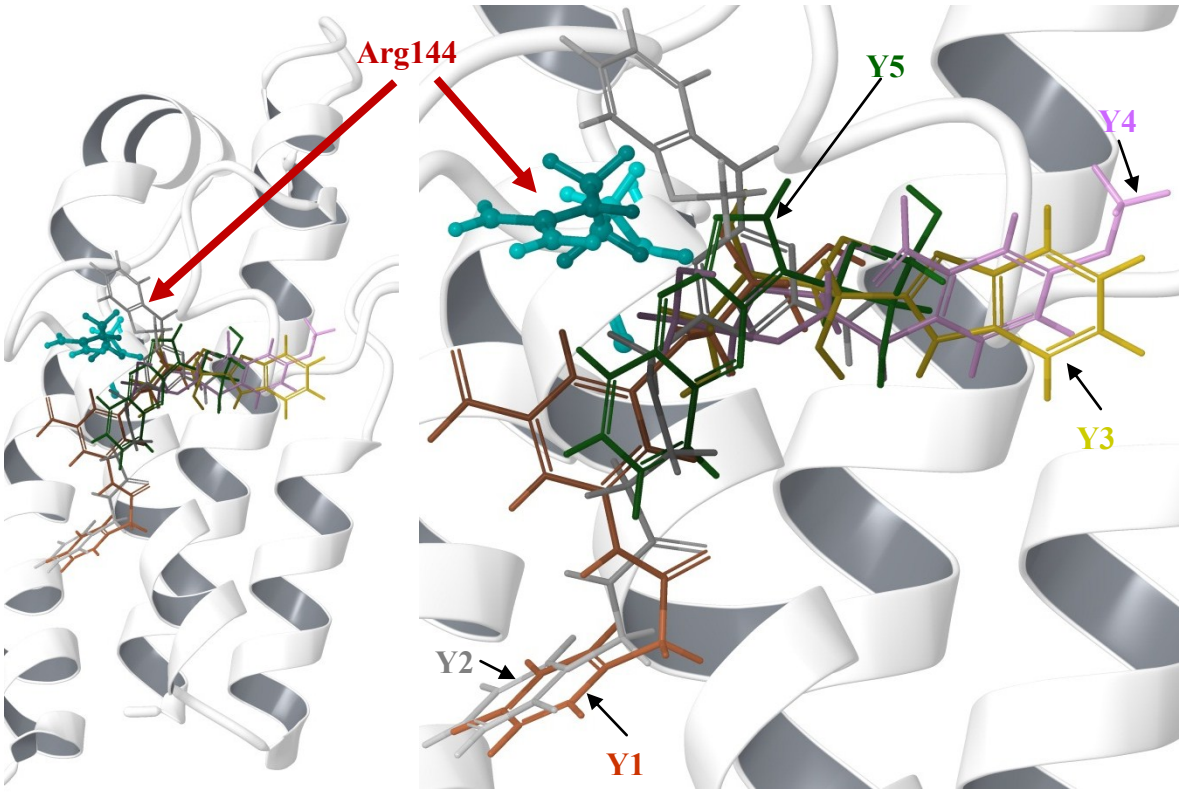


Figure 4.24. The proposed YopE inhibitors in the YopE pocket.

The binding modes of proposed molecules can be divided into two groups, depending on their size. Relatively small Y3 and Y4 were placed horizontally whereas larger molecules Y1 and Y5 were placed vertically with respect to YopE view in Figure 4.22. Overall, diversity in terms of structure and binding modes were observed among proposed YopE inhibitors. Vendors and IUPAC names of the proposed molecules are given in Table 4.18.

Table 4.18. Vendors and names of proposed YopE inhibitors.

Name	Vendor	IUPAC Name
Y1	Specs	N-[3-[(2S)-2,3-dihydroxypropyl]sulfanyl-5-nitrophenyl]-2-(4-fluorophenyl)acetamide
Y2	Labotest	(1R,2S)-1-(1H-pyrazolo[4,5-b]quinoxalin-3-yl)propane-1,2,3-triol
Y3	Aldrich CPR	(1R,2S,3R)-1-(1H-benzimidazol-2-yl)butane-1,2,3,4-tetrol
Y4	Asinex	N-(4-methoxy-2-nitrophenyl)-alpha-D-arabinofuranosylamine
Y5	UORSY	3-[[4-amino-6-(2-methoxyanilino)-1,3,5-triazin-2-yl]methylsulfanyl]-N-[(2-chlorophenyl)methyl]propan

4.1.6. Enrichment Studies

In order to validate the docking protocol and to see how accurate Glide XP docking scores are, enrichment calculations were carried out. The purpose of enrichment was to investigate if Glide XP docking protocol is able to distinguish between active and inactive drug-like molecules in virtual screening. Experimentally verified inhibitors of YopE were used as actives [46]. Inactive molecule set was provided from Schrödinger's own decoy set [106]. Properties of this ligand decoy set were given in Section 3.8.1.2.

All 1000 decoys and actives were docked to the receptor YopE in Glide XP mode without constraints. The output file of this docking run was used as input in the enrichment script, which calculates a set of metrics including enrichment factor (EF) and area under the receiver-operating characteristic curve (AUC). The results of the enrichment

calculation are tabulated in Table 4.19. Detailed enrichment report is given in Appendix D:.

Table 4.19. Enrichment metrics calculated with decoy set and known YopE inhibitors.

Total ligands(actives+decoys)	1015 (995+20)
ROC	0.88
Area under accumulation curve	0.87
EF 1%, 2%, 5%	25, 18, 7

In this table, ROC is the value of receiver operator characteristic area under the curve, which can be between 0 and 1. Ideal screening corresponds to 1 ROC value and 0.5 reflects random behavior. Area under the curve (AUC) is between 0 and 1, 1 being ideal screening performance. Higher value in these parameters is an indication that a known active will rank higher than an arbitrary decoy in docking [106]. ROC and AUC values were determined as 0.88 and 0.87 respectively, meaning that Glide XP docking protocol successfully identified known YopE inhibitors among drug-like decoys. Enrichment factors are also represented, but usually they are used in comparing other docking methods. Thus, an implication cannot be made with the EF values alone.

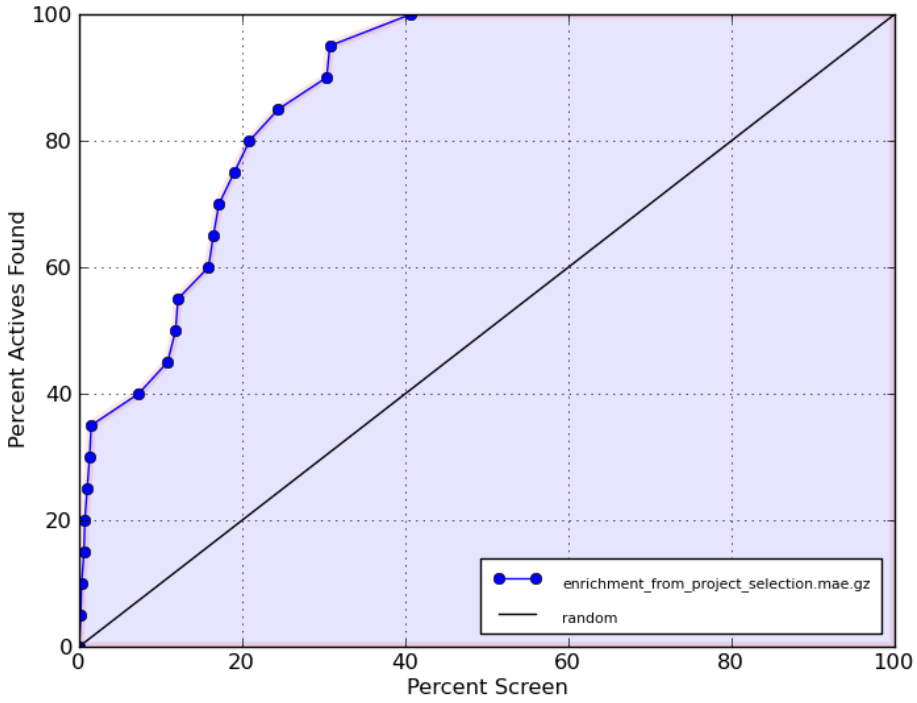


Figure 4.25. ROC calculated with decoy set and known YopE inhibitors.

4.2. SopE Results

In order to find potent inhibitors against SopE, similar to YopE case, structure-based virtual screening was combined with pre-filtering of small molecule database with pharmacophore building. Instead of ligand-based pharmacophore modeling, which could not be performed since there were no previously identified ligands for SopE, structure-based pharmacophore building (E-pharm) was carried out to create hypotheses, where known binding site information (GAGA loop of SopE) was used. Nevertheless, some structures were needed in the binding site, on which the pharmacophore model could be based. For this purpose, fragment molecules provided by Schrödinger [104] were docked to SopE's binding site with Glide XP. Using protein-fragment energetic terms computed by the Glide XP scoring function, suitable pharmacophore features and best hypotheses were determined for small molecule database filtering. Compounds in small molecule database that passed pre-filtering were docked to target SopE in standard (SP) and extra precision (XP) mode, successively. With scoring, ranking and post-docking analysis, number of potential leads was reduced for visual inspection. Workflow of the virtual screening methodology used to find SopE inhibitors is presented in Figure 4.26.

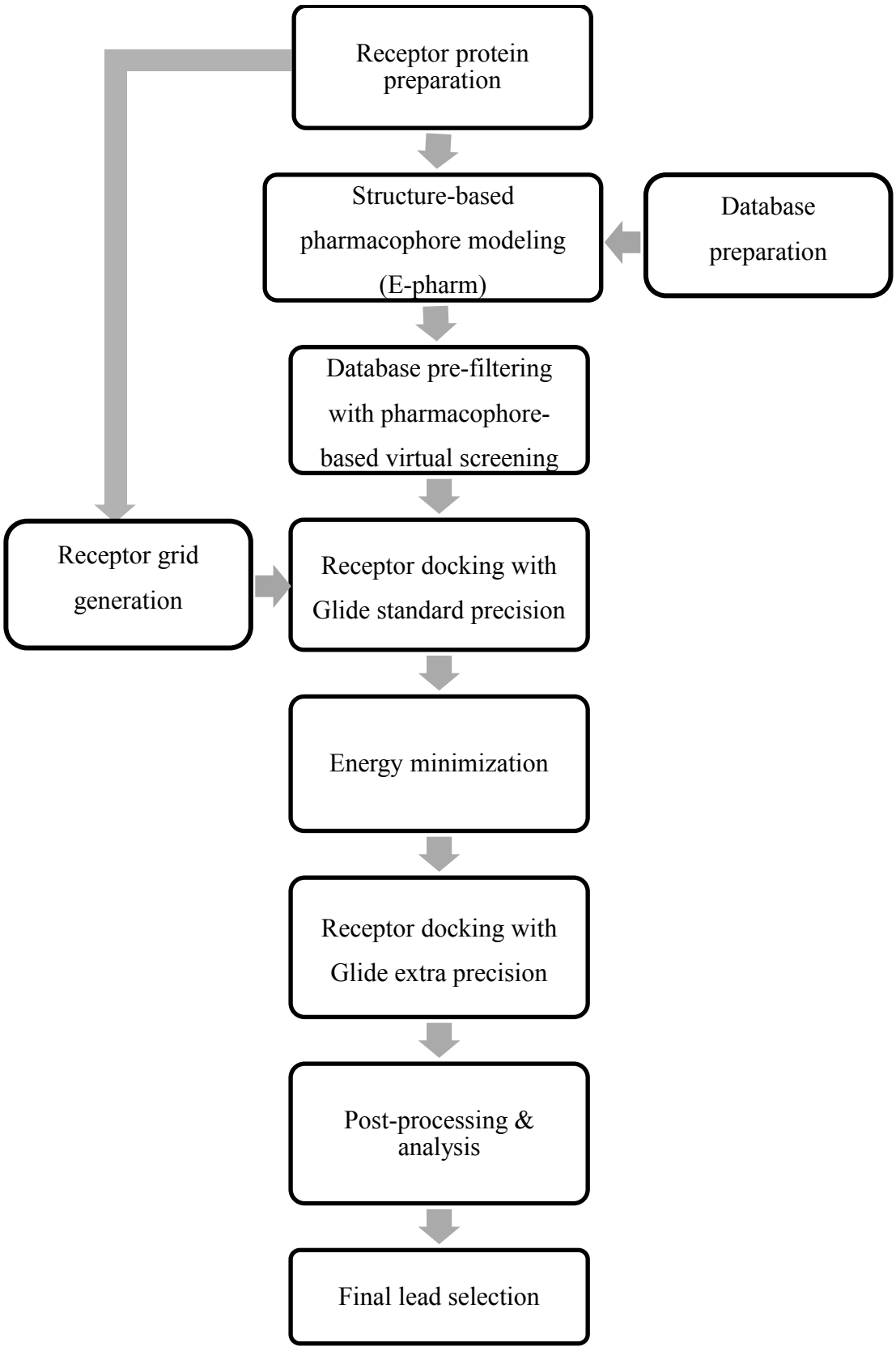


Figure 4.26. Schematic workflow of virtual screening and docking of SopE.

4.2.1. Remarks on Receptor Preparation and Grid Generation

3D structure of SopE (PDB code: 1gzs) was initially prepared with Protein Preparation Wizard module in Maestro workspace. PDB file originally included two chains of Cdc42 (chain A, C) and two chains of SopE (chain B, D) and a sulfate ion. Structure was simplified by deleting all chains of Cdc42, sulfate ion and one chain of SopE. Hydrogen atoms were added to protein's heavy atoms, and bond orders were automatically assigned. The structure was subjected to the Prime refinement tool in Protein Preparation Wizard to predict missing side chains and loops in the receptor. No residues or loops were missing in PDB file; therefore no changes were induced by this script. There were also no metal ions, cofactors were in structure; therefore their treatment was omitted. All 45 water molecules in the remaining structure were deleted since there was no information on the presence of structural waters that could mediate receptor-ligand interactions. Hydrogen bonding network was carried out at neutral pH, and proper orientations of aspartic acid, cysteine, glutamic acid, histidine and lysine residues were determined. Final energy minimization was done using Impref module in Protein Preparation Wizard. Figure 4.27 and Figure 4.28 show structure of YopE before and after protein preparation, respectively.

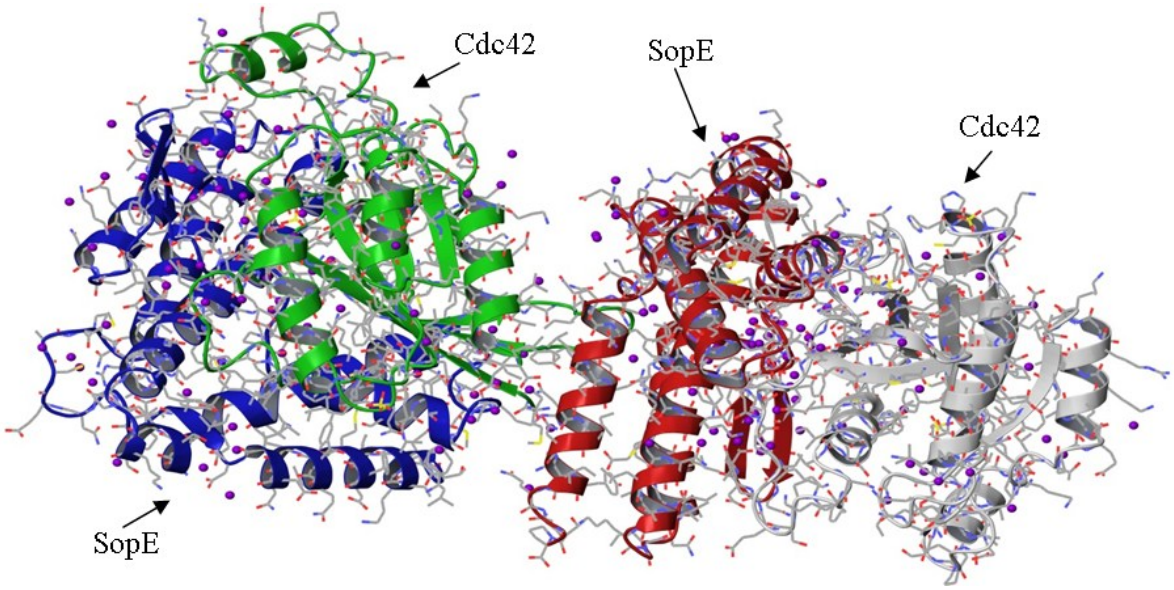


Figure 4.27. Structure of SopE before preparation. Both ribbon and tube molecular representations are displayed. Secondary structures are colored by chain name and waters are represented as ball & stick.

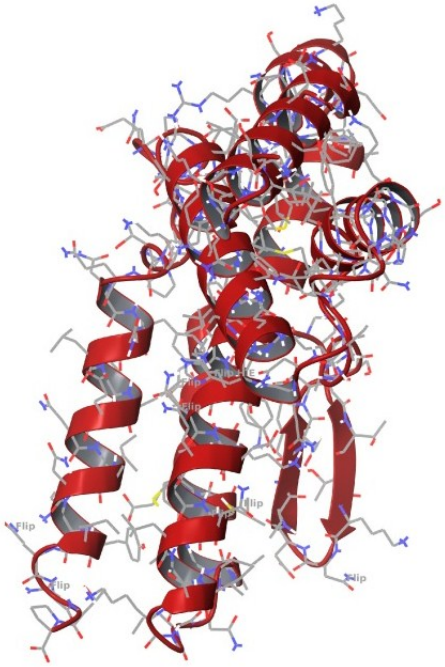


Figure 4.28. Structure of SopE after preparation. Both ribbon and tube molecular representations are displayed. Secondary structures are colored by chain name and waters are represented as ball & stick.

After protein preparation was carried out, the receptor grid was generated, where an energetically favorable site of interest on the receptor SopE was identified. Using Receptor Grid Generation module in Glide, binding site was presented as an enclosed cubic box, for which a center region in protein and side length was provided. Binding site was constructed such that it contained the GAGA loop (residue 166-169) at its center GAGA loop is the catalytic core of SopE and has the most prominent contribution to SopE-mediated nucleotide release from small GTPases [42]. Since there were no known or native ligands against SopE, side length of the cubic box was increased to 25 Å, such that it is large enough to enclose docked small library compounds. Three hydrogen bond constraints were defined on the backbones of residues Gly166, Gly168 and Ala169, since these residues are known to contact with switch regions of Cdc42 [42]. Rotation of hydroxyl groups was not allowed since an implication on the importance of these groups' flexibility was not found. Default van der Waals radius scaling settings were employed for generation of the receptor grid. Enclosing box that was used for grid generation is represented in Figure 4.29 and Figure 4.30.

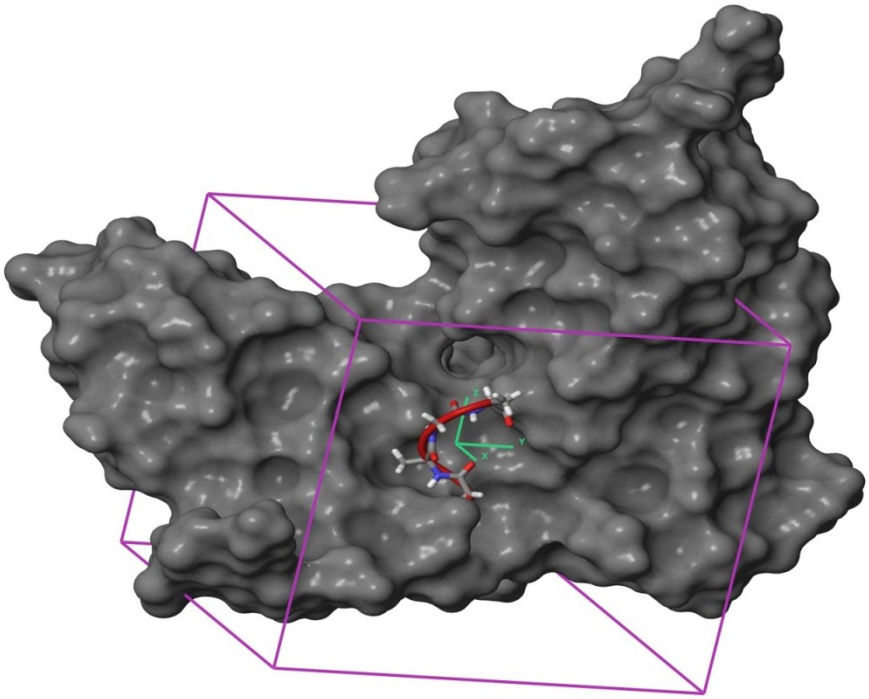


Figure 4.29. Enclosed grid box of SopE in molecule surface representation. Backbone and side chain of GAGA loop are also shown.

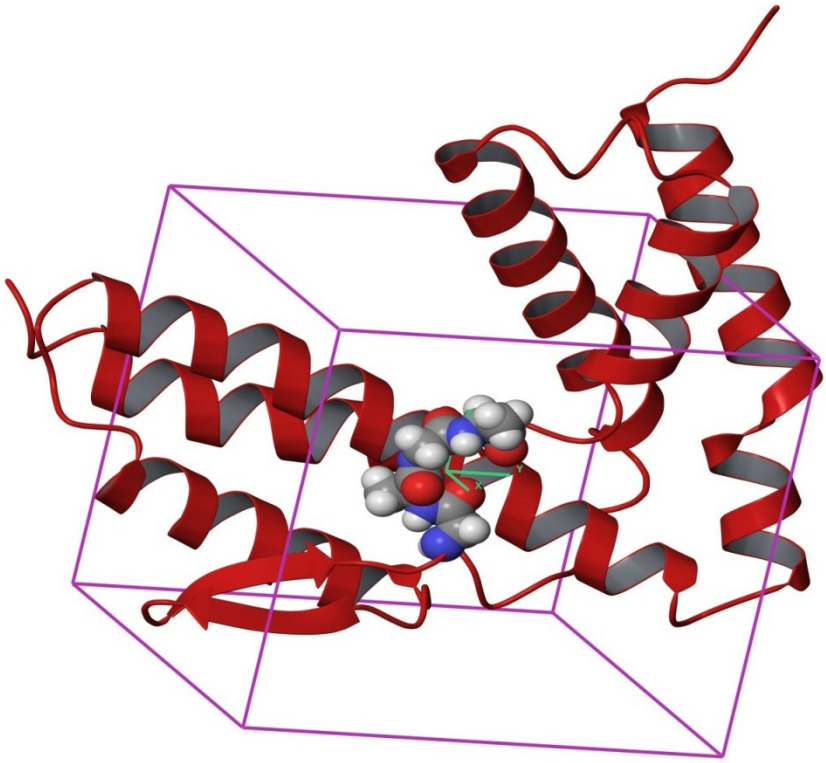


Figure 4.30. Enclosed grid box of SopE in ribbon representation. GAGA loop is shown as van der Waals spheres.

4.2.2. Pharmacophore Hypothesis Selection

Pharmacophore building, which facilitates the filtering and screening process, requires an initial set of known ligands. However, there are no known ligands identified for SopE to date. Therefore, as a first estimate, YopE-inhibitors derived hypothesis AADDR.49 was utilized for pharmacophore building toward SopE. Using this hypothesis, small molecule database generated from ZINC were pre-filtered. Molecules matching the hypothesis site points were sorted according to their fitness values in descending order, and the first 10,000 molecules were then used in for structure-based virtual screening. An unconstrained Glide standard precision docking was carried out with SopE grid. Top 10% molecules with highest Glide docking score were used in Glide extra precision docking. However, key interactions between hits and GAGA loop or switch regions of SopE were not observed. Therefore, it was concluded that pre-filtering database with hypothesis generated from YopE inhibitors is not suitable for finding potential SopE ligands. Additionally, initial 23 inhibitors of YopE were also docked to SopE without constraints (data not shown). Docking results in terms of GlideScores showed poor binding affinity of these compounds towards SopE. Interactions between GAGA loop residues of SopE and 23 compounds were not observed. As a result, the use of these 23 compounds was not pursued for SopE virtual screening.

Instead, a different procedure, named E-pharmacophores (or E-pharm) was applied, for which information about known ligands is not necessary. In contrast to ordinary pharmacophore modeling, this procedure makes use of receptor binding site, and generates pharmacophore sites according to the protein's residues can make interactions upon binding. In order to identify such residues, E-pharm procedure requires an initial docking stage. Schrödinger Company provides a fragment library, composed of unique small fragments, that is ready-to-use for this purpose [104]. After fragment docking, E-pharm module identifies possible pharmacophore sites on fragments utilizing docking energetic terms.

To define necessary pharmacophore features (i.e. hydrogen bond donor and acceptor, hydrophobic group, charged group), an unconstrained fragment docking by Glide XP was carried out with a set of 667 unique small fragments (6-37 atoms, MW range 32-226)

derived from the molecules in the medicinal chemistry literature and named Glide Fragment Library [104]. From there, protein-ligand interactions were characterized based on energetic contributions of each fragment to the docking score. Figure 4.31 shows the binding modes of all fragments on the SopE active site.

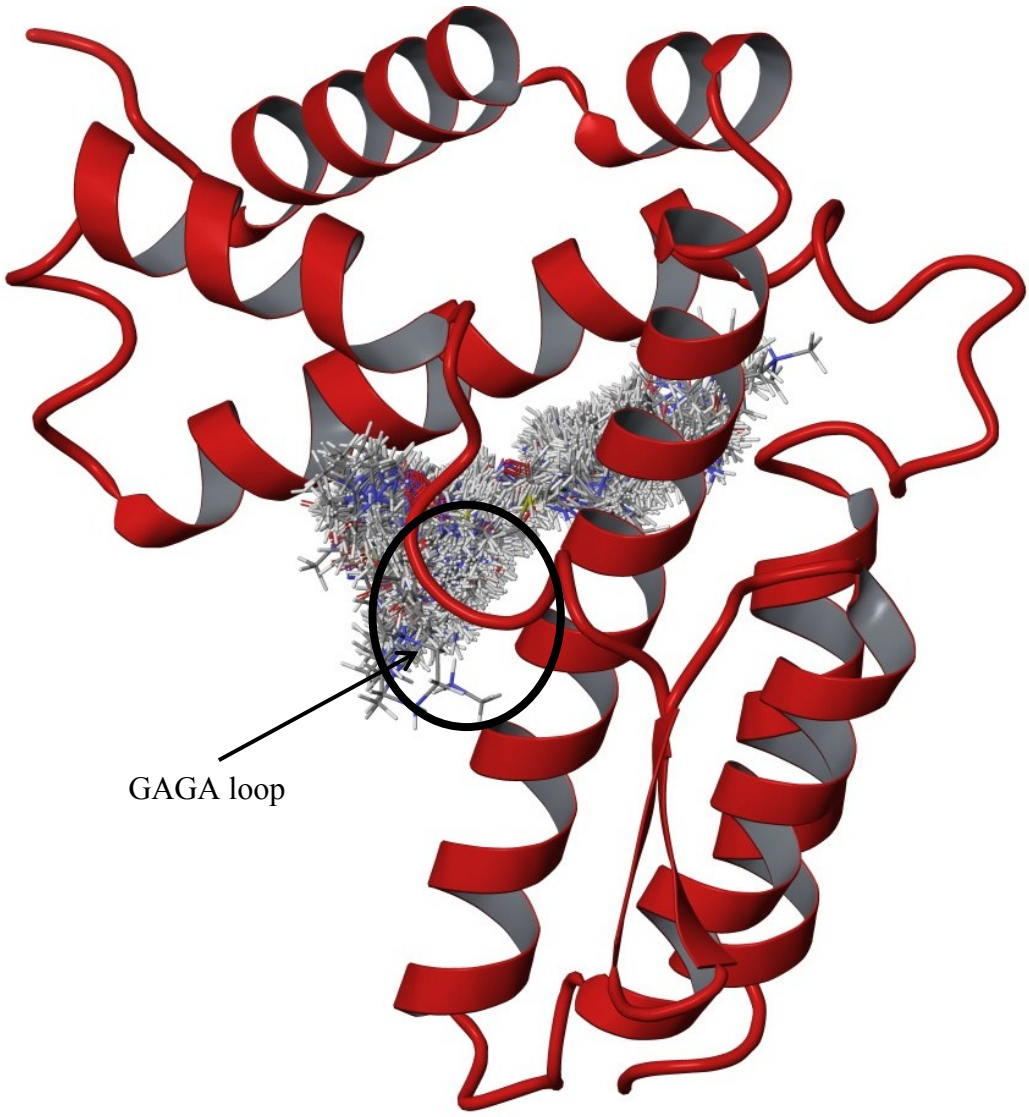


Figure 4.31. Binding modes of all fragments on SopE generated by Glide XP docking. SopE structure is represented in ribbons and 667 fragments are represented in wires.

By providing the pose viewer file from Glide XP docking as an input to E-pharm script, sites on fragments that are energetically favorable to binding were determined. Each site represented a pharmacophore feature. Based on their energetic contributions to the

docking score, all sites were ranked individually and listed as an output. A total of 21 pharmacophore sites that significantly contributed to the docking score were detected.

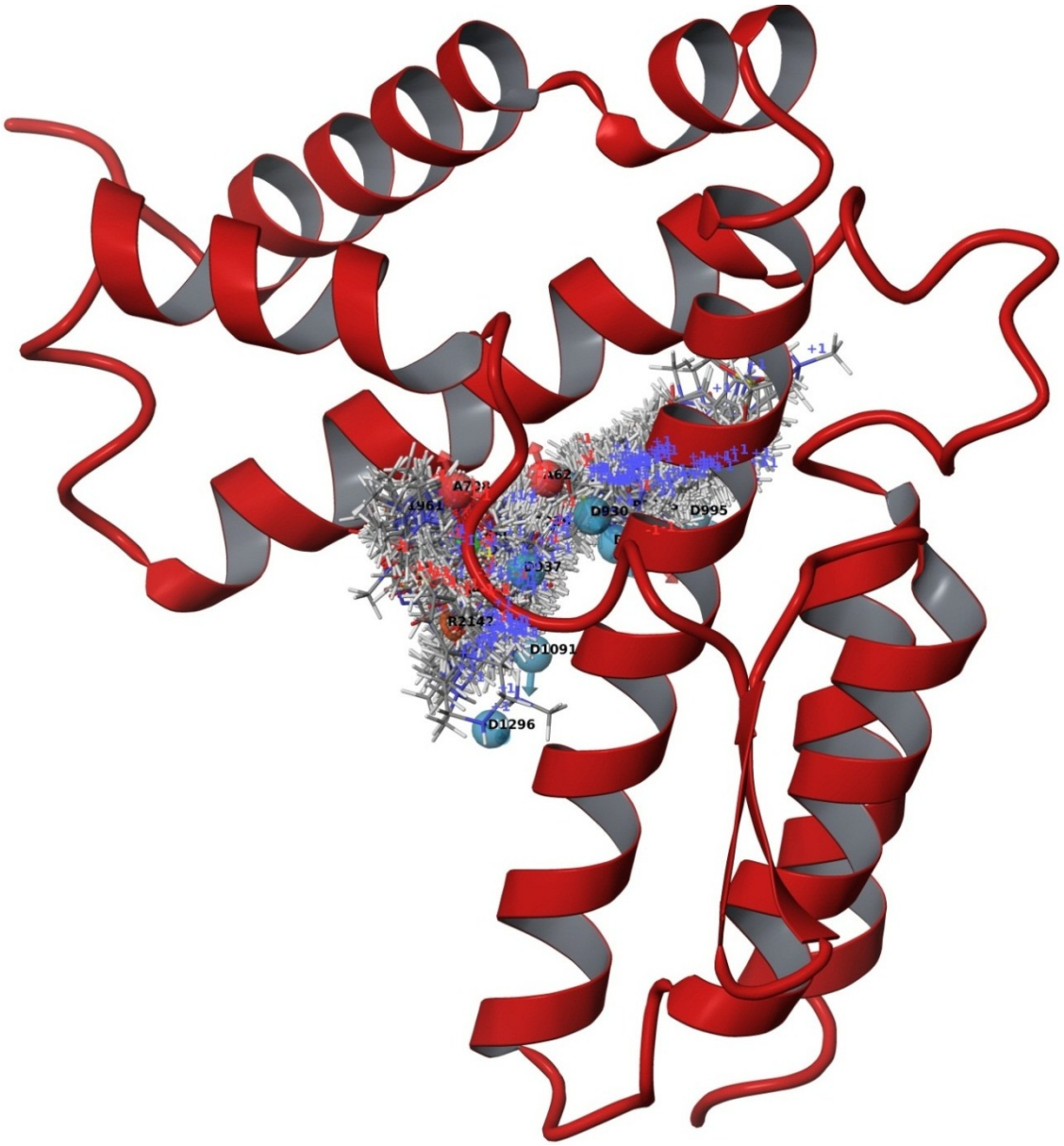


Figure 4.32. Site point location and types superimposed on fragments at the binding site of SopE. All site points were determined by E-pharm.

Figure 4.32 shows the overlap of pharmacophore site points with fragment binding modes. Ranking of all possible pharmacophore sites as well as energetic contributions, coordinates and feature types are listed in Table 4.20. Numbers in feature labels are indexes which are automatically assigned by E-pharm during feature generation.

Table 4.20. Pharmacophore features and ranks by E-pharm.

Rank	Feature label	Score	X	Y	Z	Type
1	D995	-1.80	-42.79	58.44	53.22	D
2	A708	-1.70	-30.15	57.8	55.76	A
3	N1548	-1.59	-35.87	53.23	58.95	N
4	D1296	-1.52	-31.51	48.41	53.15	D
5	A42	-1.49	-32.33	53.73	56.24	A
6	D1091	-1.29	-34.84	51.76	54.15	D
7	R2124	-0.98	-42.51	61.08	52.86	R
8	A62	-0.98	-34.14	58.86	54.35	A
9	R2224	-0.95	-41.82	59.85	57.27	R
10	R2233	-0.90	-39.8	58.26	53.84	R
11	D939	-0.90	-39.2	58.7	57.22	D
12	R1961	-0.81	-28.68	56.67	57.28	R
13	R2033	-0.77	-35.52	52.49	56.7	R
14	A81	-0.70	-43.29	56.69	55.55	A
15	D930	-0.70	-29.96	57.53	48.45	D
16	D937	-0.70	-31.93	54.98	52.91	D
17	H1450	-0.70	-31.37	55.91	55.68	H
18	D1282	-0.70	-37.06	56.69	52.44	D
19	R2142	-0.53	-29.74	52.3	54.63	R
20	D999	-0.34	-35.34	56.55	56.15	D
21	R2236	-0.20	-33.94	56.87	54.72	R

Since pharmacophore sites within a single hypothesis must be between three and seven, only a small portion of pharmacophore sites was used upon hypothesis generation. However, deciding on which sites to include in the hypothesis was not straightforward. For example, top four pharmacophore sites in Table 4.20 could not be combined since D995 and D1296 were remotely placed with respect to each other at the binding site (Figure 4.33).

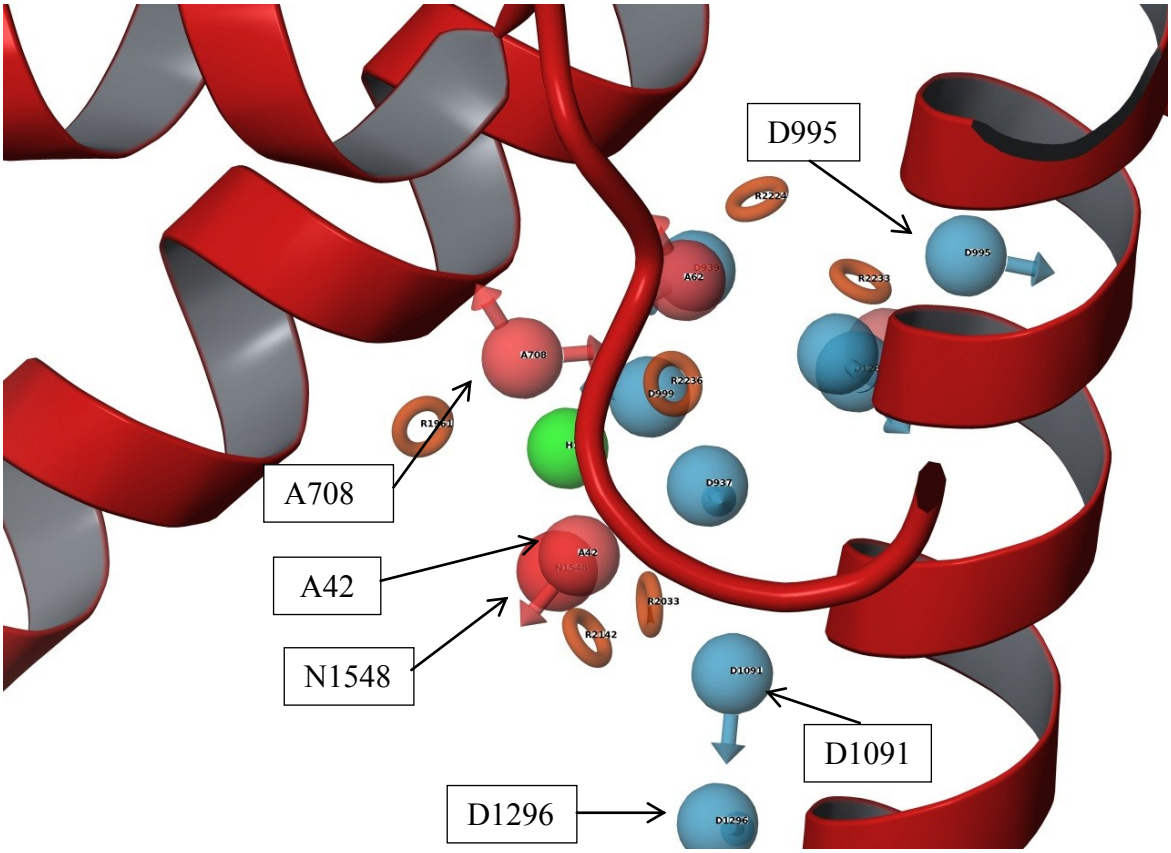


Figure 4.33. All possible pharmacophore site points located on the binding site of SopE.

A hypothesis that comprises distant site points would not yield many matching hits in a small molecule library. Similarly, since GAGA loop is very important in SopE infection mechanism, proximity of site points to the GAGA loop was also considered. Even though the energetic contribution of 4th ranking D1296 was slightly larger than 5th ranking A42, inclusion of site point A42 was favored since it was in the vicinity of the GAGA loop (Figure 4.34). Overall, it was aimed to cluster high ranking site points that were in close proximity with each other, and also with the GAGA loop. Thus, four site points, A708, N1548, A42 and D1091 were selected to create the hypothesis AADN (Figure 4.34).

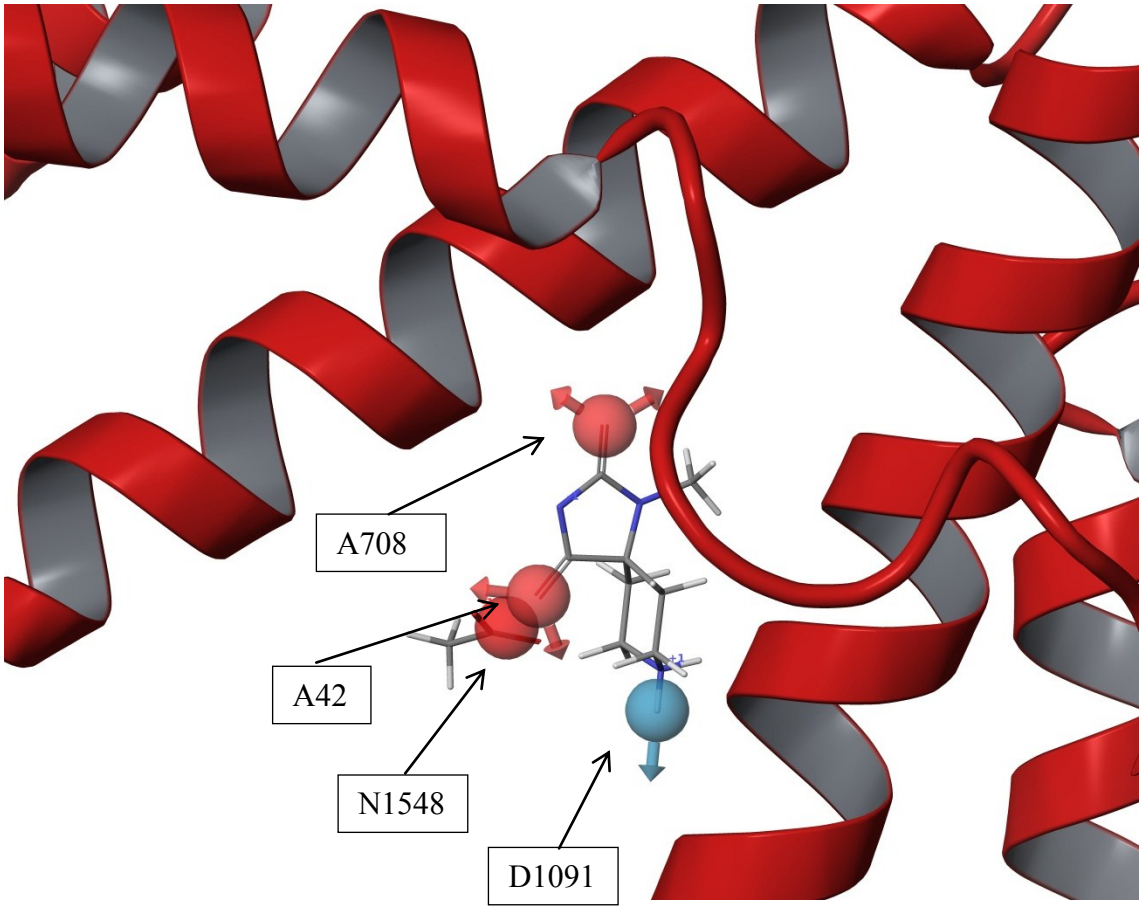


Figure 4.34. Selected hypothesis site points and their corresponding fragments located on the binding site of SopE.

After hypothesis determination, small molecule database was pre-filtered with 3D pharmacophore search. 30106 molecules out of 2.5 million matched the hypothesis. Number of molecules that matched the hypothesis was further reduced according to their fitness values. Maximum observed fitness value was 2.05, which was lower than YopE case. Therefore, fitness threshold was reduced to 0 fitness value for SopE, and hits having less fitness to hypothesis than that were rejected (minimum fitness value:-1). Figure 4.35 shows fitness of library molecules that matched the hypothesis AADN with a fitness value above 0. The small molecule database generated from ZINC was further reduced to 9970 compounds and structure-based virtual screening was continued with these compounds.

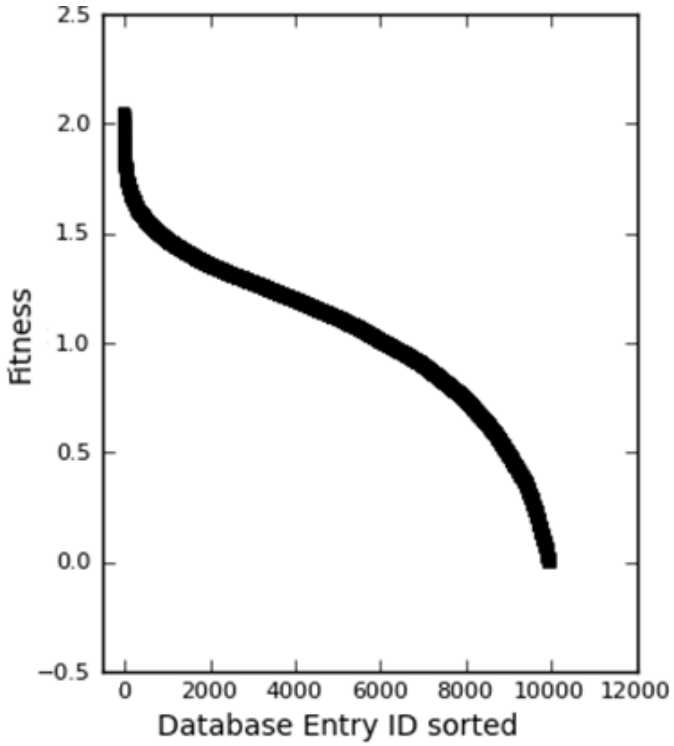


Figure 4.35. Scatter plot of fitness scores of 9970 small molecule database compounds filtered by E-pharm hypothesis AADN.

4.2.3. Docking Results and Post-Dock Analysis

Pharmacophore pre-filtering with E-pharm hypothesis AADN yielded 9970 compounds from the small molecule database generated from ZINC. These compounds were docked to the receptor SopE with Glide program to search for interactions between ligands and the protein. The position and orientation of each ligand relative to the receptor protein was determined and scored with Glide's internal scoring function GlideScore. As explained in Section 3.8, Glide standard and extra precision docking modes were used, successively. Remaining 9970 compounds were docked with Glide SP mode, and top 10% of poses were re-docked to SopE with more expensive Glide XP precision mode.

4.2.3.1. Standard Precision Docking. Glide SP docking was carried out with the previously prepared receptor grid and .maegz file (Maestro file extension) containing structures of 9970 compounds. Receptor grid was generated using SopE structure described in Section 4.2.1 with a binding site defined by a 25 Å box around GAGA loop (residue 166-169). Three hydrogen bond docking constraint were included on Gly166, Gly168, Ala169.

Constraint criterion was specified such that at least one constraint out of three is satisfied. Compounds were docked using the standard precision setting with the flexible docking option enabled. Default values were used for all other Glide docking settings used. Figure 4.36 shows plot of Glide SP scores of first 997 compounds, corresponding to top 10% of input molecules.

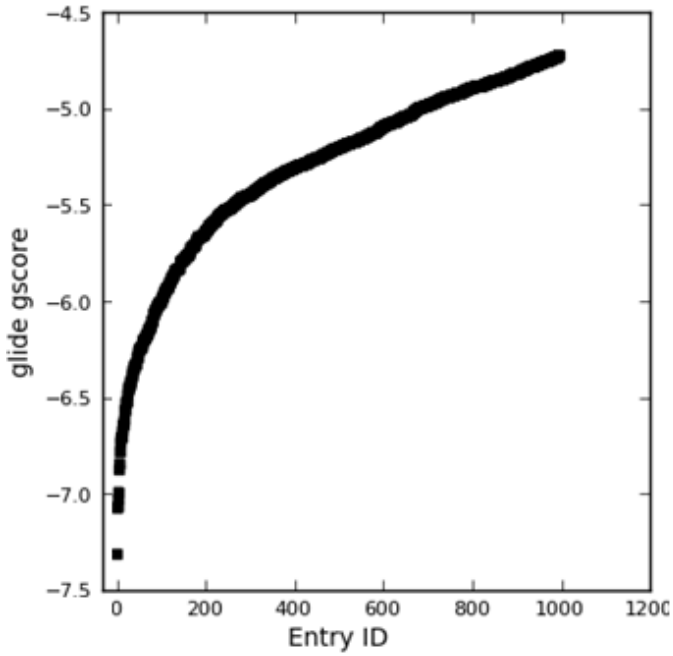


Figure 4.36. Plot of GlideScores of 997 small molecule database compounds docked to SopE in Glide SP mode. Docking scores are presented in kcal/mol.

GlideScores of top 997 compounds vary between -7.4 and -4.6 kcal/mol in Glide SP mode. A more negative GlideScore corresponds to a higher predicted binding affinity. Components of GlideScore are explained in detail in Section 3.8. Virtual screening protocol was preceded with top 10% compounds from Glide SP docking results. Using Premin utility in MacroModel, structures of docked ligands were relaxed and energetically minimized before Glide XP docking.

4.2.3.2. Extra Precision Docking. After energy minimization was carried out to top 997 compounds obtained from the Glide SP docking, all compounds were re-docked to SopE structure in Glide XP mode. The same receptor grid and hydrogen bond constraints on residues Gly166, Gly168, Ala169 were used for this mode. Similar to SP mode, constraint

criterion was specified such that at least one constraint out of three is satisfied. Compounds were docked using the extra precision setting with the flexible docking option enabled. Default values were used for all other Glide docking settings. Only 484 out of 997 compounds were successfully docked to SopE satisfied the constraint criterion. Figure 4.37 shows the plot of Glide XP scores of these 484 compounds.

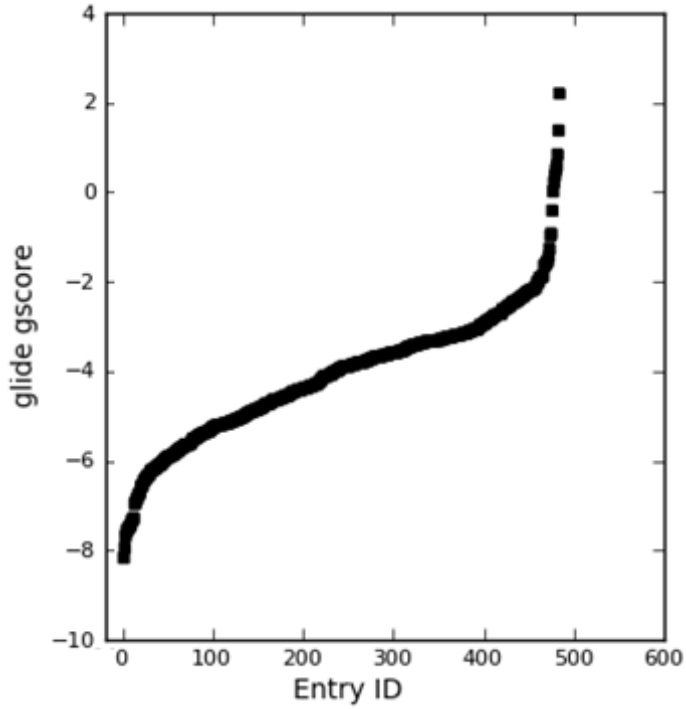


Figure 4.37. Plot of GlideScores of 484 small molecule database compounds docked to SopE in Glide XP mode. Docking scores are presented in kcal/mol.

GlideScores of all 484 compounds vary between -8.2 and -0.4 kcal/mol in Glide XP mode. Detailed analysis of docking parameters was carried out for hits with GlideScore -6.5 kcal/mol or above. This criterion yielded 20 hits out of 484 compounds. The quality of each docking was presented with a score and ranked accordingly. Docking results for these 20 hits were tabulated in Table 4.21. The docking score was calculated with GlideScore, which is based on ChemScore empirical scoring function. Glide scoring function is given in Equation 3.1.

Table 4.21. Glide XP results of top 20 hits (receptor: SopE, hypothesis: AADN). All values are shown in kcal/mol.

	Title	Glide score	Glide lipo	Glide hbond	Glide rewards	Glide evdw	Glide ecoul	Glide rotb	Glide esite	Glide emodel
1	ZINC03860635	-8.171	0.000	-0.159	-2.104	-9.675	**-21.016**	1.146	-8.29E-02	-34.444
2	ZINC10274669	-7.948	-0.769	-0.689	-0.889	**-28.284**	-15.035	0.736	-3.00E-01	-57.122
3	ZINC00370772	-7.681	-0.296	-0.510	**-2.298**	-14.588	-14.932	1.118	-2.87E-01	-37.519
4	ZINC09214236	-7.567	-0.713	-0.701	-1.982	-19.439	-14.445	0.548	-3.77E-01	-33.284
5	ZINC03874923	-7.567	-1.214	-0.931	-0.820	-22.435	**-21.508**	1.791	-3.51E-01	-50.343
6	ZINC17005625	-7.499	-0.202	-0.304	-2.050	-20.443	-12.938	0.274	-4.55E-01	-37.686
7	ZINC17020721	-7.489	-0.331	-0.315	-1.982	-20.335	-14.328	0.274	-2.88E-01	-38.399
8	ZINC19910989	-7.456	**-1.615**	-0.679	-1.002	-27.996	-18.820	1.179	-1.40E-01	**-60.893**
9	ZINC05033974	-7.410	-0.195	-0.581	-1.586	-19.878	**-21.332**	0.180	**-5.03E-01**	-44.115
10	ZINC09214236	-7.330	-0.410	-0.792	-2.031	-17.364	-14.950	0.548	-2.76E-01	-36.937
11	ZINC00375097	-7.324	-0.416	-0.224	**-2.324**	-19.409	-8.529	0.398	-7.96E-02	-39.996
12	ZINC05033974	-7.274	-0.280	-0.542	-1.554	-14.495	-20.791	0.180	-4.46E-01	-41.142
13	ZINC00388310	-6.960	-0.340	-0.371	-2.160	-16.710	-17.806	0.472	-4.79E-02	-43.441
14	ZINC00754304	-6.949	-0.274	-0.353	-1.204	-20.021	-18.936	0.143	-1.03E-01	-52.592
15	ZINC09214236	-6.863	-0.189	-0.320	-2.063	-19.842	-11.658	0.548	-4.71E-01	-32.037
16	ZINC19910989	-6.841	**-1.464**	**-1.090**	-0.792	**-28.583**	-18.217	1.179	-2.32E-01	-57.662
17	ZINC00097607	-6.763	-0.542	-0.701	-1.668	-16.003	-17.062	0.270	-1.69E-01	-37.690
18	ZINC39345243	-6.749	-1.153	**-0.976**	-0.746	-24.059	-19.919	0.892	-2.99E-01	-51.248
19	ZINC00155974	-6.700	-0.718	-0.356	-2.065	-17.510	-16.684	0.418	-4.84E-02	-41.773
20	ZINC17076964	-6.573	-1.109	-0.556	-0.849	-25.859	-18.860	0.961	-2.98E-01	-51.235

Hits in Table 4.21 were sorted according to their GlideScore values. First column of the table shows GlideScore values of each ligand, and the following seven columns include components of the GlideScore function, all values in kcal/mol. In the last column, another scoring function, Emodel, is listed. Hits can be ranked according to either GlideScore or Emodel. For both scoring functions, lower values are better. Difference between GlideScore and Emodel scoring functions is that Emodel is greatly dependent on internal strain energy of the generated poses. Therefore, it is reported that Emodel is not appropriate for comparing chemically distinct structures [87]. Additionally, it is reported that empirical GlideScore function should be used for ranking in virtual screening applications since it is more optimized for docking accuracy and database enrichment [87]. Therefore, GlideScore was used for comparing poses of different hits, and Emodel was used for deciding between different conformers (or poses) of a particular ligand.

Highest GlideScore components within each column are indicated in bold in Table 4.21. Similar to YopE case, individual values of GlideScore components are not directly correlated to ranking of corresponding ligand. Their contributions to final score are quite diverse throughout the hits. For example, for the ligand with the highest GlideScore, ZINC03860635, the Coulomb energy is highest among all contributions.

Five molecules out of final 20 hits were selected as potent inhibitors of SopE based on the GlideScore as well as additional criteria, such as absorption, distribution, metabolism and excretion (ADME) considerations and druglikeness of ligands and strain energy differences of ligands. These post-docking analyses are explained in the following sections. Consequently, hits that violate Lipinski's rule of five, have high ligand strain or show poor ADME properties, were eliminated. Eventually, five molecules were selected and they were visually inspected.

4.2.3.3. ADME and Molecular Properties of Final Hits. QikProp was used for this step, which predicts a large number of descriptors, such as pharmacodynamic properties and oral bioavailability characteristics for candidate ligand molecules. QikProp also recommends ranges for each descriptor, according to their similarity with 95% of known drugs [96].

Table 4.22 shows some of the properties of SopE hits determined by QikProp, which are #stars, SASA and its components, PSA, QPlogS and human oral absorption. Value of #stars indicates number of property that is not in 95% range of properties determined by known drugs. Recommended value for #stars is between 0-5, and high numbers indicate that the molecule is less drug-like. Number of stars is determined from a number of descriptors, such as MW, SASA, FOSA, FISA, PISA, WPSA, donorHB, accptHB, QPlogPo/w [96]. All hits in Table 4.22 were in recommended range except ZINC03860635, ZINC03874923 and ZINC00375097. For example, ZINC03860635, which has the highest GlideScore, had two star values, meaning that it has two outlier properties.

SASA is the total solvent accessible surface area in square angstroms, which has four different components. These are hydrophobic component (FOSA), hydrophilic component (FISA), π component (PISA) and weak polar component (WPSA). Recommended range of SASA value is between 300 and 1000 $Å^2$ [96]. SASA values of hits in Table 4.22 are in the acceptable range, ZINC10274669, ZINC19910989 and ZINC03874923 having the highest values. These values are indicated in bold in Table 4.22.

PSA is the van der Waals surface area of polar nitrogen and oxygen atoms and QPlogS is predicted aqueous solubility. Range for PSA is between 7 and 200 whereas QPlogS range is between -6.5 and -0.5 [96]. All hits in Table 4.22 are within these ranges. Last column represents predicted human oral absorption (HOA) on percent scale. QikProp suggest that molecules with more than 80% HOA exhibit good absorption and molecules having less than 25% HOA have poor absorption [96]. ZINC03860635 and ZINC03874923 have poor predicted human oral absorptions, with 0.7% and 0.1%, respectively. ZINC00375097 was also observed to lack hydrophobic solvent accessible surface area with 0 FOSA value. Therefore, these three molecules were eliminated. Highest HOA values were also indicated in bold.

Table 4.22. Pharmacokinetic properties of top 20 hits (receptor: SopE, hypothesis: AADN). Eliminated molecules are indicated by strikethrough.

Molecule	#stars	SASA	FOSA	FISA	PISA	WPSA	PSA	QPlogS	Human Oral Absorption
~~ZINC03860635~~	~~2~~	~~374.2~~	~~48.9~~	~~325.3~~	~~0.0~~	~~0.0~~	~~172.1~~	~~-0.33~~	~~0.7~~
ZINC10274669	0	**756.6**	334.0	283.3	44.0	95.3	170.9	-5.28	60.6
ZINC00370772	1	415.4	108.6	223.4	83.4	0.0	109.8	-1.21	54.1
ZINC09214236	1	475.6	156.6	224.0	94.9	0.0	121.1	-2.08	54.3
~~ZINC03874923~~	~~1~~	~~**701.7**~~	~~447.2~~	~~253.6~~	~~0.0~~	~~0.9~~	~~170.4~~	~~-2.12~~	~~0.1~~
ZINC17005625	1	471.2	162.5	211.7	97.0	0.0	120.0	-2.20	56.9
ZINC17020721	1	463.4	139.2	217.8	106.4	0.0	118.4	-2.13	55.4
ZINC19910989	1	694.2	97.3	212.7	384.1	0.0	147.5	-4.56	68.5
ZINC05033974	1	505.9	214.7	189.3	101.9	0.0	134.9	-2.69	58.3
ZINC09214236	1	441.4	160.4	197.5	83.5	0.0	115.4	-1.77	59.2
~~ZINC00375097~~	~~1~~	~~413.6~~	~~0.0~~	~~151.3~~	~~262.2~~	~~0.0~~	~~73.4~~	~~-1.97~~	~~**78.8**~~
ZINC05033974	1	502.7	221.9	201.7	79.2	0.0	136.0	-2.65	55.7
ZINC00388310	0	370.5	20.8	238.7	111.0	0.0	112.1	-0.90	44.9
ZINC00754304	0	593.0	215.0	163.4	214.6	0.0	110.4	-3.57	51.9
ZINC09214236	1	469.6	145.0	226.7	97.9	0.0	120.0	-2.02	53.8
ZINC19910989	0	**715.3**	103.4	219.2	392.7	0.0	151.0	-4.67	66.5
ZINC00097607	0	455.0	165.4	161.6	98.7	29.4	90.1	-2.16	45.9
ZINC39345243	0	688.5	502.8	145.5	40.2	0.0	137.6	-4.23	**79.1**
ZINC00155974	0	398.6	113.2	188.6	96.8	0.0	98.1	-1.29	57.9
ZINC17076964	0	613.3	291.0	126.9	151.6	43.8	115.0	-4.11	**86.4**

Table 4.23 shows the druglikeness properties of top 20 hits determined by Lipinski's rule of five. According to this rule, in order for a compound to be drug-like and orally active, it should have a molecular weight less than 500 daltons, hydrogen bond donor equal to or less than five, hydrogen bond acceptor equal to or less than 10 and partition coefficient (QPlogPo/w) less than five [96]. Since hydrogen bond acceptor and donor values of hits were determined from a number of configurations, they were calculated as

non-integers. QPlogPo/w (or LogP) is the predicted partition coefficient of a compound between octanol and water solution, which gives an estimate about compound's lipophilicity. Up to certain limit, compounds with higher lipophilicity have higher ability to permeate across biological membranes, which is necessary for a drug candidate. High lipophilicity, on the other hand, can result in a poor aqueous solubility. Lipinski determined the upper favorable limit of partition coefficient (QPlogPo/w) as five [79]. No violations of hits to the rule of five have been observed; therefore no compounds were eliminated with these criteria. Within desired limits, ZINC19910989 and ZINC17076964 showed the highest lipophilicity, which are indicated in bold in Table 4.23.

Table 4.23. Druglikeness of top 20 hits according to Lipinski's rule (receptor: SopE, hypothesis: AADN). Eliminated molecules are indicated by strikethrough.

Molecule	mol_MW	donorHB	accptHB	QPlogPo/w	RuleOfFive
~~ZINC03860635~~	~~210.14~~	~~4~~	~~8.8~~	~~-1.413~~	~~0~~
ZINC10274669	469.55	5	9	1.738	0
ZINC00370772	211.22	5	7.4	-1.102	0
ZINC09214236	253.26	5	9.1	-1.056	0
~~ZINC03874923~~	~~391.48~~	~~4.5~~	~~9~~	~~-1.496~~	~~0~~
ZINC17005625	253.26	5	9.1	-0.962	0
ZINC17020721	253.26	5	9.1	-1.049	0
ZINC19910989	386.40	1.5	6.2	**3.187**	0
ZINC05033974	323.31	4	9.2	0.458	0
ZINC09214236	253.26	5	9.1	-0.975	0
~~ZINC00375097~~	~~203.20~~	~~3~~	~~3.25~~	~~1.031~~	~~0~~
ZINC05033974	323.31	4	9.2	0.372	0
ZINC00388310	184.15	4	5.2	-0.399	0
ZINC00754304	368.39	4	5.75	0.456	0
ZINC09214236	253.26	5	9.1	-1.062	0
ZINC19910989	386.40	1.5	6.2	**3.231**	0
ZINC00097607	255.29	2	4.5	-0.627	0
ZINC39345243	392.45	2	9.2	2.74	0
ZINC00155974	198.18	3	5.2	0.363	0
ZINC17076964	376.43	2	6.75	3.433	0

4.2.3.4. Strain Energy Calculation. Glide docking enables ligand flexibility whereas the receptor remains essentially frozen. The docking allows ligands to be strained, so that they can fit better into the binding site. Therefore, after docking, the script called "strain-rescore", which calculates energies of free and docked conformations were used. Hits having more than 4 kcal/mol energy difference between their conformations received penalty, which is added to the GlideScore. Bound and free energy of hits, strain energy differences, penalties and adjusted GlideScores are listed in Table 4.24.

Table 4.24. Strain energy penalties of top 20 (receptor: SopE, hypothesis: AADN). Eliminated molecules are indicated by strikethrough. All values are shown in kcal/mol.

#	Title	Bound energy	Free energy	Strain energy	Strain penalty	Strain GlideScore	Glide emodel
~~1~~	~~ZINC03860635~~	~~16.182~~	~~5.816~~	~~10.365~~	~~1.591~~	-6.580	~~-34.444~~
~~2~~	~~ZINC10274669~~	~~59.025~~	~~26.953~~	~~32.072~~	~~7.018~~	-0.930	~~-57.122~~
3	ZINC00370772	49.776	45.792	3.984	0.000	**-7.681**	-37.519
4	ZINC09214236	64.928	64.928	0.000	0.000	**-7.567**	-33.284
~~5~~	~~ZINC03874923~~	~~11.363~~	~~-5.676~~	~~17.039~~	~~3.260~~	-4.307	~~-50.343~~
6	ZINC17005625	68.736	61.171	7.565	0.891	-6.608	-37.686
7	ZINC17020721	62.430	62.415	0.015	0.000	**-7.489**	-38.399
8	ZINC19910989	16.478	13.041	3.437	0.000	**-7.456**	-60.893
~~9~~	~~ZINC05033974~~	~~67.590~~	~~54.666~~	~~12.924~~	~~2.231~~	-5.179	~~-44.115~~
10	ZINC09214236	65.940	64.687	1.253	0.000	**-7.330**	-36.937
~~11~~	~~ZINC00375097~~	~~11.928~~	~~11.737~~	~~0.191~~	~~0.000~~	-7.324	~~-39.996~~
12	ZINC05033974	64.699	63.128	1.572	0.000	**-7.274**	-41.142
13	ZINC00388310	13.084	10.852	2.232	0.000	-6.960	-43.441
14	ZINC00754304	45.244	38.999	6.245	0.561	-6.388	-52.592
15	ZINC09214236	69.688	64.003	5.685	0.421	-6.442	-32.037
16	ZINC19910989	13.163	8.705	4.459	0.115	-6.726	-57.662
17	ZINC00097607	19.312	13.621	5.691	0.423	-6.340	-37.690
18	ZINC39345243	55.915	46.046	9.869	1.467	-5.282	-51.248
19	ZINC00155974	16.180	14.826	1.353	0.000	-6.700	-41.773
20	ZINC17076964	40.677	32.146	8.531	1.133	-5.44	-51.235

As a result, 11 out of 20 hits received strain penalties. ZINC10274669 had the highest strain penalty with 7 kcal/mol, which is almost as much as its original docking score. ZINC03860635, ZINC03874923 and 9th ranking ZINC05033974 also received relatively high strain penalties, which lowered their GlideScores considerably.

Values indicated in bold in Table 4.24 show the hits with GlideScores above -7 kcal/mol after strain correction. As a result, molecules ZINC00370772, ZINC09214236, ZINC17020721, ZINC19910989 and ZINC05033974 were chosen. More than one poses of two different hits were observed in the table. ZINC09214236 has three poses with rankings 4, 10 and 15. ZINC19910989 has two poses with rankings 8 and 16, whereas ZINC05033974 has three poses with rankings 9, 12 and 19. Selection of poses of a particular molecule was performed based on their Emodel values. For ZINC09214236 molecule, 10th ranking pose has the highest model score (-36.9 kcal/mol), therefore it is selected for visual inspection. Similarly, for ZINC19910989, 8th scoring pose was selected. For ZINC05033974, 9th rank, which has the highest model value, was not selected since it has too much strain. Remaining 12th and 19th scoring poses of this molecule were compared. Since their Emodel values of ZINC05033974 poses were very close, 12th ranked pose of ZINC05033974 selected due to its higher GlideScore.

4.2.3.5. Binding Free Energy Calculation. Prime MMGB-SA was used to estimate relative binding affinity of top 20 hits. The aim of this calculation was to see if there is a correlation between estimated free energy of binding and docking score. For this purpose, pose viewer file of hits and the receptor SopE is provided to MM-GBSA module as an input.

Binding free energies of each ligand along with its components are calculated as explained in Section 3.8.1.3and represented in Table 4.25. A scatter plot including binding free energies versus GlideScores of top 20 hits were shown in Figure 4.38. No direct correlation was observed between binding free energies and docking scores. Additionally, free binding energies were not consistent with rankings of the hits. Therefore, this approach was not considered in decision-making step of proposed molecules. Additionally, correlation between Emodel and binding free energies were plotted (Figure 4.39). Positive correlation was observed in this case.

Table 4.25. Binding free energies of top 20 (receptor: SopE, hypothesis: AADN). Eliminated molecules are indicated by strikethrough. All values are shown in kcal/mol.

Title	DG bind Coulomb	DG bind Covalent	DG bind vdW	DG bind SolvGB	DG bind Lipo	DG bind total
~~ZINC03860635~~	~~-49.07~~	~~1.64~~	~~-23.29~~	~~52.59~~	~~-5.55~~	~~-32.33~~
ZINC10274669	-38.14	3.45	-33.41	33.53	-18.53	-56.18
ZINC00370772	0.33	2.45	-24.88	5.27	-7.58	-26.71
ZINC09214236	1.18	3.88	-32.36	7.48	-12.54	-38.38
~~ZINC03874923~~	~~-34.02~~	~~12.03~~	~~-40.79~~	~~46.62~~	~~-28.79~~	~~-49.63~~
ZINC17005625	-12.00	9.47	-21.93	10.80	-12.36	-30.65
ZINC17020721	-2.41	1.71	-27.33	9.95	-8.50	-29.02
ZINC19910989	-40.38	9.21	-39.50	49.59	-22.78	-48.45
~~ZINC05033974~~	~~-19.91~~	~~3.64~~	~~-28.37~~	~~24.15~~	~~-7.29~~	~~-35.59~~
ZINC09214236	-4.98	4.51	-28.92	3.86	-9.78	-40.76
~~ZINC00375097~~	~~-0.60~~	~~1.48~~	~~-19.63~~	~~-3.47~~	~~-6.70~~	~~-34.24~~
ZINC05033974	-33.32	3.59	-21.27	24.56	-5.53	-42.27
ZINC00388310	-48.35	2.99	-17.16	40.06	-5.59	-36.51
ZINC00754304	-48.85	8.74	-25.93	39.97	-12.53	-43.97
ZINC09214236	-9.46	8.21	-24.41	6.70	-12.66	-36.71
ZINC19910989	-42.15	15.97	-43.21	43.60	-21.37	-52.96
ZINC00097607	-37.70	9.19	-21.11	31.42	-12.02	-36.35
ZINC39345243	-41.28	12.44	-37.86	43.29	-19.31	-49.74
ZINC00155974	-46.77	2.94	-17.93	37.77	-8.14	-38.03
ZINC17076964	-41.27	10.35	-35.27	41.68	-18.77	-49.76

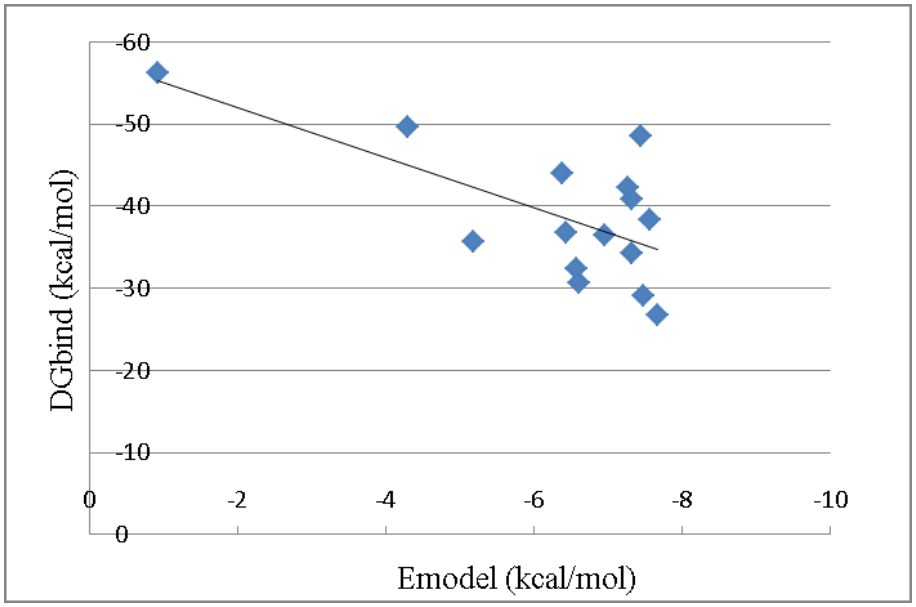


Figure 4.38. Correlation between GlideScores and MM-GBSA predicted binding affinities for top 20 hits of SopE.

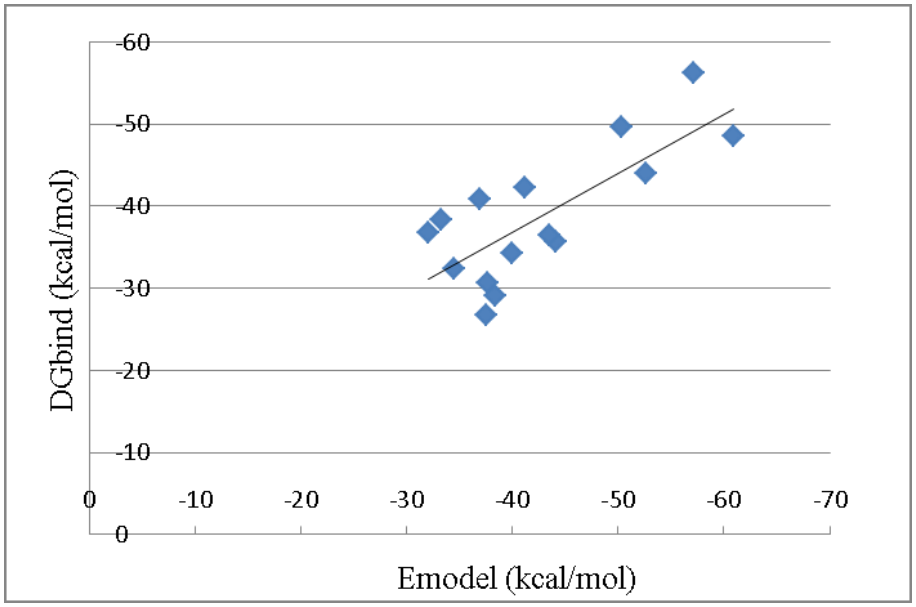


Figure 4.39. Correlation between Emodel and MM-GBSA predicted binding affinities for top 20 hits of SopE.

4.2.4. Visual Inspection of Proposed Hits

Binding modes of selected hits and their interactions with the receptor SopE were analyzed in this section. The list of proposed molecules for inhibition of SopE is given in

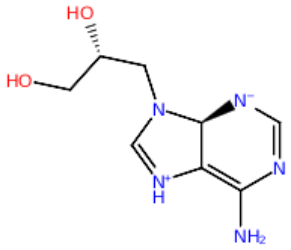
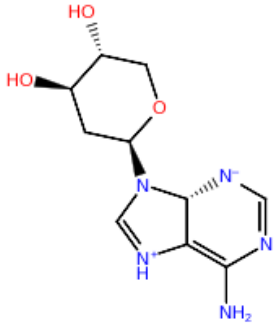
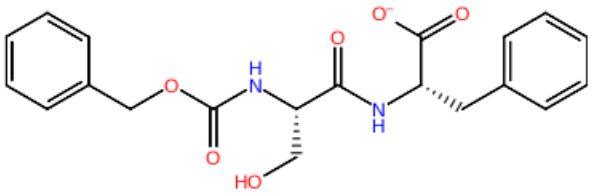
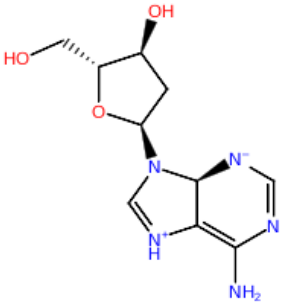
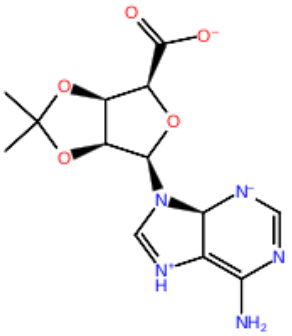
Table 4.26. Table 4.27 shows the chemical formulas and 2D structures of the proposed molecules.

Table 4.26. Summary of proposed molecules for SopE. All values are shown in kcal/mol.

Initial rank	Name	Database Title	Strain GlideScore	Glide emodel
3	S1	ZINC00370772	-7.681	-37.519
7	S2	ZINC17020721	-7.489	-38.399
8	S3	ZINC19910989	-7.456	-60.893
10	S4	ZINC09214236	-7.330	-36.937
12	S5	ZINC05033974	-7.274	-41.142

The binding mode and ligand interaction maps of these molecules were represented in Figure 4.40-Figure 4.44. These figures show the binding mode of the five proposed molecules at the binding site of SopE along with their interaction diagrams. Backbone interactions are shown with solid line whereas side chain interactions are shown with dashed lines. The residues are shown in their three-letter codes and are colored according to their amino acid types. Hydrogen bonds are represented as pink lines in interaction diagrams. Solvent exposed regions are indicated with yellow spheres. Hydrophobic contacts are not explicitly shown in figures, but they are explained within the text. Before visual inspection of proposed molecules, their structures were listed. Table 4.27 shows the chemical formulas and structures of the proposed molecules.

Table 4.27. Chemical formulas and structures of proposed SopE inhibitors.

Title	Chemical formula	2D Structure
S1	$C_8H_{13}N_5O_2$	
S2	$C_{10}H_{15}N_5O_3$	
S3	$C_{20}H_{21}N_2O_6$	
S4	$C_{10}H_{15}N_5O_3$	
S5	$C_{13}H_{16}N_5O_5$	

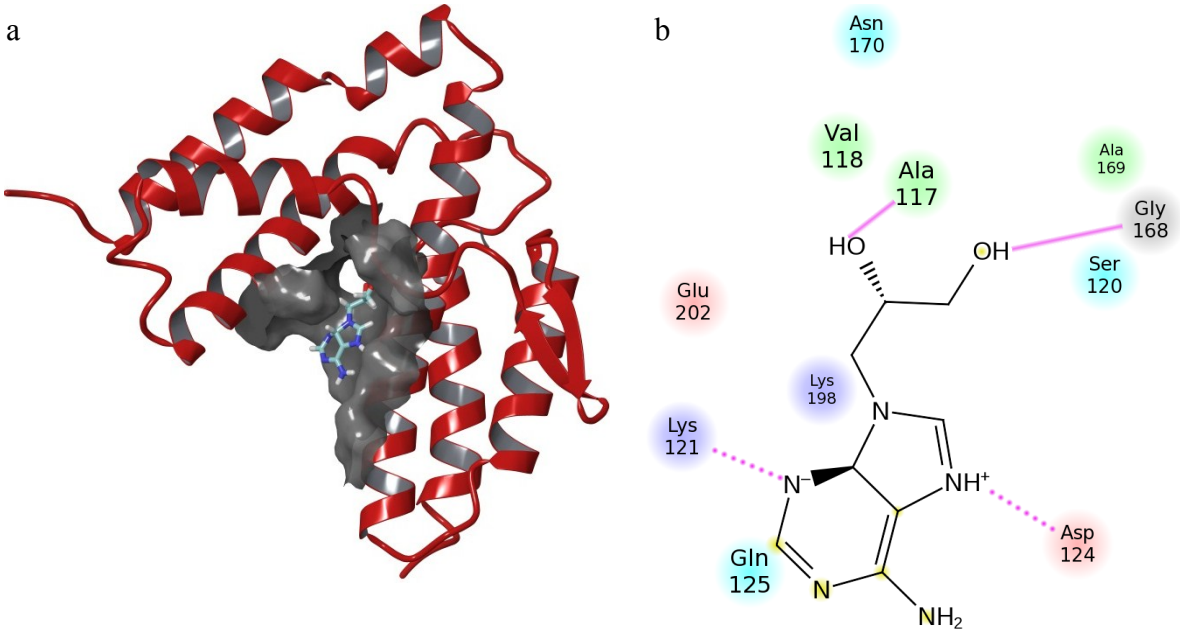


Figure 4.40. (a) Binding mode (b) Ligand interaction map of S1.

Compound S1, with molecular formula $C_8H_{13}N_5O_2$, has two ring systems, 15 heavy atoms and three rotatable bonds. S1 makes a total of four hydrogen bonds with SopE residues. There are two hydroxyl groups in the structure of S1. One hydroxyl group makes hydrogen bond with backbone of Gly168 and the other one makes hydrogen bond with backbone of Ala117 of SopE. Charged –NH group of ligand makes hydrogen bond with the side chain of Asp124 residue. Nitrogen atom in the larger ring system makes a hydrogen bond with the side chain of Lys121 residue of SopE. Ligand is buried in the binding site and does not have a notable solvent exposed region. No hydrophobic contacts were observed between S1 and SopE. Out of three constraints on GAGA loop (Gly166, Gly168, Gly169), S1 made one interaction with Gly168 residue.

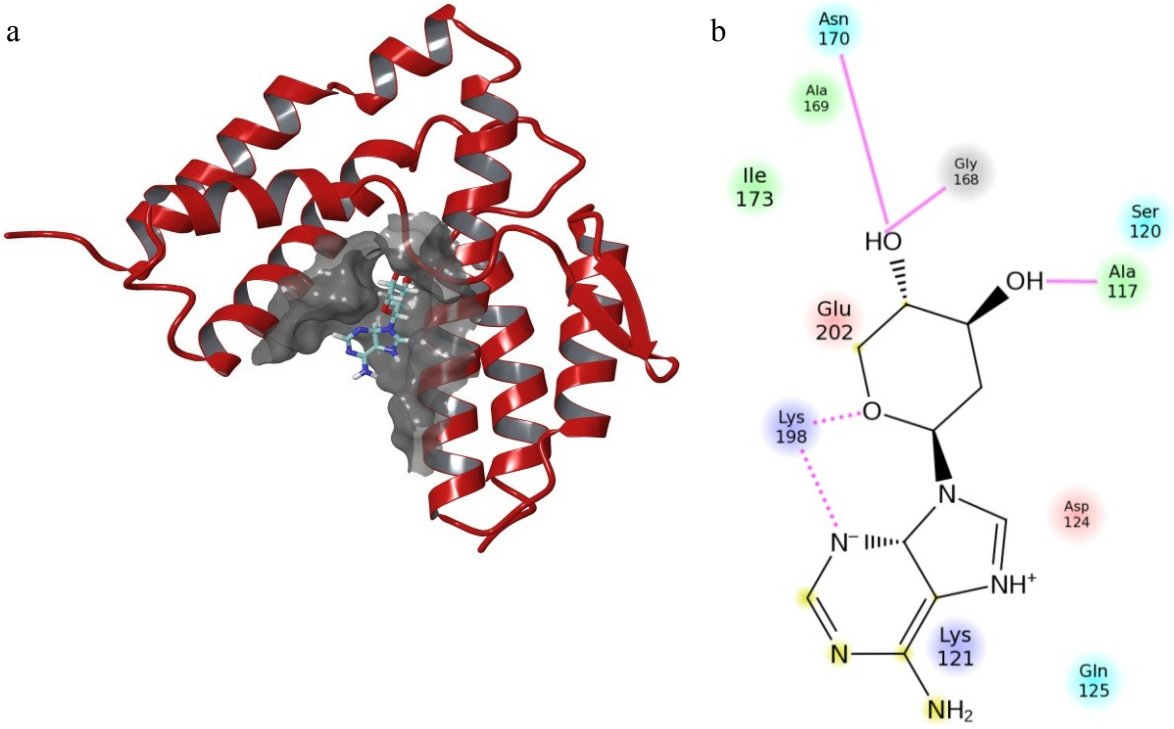


Figure 4.41. (a) Binding mode (b) Ligand interaction map of S2.

Compound S2, with molecular formula $C_{10}H_{15}N_5O_3$, has three ring systems, 18 heavy atoms and one rotatable bond. S2 makes a total of five hydrogen bonds with SopE residues. There are two hydroxyl groups in the structure of S2. One hydroxyl group makes hydrogen bond with backbone of Ala117. Other hydroxyl group has two hydrogen bond interactions with SopE residues, one with the backbone of Gly168 and one with the backbone of Asn170 residue. Side chain atoms of Lys198 of SopE make two hydrogen bonds with two different ring systems of the ligand. Lys198 interacts with charged nitrogen atom and oxygen atom. Ligand is buried in the binding site and has a solvent exposed region around its amine group. No hydrophobic contacts were observed between S2 and SopE. Out of three constraints on GAGA loop (Gly166, Gly168, Gly169), S2 made one interaction with Gly168 residue.

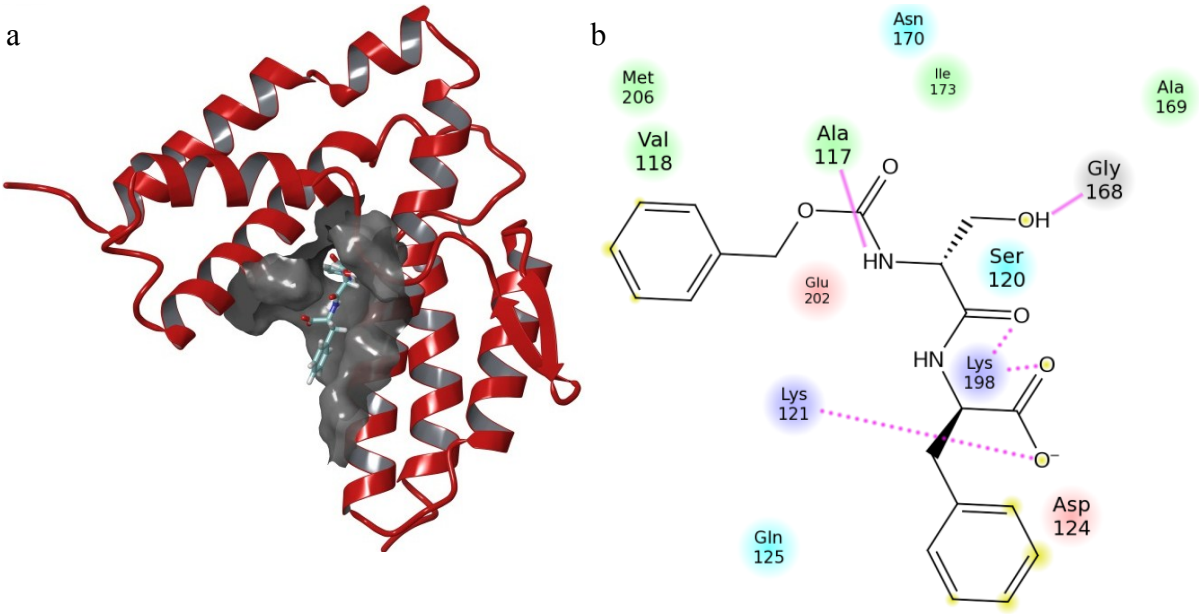


Figure 4.42. (a) Binding mode (b) Ligand interaction map of S3.

Compound S3, with molecular formula $C_{20}H_{21}N_2O_6$, has two ring systems, 28 heavy atoms and 10 rotatable bonds. ZINC19910989 makes a total of five hydrogen bonds with SopE residues. One of the hydroxyl groups in the structure of S3 makes hydrogen bond with the backbone of Gly168. Side chain atoms of Lys198 make two hydrogen bond interactions with uncharged lone O atoms of the ligand. Charged oxygen atom close to Asp124 makes a hydrogen bond with the side chain of Lys121. Backbone of Ala117 interacts with –NH group of the ligand. One ring of the ligand has hydrophobic contacts with Val118 and Met206 of SopE. Ala117 also has hydrophobic contacts with the middle region of the S3. Ligand is buried in the binding site and does not have a notable solvent exposed region. Out of three constraints on GAGA loop (Gly166, Gly168, Gly169), S3 made one interaction with Gly168 residue.

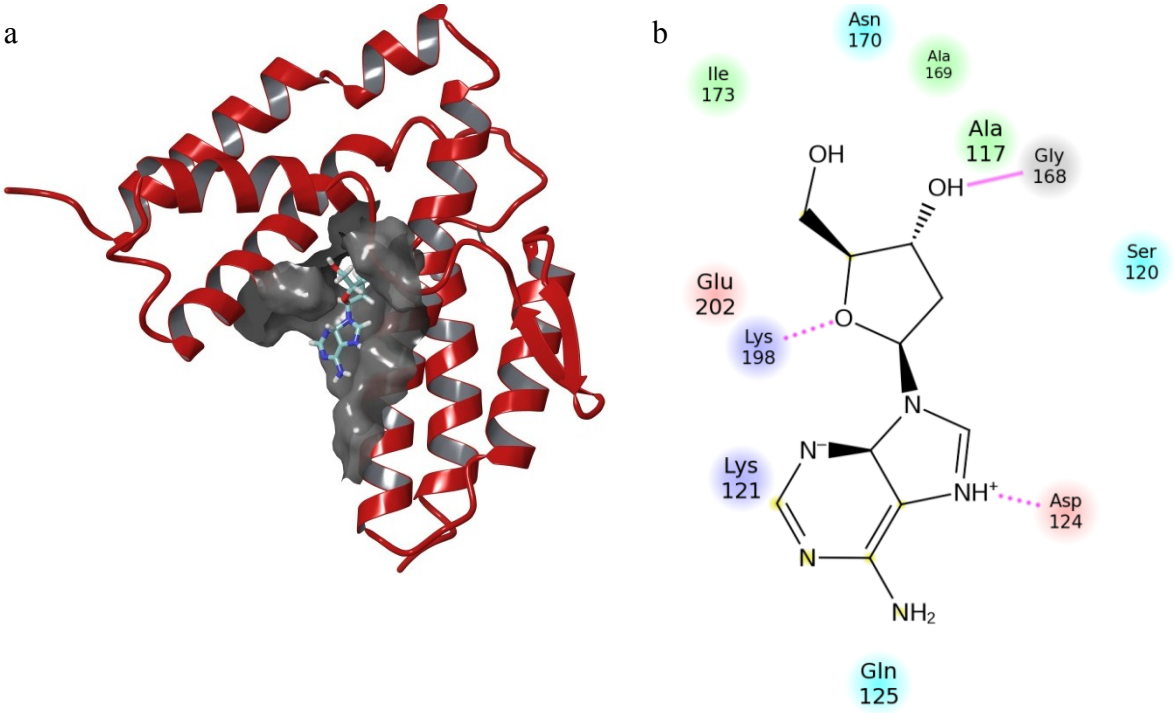


Figure 4.43. (a) Binding mode (b) Ligand interaction map of S4.

Compound S4, with molecular formula $C_{10}H_{15}N_5O_3$, has three ring systems, 18 heavy atoms and two rotatable bonds. S4 makes a total of three hydrogen bonds with SopE residues. There are one hydroxyl group in the structure of S4, and this –OH group makes a hydrogen bond interaction with the backbone of Gly168. Backbone of Asp124 of SopE makes hydrogen bond with charged –NH group of the ligand. Oxygen atom in tetrahydrofuran ring makes hydrogen bond with the side chain of Lys198. Hydrophobic contacts were observed between Ala117 and the ligand. Ligand is buried in the binding site and does not have a notable solvent exposed region. Out of three constraints on GAGA loop (Gly166, Gly168, Gly169), S4 made one interaction with Gly168 residue.

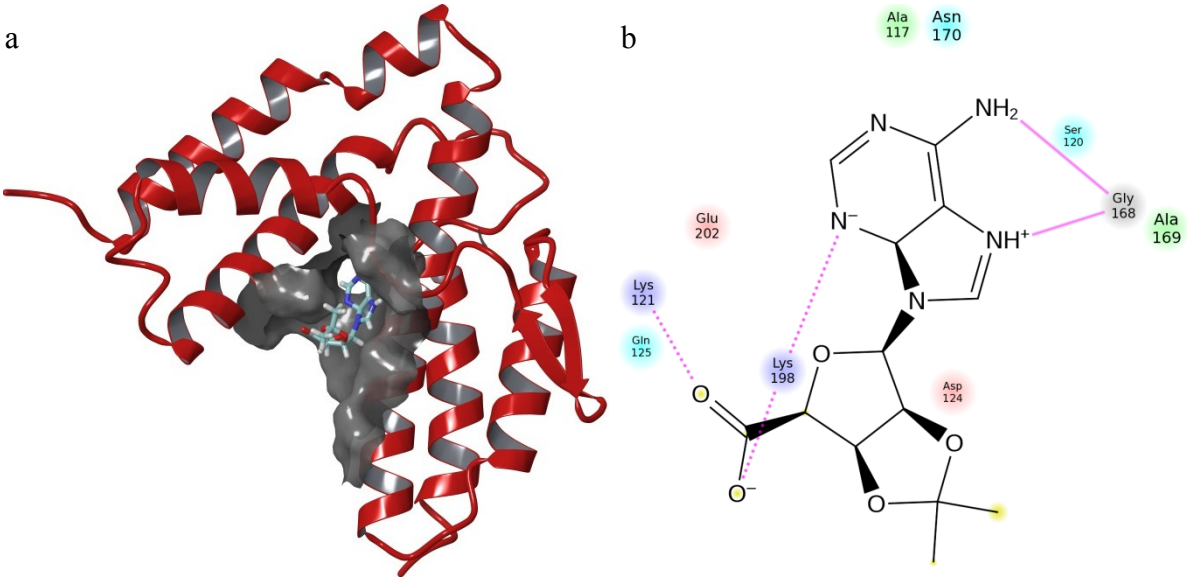


Figure 4.44. (a) Binding mode (b) Ligand interaction map of S5.

Compound S5, with molecular formula $C_{13}H_{16}N_5O_5$, has three ring systems, 23 heavy atoms and two rotatable bonds. S5 makes a total of five hydrogen bonds with SopE residues. Two hydrogen bonds are present between backbone atoms of Gly168 and S5. Gly168 interacts both with amine group and charged –NH group of the ligand. Side chain of Lys121 makes hydrogen bond interactions with the lone oxygen atom. Similar to Gly168, Lys198 also makes two hydrogen bonds. One negatively charged nitrogen atom in benzene ring and one negatively charged oxygen atom interact with side chain atoms of Lys198 residue. Hydrophobic contacts were observed between Ala117 and the ligand. Ligand is buried in the binding site and has a solvent exposed region around its neopentane group. Out of three constraints on GAGA loop (Gly166, Gly168, Gly169), S5 made one interaction with Gly168 residue. Summary of all interactions between SopE residues and proposed molecules were represented in Table 4.28.

Table 4.28. Summary of H-bond interactions between and proposed molecules of SopE.

Title/ Residue	Ala117	Lys121	Asp124	Gly168	Asn170	Lys198
S1	+	+	+	+		
S2	+			+	+	++
S3	+	+		+		++
S4			+	+		+
S5		+		++		++

In this table, "+" sign indicates presence of a hydrogen bond and number of signs indicates the number of hydrogen bonds between ligand and a particular receptor residue. Structure-based virtual screening of SopE was carried out with at least one hydrogen bond criterion on GAGA loop residues, Gly166, Ala167, Gly168 and Ala169. All proposed molecules favored the interaction with Gly168 residue. GAGA loop is reported to be an important promoter of nucleotide release [42]. Importance of GAGA loop was investigated with a mutation of Gly168 to alanine, and it was observed that SopE was no longer able to mediate nucleotide release with this mutation [42]. Therefore, proposed molecules' tendency to interact with Gly168 was interpreted as a positive outcome. Additionally, Asp124 and Lys198 were also reported to be important residues for stabilizing switch 1 regions of small GTPases.

Overall, visual inspection showed that proposed molecules represent favorable interactions with SopE using Glide docking protocol. It was observed that Gly168 residue preferred interactions with either hydroxyl group or amine group in ring systems. Lys198 interactions were mostly with lone oxygen or nitrogen atoms of ligands as well as nitrogen atoms. Two interactions between charged –NH groups and Asp124 residue of SopE was observed, whereas Ala117 did not prefer specific atom or group type upon ligand binding. Ala117 residue also participated in hydrophobic interactions. Four out of five proposed SopE inhibitors include an aminopurine ring, which may be a functionally important group upon SopE inhibition. These aminopurine rings bind to Lys 121, Asp124, Gly168 and Lys198 residues of SopE within different ligands. When the proposed molecules were superimposed based on their interaction with SopE (Figure 4.45), it was seen that their binding modes are similar. Proposed molecules with aminopurine groups overlaid well and similar spatial arrangement of S1, S2 and S4 was observed Figure 4.46a). In contrast, aminopurine group of S5 was faced towards GAGA loop (Figure 4.46b). S3, which does not contain this group, also overlaid well with S1, S2 and S4 (Figure 4.46c). It was observed that the region of S3 that interacts with SopE residues aligned with aminopurine groups of other proposed molecules.

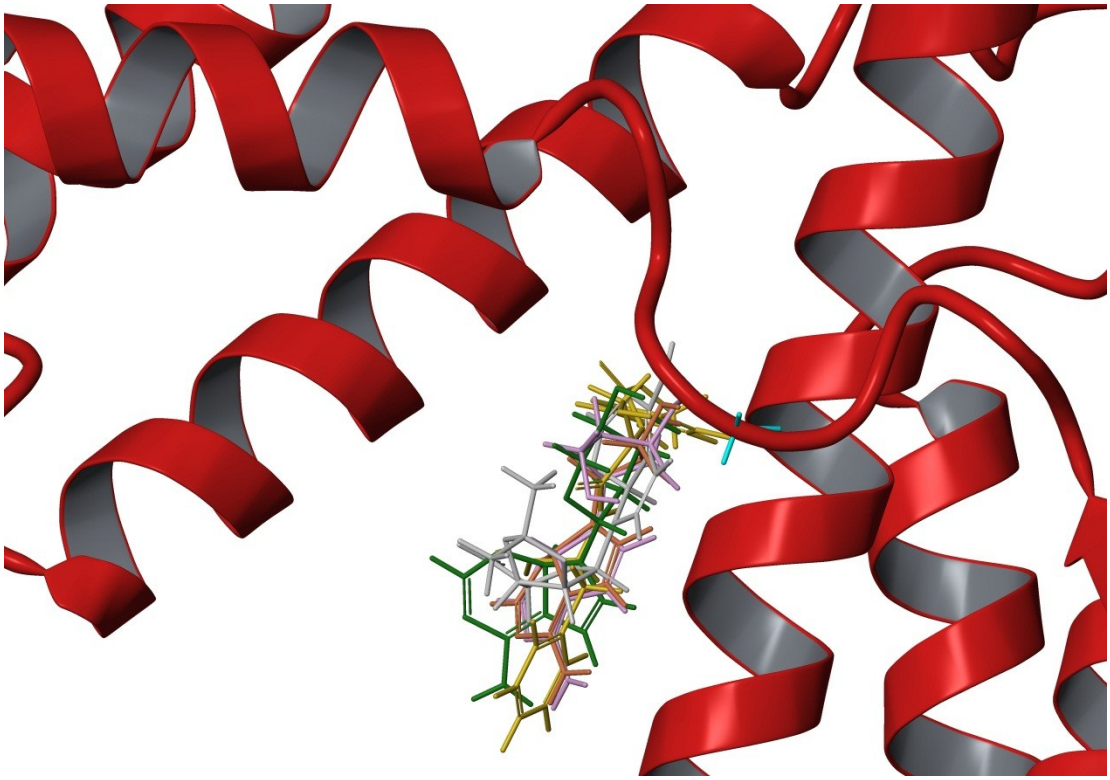


Figure 4.45. Proposed SopE inhibitors in the SopE pocket.

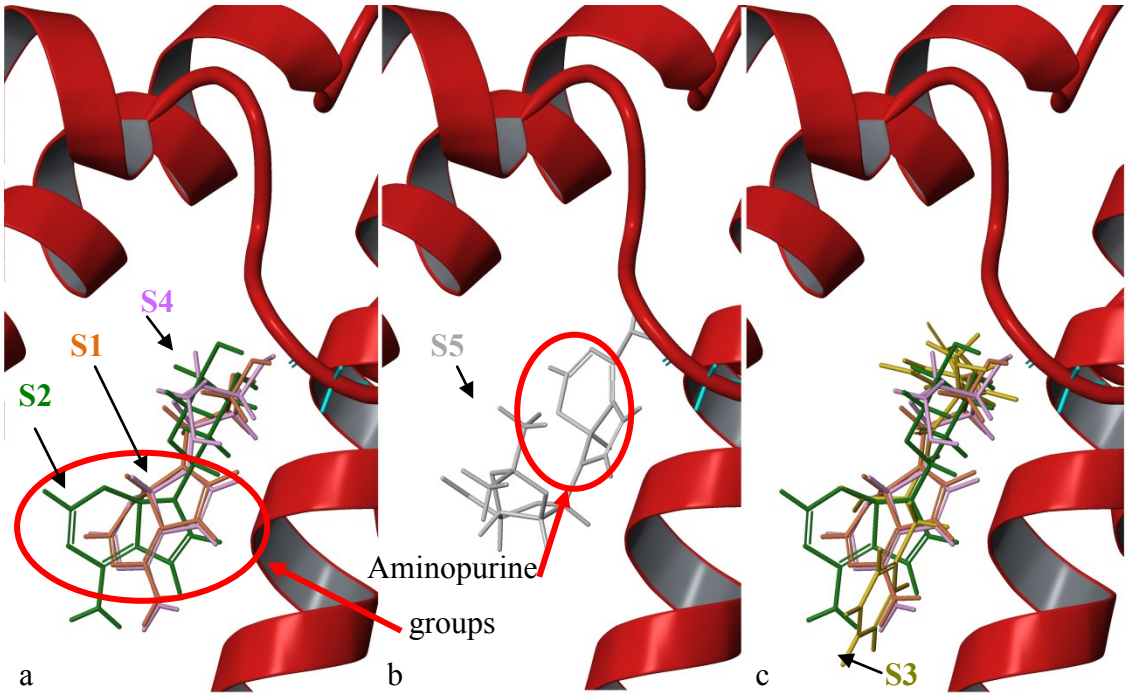


Figure 4.46. (a) S1, S2 and S4 (b) S5 (c) S1, S2, S3 and S4 in the SopE pocket.

Vendors and IUPAC names of the proposed molecules are given in Table 4.29. One of the proposed molecules, S1 or ZINC00370772, was found to participate in the therapy of viral diseases. This molecule was denoted as a novel nucleoside analog and reported to inhibit the replication *in vitro* of several DNA and RNA viruses, including vaccinia, herpes simplex (types 1 and 2), measles, and vesicular stomatitis [107].

Table 4.29. Vendors and names of proposed SopE inhibitors.

Name	Vendor	IUPAC Name
S1	Labotest	3-(6-amino-9H-purin-9-yl)propane-1,2-diol
S2	NCI Plated 2007	(3S,4S,6S)-6-(6-aminopurin-9-yl)tetrahydropyran-3,4-diol
S3	Frontier Scientific Services	2-(2-{[(benzyloxy)carbonyl]amino}-3-hydroxypropanamido)-3-phenylpropanoic acid
S4	Vitas-M	5-(6-aminopurin-9-yl)-2-(hydroxymethyl)tetrahydrofuran-3-ol
S5	ChemBridge	1-(6-Amino-9H-purin-9-yl)-1-deoxy-2,3-O-(1-methylethylidene)-

4.3. Final Remarks

Although virtual screening with target YopE yielded molecules with favorable interactions, docking results were lower than SopE by 1 kcal/mol, approximately. The reason for this outcome was investigated with binding site analysis of the target proteins. For this purpose, SiteMap module in Maestro graphical interface was used, which carries out binding site identification and gives estimates about the target's druggability [108]. SiteMap basically searches regions near the protein surface that could be a potential binding site for ligands. SiteMap locates these regions by generating hydrophobic and hydrophilic contour maps of the protein and calculating energy potentials. SiteMap regions of SopE and YopE are represented in Figure 4.47 and Figure 4.48, respectively. In these figures, green grids denote hydrophilic sites whereas yellow grids denote hydrophobic sites, blue grids denote hydrogen bond donors and red grids denote hydrogen bond acceptors. Arg144 residue in YopE and GAGA loop residues in SopE are indicated with red circles.

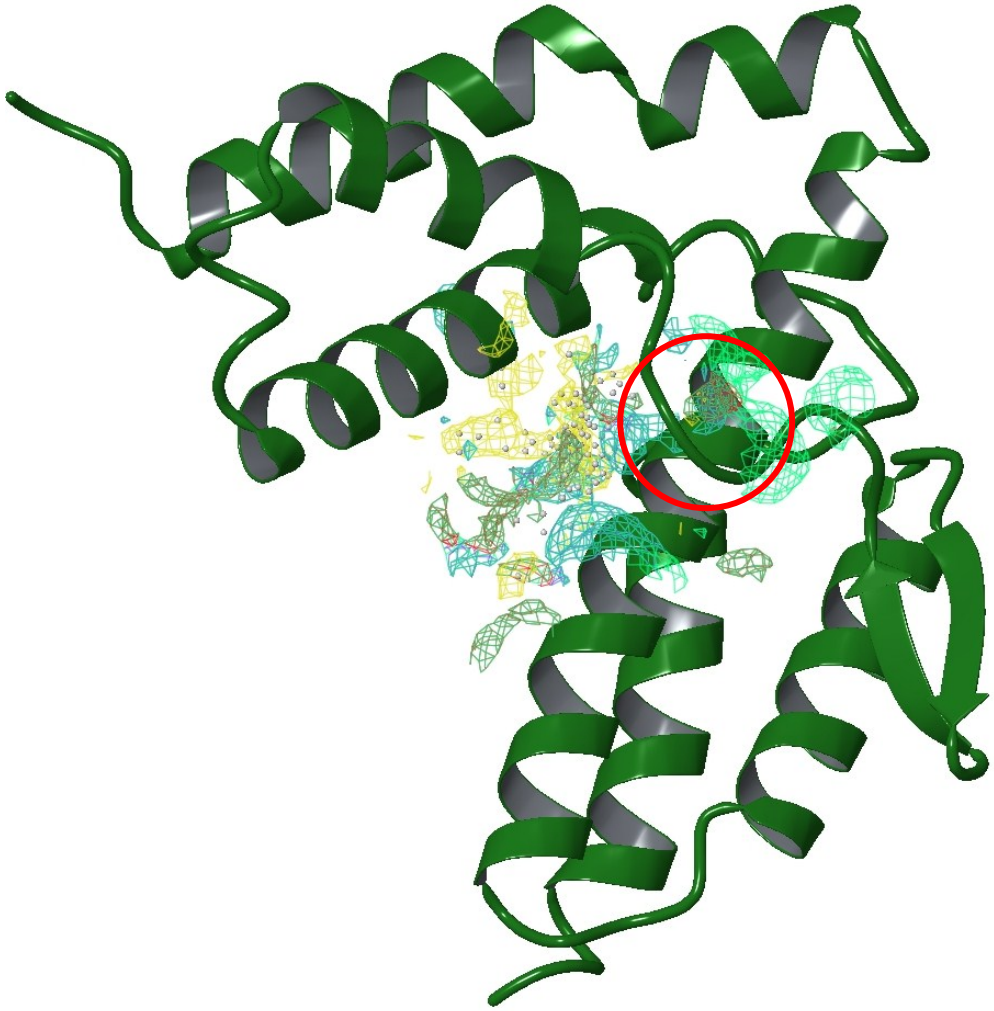


Figure 4.47. Binding site properties of SopE determined by SiteMap.

SiteMap determined the binding sites of both YopE and SopE correctly. However, in YopE, there are considerably fewer hydrophobic regions, compared to SopE. According to SiteMap, hydrophobicity plays an important role in druggability because undruggable and difficult sites typically are much less hydrophobic than druggable sites [109]. In this context, SiteMap classified YopE as a less druggable target than SopE. Therefore, docking results could have been influenced by binding characteristics of the target proteins. However, it should be noted that inhibitors that inhibit YopE with IC50 values around 20 μM have been reported in the literature [46].

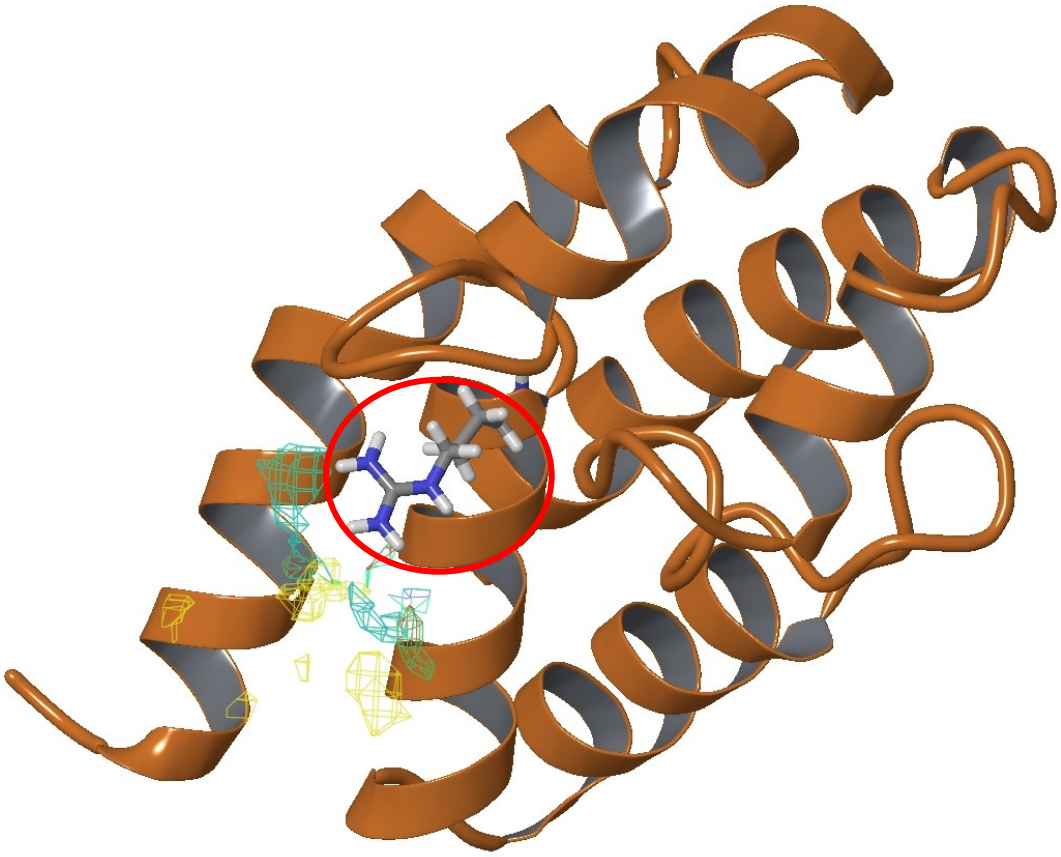


Figure 4.48. Binding site properties of YopE determined by SiteMap.

Surface exposure is another important issue that could influence docking results. SiteMap also defines solvent exposed and shallow binding sites as difficult druggable targets [109]. Close-up molecular surface views of the binding sites are represented in Figure 4.49 and Figure 4.50for SopE and YopE, respectively. YopE binding surface is indeed shallow, without buried surfaces or deep cavities. SopE, by comparison, has a more buried and enclosed binding site surface with less exposure.

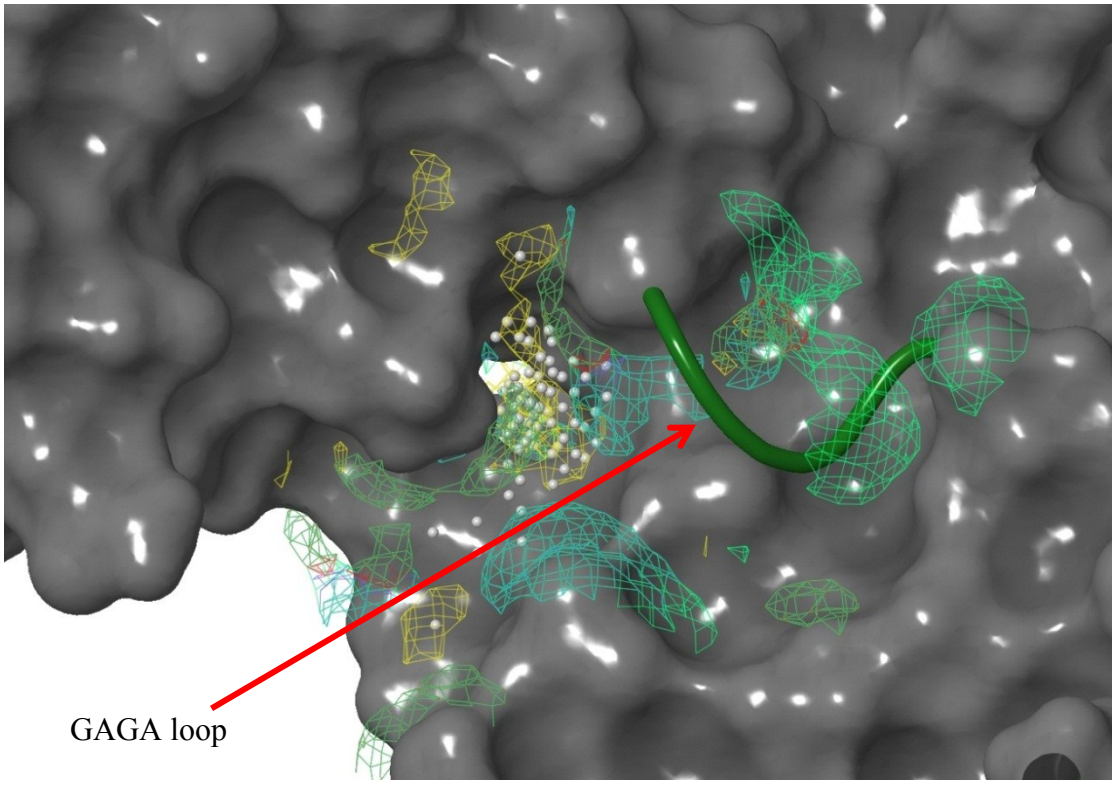


Figure 4.49. Molecular surface view of the binding site of SopE.

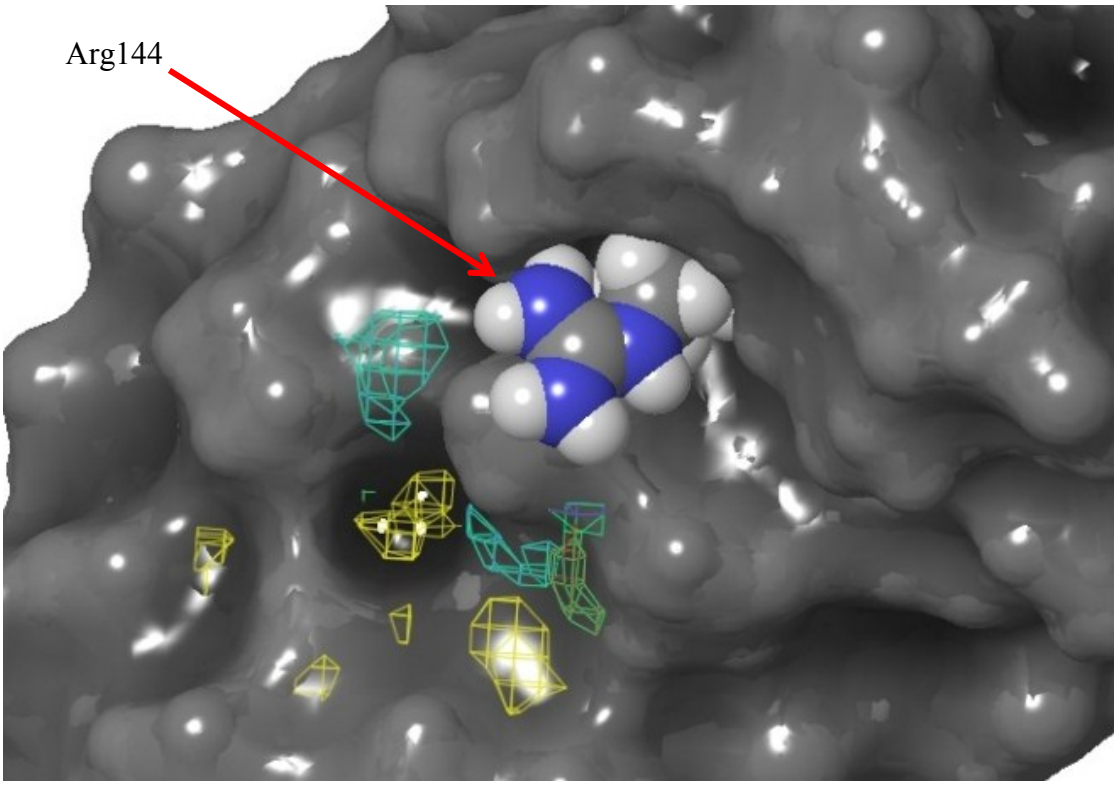


Figure 4.50. Molecular surface view of the binding site of YopE.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

In this study, the aim was to discover small molecules with the inhibition potential against YopE and SopE, both of which are outer *e*ffector proteins of gram-negative bacteria *Yersinia* and *Salmonella*. Inhibition of these proteins was aimed since these proteins secrete their bacterial cytotoxins into the cells and inhibit or activate small Rho GTPases by mimicking their regulator proteins. Proper functioning of GTPases is crucial since they control and regulate signaling events within the cell.

Potent inhibitors of YopE and SopE were investigated with computer-aided drug discovery tools. More precisely, ligand-based and structure-based virtual screening approaches were used employing Schrödinger Suite 2011 modules. For ligand-based approach regarding YopE, pharmacophore hypotheses were built via Phase, utilizing the known YopE inhibitor structures [46]. Since activities of known inhibitors were available, 3D-QSAR method was also incorporated to pharmacophore building step. For pharmacophore building approach regarding SopE, pharmacophore hypotheses were built using fragment library molecules. Small molecule database was generated from ZINC big vendor catalog molecules. Druglikeness and bioavailability issues are also considered in database generation step via Phase, such as applying Lipinski rule or eliminating molecules with reactive functional groups. With selected pharmacophore hypotheses AADDR.49 for YopE and AADN for SopE, database molecules were screened and pre-filtered, according to their 3D similarity to hypotheses.

Structure-based virtual screening approach was proceeded with filtered database compounds. 3D structures of receptor proteins YopE and SopE were obtained and prepared for molecular docking. Binding sites and important residues on the receptors were determined. Filtered database molecules were docked to receptors in two successive docking modes, allowing ligand flexibility. Binding poses of input molecules were determined and scored with empirical GlideScore. First, standard precision Glide docking

mode was performed and top 10 percent of molecules according to the GlideScore were re-docked to receptors using extra precision Glide docking mode.

Top 15 molecules from YopE and top 20 molecules from SopE were further investigated. These molecules were subjected to a number of post-docking analyses, including strain energy calculation, assessment of druglikeness, binding free energy calculation. Final molecules were selected upon their docking score, strain energy difference and druglikeness. For selecting different conformations of a particular molecule, their Emodel values were also compared. In addition, to validate the docking protocol, enrichment study was also conducted using known inhibitors of YopE and Schrödinger decoy set. Enrichment calculations yielded a ROC value of 0.88, indicating that Glide docking protocol is successful.

Ultimately, for each target protein, five molecules with a possible inhibition potential were proposed. Proposed molecules for YopE inhibition were denoted as Y1, Y2, Y3, Y4 and Y5; whereas the ones for SopE inhibition were denoted as S1, S2, S3, S4 and S5. 2D structures, IUPAC names and supplier names of the proposed molecules for YopE and SopE can be seen from Table 4.16, Table 4.18, Table 4.27 and Table 4.29, respectively. Proposed YopE inhibitors were observed to be structurally diverse, and therefore can lead to identification novel classes of compounds with activity against YopE, other than initial 23 YopE inhibitors which belong to a class of acetylated hydrazones of salicylaldehydes. On the other hand, four out of five proposed SopE inhibitors have aminopurine ring, which can be important for SopE inhibition. One of the proposed molecules, S1, was found to participate in the therapy of viral diseases as nucleoside analog. The other molecules were not linked to any pharmacological study. Proposed molecule interactions with receptor proteins YopE and SopE were investigated. All proposed molecules exhibited favorable interactions with YopE and SopE, which are summarized in Table 4.17 and Table 4.28, respectively. Overall, it was concluded that a successful virtual screening study targeting YopE and SopE was conveyed.

5.2. Recommendations for Future Studies

Although the study yielded satisfactory results, a number of improvements could be made, in terms of both method and result analysis. Recommendations will be given in three groups, regarding ligand-based approach, structure-based approach and result analysis.

In ligand-based screening, only one pharmacophore hypothesis was selected towards each target protein in this study. More than one hypothesis can be selected for database pre-filtering. By processing parallel virtual screening workflows for each hypothesis, the docking results can be compared or merged, which can increase the possibility of finding high affinity molecules. Additionally, for YopE, structure-based pharmacophore building approach with E-pharmacophores script can be used. For this purpose, known 23 inhibitors can be docked to YopE, and the output of the docking results can yield energetically favorable pharmacophore sites on inhibitor structures. By applying both ordinary pharmacophore building and structure-based pharmacophore building with the same YopE inhibitors, the effect of incorporating receptor knowledge to pharmacophores can be investigated. Additionally, instead of E-pharm, molecular interaction fields (MIF) method can be studied and applied.

In structure-based screening, ligand flexibility was allowed, but receptor structures were kept frozen during docking. Instead, induced-fit docking, which allows receptor atoms to be flexible as well as ligands, can be employed. Induced-fit requires more computational screening time; therefore it can be used after Glide XP mode, with only a small number of molecules that have high docking score. Additionally, different constraints on receptor residues can be provided during docking.

Besides virtual screening approaches, molecular dynamics (MD) simulations can also be performed to improve result analysis. For example, MD simulations can be done on apo states and selected high scoring ligand-bound states of receptors. By comparing simulations of the different ligand bound states, implications can be made about protein-ligand complex stability.

In result analysis, after Glide XP docking, top molecules based on the docking score were further investigated. But instead, all docking results can be clustered based on their structural and pharmacodynamic properties, using Canvas module of Shrödinger [110]. After similarity clustering, compounds with the highest docking score can be selected from each cluster. Hence, a more structurally diverse set of proposed molecules can be obtained.

Finally, proposed molecules can be purchased from their commercial suppliers, and their inhibitory performances can be tested *in vitro* by preparing biological assays of the target proteins.

APPENDIX A: REACTIVE FUNCTIONAL GROUPS

The reactive functional groups determined by QikProp are presented in this section. Database molecules including these structural features were eliminated during 3D database preparation step via Phase. These functional groups are as follows:

- Acyl halide
- Hetero-halogen bond
- NAS substrate
- Alkyl halide
- Halogen alpha to W-group
- Heteroatom in 3-ring
- Activated cyclopropane
- Aluminum present (toxic)
- Silicon present (toxic)
- Hetero-hetero single bond
- Azo, diazo, or azide
- Acceptor carbonyl or derivative
- Anhydride or analog
- Unhindered ester
- Sulfonate or relative
- Phosphonate or relative
- Acetal or analog
- Carbonate
- Thiol—oxidation possible
- Carbonyl in 3-ring

APPENDIX B: GLIDE XP RESULTS OF YOPE

The ligand interaction diagrams of unconstrained docking of 23 compounds to YopE and GlideScores were given in below figures and Table B.1.

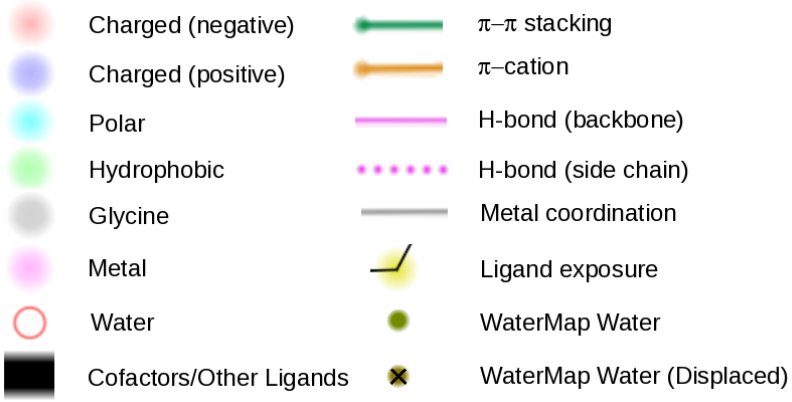


Figure B.1. Ligand interaction map legend.

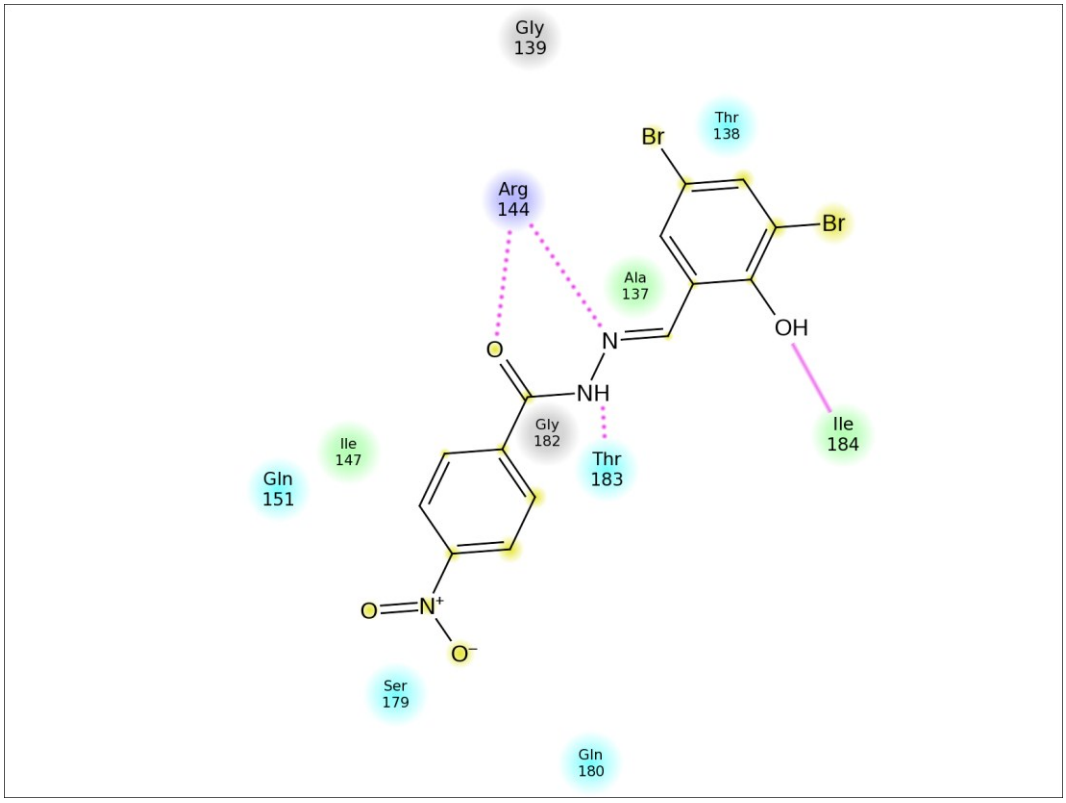


Figure B.2. Ligand interaction map of compound 1.

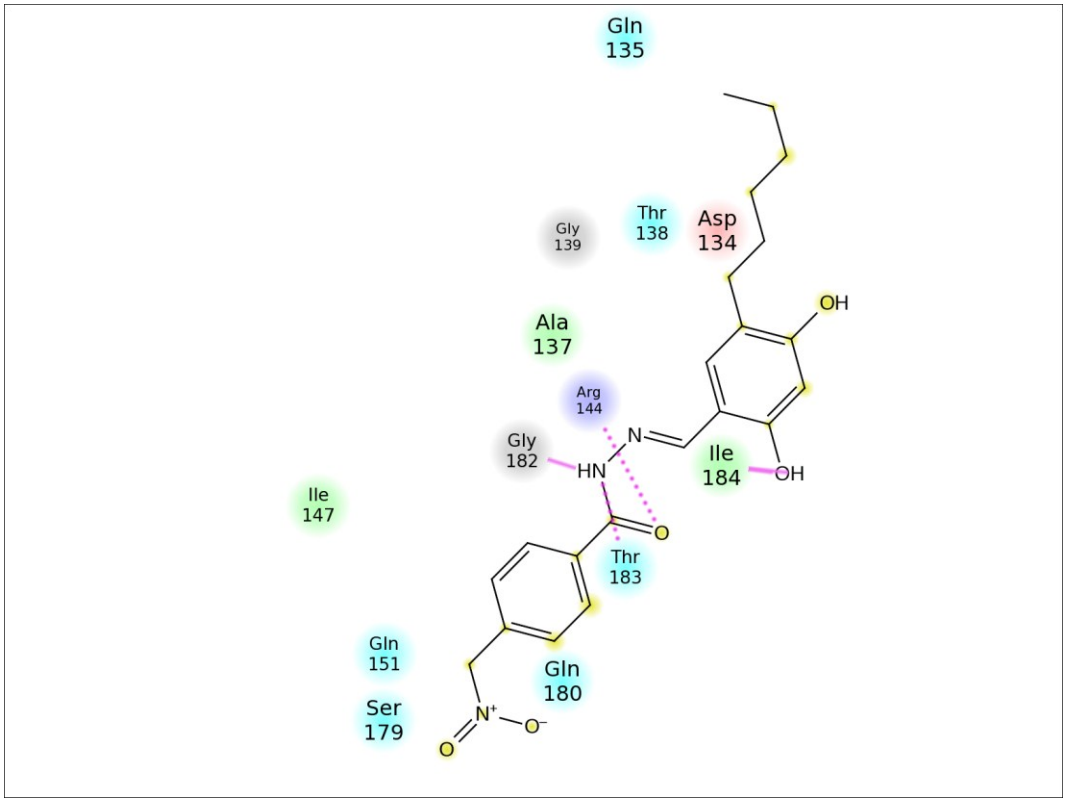


Figure B.3. Ligand interaction map of compound 2.

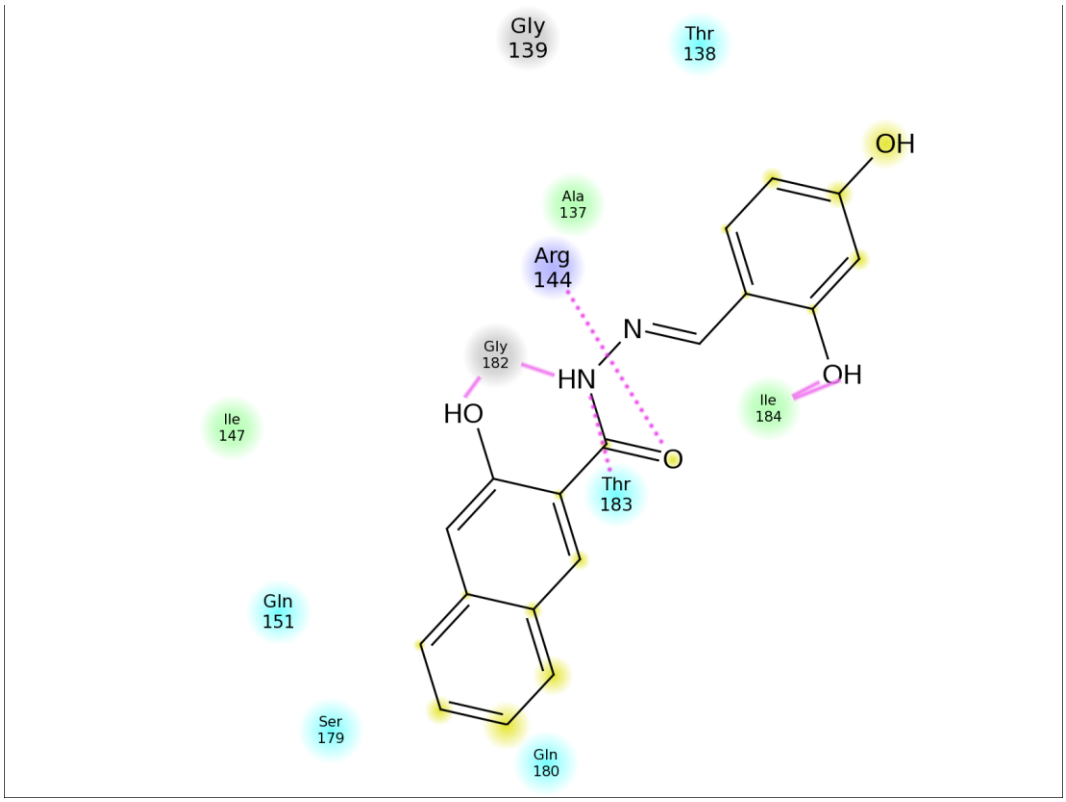


Figure B.4. Ligand interaction map of compound 3.

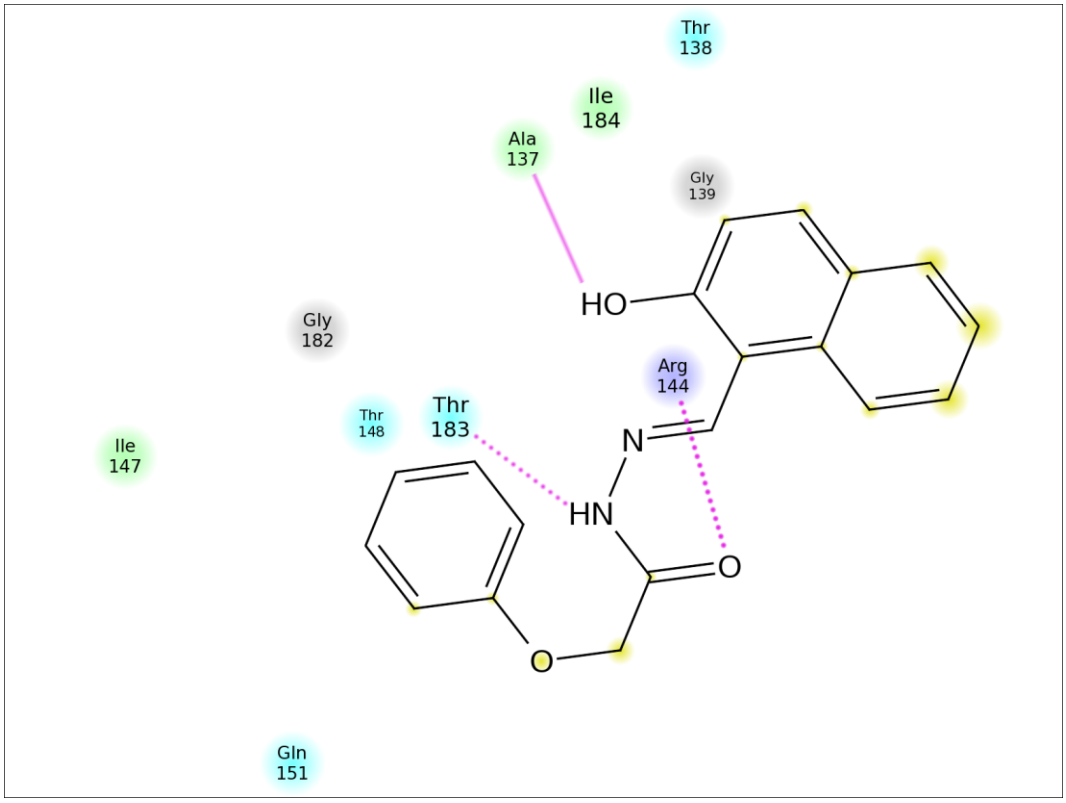


Figure B.5. Ligand interaction map of compound 4.

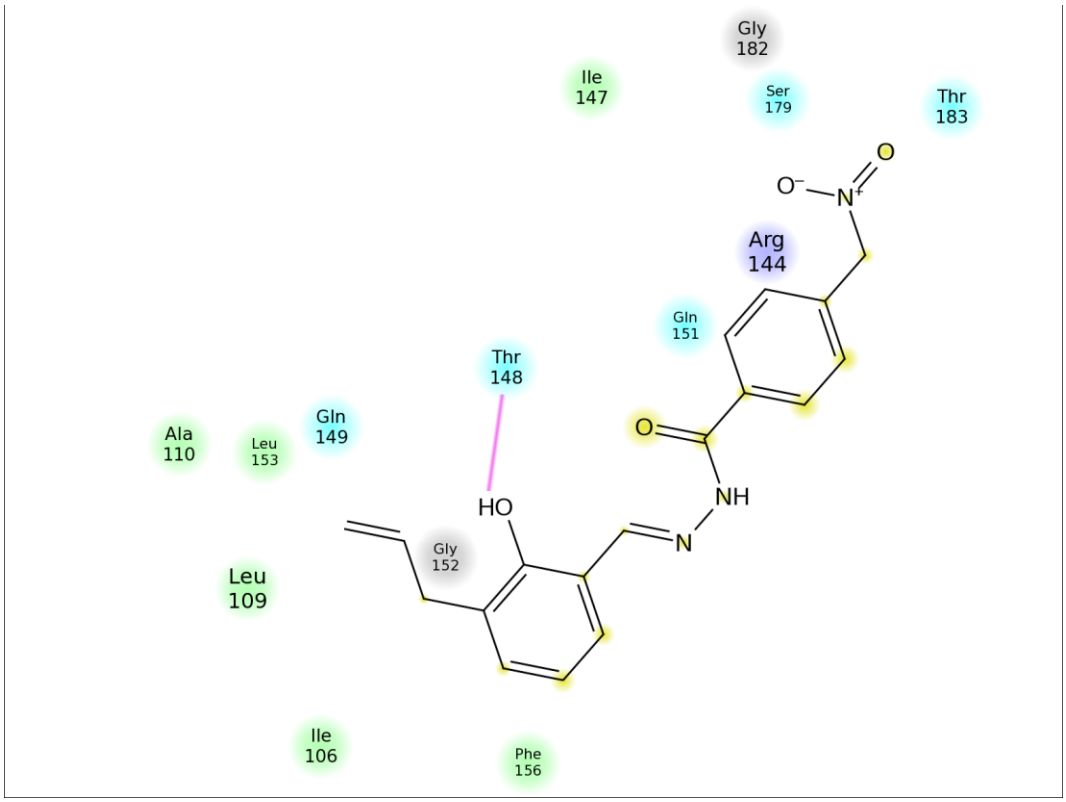


Figure B.6. Ligand interaction map of compound 5.

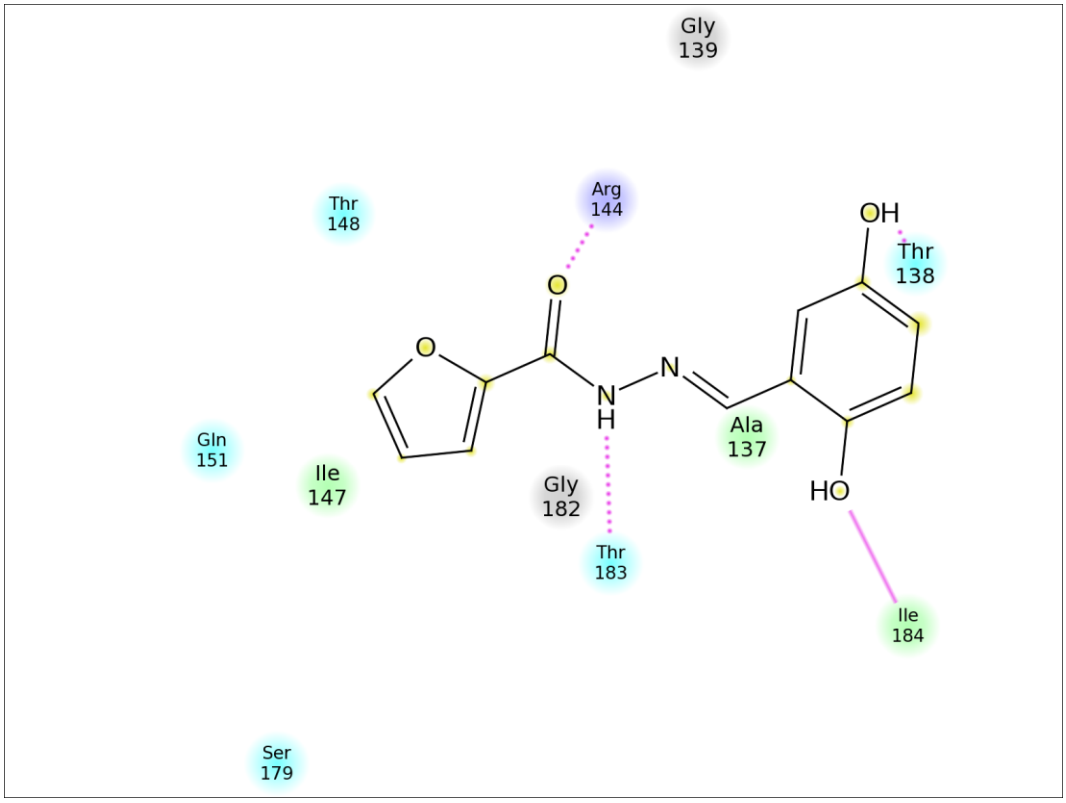


Figure B.7. Ligand interaction map of compound 6.

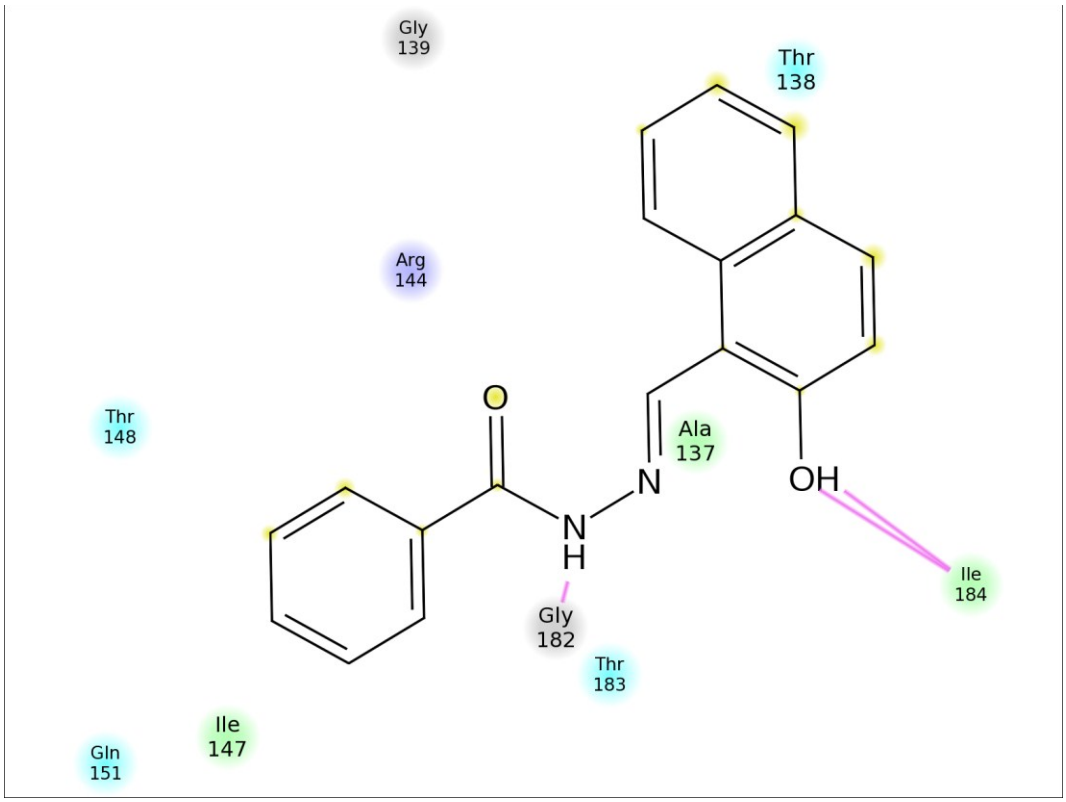


Figure B.8. Ligand interaction map of compound 7.

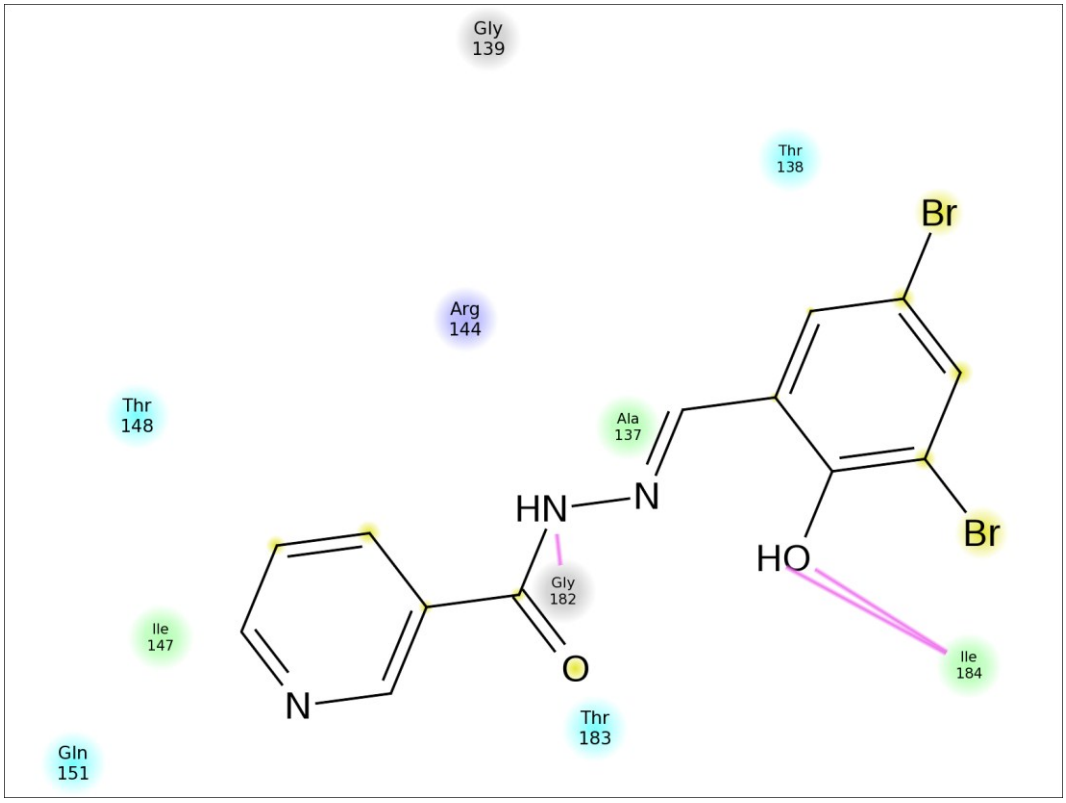


Figure B.9. Ligand interaction map of compound 8.

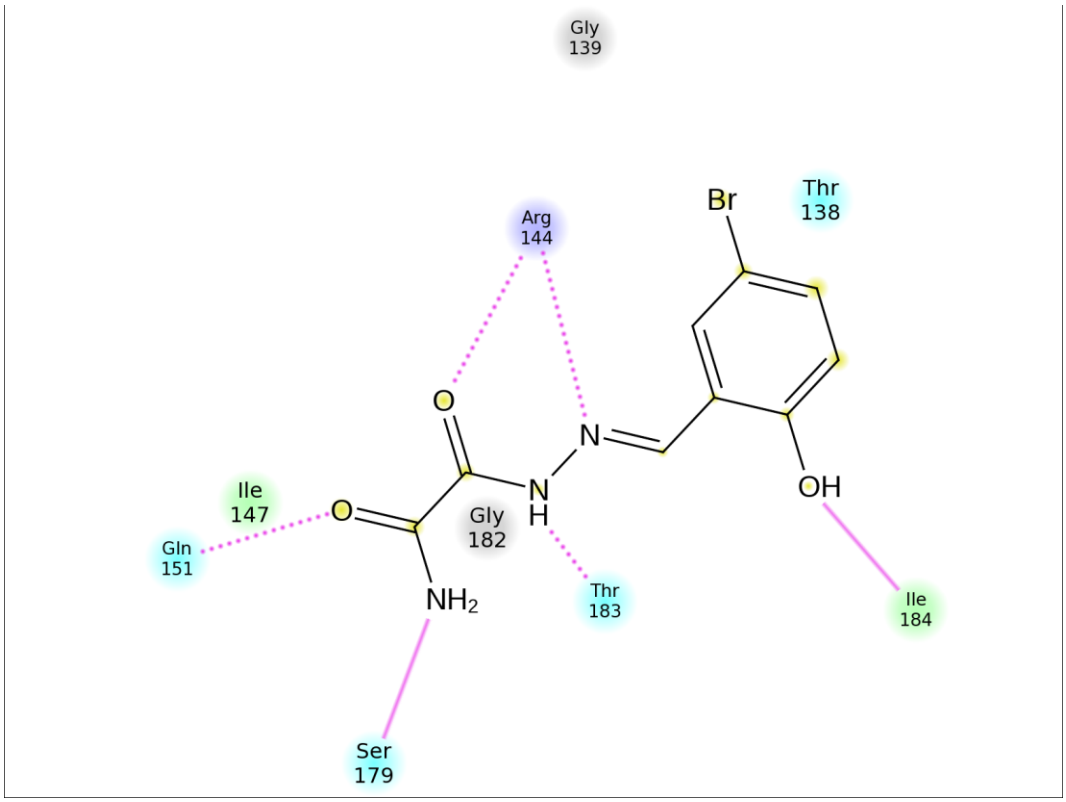


Figure B.10. Ligand interaction map of compound 9.

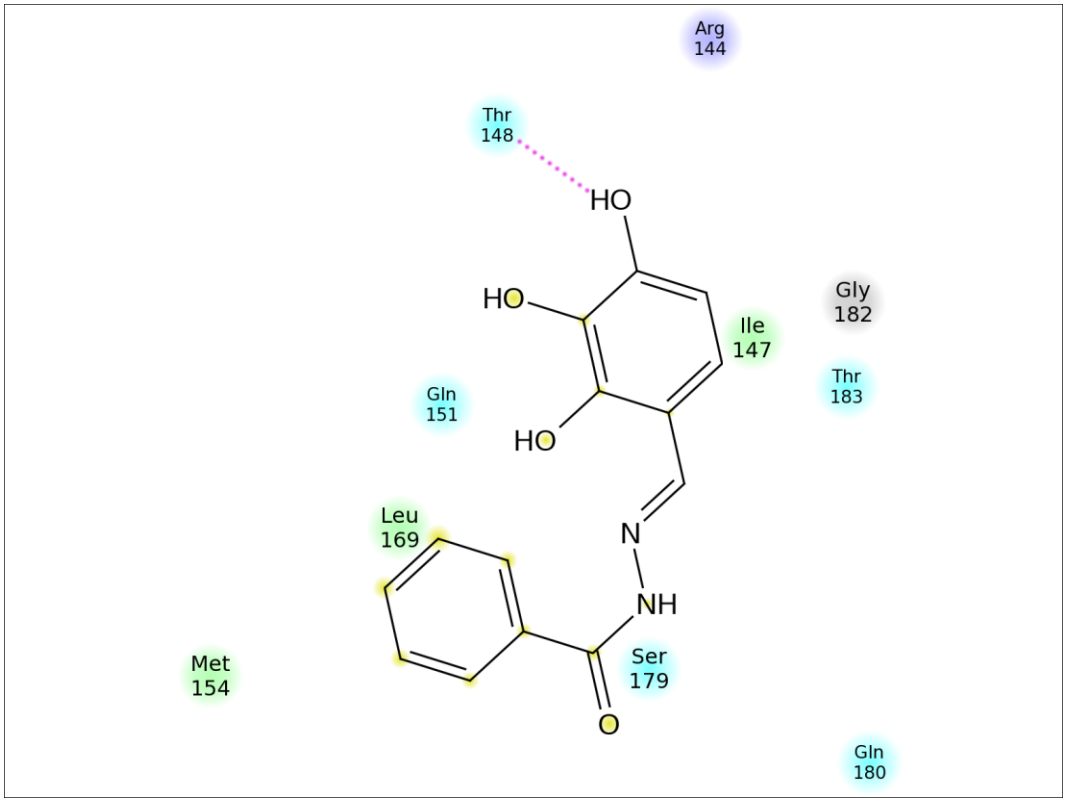


Figure B.11. Ligand interaction map of compound 10.

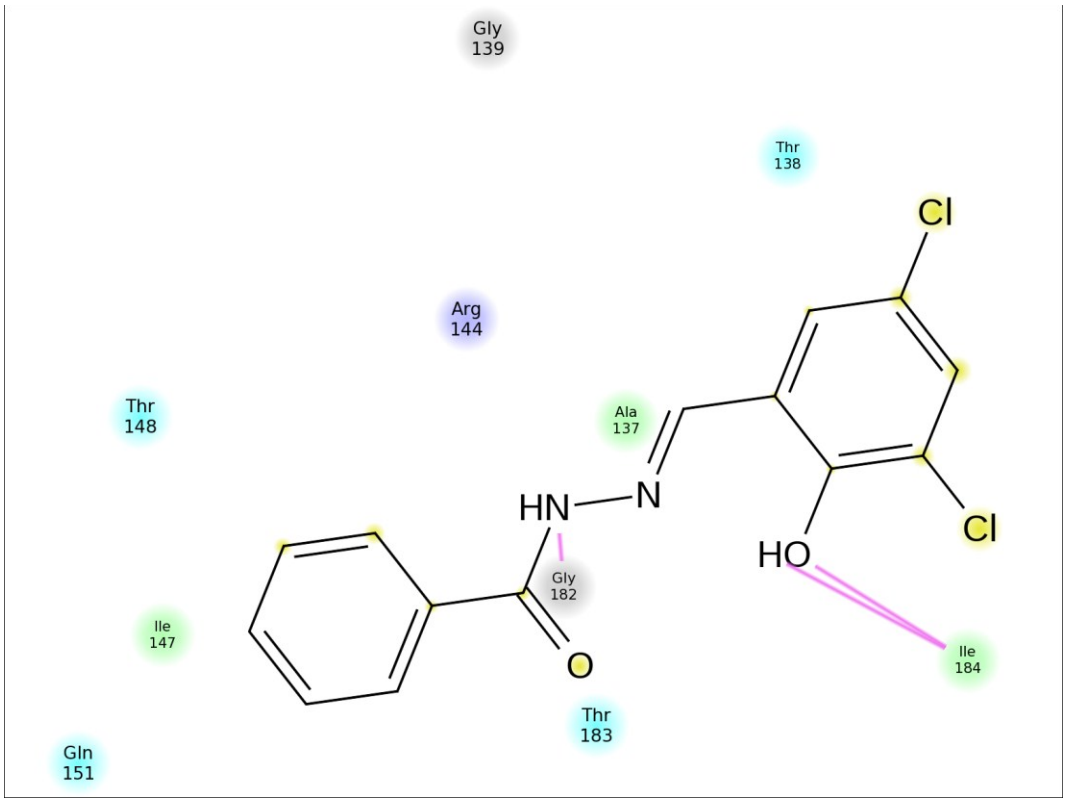


Figure B.12. Ligand interaction map of compound 11.

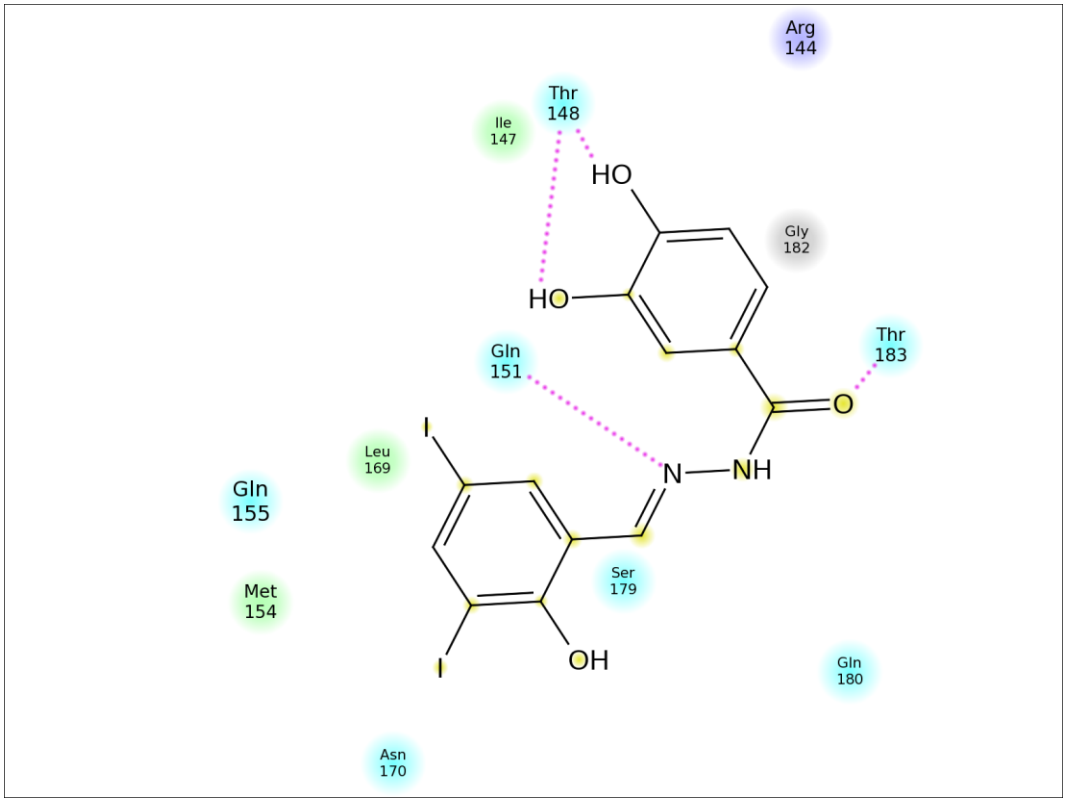


Figure B.13. Ligand interaction map of compound 12.

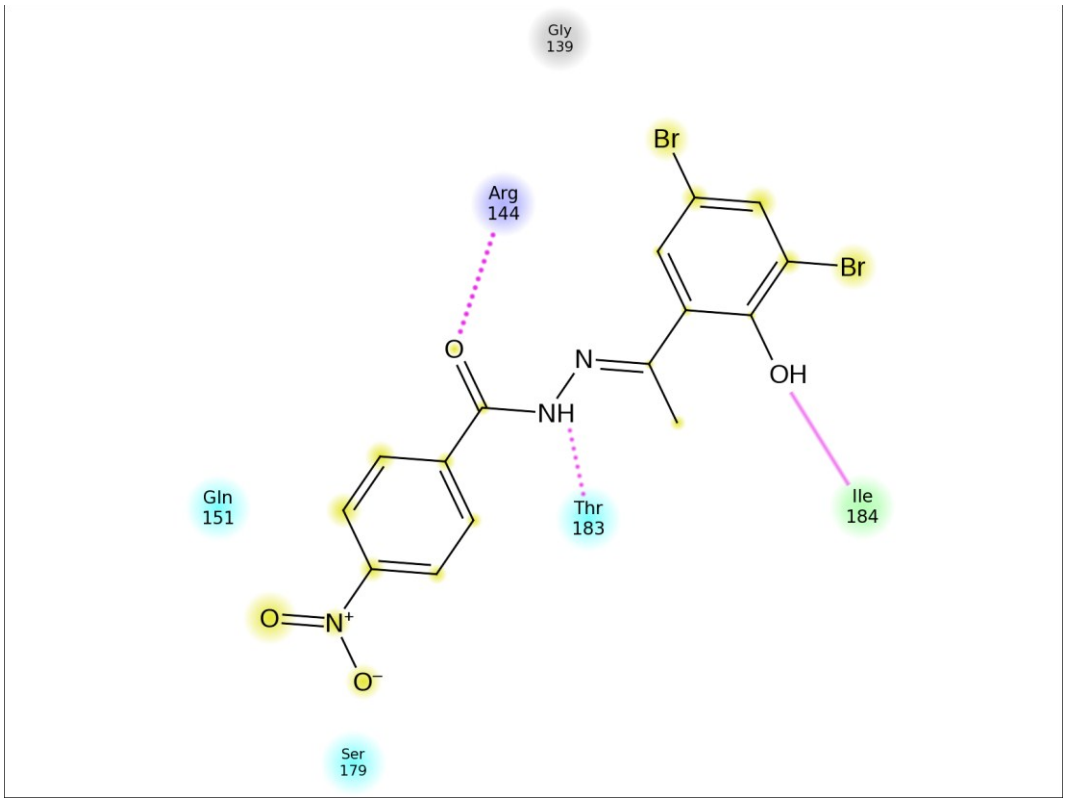


Figure B.14. Ligand interaction map of compound 13

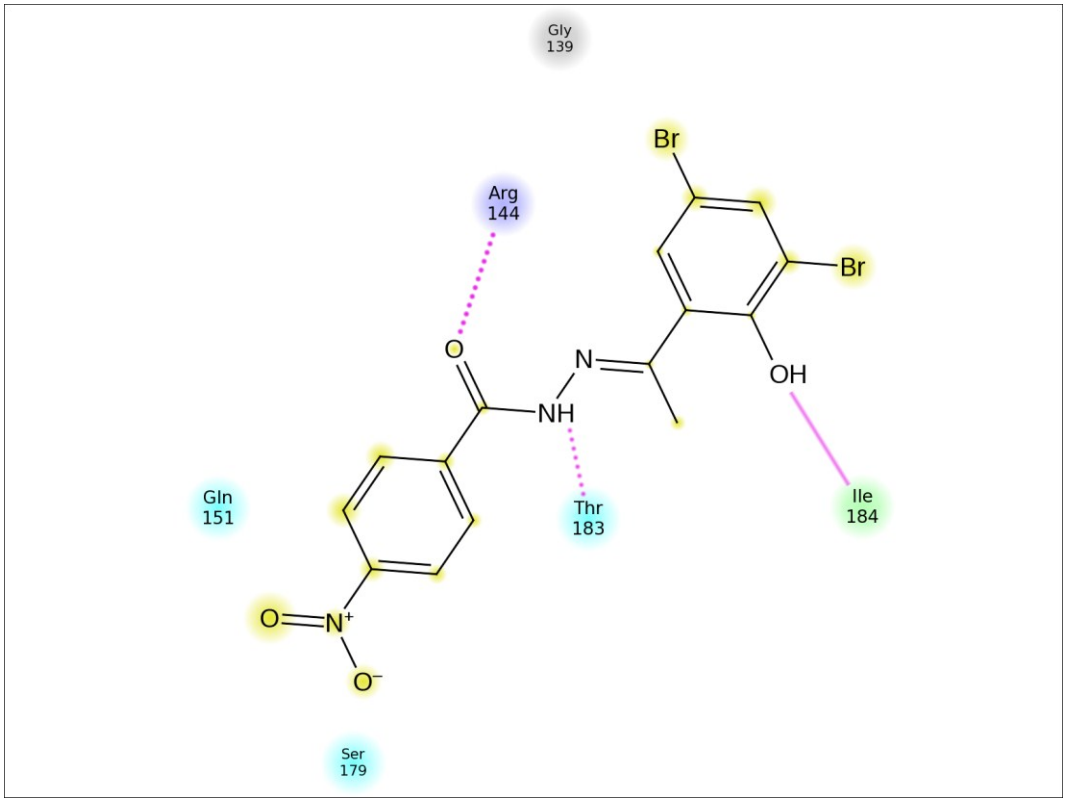


Figure B.15. Ligand interaction map of compound 14.

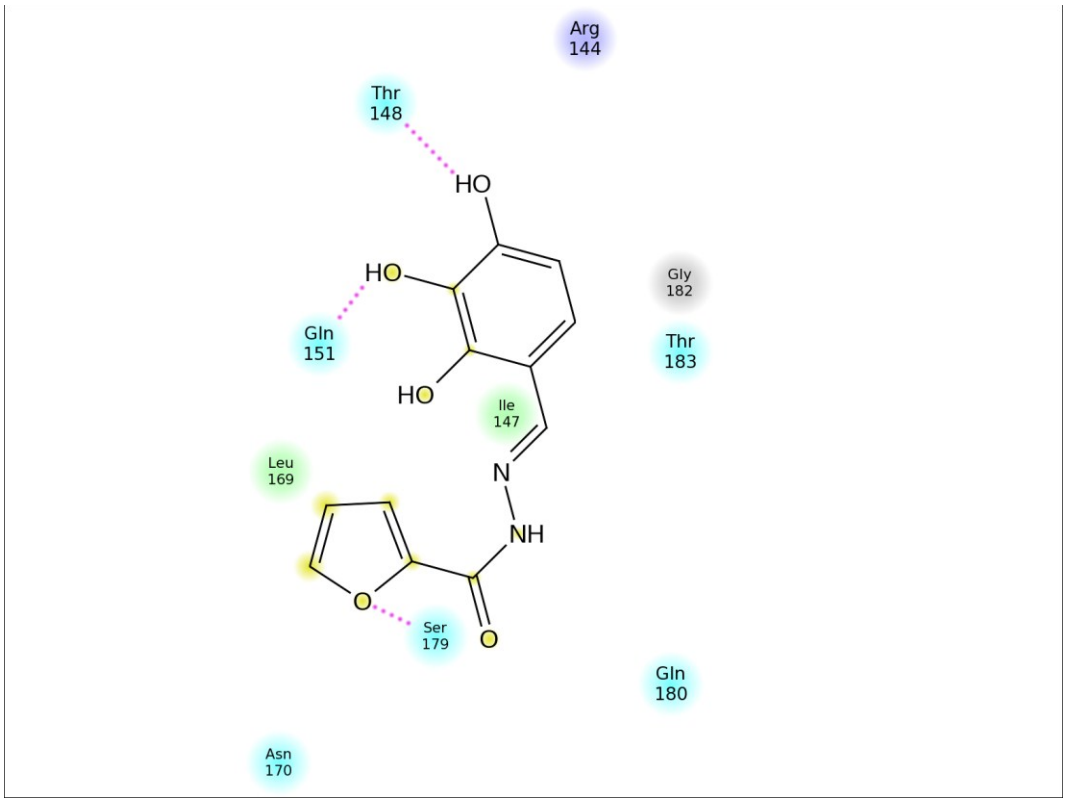


Figure B.16. Ligand interaction map of compound 15.

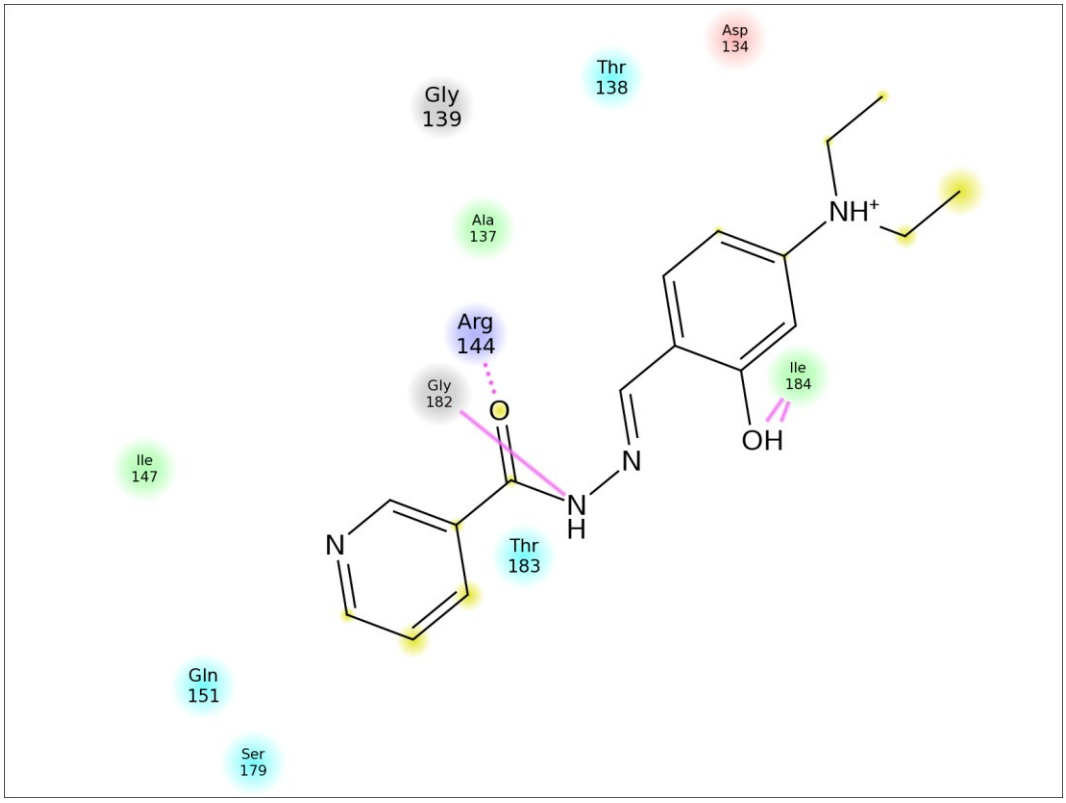


Figure B.17. Ligand interaction map of compound 16.

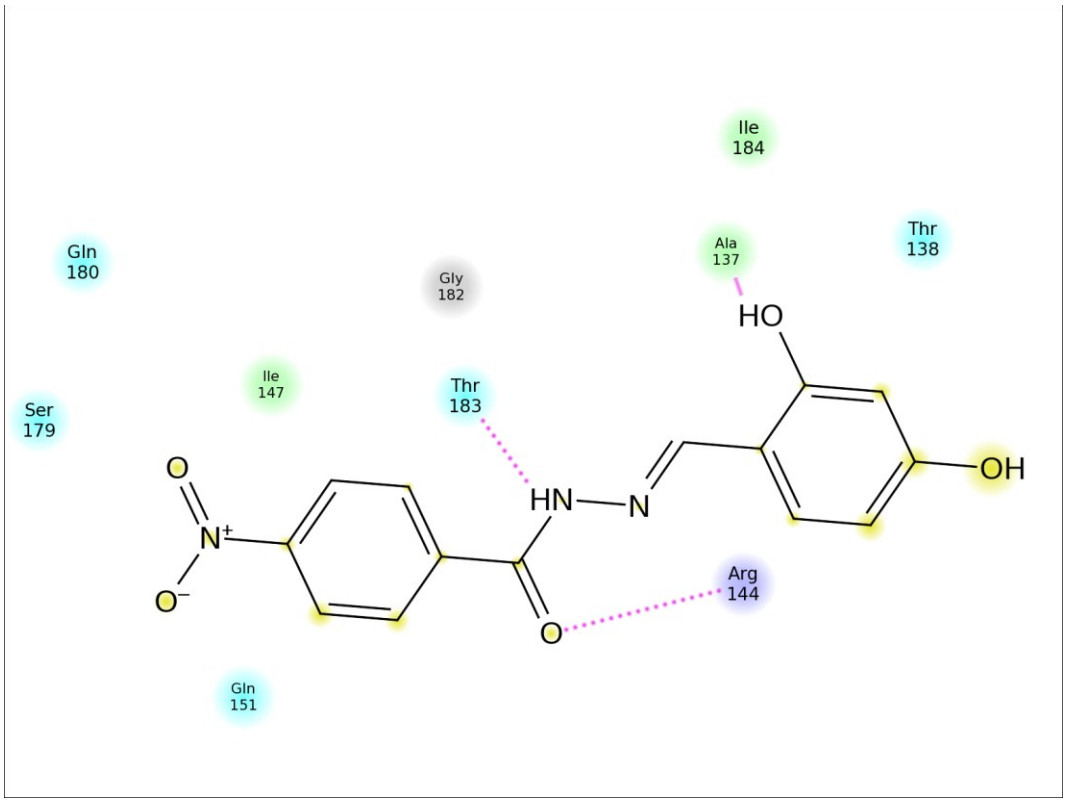


Figure B.18. Ligand interaction map of compound 17.

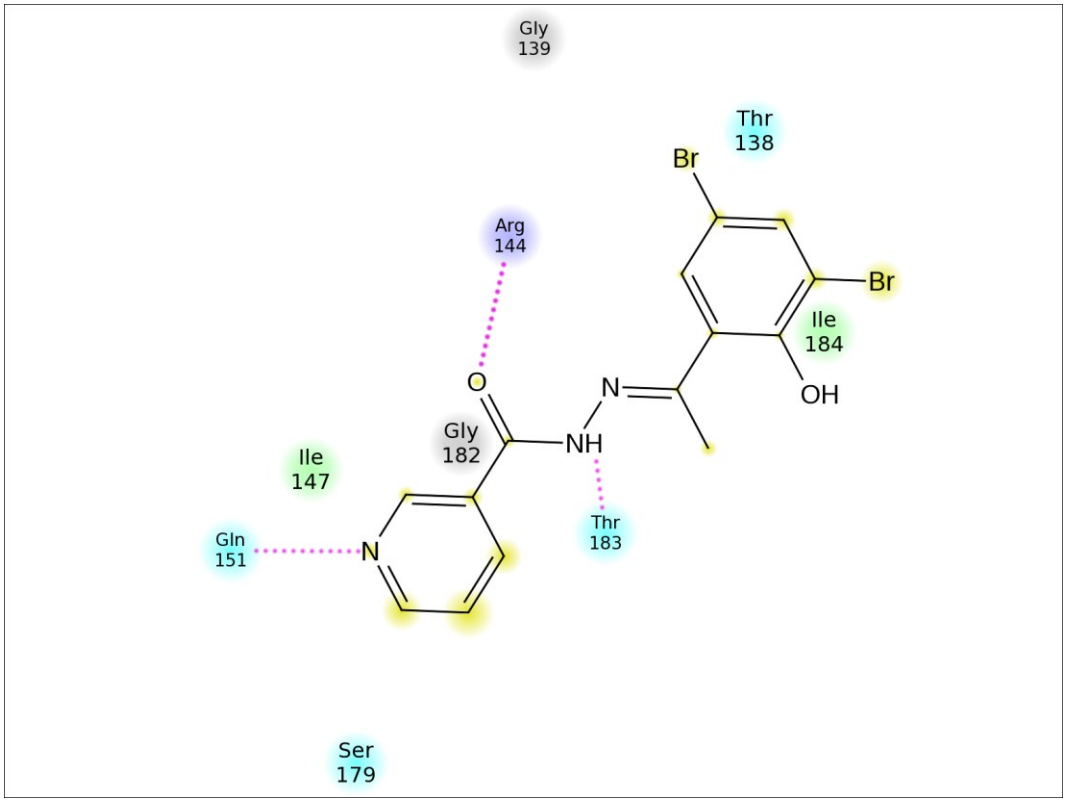


Figure B.19. Ligand interaction map of compound 18.

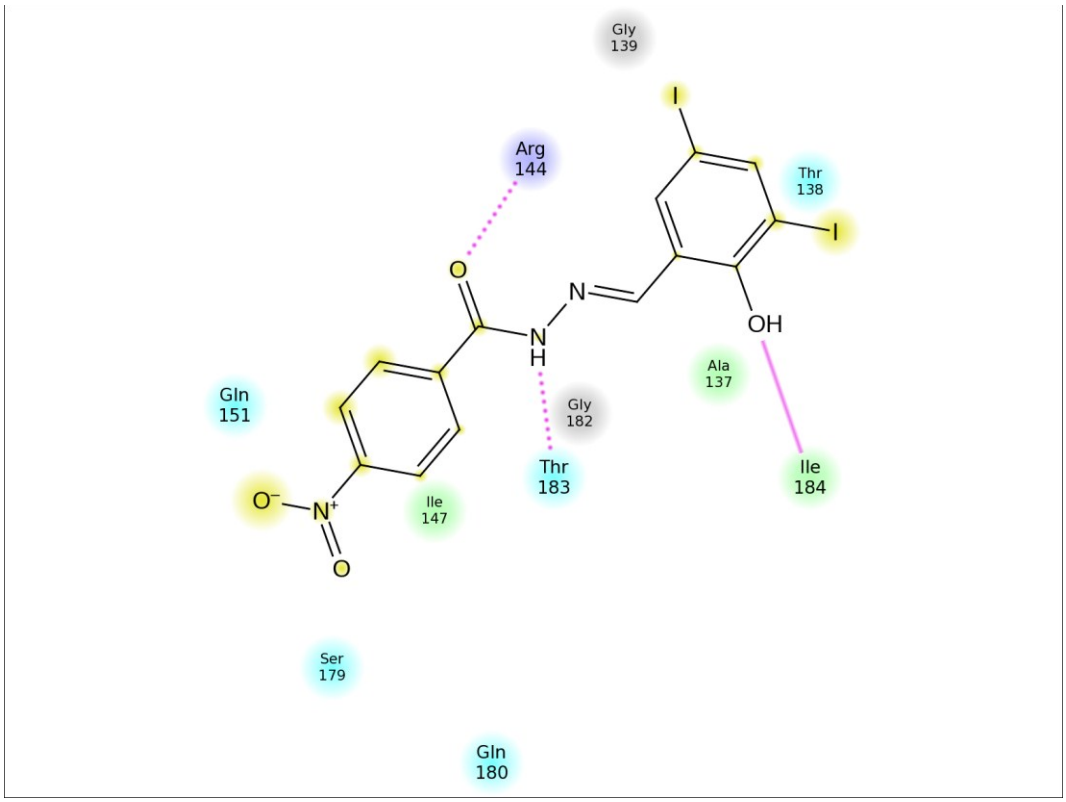


Figure B.20. Ligand interaction map of compound 19.

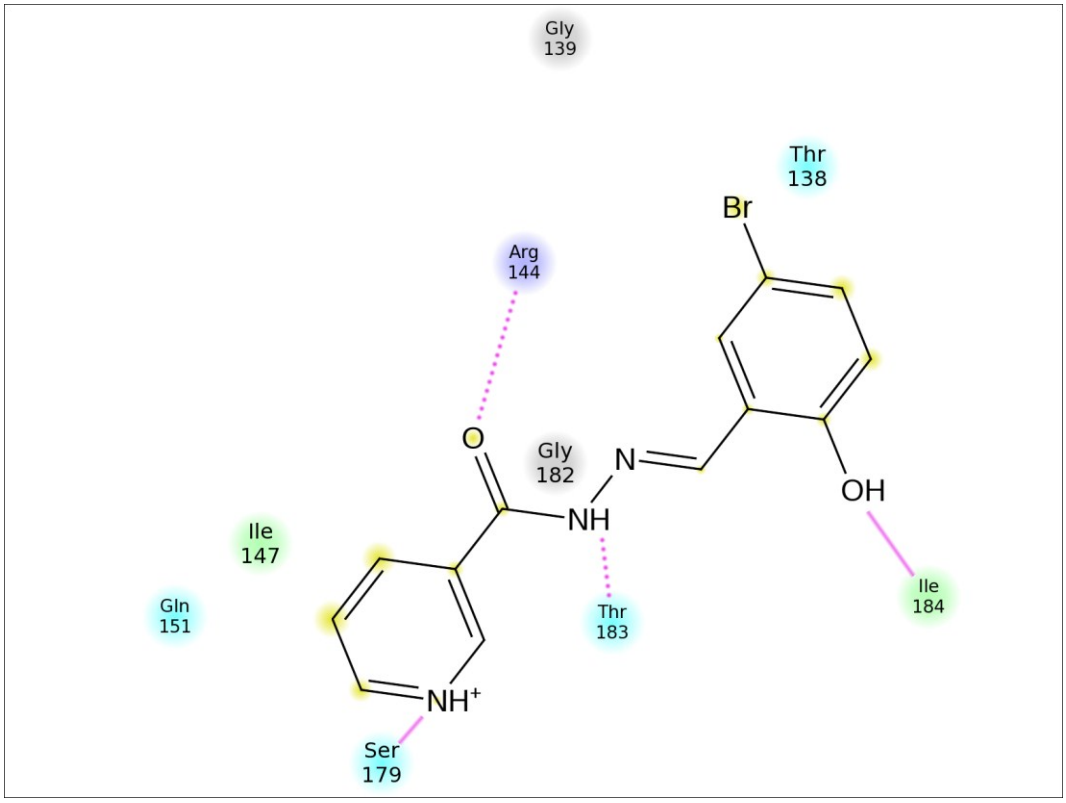


Figure B.21. Ligand interaction map of compound 20.

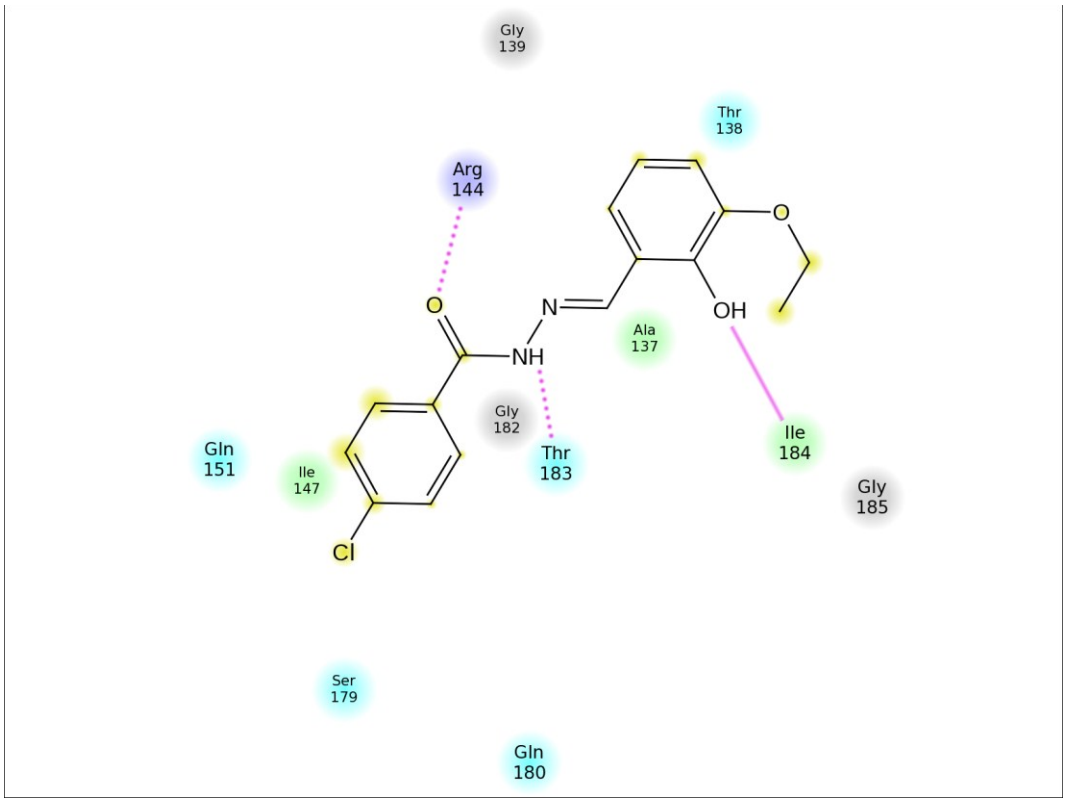


Figure B.22. Ligand interaction map of compound 21.

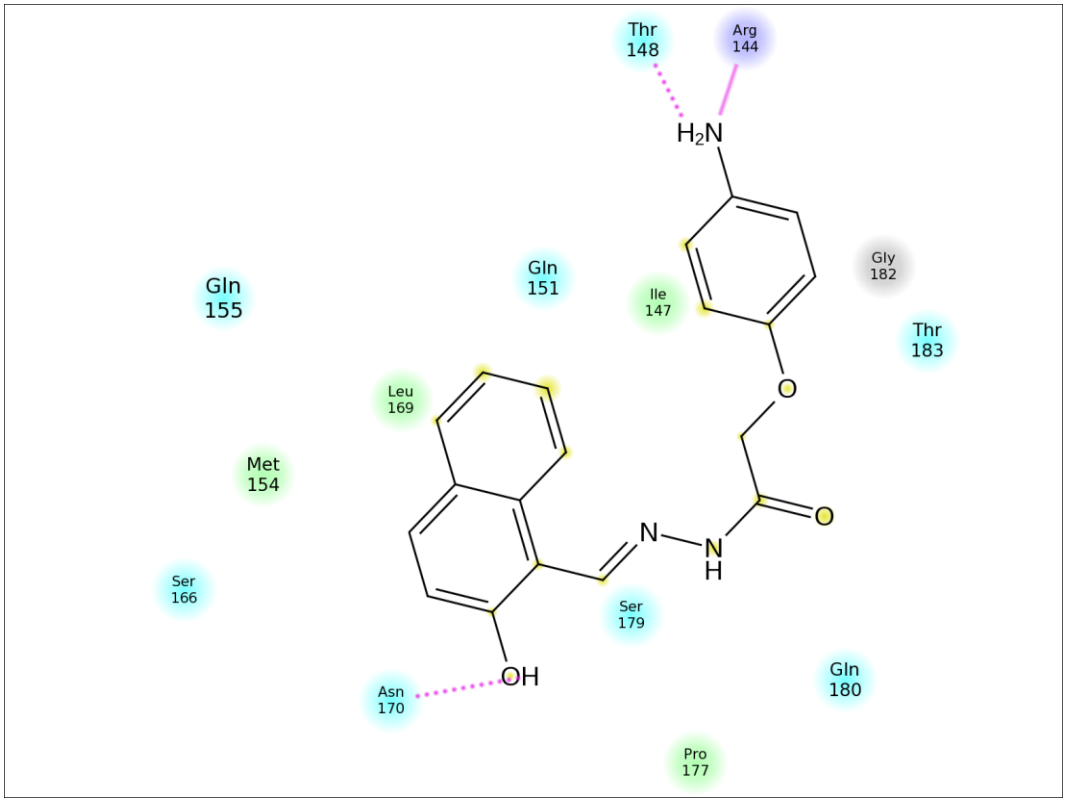


Figure B.23. Ligand interaction map of compound 22.

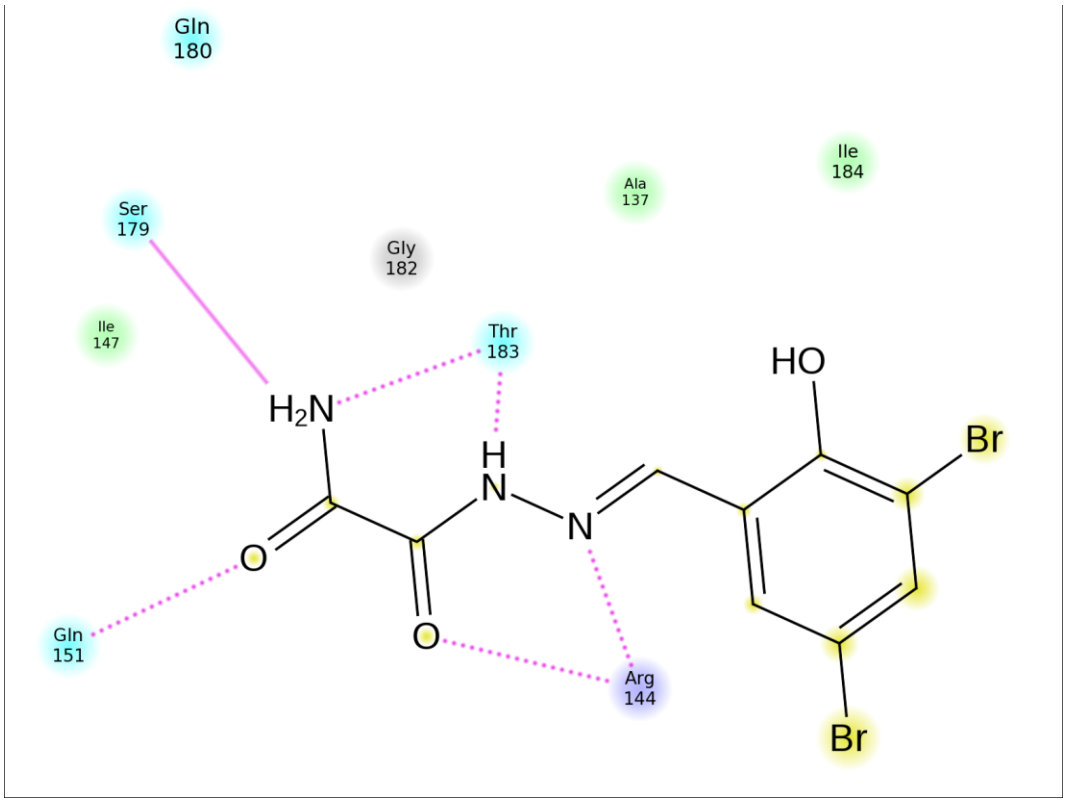


Figure B.24. Ligand interaction map of compound 23.

Table B.1. Glide XP scores of known inhibitors against YopE without constraint.

Compound	GlideScore (kcal/mol)	Compound	GlideScore (kcal/mol)
1	-3.742	13	-3.899
2	-4.229	14	-4.234
3	-5.655	15	-4.426
4	-4.613	16	-5.096
5	-3.680	17	-5.804
6	-5.454	18	-4.828
7	-4.842	19	-3.670
8	-4.787	20	-5.172
9	-5.512	21	-4.432
10	-5.630	22	-5.343
11	-4.692	23	-5.989
12	-6.136	avg	-4.910

APPENDIX C: PHASE HYPOTHESES SCORES OF YOPE

Table C.1. Scores of all 5-point hypotheses generated from YopE inhibitors.

ID	Survival	Site	Vector	Volume	Selectivity
AAADR.169	3.670	0.89	0.977	0.806	1.147
AAADR.71	3.664	0.89	0.977	0.802	1.120
AADDR.41	3.661	0.89	0.984	0.792	1.223
AADDR.49	3.647	0.87	0.982	0.791	1.193
AAADR.79	3.221	0.64	0.947	0.638	1.139
AADDR.37	3.189	0.61	0.947	0.631	1.209
AADDR.161	3.087	0.57	0.819	0.696	1.232
AADDR.43	3.002	0.67	0.882	0.453	1.235
AAADR.81	2.995	0.67	0.877	0.452	1.157
AADDR.48	2.953	0.44	0.889	0.621	1.200
AADDR.40	2.934	0.62	0.875	0.442	1.231
AAADR.77	2.932	0.62	0.872	0.44	1.155
AAADR.20	2.916	0.58	0.872	0.468	1.124
AAADR.19	2.915	0.58	0.87	0.467	1.125
AADDR.46	2.890	0.55	0.869	0.468	1.188
AADDR.42	2.861	0.53	0.948	0.387	1.223
AAADR.170	2.859	0.53	0.944	0.389	1.149
AAADR.74	2.795	0.45	0.929	0.421	1.126
AAADR.80	2.793	0.46	0.928	0.407	1.138
AADDR.39	2.79	0.46	0.916	0.411	1.207
AADDR.53	2.786	0.45	0.953	0.384	1.222
AAADR.75	2.665	0.52	0.662	0.480	1.177
AADDR.52	2.602	0.41	0.821	0.375	1.206

APPENDIX D: ENRICHMENT REPORT OF YOPE

Enrichment Report

Total actives: 20

Total ligands(actives+decoys): 1015

Number of ranked actives: 20

BEDROC(alpha=160.9, alpha*Ra=3.1704): 0.447

BEDROC(alpha=20.0, alpha*Ra=0.3941): 0.409

BEDROC(alpha=8.0, alpha*Ra=0.1576): 0.536

ROC: 0.88

RIE: 6.76

Area under accumulation curve: 0.87

Ave. Number of outranking decoys: 123

Minimum Tc over all active pairs: n/a

Count and percentage of actives in top N% of decoy results.

% Decoys	1%	2%	5%	10%	20%
# Actives	7	7	7	8	16
% Actives	35.0	35.0	35.0	40.0	80.0

Count and percentage of actives in top N% of results.

% Results	1%	2%	5%	10%	20%
# Actives	5	7	7	8	15
% Actives	25.0	35.0	35.0	40.0	75.0

Enrichment Factors with respect to N% sample size.

% Sample	1%	2%	5%	10%	20%
EF	25	18	7	4	3.8
EF*	35	17	7	4	4
EF'	44	32	14	7.7	5.4

DEF | n/a| n/a| n/a| n/a| n/a|
DEF* | n/a| n/a| n/a| n/a| n/a|
DEF' | n/a| n/a| n/a| n/a| n/a|
Eff | 0.944| 0.892| 0.75| 0.6| 0.6|

Enrichment Factors with respect to N% actives recovered.

% Actives	40%	50%	60%	70%	80%	90%	100%
EF	5.5	4.2	3.8	4.1	3.8	3	2.5
EF*	6	4.5	4	4.4	4.1	3.1	2.5
EF'	11	7.1	6.1	5.8	5.4	4.6	3.9
FOD	0.01	0.03	0.05	0.06	0.08	0.1	0.1

Rank	Title	Rank	Title
2	compound12	312	compound2
3	compound23	413	compound13
6	compound3		
7	compound10		
10	compound9		
13	compound15		
15	compound22		
74	compound6		
110	compound20		
120	compound16		
123	compound17		
161	compound7		
167	compound18		
174	compound8		
193	compound11		
212	compound4		
248	compound21		
308	compound14		

REFERENCES

1. Brändén, C.-I. and J. Tooze, *Introduction to Protein Structure*, 2nd ed. New York, Garland Publications, 1999.

2. Rossman, K. L., C. J. Der, and J. Sondek, "Gef Means Go: Turning on Rho Gtpases with Guanine Nucleotide-Exchange Factors", *Nature Reviews Molecular Cell Biology,* Vol. 6, pp. 167-80, 2005.

3. Van Aelst, L. and C. D'Souza-Schorey, "Rho Gtpases and Signaling Networks", *Genes & Development,* Vol. 11, pp. 2295-322, 1997.

4. Raftopoulou, M. and A. Hall, "Cell Migration: Rho Gtpases Lead the Way", *Developmental Biology,* Vol. 265, pp. 23-32, 2004.

5. Li, J., K. L. O'Connor, M. R. Hellmich, G. H. Greeley, Jr., C. M. Townsend, Jr., and B. M. Evers, "The Role of Protein Kinase D in Neurotensin Secretion Mediated by Protein Kinase C-Alpha/-Delta and Rho/Rho Kinase", *Journal of Biological Chemistry,* Vol. 279, pp. 28466-74, 2004.

6. Etienne-Manneville, S. and A. Hall, "Rho Gtpases in Cell Biology", *Nature,* Vol. 420, pp. 629-35, 2002.

7. Chimini, G. and P. Chavrier, "Function of Rho Family Proteins in Actin Dynamics During Phagocytosis and Engulfment", *Nature Cell Biology,* Vol. 2, pp. E191-6, 2000.

8. Schmidt, A. and A. Hall, "Guanine Nucleotide Exchange Factors for Rho Gtpases: Turning on the Switch", *Genes & Development,* Vol. 16, pp. 1587-609, 2002.

9. Finlay, B. B., "Bacterial Virulence Strategies That Utilize Rho Gtpases", *Current Topics in Medicinal Chemistry,* Vol. 291, pp. 1-10, 2005.

10. Barczak, A. K. and D. T. Hung, "Productive Steps toward an Antimicrobial Targeting Virulence", *Current Opinion in Microbiology,* Vol. 12, pp. 490-6, 2009.

11. Ryan, K. J. and J. C. Sherris, *Sherris Medical Microbiology*, 3rd ed. Norwalk, Conn., Appleton & Lange, 1994.

12. Galan, J. E. and A. Collmer, "Type Iii Secretion Machines: Bacterial Devices for Protein Delivery into Host Cells", *Science,* Vol. 284, pp. 1322-8, 1999.

13. Aktories, K., G. Schmidt, and I. Just, "Rho Gtpases as Targets of Bacterial Protein Toxins", *Biological Chemistry,* Vol. 381, pp. 421-6, 2000.

14. Galan, J. E. and H. Wolf-Watz, "Protein Delivery into Eukaryotic Cells by Type Iii Secretion Machines", *Nature,* Vol. 444, pp. 567-573, 2006.

15. Hardt, W. D., L. M. Chen, K. E. Schuebel, X. R. Bustelo, and J. E. Galan, "S. Typhimurium Encodes an Activator of Rho Gtpases That Induces Membrane Ruffling and Nuclear Responses in Host Cells", *Cell,* Vol. 93, pp. 815-26, 1998.

16. Adkins, I., M. Koberle, S. Grobner, E. Bohn, I. B. Autenrieth, and S. Borgmann, "Yersinia Outer Proteins E, H, P, and T Differentially Target the Cytoskeleton and Inhibit Phagocytic Capacity of Dendritic Cells", *International Journal of Medical Microbiology,* Vol. 297, pp. 235-44, 2007.

17. Boyle, E. C., N. F. Brown, and B. B. Finlay, "Salmonella Enterica Serovar Typhimurium Effectors Sopb, Sope, Sope2 and Sipa Disrupt Tight Junction Structure and Function", *Cellular Microbiology,* Vol. 8, pp. 1946-57, 2006.

18. Hapfelmeier, S., K. Ehrbar, B. Stecher, M. Barthel, M. Kremer, and W. D. Hardt, "Role of the Salmonella Pathogenicity Island 1 Effector Proteins Sipa, Sopb, Sope, and Sope2 in Salmonella Enterica Subspecies 1 Serovar Typhimurium Colitis in Streptomycin-Pretreated Mice", *Infection and Immunity,* Vol. 72, pp. 795-809, 2004.

19. Toksoz, D. and K. D. Merdek, "The Rho Small Gtpase: Functions in Health and Disease", *Histology and Histopathology,* Vol. 17, pp. 915-27, 2002.

20. Bernstein, F. C., T. F. Koetzle, G. J. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures", *Journal of Molecular Biology,* Vol. 112, pp. 535-42, 1977.

21. Alberts, B., *Molecular Biology of the Cell*, 4th ed. New York, Garland Science, 2002.

22. Wennerberg, K., K. L. Rossman, and C. J. Der, "The Ras Superfamily at a Glance", *Journal of Cell Science,* Vol. 118, pp. 843-6, 2005.

23. Von Pawel-Rammingen, U., M. V. Telepnev, G. Schmidt, K. Aktories, H. Wolf-Watz, and R. Rosqvist, "Gap Activity of the Yersinia Yope Cytotoxin Specifically Targets the Rho Pathway: A Mechanism for Disruption of Actin Microfilament Structure", *Molecular Microbiology,* Vol. 36, pp. 737-48, 2000.

24. Ahmadian, M. R., P. Stege, K. Scheffzek, and A. Wittinghofer, "Confirmation of the Arginine-Finger Hypothesis for the Gap-Stimulated Gtp-Hydrolysis Reaction of Ras", *Nature Structural Biology,* Vol. 4, pp. 686-9, 1997.

25. Ellenbroek, S. I. and J. G. Collard, "Rho Gtpases: Functions and Association with Cancer", *Clinical and Experimental Metastasis,* Vol. 24, pp. 657-72, 2007.

26. Boettner, B. and L. Van Aelst, "The Role of Rho Gtpases in Disease Development", *Gene,* Vol. 286, pp. 155-74, 2002.

27. Galan, J. E., "Salmonella Interactions with Host Cells: Type Iii Secretion at Work", *Annual Review of Cell and Developmental Biology,* Vol. 17, pp. 53-86, 2001.

28. Wang, Y., L. Zhang, W. L. Picking, W. D. Picking, and R. N. De Guzman, "Structural Dissection of the Extracellular Moieties of the Type Iii Secretion Apparatus", *Molecular BioSystems,* Vol. 4, pp. 1176-80, 2008.

29. Erickson, J. A., M. Jalaie, D. H. Robertson, R. A. Lewis, and M. Vieth, "Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy", *Journal of Medicinal Chemistry,* Vol. 47, pp. 45-55, 2004.

30. Sabina, Y., A. Rahman, R. C. Ray, and D. Montet, "Yersinia Enterocolitica: Mode of Transmission, Molecular Insights of Virulence, and Pathogenesis of Infection", *The Journal of Pathology,* Vol. 2011, p. 429069, 2011.

31. Viboud, G. I. and J. B. Bliska, "Yersinia Outer Proteins: Role in Modulation of Host Cell Signaling Responses and Pathogenesis", *Annual Review of Microbiology,* Vol. 59, pp. 69-89, 2005.

32. Zhang, Y. and J. B. Bliska, "Role of Macrophage Apoptosis in the Pathogenesis of Yersinia", *Current Topics in Microbiology and Immunology,* Vol. 289, pp. 151-73, 2005.

33. Pujol, C. and J. B. Bliska, "Turning Yersinia Pathogenesis Outside In: Subversion of Macrophage Function by Intracellular Yersiniae", *Clinical Immunology,* Vol. 114, pp. 216-26, 2005.

34. Grassl, G. A. and B. B. Finlay, "Pathogenesis of Enteric Salmonella Infections", *Current Opinion in Gastroenterology,* Vol. 24, pp. 22-6, 2008.

35. Layton, A. N. and E. E. Galyov, "Salmonella-Induced Enteritis: Molecular Pathogenesis and Therapeutic Implications", *Expert Reviews in Molecular Medicine,* Vol. 9, pp. 1-17, 2007.

36. Jones, M. A., S. D. Hulme, P. A. Barrow, and P. Wigley, "The Salmonella Pathogenicity Island 1 and Salmonella Pathogenicity Island 2 Type Iii Secretion

Systems Play a Major Role in Pathogenesis of Systemic Disease and Gastrointestinal Tract Colonization of Salmonella Enterica Serovar Typhimurium in the Chicken", *Avian Pathology,* Vol. 36, pp. 199-203, 2007.

37. Cornelis, G. R., "The Yersinia Ysc-Yop 'Type Iii' Weaponry", *Nature Reviews Molecular Cell Biology,* Vol. 3, pp. 742-52, 2002.

38. Schotte, P., G. Denecker, A. Van Den Broeke, P. Vandenabeele, G. R. Cornelis, and R. Beyaert, "Targeting Rac1 by the Yersinia Effector Protein Yope Inhibits Caspase-1-Mediated Maturation and Release of Interleukin-1beta", *The Journal of Biological Chemistry,* Vol. 279, pp. 25134-42, 2004.

39. Sory, M. P., A. Boland, I. Lambermont, and G. R. Cornelis, "Identification of the Yope and Yoph Domains Required for Secretion and Internalization into the Cytosol of Macrophages, Using the Cyaa Gene Fusion Approach", *Proceedings of the National Academy of Sciences,* Vol. 92, pp. 11998-2002, 1995.

40. Schesser, K., E. Frithz-Lindsten, and H. Wolf-Watz, "Delineation and Mutational Analysis of the Yersinia Pseudotuberculosis Yope Domains Which Mediate Translocation across Bacterial and Eukaryotic Cellular Membranes", *Journal of Bacteriology,* Vol. 178, pp. 7227-33, 1996.

41. Evdokimov, A. G., J. E. Tropea, K. M. Routzahn, and D. S. Waugh, "Crystal Structure of the Yersinia Pestis Gtpase Activator Yope", *Protein Science,* Vol. 11, pp. 401-8, 2002.

42. Buchwald, G., A. Friebel, J. E. Galan, W. D. Hardt, A. Wittinghofer, and K. Scheffzek, "Structural Basis for the Reversible Activation of a Rho Protein by the Bacterial Toxin Sope", *EMBO Journal,* Vol. 21, pp. 3286-95, 2002.

43. Keyser, P., M. Elofsson, S. Rosell, and H. Wolf-Watz, "Virulence Blockers as Alternatives to Antibiotics: Type Iii Secretion Inhibitors against Gram-Negative Bacteria", *Journal of Internal Medicine,* Vol. 264, pp. 17-29, 2008.

44. Pan, N., C. Lee, and J. Goguen, "High Throughput Screening for Small-Molecule Inhibitors of Type Iii Secretion in Yersinia Pestis", *Advances in Experimental Medicine and Biology,* Vol. 603, pp. 367-75, 2007.

45. Hudson, D. L., A. N. Layton, T. R. Field, A. J. Bowen, H. Wolf-Watz, M. Elofsson, M. P. Stevens, and E. E. Galyov, "Inhibition of Type Iii Secretion in Salmonella Enterica Serovar Typhimurium by Small-Molecule Inhibitors", *Antimicrobial Agents and Chemotherapy,* Vol. 51, pp. 2631-5, 2007.

46. Nordfelth, R., A. M. Kauppi, H. A. Norberg, H. Wolf-Watz, and M. Elofsson, "Small-Molecule Inhibitors Specifically Targeting Type Iii Secretion", *Infection and Immunity,* Vol. 73, pp. 3104-14, 2005.

47. Kauppi, A. M., R. Nordfelth, U. Hagglund, H. Wolf-Watz, and M. Elofsson, "Salicylanilides Are Potent Inhibitors of Type Iii Secretion in Yersinia", *Advances in Experimental Medicine and Biology,* Vol. 529, pp. 97-100, 2003.

48. Kauppi, A. M., R. Nordfelth, H. Uvell, H. Wolf-Watz, and M. Elofsson, "Targeting Bacterial Virulence: Inhibitors of Type Iii Secretion in Yersinia", *Chemistry & Biology,* Vol. 10, pp. 241-9, 2003.

49. Drews, J., "Drug Discovery: A Historical Perspective", *Science,* Vol. 287, pp. 1960-1964, 2000.

50. Archer, J. R., "History, Evolution, and Trends in Compound Management for High Throughput Screening", *Assay and Drug Development Technologies,* Vol. 2, pp. 675-681, 2004.

51. Harding, D., M. Banks, S. Fogarty, and A. Binnie, "Development of an Automated High-Throughput Screening System: A Case History", *Drug Discovery Today,* Vol. 2, pp. 385-390, 1997.

52. Lilburn, T. G. and Y. F. Wang, "Systems Biology and Computer-Aided Drug Discovery", *Current Computer-Aided Drug Design,* Vol. 2, pp. 267-274, 2006.

53. Clark, D. E., "What Has Computer-Aided Molecular Design Ever Done for Drug Discovery?", *Expert Opinion on Drug Discovery,* Vol. 1, pp. 103-110, 2006.

54. Rester, U., "From Virtuality to Reality - Virtual Screening in Lead Discovery and Lead Optimization: A Medicinal Chemistry Perspective", *Current Opinion in Drug Discovery & Development,* Vol. 11, pp. 559-568, 2008.

55. Cerqueira, N. M. F. S. A., S. F. Sousa, P. A. Fernandes, and M. J. Ramos, "Virtual Screening of Compound Libraries", *Ligand-Macromolecular Interactions in Drug Discovery: Methods and Protocols,* Vol. 572, pp. 57-70, 2010.

56. Cramer, R. D., D. E. Patterson, R. D. Clark, F. Soltanshahi, and M. S. Lawless, "Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research", *Journal of Chemical Information and Computer Sciences,* Vol. 38, pp. 1010-1023, 1998.

57. Barbosa, A. J. M. and A. Del Rio, "Freely Accessible Databases of Commercial Compounds for High-Throughput Virtual Screenings", *Current Topics in Medicinal Chemistry,* Vol. 12, pp. 866-877, 2012.

58. Koppen, H., "Virtual Screening - What Does It Give Us?", *Current Opinion in Drug Discovery & Development,* Vol. 12, pp. 397-407, 2009.

59. Pan, S. Y., S. Pan, Z. L. Yu, D. L. Ma, S. B. Chen, W. F. Fong, Y. F. Han, and K. M. Ko, "New Perspectives on Innovative Drug Discovery: An Overview", *Journal of Pharmacy and Pharmaceutical Sciences,* Vol. 13, pp. 450-471, 2010.

60. Norinder, U., "The Advantages of Using Rational Drug Design in Modern Drug Discovery: How to Integrate Computer-Aided Drug Design and Modern

Biotechnology", *Computer Aided Drug Design in Industrial Research,* Vol. 15, pp. 99-109, 1995.

61. Clark, R. D. and D. C. Roe, "Ligand- and Structure-Based Virtual Screening", *Handbook of Chemoinformatics Algorithms,* pp. 145-171, 2010.

62. Cavasotto, C. N. and A. J. W. Orry, "Ligand Docking and Structure-Based Virtual Screening in Drug Discovery", *Current Topics in Medicinal Chemistry,* Vol. 7, pp. 1006-1014, 2007.

63. Moro, S., M. Bacilieri, and F. Deflorian, "Combining Ligand-Based and Structure-Based Drug Design in the Virtual Screening Arena", *Expert Opinion on Drug Discovery,* Vol. 2, pp. 37-49, 2007.

64. Klebe, G., "Virtual Ligand Screening: Strategies, Perspectives and Limitations", *Drug Discovery Today,* Vol. 11, pp. 580-594, 2006.

65. Ekins, S., J. Mestres, and B. Testa, "In Silico Pharmacology for Drug Discovery: Methods for Virtual Ligand Screening and Profiling", *British Journal of Pharmacology,* Vol. 152, pp. 9-20, 2007.

66. Ekins, S., J. Mestres, and B. Testa, "In Silico Pharmacology for Drug Discovery: Applications to Targets and Beyond", *British Journal of Pharmacology,* Vol. 152, pp. 21-37, 2007.

67. Talele, T. T., S. A. Khedkar, and A. C. Rigby, "Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic", *Current Topics in Medicinal Chemistry,* Vol. 10, pp. 127-41, 2010.

68. Svensson, F., A. Karlen, and C. Skold, "Virtual Screening Data Fusion Using Both Structure- and Ligand-Based Methods", *Journal of Chemical Information and Modeling,* Vol. 52, pp. 225-32, 2012.

69. Ripphausen, P., B. Nisius, and J. Bajorath, "State-of-the-Art in Ligand-Based Virtual Screening", *Drug Discovery Today,* Vol. 16, pp. 372-6, 2011.

70. Jahn, A., G. Hinselmann, N. Fechner, and A. Zell, "Optimal Assignment Methods for Ligand-Based Virtual Screening", *Journal of Chemical Information and Modeling,* Vol. 1, p. 14, 2009.

71. Hu, G., G. Kuang, W. Xiao, W. Li, G. Liu, and Y. Tang, "Performance Evaluation of 2d Fingerprint and 3d Shape Similarity Methods in Virtual Screening", *Journal of Chemical Information and Modeling,* Vol. 52, pp. 1103-13, 2012.

72. Stahura, F. L. and J. Bajorath, "New Methodologies for Ligand-Based Virtual Screening", *Current Pharmaceutical Design,* Vol. 11, pp. 1189-202, 2005.

73. Miller, M. A., "Chemical Database Techniques in Drug Discovery", *Nature Reviews Drug Discovery,* Vol. 1, pp. 220-7, 2002.

74. Willett, P., "Similarity-Based Virtual Screening Using 2d Fingerprints", *Drug Discovery Today,* Vol. 11, pp. 1046-53, 2006.

75. Alvesalo, J. K., A. Siiskonen, M. J. Vainio, P. S. Tammela, and P. M. Vuorela, "Similarity Based Virtual Screening: A Tool for Targeted Library Design", *Journal of Medicinal Chemistry,* Vol. 49, pp. 2353-6, 2006.

76. Hert, J., P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, "Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures", *Organic & Biomolecular Chemistry,* Vol. 2, pp. 3256-66, 2004.

77. Dixon, S. L., A. M. Smondyrev, and S. N. Rao, "Phase: A Novel Approach to Pharmacophore Modeling and 3d Database Searching", *Chemical Biology & Drug Design,* Vol. 67, pp. 370-2, 2006.

78. Spitzer, G. M., M. Heiss, M. Mangold, P. Markt, J. Kirchmair, G. Wolber, and K. R. Liedl, "One Concept, Three Implementations of 3d Pharmacophore-Based Virtual Screening: Distinct Coverage of Chemical Search Space", *Journal of Chemical Information and Modeling,* Vol. 50, pp. 1241-7, 2010.

79. Oprea, T. I., "Property Distribution of Drug-Related Chemical Databases", *Journal of Computer-Aided Molecular Design,* Vol. 14, pp. 251-64, 2000.

80. Dixon, S. L., A. M. Smondyrev, E. H. Knoll, S. N. Rao, D. E. Shaw, and R. A. Friesner, "Phase: A New Engine for Pharmacophore Perception, 3d Qsar Model Development, and 3d Database Screening: 1. Methodology and Preliminary Results", *Journal of Computer-Aided Molecular Design,* Vol. 20, pp. 647-71, 2006.

81. Hasegawa, K. and K. Funatsu, "Partial Least Squares Modeling and Genetic Algorithm Optimization in Quantitative Structure-Activity Relationships", *Sar and Qsar in Environmental Research,* Vol. 11, pp. 189-209, 2000.

82. Akamatsu, M., "Current State and Perspectives of 3d-Qsar", *Current Topics Medicinal Chemistry,* Vol. 2, pp. 1381-94, 2002.

83. Ghosh, S., A. Nie, J. An, and Z. Huang, "Structure-Based Virtual Screening of Chemical Libraries for Drug Discovery", *Current Opinion in Chemical Biology,* Vol. 10, pp. 194-202, 2006.

84. Cavasotto, C. N. and A. J. Orry, "Ligand Docking and Structure-Based Virtual Screening in Drug Discovery", *Current Topics in Medicinal Chemistry,* Vol. 7, pp. 1006-14, 2007.

85. Kitchen, D. B., H. Decornez, J. R. Furr, and J. Bajorath, "Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications", *Nature Reviews Drug Discovery,* Vol. 3, pp. 935-49, 2004.

86. Cross, J. B., D. C. Thompson, B. K. Rai, J. C. Baber, K. Y. Fan, Y. Hu, and C. Humblet, "Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy", *Journal of Chemical Information and Modeling,* Vol. 49, pp. 1455-74, 2009.

87. Friesner, R. A., J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy", *Journal of Medicinal Chemistry,* Vol. 47, pp. 1739-49, 2004.

88. Lyne, P. D., "Structure-Based Virtual Screening: An Overview", *Drug Discovery Today,* Vol. 7, pp. 1047-55, 2002.

89. Warren, G. L., C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head, "A Critical Assessment of Docking Programs and Scoring Functions", *Journal of Medicinal Chemistry,* Vol. 49, pp. 5912-31, 2006.

90. Bissantz, C., G. Folkers, and D. Rognan, "Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations", *Journal of Medicinal Chemistry,* Vol. 43, pp. 4759-67, 2000.

91. Halperin, I., B. Ma, H. Wolfson, and R. Nussinov, "Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions", *Proteins,* Vol. 47, pp. 409-43, 2002.

92. Schrödinger Suite 2011 Schrödinger Suite; Epik version 2.2, Schrödinger, LLC, New York, NY, 2011; Impact version 5.7, Schrödinger, LLC, New York, NY, 2011; Prime version 2.3, Schrödinger, LLC, New York, NY, 2011.

93. LigPrep, version 2.5, Schrödinger, LLC, New York, NY, 2011.

94. ConfGen, version 2.3, Schrödinger, LLC, New York, NY, 2011.

95. Phase, version 3.3, Schrödinger, LLC, New York, NY, 2011.

96. QikProp, version 3.4, Schrödinger, LLC, New York, NY, 2011.

97. Glide, version 5.7, Schrödinger, LLC, New York, NY, 2011.

98. MacroModel, version 9.9, Schrödinger, LLC, New York, NY, 2011.

99. Prime, version 2.1, Schrödinger, LLC, New York, NY, 2009.

100. Maestro, version 9.2, Schrödinger, LLC, New York, NY, 2011.

101. Epik, version 2.2, Schrödinger, LLC, New York, NY, 2011.

102. Hibbert, D. B., P. Minkkinen, N. M. Faber, and B. M. Wise, "Iupac Project: A Glossary of Concepts and Terms in Chemometrics", *Analytica Chimica Acta,* Vol. 642, pp. 3-5, 2009.

103. Shannon, W. D., M. A. Province, and D. C. Rao, "Tree-Based Recursive Partitioning Methods for Subdividing Sibpairs into Relatively More Homogeneous Subgroups", *Genetic Epidemiology,* Vol. 20, pp. 293-306, 2001.

104. Schrödinger Fragment Library (2009) Schrödinger, Inc. http://www.schrodinger.com/productpage/14/5/73/ [cited at 12/04/2012].

105. Eldridge, M. D., C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee, "Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes", *Journal of Computer-Aided Molecular Design,* Vol. 11, pp. 425-45, 1997.

106. Halgren, T. A., R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening", *Journal of Medicinal Chemistry,* Vol. 47, pp. 1750-9, 2004.

107. E, D. E. C., J. Descamps, D. E. S. P, and A. Holyacute, "(S)-9-(2,3-Dihydroxypropyl)Adenine: An Aliphatic Nucleoside Analog with Broad-Spectrum Antiviral Activity", *Science,* Vol. 200, pp. 563-5, 1978.

108. SiteMap, version 2.5, Schrödinger, LLC, New York, NY, 2011.

109. Halgren, T. A., "Identifying and Characterizing Binding Sites and Assessing Druggability", *Journal of Chemical Information and Modeling,* Vol. 49, pp. 377-89, 2009.

110. Canvas, version 1.4, Schrödinger, LLC, New York, NY, 2011.