MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado em Programa de Pós-Graduação em Biologia Computacional e Sistemas

PLANEJAMENTO RACIONAL DE FÁRMACOS APLICADO À BUSCA E OTIMIZAÇÃO DE INIBIDORES DO HIV-1 E DOENÇA DE CHAGAS

LUCIANNA HELENE SILVA DOS SANTOS

Rio de Janeiro
Abril de 2016

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

LUCIANNA HELENE SILVA DOS SANTOS

Planejamento racional de fármacos aplicado à busca e otimização de inibidores do HIV-1 e doença de Chagas

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em biologia computacional e sistemas

Orientador (es): Prof. Dr. Ernesto Raúl Caffarena
Prof. Dra. Rafaela Salgado Ferreira

RIO DE JANEIRO
Abril de 2016

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

S237 Santos, Lucianna Helene Silva dos

Planejamento racional de fármacos aplicado à busca e otimização de inibidores do HIV-1 e Doença de Chagas / Lucianna Helene Silva dos Santos. – Rio de Janeiro, 2016.
xix, 159 f. : il. ; 30 cm.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2016.
Inclui bibliografia.

1. Planejamento de fármacos. 2. Modelagem molecular. 3. HIV-1. 4. Doença de Chagas. 5. Doença do sono. 6. Dinâmica molecular. 7. Docking molecular. 8. Triagem virtual. I. Título.

CDD 615.19

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: LUCIANNA HELENE SILVA DOS SANTOS

PLANEJAMENTO RACIONAL DE FÁRMACOS APLICADOS À BUSCA E OTIMIZAÇÃO DE INIBIDORES DO HIV-1 E DOENÇA DE CHAGAS

ORIENTADOR (ES): Prof. Dr. ERNESTO RAÚL CAFFARENA
Prof. Dra. RAFAELA SALGADO FERREIRA

Aprovada em: _____/_____/_____

EXAMINADORES:

Prof. Dr. Floriano Paes Silva Junior **- Presidente** (IOC-Fiocruz)
Prof. Dr. Claudio Norberto Cavasotto (IBioBA-MPSP)
Prof. Dr. Carlos Alberto Montanari (USP)
Prof. Dr. Laurent Emmanuel Dardenne – Suplente (LNCC)
Prof. Dra. Ana Carolina Ramos Guimarães – Suplente (IOC-Fiocruz)

Rio de Janeiro, 05 de abril de 2016

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

Anexar a cópia da Ata que será entregue pela SEAC já assinada.

À minha família e amigos.

AGRADECIMENTOS

Agradeço primeiramente ao Prof. Dr. Ernesto Raúl Caffarena, pela orientação, dedicação e oportunidade de aprendizado, sobretudo pela amizade, confiança e tempo dedicado as dúvidas surgidas pelo caminho.

À Profª. Dra. Rafaela Salgado Ferreira, pelas sugestões e incentivo que tornaram possível esse trabalho.

Aos amigos e colegas, em especial, André, Amanda, Deborah, Janaina, Raul e Vanessa, pela amizade que compartilhamos, por toda ajuda, força e pelos bons momentos de descontração tão importantes durante essa jornada.

Um agradecimento em especial ao Rafael, uma das pessoas mais engraçadas que já conheci, pelos conselhos e por ser grande encorajador do meu humor peculiar.

Ao corpo docente da BCS, em especial a Dra. Aline Nobre, Dra. Ana Carolina Guimarães, Dr. Paulo Ricardo Batista e Dr. Mauricio Costa, pelo convívio, apoio e compreensão.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro durante esses quatro anos de trabalho.

Ao Projeto Casadinho – BCS e UFMG, por ter proporcionado realização de parte desse trabalho.

Ao Dr. Klaus Liedl da Universidade de Innsbruck por me receber em seu instituto e me dar todo o suporte para realização de parte desse trabalho.

Aos Dr. Julian Fusch e Dra. Birgit Waldner pelas sugestões e apoio ao trabalho.

Agradeço aos meus pais, pelo apoio aos meus estudos, aos ensinamentos e pelo exemplo de vida, de carinho, de trabalho, de honestidade e simplicidade.

À minha irmã, pela amizade, companheirismo e apoio em todos as ocasiões.

A toda a minha família, pela preocupação e incentivo nas horas que mais precisei.

“Once I got home, I sulked for a while.
All my brilliant plans foiled by
thermodynamics. Damn you, Entropy!”
- Andy Weir, The Martian.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

PLANEJAMENTO RACIONAL DE FÁRMACOS APLICADOS À BUSCA E OTIMIZAÇÃO DE INBIDORES DO HIV-1 E DOENÇA DE CHAGAS

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

LUCIANNA HELENE SILVA DOS SANTOS

O sucesso no planejamento de fármacos de novas moléculas bioativas está relacionado com o entendimento do alvo molecular e seu valor para com a doença que deseja ser combatida. Entender as interações importantes serve como guia para planejar e testar potenciais ligantes, e consequentemente, para selecionar características estruturais que serão incluídas na síntese combinatória de bibliotecas de compostos. Tendo isso em vista, nós propomos este trabalho em duas partes onde os usos de abordagens computacionais de modelagem molecular são aplicados em distintos alvos moleculares. Na primeira parte, triagem virtual para a descoberta de novos compostos no sítio alostérico da transcriptase reversa do HIV-1 foi aplicado. Uma combinação de técnicas computacionais, envolvendo *docking* molecular, métodos de avaliação de desempenho, e métodos baseados na similaridade de ligantes, tornou possível a identificação de compostos candidatos extraídos de uma biblioteca de mais de dois milhões de compostos comercialmente adquiríveis. Propomos nesse trabalho vinte compostos que alcançaram boa pontuação de *docking* e possuem baixa similaridade entre si, tornando os distintos, para subsequentes avaliações. As interações desses compostos com o sítio de ligação mostraram-se similares com as ligações já determinadas em ligantes conhecidos, sugerindo que o método utilizado neste trabalho é apropriado na escolha de potenciais candidatos. Na segunda parte, é feita a investigação do comportamento dinâmico e das energias de um conjunto de inibidores não covalentes das enzimas cruzaína e rodesaína. Estados de protonação diferentes tanto dos resíduos catalíticos das enzimas (His162 e Cys25) quanto do ligante foram investigados por meio de dinâmica molecular de modo a elucidar o possível modo de ligação do inibidor não-covalente B95 e de uma série de análogos desse inibidor. As análises de dinâmica molecular apontam para a protonação de ambos resíduos catalíticos, conhecida como par iônico, junto com a protonação do ligante como o sistema mais favorável em um possível modo de ligação. Para os compostos análogos cálculo de energia livre foram realizados mostrando boa concordância entre os dados calculados e dados experimentais para uma das enzimas, a cruzaína.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

COMPUTER-AIDED RATIONAL DRUG DESIGN APPLIED TO THE DISCOVERY AND OPTIMIZATION OF HIV-1 AND CHAGAS DISEASE INHIBITORS

ABSTRACT

PHD THESIS IN COMPUTATIONAL BIOLOGY AND SYSTEMS

LUCIANNA HELENE SILVA DOS SANTOS

A successful drug design of new bioactive molecules is related to the understanding of the molecular target and its value to the studied disease. The knowledge of important interactions might serve as a guide to plan and test potential ligands, and consequently, to select structural features to be included in the combinatorial synthesis of compound libraries. With this in mind, we propose this work divided in two parts where the use of computational drug design strategies was applied to distinct molecular targets. In the first part, we performed virtual screening of a large library of commercially-available compounds to discover new lead candidates of the HIV-1 reverse transcriptase enzyme into the allosteric binding site. A combination of computational approaches, involving molecular docking, enrichment metrics, and compound similarity methods, made possible the identification of a set of candidates from a library of over two million commercially-available compounds. We propose in this work twenty compounds that have achieved good docking score and have low similarity to each other, making them all distinct for subsequent evaluations. The compounds' ligand-receptor interactions were similar to the ones found in known inhibitors of reverse transcriptase which suggests that the method used in this work is suitable in choosing potential candidates. In the second part, we investigate the dynamic and energy behavior of a set noncovalent inhibitors of the enzymes cruzain and rhodesain. Different protonation states of the enzymes' catalytic residues (His 162 and Cys 25) and the ligand were tested by molecular dynamics simulations, to elucidate a possible binding mode of the noncovalent inhibitor B95 and a series of analogues of this ligand. The molecular dynamic analysis indicated that the protonation of both catalytic residues, known as ion pair, together with the protonation of the ligand was the most favorable in a possible binding mode. For the analogues compounds, free energy calculations were done, and the cruzain systems showed good agreement between calculated relative free energy of binding and experimental relative free energy of binding.

PREFACE

The work presented in the thesis originated as two separate projects accomplished throughout the course of the author's Ph.D. program. The extensive investigation of the targets and the use of different molecular modeling methods in the projects led to the combination of both projects in two separate parts in the thesis.

The first project, entitle "Virtual screening application for discovering HIV-1 reverse transcriptase inhibitors", corresponds to the first section of the thesis. The initial part of this project was achieved at Universidade Federal de Minas Gerais (UFMG) supported by Projeto Casadinho. Projeto Casadinho promotes the interaction between UFMG and Instituto Oswaldo Cruz. The project was entirely supervised by Dr. Rafaela S. Ferreira and Dr. Ernesto R. Caffarena.

The second project, entitle "Energy of Binding Predictions of Rhodesain and Cruzain Inhibitors by Computer Simulations", corresponds to the second and final section of the thesis. This work was conducted in the Center for Molecular Biosciences Faculty for Chemistry and Pharmacy, University of Innsbruck. The Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES) is also acknowledged for sponsoring the author (Programa de Bolsas de Doutorado Sanduiche no Exterior - PDSE, Sistema de Concessões de Bolsa no Exterior - BEX process 010357/2014-09) to carry out part of the work at University of Innsbruck. Dr. Klaus Liedl was the main foreigner supervisor with collaborations of Dr. Julian Fusch and Dr. Birgit Waldner. The directors, Dr. Ernesto R. Caffarena and Dr. Rafaela S. Ferreira were throughout involved in the project progress also.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIDS	Acquired immune deficiency syndrome
AUC	Area under the ROC curve
dsDNA	double-stranded DNA
EF	Enrichment factor
EFZ	Efavirenz
Elec	Electrostatic
H162q-Cys25n	His162-NH+/Cys25-SH
H162q-Cys25q	His162-NH+/Cys25-S- (ionic pair)
HAT	Human African Trypanosomiasis
H-bond	Hydrogen bond
HIV-1	Human immunodeficiency virus type 1
MC	Monte Carlo
MD	Molecular dynamics
MM-GB/SA	Molecular mechanics Generalized Born surface area
MM-PB/SA	Molecular mechanics Poisson-Boltzmann surface area
NNBP	Non-nucleoside binding pocket
NNRTIs	Non-nucleoside reverse transcriptase inhibitors
NRTIs	Nucleoside reverse transcriptase inhibitors
NtRTIs	Nucleotide reverse transcriptase inhibitors
NVP	Nevirapine
RMSDh	Hungarian (symmetry-corrected) heavy atom RMSD
RMSDm	Minimum-distance heavy-atom RMSD
RMSDs	Standard heavy-atom RMSD
ROC	Receiver operating characteristic curves
RPV	Rilpivirine
R_s	Spearman correlation coefficient
RT	Reverse transcriptase
ssRNA	single-stranded RNA
TI	Thermodynamic integration
vdWaals	van der Waals
VS	Virtual screening

1 PART I

1.1 INTRODUCTION

1.1.1 Human immunodeficiency virus type 1 (HIV-1)

The human immunodeficiency virus (HIV) was established in 1983 as the causative agent of the acquired immune deficiency syndrome (AIDS) (Barre-Sinoussi et al. 1983), which remains as a global health care issue. HIV has two known variants: HIV-1, which causes HIV infections worldwide; and HIV-2, mostly confined to the West Africa (Reeves and Doms 2002). In the natural course of HIV-1 infection, CD4$^+$ T cells, essential for adaptive immunity, are severely deteriorated (Laskey and Siliciano 2014). When the number of these vital cells in the human immune system declines below a critical level, cell-mediated immunity is lost, and the body becomes progressively more susceptible to life-threatening opportunistic infections and cancers to prosper.

The HIV-1 life cycle starts with the infection of CD4$^+$ T cells by the virus. Viral entry is facilitated by the binding of the viral Env glycoprotein to two cell surface proteins: CD4 and a co-receptor (Laskey and Siliciano 2014). The co-receptor for HIV-1 entry is either CC-chemokine receptor 5 (CCR5) or CXC-chemokine receptor 4 (CXCR4) (Choe et al. 1996). Since HIV-1 is a retrovirus, its virions contain two copies of a single-stranded RNA (ssRNA) genome, which is reverse transcribed to double-stranded DNA (dsDNA) by the viral enzyme reverse transcriptase and is integrated into the host genome by the viral enzyme integrase (Esposito et al. 2012). The integrated viral genome, the so-called provirus, functions as a cellular gene: in activated CD4$^+$ T cells, the provirus is transcribed and translated to produce viral proteins, which combine with the viral genomic RNA to procedure new virions. After the production and release of new virions from the cell surface, the viral enzyme protease enables virion maturation by cleaving viral polyproteins into functional subunits to produce infectious particles (Laskey and Siliciano 2014).

Thirty years of research and technological innovation have allowed validation of several steps of the HIV life cycle as intervention points for antiretroviral therapies. The highly active antiretroviral therapy (HAART) is the standard treatment for HIV-infected patients and consists in the combination of three or more HIV drugs to achieve maximal virological response and reduce the potential development of antiviral resistance (Asahchop et al. 2012). Currently, twenty-six antiretroviral drugs have been approved by the United States Food and Drug Administration (FDA) (FDA 2013). These compounds are classified into six categories according to their target: nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), protease inhibitors (PIs), cell entry inhibitors or fusion inhibitors (FIs), co-receptor inhibitors (CRIs), and integrase inhibitors (INIs).

Although the presently available antiretroviral therapy proved that HIV infection is treatable, some challenges remain (Broder 2010). One important factor is the constant occurrence of new infections in many regions of the world. According to the Joint United Nations Programme on HIV/Acquired Immune Deficiency Syndrome (UNAIDS), approximately 35 million people were living with HIV and an estimated 2.3 million new HIV infections happened globally in 2012 (UNAIDS 2013). The life-long treatment brings another challenge. It can lead to long-term cardiac and metabolic complications such as dyslipidemias, insulin resistance, lipodystrophy, heart diseases and other related disorders (Filardi et al. 2008, Group et al. 2008, Silverberg et al. 2009). Also, treatment can be impaired by the development of drug resistance strains when viral suppression is not maintained (Scarth et al. 2011). A vast number of viruses are produced daily in an infected individual and genetic variation within individuals has contributed to the emergence of diverse HIV-1 subtypes, complicating extensively the development of active drugs (Sarafianos et al. 2004). Therefore, current antiretroviral research efforts have been aiming at refining current therapies and discovering new drugs with lower toxicity and favorable resistance profile (Ghosh et al. 2008, Ghosh et al. 2011, Cao et al. 2014, Maga et al. 2010, Michailidis et al. 2014, Quashie et al. 2012).

1.1.2 HIV-1 RT enzyme

The HIV-1 enzyme Reverse Transcriptase (RT) is a primary target for antiretroviral drugs. Today, a total of thirteen inhibitors act against it, including the very first drug used in HIV treatment, the NNRTI zidovudine (AZT) (Esposito et al. 2012). RT is the enzyme that converts the ssRNA into dsDNA provirus, which is afterwards imported into the cell nucleus to be integrated into the host chromosome (Esposito et al. 2012) with the help of integrase, another HIV enzyme. The catalytic steps are performed in the following order: (i) RNA-dependent DNA polymerization to synthesize an antisense (-) DNA strand, complementing the viral sense (+) RNA strand; (ii) RNase H cleavage of the RNA strand from the RNA-DNA complex; and (iii) the DNA-dependent DNA polymerization to synthesize dsDNA using the remaining antisense

(-) DNA strand as the template (Das et al. 2012). Other crucial activities of the retrotranscription process can be attributed to this highly effective enzyme: an endonucleolytic ribonuclease H (RNase H) activity and strand transfer (Liu et al. 2008).

The mechanism of viral DNA synthesis is quite similar to other DNA polymerases (Steitz 1999), but not as accurate, since viral RTs lack the 3' → 5' exonucleolytic proofreading activity (Menendez-Arias 2009). Proofreading DNA polymerases, such as DNA polymerases δ, γ, and/or ε, are characterized by an average error rates about 10^{-6} to 10^{-7} (Matsuda et al. 2003), whereas those of RT lean towards to 10-100 times higher (Menendez-Arias 2009). It is believed that the error-prone nature of RT, alongside mutations introduced by the host RNA polymerase, results in retroviral variation (Das and Arnold 2013). These factors might allow HIV to mutate rapidly, producing drug resistance strains in weeks after the treatment begins (Frankel and Young 1998).

RT is a heterodimer (Figure 1) composed of two subunits of 560 and 440 amino acid residues, referred to as p66 and p51, respectively (Menendez-Arias 2013). These subunits share almost the same amino acid sequences. However, p51 lacks the catalytic activity and the RNase H domain, performing a structural support role and assisting p66 loading onto nucleic acid (Kohlstaedt et al. 1992). Unlike p51, p66 has a more flexible structure and contains the polymerase and RNase H active sites (Steitz et al. 1993). The polymerase domain resembles the shape of the human right hand (Steitz 1999) where the p66 subdomains are designated as “fingers” (residues 1-85 and 118-155), “palm” (residues 86-117 and 156-236) and “thumb” (residues 237-318), as well as a “connection” subdomain (residues 319-426) that joins the DNA polymerase and the RNase H domains (Menendez-Arias 2013). The polymerase active site is located in the middle of the palm, fingers, and thumb subdomains. In addition, the p66 subunit contains a nucleic acid binding cleft, as well as active-site carboxylates (residues Asp110, Asp185, and Asp186), that bind the divalent magnesium ion (Mg^{2+}) required for catalysis (Mendieta et al. 2008). Nonetheless, all the commercially available RT-targeting drugs affect the polymerase activity inhibiting its function, some RNase H inhibitors have recently been designed and studied (Tramontano and Di Santo 2010, Distinto et al. 2013).

RT has been the focus of extensive research, including several structural biology studies that led to the determination of numerous crystallographic structures. Currently, over one hundred RT crystal structures are available in the RCSB Protein Data Bank repository (Berman et al. 2000). The available RT crystal structures provide insight into the conformational flexibility of the protein, including the conformational changes induced by inhibitors and DNA binding (Titmuss et al. 1999). For instance, the formation of the non-nucleoside binding pocket (NNBP) is induced by the presence of an NNRTI, i.e. it only exists in RT structures complexed with this kind of inhibitors. The “open” (when the fingers and thumb subdomains are far apart) and “closed” (when the fingers and thumb subdomains are closer to the palm subdomain) conformations can be found in crystal structures with bound and unbound DNA, respectively.

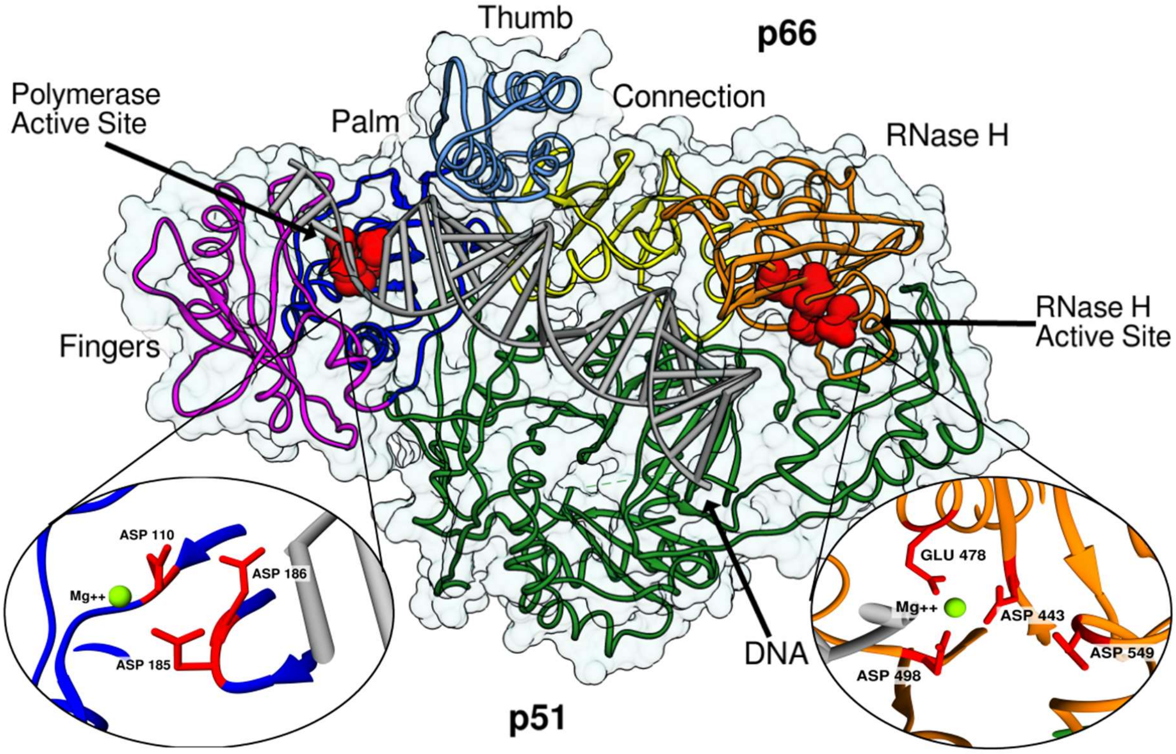


Figure 1: Structure of HIV-1 RT in complex with DNA (PDB code: 1T05 (Tuske et al. 2004)). The two domains are the p66 (colored) and the p51 (green). The polymerase domain displays a highly conserved structure that resembles the shape of the human right hand, consisting of fingers domain (magenta), palm domain (blue), thumb domain (light blue). The p66 subunit also includes the connection domain (yellow) and RNase H domain (orange). The polymerase active site is located in the center of palm, fingers, and thumb subdomains. The three catalytic aspartic acid residues (110, 185 and 186), shown in red, are located in the palm subdomain and bind the cofactor divalent ions (Mg2+). The RNase H domain is situated at the p66 C-terminus, approximately 60 Å from polymerase active site. The RNase H active site contains a DDE motif comprising the carboxylates residues ASP443, GLU478, ASP498, and ASP549 that can coordinate divalent Mg2+.

1.1.2.1 HIV-1 RT inhibitors

The two classes of RT inhibitors include nucleoside analogs RT inhibitors (NRTIs) and non-nucleoside analogs RT inhibitors (NNRTIs). The NRTIs are composed of modified nucleosides that mimic and compete with natural substrates for

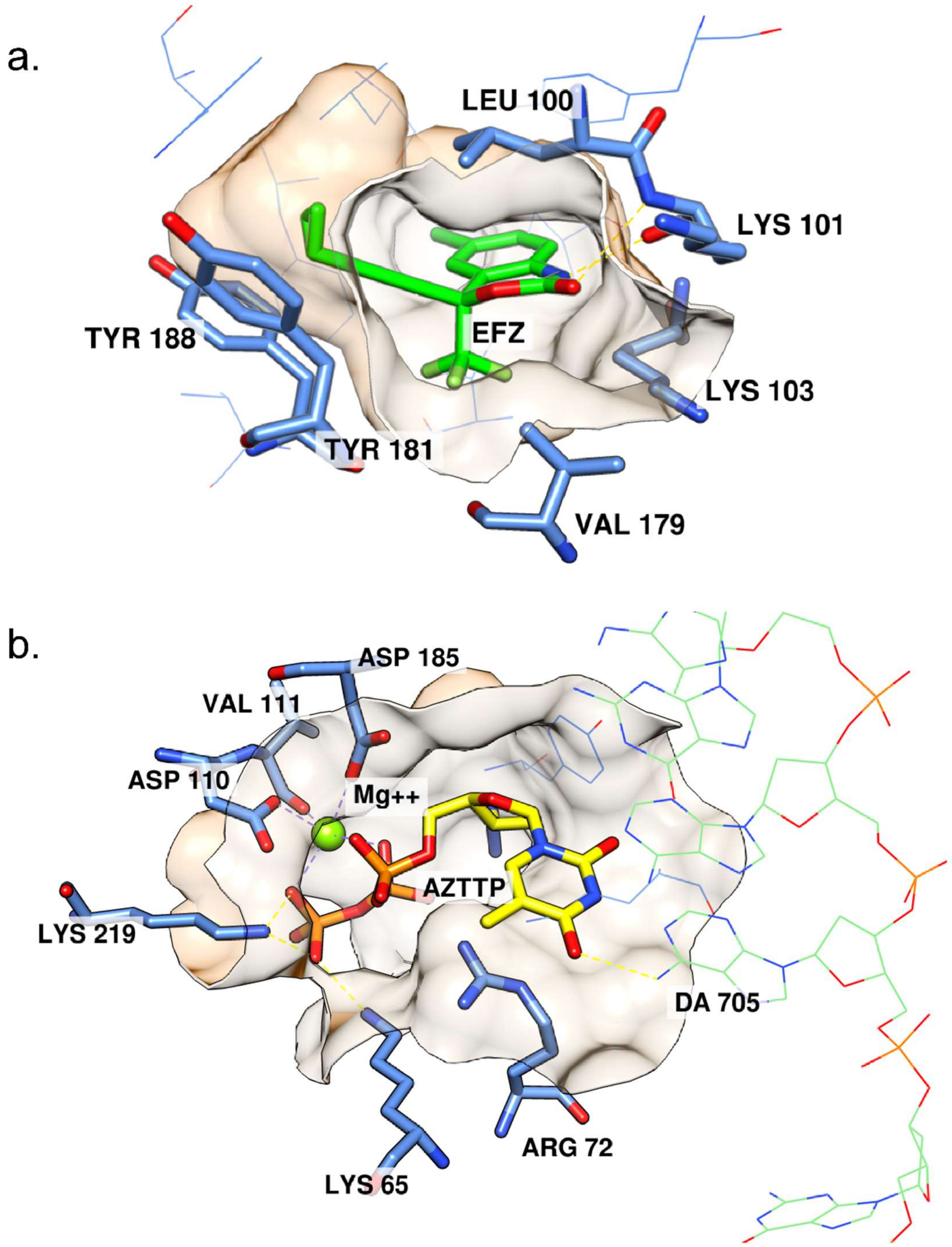


Figure 2: Graphic illustration of a NNRTI and NRTI in their binding site. (a) Efavirenz (green) within the NNRTI allosteric binding site (PDB code: 1FK9 (Ren et al. 2000)) and (b) AZT (yellow) within the NRTI binding site (PDB code: 3V4I (Das et al. 2012)).

binding and incorporation at the polymerase site (Figure 2-b) (Mehellou and De Clercq 2010). They act as chain terminators due to the lack of a 3’-OH group on their sugar moiety. Similarly to their natural counterparts, the NRTIs need to be converted in 5’-triphosphate nucleotides by host-cell kinases to compete with the analogous deoxynucleotide-triphosphates (dNTPs), and consequently be incorporated into the growing DNA strand (Esposito et al. 2012).

The current clinically available NRTIs are structurally similar to pyrimidine and purine analogues, including thymidine analogues zidovudine (AZT, Retrovir®) and stavudine (d4T, Zerit®); together with cytidine analogues zalcitabine (ddC, Hivid®), lamivudine (3TC, Epivir®) and emtricitabine (FTC, Emtriva®). Purine analogues include the inosine analogue didanosine (ddI, Videx®) along with the carbocyclic nucleoside analogue abacavir (ABC, Ziagen®), a guanine analogue when in its active form (Figure 3) (Mehellou and De Clercq 2010).

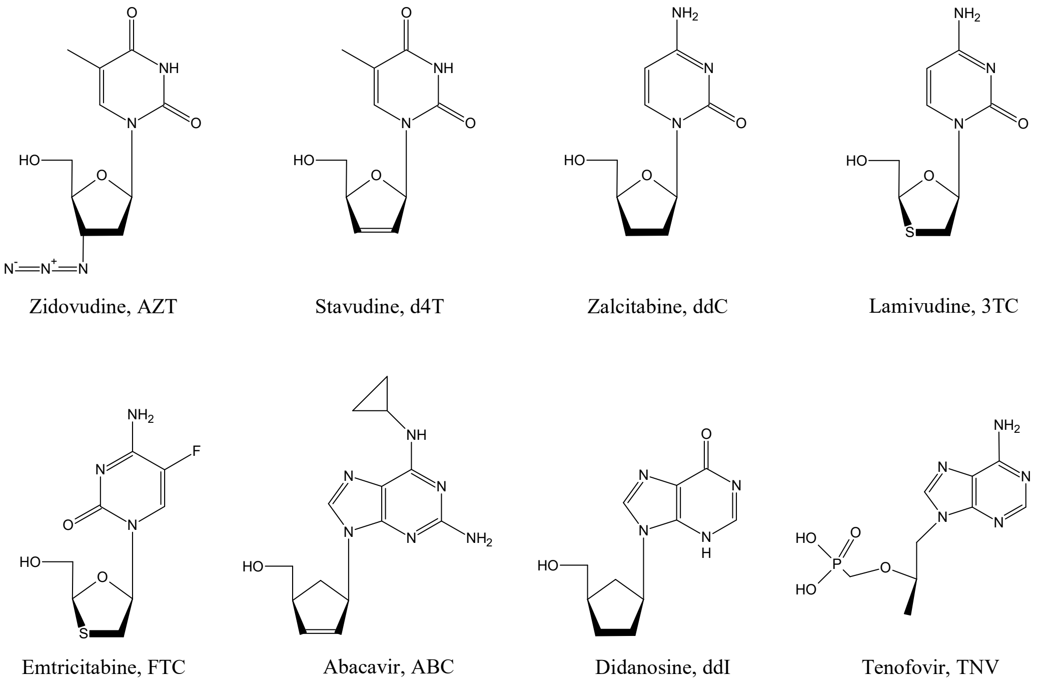
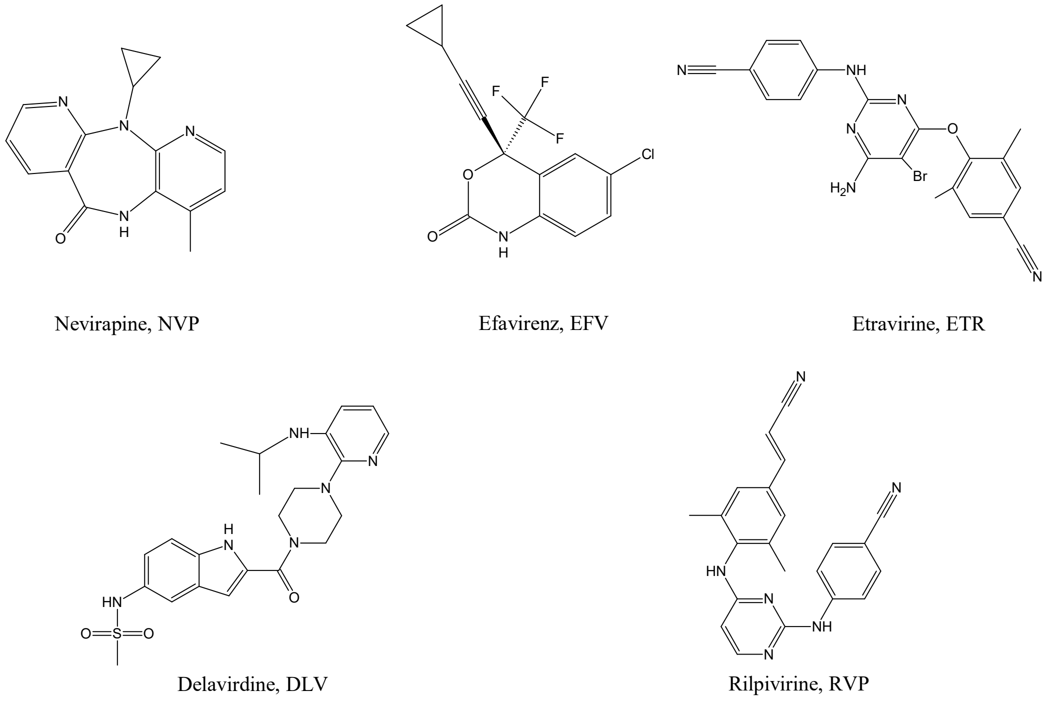


Figure 3: 2D chemical structures of eight approved nucleoside and nucleotide reverse transcriptase inhibitors (N[t]RTI).

In the NRTI class, there are RT inhibitors that already have a phosphate group incorporated into its structure. They are also known as nucleotide RT inhibitors (NtRTIs), such as tenofovir (TFV) (Figure 3), formulated as tenofovir disoproxil fumarate (TDF, Viread®), they require only two phosphorylation steps to achieve their active triphosphate derivatives (Squires 2001). However, their mode of action is the same as for the NRTIs.

The NNRTIs are allosteric inhibitors of DNA polymerization. These compounds bind in a non-competitive manner to a hydrophobic pocket (Figure 2-a) located approximately 10 Å away from the polymerase active site, causing conformation changes that impair DNA synthesis (Squires 2001). During the DNA synthesis, the RT fits a "closed" conformation bringing the fingers and thumb subdomains closer to the palm one and allowing the binding of nucleic acids. The presence of an NNRTI leads to an open conformation that restricts the thumb to a hyperextension position, which prevents the polymerization (Asahchop et al. 2012, Das et al. 2012).The currently approved NNRTIs are nevirapine (NVP, Viramune®), efavirenz (EFV, Sustiva®), delavirdine (DLV, Rescriptor®), etravirine (ETR, Intelence®) and rilpivirine (RPV, Edurant®) (Figure 4). The NNBP consists of hydrophobic residues with significant aromatic character (Tyr 181, Try 188, Phe 227, Trp 229, and Tyr 232 of p66) and hydrophilic residues (Lys 101, Lys 103, Ser 105, Asp 192, and Glu 224 of p66, and Glu 138 of p51) (Sluis-Cremer et al. 2004). The solvent accessible entrance is formed by the residues Leu 100, Lys 101, Lys 103, Val 179, Tyr 181, and Glu 138 (Figure 2-a). However, this open state of the binding pocket is only noticeable when the structure is co-crystallized with NNRTIs, mainly due to significant torsional shifts of the Y181 and Y188 residues to accommodate the ligand (Hsiou et al. 1996). In the absence of a ligand, the binding pocket is blocked given that the side chains of Tyr 181 and Try 188 are situated at the hydrophobic core, representing a closed state of the pocket. Previous docking studies showed that the difference in geometries can affect the accuracy of ligand binding energies when docking other NNRTIs into the inhibitor binding pocket (Titmuss et al. 1999, Kroeger Smith et al. 1995).



Nevirapine, NVP

Efavirenz, EFV

Etravirine, ETR

Delavirdine, DLV

Rilpivirine, RVP

Figure 4: 2D Chemical structures of five approved non-nucleoside reverse transcriptase inhibitors (NNRTI).

Despite their popularity and the number of drugs already approved for this class, most RT inhibitors have their antiviral potency limited by several factors such as mutations in the binding site, drug-drug harmful interactions, toxicity and long-term complications (Cihlar and Ray 2010, Ho and Hitchcock 1989, Waters et al. 2007, Johnson et al. 2008). Consequently, new inhibitors are being sought out.

1.1.3 Computer-aided drug design methods

Presently, computational methods are a significant part of the drug design process, and this kind of modeling is often denoted as computer-aided drug design (CADD). Computational methods can offer detailed information about the interaction between compounds and targets, increasing the efficiency and lowering the cost of research in several stages of drug discovery (Kirchmair et al. 2011). Choosing the most appropriate computational technique to apply when planning novel drugs depends on the understanding of the target of interest (Jorgensen 2004). So far, various computational methods have been employed to the development of anti-viral drugs (reviewed by Kirchmair et al. (2011) and Wlodawer (2002)). It is noteworthy that some approved drugs, for the treatment of an assortment of diseases, owe their discovery in part to CADD methods (recently reviewed by Sliwoski et al. (2014)). This group includes anti-HIV drugs such as protease inhibitors saquinavir (SQV, Invirase®), ritonavir (RTV, Norvir®) and indinavir (IDV, Crixivan®); integrase inhibitor raltegravir (Isentress®); reverse transcriptase inhibitor rilpivirine (RPV, Edurant®); and fusion inhibitor enfuvirtide (T-20, Fuzeon®).

Computational studies, frequently applied in CADD, such as molecular docking, molecular dynamics (MD), free energy calculations, quantitative structure-activity relationships (QSARs), pharmacophore modeling and ADMET have been performed using the RT and its inhibitors as targets. A successful example of the multidisciplinary effort in drug discovery, when modeling RT inhibitors, is the 2011 FDA-approved NNRTI rilpivirine (RPV). RPV was developed by combining chemical synthesis with broad antiviral screening; bioavailability and safety assessments in animals; and molecular modeling, including analysis of three-dimensional structures and ligand-target relationships by molecular docking (Janssen et al. 2005).

In this work, we applied virtual screening in a library of lead compounds to search for potential RT inhibitors.

1.1.3.1 Molecular docking

Molecular docking can provide a better understanding of the interactions between protein and a ligand. Docking begins by sampling ligands' orientations and conformations within the target binding site (Meng et al. 2011, Yuriev and Ramsland

2013). Afterward, the best poses for each ligand are determined, and the compounds are ranked according to a scoring function (Lahti et al. 2012). One of the earliest docking methods was constructed based on the lock-and-key theory of ligand-protein binding, where both the protein and ligand structures are treated as rigid bodies (Kuntz et al. 1982).

Currently, the most popular docking programs account for ligand flexibility when binding to rigid targets, such as Autodock (Goodsell et al. 1996), DOCK (Ewing et al. 2001), FlexX (Kramer et al. 1999), Glide (Friesner et al. 2004), GOLD (Verdonk et al. 2003), and Surflex (Spitzer and Jain 2012), to name a few. Numerous studies (Titmuss et al. 1999, Zhou et al. 2002, Ivetac and McCammon 2011, Ragno et al. 2005, Sherman et al. 2006) have reported the use of molecular docking, by itself or in combination with other molecular modeling techniques, upon HIV-1 RT. A very promising use of molecular docking technique is the screening of compounds databases.

1.1.3.2 Virtual screening

Virtual screening (VS) is an important drug discovery tool, which allows identification of lead compounds among large databases, thanks to its ability to discriminate between true and false positives (Cummings et al. 2005). Several VS approaches have been described, among which the most common one uses molecular docking as a fast and more cost-effective alternative than experimental high-throughput screening (HTS). VS aims to reduce a vast virtual library of approximately 10^5-10^6 chemical compounds, to a more manageable number for the screening of compounds for biological targets and further synthesis of analogs, which could lead to potential drug candidates.

1.2 AIMS

1.2.1 Main goal

To search low weight molecules capable of acting as lead candidates and determine their possible binding modes from a compound library in the HIV-1 RT enzyme through the use of a combination of computational methods.

1.2.2 Particular goals

- To assess DOCK's algorithm and primary score function performance in the NNBP using pose reproduction in multiple crystal structures.
- To investigate suitable RT crystal structures for virtual screening of potential drug candidates through enrichment methods.
- To perform screening of a compound library using molecular docking.
- To cluster molecular docking outcomes into discrete similarity groups using the ChemmineR package.
- To examine molecular docking outcomes in the NNBP.

1.3 MATERIAL AND METHODS

1.3.1 Crystal structures selection

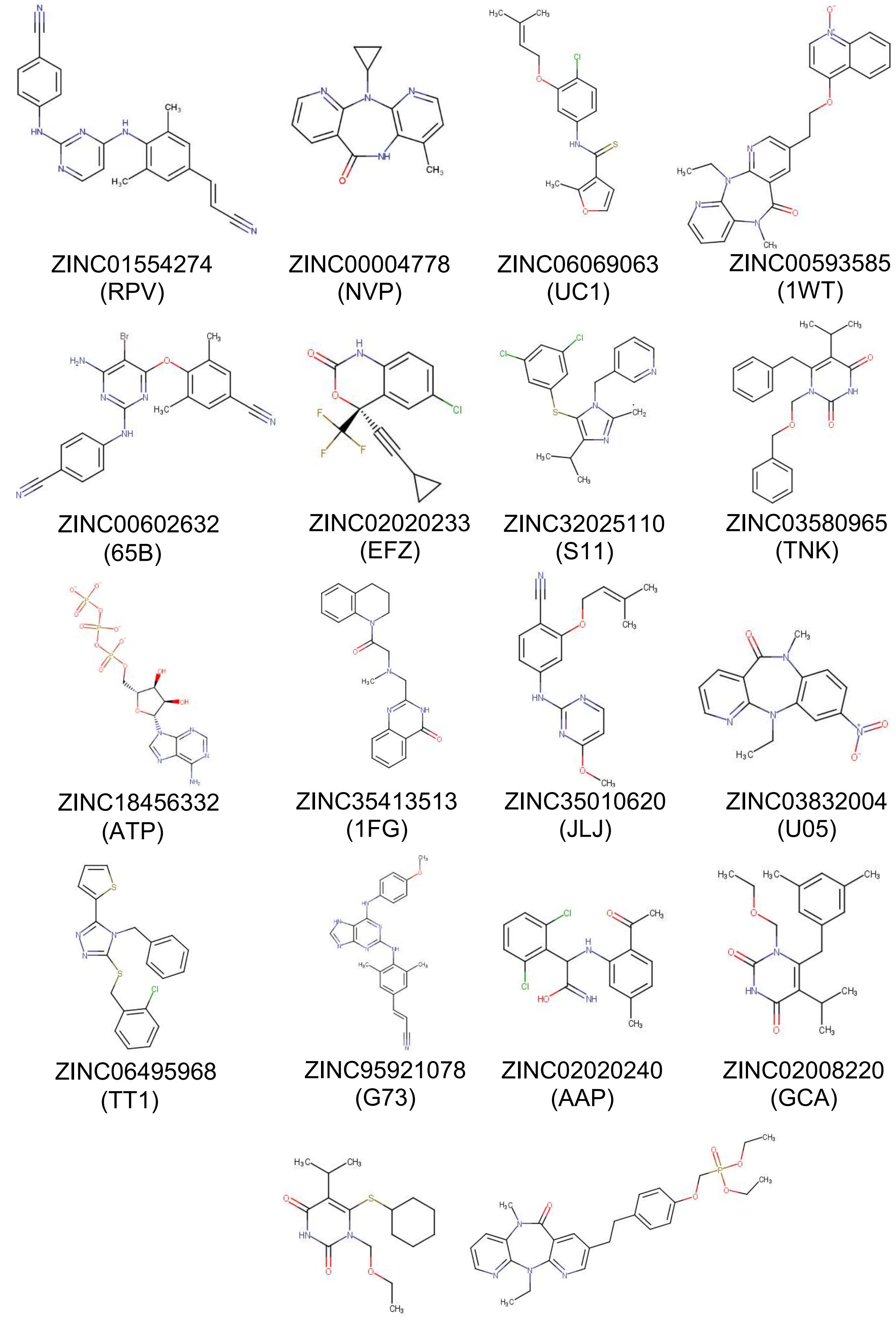
Out of hundreds of RT structures available in PDB (as of May 2014), thirty-three with X-ray resolutions up to 2.5 Å and in complex with an NNRTI were filtered. The structure with the highest resolution (1.5 Å), PDB code 4G1Q (Bauman et al. 2013), was chosen as the reference structure to analyze the NNRTI binding site and surroundings of the different structures. Previous studies defined that residues: Leu 100, Lys 101, Lys 103, Ser 105, Val 179, Tyr 181, Tyr 188, Asp 192, Glu 224, Phe 227, Trp 229, Tyr 232, and Glu 138, form the NNBP. Since conformational changes induced by inhibitors are common in the binding site, the root-mean-square deviation (RMSD) of each NNBP residue was calculated between the reference and the structures was calculated to identify the ones with distinct variation. After this analysis, ten structures were chosen for docking studies: 4G1Q, 1C1C (Hopkins et al. 1999), 1VRT (Ren et al. 1995), 3MEC (Lansdon et al. 2010), 3MEE (Lansdon et al. 2010), 4I7F (Parrish et al. 2013), 4IG0 (Bauman et al. 2013), 4KV8 (LaPlante et al. 2013), 1FK9 (Ren et al. 2000) and 1EP4 (Ren et al. 2000).

1.3.2 Known ligands and decoys compounds

A total of 18 known NNRTIs (including the ligands bound to the selected structures) (Figure 5) were combined with 980 DUD-e-like decoys (Mysinger et al. 2012) into a compound set drawn from ZINC (Irwin et al. 2012). This ZINC feature, the so-called decoy maker, selects, for each ligand, at least 50 compounds with similar properties (molecular weight, logP, net charge, H-bond donors, and H-bond acceptors) to those of the known ligands though chemically different to act as decoys. Bond orders, stereochemistry, hydrogen atoms, and protonation states of the known ligands and decoy set were preserved as assigned by ZINC. Properties of the 18 known ligands, taken from ZINC, are displayed in Table 1. However, the ligands from the ten selected crystal structures were also prepared with Chimera and assigned AMBER ff12SB (Maier et al. 2015) partial charges to perform redocking and cross-docking procedures only.

Table 1: Properties of 18 known NNRTI taken from ZINC.

ZINC ID	Ligand	Heavy atoms	pH range	xlogP	Apolar desolvation (kcal/mol)	Polar desolvation (kcal/mol)	H-bond donors	H-bond acceptors	Net charge	tPSA (Å²)	Molecular weight (g/mol)	Rotatable bonds
ZINC01554274	RPV	28	7	5.46	3.57	-12.57	2	6	0	97	366.43	5
ZINC00004778	NVP	20	7	1.39	6.59	-11.74	1	5	0	64	266.30	1
ZINC06069063	UC1	22	7	5.32	9.39	-12.01	1	3	0	34	335.86	6
ZINC00593585	1WT	33	7	2.20	0.16	-29.24	0	8	0	87	441.50	5
ZINC00602632	65B	28	7	5.03	0.49	-10.21	3	7	0	120	435.29	4
ZINC02020233	EFZ	21	7	4.53	7.44	-10.07	1	3	0	38	315.68	1
ZINC32025110	S11	29	7	4.50	8.39	-9.99	2	6	0	83	451.38	8
ZINC03580965	TNK	27	7	4.12	8.72	-13.33	1	5	0	64	364.44	7
ZINC18456332	ATP	31	7	-3.54	-3.61	-227.01	5	18	-3	288	504.16	8
ZINC35413513	1FG	27	7	2.45	7.93	-19.64	1	6	0	69	362.43	4
ZINC35010620	JLJ	23	7	3.92	8.16	-10.25	1	6	0	80	310.36	6
ZINC03832004	U05	22	7	2.37	1.52	-11.35	0	7	0	85	298.30	2
ZINC06495968	TT1	26	7	6.10	1.34	-11.2	0	3	0	30	397.96	6
ZINC95921078	G73	31	7	6.38	13.78	-14.99	3	8	0	112	411.47	6
ZINC02020240	AAP	23	7	3.64	-1.52	-13.98	3	4	0	72	351.23	5
ZINC02008220	GCA	24	7	3.73	-0.26	-11.64	1	5	0	64	330.43	6
ZINC02008218	612	22	7	3.33	6.22	-8.68	1	5	0	64	326.46	6
ZINC95590437	NVE	37	7	3.59	13.34	-22.96	0	9	0	97	524.56	11



ZINC01554274 (RPV)

ZINC00004778 (NVP)

ZINC06069063 (UC1)

ZINC00593585 (1WT)

ZINC00602632 (65B)

ZINC02020233 (EFZ)

ZINC32025110 (S11)

ZINC03580965 (TNK)

ZINC18456332 (ATP)

ZINC35413513 (1FG)

ZINC35010620 (JLJ)

ZINC03832004 (U05)

ZINC06495968 (TT1)

ZINC95921078 (G73)

ZINC02020240 (AAP)

ZINC02008220 (GCA)

ZINC02008218 (612)

ZINC95590437 (NVE)

Figure 5: 2D chemical structure of known NNRTI compounds from ZINC.

1.3.3 Receptor preparation

Receptor structures were prepared with the Dock Prep module in Chimera (Pettersen et al. 2004). This module consists of a graphical interface that performs several tasks such as solvent deletion, repairing of truncated sidechains, deletion of alternate positions, hydrogen addition, partial charges assignment, and Mol2 output files. For all receptor structures, water molecules, ions, and ligand compounds were removed. Propka (Li et al. 2005) was used to examine the correct protonation state of ionizable residues in the proteins at pH=7. Standard receptor residues were assigned AMBER parm99 atomic partial charges (Cornell et al. 1996). Each crystal structure NNBP was identified according to the location of its ligand and prepared separately in several steps necessary to perform DOCK6.6 (Brozell et al. 2012) calculations. The binding site preparation began with the calculation of the solvent accessible surface of each receptor, devoid of hydrogen atoms, using a probe radius of 1.4 Å with the Write DMS module in Chimera. This module provides a DMS file containing dot molecular surfaces, where at each surface point the surface normal vector was computed. The DMS file was used as input to the program SPHGEN (Kuntz et al. 1982), available with standard DOCK distribution. The program SPHGEN produces a negative image of the surface as a set of overlapping spheres from the molecular surface and the normal vectors (Kuntz et al. 1982). Spheres are generated over the entire surface, producing approximately one sphere per surface point (Figure 6). However, a filtering rule is used to keep only the largest sphere associated with each surface atom. The resulting spheres are then clustered using a single-linkage algorithm. Each cluster represents an evagination in the receptor. SPHGEN was assigned to generate spheres with a minimum radius of 1.4 Å and a maximum of 4.0 Å to all surface points of the receptor.

The coordinates of the crystallographic ligand were then used to select a subset of spheres within a radius of 7.0 Å from the ligand with the program sphere_selector, which is also distributed as an accessory with DOCK.

1.3.4 Grid generation

Normally in computational docking, ligand conformations are generated by matching the distances between points in the ligand and accessible points in the receptor. When docking, the ligand conformations are scored using a precalculated grid of energies, in an attempt to allow rapid evaluation of the conformation. To visualize, define the location, and size of the grid, a cubic box around the binding site

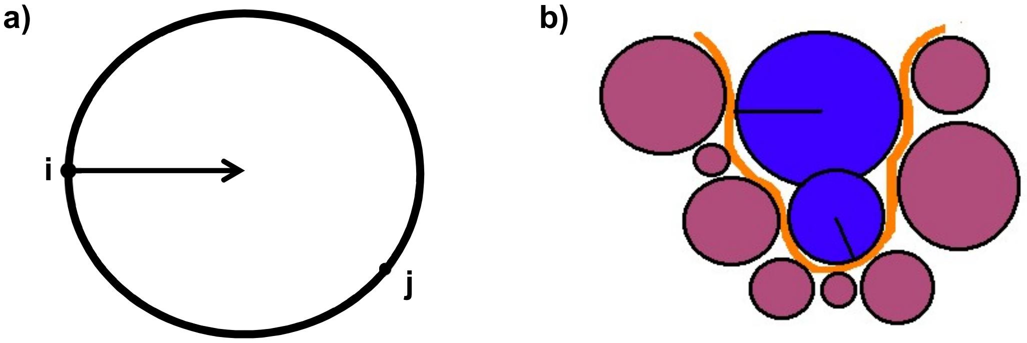


Figure 6: Illustrative representation of sphere generation in DOCK. a) The spheres are generated tangentially to the surface points i, j with the center on the surface normal of point i. (b) Illustrative representation of a small binding pocket formed by eight atoms (purple). The spheres (blue) are generated using points from the molecule. Adapted from: http://dock.compbio.ucsf.edu/Dock_6/tutorials/sphere_generation/generating_spheres.htm.

center of the cubic box was the selected spheres and extra margins of 5 Å in all six directions were chosen. The grids were computed with DOCK's GRID program using a 0.3 Å grid spacing, a 9,999 Å distance cutoff, 6–12 Lennard-Jones exponents, and a 4r distance-dependent dielectric constant.

DOCK has two types of scoring available: contact and energy scoring. In this work Energy scoring was the scoring mode chosen when generating the grids. DOCK's energy scoring component is a type of force field scoring. Force field scores are approximate molecular mechanics interaction energies. In this case, it consists of Lennard-Jones and electrostatic components:

$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left(\frac{A_{ij}}{r_{ij}^{a}} - \frac{B_{ij}}{r_{ij}^{b}} + 332 \frac{q_i q_j}{D r_{ij}} \right) \tag{1}$$

where each term is a double sum over ligand atoms i and receptor atoms j. A bump grid was also calculated with GRID with a van der Waals overlap allowance of 0.75, i.e., if the sum of the van der Waals radii of any two atoms approaches closer than the allowed overlap, the grid point is flagged as a bump and stored in another grid file. The bump grid is used to identify orientations during docking whether a ligand atom is in severe steric overlap with a receptor atom.

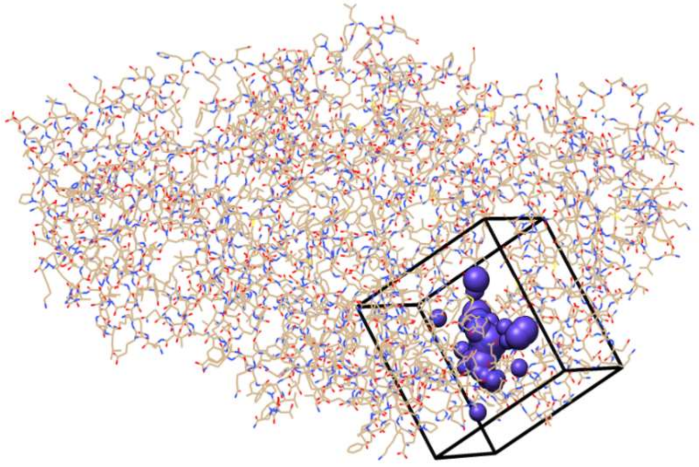


Figure 7: Graphic representation of the cubic box (in black) where the grid is generated.

1.3.5 Molecular docking

Ligand docking was performed in all structures with DOCK6.6. The ligand flexibility sampling algorithm, called anchor-and-grow, is an incremental construction method where the largest rigid substructure of the ligand is firstly recognized (anchor) and then rigidly oriented in the binding site. Later, each layer of flexible bonds is then grown from each cluster, minimized, ranked, and clustered again. This process is repeated until the molecule is fully built. The maximum number of orientations for each ligand was set at 500 and only the best pose, meaning the one with the lowest energy, was retained for each docking run. The grid-based score was chosen as the primary scoring function (For further details, see Appendix A).

The final ligand orientations were submitted to rescore with the Amber Score function (available as a DOCK6.6 scoring function). AMBER score enables not only ligand flexibility but also all or a part of the receptor to be flexible, to reproduce the so-called "induced-fit". Three AMBER score movable region protocols can be employed: ligand, nothing, and distance. For the ligand option, only the ligand is allowed to move during minimization and MD simulation. No minimization or MD simulation occurs for the nothing option, and the ligand is not flexible during the AMBER score energy protocol. The distance movable region option selects residues that are allowed to move by receptor-ligand distance. If any atom in a receptor residue is within the cutoff distance of the ligand, then the whole residue is selected. Amber input files for each receptor, ligand and, corresponding complex were prepared with the help of the Perl script *prepare_amber.pl* provided with DOCK distribution. The script employs other scripts and programs, such as antechamber (Wang et al. 2006) to calculate the AM1-BCC charges (Jakalian et al. 2002) for the ligands, and tLEaP (Case et al.) to assign the parm94 (Cornell et al. 1995) parameter set for protein atoms and the general Amber force field (GAFF) (Wang et al. 2004) parameter set for ligand atoms. We calculated Amber Score energies with the nothing option, since we aimed to rescore with multiple score functions (For further details, see Appendix A).

Docking outcomes were labeled according to the definition of Allen, W. J., et al. (Allen et al. 2015). When the top-scoring pose produced from the docking run was within 2.0 Å RMSD from the crystallographic ligand position, it was named a successful pose. From the successful poses, docking success rate was calculated as

$$Docking\ Success\ (\%) = \frac{Successful\ poses}{Number\ of\ docking\ calculations} \times 100 \tag{2}$$

A number of docking calculations include only runs that produced poses. If a top-scoring pose produced RMSD over the 2.0 Å RMSD threshold, it was named a docking failure. The sum of docking success and docking failures equals 100%.

In addition, a correlation coefficient between RMSD and the scoring functions were computed. A correlation coefficient illustrates the linear relation between two quantitative measures. The Spearman correlation coefficient (R_S) was calculated using the *stats* package of program R. This coefficient defines the correlation between two sets of ranking variables. For a sample of size n, the n, the n raw scores X_i, Y_i are converted to rank x_i, y_i, and R_S is:

$$R_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3}$$

where $d_i = x_i - y_i$, is the difference between ranks.

1.3.6 Ligand RMSD

An initial evaluation of DOCK performance was to employ the process of docking each ligand into the native protein, known as Redocking, and afterward into a non-native RT structure, known as Cross-docking. To both procedures, the heavy atom RMSD between the final ligand pose and corresponding ligand from the crystal structure were calculated. RMSD calculation is a commonly used method for measuring geometric similarity concerning two poses of the same ligand (Allen et al. 2015). DOCK reports three types of RMSD:

(i) Standard heavy-atom RMSD (RMSDs) – standard pair-wise RMSD calculation between non-hydrogen atoms of a reference conformation A and a pose conformation B for a ligand with N total heavy atoms of index i.

$$RMSDs(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|a_i - b_i\|^2} \tag{4}$$

The variables a_i and b_i are Cartesian coordinates of corresponding atoms between the two ligands.

(ii) Minimum-distance heavy-atom RMSD (RMSDm) – this measure is based on the method implemented in Autodock Vina (Trott and Olson 2010) where atom pairings between reference conformation A and pose conformation B are determined by the minimum distance to any atom of the same element type.

$$HA\_RMSDmin(A,B) = max\{HA\_RMSDm(A,B), HA\_RMSDm(B,A)\} \tag{5}$$

$$RMSDm(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \min_{j} \|a_i - b_i\|^2} \tag{6}$$

In this method, a one-to-one atom correspondence is not always preserved, since multiple atoms from one molecule can be matched to a single atom from the other molecule.

(iii) Hungarian (symmetry-corrected) heavy-atom RMSD (RMSDh) – the last RMSD measure is based on the Hungarian algorithm (Kuhn 1955) implementation (Allen and Rizzo 2014). The algorithm solves the optimal assignment between a set of reference conformation A atoms and a set of pose conformation B atoms of the same size. For groups of atoms of the same atom type, a cost matrix M is populated where each matrix element m_{ij} is equal to the distance-squared between reference atom a_i and pose atom b_j.

$$RMSDh(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \operatorname*{cor}_{j} \left\|a_i - b_j\right\|^2} \tag{7}$$

This algorithm is employed to determine one-to-one assignments in a way that the total distance between the atoms from the selected molecules is minimized.

1.3.7 Scoring functions

Three scoring functions were employed in this work: Grid Score, Amber Score, and DrugScore eXtended (DSX). Pose sampling was generated with DOCK6.6 and scored using DOCK's Grid Score. DOCK's outcomes were subsequently rescored with Amber Score and DSX. A combination of the three scoring functions, using consensus scoring (Charifson et al. 1999), were used to reevaluated the top-scoring compounds.

Grid Score is a grid-based energy score and is based on the implementation of force field scoring. It evaluates intermolecular non-bonded van der Waals and Coulombic energies (scaled by a distance-dependent dielectric) between receptor and ligand (Meng et al. 1992). The van der Waals components are generalized to handle any combination of repulsive and attractive exponents that can be carefully chosen by the user at the grid generation part of the molecular docking process.

Amber score is a simple MM-GB/SA approach with the traditional all-atom AMBER force fields (Pearlman et al. 1995) and the GAFF. Electrostatic and van der Waals energy terms represent the interaction between the ligand and the receptor, and the solvation energy is calculated using a Generalized Born (GB) solvation model chosen by the user. The Amber score is calculated as:

$$E_{binding} = E_{complex} - \left(E_{receptor} + E_{ligand}\right) \tag{8}$$

where $E_{complex}$, $E_{receptor}$, and E_{ligand} are, respectively, the internal energies of the complex, receptor, and ligand (all solvated) as approximated by the AMBER force field with the chosen MM-GB/SA solvation term. All or a part of the receptor-ligand complex can be selected to be flexible when using Amber score. However, flexibility increases the computational cost.

DSX is a knowledge-based scoring function that enables scoring (putative) protein-ligand complexes (Neudert and Klebe 2011), whose pair potentials are based on the DrugScore formalism (Gohlke et al. 2000, Velec et al. 2005). Usage of DSX is available by a web-based user interface or by standalone version. This function uses the statistical distribution of non-bonded interactions, obtained from a database of known receptor-ligand complexes, and statistically derived torsional angle potentials, which allow fast relaxation of docking poses that may improve the ranking of a set of ligands according to their binding affinities. The master equation for knowledge based scoring functions is

$$score(i) = -\ln\left(\frac{\rho(i)}{\rho_{ref}}\right) \tag{9}$$

where the fraction $\rho(i)$ is a state-dependent density function, that is also a probability function (Neudert and Klebe 2011). The total score for a given complex of protein atoms a_p and ligand atoms a_l is calculated as

$$total\ score_{pair} = \sum_{a_p} \sum_{a_l} score\left(p(a_p), l(a_l), r(a_p, a_l)\right) \tag{10}$$

$$score_{pair}(p, l, r) = -\ln\left(\frac{\rho(p, l, r)}{\rho_{ref}}\right) \tag{11}$$

where $p(a_p)$, and $l(a_l)$ are the atom types, $r(a_p, a_l)$ is the distance of a_p and a_l, and ρ_{ref} can be seen as a kind of weighting function for $\rho(i)$ (Neudert and Klebe 2011). Most knowledge-based functions are based in Eq. (11), however they differ in the definition of the density functions.

Consensus scoring with the three chosen scoring functions was applied using the rank-by-rank method. This method was chosen since each scoring function assesses protein-ligand complexes from its own perspective, thus absolute score values usually differ from one another (Oda et al. 2006). For instance, if one score is bigger than the others, it will influence the average-based consensus score. Therefore, in a ranking method, the compounds are organized according to the score values, and the average of the ranks calculated from the individual scoring are employed, rather than the raw scores (Zhong et al. 2010).

1.3.8 ROC analysis and enrichment metrics

To investigate the influence of receptor conformation and the methodological predictive power in the ranking of known ligands when faced with a significant number of decoys, enrichment statistics and receiver operating characteristic curves (ROC) were employed. After the docking of the screening library of actives and inactives, ROC method was applied to assess the evolution of known compounds rate, identified as true positive rate (TRP), versus the decoys rate, as the false positive rate (FRP), in the final ranking (Triballeau et al. 2005) with the use of a curve (Figure 8).

TPR is the percentage of truly active compounds being selected from the screening library and is calculated by the number of true positive (TP) results divided by the sum of true positives and false negatives (FN):

$$\frac{TP}{TP + FN} \tag{12}$$

Whereas FPR is the percentage of truly inactive compounds being incorrectly identified as actives and is calculated by one minus the number of true negative results (TN) divided by the sum of true negatives and false positives (FN):

$$1-\left(\frac{TN}{TN+FP}\right) \tag{13}$$

From ROC curves, the performance was evaluated numerically as the area under the ROC curve (AUC), a metric that allows comparing the relative predictive performances between different receptors providing a measure of global enrichment. An ideally perfect performance has an AUC value of 1.0, while a random selection performance has an AUC of 0.5. ROC curves and ROC AUC were calculated with ROCR (Sing et al. 2005) package.

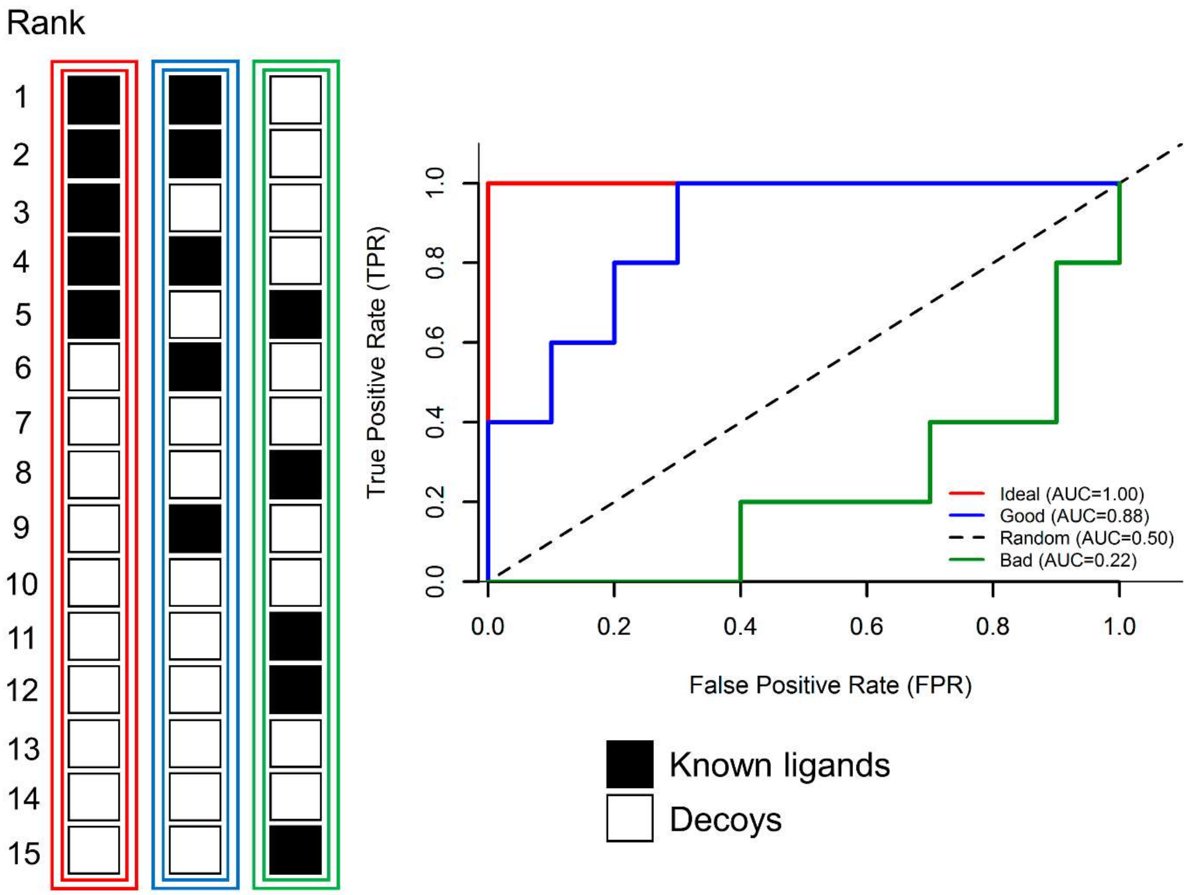


Figure 8: Schematic procedure for ROC curve plotting. Evaluation of five know ligands (black squares) against ten decoys (white squares) in a docking protocol through ROC curves. Three different docking outcomes ranked based on score are described. Perfect enrichment (red box and line) with AUC of 1, good enrichment (blue box and line) with AUC of 0.88, and bad enrichment (green box and line) with AUC of 0.22. The dashed line represents random enrichment with AUC of 0.5.

To focus on early enrichment LogAUC (Mysinger and Shoichet 2010), calculated from a ROC curve plot with a base 10 semilog x-axis was also considered. Enrichment is defined as the proportion of the observed fraction of active compounds in the top few percent of a screening essay to that expected by random selection (Jain and Nicholls 2008). A real early enrichment in a screening of a large ligand library might infer that the selection of known compounds is been prioritized in ranking when faced with inactive ones (Figure 9). Enrichment factors (EF) after the $x\%$ of the screening library were calculated as:

$$EF_{x\%} = \frac{Ligand_{x\%} / N_{x\%}}{Ligand_{total} / N_{total}} \tag{14}$$

where, $Ligand_{x\%}$ is the number of known compounds found at $x\%$ of database screened; $N_{x\%}$ is the number of compounds screened at $x\%$ of the database; $Ligand_{total}$ is the number of actives in the entire database, and N_{total} is the number of compounds in the entire database. Enrichment percentages were reported in the 0.5%, 1%, 2% and 5% of the database screened. Maximum enrichment (E_{max}) is given by the total number of active compounds and the total number of compounds in the database.

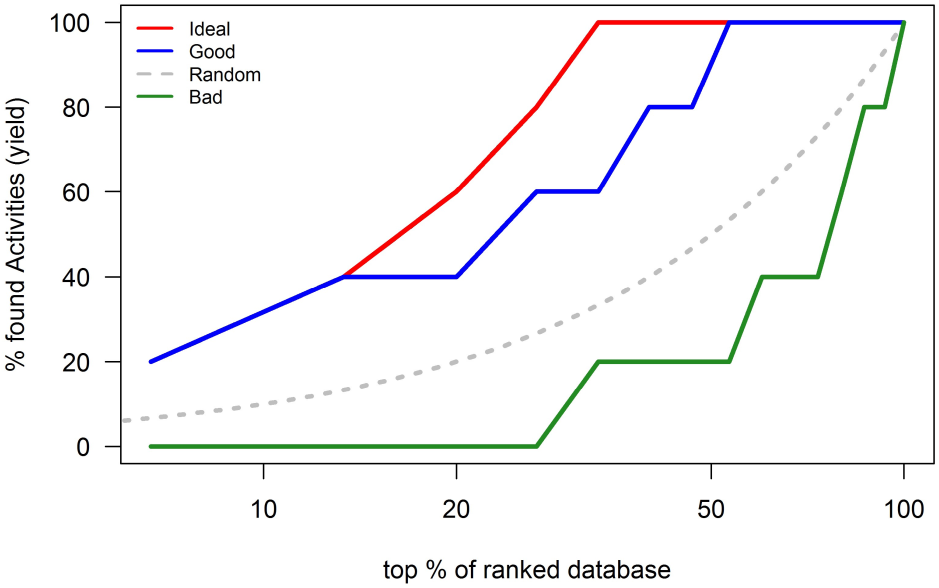


Figure 9: Graphical enrichment curve of the schematic procedure discussed in Figure 8. An ideal run would provide a curve where a high percentage of known ligands are ranked high in a low percentage of the database screened (red). Any run found above the random line (dashed gray line) would provide some good enrichment, whereas under it would show bad enrichment. In a bad enrichment curve is clear to see that decoys are receiving higher scores than known ligands.

1.3.9 Ligand subset

The lead-like now ligand subset from the ZINC database was chosen. ZINC lead-like compounds are large enough to be detected in high-throughput spectrophotometric or other assays, smaller and more soluble than most drugs. At the time of this work, the ZINC lead-like now subset was composed of 2,797,315 compounds and submitted to molecular docking using DOCK6.6. The screening was performed using the reference structure, in a way to filter and restrict the high number of compounds to a more manageable set for further rescoring.

1.3.10 Compound selection

After the docking, the compounds were ranked by their Grid score. The mean, median, 25th and 75th quartiles, min and max, were calculated to provide a better view of the data. Positives scores were discarded. We used the R package ChemmineR (Cao et al. 2008) to cluster the chosen compounds.

ChemmineR is a cheminformatics package for analysing drug-like small molecule data in R, which can perform clustering of compounds into discrete similarity groups among other functions. This mathematical function uses single linkage for cluster joining with multiple cut-offs of a chosen similarity method. Instead of working with the complete molecule, atom pairs descriptors (Carhart et al. 1985) for the compounds in the sample was calculated. Atom pairs are 2D topological descriptors that count the distance between two atoms in the shortest path of bonds. With the calculation of atom pairs, duplicated or identical compounds were removed from the sets. The chosen atom pairs were converted into binary atom pair fingerprints (Chen and Reynolds 2002) of fixed length. In general, fingerprints are binary representations of attributes where each bit denotes the absence or presence of a characteristic in a molecule (Chen and Reynolds 2002). Computations on this compact data structure are more time and memory efficient than on their relatively complex atom pair counterparts. The use of fingerprint permits the use of similarity methods. We have chosen the implemented method that calculates Tanimoto coefficient (Jaccard 1901, Tanimoto 1957) as a similarity method. Tanimoto coefficient for dichotomous variables is

$$T_{A,B} = \frac{c}{a + b - c} \tag{15}$$

where a is the number of on bits in molecule A, b is the number of on bits in molecule B, while c is the number of bits that are on in both molecules (Bajusz et al. 2015). The Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones. For instance, the same molecule compared with itself will produce a Tanimoto coefficient of 1. However, a Tanimoto coefficient of 1 does not automatically imply that two compounds are identical (Backman et al. 2011). It suggests an identical structural descriptors or identical on-bits between them, since features of the two molecules that have not been determined in the fingerprint could be different (Backman et al. 2011).

The compounds resulted from clustering were submitted to toxicity filtering using specific organic toxic roots using the ChemBioServer (Athanasiadis et al. 2012) The ChemBioServer hosts a group of tools aimed to facilitate computational compound screening and analysis. The final compounds were ranked according to rescoring using the chosen score functions and visually analysed with Chimera.

1.4 RESULTS AND DISCUSSION

1.4.1 Identifying a structure

RT popularity as a drug target, confirmed by a large number of crystal structures available in the Protein Data Bank (Berman et al. 2000), is of considerable aid to the computational discovery of new RT inhibitors. This structural diversity provides information about the binding pocket and binding mode of its inhibitors. Taking advantage of these structures, an initial validation protocol was applied to retrospectively evaluate the screening performance of each high-resolution structure in VS.

The protocol started with an initial search for RT structures in the PDB database, where over one hundred crystal structures were found. Most RT crystal structures found in the PDB are complexed with NNRTIs, with only seven structures bound to NRTIs. From this result, we filtered twenty-three structures with X-ray resolutions up to 2.5 Å, bound to an NNRTI compound, and with no mutations in the residues composing the NNBP. Also, two structures with unbound NNRTIs were included. As a reference, we chose the structure (PDB code 4G1Q) with the lowest resolution (1.5 Å), bound to the known NNRTI, RPV. The reference structure was used to compare the differences between the structures NNBP. The heavy atom RMSD between the NNBP residues of the reference and the structures was analyzed (Table 2). As it was expected, a slight variation in the RMSD could be noticed when comparing the reference with structures bound to compounds other than RPV (Figure 10). In these structures, the residues Tyr 181 and Tyr 188 appeared to be the ones with the largest deviation. It is worth mentioning that the unbound structures showed the highest RMSD deviation between all residues of the reference structure. However, the lack of bound compounds in these structures led to a semi-closed conformation of the NNBP (Figure 11). Out of the twenty-five RT structures, we selected a manageable group of ten structures to perform redocking and cross-docking experiments using DOCK 6.6. The selection criterion was based on the differences in the heavy atom RMSD values of the NNBP residues from each structure when compared to the reference structure NNBP residues. The group consisted of two structures bound to RPV (the reference, and 3MEE) and eight bound to different NNRTIs (1C1C, 1VRT, 3MEC, 4I7F, 4IG0, 4KV8, 1FK9, and 1EP4).

Redocking outcomes were considered successful (docking success rate of 80%) and can be seen in Table 3 diagonal elements. Out of the ten ligands evaluated, only two presented top-scoring poses over 2.0 Å RMSD from the crystallographic position (Figure 12). For the ligands S11 and NVE, from 1EP4 and 4I7F structures respectively, the predicted poses displayed part or the entire compound outside of the NNBP. These results may be due to limitations in DOCK's anchor-and-grow algorithm since the algorithm has been validated only for binding mode prediction on sets of ligands with no more than seven rotatable bonds (Moustakas et al. 2006). S11 and NVE have eight and eleven rotatable bonds, respectively.

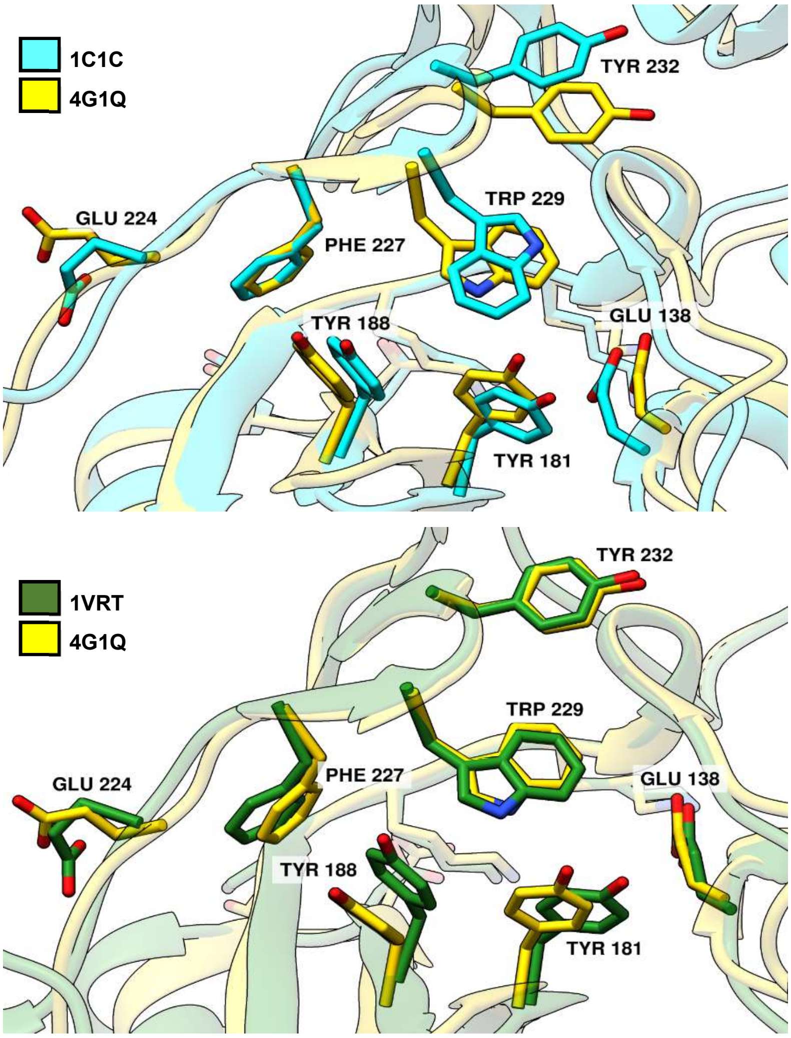


Figure 10: Comparison of the reference structure-binding site, 4G1Q (yellow), between structures 1C1C (cyan) and 1VRT (green).

Table 2: Heavy atom RMSD between the residues of the reference structure, 4G1Q, and the other structures.

		RMSD (Å)												
PDB	**Ligand**	**Leu**	**Lys**	**Lys**	**Ser**	**Val**	**Tyr**	**Tyr**	**Asp**	**Glu**	**Phe**	**Trp**	**Try**	**Ser**
		100	**101**	**103**	**105**	**179**	**181**	**188**	**192**	**224**	**227**	**229**	**232**	**138**
1C1B	GCA	0.76	1.60	0.41	0.79	1.52	1.58	1.50	0.41	1.88	0.45	0.87	1.73	1.84
1C1C	612	1.14	1.77	0.42	0.91	1.49	1.73	1.65	0.55	1.75	0.44	2.65	2.20	2.22
1EP4	S11	0.86	1.64	1.72	1.45	0.98	1.52	1.74	1.86	1.81	0.88	0.91	1.60	1.54
1FK9	EFZ	0.69	0.96	0.65	1.10	1.41	1.99	1.68	0.74	1.97	0.81	0.58	1.46	1.18
1RTH	U05	0.59	1.21	0.60	1.11	1.48	0.79	1.58	0.27	1.82	1.03	0.48	1.85	1.29
1RTJ	--	1.71	1.90	1.45	0.79	1.67	5.06	5.59	1.11	2.47	1.25	3.54	1.29	2.84
1VRT	NVP	0.29	0.74	0.72	0.82	1.39	1.08	2.09	0.90	1.94	0.81	0.80	0.33	0.75
1VRU	AAP	0.48	1.02	0.74	0.81	1.25	1.00	1.96	0.65	1.72	0.63	0.80	0.44	2.67
2RKI	TT1	0.77	1.72	1.74	0.93	1.02	0.93	2.01	0.75	1.93	0.58	0.81	0.65	1.82
2ZD1	RPV	0.13	0.10	0.21	0.22	0.27	0.19	1.39	0.25	1.58	1.46	0.13	1.41	0.14
3DLK	--	1.35	1.68	1.41	1.68	2.41	4.99	5.58	2.49	5.29	3.09	2.08	2.58	2.54
3MEC	ETR	0.39	1.33	0.55	0.86	1.41	1.80	2.73	1.45	3.19	1.43	0.93	0.68	0.76
3MEE	RPV	0.27	0.55	0.67	0.67	1.41	1.09	1.92	0.59	2.18	0.95	0.54	0.71	0.40
4I2P	G73	0.36	0.60	0.63	0.34	0.19	0.30	1.41	0.72	1.69	1.45	0.21	1.39	0.57
4I7F	NVE	0.45	1.45	0.98	0.84	1.16	1.22	1.98	0.75	2.05	0.90	0.47	0.72	1.32
4ICL	RPV	0.05	0.05	0.09	0.11	0.08	0.14	0.13	0.09	0.16	0.08	1.20	0.07	0.11
4ID5	RPV	0.07	0.07	0.12	0.18	0.11	0.09	1.38	0.21	0.44	0.18	0.10	0.07	0.12

PDB	**Ligand**	**RMSD (Å) (cont.)**												
		Leu	**Lys**	**Lys**	**Ser**	**Val**	**Tyr**	**Tyr**	**Asp**	**Glu**	**Phe**	**Trp**	**Try**	**Ser**
		100	**101**	**103**	**105**	**179**	**181**	**188**	**192**	**224**	**227**	**229**	**232**	**138**
4IDK	RPV	0.09	0.11	0.16	0.11	0.14	0.12	0.13	0.29	1.49	0.16	0.17	0.17	0.18
4IFV	RPV	0.08	0.11	0.16	0.20	0.23	0.13	0.23	0.33	0.95	0.16	0.17	0.17	0.17
4IFY	RPV	0.07	0.09	0.12	0.16	0.14	0.15	0.09	0.25	1.10	0.20	0.15	0.12	0.14
4IG0	1FG	0.59	0.41	0.90	1.01	1.15	1.58	1.65	0.69	3.02	0.99	0.76	0.67	0.51
4IG3	RPV	0.10	0.17	0.32	0.43	0.42	0.21	0.29	0.36	3.12	0.30	0.16	0.27	0.27
4KFB	RPV	0.06	0.14	0.17	0.11	0.28	0.39	1.39	0.20	0.37	0.15	0.15	0.15	0.41
4KO0	JLJ	0.49	0.32	0.23	0.55	0.93	0.95	1.21	0.40	0.69	0.34	0.59	0.38	0.71
4KV8	1WT	0.45	1.65	1.03	0.80	1.24	1.19	1.99	1.18	2.27	0.98	0.47	0.61	0.95

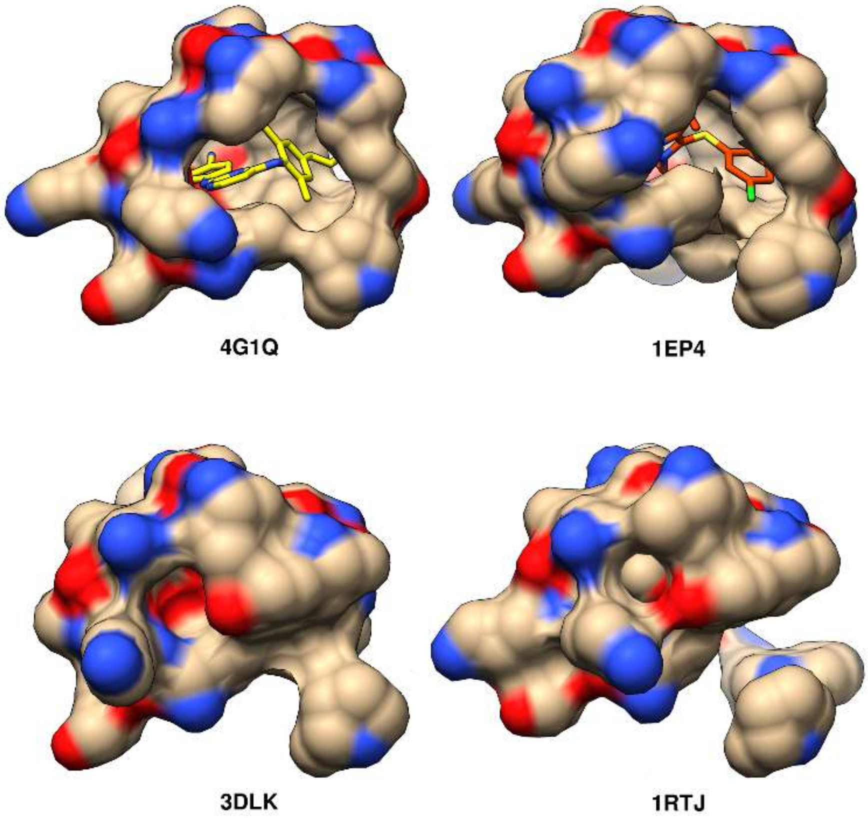


Figure 11: Surface representation of four different NNBP. An open conformation is noticeable in the presence of a bound NNRTI. Structures without a bound NNRTI have a closed conformation of the NNBP.

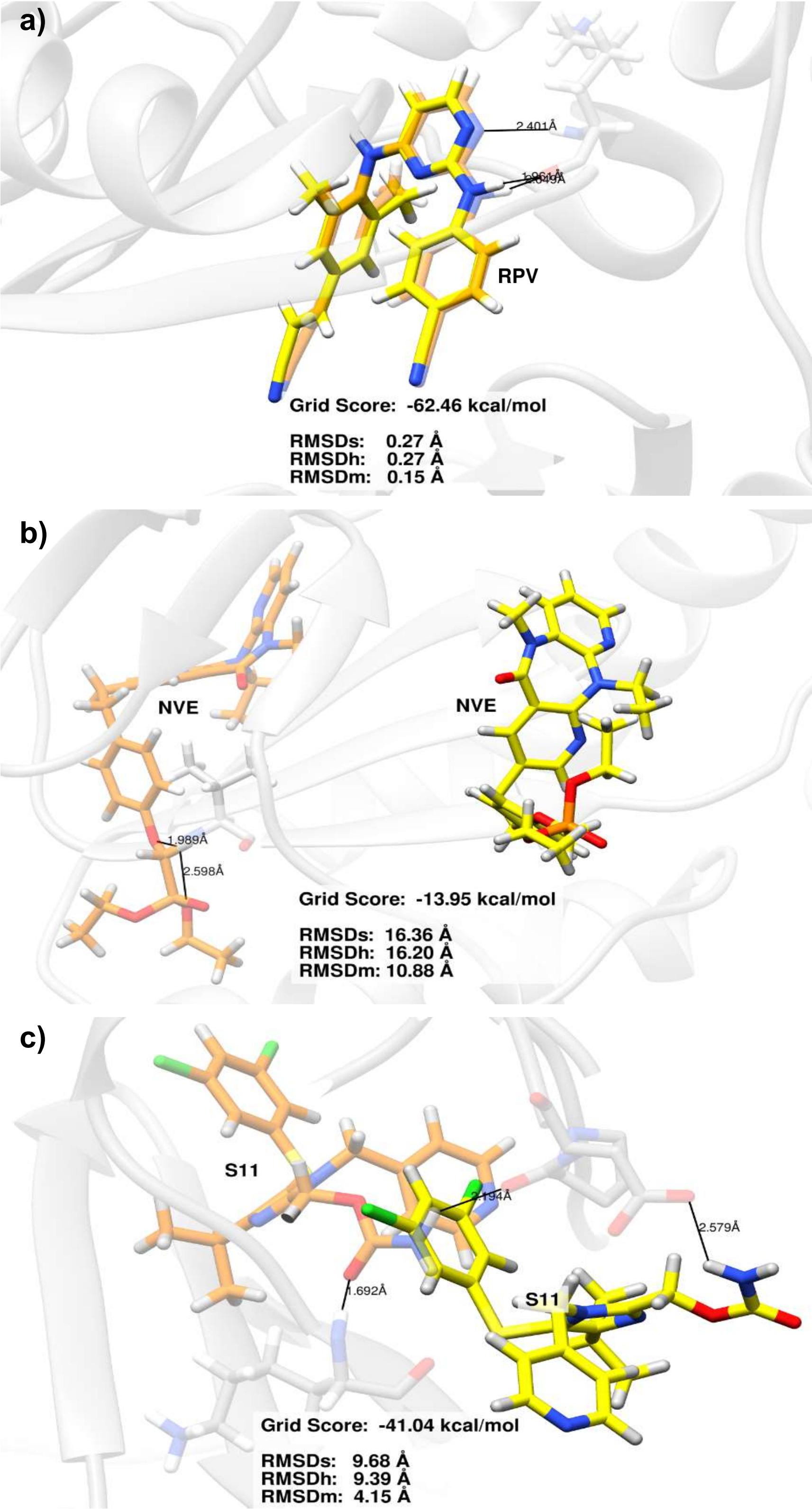


Figure 12: Docking poses compared to the crystallographic position of the ligand in the binding site. Docking pose outcomes are showed in yellow against crystallographic ligands in orange from crystal structures. a) Good pose reproduction was reached when redocking the ligand from the reference structure, 4G1Q. b) The crystallographic position could not be reproduced in the case of the 4I7F structure. c) Part of the docking pose was placed in the NNBP, however pose reproduction could not be reached from structure 1EP4.

Cross-docking results did not show the same precision (docking success rate of 23%), as seen in Table 3 off-diagonal elements. For some structure-ligand pairs, no ligand pose was found. Since this problem was extended to other ligands besides S11 and NVE, the rotatable bond algorithm behavior could not be entirely at fault. The induced fit effects that the NNBP suffers when ligand binding happen could be the source of cross-docking fail. In some cases, the poses have been adjusted not in the NNBP, resulting in a large RMSD value, when docking into the non-native structure. However, a few poses under 2.0 Å RMSD from the aligned ligand position were obtained (Figure 13). A small improvement in RMSD values was obtained when symmetry-corrected RMSD is taken into account (docking success rate of 26%) as seen in Table 4.

Table 3: Standard heavy-atom RMSD (RMSDs) matrix. Matrix diagonal represents Redocking RMSDs, whereas off-diagonal elements represent cross-docking RMSD values.

	RMSDs (Å)									
	4G1Q	**1C1C**	**3MEC**	**3MEE**	**1EP4**	**1FK9**	**1VRT**	**4IG0**	**4I7F**	**4KV8**
RPV	**0.3**	14.6	2.2	2.0	15.1	1.9	4.6	10.2	13.7	13.5
612	--	**0.9**	14.6	--	15.2	5.7	4.4	13.8	13.5	--
ETR	12.0	13.4	**0.3**	1.9	--	5.4	12.2	13.8	16.5	13.9
RPV	6.8	6.3	1.4	**1.8**	16.9	1.5	4.0	9.4	11.2	9.3
S11	4.5	5.6	11.4	8.0	**9.7**	5.5	1.3	6.1	3.1	1.5
EFZ	7.5	13.0	15.5	7.9	--	**0.8**	0.5	5.0	15.1	--
NVP	7.1	1.2	2.2	7.2	13.6	1.1	**0.5**	9.0	13.9	12.2
1FG	2.2	5.3	0.8	8.1	6.3	5.7	1.6	**0.5**	12.6	9.2
NVE	8.0	1.3	--	7.8	15.5	1.5	1.8	6.6	**16.4**	0.3
1WT	8.2	1.2	--	6.1	15.9	1.1	1.3	6.2	6.1	**0.3**

Table 4: Symmetry-corrected heavy-atom RMSD (RMSDh) matrix. Matrix diagonal represents Redocking RMSDs, whereas off-diagonal elements represent cross-docking RMSD values.

	RMSDh (Å)									
	4G1Q	**1C1C**	**3MEC**	**3MEE**	**1EP4**	**1FK9**	**1VRT**	**4IG0**	**4I7F**	**4KV8**
RPV	**0.3**	14.3	1.5	1.1	14.9	1.8	3.8	8.3	13.5	13.3
612	**--**	**0.5**	14.5	**--**	15.2	5.2	3.7	13.2	13.4	**--**
ETR	11.9	12.8	**0.3**	0.9	**--**	5.0	12.2	13.3	16.0	13.7
RPV	5.9	4.3	1.4	**0.8**	16.3	1.1	3.9	7.3	10.9	8.9
S11	4.5	4.0	10.4	7.8	**9.4**	5.2	1.3	5.4	2.7	1.5
EFZ	6.5	12.7	15.0	6.6	**--**	**0.5**	0.5	4.8	13.5	**--**
NVP	6.2	1.2	2.2	6.3	13.4	0.7	**0.5**	6.8	13.6	12.1
1FG	1.5	3.9	0.8	6.5	5.1	5.3	1.6	**0.5**	10.9	8.8
NVE	7.6	1.1	**--**	7.3	15.2	1.3	1.8	5.4	**16.2**	0.3
1WT	7.7	0.9	**--**	5.7	15.6	1.1	1.3	5.5	6.0	**0.3**

Table 5: Minimum-distance heavy-atom RMSD (RMSDm) matrix. Matrix diagonal represents Redocking RMSDs, whereas off-diagonal elements represent cross-docking RMSD values.

	RMSDm (Å)									
	4G1Q	**1C1C**	**3MEC**	**3MEE**	**1EP4**	**1FK9**	**1VRT**	**4IG0**	**4I7F**	**4KV8**
RPV	**0.1**	11.2	1.0	0.8	9.1	1.0	0.8	4.1	9.6	9.6
612	**--**	**0.3**	8.6	**--**	10.9	3.1	1.6	9.9	9.9	**--**
ETR	6.0	8.6	**0.3**	0.4	**--**	2.7	8.5	8.6	10.5	8.9
RPV	0.5	0.9	0.6	**0.4**	9.6	0.8	1.9	3.2	6.5	5.3
S11	1.0	0.5	6.3	2.1	**4.1**	2.4	0.3	1.0	0.8	0.8
EFZ	1.2	8.5	8.9	0.9	**--**	**0.3**	0.2	2.5	7.7	**--**
NVP	1.0	0.5	0.8	0.8	7.5	0.3	**0.3**	3.8	8.0	8.3
1FG	0.7	1.2	0.2	1.2	1.5	2.7	0.8	**0.3**	3.5	5.3
NVE	2.8	0.7	**--**	2.1	10.6	1.0	1.0	1.5	**10.9**	0.2
1WT	2.8	0.6	**--**	0.8	9.8	0.7	0.7	0.8	1.9	**0.3**

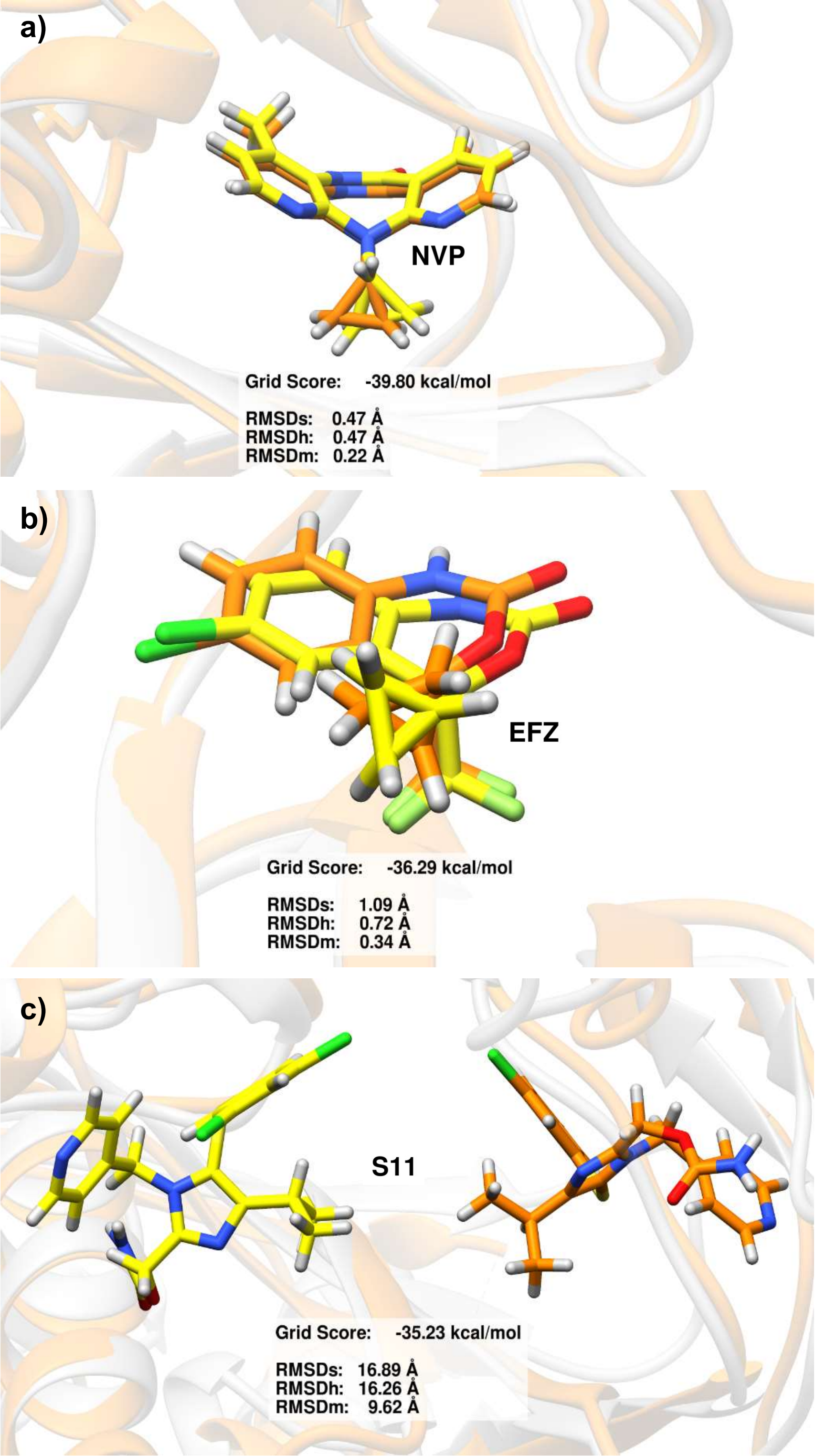


Figure 13: Examples of the cross-docking experiments in the NNBP. Docking outcome are shown in yellow. Native (grey) and non-native structure (orange) are superimpose to better visualization. a) Good reproduction of NVP in the non-native structure 1FK9. b) An instance where an improvement of RMSD was made by the symmetry-correlated RMSD. Pose of EFZ in the non-native structure 1VRT. c) A cross-docking failure where the pose could not be achieved in the NNBP and RMSD was higher than the threshold. Pose outcome of S11 in the non-native structure 3MEE.

Minimum-distance RMSD also corrects ligand symmetry leading to poses values equal to or less than the standard RMSD and symmetry-corrected RMSD (Table 5). When RMSDm was chosen, the RMSD results were improved (docking success rate of 50%). However, symmetry-corrected RMSD is preferred when one-to-one correspondence between pose and reference atoms is evaluated (Allen et al. 2015). The overall docking success rate for RMSDs was 29%, for RMSDh was 32%, and for RMSDm was 53%.

A rmsd-score correlation for each ligand using the primary score function was calculated using the Spearman correlation coefficient. A good rmsd-score correlation does not make it certain that better success rates can be achieved to identify the correctly docked conformations. However, it could imply that the score function might find a global minimum when conformation sampling is performed (Wang et al. 2003). Scoring with DOCK’s primary score function, Grid Score, is showed in Table 6. Grid Score did not give a good rmsd-score correlation in all three types of RMSD. RMSDs reached a $\mathrm{R_S} = 0.34$, RMSDh was $\mathrm{R_S} = 0.36$, and RMSDm achieved a slight improvement with $\mathrm{R_S} = 0.40$ (Figure 14). In theory, scoring functions are expected to recognize the correct binding pose for compounds and rank them to separate ligands from non-ligands (Zhong et al. 2010). Moreover, scoring functions may perform better on some classes of proteins than on others (Neudert and Klebe 2011). An alternative to overcome individual scoring functions limitations is the use of multiple scoring functions upon rescoring (Wang and Wang 2001).

Amber Score and DSX were applied to docking outcomes. The submission of docking outcomes to another score function is called rescoring. Each rescoring of docking outcomes is a consensus scoring since it is a combination of the primary scoring function used in docking and the function applied subsequently (Neudert and Klebe 2011). Rescoring with DOCK’s score function, Amber Score, is showed in Table 7. Amber Score is a more advanced force-field based score function than Grid Score. However, an increase in interaction energy between the ligand and receptor in the majority of cases was observed. This less negative energy was expected since Amber Score is more refined in shielding electrostatics than Grid Score (Lang et al. 2009). Amber Score accomplished better rmsd-score correlation than Grid Score, RMSDs reached a $\mathrm{R_S} = 0.66$, RMSDh $\mathrm{R_S} = 0.68$, and RMSDm $\mathrm{R_S} = 0.67$ (Figure 14). Whereas, DSX rmsd-score correlation was $\mathrm{R_S} = 0.70$ for RMSDs, $\mathrm{R_S} = 0.73$ for RMSDh, and $\mathrm{R_S} = 0.76$ for RMSDm (Figure 14). DSX improvement may be due to the

Table 6: Redocking and cross-docking scoring with DOCK's primary score function Grid Score.

	Grid Score (kcal/mol)									
	4G1Q	**1C1C**	**3MEC**	**3MEE**	**1EP4**	**1FK9**	**1VRT**	**4IG0**	**4I7F**	**4KV8**
RPV	**-62.5**	-26.5	-55.2	-64.3	-33.1	-41.6	-34.0	-39.7	-45.2	-30.6
612	--	**-56.6**	-17.5	--	5.5	-28.2	-20.0	-24.2	-37.8	--
ETR	-33.1	-29.7	**-57.0**	-31.3	--	-37.8	-22.8	-38.4	-37.5	-37.2
RPV	-45.0	-27.9	-48.3	**-53.2**	-35.2	-39.0	-26.7	-44.8	-49.7	-41.5
S11	-48.7	-44.0	-31.1	-46.3	**-41.0**	-33.0	-37.1	-47.1	-58.0	-50.3
EFZ	-39.6	-31.7	-8.2	-37.9	--	**-47.4**	-39.8	-25.7	45.5	--
NVP	-40.9	-28.8	-9.3	-45.2	-39.0	-36.3	**-47.3**	-43.7	-49.6	-27.7
1FG	-44.4	-50.2	-44.4	-47.0	-51.3	-39.4	-39.0	**-57.9**	-62.1	-48.4
NVE	-38.0	-37.5	--	-41.7	-33.0	-37.0	-33.3	-25.2	**-14.0**	-30.2
1WT	24.3	-36.1	--	-2.4	-30.8	-33.7	-37.2	-51.3	3.4	**-69.1**

Table 7: Redocking and cross-docking rescoring with DOCK's score function Amber Score.

	Amber Score (kcal/mol)									
	4G1Q	**1C1C**	**3MEC**	**3MEE**	**1EP4**	**1FK9**	**1VRT**	**4IG0**	**4I7F**	**4KV8**
RPV	**-55.6**	-12.3	-40.1	-53.6	-17.2	-33.3	-26.4	3.0	-12.9	-5.2
612	--	**-47.2**	0.1	--	52.9	-18.4	-2.7	18.4	12.0	--
ETR	-4.5	-10.5	**-48.3**	-25.8	--	-22.5	-2.5	-17.5	-9.2	-11.9
RPV	-23.9	-7.8	-32.8	**-45.4**	-4.5	-25.3	-7.6	5.8	-3.1	-19.1
S11	-29.9	-27.0	-15.8	-30.8	**-24.3**	-22.7	-26.9	-18.0	-33.4	-34.2
EFZ	-17.4	-10.8	6.0	-8.9	--	**-35.4**	-35.1	41.0	-15.7	--
NVP	-26.8	-9.0	1.0	-20.6	-13.3	-27.7	**-39.2**	1.5	-17.1	-10.2
1FG	-40.3	-33.8	-30.8	-29.1	-22.3	-23.8	-28.6	**-38.0**	-32.6	-32.3
NVE	-22.2	-30.3	--	-24.0	-1.6	-25.9	-24.9	4.0	**12.2**	-20.0
1WT	34.0	-25.2	--	15.2	5.0	-20.8	-28.9	-22.2	46.8	**-54.6**

Table 8: Redocking and cross-docking rescoring with knowledge-based score function DSX.

	DSX (a.u.)									
	4G1Q	**1C1C**	**3MEC**	**3MEE**	**1EP4**	**1FK9**	**1VRT**	**4IG0**	**4I7F**	**4KV8**
RPV	**-149**	-44	-141	-151	-52	-112	-89	-74	-96	-74
612	**--**	**-141**	-44	**--**	-13	-102	-103	-49	-86	**--**
ETR	-62	-60	**-151**	-144	**--**	-116	-51	-86	-70	-75
RPV	-113	-117	-144	**-143**	-72	-103	-102	-98	-135	-101
S11	-131	-126	-69	-111	**-101**	-93	-95	-130	-151	-155
EFZ	-110	-57	-15	-105	**--**	**-128**	-106	-102	-81	**--**
NVP	-106	-127	-109	-108	-71	-108	**-115**	-107	-95	-51
1FG	-121	-127	-135	-108	-138	-106	-95	**-142**	-138	-123
NVE	-120	-119	**--**	-113	-46	-105	-99	-127	**-41**	-164
1WT	-93	-116	**--**	-116	-49	-98	-107	-134	-121	**-183**

scoring function formulation. DSX was designed to complement functions used in docking, since it does not calculate binding energies but relies on probabilities for given geometries. Rescoring with DSX function is shown in Table 8. Rescoring with both Amber Score and DSX seemed to complement Grid Score outcomes, and it was carried out to all docking experiments throughout this work.

Besides the rmsd-score correlation of the individual score functions, we investigated the best combination to perform consensus scoring with the primary score function Grid Score (Figure 15). Statistically, consensus scoring might be more robust and accurate than using a single scoring approach (Wang and Wang 2001). All combinations of consensus scoring showed superior rmsd-score correlation than Grid Score itself. This fact might indicate a complementarity between the selected score functions. Consensus score using the combination Grid Score and Amber Score presented correlations close to those of Amber Score itself. Correlations were $\mathrm{R_S} =$ 0.54 for RMSDs, $\mathrm{R_S} = 0.56$ for RMSDh, and $\mathrm{R_S} = 0.58$ for RMSDm (Figure 15). Consensus score using the combination Grid Score and DSX showed better correlation than the combination Grid Score and Amber Score, but worse than DSX alone. Correlations were $\mathrm{R_S} = 0.57$ for RMSDs, $\mathrm{R_S} = 0.60$ for RMSDh, and $\mathrm{R_S} = 0.65$ for RMSDm (Figure 15). Correlation with all three score functions combined had the best performance of all consensus score and it was only slightly below the correlation of DSX alone. Correlations were $\mathrm{R_S} = 0.64$ for RMSDs, $\mathrm{R_S} = 0.67$ for RMSDh, and

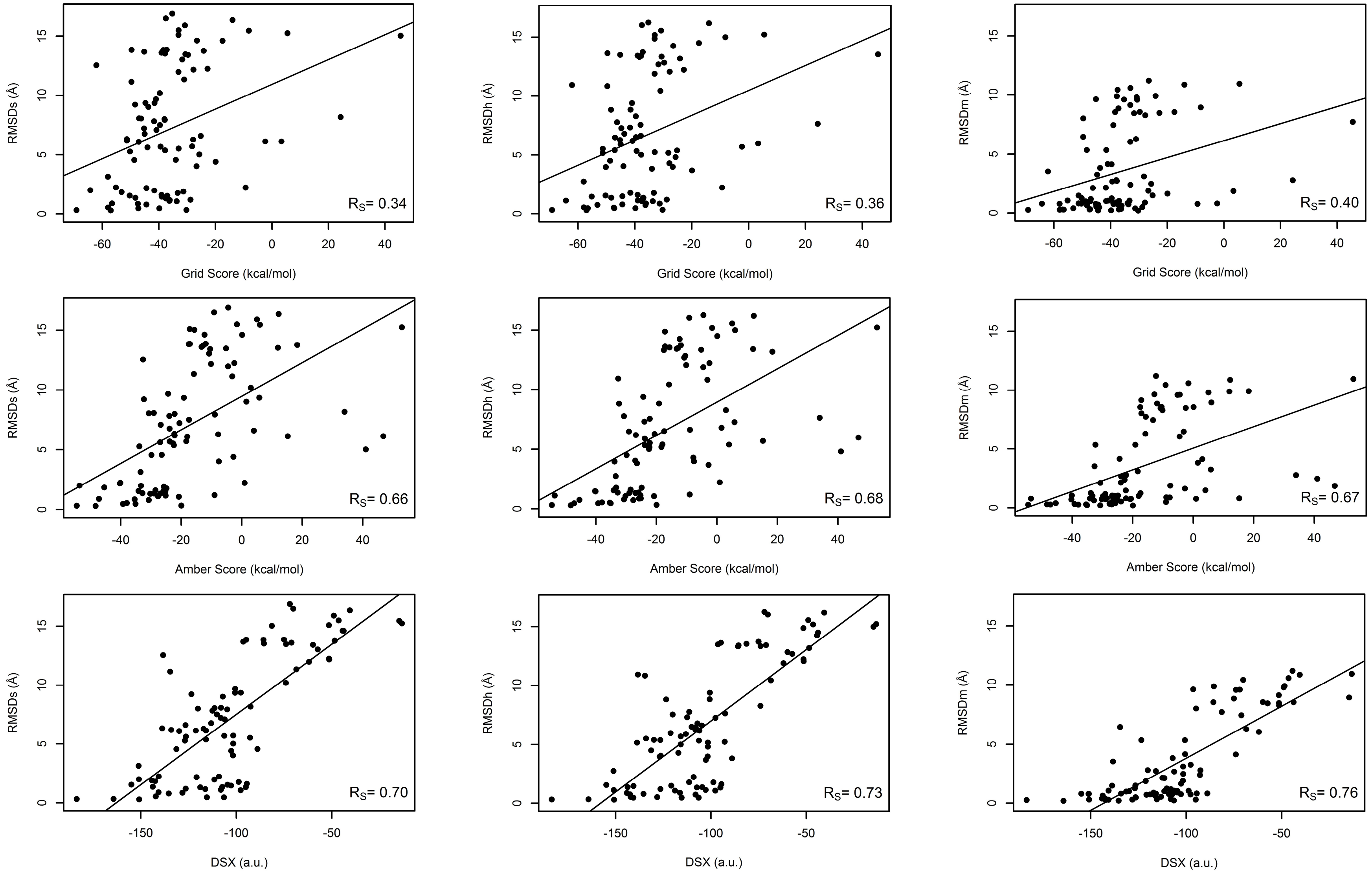


Figure 14: Correlations between RMSD values (Å) and binding scores of the three score functions, Grid Score, Amber Score, and DSX.

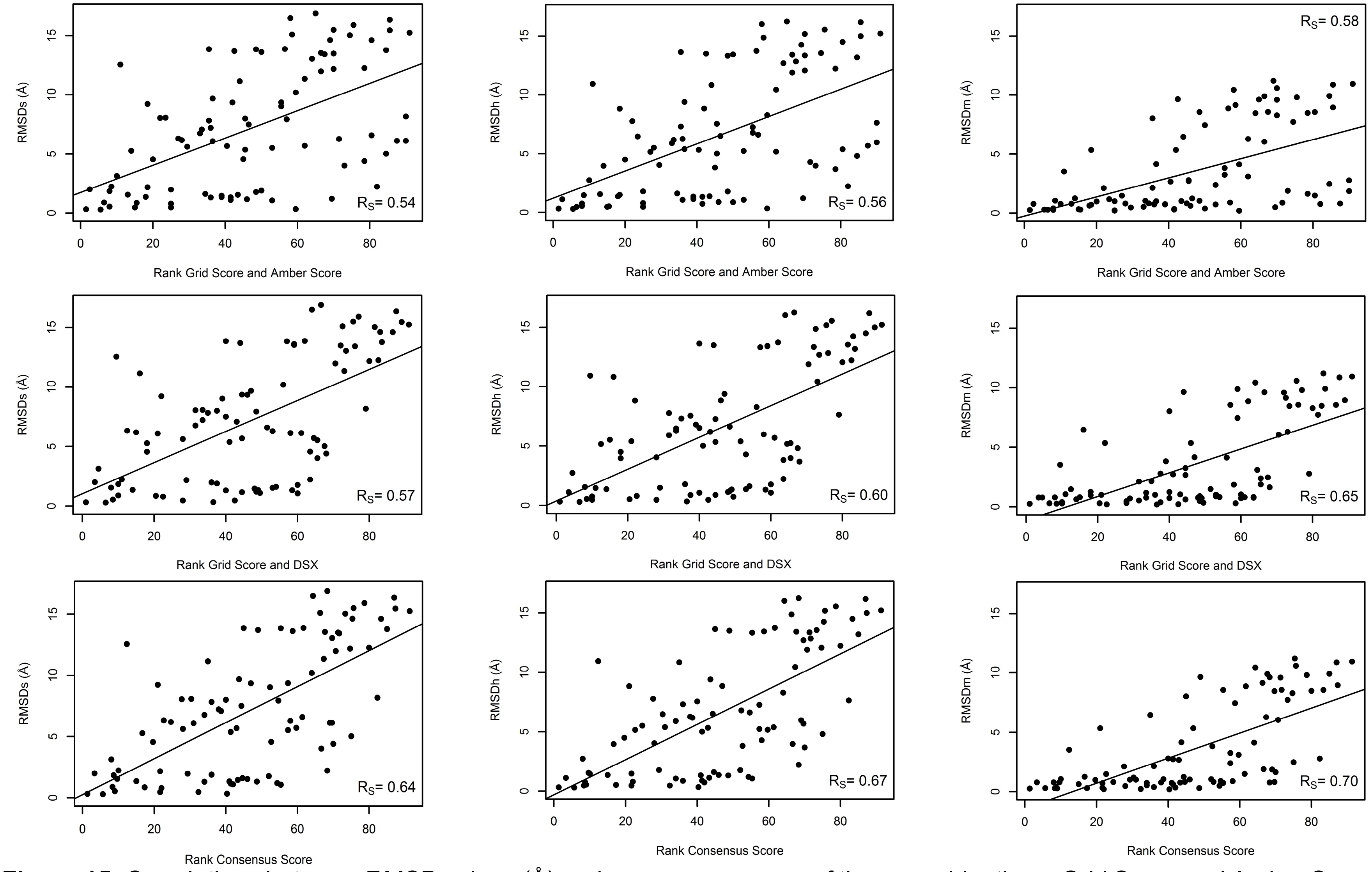


Figure 15: Correlations between RMSD values (Å) and consensus scores of three combinations, Grid Score and Amber Score, Grid Score and DSX, and all three score functions together.

$R_S = 0.70$ for RMSDm (Figure 15). Therefore, we have chosen for consensus scoring the combination of all three functions.

We carried out investigating the high-resolution structures and DOCK's performance in virtual screening. Docking of the screening library formed by known ligands and decoys were performed into all ten structures. Though, not all compounds of the screening library produced a pose, including known ligands. For instance, the structure of PDB code 1VRT could only produce poses for 5% of the entire screening library, and containing only 11% of the known ligands. Since receptor files were the same ones used in the redocking and cross-docking procedures, failure to produce poses might be related to ZINC's ligand assigned properties. Percentages of the docked screening library compounds in the selected structures can be seen at Table 9. From the docking results of each selected structure enrichment calculations were done.

Enrichment outcomes are plotted as ROC curves (Figure 16 and Figure 17) and as enrichment curves (Figure 18, Figure 19 and Figure 20). The results are numerically presented in Tables 10 and 11. When only Grid Score was considered only three structures demonstrated AUC values of 0.60 or over, the reference structure, 1C1C and 1EP4 (Table 9). However, early enrichment values were not adequate. Early enrichment is an interesting feature to bear in mind, since enrichment focuses on the rank of known ligands among decoys. On the other hand, early enrichment is limited by the fact that it does not entail that docked poses be correct regarding their experimental binding geometry. Structures 3MEC, 4I7F, and 4KV8 displayed reasonable early enrichment, but AUC values were not acceptable. Early enrichment for top 0.5%, 1%, 2%, and 5% of the docked database is shown in Table 11. When rescoring with Amber Score was considered two additional structures reached AUC values of 0.60 or over, 1C1C and 4KV8. Nevertheless, early enrichment was poorer than Grid Score. Rescoring with DSX, resulted in seven structures with AUC values of 0.60 or higher. Early enrichment was also improved for the reference, 1EP4, and 4KV8 structures. When consensus score was considered, eight out of the ten structures reached AUC of 0.60 or over, but early enrichment was not improved to most structures. The lack of early enrichment improvement could be to the influence of high-rank values in the averaged rank calculated for consensus score. The number of compounds and known ligands docked varied from structure to structure the average calculated from the molecules rank could be influenced by high rank values from the structures.

Considering the individual performance of the structures, a final rescoring of the screening library was done for each compound against groups of structures to assess the possibility of performing ensemble docking. Since consensus scoring produced the best outcome, it was chosen as the score to calculate the averages of the rank of each compound across the structures of each group. The first group comprised all ten structures. The second group comprised seven structures (4G1Q, 1C1C, 3MEC, 3MEE, 1EP4, 4IG0, and 4KV8) with AUC of 0.60 or over. The third group had four (4G1Q, 1C1C, 3MEE and 1EP4) structures with AUC of 0.65 or over. A fourth group involved two structures (4G1Q and 1C1C) with AUC over 0.70. A fifth and last group included three structures (1FK9, 4I7F, and 4IG0) with AUC up to 0.60. These tests produced low AUC values with poor early enrichment (Table 12). The problem here could be that the number of known ligands docked varied from structure to structure. Even if a compound did not achieve a pose in a particular structure, the average of ranks is still done by the total number of structures. However, consensus score provided good results in individual structures; it might be unsuited for these group of structures. Therefore, we have chosen the reference structure to perform VS of the screening library.

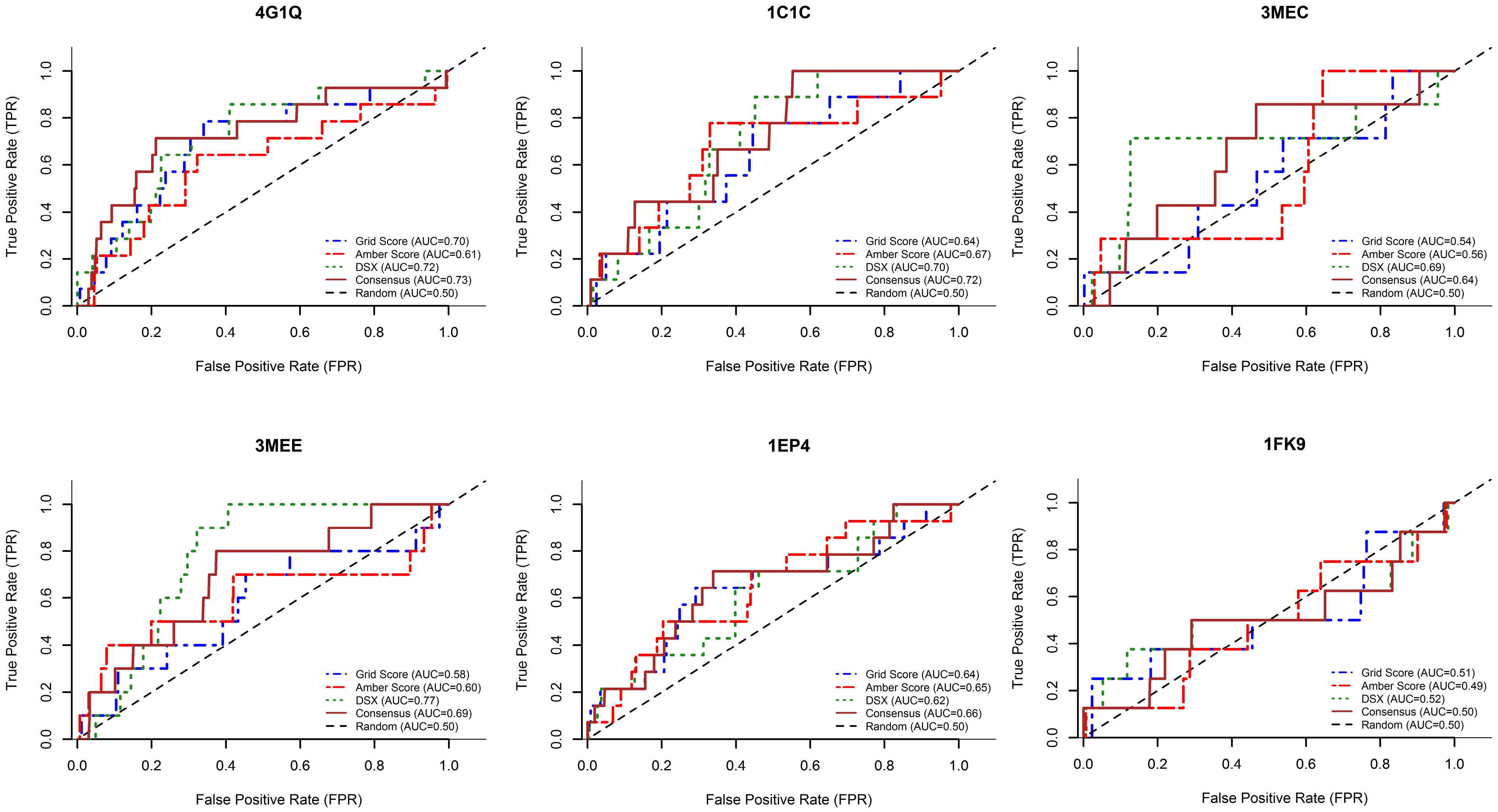


Figure 16: ROC curves of the scoring functions performance, Grid Score (blue), Amber Score (red), DSX (green), and consensus score (brown) from the structures: 4G1Q, 1C1C, 3MEC, 3MEE, 1EP4, and 1FK9.

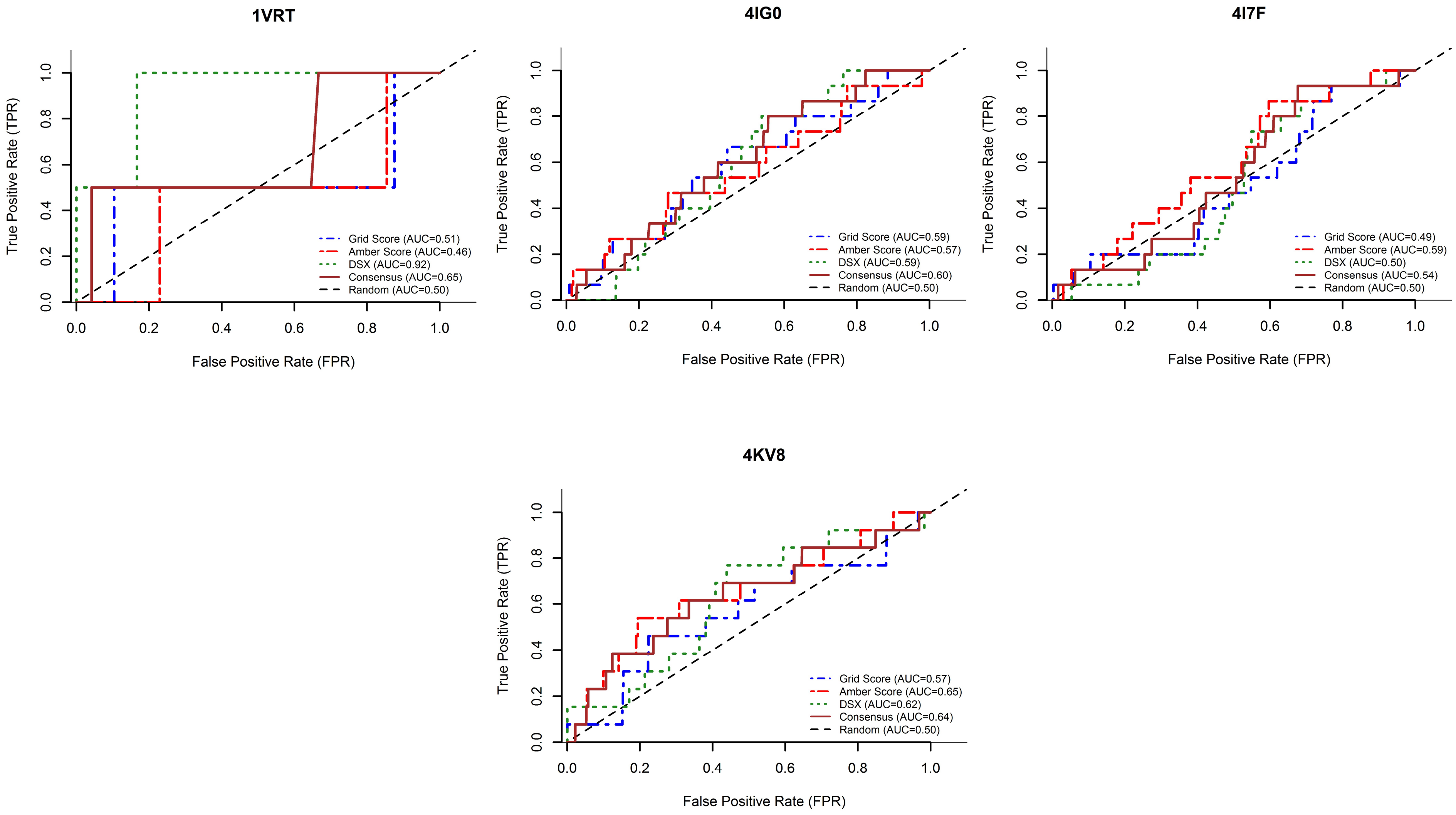
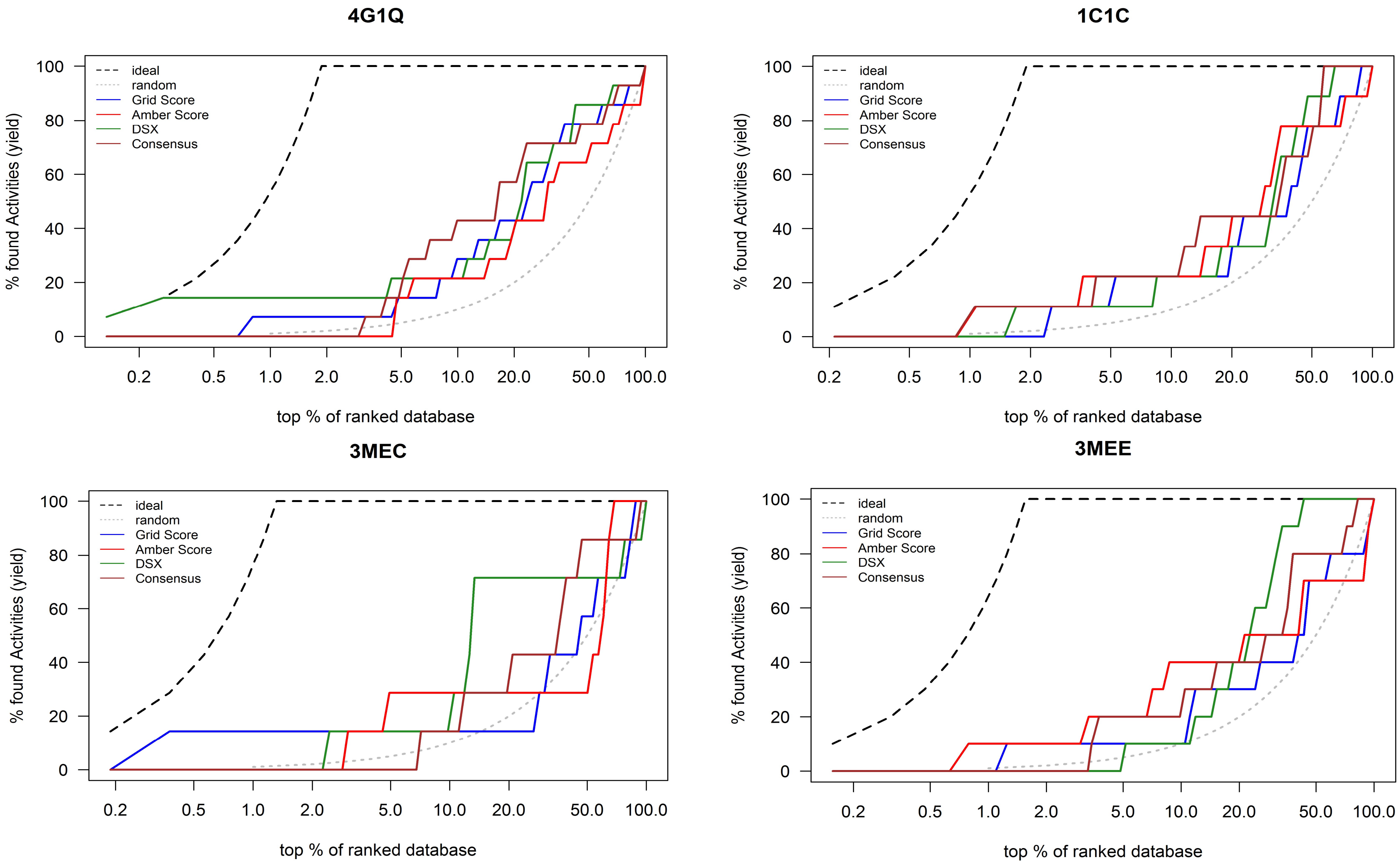


Figure 17: ROC curves of the scoring functions performance, Grid Score (blue), Amber Score (red), DSX (green), and consensus score (brown) from the structures: 1VRT, 4IG0, 4I7F, and 4KV8.


4G1Q

1C1C

3MEC

3MEE

Figure 18: Enrichment plots of the scoring functions performance, Grid Score (blue), Amber Score (red), DSX (green), and consensus score (brown) from the structures: 4G1Q, 1C1C, 3MEC, and 3MEE.

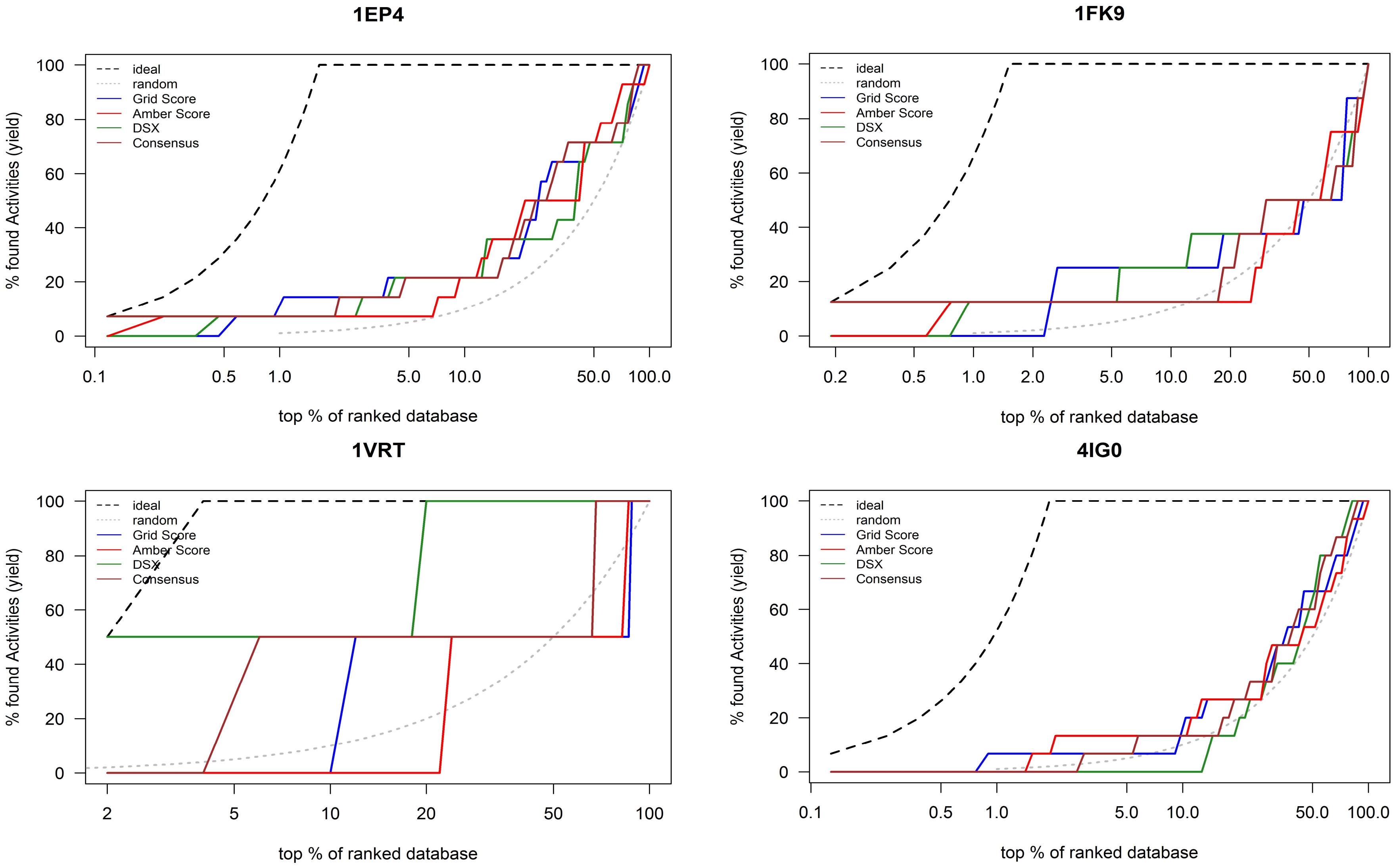


Figure 19: Enrichment plots of the scoring functions performance, Grid Score (blue), Amber Score (red), DSX (green), and consensus score (brown) from the structures: 1EP4, 1FK9, 1VRT, and 4GI0.

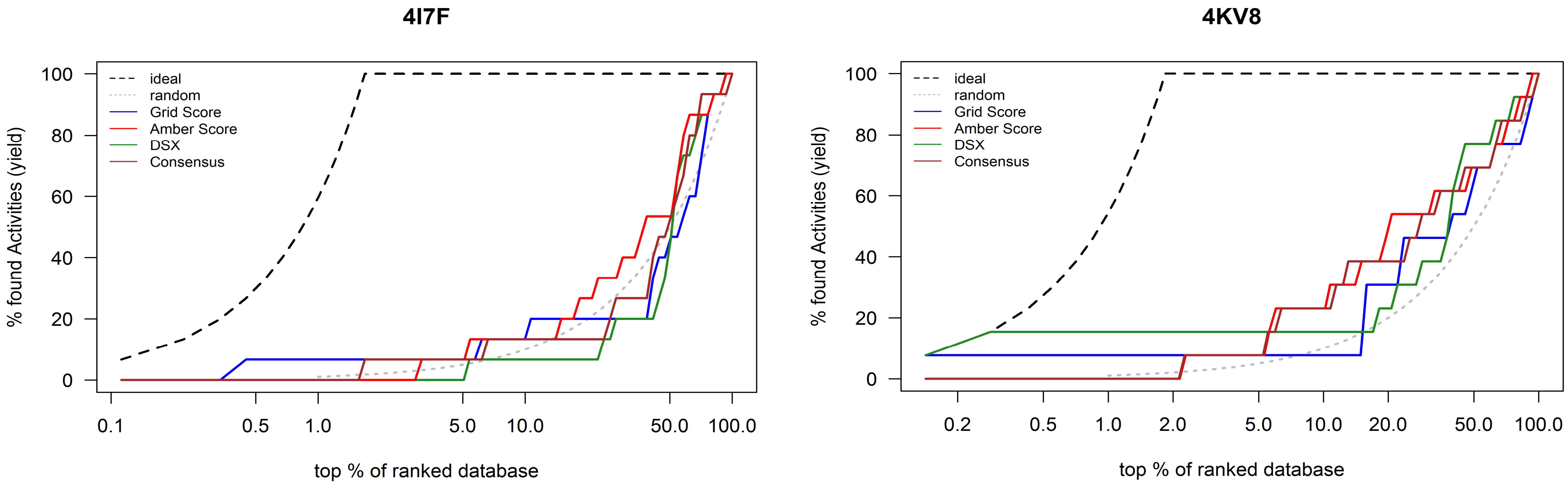


Figure 20: Enrichment plots of the scoring functions performance, Grid Score (blue), Amber Score (red), DSX (green), and consensus score (brown) from the structures: 4I7F, and 4KV8.

Table 9: Percentage of docked compounds from the screening library in the selected structures.

PDB	Docked % of Known Ligands	Docked % of Decoys	% of initial library
4G1Q	77.8	74.7	74.7
1C1C	50.0	47.2	47.3
3MEC	38.9	53.6	53.3
3MEE	55.6	64.4	64.2
1EP4	77.8	85.9	85.8
1FK9	44.4	53.0	52.8
1VRT	11.1	4.9	5.0
4IG0	83.3	78.2	78.3
4I7F	83.3	89.5	89.4
4KV8	72.2	70.7	70.7

Table 10: AUC and logAUC for the score functions to the ten selected structures.

	Grid Score		**Amber Score**		**DSX**		**Consensus Score**	
PDB	AUC	logAUC	AUC	logAUC	AUC	logAUC	AUC	logAUC
4G1Q	0.70	0.25	0.61	0.20	0.72	0.31	0.73	0.27
1C1C	0.64	0.21	0.67	0.25	0.70	0.23	0.72	0.26
3MEC	0.54	0.22	0.56	0.19	0.69	0.26	0.64	0.19
3MEE	0.58	0.20	0.60	0.25	0.77	0.24	0.69	0.23
1EP4	0.64	0.25	0.65	0.24	0.62	0.23	0.66	0.26
1FK9	0.51	0.20	0.49	0.17	0.52	0.22	0.50	0.21
1VRT	0.51	0.17	0.46	0.12	0.92	0.41	0.65	0.26
4IG0	0.59	0.19	0.57	0.19	0.59	0.15	0.60	0.18
4I7F	0.49	0.16	0.59	0.18	0.50	0.12	0.54	0.15
4KV8	0.57	0.21	0.65	0.23	0.62	0.26	0.64	0.22

Table 11: Early enrichment for 0.5%, 1%, 2%, and 5% of the ranked database.

		Grid Score				Amber Score				DSX				Consensus Score			
PDB	**E_{max}**	0.5%	1%	2%	5%	0.5%	1%	2%	5%	0.5%	1%	2%	5%	0.5%	1%	2%	5%
4G1Q	53.3	0.0	7.1	3.6	2.9	0.0	0.0	0.0	2.9	28.6	14.3	7.1	4.3	0.0	0.0	0.0	3.3
1C1C	52.4	0.0	0.0	0.0	2.2	0.0	7.4	5.6	4.4	0.0	0.0	5.6	2.2	0.0	8.0	5.6	4.4
3MEC	76.0	28.6	14.3	7.1	2.9	0.0	0.0	0.0	5.7	0.0	0.0	0.0	2.9	0.0	0.0	0.0	0.0
3MEE	64.1	0.0	0.0	5.0	2.0	0.0	10.0	5.0	4.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	4.0
1EP4	61.1	4.0	11.1	7.1	4.3	14.3	7.1	3.6	1.4	14.3	7.1	3.6	4.3	14.3	7.1	4.0	4.3
1FK9	65.9	0.0	0.0	0.0	5.0	0.0	12.5	6.3	2.5	0.0	12.5	6.3	2.5	25.0	12.5	6.3	2.5
1VRT	25.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	25.0	25.0	10.0	0.0	0.0	0.0	5.0
4IG0	52.1	0.0	6.7	3.3	1.3	0.0	0.0	4.9	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3
4I7F	59.5	13.3	6.7	3.3	1.3	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	3.3	1.3
4KV8	54.3	15.4	7.7	3.8	1.5	0.0	0.0	0.0	1.5	30.8	15.4	7.7	3.1	0.0	0.0	0.0	1.5

Table 12: AUC for selected structure groups rescoring.

Group	**AUC**
1 (All structures)	0.36
2 (4G1Q, 1C1C, 3MEC, 3MEE, 1EP4, 4IG0, and 4KV8)	0.33
3 (4G1Q, 1C1C, 3MEE, and 1EP4)	0.32
4 (4G1Q and 1C1C)	0.35
5 (1FK9, 4I7F, and 4IG0)	0.41

1.4.2 Clustering compounds

From the screening library of 2,797,315 compounds docked into the reference structure 4G1Q, a total of 2,656,863 compounds achieved a pose and were scored by Grid Score. The average score of these compounds was -38.47 kcal/mol and the median -40.44 kcal/mol. When positive scores were eliminated (28,866 compounds), the average was -39.01 kcal/mol and the median -40.52 kcal/mol. Thus, we have chosen the median of -40.52 as the cut-off to select a subset of ligands for further analysis. A total of 1,317,146 compounds achieved Grid Score of -40.52 kcal/mol or over. However, the compound that reached the minimum Grid Score (-153.90 kcal/mol), and the next three top-scoring compounds (Grid Scores of -148.06, -109.84, and -94.87 kcal/mol), were discarded due to structural errors (Figure 21). These structural errors lead to close contacts between the protein residues and the compounds, triggering in turn the favorable scores.

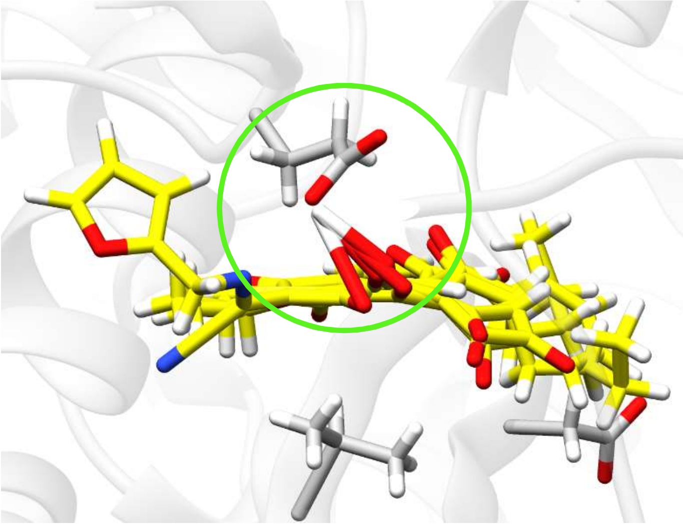


Figure 21: Structural error presented in four top scoring compounds.

From the remaining compounds, we have chosen two samples of 5,000 compounds each. One sample containing compounds ranked by the lowest Grid Score value (-71.67 kcal/mol), and another sample containing the same amount of compounds, however randomly selected. The random selection of compounds was done in order to explore the possibility to find out a new sample of good candidates in the remaining database. The Grid Score average for the molecules ranked by Score sample was -56.60 kcal/mol and median -57.10 kcal/mol. As for the randomly selected sample, Grid Score average was -44.79 kcal/mol and median -44.20 kcal/mol.

With the help of the R package ChemmineR, we clustered the compounds from selected the sets. With the calculation of atom pairs, duplicated or identical compounds were removed from the sets. This resulted in 4,288 compounds to the set ranked by Grid Score and 4,994 to the random set. Afterward, clustering of the sets were done to identify discrete similarity groups using the binning clustering function. We have chosen the implemented method that calculates Tanimoto coefficient as similarity method with a cut-off of 0.55. The cut-off of Tanimoto coefficient of 0.55 was the one that provided a manageable number of compounds to be investigated for both sets (Table 13 and Table 14).

Table 13: Clustering using Tanimoto coefficient of 0.55 for the ranked by Grid Score compounds.

Cluster size	Count
1	58
2	9
3	1
49	1
4160	1

Table 14: Clustering using Tanimoto coefficient of 0.55 for the randomly selected compounds.

Cluster size	Count
1	91
2	3
7	1
4890	1

We have chosen all compounds where cluster size was one and a representative compound from each remaining cluster, totalizing 70 compounds for the set ranked by Grid Score, and 96 for the randomly selected one.

The ChemBioServer toxicity report flagged out two compounds containing organic toxic roots for the ranked by Grid Score set. One compound presented in its composition diazene and aminothiazole, the latter was also found in the other compound. For the randomly selected set, the server signaled ten compounds. These compounds displayed in their composition a varied of toxicity motifs including Acrylonitrile-Michael acceptor, bromoethane-Michael acceptor, diazene, aminothiazole, hydroquinone, benzo-dioxane, and catechol. All flagged compounds were removed.

In the end, 68 compounds from the ranked by Grid Score set and 86 compound from the randomly selected set remained. Visual examination of the compounds confirmed that they all were within the NNBP. Using Chimera, we calculated the hydrogen bond interactions between the ligands and the receptor for all compounds in both sets. To further filter the remaining compounds, we selected only compounds with one or more hydrogen bonds. This rule resulted in 40 molecules in the ranked by Grid Score set and 20 molecules in the randomly selected one. Since the compounds were ranked according to consensus score of the three chosen scoring functions, we have chosen for each set ten compounds to further analysis (Figure 22 and Figure 24).

For the ranked by Grid Score set, all ten compounds displayed hydrogen bond interaction with Lys 101 in the NNBP (Table 15). Hydrogen bond interactions with Lys 101 are found between the crystallographic ligands RPV, EFZ, 1FG, 612, and JLJ (Table 16). A closer investigation of the ranked compounds shown that they displayed not only hydrogen bond with Lys 101, but also stacking interactions with Tyr 181 and Tyr 188. Interactions with residues Lys 103, Trp 229, and Val 179 were also observed. These interactions repeated in most compounds with some variation of residues (Figure 23 and Table 15).

Since interactions between the ligands and the protein were similar among the compounds, we analyzed their properties from the ZINC database and the chemical structure displayed by them. Chemical structures also presented similarities with known RT inhibitors, such as compound 2 benzonitrile, also found in RPV. However, compounds 1 and 9 were not commercially available (Table 17).

For the randomly selected set, all but two compounds displayed hydrogen bond interaction with Lys 101 in the NNBP (Table 18). Compound 8 interacted by hydrogen bonding with Tyr 318, whereas compound 10 formed hydrogen bond interaction with residue Ile 180. Most of the same observed interactions in the other set were also present in the randomly selected set (Figure 25 and Table 18). Compounds 5 and 6 presented a cyclopropyl, also found in EFZ. However, compounds 1, 8, and 10 are charged ligands. Charged ligands are not common in NNRTIs, probably due to the hydrophobic nature of the pocket (Table 19).

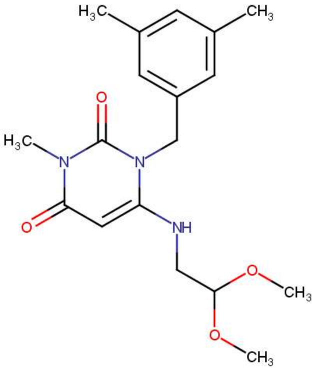
In the end, eight compounds (compounds 2, 3, 4, 5, 6 ,7, 8, and 10) from the ranked by Grid Score set remained and seven (compounds 2, 3, 4, 5, 6, 7, and 9) from the randomly selected set. It seems that the performed approach is capable of choosing interesting compounds. However, to establish which compound would be a drug candidate, further examination with other methods are needed.

Table 15: Hydrogen bond and other interactions with NNBP residues from the selected compounds from the ranked by Grid Score set.

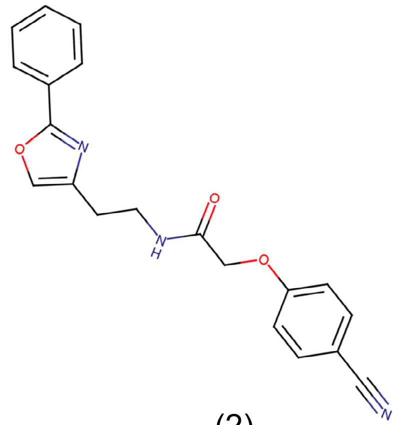
Rank	ZINC ID	H-bond Donor	H-bond Acceptor	Distance (Å)	Other Interactions
1	ZINC01495366	Lys 101.A.N-H	1.O	2.1	Lys 103, Val 179, Tyr 181,Tyr 188,Trp 229,Pro 236,Glu 138.B
2	ZINC58331692	Lys 101.A.N-H	2.O	2.3	Val 179, Tyr 181,Tyr 188,Trp 229, His 235
3	ZINC19497532	3.N-H Lys 101.A.N-H	Lys 101.A.O 3.N	1.8 2.3	Lys 103, Val 106, Tyr 181, Leu 234, His 235
4	ZINC17068270	Lys 101.A.N-H 4.N-H	4.N Lys 101.A.O	2.2 2.3	Leu 100, Val 179, Tyr 181, Tyr 188,Trp 229
5	ZINC58168359	5.N-H	Lys 101.A.O	2.0	Val 179,Tyr 181,Tyr 188,Trp 229, Tyr 318
6	ZINC71499109	6.O-H	Pro 236.A.O	2.2	Lys 103, Val 179, Tyr 181, Tyr 188, Trp 229, Pro 236
7	ZINC04013923	7.N-H 7.O	Lys 101.A.O Glu 690.B.O	1.9 3.1	Leu 100, Val 179, Tyr 181, Tyr 188,Trp 229, Leu 234, His 235
8	ZINC95425644	8.N4-H 8.N6-H19 8.N6-H18	Lys 101.A.O His 235.O Pro 236.O	2.0 2.0 2.1	Lys 103, Val 179, Try 181, Tyr 188, His 235, Pro 236, Tyr 318
9	ZINC20318147	Lys 101 A.N-H	9.O	2.6	Lys 103, Val 106, Val 179, Try 181, Tyr 188, Pro 236, Tyr 318
10	ZINC40493497	Lys 101 A.N-H	10.O	2.6	Lys 103, Val 106, Val 179, Try 181, Tyr 188, Trp 229, His 235, Pro 236

Table 16: Hydrogen bond and other interactions with NNBP residues from the crystallographic structures.

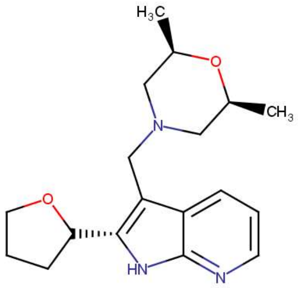
PDB	Ligand	H-bond Donor	H-bond Acceptor	Heavy-atom Distance (Å)
1C1B	GCA	GCA.N-H	Lys 101.A.O	2.9
1C1C	612	612.N-H	Lys 101.A.O	2.6
1EP4	S11	S11.N-H	Pro 236.A.O	3.0
		Lys 103.A.N-H	S11.O	2.7
1FK9	EFZ	Lys 101.A.N-H	EFZ.O	3.2
		EFZ.N-H	Lys 101.A.O	2.8
1VRU	AAP	AAP.N-H	Val 179.A.O	3.2
2RKI	TT1	Lys 103.A.N-H	TT1.N	2.9
3MEC	65B	Lys 101.A.N-H	65B.N	3.4
		65B.N-H18	Glu 138.B.O	3.3
		G5B.N-H5	Lys 101.A.O	2.7
4G1Q	T27	Lys 101.A.N-H	T27.N	3.2
		T27.N-H4	Lys 101.A.O	2.8
4I2P	G73	G73.N-H	Lys 101.A.O	2.9
4I7F	NVE	Val 106.A.N-H	NVE.O1	2.9
		Val 106.A.N-H	NVE.O2	3.3
4IG0	1FG	Lys 101.A.N-H	1FG.O	3.0
4KO0	JLJ	Lys 101.A.N-H	JLJ.N	3.1
		JLJ.N-H	Lys 101.A.O	2.6



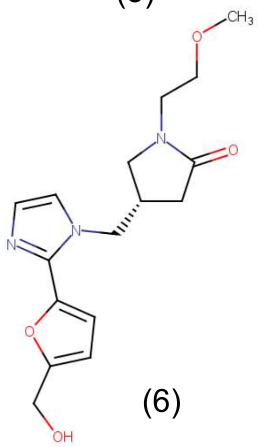
(1)

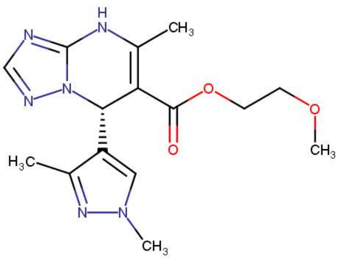


(2)

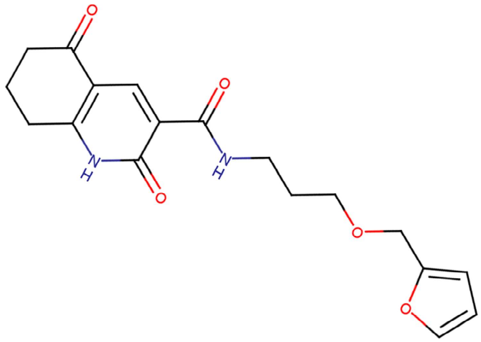


(3)



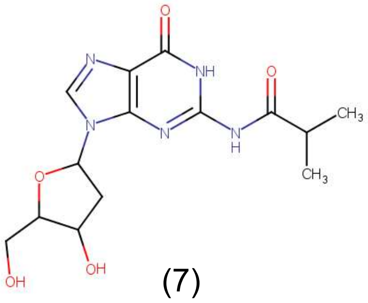


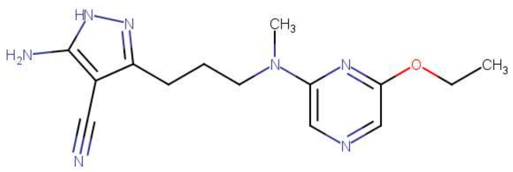
(4)



(5)

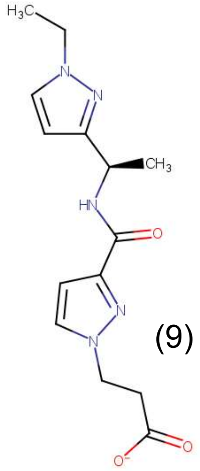
(6)



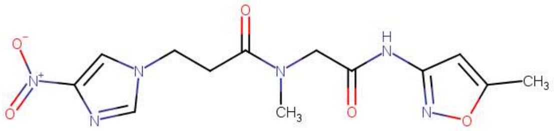


(8)

(7)



(9)



(10)

Figure 22: 2D chemical structure of the ten selected compounds from the ranked by Grid Score set. Compounds names are: (1) 6-(2,2-dimethoxyethylamino)-1-(3,5-dimethylbenzyl)-3-methyl-pyrimidine-2,4-quinone; (2) 2-(4-cyanophenoxy)-N-[2-(2-phenyloxazol-4-yl)ethyl]acetamide; (3) 3-[[(2R,6S)-2,6-dimethylmorpholin-4-yl]methyl]-2-[(2R)-tetrahydrofuran-2-yl]-1H-pyrrolo[2,3-b]pyridi; (4) 2-methoxyethyl; (5) N-[3-(2-furylmethoxy)propyl]-2,5-dioxo-1,6,7,8-tetrahydroquinoline-3-carboxamide; (6) (4S)-4-[[2-[5-(hydroxymethyl)-2-furyl]imidazol-1-yl]methyl]-1-(2-methoxyethyl)pyrrolidin-2-one; (7) N-[9-[4-hydroxy-5-(hydroxymethyl)tetrahydrofuran-2-yl]-6-oxo-1H-purin-2-yl]-2-methyl-propanamide; (8) 5-amino-3-[3-[(6-ethoxypyrazin-2-yl)-methyl-amino]propyl]-1H-pyrazole-4-carbonitrile; (9) 3-[3-[[(1R)-1-(1-ethylpyrazol-3-yl)ethyl]carbamoyl]pyrazol-1-yl]propanoic; (10) N-methyl-N-[2-[(5-methylisoxazol-3-yl)amino]-2-oxo-ethyl]-3-(4-nitroimidazol-1-yl)propanamide.

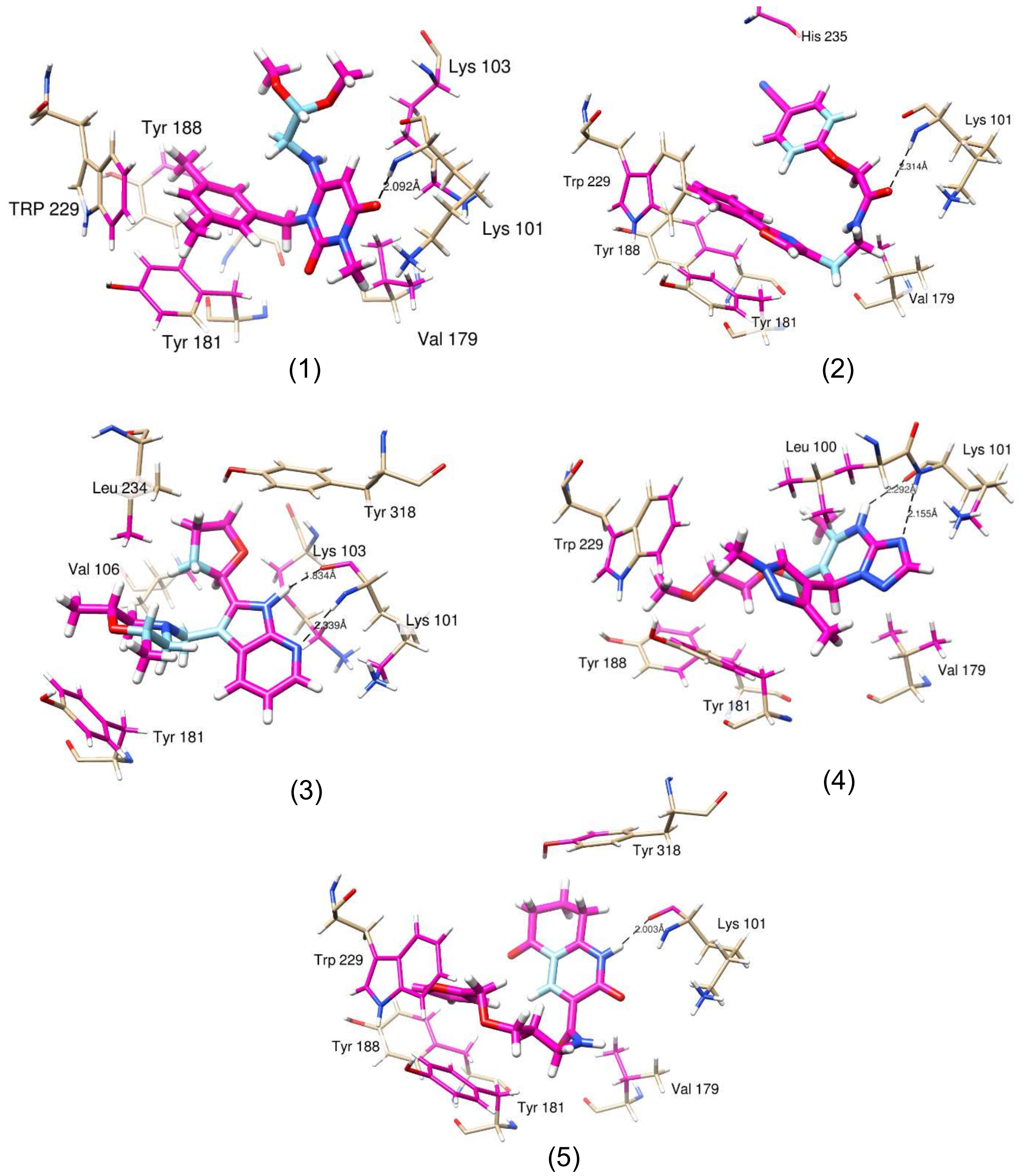


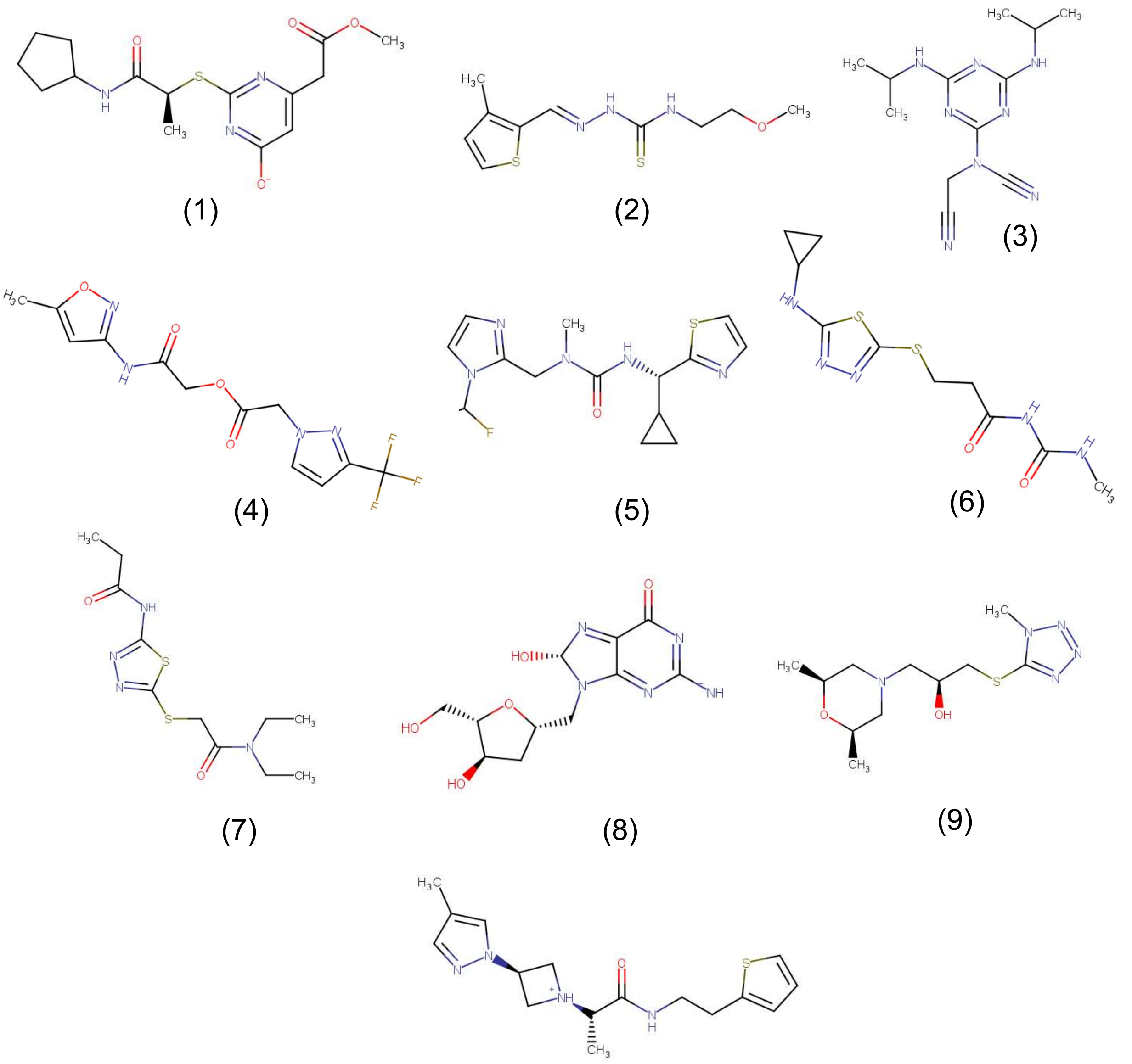
Figure 23: Compounds (1), (2), (3), (4), and (5) interactions with protein residues of the NNBP. Black dashed lines show hydrogen bond interactions and the pink coloring shows other interactions regions.

Table 17: Properties from ZINC of the ten compounds from the ranked by Grid Score set.

Rank	ZINC ID	Heavy atoms	Benign functionality	xlogP	Apolar desolvation (kcal/mol)	Polar desolvation (kcal/mol)	H-bond donors	H-bond acceptors	Net charge	Molecular weight (g/mol)	Rotatable bonds	Vendors
1	ZINC01495366	25	Yes	1.80	-0.36	-12.37	1	7	0	347.415	7	None
2	ZINC58331692	26	Yes	3.11	6.78	-21.89	1	6	0	347.374	7	5
3	ZINC19497532	23	Yes	2.15	4.92	-7.57	1	5	0	315.417	3	3
4	ZINC17068270	24	Yes	0.28	8.42	-13.98	1	9	0	332.364	6	7
5	ZINC58168359	25	Yes	1.10	4.31	-22.79	2	7	0	344.367	7	5
6	ZINC71499109	23	Yes	0.67	3.89	-17.29	1	7	0	319.361	7	6
7	ZINC04013923	24	Yes	-0.82	0.39	-23.12	4	10	0	337.336	4	4
8	ZINC95425644	22	Yes	1.55	3.16	-12.47	3	8	0	301.354	7	3
9	ZINC20318147	22	Yes	0.36	4.49	-50.24	1	8	-1	304.33	7	None
10	ZINC40493497	24	No	-0.30	8.09	-25.0	1	11	0	336.308	7	6

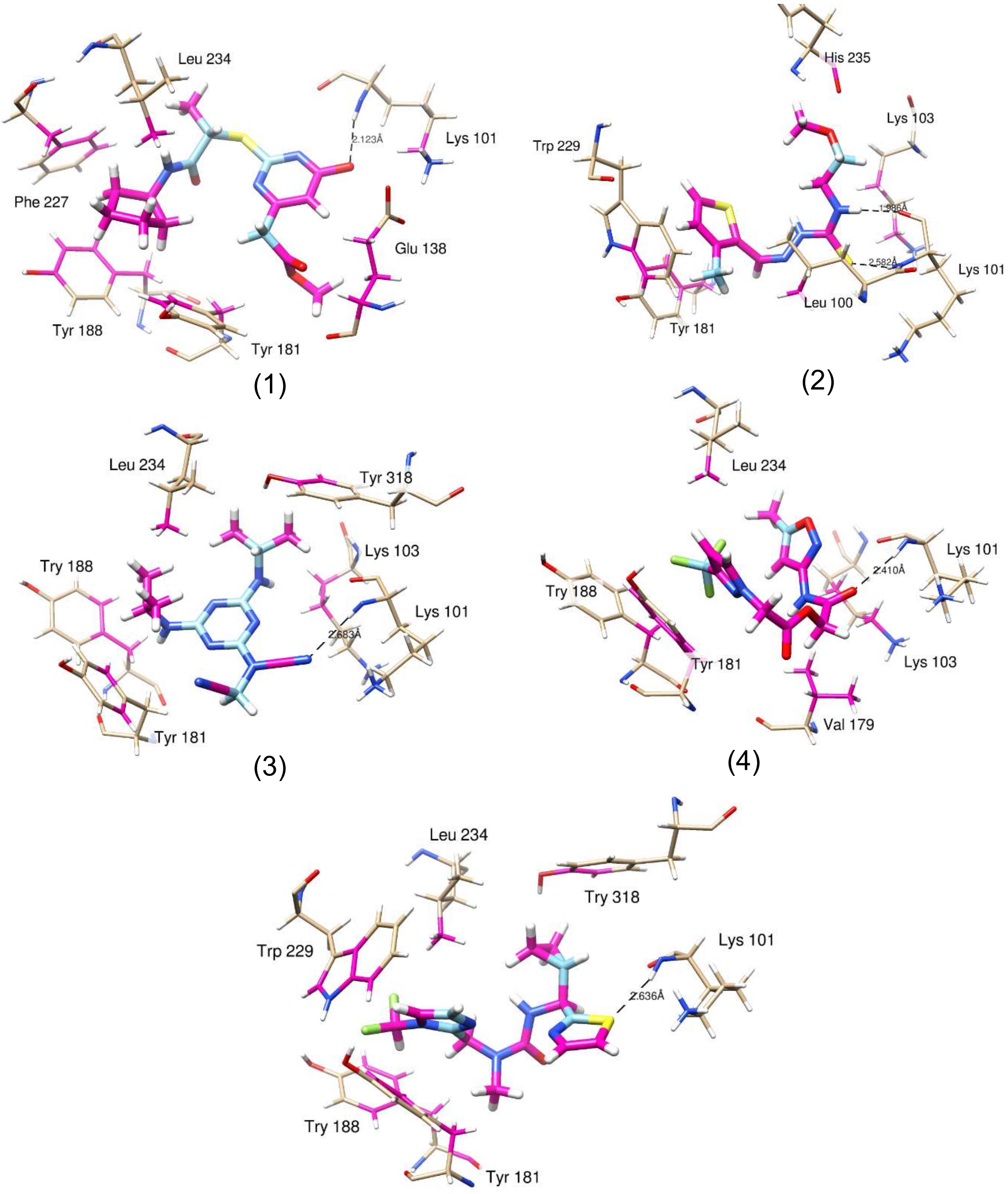
Table 18: Hydrogen bond and other interactions with NNBP residues from the selected compounds from the randomly selected set.

Rank	ZINC ID	H-bond Donor	H-bond Acceptor	Distance (Å)	Other Interactions
1	ZINC07047922	Lys 101.A.N-H	1.O	2.6	Leu 100, Lys 103, Val 179, Tyr 181,Tyr 188,Trp 227,Leu 234,Glu 138.B
2	ZINC03261798	Lys 101.A.N-H 2.N	2.S Lys 101.A.O	2.0	Leu 100, Lys 103,Tyr 181,Tyr 188,Trp 229, His 235
3	ZINC00346063	101.A.N-H	3.N	2.7	Lys 103, Val 106, Tyr 181, Leu 234, His 235, Tyr 318
4	ZINC11394129	Lys 101.A.N-H	4.N	2.4	Leu 100, Lys 103, Val 179, Tyr 181, Tyr 188, Leu 234
5	ZINC74708975	Lys 101.A.N-H	5.S	2.4	Lys 103,Tyr 181,Tyr 188,Trp 229, Leu 234, Tyr 318
6	ZINC12918855	6.N-H	Lys 101.A.O	2.1	Lys 103, Tyr 181, Tyr 188, Trp 229, Tyr 318
7	ZINC00416644	7.N-H	Lys 101.A.O	2.1	Lys 103, Tyr 181, Tyr 188,Trp 229, Leu 234, His 235
8	ZINC27644375	Tyr 318.A.OH	8.N	2.5	Leu 100, Val 179, Try 181, Tyr 188, Phe 227, Trp 229
9	ZINC66476555	Lys 101 A.N-H	9.N	2.2	Lys 103,Val 106, Val 179, Try 181, Tyr 188, Leu 234, Glu 138.B
10	ZINC69571460	Ile 180.A.N	10.O	2.5	Leu 100, Thr 139, Val 179, Ile 180,Tyr 181,Phe 227, His 235



(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

Figure 24: 2D chemical structure of the ten selected compounds from the randomly selected set. Compounds names are: (1) 2-[(2S)-1-(cyclopentylamino)-1-oxopropan-2-yl]sulfanyl-6-(2-methoxy-2-oxoethyl)pyrimidin-4-olate; (2) 1-(2-methoxyethyl)-3-[(E)-(3-methyl-2-thienyl)methyleneamino]thiourea; (3) [4,6-bis(isopropylamino)-s-triazin-2-yl]-(cyanomethyl)cyanamide; (4) [2-[(5-methyl-1,2-oxazol-3-yl)amino]-2-oxoethyl] 2-[3-(trifluoromethyl)pyrazol-1-yl]acetate; (5) 3-[(S)-cyclopropyl(thiazol-2-yl)methyl]-1-[[1-(difluoromethyl)imidazol-2-yl]methyl]-1-methyl-urea; (6) 3-[[5-(cyclopropylamino)-1,3,4-thiadiazol-2-yl]sulfanyl]-N-(methylcarbamoyl)propanamide; (7) N-[5-[[2-(diethylamino)-2-keto-ethyl]thio]-1,3,4-thiadiazol-2-yl]propionamide; (8) 8-hydroxy-2-deoxy Guanosine; (9) 1-(2,6-dimethylmorpholin-4-yl)-3-(1-methyltetrazol-5-yl)sulfanylpropan-2-ol; (10) (2S)-2-[3-(4-methylpyrazol-1-yl)azetidin-1-yl]-N-[2-(2-thienyl)ethyl]propanamide.



(5)

Figure 25: Compounds (1), (2), (3), (4), and (5) interactions with protein residues of the NNBP. Black dashed lines show hydrogen bond interactions and the pink coloring shows other interactions regions.

Table 19: Properties from ZINC of the ten compounds from the randomly selected set.

Rank	ZINC ID	Heavy atoms	Benign functionality	xlogP	Apolar desolvation (kcal/mol)	Polar desolvation (kcal/mol)	H-bond donors	H-bond acceptors	Net charge	Molecular weight (g/mol)	Rotatable bonds	Vendors
1	ZINC07047922	23	Yes	2.76	3.82	-50.72	1	7	-1	338.409	7	4
2	ZINC03261798	16	No	2.52	-3.89	-10.84	2	4	0	257.384	7	5
3	ZINC00346063	20	No	2.36	3.03	-11.79	2	8	0	274.332	6	8
4	ZINC11394129	23	Yes	1.02	5.97	-19.94	1	8	0	332.238	7	6
5	ZINC74708975	23	Yes	1.78	6.15	-14.26	1	6	0	341.387	6	5
6	ZINC12918855	19	No	0.63	3.04	-25.59	3	7	0	301.397	6	5
7	ZINC00416644	19	Yes	1.53	-0.77	-17.08	1	6	0	302.425	7	2
8	ZINC27644375	21	Yes	-2.35	-10.66	-53.93	4	10	-1	296.263	3	1
9	ZINC66476555	19	Yes	0.19	0.04	-10.61	1	7	0	287.389	5	1
10	ZINC69571460	22	Yes	1.78	7.36	-48.32	2	5	1	319.454	6	1

1.5 PERSPECTIVES

We plan to apply further analysis using molecular dynamics to determine the dynamic and energy behavior of the remaining compounds. Using methods of free energy calculations, it could be possible to establish a new rank to the compounds. With that, we could propose with slight more assertion lead candidates for experimental testing.

Since we are selecting compounds that are commercially-available, experimental assessment could be for easy access.

1.6 CONCLUSIONS

- Prior assessment of a molecular docking program performance seemed to be advantageous in the case of the chosen target HIV-1 RT. Using well established methods, such as redocking, cross-docking, and enrichment metrics, we were able to identify the strengths and weaknesses of the selected docking program, DOCK6.6.
- DOCK6.6 had a good pose reproduction performance, and a fair cross-docking one. On the other hand, the implemented algorithm, anchor-and-grow, is capable of generating useful conformations.
- DOCK's primary score function, Grid Score, did not display appealing results. With the use of one of DOCK's score function, Amber Score, alongside another score function DSX, we were able to notice improvement in the rank of scoring compounds. Using these knowledge, we performed screening of a large compound library.
- The use of molecular docking method in our work ensured that the subsequent chosen compounds, from the similarity evaluation, would fit and possible have interactions with the NNBP.
- We were able to propose a representative number of compounds from the total initial screening library that might be potential drug candidates. However, the approximations used in our compound selection might have also disregarded possible candidates.

1.7 REFERENCES

Allen WJ, Balius TE, Mukherjee S, Brozell SR, Moustakas DT, Lang PT, et al. DOCK 6: Impact of new features and current docking performance. Journal of computational chemistry. 2015;36(15):1132-56.

Allen WJ, Rizzo RC. Implementation of the hungarian algorithm to account for ligand symmetry and similarity in structure-based design. Journal of chemical information and modeling. 2014;54(2):518-29.

Asahchop EL, Wainberg MA, Sloan RD, Tremblay CL. Antiviral drug resistance and the need for development of new HIV-1 reverse transcriptase inhibitors. Antimicrobial agents and chemotherapy. 2012;56(10):5000-8.

Athanasiadis E, Cournia Z, Spyrou G. ChemBioServer: A web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery. Bioinformatics. 2012;28(22):3002-3.

Backman TW, Cao Y, Girke T. ChemMine tools: an online service for analyzing and clustering small molecules. Nucleic acids research. 2011;39(suppl 2):W486-W91.

Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of Cheminformatics. 2015.

Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science. 1983;220(4599):868-71.

Bauman JD, Patel D, Dharia C, Fromer MW, Ahmed S, Frenkel Y, et al. Detecting allosteric sites of HIV-1 reverse transcriptase by X-ray crystallographic fragment screening. Journal of medicinal chemistry. 2013;56(7):2738-46.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000;28(1):235-42.

Broder S. The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. Antiviral research. 2010;85(1):1-18.

Brozell SR, Mukherjee S, Balius TE, Roe DR, Case DA, Rizzo RC. Evaluation of DOCK 6 as a pose generation and database enrichment tool. Journal of computer-aided molecular design. 2012;26(6):749-73.

Cao L, Song W, De Clercq E, Zhan P, Liu X. Recent progress in the research of small molecule HIV-1 RNase H inhibitors. Current medicinal chemistry. 2014;21(17):1956-67.

Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T. ChemmineR: a compound mining framework for R. Bioinformatics. 2008;24(15):1733-4.

Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: definition and applications. Journal of Chemical Information and Computer Sciences. 1985;25(2):64-73.

Case D, Berryman J, Betz R, Cerutti D, Cheatham III T, Darden T, et al. AMBER 2015; University of California: San Francisco, CA, 2015. There is no corresponding record for this reference.

Charifson PS, Corkery JJ, Murcko MA, Walters W. A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. Journal of medicinal chemistry. 1999;42(25):5100-9.

Chen X, Reynolds CH. Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. Journal of chemical information and computer sciences. 2002;42(6):1407-14.

Choe H, Farzan M, Sun Y, Sullivan N, Rollins B, Ponath PD, et al. The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. Cell. 1996;85(7):1135-48.

Cihlar T, Ray AS. Nucleoside and nucleotide HIV reverse transcriptase inhibitors: 25 years after zidovudine. Antiviral research. 2010;85(1):39-58.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc. 1995;117(19):5179-97.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). J Am Chem Soc. 1996;118(9):2309-.

Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP. Comparison of automated docking programs as virtual screening tools. Journal of medicinal chemistry. 2005;48(4):962-76.

Das K, Arnold E. HIV-1 reverse transcriptase and antiviral drug resistance. Part 1. Current opinion in virology. 2013;3(2):111-8.

Das K, Martinez SE, Bauman JD, Arnold E. HIV-1 reverse transcriptase complex with DNA and nevirapine reveals non-nucleoside inhibition mechanism. Nat Struct Mol Biol. 2012;19(2):253-9.

Distinto S, Maccioni E, Meleddu R, Corona A, Alcaro S, Tramontano E. Molecular aspects of the RT/drug interactions. Perspective of dual inhibitors. Current pharmaceutical design. 2013;19(10):1850-9.

Esposito F, Corona A, Tramontano E. HIV-1 Reverse Transcriptase Still Remains a New Drug Target: Structure, Function, Classical Inhibitors, and New Inhibitors with Innovative Mechanisms of Actions. Molecular biology international. 2012;2012:586401.

Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. Journal of computer-aided molecular design. 2001;15(5):411-28.

FDA. 2013 [cited 2013 December 2014]. Available from: http://www.fda.gov/forpatients/illness/hivaids/ucm118915.htm.

Filardi PP, Paolillo S, Marciano C, Iorio A, Losco T, Marsico F, et al. Cardiovascular effects of antiretroviral drugs: clinical review. Cardiovascular & hematological disorders drug targets. 2008;8(4):238-44.

Frankel AD, Young JAT. HIV-1: Fifteen proteins and an RNA. Annu Rev Biochem. 1998;67:1-25.

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of medicinal chemistry. 2004;47(7):1739-49.

Ghosh AK, Chapsal BD, Parham GL, Steffey M, Agniswamy J, Wang YF, et al. Design of HIV-1 protease inhibitors with C3-substituted hexahydrocyclopentafuranyl urethanes as P2-ligands: synthesis, biological evaluation, and protein-ligand X-ray crystal structure. Journal of medicinal chemistry. 2011;54(16):5890-901.

Ghosh AK, Chapsal BD, Weber IT, Mitsuya H. Design of HIV protease inhibitors targeting protein backbone: an effective strategy for combating drug resistance. Accounts of chemical research. 2008;41(1):78-86.

Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol. 2000;295(2):337-56.

Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: applications of AutoDock. Journal of molecular recognition : JMR. 1996;9(1):1-5.

Group DADS, Sabin CA, Worm SW, Weber R, Reiss P, El-Sadr W, et al. Use of nucleoside reverse transcriptase inhibitors and risk of myocardial infarction in HIV-infected patients enrolled in the D:A:D study: a multi-cohort collaboration. Lancet. 2008;371(9622):1417-26.

Ho HT, Hitchcock MJ. Cellular pharmacology of 2',3'-dideoxy-2',3'-didehydrothymidine, a nucleoside analog active against human immunodeficiency virus. Antimicrobial agents and chemotherapy. 1989;33(6):844-9.

Hopkins AL, Ren J, Tanaka H, Baba M, Okamato M, Stuart DI, et al. Design of MKC-442 (emivirine) analogues with improved activity against drug-resistant HIV mutants. Journal of medicinal chemistry. 1999;42(22):4500-5.

Hsiou Y, Ding J, Das K, Clark AD, Jr., Hughes SH, Arnold E. Structure of unliganded HIV-1 reverse transcriptase at 2.7 A resolution: implications of conformational changes for polymerization and inhibition mechanisms. Structure. 1996;4(7):853-60.

Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. Journal of chemical information and modeling. 2012;52(7):1757-68.

Ivetac A, McCammon JA. Molecular recognition in the case of flexible targets. Current pharmaceutical design. 2011;17(17):1663-71.

Jaccard P. Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines: Rouge; 1901.

Jain AN, Nicholls A. Recommendations for evaluation of computational methods. J Comput Aid Mol Des. 2008;22(3-4):133-9.

Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. Journal of computational chemistry. 2002;23(16):1623-41.

Janssen PA, Lewi PJ, Arnold E, Daeyaert F, de Jonge M, Heeres J, et al. In search of a novel anti-HIV drug: multidisciplinary coordination in the discovery of 4-[[4-[[4-[(1E)-

2-cyanoethenyl]-2,6-dimethylphenyl]amino]-2- pyrimidinyl]amino]benzonitrile (R278474, rilpivirine). Journal of medicinal chemistry. 2005;48(6):1901-9.

Johnson JA, Li JF, Wei X, Lipscomb J, Irlbeck D, Craig C, et al. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. PLoS medicine. 2008;5(7):e158.

Jorgensen WL. The many roles of computation in drug discovery. Science. 2004;303(5665):1813-8.

Kirchmair J, Distinto S, Liedl KR, Markt P, Rollinger JM, Schuster D, et al. Development of anti-viral agents using molecular modeling and virtual screening techniques. Infect Disord Drug Targets. 2011;11(1):64-93.

Kohlstaedt LA, Wang J, Friedman JM, Rice PA, Steitz TA. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science. 1992;256(5065):1783-90.

Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins. 1999;37(2):228-41.

Kroeger Smith MB, Rouzer CA, Taneyhill LA, Smith NA, Hughes SH, Boyer PL, et al. Molecular modeling studies of HIV-1 reverse transcriptase nonnucleoside inhibitors: total energy of complexation as a predictor of drug placement and activity. Protein science : a publication of the Protein Society. 1995;4(10):2203-22.

Kuhn HW. The Hungarian method for the assignment problem. Naval research logistics quarterly. 1955;2(1-2):83-97.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol. 1982;161(2):269-88.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. Journal of molecular biology. 1982;161(2):269-88.

Lahti JL, Tang GW, Capriotti E, Liu TY, Altman RB. Bioinformatics and variability in drug response: a protein structural perspective. J R Soc Interface. 2012;9(72):1409-37.

Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: Combining techniques to model RNA–small molecule complexes. Rna. 2009;15(6):1219-30.

Lansdon EB, Brendza KM, Hung M, Wang R, Mukund S, Jin D, et al. Crystal structures of HIV-1 reverse transcriptase with etravirine (TMC125) and rilpivirine (TMC278): implications for drug design. Journal of medicinal chemistry. 2010;53(10):4295-9.

LaPlante SR, Bilodeau F, Aubry N, Gillard JR, O'Meara J, Coulombe R. N-versus O-alkylation: Utilizing NMR methods to establish reliable primary structure determinations for drug discovery. Bioorganic & medicinal chemistry letters. 2013;23(16):4663-8.

Laskey SB, Siliciano RF. A mechanistic theory to explain the efficacy of antiretroviral therapy. Nature reviews Microbiology. 2014;12(11):772-80.

Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins: Structure, Function, and Bioinformatics. 2005;61(4):704-21.

Liu S, Abbondanzieri EA, Rausch JW, Le Grice SFJ, Zhuang X. Slide into action: dynamic shuttling of HIV reverse transcriptase on nucleic acid substrates. Science (New York, N Y). 2008;322(5904):1092-7.

Maga G, Radi M, Gerard MA, Botta M, Ennifar E. HIV-1 RT Inhibitors with a Novel Mechanism of Action: NNRTIs that Compete with the Nucleotide Substrate. Viruses. 2010;2(4):880-99.

Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. Journal of chemical theory and computation. 2015;11(8):3696-713.

Matsuda T, Vande Berg BJ, Bebenek K, Osheroff WP, Wilson SH, Kunkel TA. The base substitution fidelity of DNA polymerase beta-dependent single nucleotide base excision repair. J Biol Chem. 2003;278(28):25947-51.

Mehellou Y, De Clercq E. Twenty-six years of anti-HIV drug discovery: where do we stand and where do we go? Journal of medicinal chemistry. 2010;53(2):521-38.

Mendieta J, Cases-Gonzalez CE, Matamoros T, Ramirez G, Menendez-Arias L. A Mg2+-induced conformational switch rendering a competent DNA polymerase catalytic complex. Proteins. 2008;71(2):565-74.

Menendez-Arias L. Mutation rates and intrinsic fidelity of retroviral reverse transcriptases. Viruses. 2009;1(3):1137-65.

Menendez-Arias L. Molecular basis of human immunodeficiency virus type 1 drug resistance: Overview and recent developments. Antiviral research. 2013;98(1):93-120.

Meng EC, Shoichet BK, Kuntz ID. Automated Docking with Grid-Based Energy Evaluation. Journal of Computational Chemistry. 1992;13(4):505-24.

Meng XY, Zhang HX, Mezei M, Cui M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. Curr Comput-Aid Drug. 2011;7(2):146-57.

Michailidis E, Huber AD, Ryan EM, Ong YT, Leslie MD, Matzek KB, et al. 4'-Ethynyl-2-fluoro-2'-deoxyadenosine (EFdA) inhibits HIV-1 reverse transcriptase with multiple mechanisms. J Biol Chem. 2014;289(35):24533-48.

Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N, et al. Development and validation of a modular, extensible docking program: DOCK 5. Journal of computer-aided molecular design. 2006;20(10-11):601-19.

Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. Journal of medicinal chemistry. 2012;55(14):6582-94.

Mysinger MM, Shoichet BK. Rapid Context-Dependent Ligand Desolvation in Molecular Docking. Journal of chemical information and modeling. 2010;50(9):1561-73.

Neudert G, Klebe G. DSX: A Knowledge-Based Scoring Function for the Assessment of Protein-Ligand Complexes. Journal of chemical information and modeling. 2011;51(10):2731-45.

Neudert G, Klebe G. DSX: a knowledge-based scoring function for the assessment of protein–ligand complexes. Journal of chemical information and modeling. 2011;51(10):2731-45.

Oda A, Tsuchida K, Takakura T, Yamaotsu N, Hirono S. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. Journal of chemical information and modeling. 2006;46(1):380-91.

Parrish J, Tong L, Wang M, Chen X, Lansdon EB, Cannizzaro C, et al. Synthesis and biological evaluation of phosphonate analogues of nevirapine. Bioorganic & medicinal chemistry letters. 2013;23(5):1493-7.

Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, DeBolt S, et al. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Computer Physics Communications. 1995;91(1):1-41.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry. 2004;25(13):1605-12.

Quashie PK, Sloan RD, Wainberg MA. Novel therapeutic strategies targeting HIV integrase. BMC Med. 2012;10:34.

Ragno R, Frasca S, Manetti F, Brizzi A, Massa S. HIV-reverse transcriptase inhibition: inclusion of ligand-induced fit by cross-docking studies. Journal of medicinal chemistry. 2005;48(1):200-12.

Reeves JD, Doms RW. Human immunodeficiency virus type 2. The Journal of general virology. 2002;83(Pt 6):1253-65.

Ren J, Esnouf R, Garman E, Somers D, Ross C, Kirby I, et al. High resolution structures of HIV-1 RT from four RT–inhibitor complexes. Nature structural & molecular biology. 1995;2(4):293-302.

Ren J, Milton J, Weaver KL, Short SA, Stuart DI, Stammers DK. Structural basis for the resilience of efavirenz (DMP-266) to drug resistance mutations in HIV-1 reverse transcriptase. Structure. 2000;8(10):1089-94.

Ren J, Nichols C, Bird LE, Fujiwara T, Sugimoto H, Stuart DI, et al. Binding of the second generation non-nucleoside inhibitor S-1153 to HIV-1 reverse transcriptase involves extensive main chain hydrogen bonding. Journal of Biological Chemistry. 2000;275(19):14316-20.

Sarafianos SG, Das K, Hughes SH, Arnold E. Taking aim at a moving target: designing drugs to inhibit drug-resistant HIV-1 reverse transcriptases. Curr Opin Struct Biol. 2004;14(6):716-30.

Scarth BJ, Ehteshami M, Beilhartz GL, Gotte M. HIV-1 reverse transcriptase inhibitors: beyond classic nucleosides and non-nucleosides. Future Virol. 2011;6(5):581-98.

Sherman W, Day T, Jacobson MP, Friesner RA, Farid R. Novel procedure for modeling ligand/receptor induced fit effects. Journal of medicinal chemistry. 2006;49(2):534-53.

Silverberg MJ, Leyden W, Hurley L, Go AS, Quesenberry CP, Jr., Klein D, et al. Response to newly prescribed lipid-lowering therapy in patients with and without HIV infection. Annals of internal medicine. 2009;150(5):301-13.

Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940-1.

Sliwoski G, Kothiwale S, Meiler J, Lowe EW, Jr. Computational methods in drug discovery. Pharmacol Rev. 2014;66(1):334-95.

Sluis-Cremer N, Temiz NA, Bahar I. Conformational changes in HIV-1 reverse transcriptase induced by nonnucleoside reverse transcriptase inhibitor binding. Current HIV research. 2004;2(4):323-32.

Spitzer R, Jain AN. Surflex-Dock: Docking benchmarks and real-world application. Journal of computer-aided molecular design. 2012;26(6):687-99.

Squires KE. An introduction to nucleoside and nucleotide analogues. Antiviral therapy. 2001;6 Suppl 3:1-14.

Steitz TA. DNA polymerases: Structural diversity and common mechanisms. Journal of Biological Chemistry. 1999;274(25):17395-8.

Steitz TA, Kohlstaedt LA, Wang J, Friedman JM, Rice PA. Crystal-Structure at 3.5 Angstrom Resolution of Hiv-1 Reverse-Transcriptase Complexed with an Inhibitor (Science, Vol 256, Pg 1783, 1992). Science. 1993;259(5093):295-.

Tanimoto TT. IBM internal report. Nov. 1957;17:1957.

Titmuss SJ, Keller PA, Griffith R. Docking experiments in the flexible non-nucleoside inhibitor binding pocket of HIV-1 reverse transcriptase. Bioorganic & medicinal chemistry. 1999;7(6):1163-70.

Tramontano E, Di Santo R. HIV-1 RT-associated RNase H function inhibitors: Recent advances in drug development. Current medicinal chemistry. 2010;17(26):2837-53.

Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. Journal of medicinal chemistry. 2005;48(7):2534-47.

Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry. 2010;31(2):455-61.

UNAIDS. Global Report: UNAIDS report on the Global AIDS Epidemic 2013. 2013.

Velec HF, Gohlke H, Klebe G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. Journal of medicinal chemistry. 2005;48(20):6296-303.

Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins. 2003;52(4):609-23.

Wang J, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. Journal of molecular graphics and modelling. 2006;25(2):247-60.

Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. Journal of computational chemistry. 2004;25(9):1157-74.

Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. Journal of medicinal chemistry. 2003;46(12):2287-303.

Wang R, Wang S. How does consensus scoring work for virtual library screening? An idealized computer experiment. Journal of chemical information and computer sciences. 2001;41(5):1422-6.

Waters LJ, Moyle G, Bonora S, D'Avolio A, Else L, Mandalia S, et al. Abacavir plasma pharmacokinetics in the absence and presence of atazanavir/ritonavir or lopinavir/ritonavir and vice versa in HIV-infected patients. Antiviral therapy. 2007;12(5):825-30.

Wlodawer A. Rational approach to AIDS drug design through structural biology. Annu Rev Med. 2002;53:595-614.

Yuriev E, Ramsland PA. Latest developments in molecular docking: 2010-2011 in review. Journal of molecular recognition : JMR. 2013;26(5):215-39.

Zhong S, Zhang Y, Xiu Z. Rescoring ligand docking poses. Curr Opin Drug Discov Devel. 2010;13(3):326-34.

Zhou Z, Madrid M, Madura JD. Docking of non-nucleoside inhibitors: neotripterifordin and its derivatives to HIV-1 reverse transcriptase. Proteins. 2002;49(4):529-42.

2 PART II

2.1 INTRODUCTION

2.1.1 Trypanosoma and Trypanosomiasis

Trypanosoma (genus *Trypanosoma*, class *Kinetoplastida*, phylum Euglenozoa) is a group of unicellular parasitic flagellate protozoa. Most trypanosomes are transmitted via a vector and require more than one obligatory host to fulfill their life cycle. Usually, the propagation of a Trypanosoma species is done by blood-feeding invertebrates, but transmission mechanism may differ among the varying species. In an invertebrate host, they are found in the intestine, but normally occupy the bloodstream or an intracellular environment in the mammalian host. Trypanosomes cause several diseases in a variety of hosts. Trypanosomiasis is the name given to a group of diseases in vertebrates caused by parasitic protozoan trypanosomes of the genus *Trypanosoma*. The diseases include Chagas disease, caused by *Trypanosoma cruzi*, and African trypanosomiasis or sleeping sickness, caused by *Trypanosoma brucei*.

2.1.1.1 Trypanosoma cruzi

The protozoan parasite, *Trypanosoma cruzi*, was identified in 1909 by Brazilian physician Carlos Chagas as the etiological agent of Chagas' disease (Chagas 1911). *Trypanosoma cruzi* is transmitted by the hematophagous *reduviid* bug (also known as the kissing bug or assassin bug). Transmission can also occur by blood transfusions and oral route, via ingestion of unwashed or undercooked food (Sajid and McKerrow 2002). The parasite can be found in different hosts, including humans, domestic animals, and rodents.

Trypanosoma cruzi manifest itself in three forms (Brener 1997, Martins et al. 2012): amastigotes, epimastigotes, and trypomastigotes. These morphologies can be identified by the position of the kinetoplast in relation to the cell's nucleus and by the emergence of the flagellum (Hoare and Wallace 1966, Contreras et al. 1988). The amastigote is the intracellular form of the parasite found in the tissues of the vertebrate host. It lacks an exterior flagellum and undulating membrane. This form, when released into the blood stream of the mammalian host, can infect new cells (Ley et al. 1988). The epimastigote is the noninfective dividing form found in the *reduviid* vectors digestive tract. It presents a free flagellum, kinetoplast, and poorly developed undulating membrane (López-Velázquez et al. 2005). The trypomastigote corresponds to the non-replicative extracellular infective form, found in both hosts. In the invertebrate ones, epimastigotes transform themselves into infective metacyclic trypomastigotes in the insect's midgut. Whereas, in the vertebrate hosts, it shifts into blood trypomastigote observed in the blood or other body fluids, for example, the cerebrospinal fluid and lymph, of the host. The trypomastigote forms possess short flagellum, narrow undulating membrane, and kinetoplast of high DNA density. The protozoan life cycle involves biological transformations between these three forms to adapt to different inner microenvironments of its mammalian hosts, including humans, and insect vectors (Vickerman 1985).

When feeding on a vertebrate host, an infected insect vector deposits its feces containing metacyclic trypomastigotes on the skin surface near the biting, that later penetrate the skin or mucosa. When reaching the tissue of the vertebrate host, the metacyclic trypomastigotes are endocytosed by the local mononuclear phagocytic system. After cell invasion, the vacuoles are disrupted, and the parasite escapes into the cytoplasm of the cell, where it replicates into round-shaped amastigotes. After several binary divisions, the amastigotes transform into infective blood trypomastigotes, which are released into the blood and tissue spaces (Martins et al. 2012). If another *reduviid* bug bites the infected vertebrate host, it may ingest the blood trypomastigotes. These trypomastigotes transform into epimastigotes in the vector's midgut, which later multiply and differentiate. Epimastigotes differentiate into metacyclic trypomastigotes in the hindgut, and part of these will expelled to the outside by the vector's feces, consequently restarting the cycle (Figure 26).

Chagas disease, also known as American trypanosomiasis, is a potentially life-threatening illness caused by the protozoan parasite *Trypanosoma cruzi*. In 2010, the World Health Organization (WHO) estimated that around 7 to 8 million people worldwide were infected with *T. cruzi,* resulting in more than 50,000 deaths every year (WHO 2013). Most cases are found in Latin American countries where the disease is endemic. However, the past decades have seen a substantial spread of this illness in the United States of America, Canada, and many European and some Western Pacific countries. The phenomenon is mainly due to population mobility between Latin America and the rest of the world.

There are two successive clinical phases of Chagas disease; an initial acute phase, and a chronic phase. In the acute phase, a high number of parasite circulate in the blood. However, in most cases symptoms are absent or mild, with less than fifty percent of people bitten by the *reduviid* bug with visible signs of the disease. These signs can be a skin lesion or a purplish swelling of the eyelid close to the bite wound or in the vicinity where the bug feces were deposited. Other symptoms are fever, headache, enlarged lymph glands, paleness, muscle pain, difficulty in breathing, swelling, and abdominal or chest pain. This acute phase lasts for about two months after infection (Coura and Borges-Pereira 2012). As the parasite sustains its life cycle by multiplying intracellularly and infecting new cells, it leads to a chronic accumulation of host tissue damage over several years, resulting in cardiac disorders, digestive (typically enlargement of the esophagus or colon), neurological or mixed alterations (Coura and Borges-Pereira 2011). If the infection is left untreated, it can cause sudden death or heart failure.

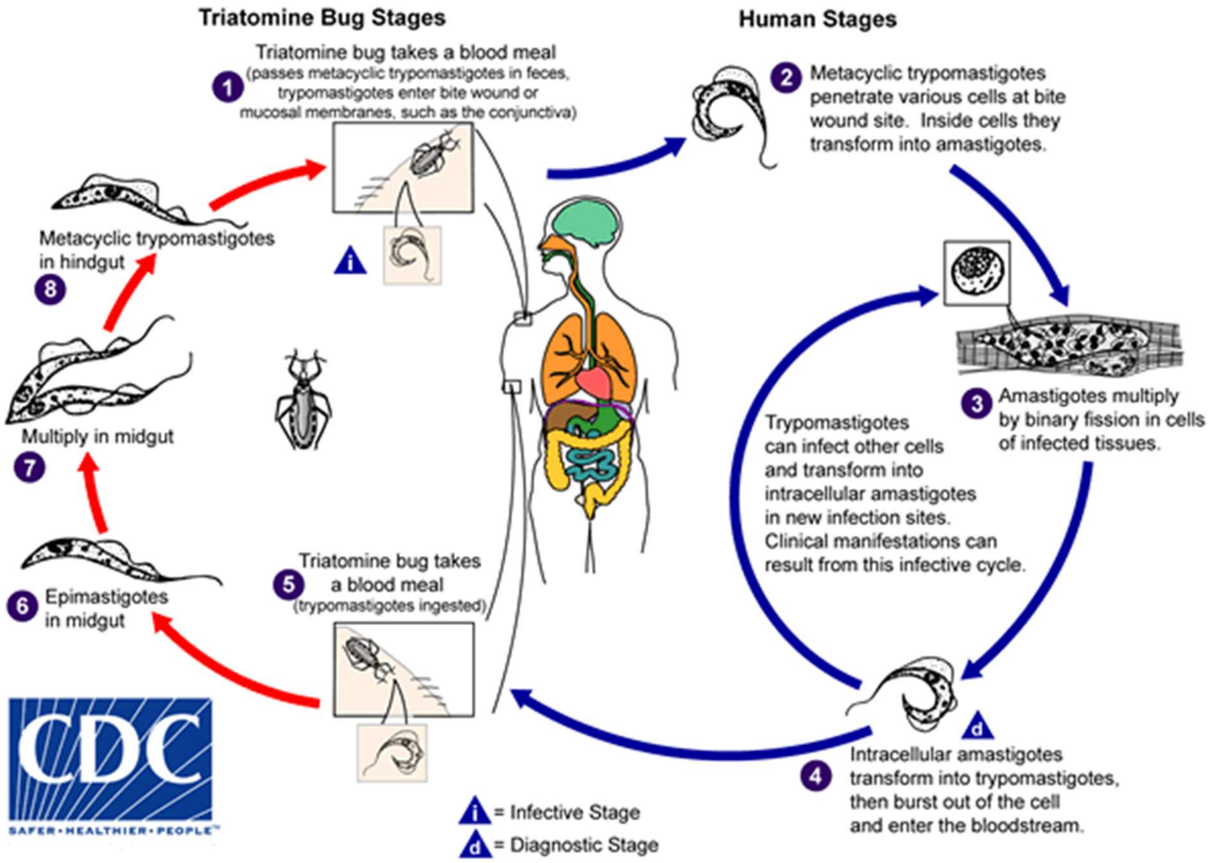
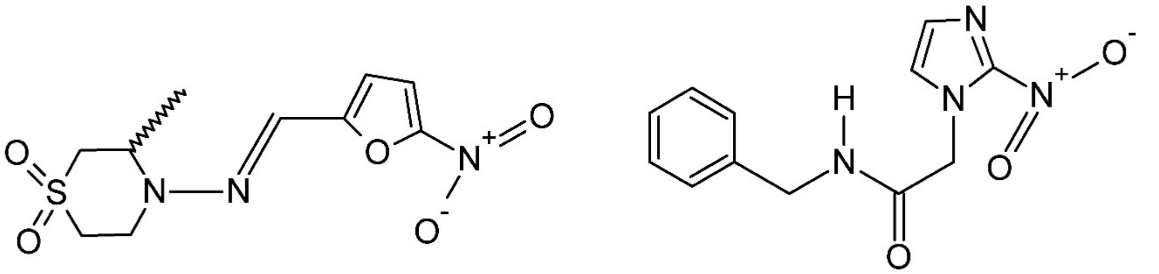

Triatomine Bug Stages
Human Stages
1 Triatomine bug takes a blood meal (passes metacyclic trypomastigotes in feces, trypomastigotes enter bite wound or mucosal membranes, such as the conjunctiva)
2 Metacyclic trypomastigotes penetrate various cells at bite wound site. Inside cells they transform into amastigotes.
3 Amastigotes multiply by binary fission in cells of infected tissues.
Trypomastigotes can infect other cells and transform into intracellular amastigotes in new infection sites. Clinical manifestations can result from this infective cycle.
4 Intracellular amastigotes transform into trypomastigotes, then burst out of the cell and enter the bloodstream.
5 Triatomine bug takes a blood meal (trypomastigotes ingested)
6 Epimastigotes in midgut
7 Multiply in midgut
8 Metacyclic trypomastigotes in hindgut
i
d
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™

Figure 26: Overview of *Trypanosoma cruzi* infective and diagnostic stages. Figure taken directly from http://www.cdc.gov/dpdx.

The current chemotherapy for Chagas disease is composed of nitrofurans like nifurtimox (Lampit, Bayer) and benznidazole (Radanil/Rochagan, Roche) (Figure 27), used to cure the acute phase of the disease. These drugs present severe side effects and limited efficacy (10-20%) for the treatment of the chronic stage of the disease (Rodrigues et al. 2002, Cazzulo 2005). Furthermore, certain *T. cruzi* strains have shown resistance to these drugs and neither compounds eradicate the parasite nor prevent damage to the heart tissue (Wilkinson et al. 2008, Lauria-Pires et al. 2000). Throughout the years, potential drugs have demonstrated efficacy *in vitro* against *T. cruzi* (Urbina et al. 1996, Urbina et al. 2000, Molina et al. 2000, Docampo 2001). However, these drugs showed not to be effective *in vivo*, including when treating the acute phase of the disease (Coura and Borges-Pereira 2012). Therefore, there is an urgent need to develop safe and efficient new anti-Chagas' drugs to overcome the issues arising from the current treatment.



Nifurtimox **Benznidazole**

Figure 27: 2D chemical structures of Nifurtimox and Benznidazole.

2.1.1.2 Trypanosoma brucei

Trypanosoma brucei is a flagellated protozoan and the etiological agent of African trypanosomiasis that infects both man and animals. The parasite is transmitted by the bite of infected tsetse flies and lives extracellularly in the blood and tissue fluids of humans. Two subspecies, morphologically indistinguishable, cause distinct disease patterns in humans: *T. b. gambiense* causes West African sleeping sickness and *T. b. rhodesiense* causes East African sleeping sickness (Ley et al. 1988). *Trypanosoma brucei brucei*, *Trypanosoma congolense* and *Trypanosoma vivax* are responsible for the infection in animals (Caffrey et al. 2001).

T. brucei undergoes morphological changes as host infection progresses (Ooi and Bastin 2013). It has a single flagellum, which is present during the cell cycle and all stages of development (Langousis and Hill 2014). *T. brucei* life cycle in mammalian hosts begins when a tsetse fly delivers growth-arrested metacyclic trypomastigotes into skin tissue through its bite. The parasites enter into the lymphatic system, moves into the bloodstream, and differentiates into proliferating long slender forms: the bloodstream trypomastigotes. Eventually, they are carried to other sites throughout the body, invade extravascular tissues, including the nervous system, and continue the replication by binary fission. The extracellular stages accomplish the entire life cycle of African trypanosomes. If a tsetse fly bites an infected host, parasites will be transported together with the infected blood into the insect midgut. The parasites transform into procyclic trypomastigotes, multiply by binary fission and leave the midgut to transform themselves into epimastigotes. The epimastigotes reach the fly's salivary glands, replicate, and eventually complete the life cycle generating metacyclic trypomastigotes that are free and adapted to survive in the mammalian host (Figure 28).

Human African trypanosomiasis (HAT) or sleeping sickness is an infection caused by the protozoan *T. brucei*. Currently, sleeping sickness is restricted to sub-Saharan Africa, where it causes morbidity and mortality in its population. The disease affects mostly rural populations living in regions where tsetse flies are found. Those who suffer from this illness depend usually on agriculture, fishing, animal husbandry or hunting. In 2013, WHO reported that the number of new cases was about 7,000, although over 50 million people in 36 countries were still at risk of acquiring the disease (WHO 2015).

HAT takes two forms, depending on the parasite involved. Sleeping sickness chronic form is caused by *T. b. gambiense* and occurs in western and central Africa (Steverding 2013). This form currently accounts for over 98% of reported cases of the infection (WHO 2015). The remaining 2% are due to the acute form of HAT caused by *T. b. rhodesiense*, found in eastern and southern Africa (WHO 2015). Throughout the course of HAT, two distinct stages can be perceived. In the first one, the parasites are restricted to the blood and lymph systems, causing irregular fever, headaches, joint pain, and itching (WHO 2015). In the second phase, the central nervous system is infected, and patients display more pronounced symptoms of the diseases, such as confusion, disturbed sleep pattern, sensory disturbances, extreme lethargy, poor coordination, and coma. If the infection is left untreated, patients infected with the *T. b. rhodesiense* form die within months and those infected with *T. b. gambiense* within years.

Chemotherapy for HAT depends on the disease stage and form of infection (Croft et al. 1997). Currently, there are four available drugs (Figure 29), where three (suramin, pentamidine, and melarsoprol) were discovered over 50 years ago (Caffrey et al. 2001). The drugs used in the first phase are safer and easier to manage than those administered for the second phase. Pentamidine and suramin are first-phase treatment drugs and are only effective against *T. b. gambiense* and *T. b. rhodesiense*, respectively. Both provoke undesirable side effects affecting the urinary tract or including allergic reactions, and difficulty in breathing (WHO 2015). Second-phase drugs include melarsoprol, active against both *T. b. gambiense* and *T. b. rhodesiense* infections; eflornithine, effective only against *T.b. gambiense*; and a combination of nifurtimox and eflornithine, recommended as first-line treatment for the *T. b. gambiense* form, but not studied for *T. b. rhodesiense*. All these drugs present major shortcomings, including poor efficacy, significant toxicity, and drug resistance has also been reported (Caffrey et al. 2001, Baker et al. 2013, Delespaux and de Koning 2007). For example, treatment with melarsoprol causes acute encephalopathy (encephalopathic syndrome) in 5-10% of patients treated, which can be fatal (1% to 5% of the cases) (Caffrey et al. 2001). Evidently, new treatment strategies for HAT are required.

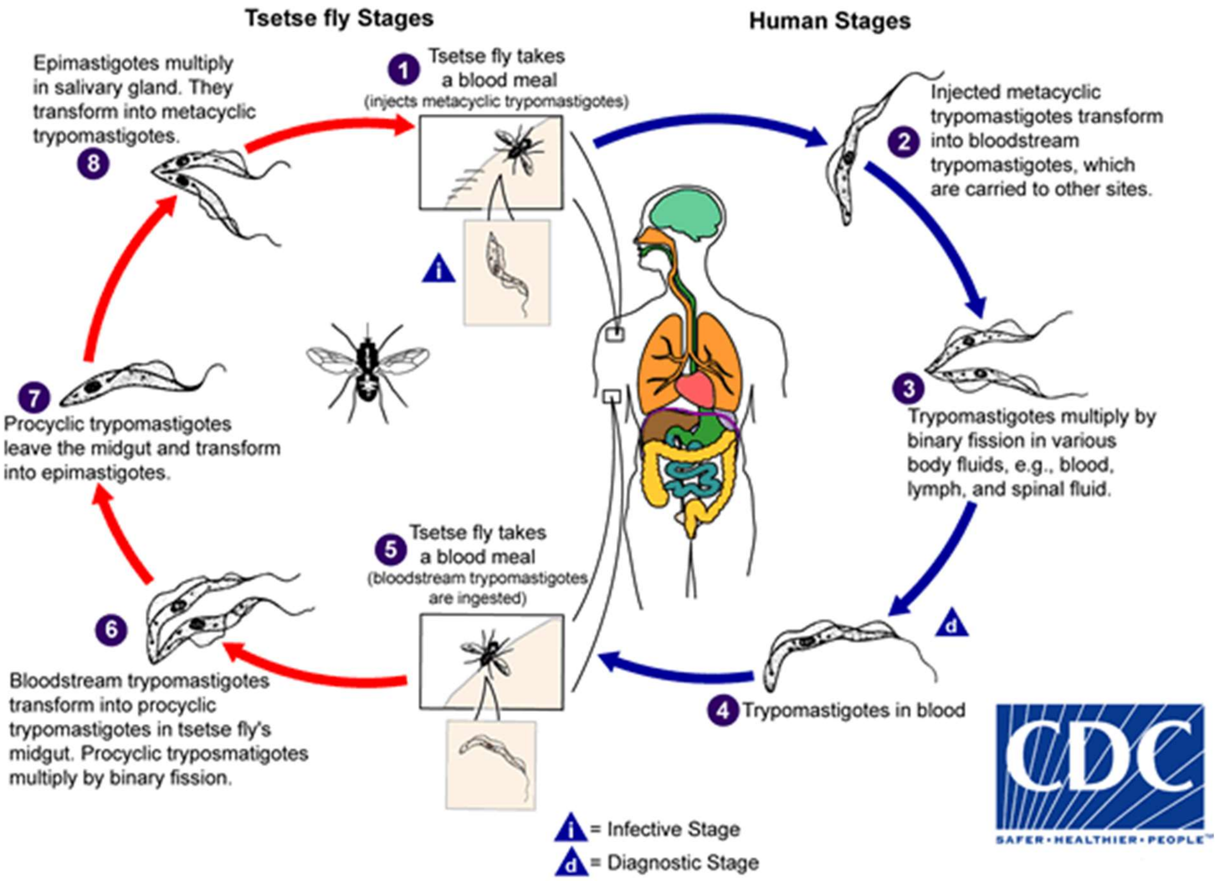

Tsetse fly Stages
Human Stages
1 Tsetse fly takes a blood meal (injects metacyclic trypomastigotes)
2 Injected metacyclic trypomastigotes transform into bloodstream trypomastigotes, which are carried to other sites.
3 Trypomastigotes multiply by binary fission in various body fluids, e.g., blood, lymph, and spinal fluid.
4 Trypomastigotes in blood
5 Tsetse fly takes a blood meal (bloodstream trypomastigotes are ingested)
6 Bloodstream trypomastigotes transform into procyclic trypomastigotes in tsetse fly's midgut. Procyclic tryposmatigotes multiply by binary fission.
7 Procyclic trypomastigotes leave the midgut and transform into epimastigotes.
8 Epimastigotes multiply in salivary gland. They transform into metacyclic trypomastigotes.
i
d
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™

Figure 28: Overview of *Trypanosoma brucei* infective and diagnostic stages. The Figure was taken directly from http://www.cdc.gov/dpdx.

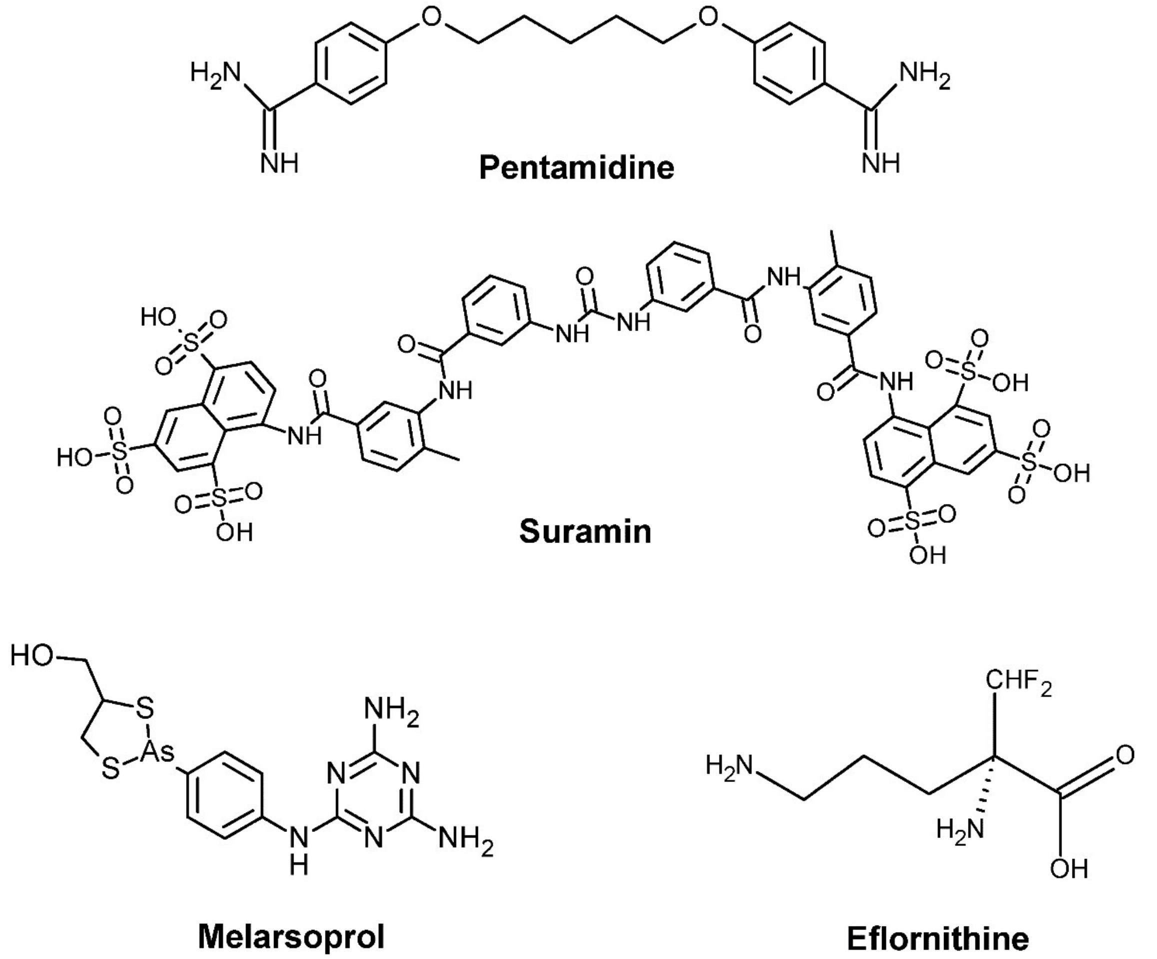


Figure 29: 2D chemical structures of pentamidine, suramin, melarsoprol, and eflornithine.

2.1.2 Overview of Proteases

Proteases (also referred to as peptidases or proteinases) are any enzyme that executes proteolysis. Proteolysis is the hydrolytic cleavage of peptide bonds under enzymatic conditions, in which the enzymes act as catalysts for the irreversible cleavage of a polypeptide chain (Smith and Simons 2004). Proteases comprise a widely studied class of enzymes and are used as drug targets for treating diseases such as diabetes, osteoporosis, various kinds of cancer, and infectious diseases (Rodenko and de Koning 2013). These enzymes perform essential functions in all living organisms. As well as mediating nonspecific protein hydrolysis, they can act as processing enzymes that execute highly selective, limited and efficient cleavage of specific substrates, which sets off irreversible decisions that influence many biological processes (Puente et al. 2003).

Proteases can be categorized according to their key catalytic group in the active site: serine, threonine, cysteine, aspartate, glutamate, or zinc in metalloproteases.

Serine, cysteine and threonine proteases act directly as nucleophiles that attack an amide carbonyl C, whereas aspartate, glutamate, and metalloproteases activate a water molecule that then acts as a nucleophile (Siklos et al. 2015). Proteases can also be classified into exopeptidases and endopeptidases, based on their substrate specificities or mechanism of catalysis. Exopeptidases truncate one or several amino acids from either the N- or the C-terminus of a peptide, whereas endopeptidases cleave within a polypeptide chain (Siklos et al. 2015).

A catalytic triad or dyad, an oxyanion hole, and specificity binding pockets categorize the active sites of proteases. The substrate specificity is determined by the structure of the binding pockets and is influenced by a defined optimum pH (Smith and Simons 2004).

For the interaction substrate-enzyme, a nomenclature according to Schechter and Berger (Schechter and Berger 1967) has become commonly used. This nomenclature designates residues carboxy-terminal to the scissile peptide bond as a prime side (P’) and amino-terminal residues as a non-prime side (P). P and P’ residues interact with complementary protease subsites called S and S’ (Schilling and Overall 2008) (Figure 30). Binding can be due to hydrophobic interactions, salt bridges, and hydrogen bonds, and the cleavage rate, or, at least, selectivity is influenced by the intensity of binding (Smith and Simons 2004).

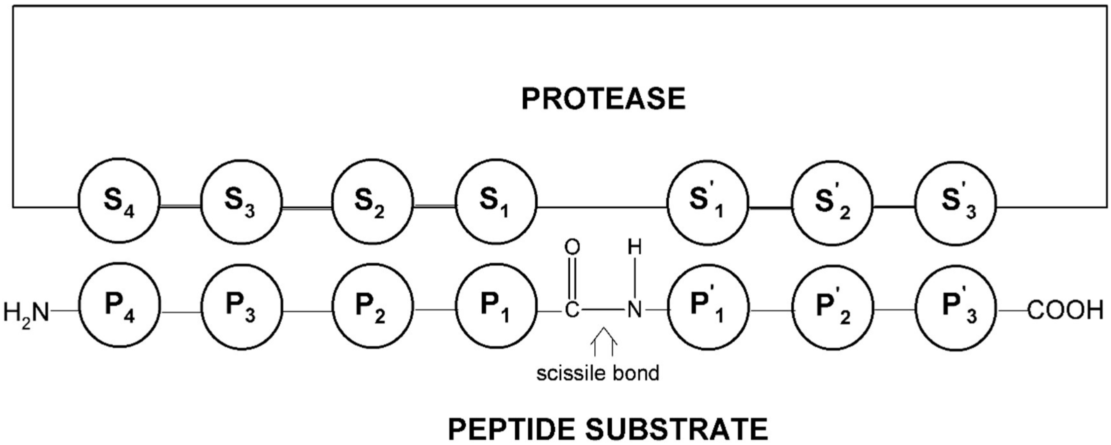


Figure 30: Substrate binding region of proteases. Enzyme subsites are designated as S1, S2, S3, S4, S1', S2', and S3'. The appropriate amino acid positions in the substrate are named as P1, P2, P3, P4, P1', P2', and P3'. The scissile bond in substrates is between P1 and P1’. Figure adapted from (Storer and Ménard 1994).

2.1.2.1 Cysteine proteases

Cysteine proteases receive their name due to the nucleophilic cysteine residue present in the active site that forms a covalent bond with the carbonyl group of the scissile peptide bond in substrates (Brömme 2001). Catalytic residues Cys 25 and His 159 (papain numbering) are evolutionarily preserved in all cysteine proteases. These residues are responsible for the enzymatic activity of cysteine proteases and thought to exist as a thiolate-imidazolium ion pair $–S^{-}…H^{+}Im–$ in the free enzyme (Storer and Ménard 1994, Polgar 2004). The formation of an unstable tetrahedral intermediate, S-acyl-enzyme moiety, is a significant step in hydrolysis (Grzonka et al. 2000). The moiety is formed via nucleophilic attack of the thiolate group of the cysteine residue acting on the carbonyl group of the hydrolyzed peptide bond with the release of the C-terminal fragment of the cleaved product. When a water molecule reacts with the moiety, the N-terminal fragment is released, and the regenerated free cysteine protease can start over a new catalytic cycle (Storer and Ménard 1994)(Figure 31).

Cysteine proteases are optimally active in slightly acidic conditions (pH 4 – 6.5) and their molecular masses are about 21-30 kDa (Grzonka et al. 2000). They are expressed either ubiquitous or tissue and cell specific, and play a major role in human physiology and pathology (Bromme and Kaleta 2002). Human cysteine proteases are involved not only in protein catabolism, but also in hormone activation, antigen presentation, and tissue remodeling (Turk et al. 2000, Nägler and Ménard 2003). Also, investigation of parasite-derived cysteine proteases has shown they play an essential role in the parasite's life cycle (Lecaille et al. 2002).

These enzymes are divided into clans, C-, CA, CD, CE, CF, and CH in the MEROPS peptidase database (Rawlings et al. 2010, Barrett and Rawlings 2001). The most explored clan is the CA, the papain-like enzymes, in which the belonging proteases consist of promising targets for therapeutic discovery in parasitic infections and human diseases.

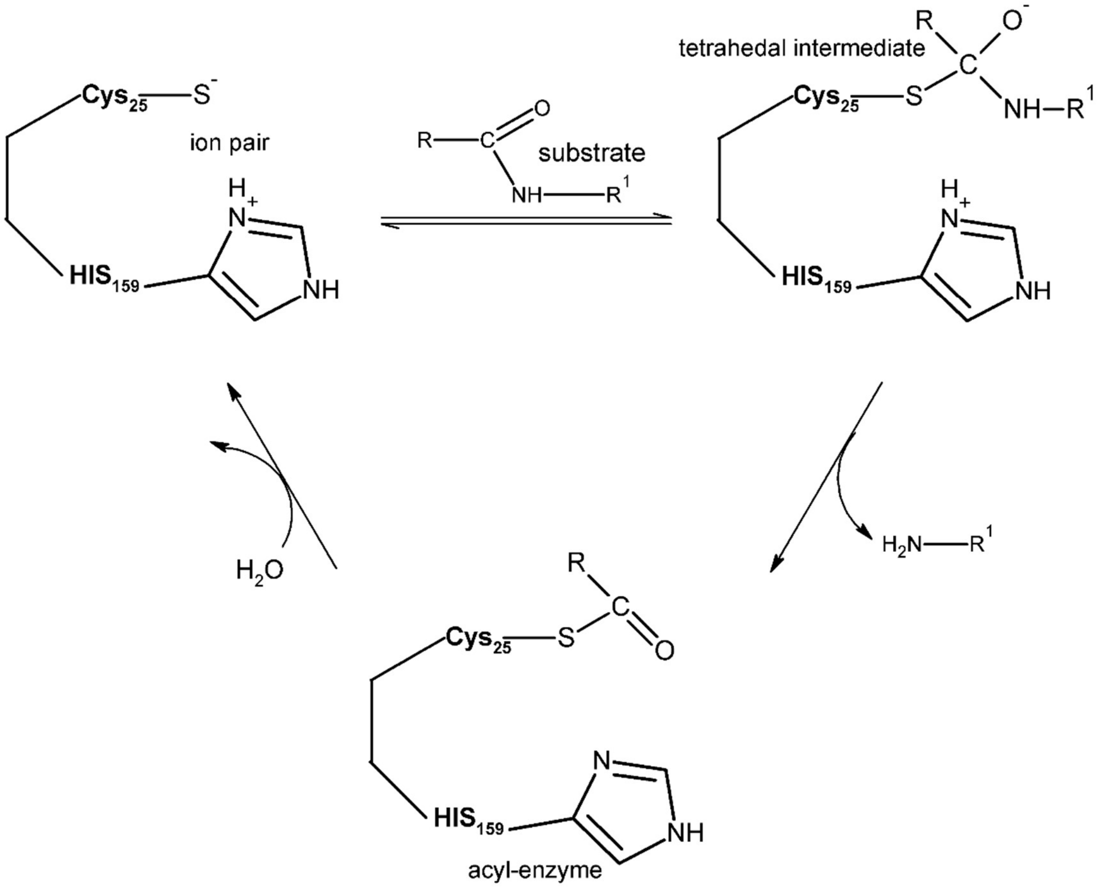


Figure 31: Simplified mechanism of substrate hydrolysis by cysteine proteases. The enzyme thiol group attacks the carbonyl carbon of a substrate's peptide bond and forms a tetrahedral intermediate that disintegrates into an acyl-enzyme powered by the release of the C-terminal portion of the substrate. Afterward, the acyl-enzyme hydrolyzes into the free enzyme and the N-terminal portion of the substrate. Figure adapted from (Storer and Ménard 1994).

2.1.2.2 Papain Family

Papain-like cysteine proteases have been classified as the clan CA, family C1 (Rawlings and Barrett 1993) and are found in viruses, plants, parasites, invertebrates, and vertebrates (Brömme 2001). Papain is a plant protease isolated from papaya fruits and gives the C1 family its name (Smith and Simons 2004). The CA-clan papain-like cysteine proteases catalytic triad consists of the referred cysteine residue, a histidine residue, and an aspartate or asparagine residue and is highly conserved among members of the enzyme family (Lecaille et al. 2002). The third member of the triad does not contribute directly to the catalysis, but acts as hydrogen bond acceptor toward the imidazole entity of the histidine, allowing consequently the formation of a permanent thiolate-imidazolium ion pair (Löser et al. 2005).

Mammalian proteases, also known as cathepsins, are not catalytically conserved. The majority are papain-like cysteine proteases, including the human isoforms B, C, F, H, K, L, O, S, V, X and W (Turk et al. 2012). Also, numerous cysteine proteases have been identified in several parasitic organisms, sharing the common amino acid sequence and fold of a papain-like structure. Location of papain-like cysteine proteases is not restricted to lysosomes; these enzymes also navigate between phagosomes and endosomes.

The mature domain of most papain-like cysteine proteases has between 214 to 260 amino acids in length, and the highest conservation is observed in the catalytic domain (Brömme 2001). In these proteases, the cysteine residue is surrounded by a highly conserved peptide sequence, CGSCWAFS (underlined is the catalytic cysteine), where only a small number of proteases have different residues in this region (Smith and Simons 2004). Furthermore, the area surrounding the two other catalytic triad residues is also preserved. A common fold is also shared between the mature enzymes, consisting of two domains of similar size, the N-terminal L (left) and the C-terminal R (right) domain. Between the two domains is located the V-shaped active site cleft, with the catalytic triad in the center, where the substrate can bind in an extended conformation (Löser et al. 2005).

In 1967, Schechter and Berger (Schechter and Berger 1967) described that the substrate pocket of papain-like cysteine proteases binds, at least, seven amino acid residues in proper subsites. With four sites binding amino acid residues N-terminal on the scissile bond (S1-S4) and three sites in the C-terminal direction (S1'-S3') (Brömme 2001). A revised proposition by Turk (Turk et al. 1998), suggested that only five subsites are essential for substrate binding. Based on kinetic and structural studies, it was revealed that S2, S1 and S1' pockets are necessary for both backbone and side-chain binding, whereas S3 and S2' are crucial for amino acid side-chain binding (Grzonka et al. 2000).

Papain-like cysteine proteases activities have been recognized as critical in the degenerative, invasive, and immune system associated disorders as well as in several parasitic infections. Parasitic papain-like cysteine proteases have been known to display virulence factors, degrade extracellular matrix proteins, and enhance the processing of proenzymes (Que and Reed 2000).

2.1.2.3 Cruzain

Cruzain is the major papain-like cysteine protease of *Trypanosoma cruzi*, the etiological agent of Chagas' disease. Originally it was named cruzipain, a term currently employed to refer to the native parasite-derived enzyme, while cruzain is used to the recombinantly-expressed protein (Sajid and McKerrow 2002). This cysteine protease is present in all stages of *T. cruzi* life cycle (Serveau et al. 1996), with higher levels in the epimastigote form (Fampa et al. 2008), and plays a number of essential biological roles (Sajid and McKerrow 2002, Steverding et al. 2006, Scharfstein et al. 2000, Aparicio et al. 2004); such as contributing to general protein turnover, nutrient processing, evasion of host immune response, cell infection, and parasite differentiation.

Cruzain structural domain contains 215 amino acids including the catalytic residues, Cys 25, His 162, Asn 182, and the oxyanion hole Glu 19 (cruzain numbering). As a member of the papain superfamily of cysteine proteases, cruzain has a sequence identity closely related to the major cysteine proteases of *Trypanosoma brucei* (around 70%), rhodesain, and to the human cathepsin F enzyme (around 50%). It has been shown that the specificity of cruzain is largely due to the composition of the S2 pocket of the substrate-binding cleft, as observed for other proteases of the same family (Sajid et al. 2011). Cruzain's S2 pocket is hydrophobic in nature and residues Met 68, Ala 138, Leu 160 and Gly 163, with Glu 208 at the base of the pocket, are present in its formation. Glu 208 plays a significant role in the physicochemical preference for substrate residues, typically amino acids with a noncharged aliphatic or aromatic side group, when at the pH optimum of cruzain (pH 5.5) (Gillmor et al. 1997, McGrath et al. 1995). However, the presence of a Glu at the S2 pocket also grants the opportunity for interaction with amines in the P2 position of the substrate, due mostly to Glu 208 restructuration of the pocket at different pH (Gillmor et al. 1997). Depending on the ligand co-crystalized, crystal structure comparisons showed that the carboxylic acid moiety of Glu 208 swings into the S2 pocket at neutral pH but is directed away at acid pH (Sajid et al. 2011).

As a target, cruzain has been validated through biochemical studies, animal models, and X-ray crystallographic structure determination. Also, new classes of cruzain compounds have been discovered throughout the years, and for some of the binding mode has been determined through crystallography (Gillmor et al. 1997, McGrath et al. 1995, Mott et al. 2009, Ferreira et al. 2010). However, none of these compounds has reached clinical trials. Inhibitor classes extend from potent peptidic compounds (McGrath et al. 1995, Kerr et al. 2009, Choe et al. 2005, Huang et al. 2003) to potent nonpeptidic compounds (Bryant et al. 2009, Brak et al. 2010, Ferreira et al. 2014). Cruzain inhibitors based on vinyl sulfones, tetrafluorphenoxymethyl ketones, and diazomethyl ketones are the most structuraly studied compounds for the enzyme. These classes contain an electrophilic warhead capable of inactivating covalently and irreversibly the enzyme as a result of nucleophilic attack by the active site cysteine (Yang et al. 2012). Also, several small reversible compounds, which usually bind noncovalently to the enzyme producing inhibition, have been identified (Ferreira et al. 2010, Ferreira et al. 2014). Despite the different binding modes, the great concern is to develop selective inhibitors able to distinguish between mammalian and parasitic proteases, since there is a high identity percentage with cysteine proteases (Ettari et al. 2013).

2.1.2.4 Rhodesain

Trypanosoma brucei in bloodstream expresses two papain-like cysteine proteases: a cathepsin L-like protease and a cathepsin B-like protease (Ehmke et al. 2013). However, *T. brucei* cathepsin L-like enzymes are the most abundant of the two proteases. Recently, Steverding et al. (Steverding et al. 2012) conducted a study in which they showed that *T. brucei* cathepsin L-like protease is the essential cysteine protease of *T. brucei* and should be considered the primary target for the treatment of sleeping sickness (Steverding et al. 2012). To distinguish the cathepsin L-like proteases from *T. b. rhodesiense* and *T. b. brucei*, they are commonly termed rhodesain and brucipain, respectively. Rhodesain and brucipain are nearly identical in sequence (98.4% identity) (Ettari et al. 2013) and the same functional role is assumed in both subspecies of *T. brucei* (Costa et al. 2012). Furthermore, rhodesain exhibits notable structural similarity with *T. cruzi* cruzain (around 70% amino acid identity). Similar to cruzain, it consists of a single polypeptide chain of 215 amino acids with the catalytic triad (Cys 25/His 162/Asn 182) located between the left (L) and right (R) domains.

Rhodesain plays essential roles in all phases of *T. b. rhodesiense* life cycle (Kerr et al. 2010). Its presence is required to allow the parasite cross the blood–brain barrier of the human host that causes the severe (and lethal) stage of the disease (Lonsdale-Eccles and Grab 2002). Rhodesain is also involved in the turnover of variant surface glycoproteins of trypanosomes (Barry and McCulloch 2001), enabling infected *T. brucei* to evade host immune responses, causing chronic infection (Ettari et al. 2013).

Like other papain-like cysteine proteases, rhodesain's S2 pocket establishes the specificity towards peptidyl substrates. Within the S2 pocket, a range of bulky hydrophobic residues is tolerated with a preference for leucine and phenylalanine in the P2 position of the substrates. The presence of an alanine residue, Ala 208, at the bottom of the S2 pocket, turns it shallower than the cruzain pocket (Kerr et al. 2010), while residues Gln 159 and Leu 160 narrow the pocket (Kerr et al. 2009).

With the help of crystal structures of rhodesain in complex with specific inhibitors (Kerr et al. 2009), structural and biochemical insights have enabled the design of rhodesain compounds. Rhodesain main classes of inhibitors include peptidic, peptidomimetic, and nonpeptidic structures, with different modes of action (Ettari et al. 2013). Promising inhibitors can be found in the groups of vinyl sulfones, aldehydes, ketone derivatives, azadipeptide nitriles, thiosemicarbazones and fumaric acid derivatives (Ettari et al. 2013).

2.1.3 Molecular modeling

Both cruzain and rhodesain are validated therapeutic targets, with inhibitor classes described, and for some of these compounds, their structures in complex with the respective proteins were determined by X-ray crystallography. Structural similarity between cruzain and rhodesain enables the development of novel inhibitors for both enzymes. Amino acid identity between both proteins is around 70%; the similarity is even higher in the active site. In this region the differences are found in the S3 subsite, where the cruzain residue Ser 186 is replaced by a Phe residue in rhodesain; and at the base of the S2 subsite, where the cruzain residue Glu 208 is replaced by an Ala residue. Consequently, inhibitors often have little distinction between both enzymes, and classes of compounds which act against both proteins have been developed (Mott et al. 2009, Greenbaum et al. 2004, Jaishankar et al. 2008)

From molecular docking studies only is not possible to understand the differences in their inhibitory capacity. Therefore, in this work we propose the use of Molecular Dynamics (MD) simulations and free energy calculations to determine the affinity of a series inhibitors candidates of both enzymes.

2.1.3.1 Molecular Dynamics

Molecular Dynamics (MD) is a widely used method to determine the dynamic, structural and thermodynamic properties of biological systems through computer simulations. This methodology uses a potential energy function, that is a description of the terms by which the particles in the simulation will interact, known as a force field (Adcock and McCammon 2006). This potential function is based on molecular mechanics and is represented by

$$E = \underbrace{E_{bonds} + E_{angles} + E_{torsions}}_{bonded} + \underbrace{E_{electrostatic} + E_{vdw}}_{non-bonded} \tag{1}$$

the sum of bonded and non-bonded potentials, defined by their atomic coordinates. The bonded terms refer to bonds, angles, and torsional energies, and the non-bonded terms describe the electrostatic and van der Waals interactions (Rapaport 2004). In a MD simulation, the interaction and the physical movements of atoms and molecules are depicted over time, usually over tens to hundreds of nanoseconds (ns), generating a trajectory interactively (Adcock and McCammon 2006). Trajectories are obtained by integrating Newton's equation of motion

$$\vec{F_i} = m_i \vec{a_i} = m_i \frac{d^2 \vec{r_i}}{dt^2} \tag{2}$$

where $\vec{F_i}$ is the resulting force acting on atom i with mass m_i. $\vec{a_i}$ is the acceleration, which can be represented as the second derivative of the coordinates $\vec{r_i}$ with respect to the time t. The force can also be derived from the gradient of the potential energy,

$$\vec{F_i} = -\nabla_i E \tag{3}$$

Combining these two equations, Newton's equation of motion can then relate to the derivative of the potential energy of the changes in positions as a function of time. The atoms motions are not independent, but they influence each other and are coupled together. Since the potential energy is a function of the atomic positions of all the atoms in the system, there is no analytical solution to the equations of motion. Therefore, they must be solved numerically.

2.1.3.2 Free energy calculations

The term "free energy calculations" is commonly used to denote a class of numerical simulations that relate, through equations of classical statistical mechanics, the free energy difference between two different states or conformations to a thermodynamic ensemble average determined by potential energy properties of those states or conformations (Beveridge and DiCapua 1989). Free energy simulation methods have been successfully applied to a variety of problems in chemistry and biology (Kollman 1993, Hansen and van Gunsteren 2014, Michel et al. 2010). All molecular behavior, from association to conformational preference, stems directly from the free energy profile (Pearlman and Rao 1998). As the calculation of the absolute free energy of a receptor-ligand complex is virtually impossible (Pearlman and Rao 1998), most free energy calculations are formulated regarding estimations of the relative free energy differences (Brandsdal et al. 2003). A difference in free energy provides the relative likelihood of directing the system to one state as opposed to another.

The difference in free energy between two neighboring states, A and B, can be calculated from (Zwanzig 1954):

$$\Delta G = G_B - G_A = -\beta^{-1} \ln\langle \exp(-\beta \Delta V) \rangle_A \tag{4}$$

where $\beta = 1/kT$ and $\langle \quad \rangle_A$ denotes an ensemble average of $\Delta V = V_B - V_A$.

From the computational point of view and convergence's sake, simulations are usually carried out by describing a series of non-physical intermediate states connecting the physical states A and B (Pearlman 2001). These intermediate states are potential energy functions that are commonly constructed as linear combinations of the states A and B:

$$V_m = (1 - \lambda_m)V_A + \lambda_m V_B \tag{5}$$

where λ varies from 0 to 1 (Brandsdal et al. 2003) (Figure 32). As the simulation progresses, the system gradually begins to look more like B and less like A. The use of the coupling parameter λ is supported by the fact that the free energy difference is defined by the initial and final states and can be calculated along any reversible path connecting those states (Brandsdal et al. 2003).

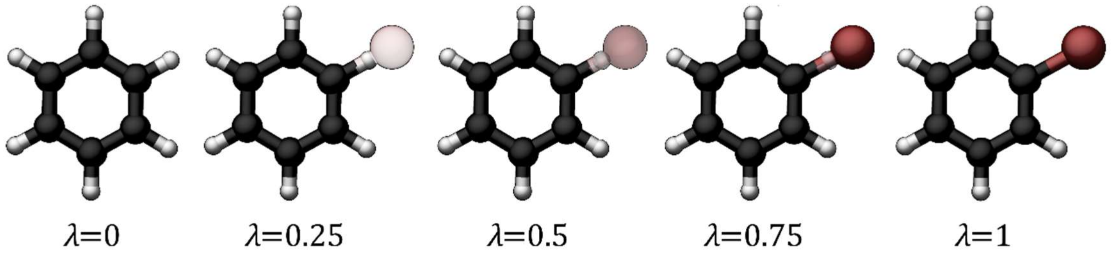


Figure 32: Alchemical transformation from an initial state (λ=0) to a final state (λ=1) divided into distinct intermediate states with a corresponding coupling parameter λ.

The total free energy change can be obtained by summing over the intermediate states along the λ variable:

$$\Delta G = G_B - G_A = -\beta^{-1} \sum_{m=1}^{n-1} \ln\langle \exp[-\beta(V_{m+1} - V_m)] \rangle_m \tag{6}$$

In general, free energy simulations are carried out in the context of a thermodynamic cycle (Figure 33). For instance, to compare the relative binding energies of two ligands to the same receptor, we can use the cycle in Figure 33. To calculate free energy differences, two transformation systems need to be prepared: one for the unbound ligands in solution (ΔG_S) and the other complexed to the receptor (ΔG_C). Since free energy is a state function, the variation in free energy can be rigorously calculated as the sum of the free energies differences between these similar intermediates (Pearlman 2001).

$$\Delta\Delta G = \Delta G_X - \Delta G_Y = \Delta G_S - \Delta G_C \qquad (7)$$

Two approaches are broadly used to produce a thermodynamically relevant ensemble: Molecular Dynamics (MD) (McCammon and Karplus 1983) and Monte Carlo (MC) (Jorgensen 1983). In MD, the ensembles are generated over time, from which the averaged quantities can be derived, whereas, in MC the averages are derived from ensembles over space, through a series of random moves along with energy based approval criteria to create a thermodynamically significant ensemble (Pearlman and Rao 1998). However, Jorgensen et al. (Jorgensen and Tirado-Rives 1996) demonstrated that for reasonably sized solutes, MC simulations can be exposed to limitations in the total amount of sampling when compared to those obtained from MD, especially for macromolecular systems that can undergo large-scale motions (Pearlman and Rao 1998).

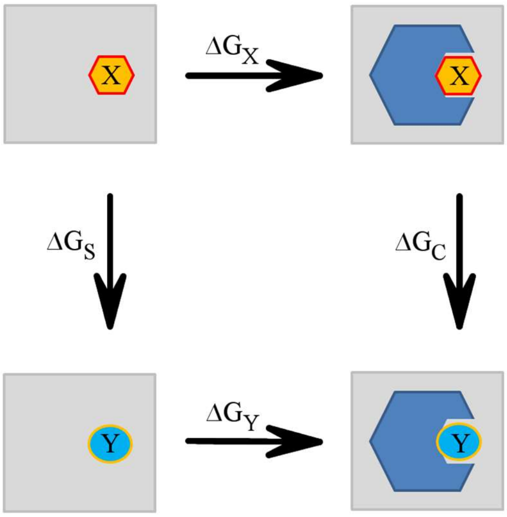


Figure 33: Thermodynamic cycle for relative free energies of binding. The receptor is in dark blue, and X and Y are two ligands.

The most applicable and possibly most accurate methods for calculating free energies are the sampling-based explicit methods, free energy perturbation (FEP) and thermodynamic integration (TI), or variants of those. Despite their accuracy, these methods are computationally expensive with slow convergence. Other methods can be applied to estimate relative and absolute free energies, however, they provide some limitations depending on the system. In the present work, TI was used to calculate the difference in the free energy of binding for a series of potential inhibitors to the enzymes cruzain and rhodesain.

TI method can be derived from the same basic classical statistical mechanical equations to determine free energies (Pearlman and Rao 1998). If the λ-steps are sufficiently small, the potential difference in the exponent of Eq. (5) can be rewritten as

$$V_{m+1} - V_m = \frac{\partial V_m}{\partial \lambda_m} \Delta\lambda \tag{8}$$

where $\Delta\lambda_m = \lambda_{m+1} - \lambda_m$ and $\Delta V = V_B - V_A = \frac{\partial V_m}{\partial \lambda_m}$. Combining Eq. (6) with Eq. (8), we have

$$\Delta G = -\beta^{-1} \sum_{m=1}^{n-1} \ln \left\langle \exp\left[-\beta \frac{\partial V_m}{\partial \lambda_m} \Delta\lambda_m\right] \right\rangle_m \tag{9}$$

For small steps in λ, Eq. (9) can be linearized by retaining only the leading terms in the Taylor expansion of the exponent and logarithm, yielding:

$$\Delta G = \sum_{m=1}^{n-1} \left\langle \frac{\partial V_m}{\partial \lambda_m} \right\rangle_m \Delta\,\lambda_m \tag{10}$$

When $\lambda \rightarrow 0$, Eq. (10) can be written as an integral over λ:

$$\Delta G = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{11}$$

Eq. (11) is the master equation for TI method. In this equation, $\lambda = 0$ corresponds to V_A and $\lambda = 1$ corresponds to V_B. In practice, the integrand in Eq. (11) is assessed at a series of discrete points or “windows”, and the integral is approximated from these points using a numerical integration method. All numeric integration methods have the form

$$\Delta G \approx \sum_{m=1}^{n} w_m \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_m \tag{12}$$

where w_m are weights and will depend on the numeric integration chosen. TI calculations use several independent MD (or MC) simulations at fixed λ values, so parallelization and the subsequent addition of λ points can be performed.

2.2 AIMS

2.2.1 Main goal

To investigate the dynamic and energy behavior of a set noncovalent inhibitors of the enzymes cruzain and rhodesain in different protonation states of the enzyme and elucidate the possible binding modes of a series of analogues derived from the crystallographic inhibitor B95 through molecular modeling methods.

2.2.2 Particular goals

- To examine and analyze structural stability of the B95 ligand (neutral and positively charged) under the influence of different protonation states of the cruzain catalytic dyad, Cys 25 and His 162.
- To define and investigate initial binding modes for the ligand B95 in its non-native structure of rhodesain by molecular docking and molecular dynamics.
- To estimate the free energy of binding of a series of B95 analogs with related structures in complex with cruzain and rhodesain through thermodynamic integration methodology. Calculations will be useful to give support to the binding modes obtained by molecular docking methods.

2.3 MATERIAL AND METHODS

2.3.1 Crystal structure selection

The crystal structures of cruzain and rhodesain were taken from the Protein Data Bank (Berman et al. 2000). Cruzain structure, PDB code 3KKU (Ferreira et al. 2010), is in complex with the non-covalent ligand B95 located in the active site of the enzyme, and has an X-ray resolution of 1.28 Å. On the other hand, the chosen rhodesain structure, PDB code 2P86 (Marion et al.), is bound to the irreversible pseudopeptide inhibitor K11002 located in the active site of the enzyme, and presents an X-ray resolution of 1.16 Å. Computational preparation of the protein consisted of hydrogen addition of residues according to acidic protonation states (pH=5.5) using the MOE (2016) program. The ligand B95 was also prepared with MOE. Hydrogens were added and two protonation states of the ligand benzimidazole ring were considered.

2.3.2 Active site amino acids

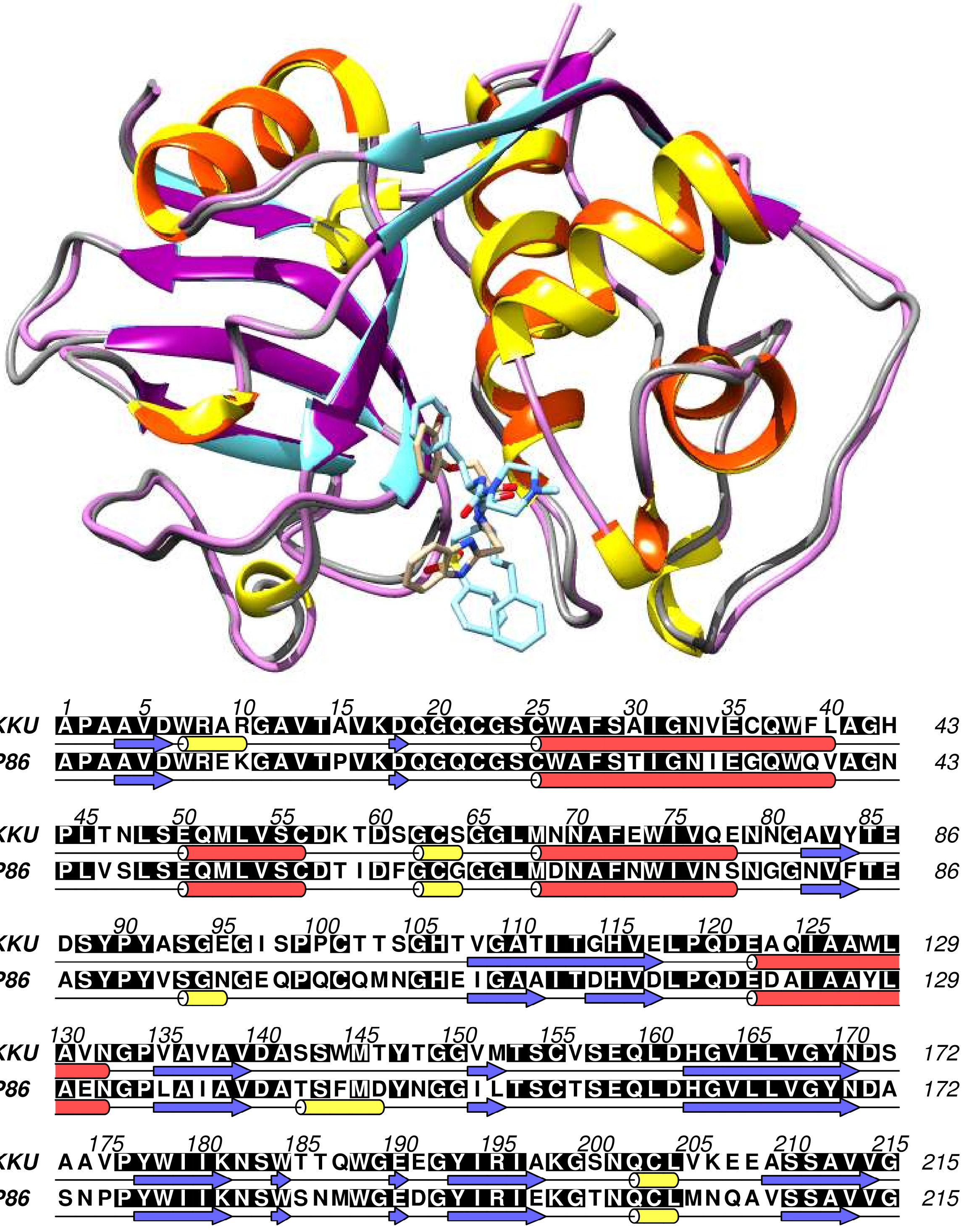
The active site of cruzain was considered as the residues within 10 Å of compound B95 center of mass: Gln 19, Gly 23, Cys 25, Trp 26, Ala 27, Ser 29, Asp 60, Ser 61, Gly 62, Cys 63, Ser 64, Gly 65, Gly 66, Leu 67, Met 68, Asn 69, Glu 117, Val 137, Ala 138, Val 139, Asp 140, Ala 141, Ser 142, Gln 159, Leu 160, Asp 161, His 162, Gly 163, Val 164, Trp 184, Ser 186, Leu 204, and Glu 208. We superimposed the crystal structures 3KKU and 2P86 to assess the active site of rhodesain (Table 20 and Figure 34). Since the region occupied by both ligands was the same, we selected the residues chosen at same position of cruzain in rhodesain as the corresponding active site. The differences between sites are found in residues Phe 61, Gly 64, Asp 69, Asp 117, Ile 137, Thr 142, Phe 186, and Ala 208.

2.3.3 Molecular docking

Since no crystal structure of the non-covalent ligand B95 bound to rhodesain is available, molecular docking was employed to obtain an initial binding mode. Docking was performed with Glide version 6.6 (Friesner et al. 2006, Friesner et al. 2004, Halgren et al. 2004) using the Virtual Screening Workflow that allows ensemble docking of a ligand library against multiple rigid receptors.

Table 20: Cruzain and rhodesain alignment results.

Max score	Total score	Query cover	E value	Identity	Positives	Gaps	RMSD
322	322	100%	2e-111	151/215 (70%)	175/215 (81%)	0/215 (0%)	0.466 Å



3KKU APAAVDWRARGAVTAVKDQGQCGSCWAFSAIGNVECQWFLAGH 43
2P86 APAAVDWREKGAVTPVKDQGQCGSCWAFSTIGNIEGQWQVAGN 43
3KKU PLTNLSEQMLVSCDKTDSGCSGGLMNNAFEWIVQENNGAVYTE 86
2P86 PLVSLSEQMLVSCDTIDFGCGGGLMDNAFNWIVNSNGGNVFTE 86
3KKU DSYPYASGEGISPPCTTSGHTVGATITGHVELPQDEAQIAAWL 129
2P86 ASYPYVSGNGEQPQCQMNGHEIGAAITDHVDLPQDEDAIAAYL 129
3KKU AVNGPVAVAVDASSWMTYTGGVMTSCVSEQLDHGVLLVGYNDS 172
2P86 AENGPLAIAVDATSFMDYNGGILTSCTSEQLDHGVLLVGYNDA 172
3KKU AAVPYWIIKNSWTTQWGEEGYIRIAKGSNQCLVKEEASSAVVG 215
2P86 SNPPYWITKNSWSNMWGEDGYIRIEKGTNQCLMNQAVSSAVVG 215

Figure 34: Superimposed structures of cruzain (purple and orange), PDB code 3KKU, and rhodesain (cyan and yellow), PDB code 2P86.

Ten representative structures were extracted from a MD simulation of rhodesain structure 2P86 without its native ligand using hierarchical clustering. The first two poses were selected and submitted to additional analysis using short 10 ns MD simulations to determine a possible B95 binding mode to rhodesain.

2.3.4 Molecular dynamics simulations

MD simulations were carried out using AMBER 14.0 (Case et al. 2014) version of PMEMD. Protein models used the Amber 99SB-ILDN force field (Hornak et al. 2006, Lindorff-Larsen et al. 2010), with the TIP3P (Jorgensen et al. 1983) model for water. Ligands were parametrized using the generalized Amber force field (GAFF) (Wang et al. 2004) with the LEaP program, whereas AM1-BCC (Jakalian et al. 2000) partial charges were assigned using the Antechamber (Wang et al. 2006) program. Also using LEaP, truncated periodic octahedral box was used with a minimum distance of 12 Å between any box edge and any solute atom. Both LEaP and Antechamber programs are part of AmberTools version 14.0 (Case et al. 2014). Before the free MD simulations, an extensive protocol involving minimization and equilibration was performed (For further details, see Appendix B).:

i) A minimization of 1000 cycles where all heavy atoms of the system were kept fixed with a force constant weight of 1000 kcal mol^{-1} Å^{-1}. The steepest descent algorithm was used to the first 500 cycles and then switched to conjugate gradient algorithm for the remaining steps. Short range vdW interactions were truncated at 8 Å.

ii) The same minimization settings as before was used, however, with all ligand and protein atoms fixed instead of only the heavy atoms.

iii) Heating in the NVT ensemble from 0K to 300K over 100 ps using the Langevin thermostat (Loncharich et al. 1992, Pastor et al. 1988), with a collision frequency set to 2 ps^{-1}. The time step for this simulation was of 1 fs. Bonds involving hydrogen atoms were constrained with SHAKE (Ryckaert et al. 1977) for nonwater molecules. Force constant weight of 1000 kcal mol^{-1} Å^{-1} was used to the ligand, protein, and heavy atoms.

iv) Equilibration in the NPT ensemble with a target pressure of 1 bar and 2ps coupling time over 200 ps using the Berendsen barostat (Berendsen et al. 1984). The time step for this simulation was of 2 fs. Bonds involving hydrogen atoms were constrained with SHAKE for nonwater molecules. Force constant weight of 1000 kcal mol^{-1} Å^{-1} were used to the ligand, protein, and heavy atoms.

v) Another simulation in the NVT ensemble with the same previous settings was performed. However, the temperature was decreased from 300K to 100K over 100 ps with 1 fs time step.

vi) Thirteen consecutive minimization simulations where force constant weight was lessened from 1000 kcal mol^{-1} Å^{-1} to 0.5 kcal mol^{-1} Å^{-1} (1000, 500, 200, 100, 50, 20, 10, 5, 4, 3, 2, 1, 0.5) were performed. Previous settings remained the same.

vii) A last NVT ensemble simulation to increase the temperature from 100K to 300K over 400 ps using the same NVT settings as before, however, all force constant weights were removed from the system.

viii) A last 100 ps NPT ensemble simulation using the same NPT settings as before, without force constant weights applied to all atoms of the system, was executed.

The final coordinates of the minimization and equilibration protocol were then used to complete 10 nanoseconds (ns) NPT ensemble production simulations using 2 fs time step, during which the temperature was kept at 300K. SHAKE and a cut-off of 8 Å for the non-bonded interaction were employed. Electrostatic interactions were computed using the Particle Mesh Ewald (PME) (Darden et al. 1993) method. Energies were recorded and coordinates were recorded every 20 ps. A succession of production simulations was executed, using the coordinates of the preceding MD simulation, until completing a total of 1000 ns.

2.3.5 MM-PB(GB)/SA calculations

The Molecular Mechanics Poisson-Boltzmann/Generalized Born Surface Area (MM-PB(GB)/SA) are post-processing end-state approaches, characterized by the use of Poisson-Boltzmann (PB) (Homeyer and Gohlke 2012) and Generalized Born (GB) (Homeyer and Gohlke 2012) methods, to compute free energies of molecules in solution. These methods provide two types of analysis, calculation of the relative stability of multiple conformations of a system and the binding free energy of the noncovalently bound, receptor-ligand complex (Miller III et al. 2012). Binding free energies of the complex are considered by subtracting the unbound receptor and ligand-free energies from the bound complex free energy, as shown in the equation:

$$\Delta G_{binding,solvated} = \Delta G_{complex,solvated} - \left[\Delta G_{receptor,solvated} + \Delta G_{ligand,solvated}\right] \quad (13)$$

The free energy change associated with eq. (13) is approximated by:

$$\Delta G_{solvated} = \langle \Delta G_{MM} \rangle + \langle \Delta G_{solvation} \rangle - T\Delta S \quad (14)$$

where ΔG_{MM} characterizes the change in the molecular mechanics energy upon complexation in the gas-phase, often the molecular mechanical (MM) energies from the force field, the solvation free energies, $\Delta G_{solvated}$ are calculated using the implicit solvent model, and $T\Delta S$ is the change of conformational entropy associated with ligand binding. ΔS, the entropic contribution, is predicted using known approximations or by using normal mode analysis (Miller III et al. 2012). The average interaction energies of receptor and ligand are usually obtained by performing calculations on an ensemble of uncorrelated snapshots collected from an equilibrated molecular MD or Monte Carlo (MC) simulation. The interaction energy and solvation free energy for the complex, receptor and ligand and the results average to obtain an estimate of the binding free energy were calculated using the MMPBSA.py script (Miller III et al. 2012) available through the AMBER distribution. The molecular mechanics free energy (ΔG_{MM}) is decomposed as:

$$\Delta G_{MM} = \Delta G_{ele} + \Delta G_{vdW} \quad (15)$$

where ΔG_{ele} is the electrostatic and ΔG_{vdW} non-electrostatic (hydrophobic) contributions. The solvation free energy $\Delta G_{solvation}$ arises from the sum of the polar (ΔG_{PB}) and nonpolar solvation (ΔG_{SA}) as shown:

$$\Delta G_{solvation} = \Delta G_{PB} + \Delta G_{SA} \quad (16)$$

ΔG_{PB} was computed by solving the linearized PB equation using Parse radii and a solvent probe radius of 1.4 Å. In this work, the dielectric constant was set to 1.0 for the interior of solutes (interior of protein) and 80.0 for the solvent. ΔG_{SA} was determined using a solvent accessible surface area (SASA) term as in:

$$\Delta G_{SA} = \gamma \times SASA + \beta \quad (17)$$

where γ is the surface tension proportionality constant and was set to 0.00542 kcal/(mol·Å^{-2}), and β is the offset value, set to 0.92 kcal/mol. Usually, the binding free energy described would require three independent MD simulations of the complex, proteins, and ligands. However, approximations are made that no significant conformational changes occur upon binding so that the snapshots for all three species can be obtained from a single trajectory, the so-called single trajectory approach. We extracted 1 snapshot every 2 ns from the 1000 ns MD simulation. However, calculation of the entropy contribution to binding were not performed.

The binding energies were also decomposed into contributions of individual residues using the MMPBSA.py script by applying the so-called per-residue decomposition.

2.3.6 Thermodynamic integration

TI simulations were carried out using AMBER 14.0 implementation of PMEMD. In PMEMD, the dual-topology approach is implemented (Kaus et al. 2013). When dual-topology is chosen, the system is prepared in a way that the two complete versions (initial state and final state) of the changing group coexist at every λ (Pearlman and Rao 1998). The functional form of the potential energy used by PMEMD is

$$V(q,\lambda) = V_{common}(q) + (1-\lambda)V_{i,perturbed}(q,\lambda) + \lambda V_{f,perturbed}(q,\lambda) \qquad (18)$$

where V_{common} is the potential for the unperturbed atoms, $V_{i,perturbed}$ and $V_{f,perturbed}$ are the potentials that correspond to the initial and final stated for the perturbed part of the system, and q denotes the $3N$ atomic coordinates (Kaus et al. 2013). Since singularity problems can occur, in which the value of $\langle \partial V(\lambda)/\partial \lambda \rangle$ diverges when λ approaches to zero or one, the use of softcore van der Waals and electrostatic terms may be included to improve the efficiency and stability of the simulations (Shirts and Pande 2005, Steinbrecher et al. 2011). For softcore simulations the potential energy form is

$$\begin{aligned} V(q,\lambda) = V_{common}(q) + V_{i,bsc}(q,\lambda) \\ + (1-\lambda)\left[V_{i,nbsc}(q,\lambda) + V_{i,perturbed}(q,\lambda)\right] + V_{f,bsc}(q,\lambda) \\ + \lambda[V_{f,nbsc}(q,\lambda) + V_{f,perturbed}(q,\lambda)] \qquad (19) \end{aligned}$$

where V_{bsc} is the potential for the bonded interactions of the softcore atoms and V_{nbsc} is the potential for the nonbonded interactions including the softcore atoms (Kaus et al. 2013). Both potentials may be used in a single-step transformation, or may use only the van der Waals softcore in a multistep transformation. In single-step transformations, electrostatic and van der Waals forces are simultaneously modified. In a multistep transformation, these properties are changed in separate calculations, where adding and removing charges receive their own step. In the first step, the atomic partial charge of the disappearing atom from the initial ligand is removed linearly from $\lambda = 0$ to $\lambda = 1$. In the second step, the ligand vdW-transformation is performed using the van der Waals softcore. The initial ligand is decoupled from its surroundings from $\lambda = 0$ to $\lambda = 1$, while simultaneously the atom from the final ligand is built up. In AMBER, the van der Waals softcore potential is a different form of the LJ-equation, specifically designed for better convergence of TI calculations in the case of appearing or disappearing atoms:

$$V_{V_0,disappearing} = 4\varepsilon(1-\lambda)\left[\frac{1}{\left[\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6}\right]$$

$$V_{V_1,appearing} = 4\varepsilon\lambda\left[\frac{1}{\left[\alpha(1-\lambda) + \left(\frac{r_{ij}}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha(1-\lambda) + \left(\frac{r_{ij}}{\sigma}\right)^6}\right] \quad (20)$$

where ε is depth of the potential well, σ is the finite distance at which the inter-particle potential is zero, and r_{ij} is the distance between particles. The parameter α, which is characteristically 0.5, is used to specify the λ-dependent limit and this to prevent singularity effects during the simulations (Shirts and Pande 2005). Finally, the atomic partial charge of the newly constructed atom is switched on linearly form $\lambda = 0$ to $\lambda = 1$ (Figure 35). The total free energy from a multistep approach is then:

$$\Delta G = \Delta G_{decharge} + \Delta G_{vdW\ bonded} + \Delta G_{charge} \quad (21)$$

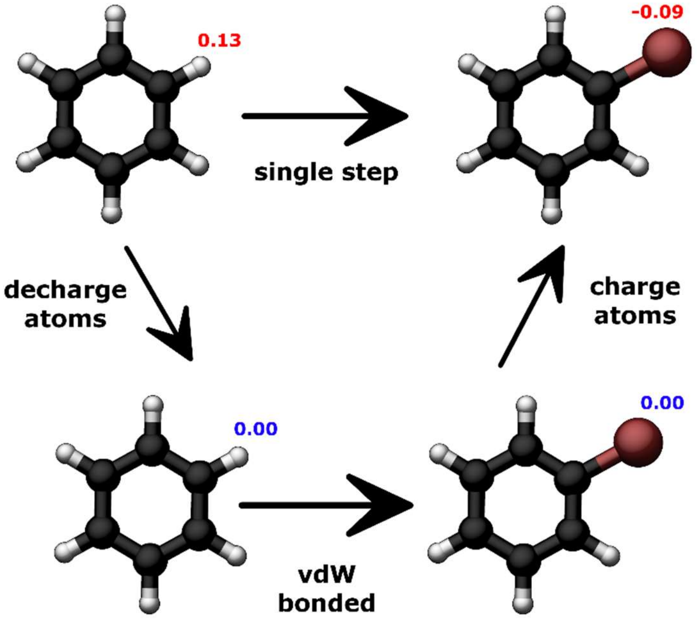


Figure 35: Schematic representation of single step and multistep approach. In a multistep approach three independent simulations are necessary: decharge the atoms, vdW transformation, and charge the atoms.

Relative binding energies of the potential cruzain and rhodesain inhibitors were calculated using a dual-topology and multistep approach. A thermodynamic cycle was used to calculate the relative free energies, where simulations of the ligands bound to the receptor and the ligands in solution were performed. Ligand and complex preparation parameters were maintained the same as the ones used in the MD simulations (For further details, see Appendix C). Both systems, ligand in solution and ligand bound to the receptor, were prepared with a truncated octahedral periodic box with a minimum distance of 12 Å between any box edge and any solute atom. For each step and simulated system, the following equilibration protocol was applied:

i) Initial coordinates were minimized using 1000 steps of steepest descent minimization at $\lambda = 0.5$.

ii) Each simulation was heated to 300K over 250 ps.

iii) Equilibration to adjust the density over 250 ps at NPT ensemble.

TI simulations were performed with PME periodic boundary conditions for long-range electrostatics, a cutoff of 8 Å for nonbonded interactions, Langevin thermostat with a collision frequency of 2 ps^{-1}, an isotropic pressure scaling and a time constant of 1 ps. For the two systems, the ligand in solution and the ligand-receptor complex, the simulation time to each λ-window was set to 0.5 ns for switching the charge on and off. The vdW-transformation step was assigned to take 1 ns to each λ-window. Simulations were performed for every $\Delta\lambda = 0.05$, resulting in 21 λ-windows for each ligand transformation. The trapezoid rule was chosen as the numeric integration method. Under the trapezoid rule, the lambda weights in Eq. (12) are

$$w_1 = w_m = \frac{1}{[2(n-1)]} \qquad and \qquad w_{m\neq 1,n} = \frac{1}{(n-1)} \tag{22}$$

Forward (A→B) and backward (B→A) transformations were calculated. The forward transformations used the same initial structure. The backward transformations used the last snapshots of the $\lambda = 1$ simulations as starting structures for the simulations. The final results are shown as the average of the forward and backward transformations.

The statistical error in TI can be calculated by the total variance for TI over the entire interval as a weighted sum of the variances:

$$\sigma_{\Delta G} = \sqrt{\sum_m w_m^2 \sigma_m^2} \tag{23}$$

where σ_m is the standard error of the mean for the $\partial V/\partial\lambda$ values of the m-th window. The standard error for each window can be estimated as:

$$\sigma = \sigma_{\partial V/\partial\lambda} \cdot \sqrt{2\tau/t_S} \tag{24}$$

where $\sigma_{\partial V/\partial\lambda}$ is the standard deviation, τ is the autocorrelation time of $\partial V/\partial\lambda$, and t_S is the total length of the simulation (Straatsma et al. 1986). The autocorrelation function used was the one defined in Steinbrecher et al (Steinbrecher et al. 2011) work as,

$$R(t) = \frac{1}{n-t}\sum_{i=1}^{n-t}\left[\left(\frac{\partial V}{\partial\lambda}\right)_i - \left\langle\frac{\partial V}{\partial\lambda}\right\rangle\right]\cdot\left[\left(\frac{\partial V}{\partial\lambda}\right)_{i+t} - \left\langle\frac{\partial V}{\partial\lambda}\right\rangle\right] \tag{25}$$

The autocorrelation function $(R(t))$ was evaluated for continuous points over the length of the simulation. From $R(t)$, the estimated correlation time (τ) was:

$$\begin{aligned}\tau &= \int_0^\infty R(t)/R(0)dt \\ &\simeq \int_0^1 R(t)/R(0)dt + \int_1^\infty e^{-t/\tau}dt \\ &= \int_0^1 R(t)/R(0)dt + \tau e^{-1/\tau}\end{aligned} \tag{26}$$

The Root-Mean-Square Error (RMSE) between the experimental and calculated free energies was also computed and determined by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{27}$$

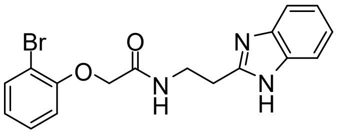
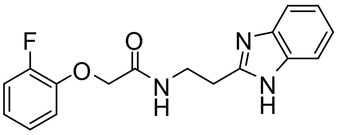
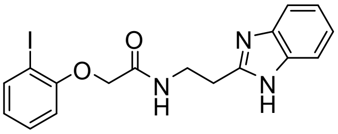
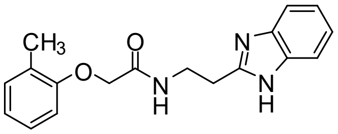
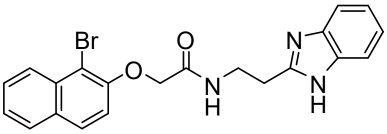
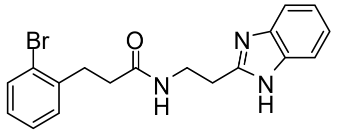
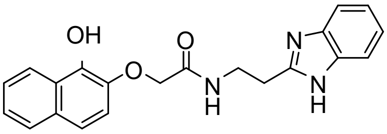
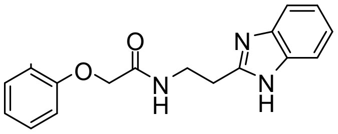
where N is the number of TI transformations, x is the value experimentally obtained experimental measure, and $\hat{x}$ is the one computed free energy by TI calculations.

In a previous work, compound B95 was identified as a powerful competitive cruzain inhibitor (Ferreira et al. 2010), inspiring the synthesis of structurally similar compounds to establish a structure-activity relationship between this class and cruzain, published in Ferreira et al. (2014). Afterward, these same compounds were also evaluated against rhodesain. Despite the high structural similarity, the compounds potency against the enzymes are variable (Table 21). These compounds were used in the TI calculations. To estimate the relative free energy of binding between a pair of molecules using TI outcomes, we used the IC_{50} values of the transforming compounds as

$$\Delta\Delta G_{experimental} = -RT \ln \frac{IC_{50,2}}{IC_{50,1}} \tag{28}$$

where $IC_{50,1}$ is the IC_{50} value of the initial compound and $IC_{50,2}$ is the IC_{50} value of the transformed compound, R is the Universal constant of gases and T the absolute temperature.

Table 21: Inhibitory activity of B95 analogs against the rhodesain and cruzain enzymes.

Code	Structure	Inhibition against Cruzain 100 µM (%)[a]	Cruzain IC_{50} (µM)[b]	Inhibition against Rhodesain 100 µM (%)[a]	Rhodesain IC_{50} (µM)[b]
MAD558 (B95)		89 ± 2	0.8	93 ± 6	2.7
MAD619		92 ± 0	5 ± 1	97 ± 13	0.82 ± 0.08
MAD644		96 ± 3	1.6	88 ± 8	9.4 ± 2.0
MAD554		88 ± 1	4	84 ± 8	6.5
MAD597		92 ± 0.6	0.2	88 ± 5	0.25 ± 0.21
MAD574		51 ± 0.5	78	86 ± 9	43.7 ± 11.0
MAD790		51	8.2 ± 0.9	82 ± 4	4.4 ± 0.7
MAD700		88	10.9 ± 1	54 ± 8	2.4 ± 0.5

[a] The percent inhibition values of the compounds are the average of at least three measurements, and the errors are within 10%. [b] The IC_{50} values of the compounds against cruzain and rhodesain were independently determined by obtaining rate measurements in triplicate for at least six inhibitor concentrations. The values represent the means of a least three individual experiments.

2.4 RESULTS AND DISCUSSION

2.4.1 Cruzain dyad simulations

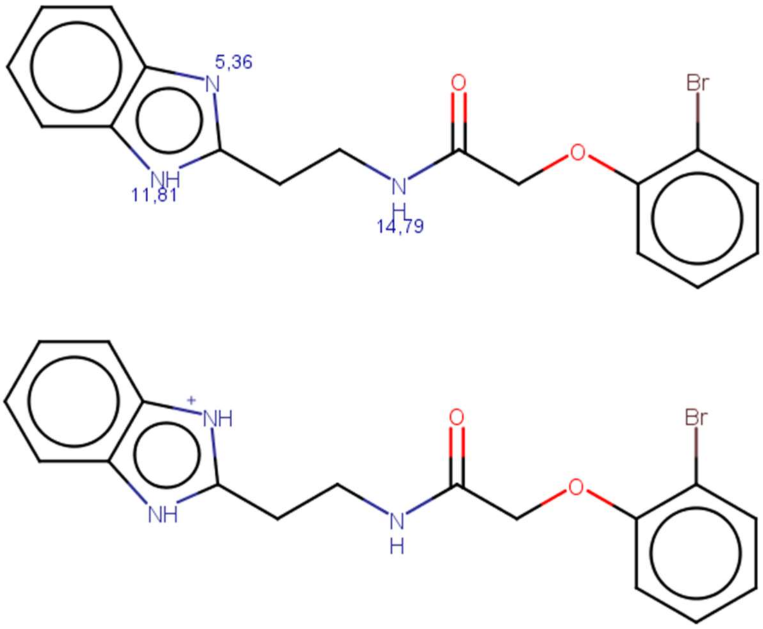
Cysteine and histidine form the catalytic dyad in the active site of cysteine proteases, including cruzain and rhodesain. The dyad is believed to exist as a thiolate-imidazolium ion pair where His162-NH$^+$/Cys25-S$^-$, in this work represented as H162q-Cys25q, in the free enzyme (Storer and Ménard 1994, Otto and Schirmeister 1997). However, recent studies suggest other catalytic mechanism (Sárkány et al. 2001, Shokhen et al. 2009). In the work of Shokhen et al. (2009), the authors suggested the catalytic dyad as His162-NH$^+$/Cys25-SH being the cysteine in its neutral state, in this work represented as H162q-Cys25n. Since X-ray crystallography cannot directly identify the actual protonation state of the catalytic dyad, unless the structures very high resolution, we have chosen to simulate both states.

We have also extended the protonation study to the ligand B95 bound to the cruzain structure found in the crystal structure 3KKU (Figure 36). We submitted the compound to pKa prediction programs and online servers in an attempt to determine its protonation state at pH of 5.5. However, the predictions were inconclusive, and an assumption of which protonation state of benzimidazole ring presents in the ligand to use was indecisive (Table 22). Therefore, we have chosen to simulate both protonation states of the ligand as well (Figure 37).

Table 22: pKa predictions of the nitrogen in the benzimidazole ring from ligand B95.

Program/Server	pKa of N
MOKA	6.05
Epik (Schrödinger)	5.95 +/-0.73
ACE and JChem acidity and basicity calculator	5.40
Marvin	5.36

In order to assess the interactions between receptor-ligand complex, 1000 ns MD simulations were performed in four systems: H162q-Cys25q with B95 protonated; H162q-Cys25q with B95 unprotonated; H162q-Cys25n with B95 protonated; and H162q-Cys25n with B95 unprotonated.



```
 pH      %-1      %-2
0.00     0.00    100.00
0.50     0.00    100.00
1.00     0.00    100.00
1.50     0.01     99.99
2.00     0.04     99.96
2.50     0.14     99.86
3.00     0.43     99.57
3.50     1.35     98.65
4.00     4.15     95.85
4.50    12.03     87.97
5.00    30.19     69.81
5.50    57.77     42.23
6.00    81.22     18.78
6.50    93.19      6.81
7.00    97.74      2.26
7.50    99.27      0.73
8.00    99.75      0.23
```

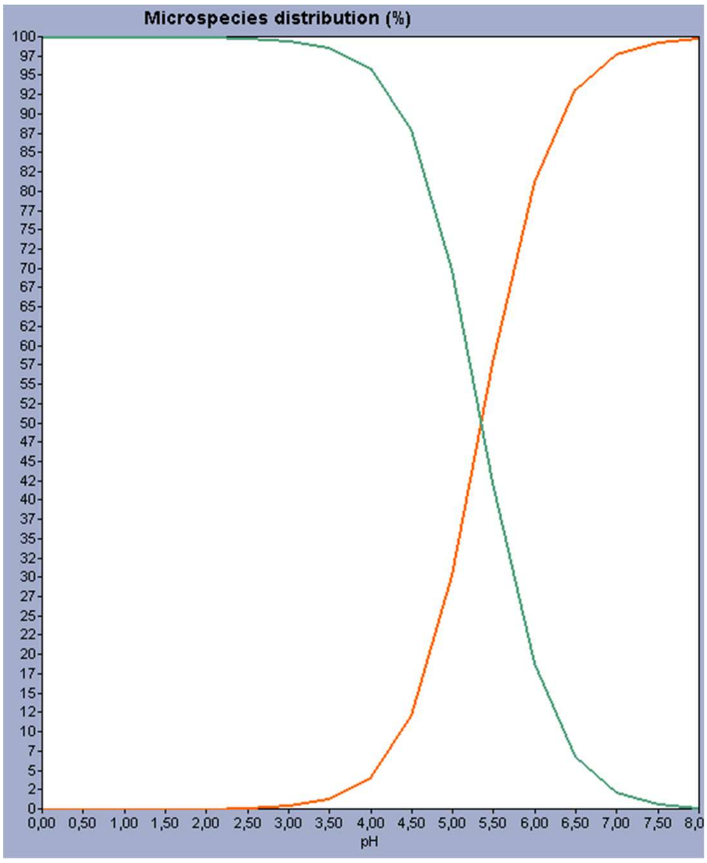


Figure 36: Outcome of protonation prediction of ligand B95 from the program Marvin. The program calculates the probability of the states according to the pH, where 1 is the unprotonated state of the ligand and 2 is the protonated state. Marvin predicted that at the pH 5.5 the ligand B95 would be in the unprotonated state.

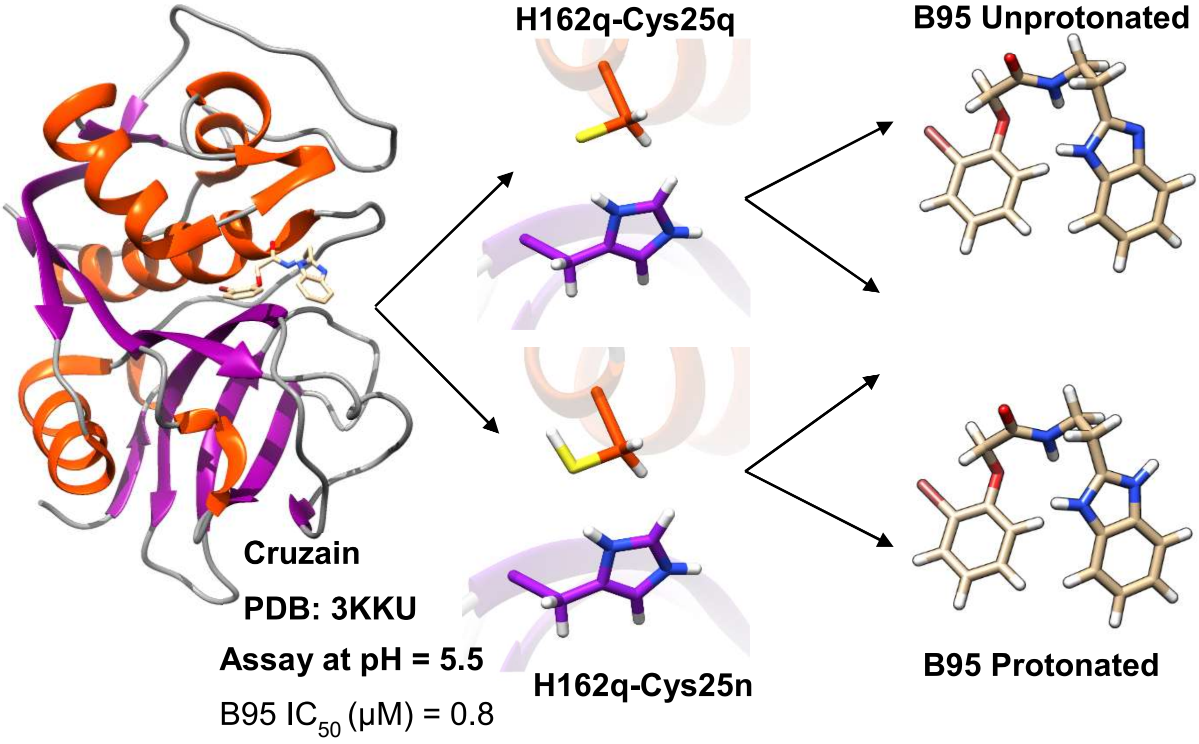


Figure 37: Graphical representation of ligand B95 in the cruzain active site and protonation states simulated.

Ligand RMSD was calculated between simulation frames and the crystallographic ligand position found in 3KKU (Figure 38). All simulations were carried through 1000 ns with the ligand within the active site of cruzain, except the simulation of B95 protonated in the H162q-Cys25n pair. In this simulation, the ligand exited the active site at approximated 500 ns, suggesting that the presence of the hydrogen atom in both the protein cysteine and protonated benzimidazole ring of the ligand is not favorable. The protonation state of B95 may also be corroborated by the fact that the protonated ligand in the H162q-Cys25q system produced conformations closer to the crystallographic position of the ligand than the unprotonated one in the same system. On the other hand, the unprotonated ligand, in the H162q-Cys25n system, was the one with lower deviation to the X-ray conformation. The peaks in the RMSD can be associated with the flexibility of the ligand. The benzimidazole part of the ligand were constantly moving, while the bromobenzene part were offered less deviation from the crystallographic positon (Figure 39).

Binding free energies were analyzed from the MD simulations, using MMPBSA.py, and provided favorable results to all systems (Table 23). The protonated ligand in the H162q-Cys25q (ionic pair) system achieved the lowest total free energy of binding, however it was not much lower than for the other systems.

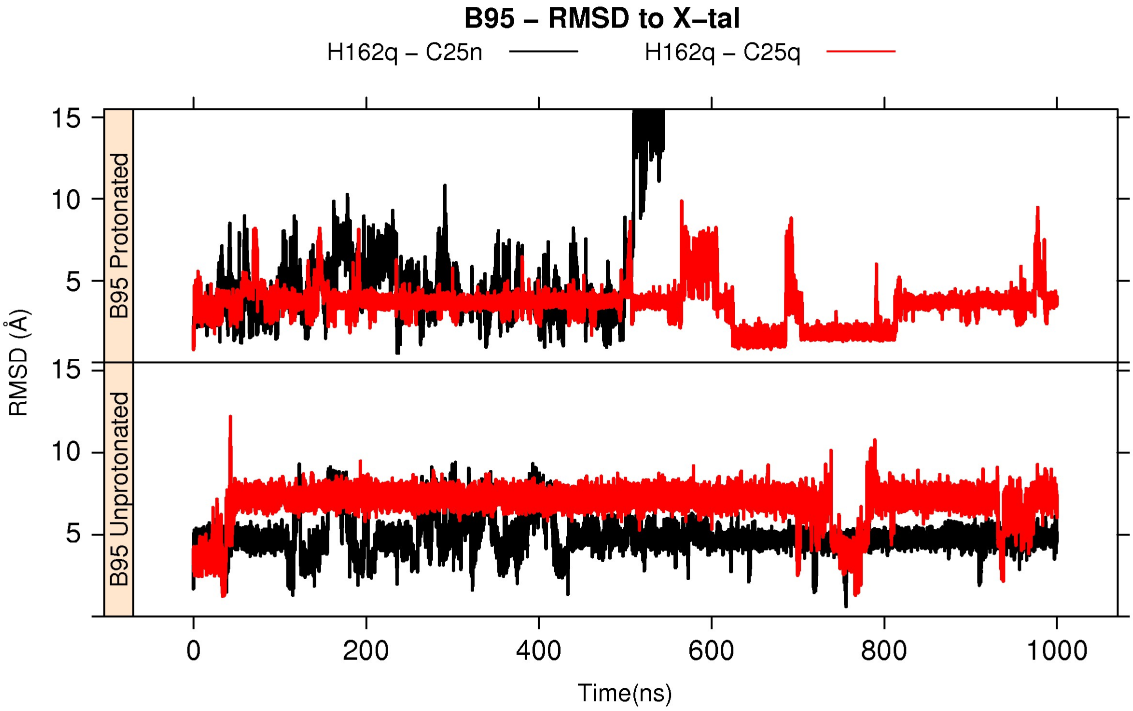


Figure 38: Heavy-atom RMSD between the simulated ligand states and the crystallographic ligand of 3KKU. a) the protonated ligand in the H162q-C25q (red) system offered a more stable and close to the crystallographic ligand than H162q-C25n (black); b) the unprotonated ligand in the H162q-C25n (black) was the one that showed conformations closer to the crystallographic ligand than the H162q-C25q (red).

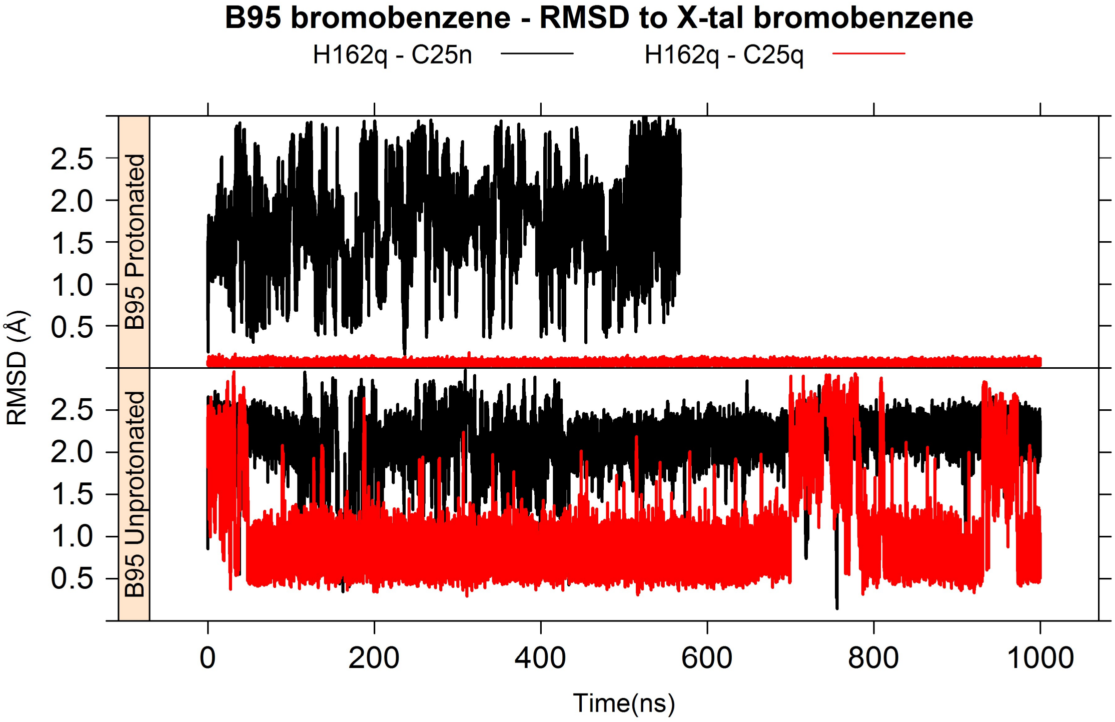


Figure 39: Heavy-atom RMSD between the bromobenzene in the ligands and the crystallographic ligand bromobenzene.

On the other hand, for the H162q-Cys25n system contributions were calculated up until the ligand left the active site, meaning that the binding free energy could be higher. What seems to distinguish between the systems can be the electrostatic energy and solvation free energy contributions. This expresses the energy that the system has to use to de-solvate the binding particles and to align their binding interfaces. In systems where the ligand was in the protonated form, the electrostatic energy was extremely negative and the solvation free energy extremely positive, almost canceling each other (Table 23). Since no entropy contribution to binding was calculated, these results do not equal to actual free energy of binding and were calculated aiming to establish a comparison between the systems.

The amino acid free energy contributions decomposition showed that Asp 161, Hip 162, and Glu 208 were key amino acids to electrostatic contribution in the binding site of B95 protonated (Figure 40). The difference between electrostatic contributions in these systems was exactly the catalytic cysteine, which offered stronger electrostatic interaction in H162q-Cys25q than in H162q-Cys25n. While systems where the ligand was unprotonated, van der Waals component to the interaction energies was the major contributor. Amino acid free energy contributions decomposition showed that Leu 67, Leu 160, Asp 161, and Hip 162 were key amino acids to van der Waals contribution to the binding site of B95 unprotonated (Figure 41). However, Leu 160 was the main contributor in the H162q-C25n, while Hip 162 was the one in the H162q-C25n. This could also be associated with the protonation state of the cysteine, as it was the single difference in this region.

Table 23: Averaged binding free energies decomposed in contributions and calculated by MMPBSA.py.

	Average energy component (kcal/mole)						
System	vdWaals	Elec	EPB	EPolar	ΔG$_{gas}$	ΔG$_{solv}$	ΔG$_{total}$
H162q-Cys25q B95 Protonated	-30.7	-289.2	294.8	-3.1	-319.9	291.7	-28.2
H162q-Cys25q B95 Unprotonated	-31.4	-19.2	30.2	-3.1	-50.6	27.2	-23.4
H162q-Cys25n B95 Protonated	-26.9	-214.7	223.7	-2.8	-241.6	220.9	-20.7
H162q-Cys25n B95 Unprotonated	-37.4	-34.0	50.9	-3.4	-71.4	47.6	-23.8

VDWAALS = van der Waals contribution from MM.
EEL = electrostatic energy as calculated by the MM force field.
EPB = the electrostatic contribution to the solvation free energy calculated by PB.
EPOLAR = nonpolar contribution to the solvation free energy calculated by an empirical model.

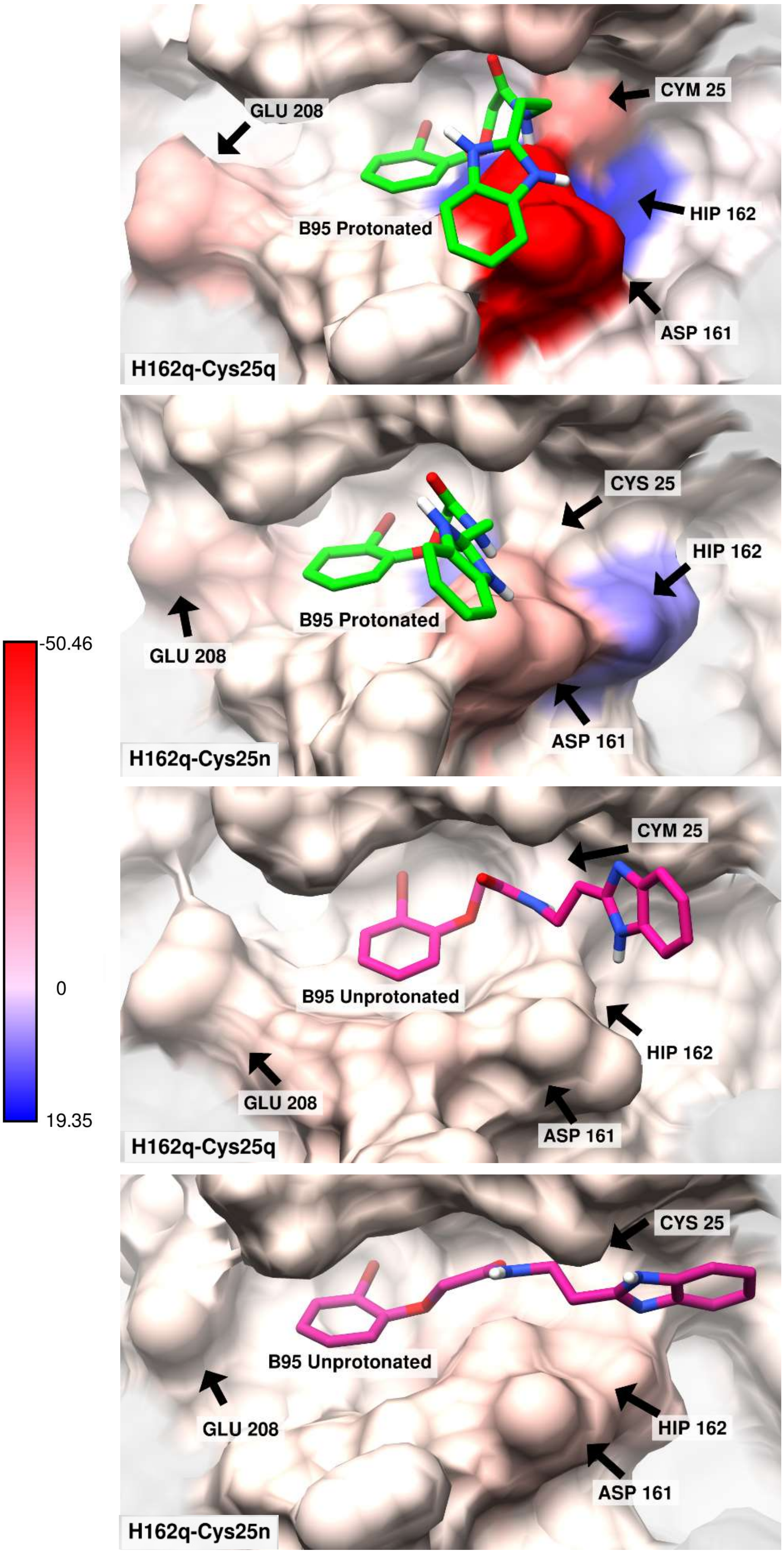


Figure 40: Electrostatic contribution decomposed by residues. Red shows negative contribution and blue shows positive contribution. Protonated ligand is presented in green and the unprotonated one in magenta.

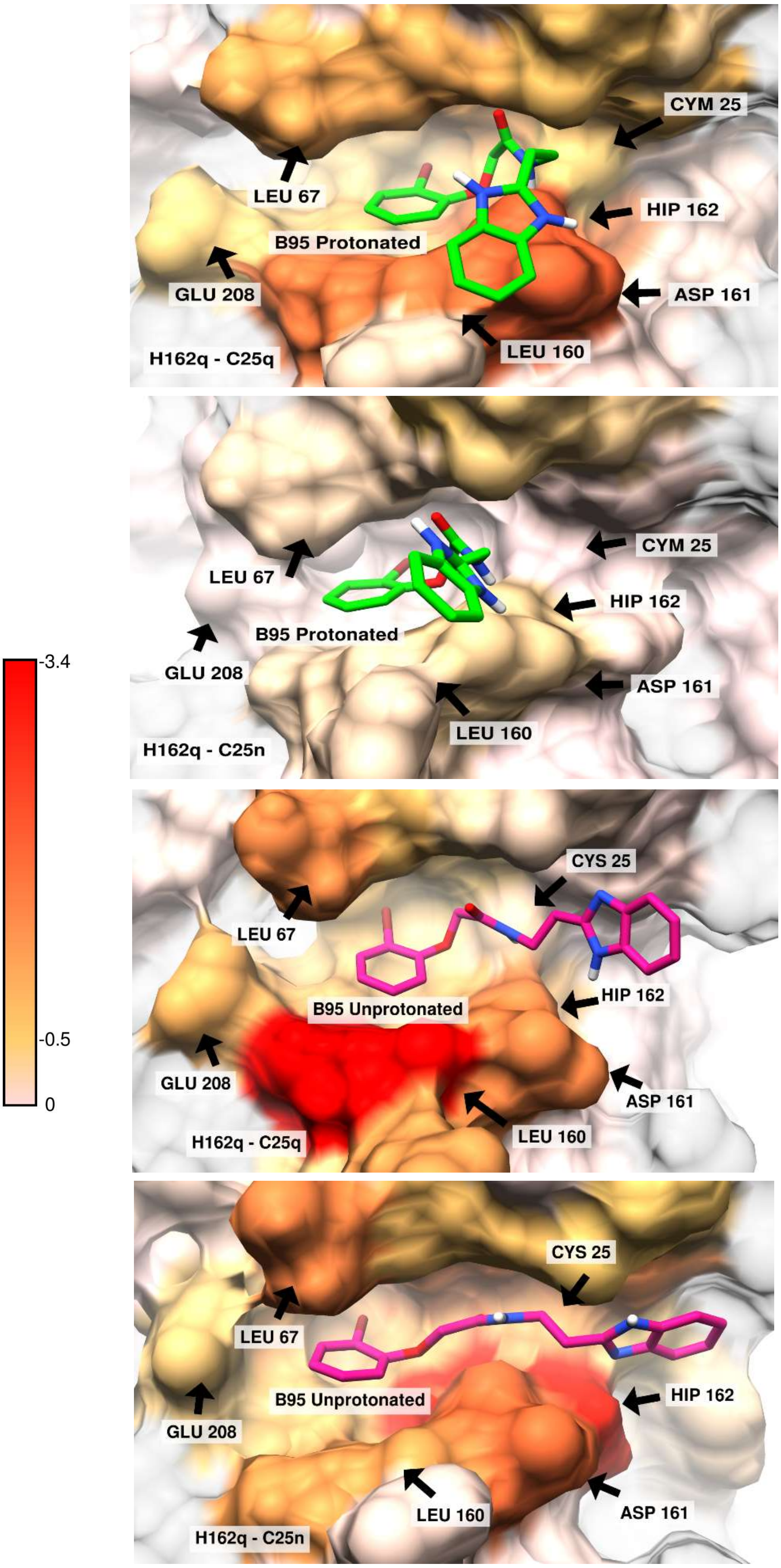


Figure 41: van der Waals contribution decomposed by residues. Red shows negative contribution and white shows no contribution. Protonated ligand is presented in green and the unprotonated one in magenta.

Hydrogen bond interactions between the ligand and residues Asp 161, Gly 66, and Ser 64 were observed for all systems. In the systems where the ligand was protonated, interactions with residue Asp 161 alternated with the ones with Gly 66 throughout the simulation (Figure 42 and Figure 43). However, in the H162q-C25q system this behavior was more noticed than in the H162q-C25n (Table 24). For the unprotonated ligand in the H162q-C25q system, besides the hydrogen bond interaction with Asp 161, it was observed a long-term hydrogen bond interaction with Leu 160 (53.4% of the simulation time) (Figure 44 and Table 24). On the other hand, this interaction was not as long in the H162q-C25n (1.8% of the simulation time), where the longest hydrogen bond interaction was between Asp 161 and His 162 (29.9% of the simulation time) (Figure 45 and Table 24).

The influence of the protonation states of cysteine and histidine could also be observed in the atomic fluctuations of the protein residues in the binding site (Figure 46). Overall, the H162q-C25n systems showed more fluctuations in the amino acids near the ligand than the H162q-C25q systems. Residues Asp 161, His 162 and Cys 25 seemed to be unaffected by the protonated state of the ligand in the H162q-C25n system. While in the H162q-C25q system, they displayed practically the same fluctuation for both protonation states of the ligand (Figure 46). This fact suggested that the state of the cysteine might generate instabilities in the binding site.

Nevertheless, the unprotonated ligand bound to the H162q-Cys25n protein might offer promising outcomes, and the protonated compound provided more favorable energy contributions when bound to the H162q-Cys25q protein. In the following simulations, we have employed the H162q-Cys25q system.

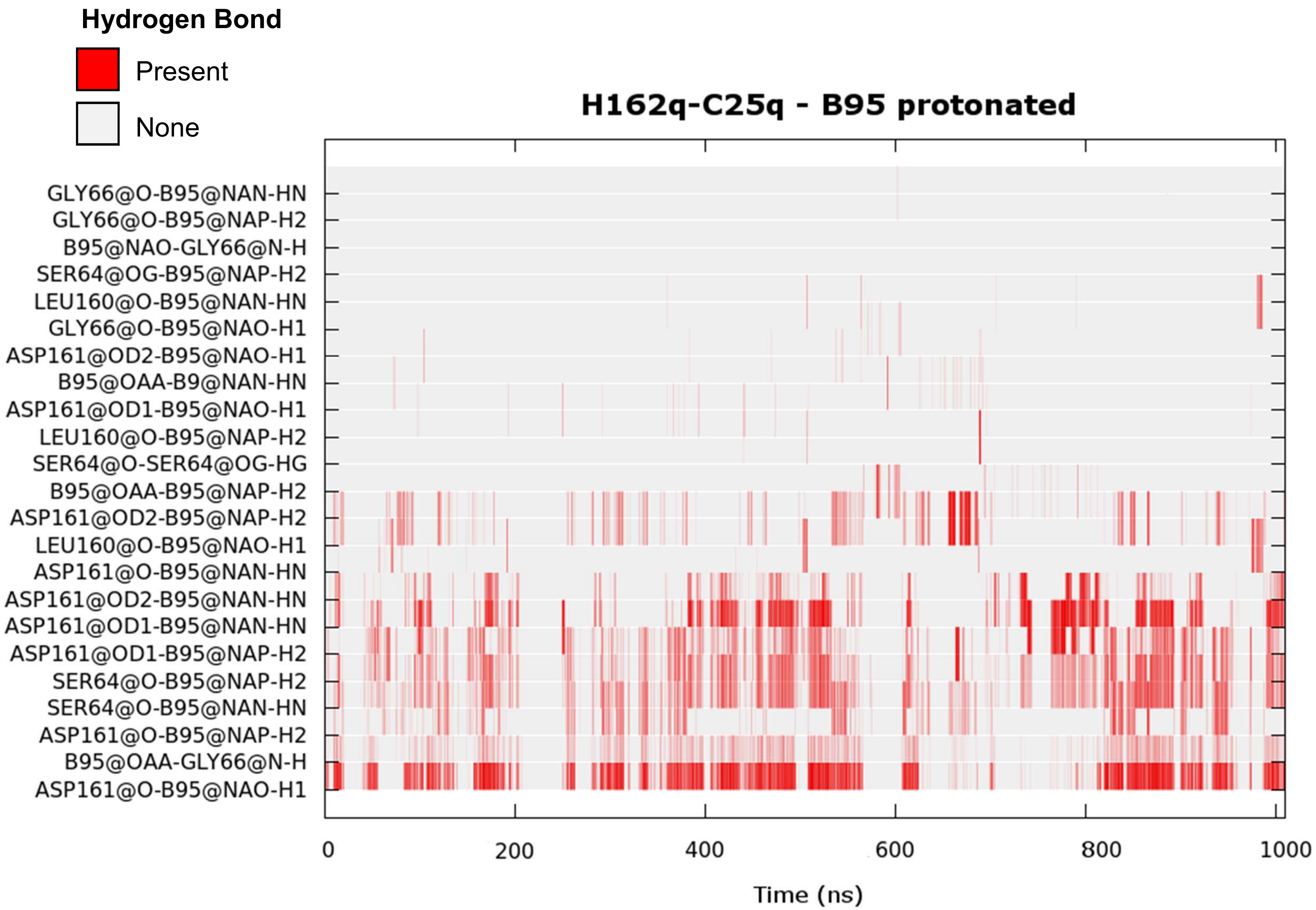


Figure 42: Hydrogen bond analysis of the H162q-C25q system with B95 protonated during the 1000 ns simulation.

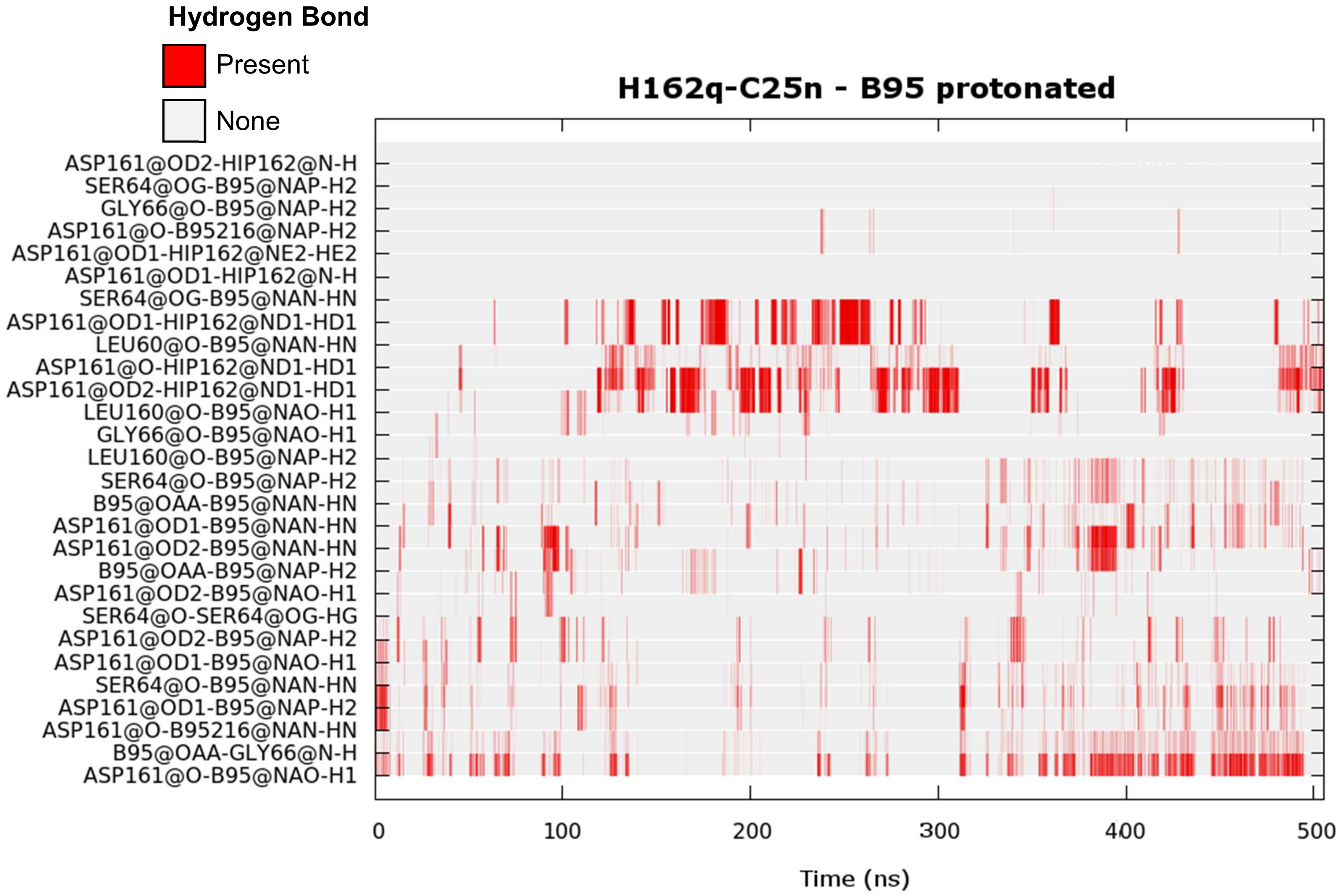


Figure 43: Hydrogen bond analysis of the H162q-C25n system with B95 protonated up until the ligand left the binding site.

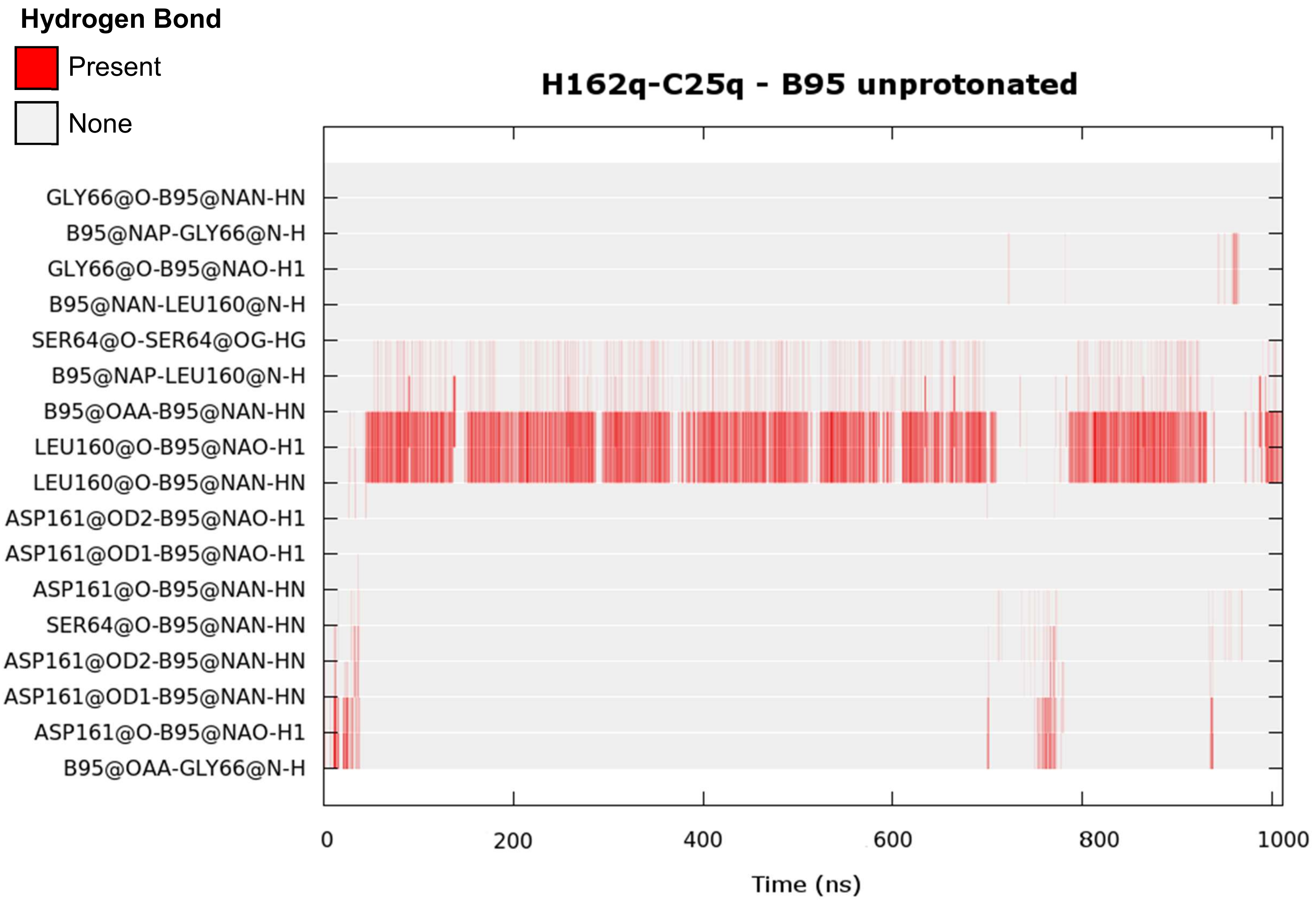


Figure 44: Hydrogen bond analysis of the H162q-C25q system with B95 unprotonated during the 1000 ns simulation.

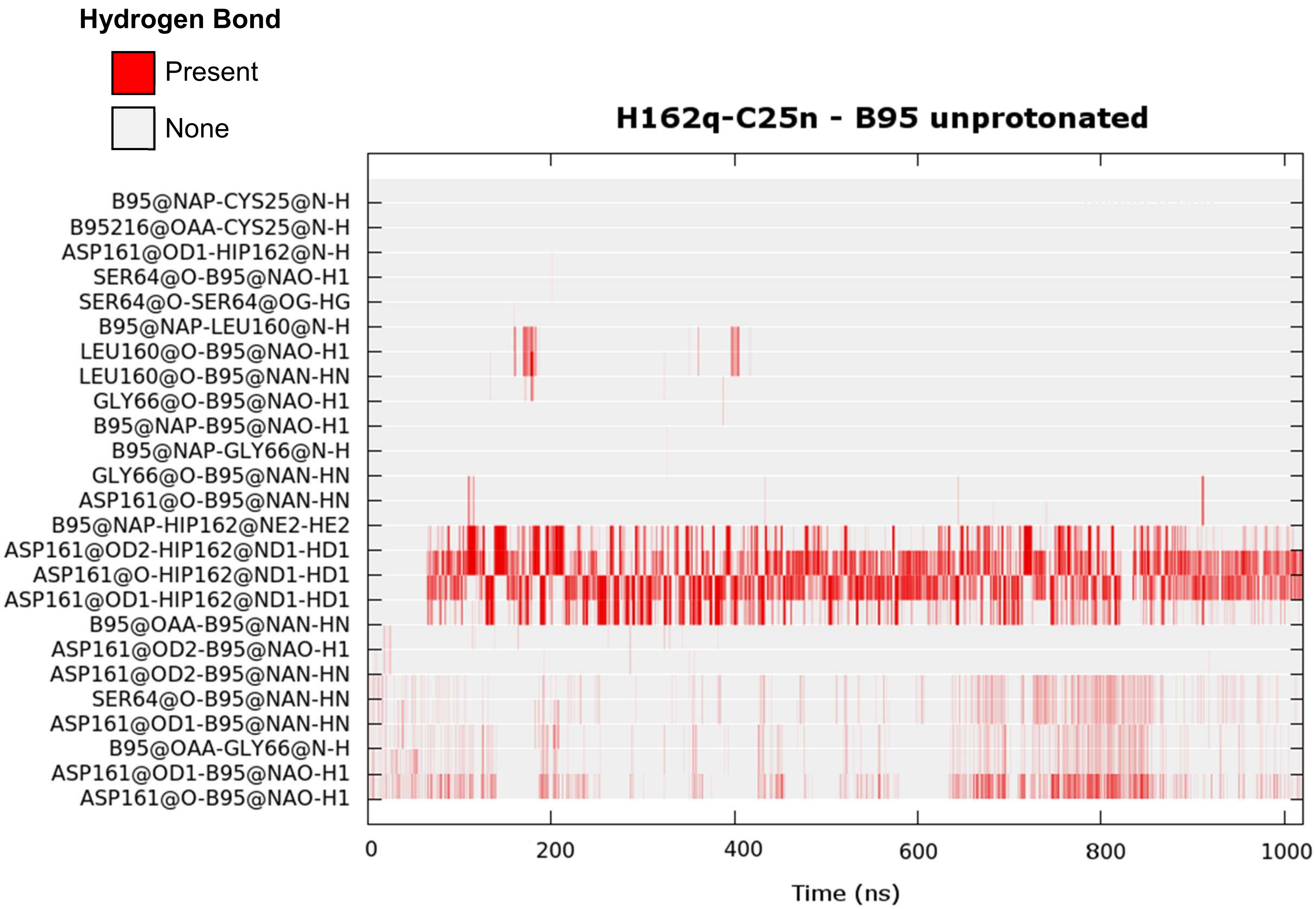


Figure 45: Hydrogen bond analysis of the H162q-C25n system with B95 unprotonated during the 1000 ns simulation.

Table 24: Important protein-ligand and protein-protein hydrogen bond interaction lifetime through simulations.

	Fraction of the simulation time (%)			
	H162q-C25q		H162q-C25n	
H-bond	B95 Protonated	B95 Unprotonated	B95 Protonated	B95 Unprotonated
ASP_161@O-B95@NAO-H1	32.5	2.2	19.0	22.4
B95@OAA-GLY_66@N-H	21.6	0.9	15.1	11.0
SER_64@O-B95@NAP-H2	19.3	-	5.9	-
ASP_161@OD1-B95@NAN-HN	14.5	0.4	5.9	0.6
ASP_161@OD2-B95@NAN-HN	12.8	-	6.7	0.3
SER_64@O-B95@NAN-HN	12.7	0.5	5.4	12.5
ASP_161@OD2-B95@NAP-H2	10.0	-	6.2	-
LEU_160@O-B95@NAO-H1	1.6	53.4	2.8	1.8
B95@NAP-LEU_160@N-H	-	14.5	0.2	-
ASP_161@O-HIP_162@ND1-HD1	-	-	10.0	29.9
ASP_161@OD1-HIP_162@ND1-HD1	-	-	16.7	26.3
ASP_161@OD2-HIP_162@ND1-HD1	-	-	18.3	24.0

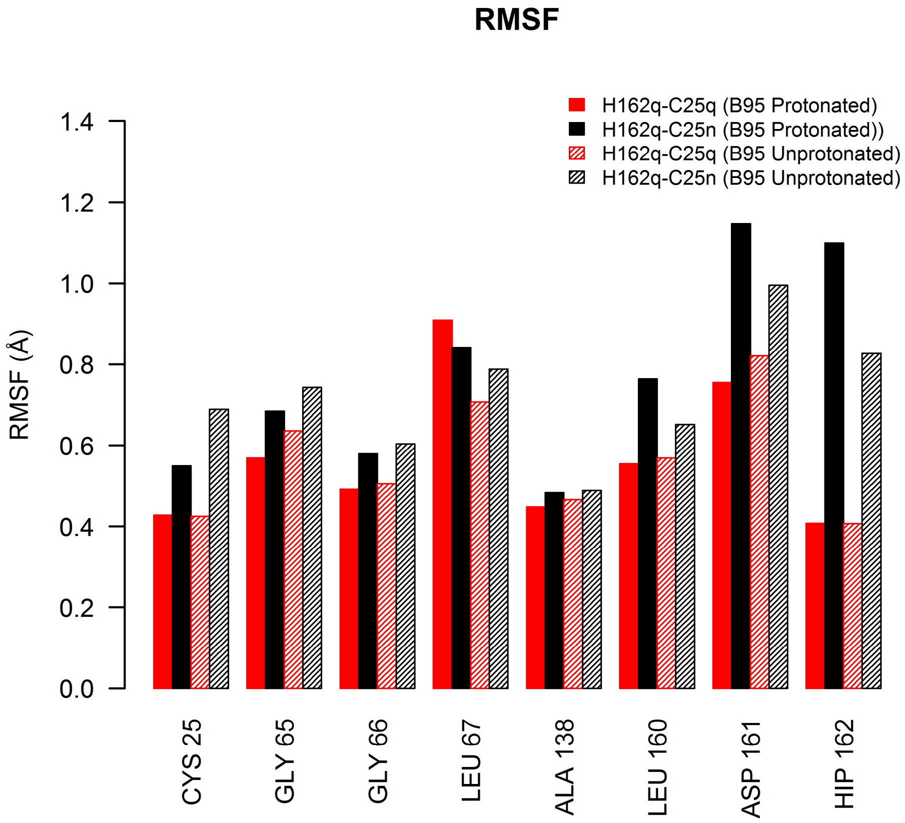


Figure 46: Graphic illustration of the atomic fluctuations of the protein residues in a distance of 6 Å of the center of mass of the ligand. Red represents the H162q-Cys25q systems and in black the H162q-Cys25n systems. Instabilities can be found in H162q-Cys25n systems, especially in residues Cys 25, Asp 161, and Hip 162.

2.4.2 Docking of B95 in Rhodesain

We have docked B95 ligand in both protonated states within the rhodesain binding site. Ten representative structures were extracted using hierarchical clustering on the frames of an MD simulation of the 2P86 crystal without the crystallographic ligand and used as receptors. From each structure, the top-scoring pose was chosen and the two best scoring poses were selected to further analysis (Table 25). However, it is worth to mention that none of the top-scoring pose reproduced the position of the crystallographic B95 in cruzain. The selected poses were submitted to MD simulations in a total of 20 ns for each one. The outcomes were compared with the simulations done with cruzain.

Table 25: GlideScore for docking B95 protonated and unprotonated in rhodesain.

	GlideScore (kcal/mol)	
Stucture	**Protonated**	**Unprotonated**
1	-7,2	-4,7
2	-5,8	-4,6
3	-6,7	-6,5
4	-5,4	-5,3
5	-5,9	-5,6
6	-6,4	-5,8
7	-5,0	-4,9
8	-7,7	-3,5
9	-5,5	-5,7
10	-6,3	-5,2

The two best poses of the protonated ligand in rhodesain belonged to structure 8 (Pose 1) and structure 1 (Pose 2). Curiously, these poses pointed toward different directions in the active site (Figure 47). Pose 1 position may be due to the binding site conformation achieved by the extracted structure from MD without the ligand in the binding site. Whereas, pose 2 presented a position much closer to the crystallographic ligand than pose 1. RMSD plots from the MD simulations showed a stable evolution of the ligands in the active site of rhodesain (Figure 48). However, pose 2 deviations were lower and closer to the crystallographic ligand than pose 1. Since, a small number of residues in the active site of rhodesain differ from those of cruzain, we investigated the

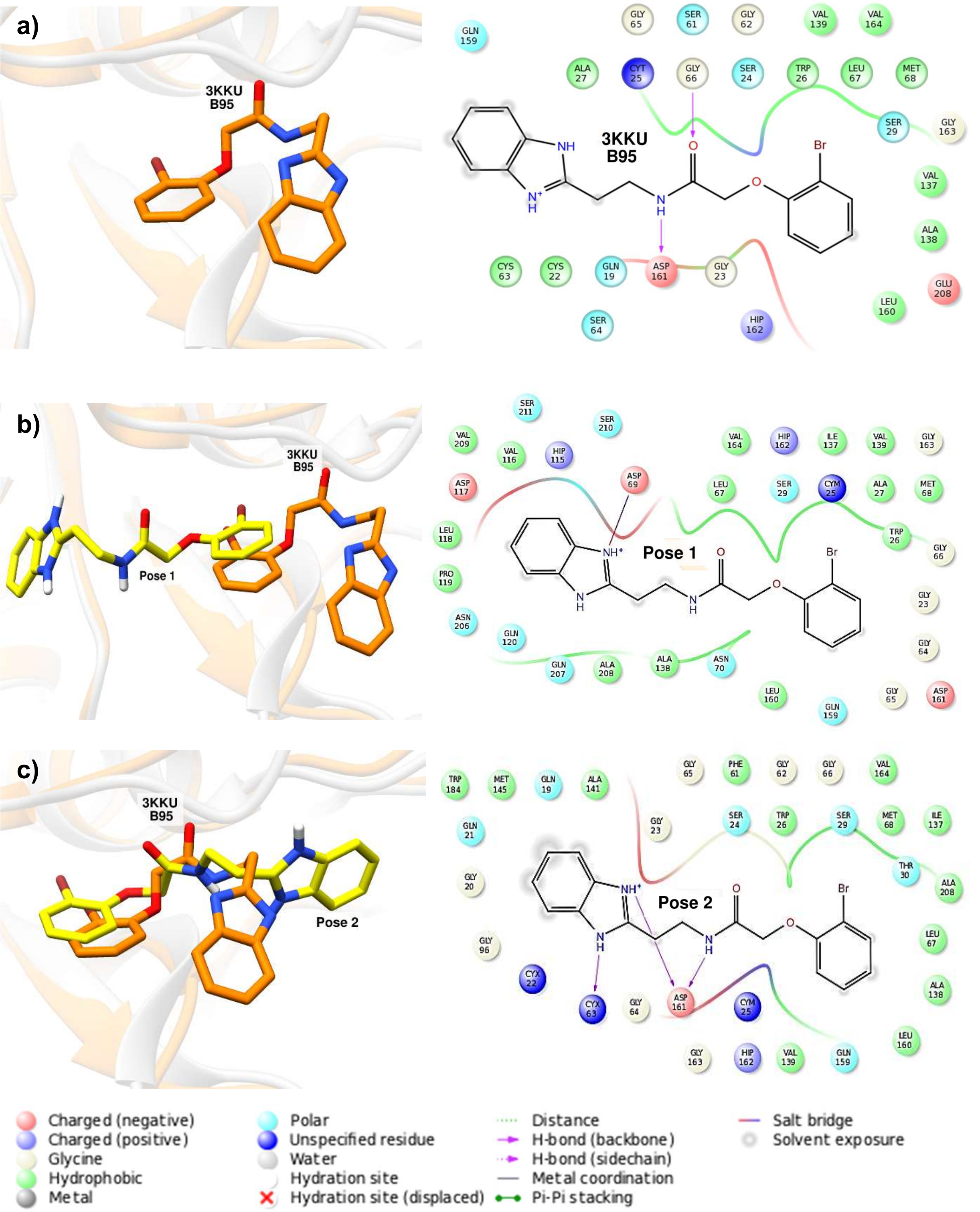


Figure 47: a) Superimposed structures of 3KKU and 2P86 with the B95 ligand in the active site. Map of interactions between the ligand and the protein. b) Pose 1 (yellow) position in relation to the B95 ligand of 3KKU (orange). The pose achieved different interactions from the crystallographic ligand. c) Pose 2 (yellow) position in relation to the B95 ligand of 3KKU (orange). Even though, pose was not similar to the crystallographic ligand interaction with amino acid Asp 161 remained.

interactions formed by ligand-receptor complex (Figure 47). The interaction with Asp 161 that the crystallographic ligand does with cruzain was maintained only in pose 2 for rhodesain, also in the simulation. Hydrogen bond interactions with amino acids Gly 64 and Gly 66 were also observed for pose 2 throughout the MD simulation.

The two best poses of the unprotonated ligand belonged to structure 3 (Pose 1) and structure 6 (Pose 2). Different from the poses with the protonated ligand, the ones with neutral ligand showed similar position between them in the active site; however, conformation was not equivalent to the one corresponding to crystallographic B95 (Figure 49). Similar behavior of the poses was also observed in the RMSD plots from the MD simulations (Figure 48). However, deviations from the crystallographic position of B95 were much higher than the cases in which the ligand was charged. Interactions with amino acids Gly 66, Leu 160, and Asp 161 were identified for both poses during the course of the simulations.

What seemed to be common for the majority of examined poses was the orientation of the bromobenzene ring. In all poses, the ring pointed to the inside of the binding pocket in a similar position B95 as in cruzain. From the cruzain simulations, it appeared that the bromobenzene ring deviated little from the initial position, especially for the protonated ligand (Figure 39). Therefore, the docked poses with this characteristic might be favorable for an initial investigation of the binding mode of B95 in rhodesain. We have chosen both poses 2, from the protonated ligand and the unprotonated ligand, to perform a 1000 ns MD simulation to compare with the already performed cruzain simulations.

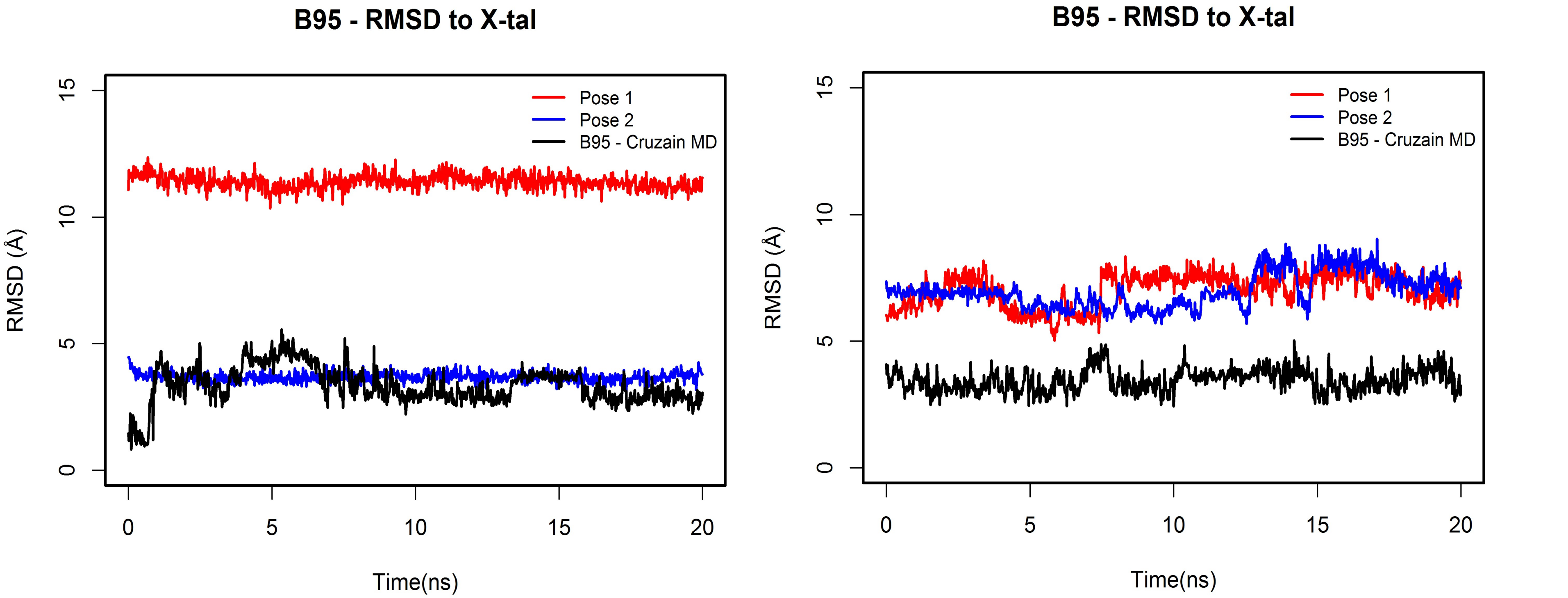


Figure 48: RMSD plots of the docking poses of the protonated and unprotonated ligands compared to the simulation in cruzain with the same ligands.

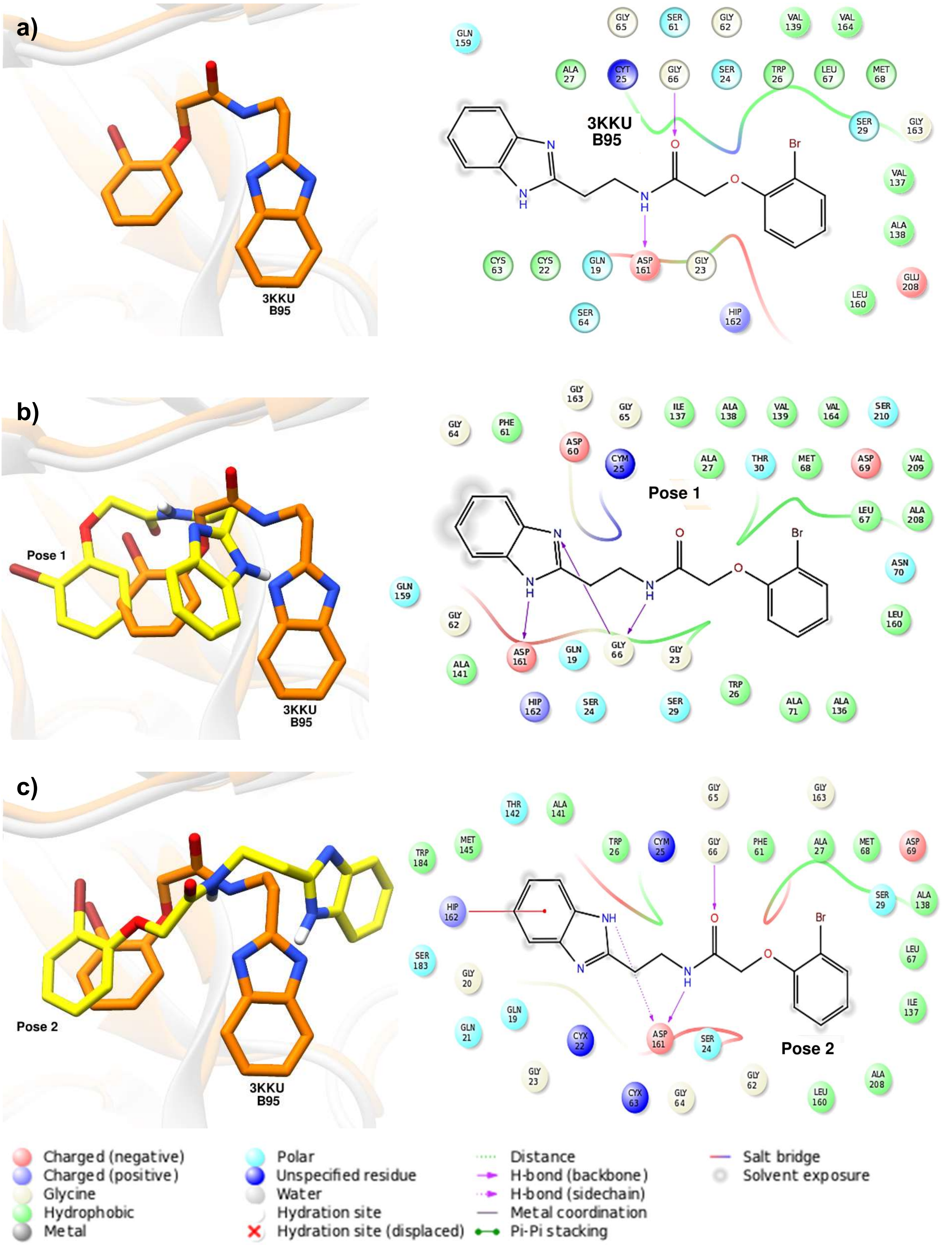


Figure 49: a) Superimposed structures of 3KKU and 2P86 with the B95 ligand in the active site. Map of interactions of the ligand with the protein. b) Pose 1 (yellow) position in relation to the ligand B95 of 3KKU (orange). The pose achieved interactions similar to the crystallographic ligand. c) Pose 2 (yellow) position in relation to the ligand B95 of 3KKU (orange). Even though, the pose was not similar to the crystallographic ligand, interaction with amino acids Asp 161 and Gly 66 remained.

2.4.3 Comparing Cruzain and Rhodesain simulations bound to B95

We performed 1000 ns simulations of both states of B95 (protonated and unprotonated) bound to rhodesain to compare their behavior across the simulation time. Ligand RMSD was calculated between simulation frames and the crystallographic ligand position found in 3KKU (Figure 50). Both ligand states remained in the active site of rhodesain. Similar to the cruzain outcomes the protonated ligand was the one to achieve lower deviation from the crystallographic ligand. Nevertheless, the unprotonated ligand evolution showed high RMSD when compared to B95 crystallographic position, it was very stable. The bromobenzene ring of the ligands were also evaluated for rhodesain simulations (Figure 51). In rhodesain, the bromobenzene ring from the protonated ligand oscillated much more than in the cruzain simulation. This might be due to the replacement of Glu 208 in cruzain for Ala 208 in rhodesain that changes the conformation of the S2 site. In both RMSD plots, it was clear to see that the unprotonated ligand maintained the conformation obtained from docking throughout the simulation, with little deviation.

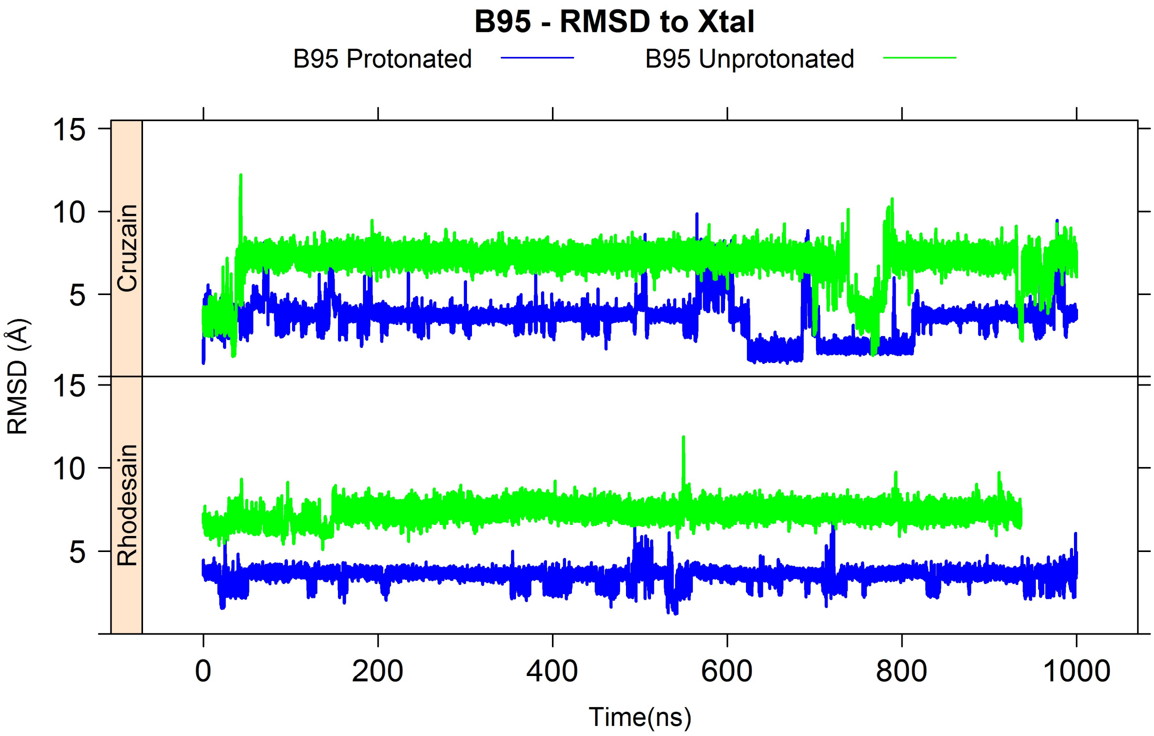


Figure 50: Heavy-atom RMSD between the simulated ligand states and the crystallographic ligand of 3KKU. Among the cruzain simulations the protonated ligand (blue) offered conformations close to the crystallographic ligand than the unprotonated ligand (green); Among rhodesain simulations the protonated ligand (blue) conformations was closer to the crystallographic ligand. However, RMSD of unprotonated ligand showed stable behavior (green).

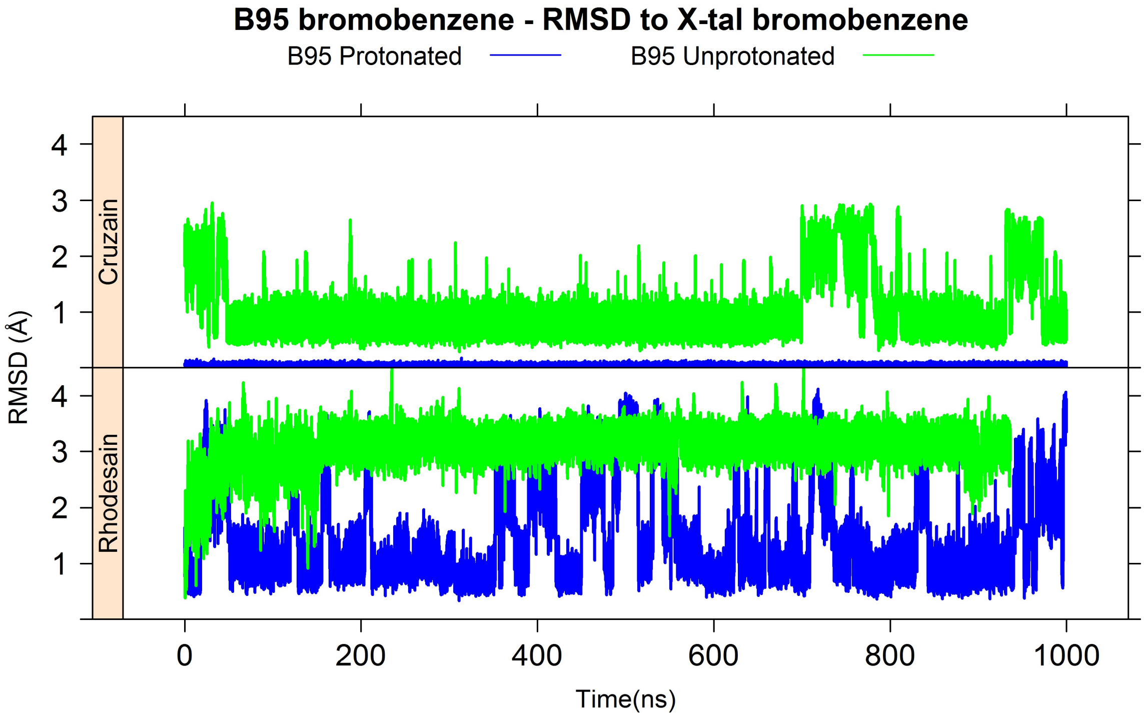


Figure 51: Heavy-atom RMSD between the bromobenzene in the ligands and the crystallographic ligand bromobenzene.

To better distinguish between the ligand conformations sampled by the MD simulations, we extracted from the simulations at regularly spaced intervals 10 clusters using the average distance between members of two cluster as a condition. These clusters had the purpose of group together similar conformations of the ligand. From the cluster, we can observe how long a particular ligand conformation prevailed along the simulation. We used the *nofit* option when clustering, so that the structures were not fitted onto each other prior to calculating RMSD. This way, the cluster is done in the conformation achieved in the MD simulation. Nevertheless, we asked for ten clusters; all simulations could be separated into two clusters at maximum. Especially in the simulations with the protonated ligand, it was clear by clustering that the ligand shifted from one conformation to the other all through the simulation (Figure 52). Whereas, simulations with unprotonated ligand did not show as many shifts (Figure 53).

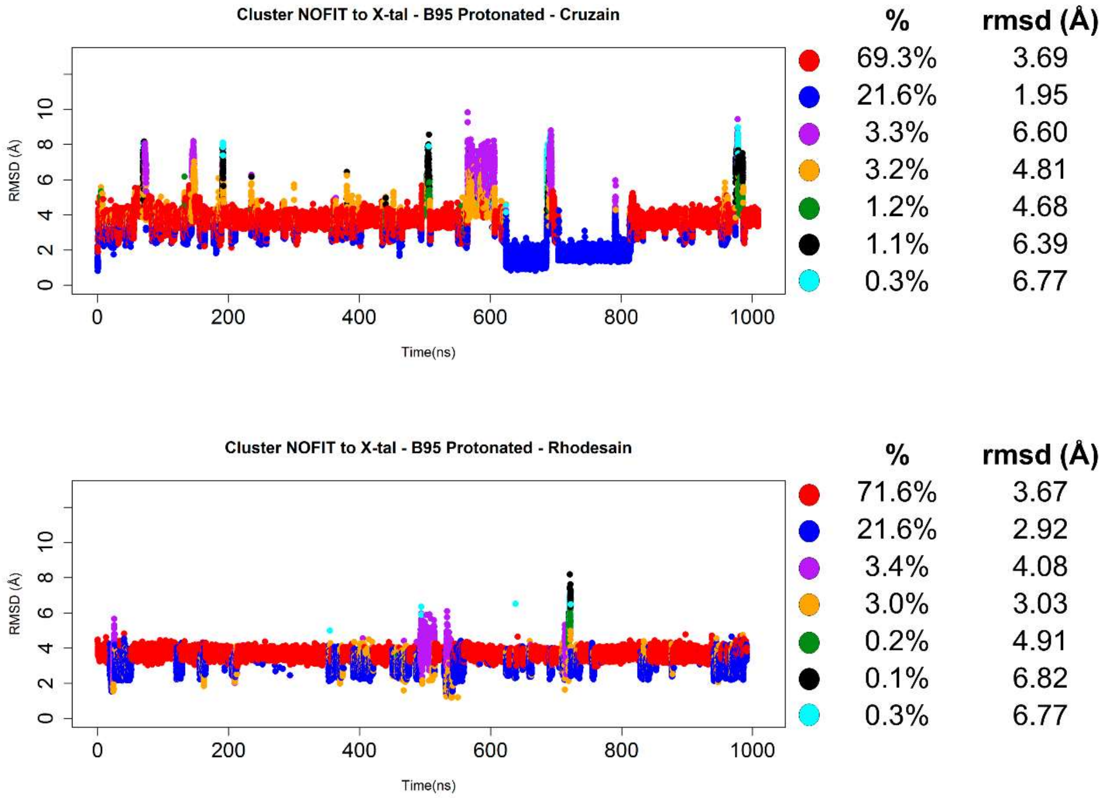


Figure 52: RMSD plot colored according to the cluster for the protonated ligand in both structures, cruzain and rhodesain. The average RMSD between the frames of the simulation and the crystallographic ligand was also calculated.

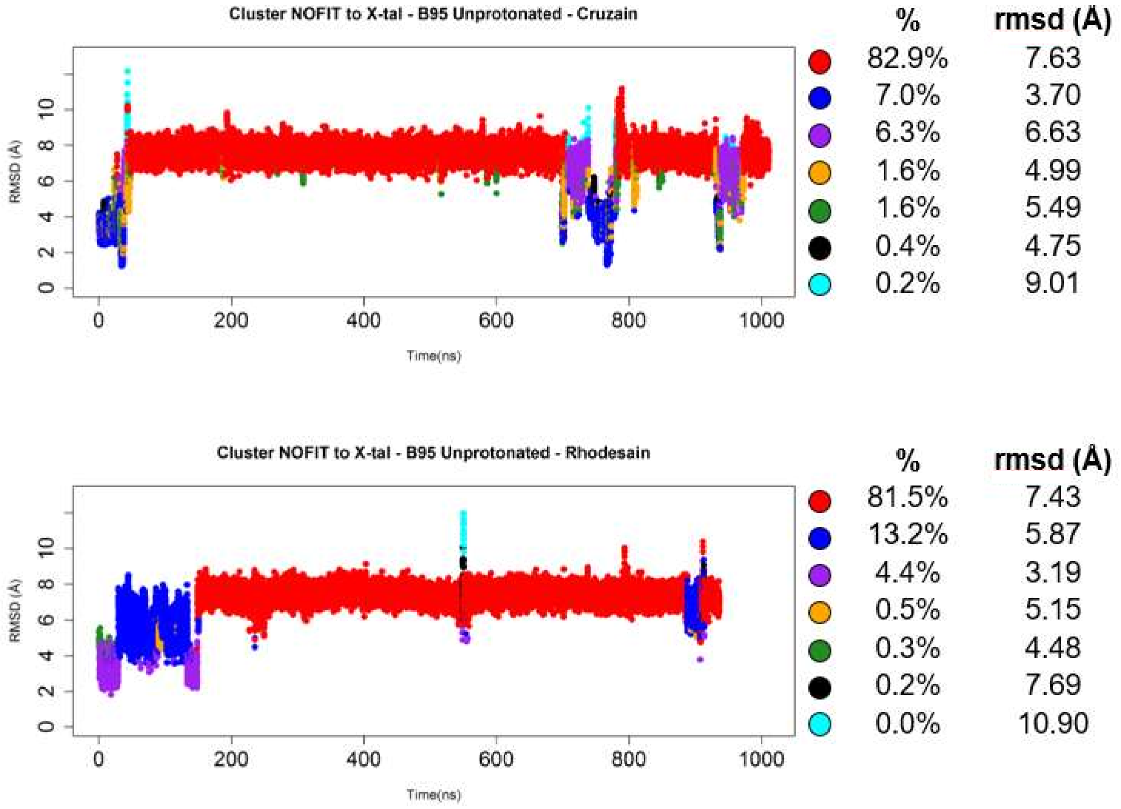


Figure 53: RMSD plot colored according to the cluster for the unprotonated ligand in both structures, cruzain and rhodesain. The average RMSD between the frames and the crystallographic ligand was also calculated.

By analyzing free energy contributions decomposition by MMPBSA.py for all neighboring residues in the active site, we observed similar behavior across the enzymes in the presence of B95 (Figure 54 and Figure 55). When ligand B95 was in its protonated state residues Asp 161 and His 162 were the major contributors to the binding free energy in both systems. As discussed before for cruzain, systems with the B95 ligand protonated had an extremely favorable electrostatic energy. This behavior was also observed for rhodesain (Table 26). When B95 was in its unprotonated state, major energy contributions detected were from residues Leu 67 and Leu 160 (Figure 55). Rhodesain system with B95 unprotonated had better vdWaals contribution than electrostatic one, similar to cruzain bound to the same ligand, although the total free energy of binding was the highest of all systems (Table 26). Once again, no entropy contribution to binding was calculated. Therefore, these results do not equal to the real binding free energy.

Table 26: Averaged binding free energies decomposed in contributions and calculated by MMPBSA.py.

	Average energy component (kcal/mole)						
System	**vdWaals**	**Elec**	**EPB**	**EPolar**	**ΔG$_{gas}$**	**ΔG$_{solv}$**	**ΔG$_{total}$**
Cruzain B95 Protonated	-30.7	-289.2	294.8	-3.1	-319.9	291.7	-28.2
Cruzain B95 Unprotonated	-31.4	-19.2	30.2	-3.1	-50.6	27.2	-23.4
Rhodesain B95 Protonated	-26.0	-317.1	325.5	-2.7	-343.2	322.8	-20.4
Rhodesain B95 Unprotonated	-29.4	-16.1	28.7	-2.9	-45.5	25.8	-19.7

VDWAALS = van der Waals contribution from MM.
EEL = electrostatic energy as calculated by the MM force field.
EPB = the electrostatic contribution to the solvation free energy calculated by PB.
EPOLAR = nonpolar contribution to the solvation free energy calculated by an empirical model.

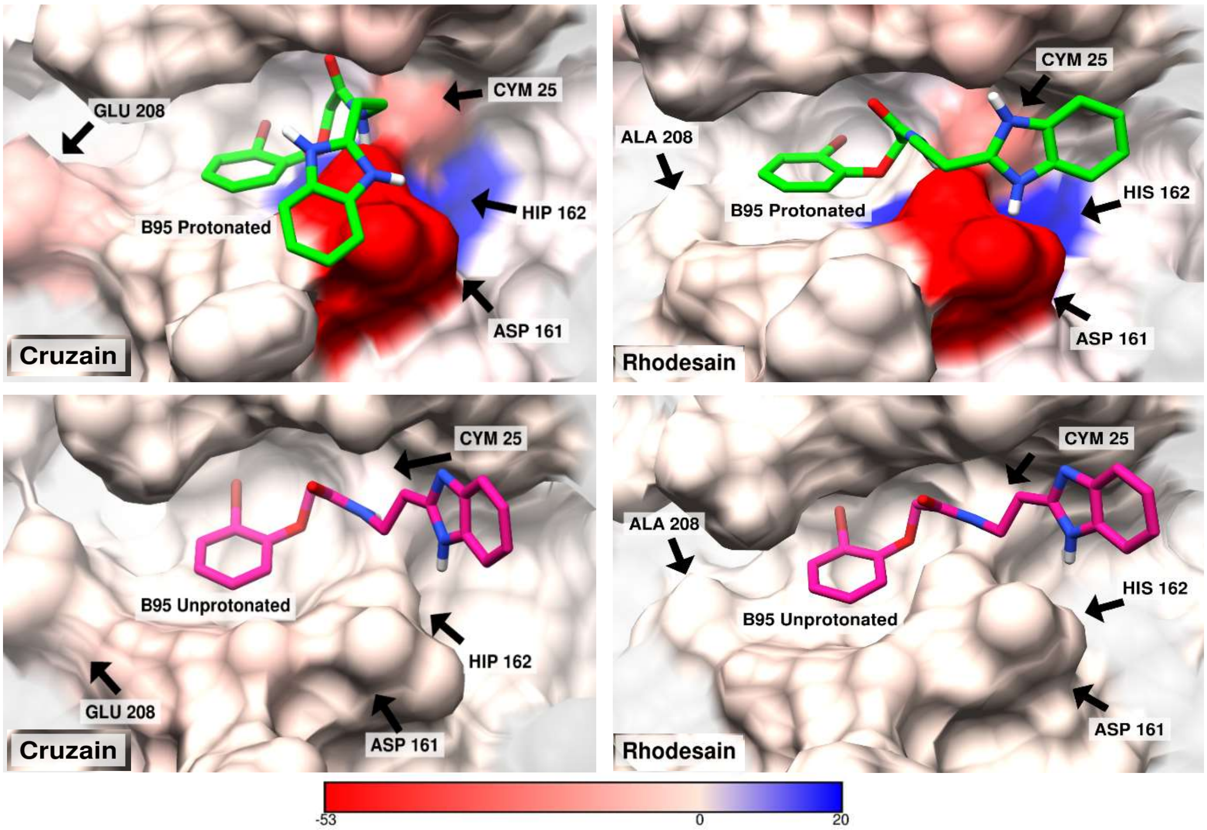


Figure 54: Electrostatic contribution decomposed by residues. Red shows negative contribution and blue shows positive contribution. Protonated ligand is presented in green and the unprotonated one in magenta

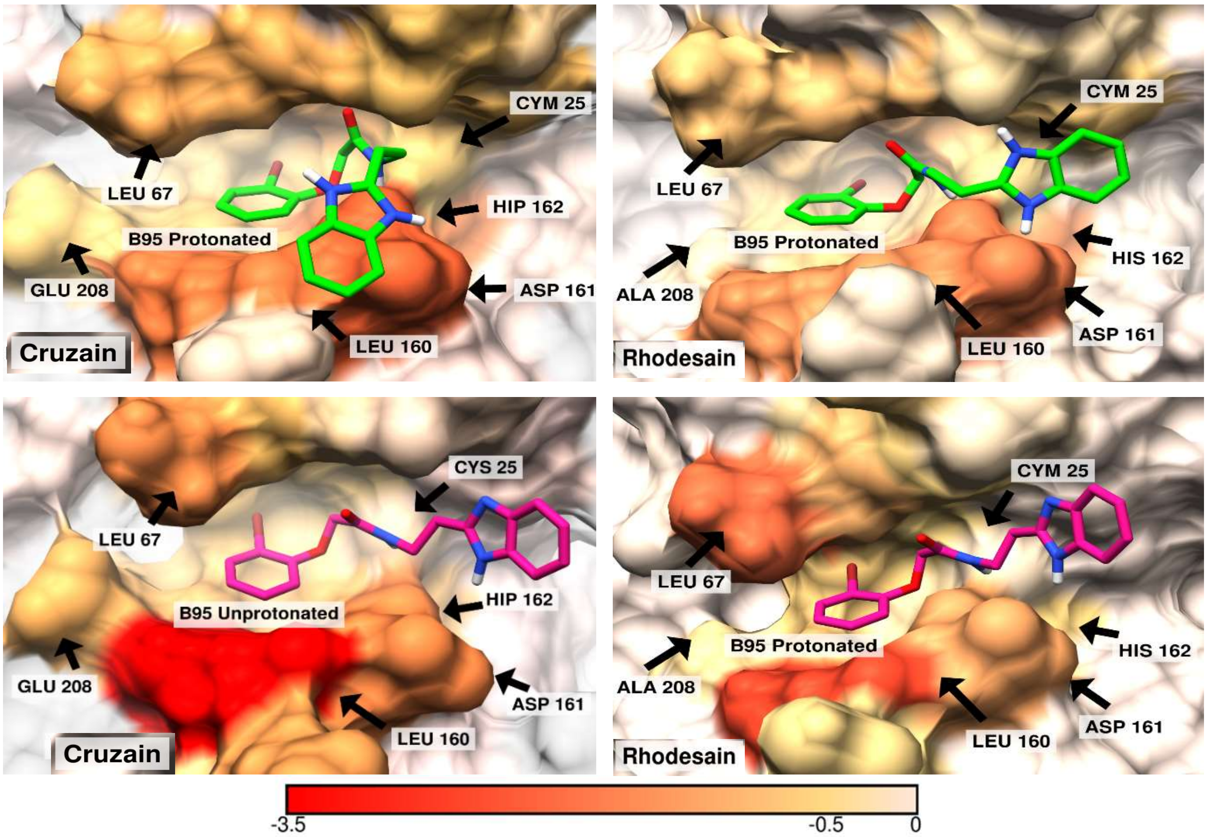


Figure 55: van der Waals contribution decomposed by residues. Red shows negative contribution and white shows no contribution. Protonated ligand is presented in green and the unprotonated one in magenta.

Hydrogen bond analysis showed similar behavior between rhodesain and cruzain systems. Likewise cruzain, when B95 was in its protonated form hydrogen bond interactions with residue Asp 161 (32.5% of the simulation time in cruzain and 55.9% in rhodesain) alternated with the ones with Gly 66 (21.6% of the simulation time in cruzain and 37.5% in rhodesain) throughout the simulation (Figure 56 and Table 27). However, hydrogen bonds interactions with the residue Gly 64 (26.8% of the simulation time), a serine in cruzain, were also present (Figure 56). Nevertheless, when B95 was unprotonated in rhodesain, hydrogen bond analysis showed similar interaction behavior as cruzain (Figure 57 and Table 27).

The influence of the protonation states of B95 could also be observed in the atomic fluctuations of the protein residues in the binding site (Figure 58). Overall, the systems behaved in similar manner. However, small instabilities could be noticed with residues Gly 65 and Gly 66 in the presence of the protonated ligand in rhodesain. This could be due the presence of Gly 64 between them, since in cruzain this residue is a serine.

At the end of all analyzes, it seemed clear that cruzain and rhodesain acted in a similar manner when bound to the same protonated state of the ligand. However, like observed for cruzain the protonated form of B95 provided slightly advantageous outcomes than the unprotonated state. Therefore, we have chosen the protonated ligand to perform TI in both enzymes.

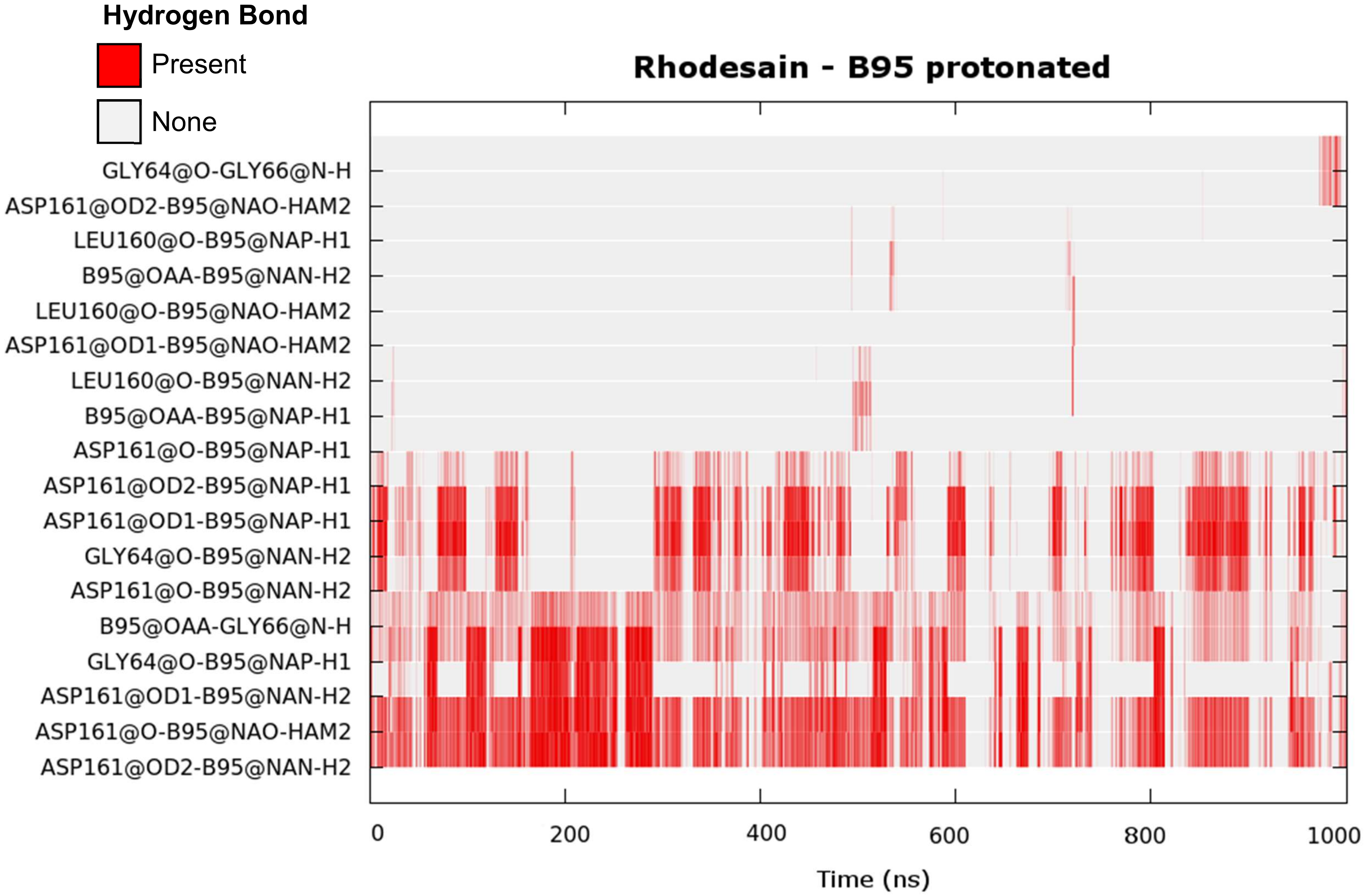


Figure 56: Hydrogen bond analysis of B95 protonated in rhodesain during the 1000 ns simulation.

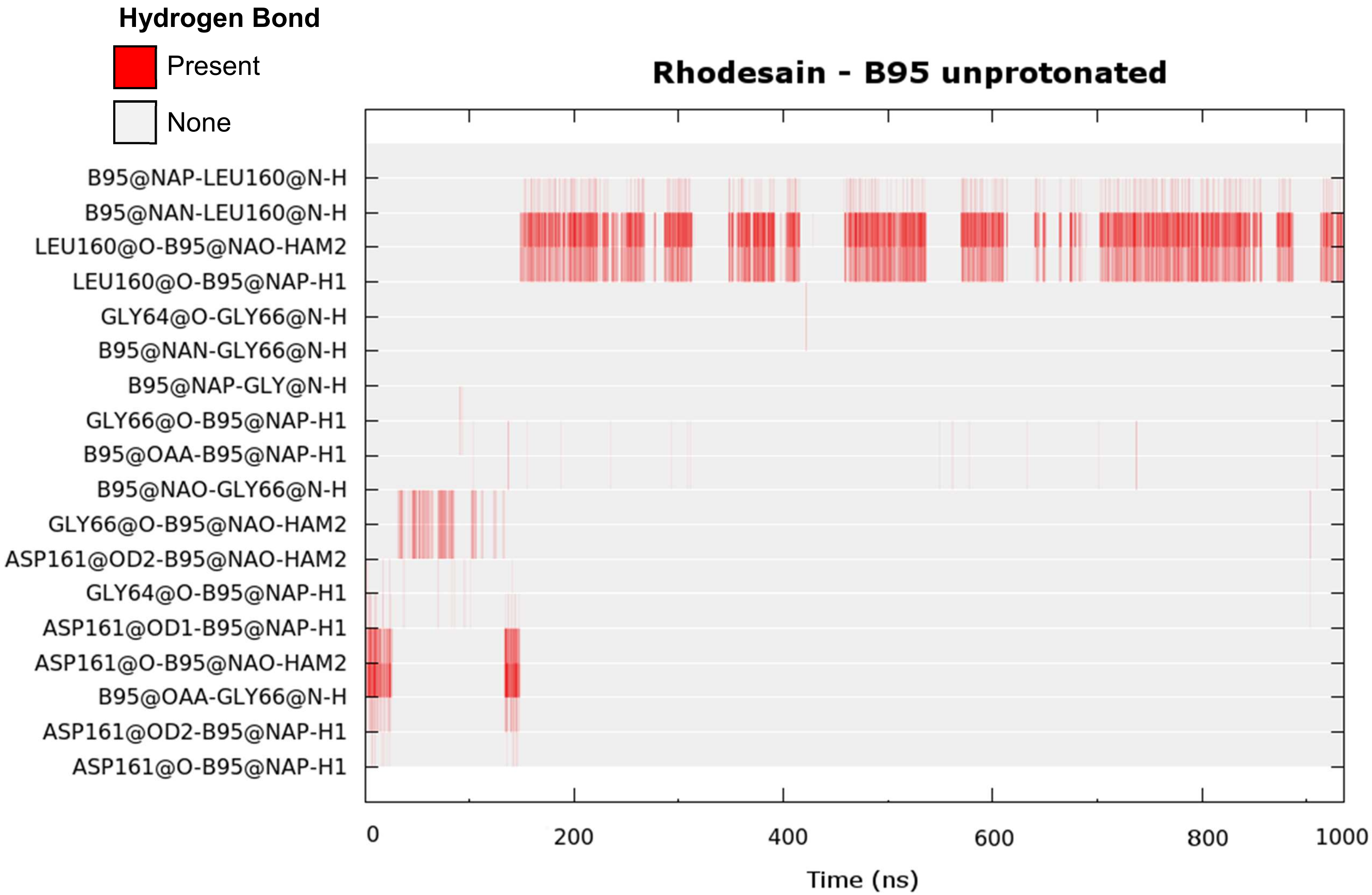


Figure 57: Hydrogen bond analysis of B95 unprotonated in rhodesain during the 1000 ns simulation.

Table 27: Important protein-ligand hydrogen bond interaction lifetime through simulations.

	Fraction of the simulation time (%)			
	Cruzain		Rhodesain	
H-bond	B95 Protonated	B95 Unprotonated	B95 Protonated	B95 Unprotonated
ASP_161@O-B95@NAO-H1	32.5	2.2	55.9	2.5
B95@OAA-GLY_66@N-H	21.6	0.9	37.5	1.3
SER_64@O-B95@NAP-H2	19.3	-	-	-
ASP_161@OD1-B95@NAN-HN	14.5	0.4	13.1	0.4
ASP_161@OD2-B95@NAN-HN	12.8	-	12.4	0.2
SER_64@O-B95@NAN-HN	12.7	0.5	-	-
ASP_161@OD2-B95@NAP-H2	10.0	-	17.1	-
LEU_160@O-B95@NAO-H1	1.6	53.4	0.2	33.8
B95@NAP-LEU_160@N-H	-	14.5	0.3	13.1
GLY_64@O-B95@NAP-H2	-	-	26.8	0.3
GLY_64@O-B95@NAN-H2	-	-	19.1	-

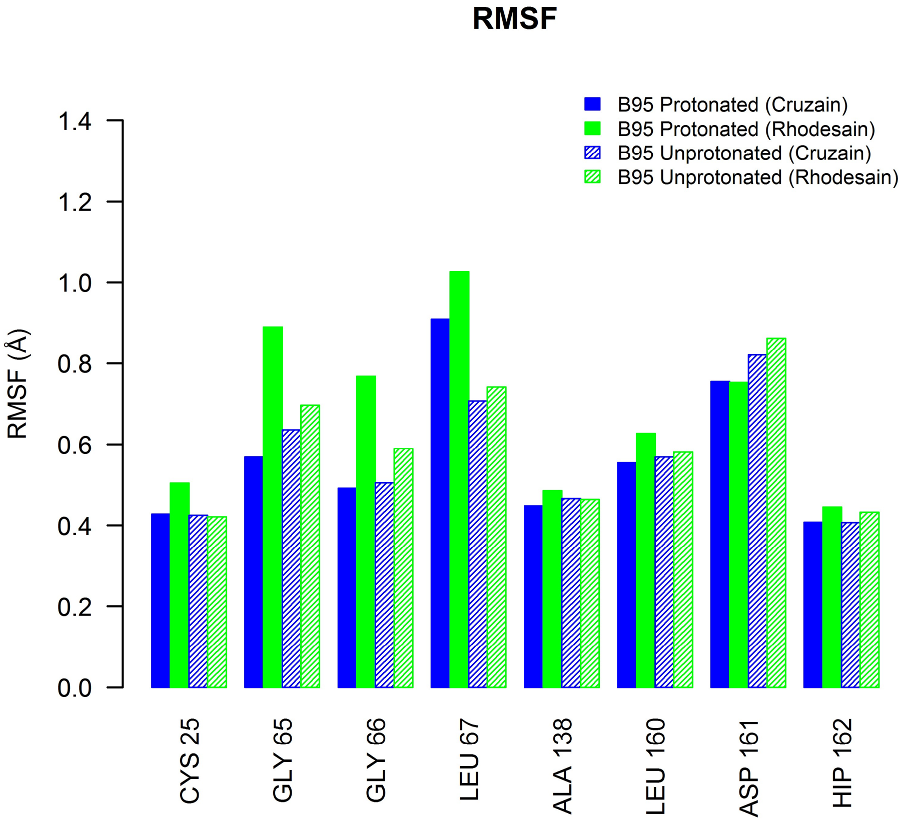


Figure 58: Graphic illustration of the atomic fluctuations of the protein residues in a distance of 6 Å of the center of mass of the ligand. Blue represents the cruzain systems and green the rhodesain systems. Small instabilities can be seen with residues Gly 65 and Gly 66 in the presence of the protonated ligand in rhodesain.

2.4.4 Thermodynamic Integration of B95 analogs

Our purpose here was to validate the experimental structure activity data of a series of B95 analogues, tested against both cruzain and rhodesain. From Table 21, we organized a workflow to calculate the relative energy of binding (Figure 59). In total, seven transformations systems were performed in cruzain and rhodesain. For cruzain, the differences in relative free energy of binding, calculated by TI, were in good agreement with experimental data, calculated from compounds IC_{50} (Table 28). The RMSE observed for cruzain was of ∼1.6 kcal/mol, while for rhodesain, RMSE was a little higher ∼2.2 kcal/mol (Figure 60). It is unclear why rhodesain TI outcomes were not as good as cruzain ones.

Table 28: Relative free energy of binding of compounds calculated by TI and compared to the experimental relative binding free energy.

	Cruzain			Rhodesain		
Transformation	**Calculated ΔΔG**	**Experimental ΔΔG**	**|Δx|[a]**	**Calculated ΔΔG**	**Experimental ΔΔG**	**|Δx|[a]**
MAD558 ↔ MAD574	1.29 ± 0.08	2.75	1.46	1.43 ± 0.14	1.67 ± 0.20	0.24
MAD558 ↔ MAD597	-0.98 ±0.12	-0.83	0.15	-1.45 ±0.08	-1.43 ± 1.53	0.02
MAD558 ↔ MAD619	1.06 ± 0.17	1.10 ± 0.12	0.04	1.02 ± 0.11	-0.72 ± 2.11	1.74
MAD558 ↔ MAD700	1.38 ± 0.13	1.56 ± 0.13	0.18	3.44 ± 0.12	0.29 ± 0.81	3.15
MAD619 ↔ MAD644	-0.60 ±0.13	-0.68	0.08	-1.63 ±0.13	1.46 ± 1.93	3.09
MAD597 ↔ MAD790	4.62 ± 0.13	2.23 ± 0.90	2.39	4.7 ± 0.15	1.36 ± 0.52	3.34
MAD619 ↔ MAD554	-0.06 ±0.11	-0.15	0.09	0.74 ± 0.12	1.24 ± 2.64	0.50

[a] Absolute error between the calculate and experimental values.

Some systems, mostly in rhodesain, presented large absolute error between the calculate and experimental values. For instance, the transformation MAD597 ↔ MAD790 absolute error of over 2.0 kcal/mol in both enzymes, might be due to the transformation group -OH. In this the transformation step of recharging the transformed group in water produced a value three units over the same transformation in the

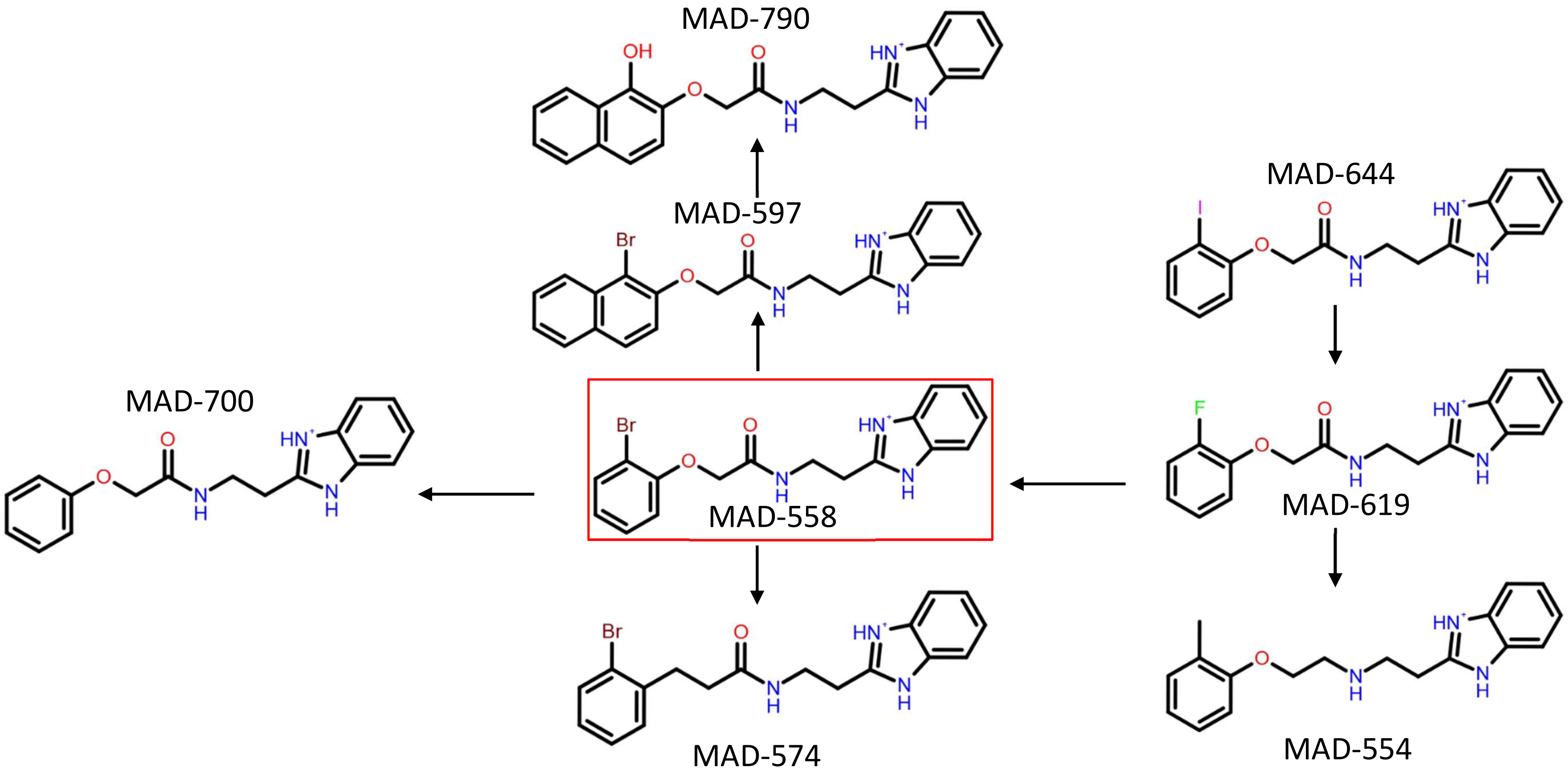


Figure 59: Schematic workflow for relative binding free energy of analogs of B95 (highlighted in red).

Table 29: Outcomes of all TI transformation steps of the compounds in Cruzain

	Ligand in Water				Complex			
Transformation	**$\Delta G_{decharge}$**	**ΔG_{vdw}**	**$\Delta G_{recharge}$**	**Total**	**$\Delta G_{decharge}$**	**ΔG_{vdw}**	**$\Delta G_{recharge}$**	**Total**
MAD558 ↔ MAD574	-13.78	-13.58	4.63	**-22.73**	-12.82	-13.72	5.1	**-21.44**
MAD558 ↔ MAD597	-9.57	5.67	1.81	**-2,09**	-8.76	4.07	1.62	**-3.07**
MAD558 ↔ MAD619	1.12	6.02	-3.03	**4.11**	1.10	6.97	-2.90	**5.17**
MAD558 ↔ MAD700	1.10	-10.79	4.45	**-5.23**	1.08	-8.83	3.89	**-3.85**
MAD619 ↔ MAD644	2.90	-1.61	-1.48	**-0.19**	2.96	-2.10	-1.65	**-0.79**
MAD597 ↔ MAD790	0.84	7.51	-10.88	**-2.52**	0.85	8.16	-6.09	**2.10**
MAD619 ↔ MAD554	2.94	-10.23	0.23	**-7.06**	2.90	-10.15	0.25	**-7.00**

Table 30: Outcomes of all TI transformation steps of the compounds in Rhodesain.

	Ligand in Water				Complex			
Transformation	**$\Delta G_{decharge}$**	**ΔG_{vdw}**	**$\Delta G_{recharge}$**	**Total**	**$\Delta G_{decharge}$**	**ΔG_{vdw}**	**$\Delta G_{recharge}$**	**Total**
MAD558 ↔ MAD574	-12.91	-6.56	4.8	**-14.67**	-11.21	-7.27	5.24	**-13.24**
MAD558 ↔ MAD597	-9.62	14.61	1.85	**6.84**	-8.55	11.42	2.52	**5.39**
MAD558 ↔ MAD619	0.94	8.45	-2.16	**7.23**	1.05	9.66	-2.46	**8.25**
MAD558 ↔ MAD700	0.93	-10.37	4.36	**-5.08**	1.01	-6.51	3.85	**-1.64**
MAD619 ↔ MAD644	3.65	-3.31	-2.8	**-2.46**	3.9	-4.82	-3.17	**-4.09**
MAD597 ↔ MAD790	1.54	2.36	-13.8	**-9.89**	1.83	4.54	-11.55	**-5.19**
MAD619 ↔ MAD554	3.71	-13.08	0.43	**-8.94**	3.8	-12.85	0.85	**-8.20**

complex system. The value implies that the -OH interaction with water exceeded the group interaction with protein residues and caused this discrepancy in the results (Table 29). The same thing might have occurred with this and the MAD558 ↔ MAD700 transformation system in rhodesain (Table 30). However, in rhodesain the MAD597 ↔ MAD790 transformation saw an increase in the vdW part of the transformation for the complex, which might suggest appearance of steric effects between the ligand and protein (Table 30). For those systems where the difference between calculated and experimental was large, we tried to add λ-windows. This provided no improvement to the calculated curves, which implies that the problem with these systems might go beyond sampling problems.

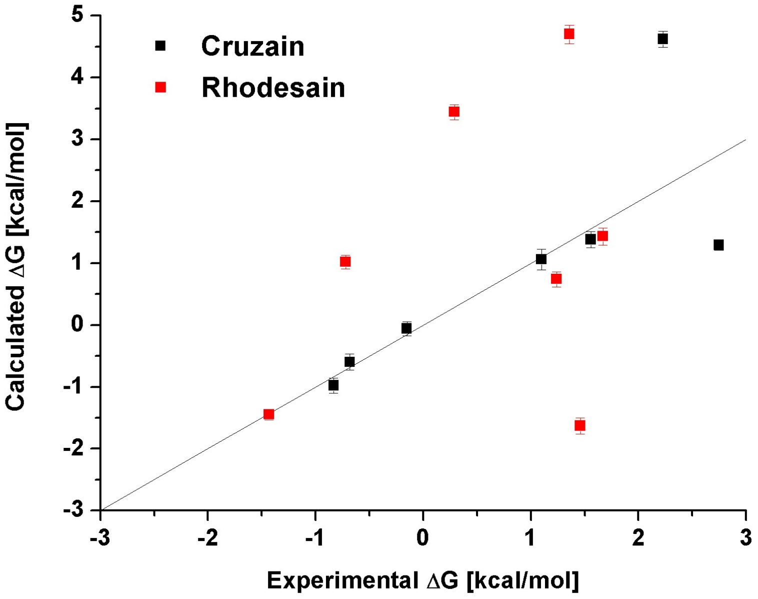


Figure 60: Comparison between the calculated relative binding free energy and experimental biding free energy. Cruzain outcomes are in black and rhodesain ones are in red. The coefficient of determination when both systems are considered is $\boldsymbol{R^2} = \mathbf{0.27}$. When only cruzain is considered is $\boldsymbol{R^2} = \mathbf{0.63}$, and only rhodesain is considered $\boldsymbol{R^2} = \mathbf{0.09}$.

Other problems, such as the chosen numerical integration or the force field used could also be at fault. Or maybe, the reason for the discrepancy might be attributed to the uncertainty of the ligand conformation chosen from docking essays.

2.5 PERSPECTIVES

From the successful prediction of the free energy of formation, particularly of the cruzain complexes, might enable the same methodology to be used to calculate the ΔG of the analogs with unknown activity, allowing prioritization of new analogs to be synthesized. After the synthesis of new molecules by collaborators, these will be assessed against cruzain and rhodesain enzymes and their respective parasites. Finally, for the most potent inhibitors there is the prospect of determining the crystal structure of inhibitor-protein complexes, providing support for the design of new molecules.

Another possibility is the employment of metadynamics simulations to explore the free energy surface to determine the existence of transition states before the ligand achieves the final conformation in the binding site.

2.6 CONCLUSIONS

- We have studied two protonation states of the catalytic dyad, His162-NH$^+$/Cys25-S$^-$ and His162-NH$^+$/Cys25-SH, in a computational context using MD simulations. As far as we know, these are the first MD simulations of the ligand B95 bound to the cruzain crystal structure used, 3KKU.
- The simulations allowed to study the protonation state of the B95 ligand. MD simulations suggested that the presence of the hydrogen atom in both the protein cysteine and protonated benzimidazole ring of the ligand is not favorable and introduces instabilities within the binding site.
- The protonation effect was corroborated by the fact that the ligand left the binding site in the His162-NH$^+$/Cys25-SH early in the simulations. Although outcomes seemed to point out to the His162-NH$^+$/Cys25-S$^-$ protonation bound to the protonated form of the ligand, the neutral ligand in the His162-NH$^+$/Cys25-SH performed well and displayed conformation of the ligand close to the crystallographic ligand B95.
- Probably, the mechanism involves an interchange of the proton between the His162-NH$^+$/Cys25-S$^-$ and the protonated ligand, or, the His162-NH$^+$/Cys25-SH and the neutral ligand. However, this assumption should be confirmed through more precise simulations such as the ones involving quantum methods.
- The ligand B95 in its non-native structure of rhodesain exhibited very similar behavior to the one presented in cruzain. The small changes in the binding site did not offer much interference to B95.
- When confronted to B95 analogues, the binding mode proposed by docking did not produced outcomes as good as cruzain.
- The RMSE observed for the TI calculations of B95 analogues in cruzain was of ~1.6 kcal/mol, while for the same systems in rhodesain, TI calculations provided a higher RMSE of ~2.2 kcal/mol.
- Reasons for the discrepancy in TI calculations might be attributed to the chosen numerical integration method, the chosen force field, and steric effects between the ligand and protein. Although both enzymes share high active site identity, their dynamic behavior is slightly different and the specificity of the B95 analogues needs to be wisely explored.

2.7 REFERENCES

Adcock SA, McCammon JA. Molecular dynamics: survey of methods for simulating the activity of proteins. Chemical reviews. 2006;106(5):1589-615.

Aparicio IM, Scharfstein J, Lima APC. A new cruzipain-mediated pathway of human cell invasion by Trypanosoma cruzi requires trypomastigote membranes. Infection and immunity. 2004;72(10):5892-902.

Baker N, de Koning HP, Mäser P, Horn D. Drug resistance in African trypanosomiasis: the melarsoprol and pentamidine story. Trends in parasitology. 2013;29(3):110-8.

Barrett AJ, Rawlings ND. Evolutionary lines of cysteine peptidases. Biological chemistry. 2001;382(5):727-33.

Barry JD, McCulloch R. Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. Advances in parasitology. 2001;49:1-70.

Berendsen HJ, Postma JPM, van Gunsteren WF, DiNola A, Haak J. Molecular dynamics with coupling to an external bath. The Journal of chemical physics. 1984;81(8):3684-90.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, et al. The protein data bank. Nucleic acids research. 2000;28(1):235-42.

Beveridge DL, DiCapua F. Free energy via molecular simulation: applications to chemical and biomolecular systems. Annual review of biophysics and biophysical chemistry. 1989;18(1):431-92.

Brak K, Kerr ID, Barrett KT, Fuchi N, Debnath M, Ang K, et al. Nonpeptidic tetrafluorophenoxymethyl ketone cruzain inhibitors as promising new leads for Chagas disease chemotherapy. Journal of medicinal chemistry. 2010;53(4):1763-73.

Brandsdal BO, Osterberg F, Almlof M, Feierberg I, Luzhkov VB, Aqvist J. Free energy calculations and ligand binding. Advances in protein chemistry. 2003;66:123-58.

Brener Z. Trypanosoma cruzi: morfologia e ciclo evolutivo. DIAS, CP; COURA, R Clínica e terapêutica da doença de Chagas FIOCRUZ. 1997:23-31.

Brömme D. Papain-like Cysteine Proteases. Current Protocols in Protein Science. 2001:21.2. 1-.2. 14.

Bromme D, Kaleta J. Thiol-dependent cathepsins: pathophysiological implications and recent advances in inhibitor design. Current pharmaceutical design. 2002;8(18):1639-58.

Bryant C, Kerr ID, Debnath M, Ang KK, Ratnam J, Ferreira RS, et al. Novel non-peptidic vinylsulfones targeting the S2 and S3 subsites of parasite cysteine proteases. Bioorganic & medicinal chemistry letters. 2009;19(21):6218-21.

Caffrey CR, Hansell E, Lucas KD, Brinen LS, Hernandez AA, Cheng J, et al. Active site mapping, biochemical properties and subcellular localization of rhodesain, the major cysteine protease of Trypanosoma brucei rhodesiense. Molecular and biochemical parasitology. 2001;118(1):61-73.

Case D, Babin V, Berryman J, Betz R, Cai Q, Cerutti D, et al. Amber 14. 2014.

Cazzulo JJ. Proteinases of Trypanosoma cruzi: potential targets for the chemotherapy of Chagas disease. Medicinal Chemistry Reviews-Online. 2005;2(6):495-504.

Chagas C. Nova entidade morbida do homem: rezumo geral de estudos etiolojicos e clinicos. Memórias do Instituto Oswaldo Cruz. 1911;3(2):219-75.

Choe Y, Brinen LS, Price MS, Engel JC, Lange M, Grisostomi C, et al. Development of α-keto-based inhibitors of cruzain, a cysteine protease implicated in Chagas disease. Bioorganic & medicinal chemistry. 2005;13(6):2141-56.

Contreras VT, Araújo-Jorge TCd, Bonaldo MC, Thomaz N, Barbosa HS, de Meirelles MdNS, et al. Biological aspects of the DM28C clone of Trypanosoma cruzi after metacylogenesis in chemically defined media. Memorias do Instituto Oswaldo Cruz. 1988;83(1):123-33.

Costa TF, dos Reis FC, Lima APC. Substrate inhibition and allosteric regulation by heparan sulfate of Trypanosoma brucei cathepsin L. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics. 2012;1824(3):493-501.

Coura JR, Borges-Pereira J. Chronic phase of Chagas disease: why should it be treated? A comprehensive review. Memorias do Instituto Oswaldo Cruz. 2011;106(6):641-5.

Coura JR, Borges-Pereira J. Chagas disease: What is known and what should be improved: a systemic review. Revista da Sociedade Brasileira de Medicina Tropical. 2012;45(3):286-96.

Croft S, Urbina J, Brun R, Hide G, Mottram J, Coombs G, et al. Chemotherapy of human leishmaniasis and trypanosomiasis. Trypanosomiasis and leishmaniasis: biology and control. 1997:245-57.

Darden T, York D, Pedersen L. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. The Journal of chemical physics. 1993;98(12):10089-92.

Delespaux V, de Koning HP. Drugs and drug resistance in African trypanosomiasis. Drug Resistance Updates. 2007;10(1):30-50.

Docampo R. Recent developments in the chemotherapy of Chagas disease. Current pharmaceutical design. 2001;7(12):1157-64.

Ehmke V, Winkler E, Banner DW, Haap W, Schweizer WB, Rottmann M, et al. Optimization of Triazine Nitriles as Rhodesain Inhibitors: Structure–Activity Relationships, Bioisosteric Imidazopyridine Nitriles, and X-ray Crystal Structure Analysis with Human Cathepsin L. ChemMedChem. 2013;8(6):967-75.

Ettari R, Tamborini L, Angelo IC, Micale N, Pinto A, De Micheli C, et al. Inhibition of rhodesain as a novel therapeutic modality for human African trypanosomiasis. J Med Chem. 2013;56(14):5637-58.

Fampa P, Lisboa C, Jansen A, Santos A, Ramirez M. Protease expression analysis in recently field-isolated strains of Trypanosoma cruzi: a heterogeneous profile of cysteine protease activities between TC I and TC II major phylogenetic groups. Parasitology. 2008;135(09):1093-100.

Ferreira RS, Dessoy MA, Pauli I, Souza ML, Krogh R, Sales AI, et al. Synthesis, Biological Evaluation, and Structure–Activity Relationships of Potent Noncovalent and Nonpeptidic Cruzain Inhibitors as Anti-Trypanosoma cruzi Agents. Journal of medicinal chemistry. 2014;57(6):2380-92.

Ferreira RS, Simeonov A, Jadhav A, Eidam O, Mott BT, Keiser MJ, et al. Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. Journal of medicinal chemistry. 2010;53(13):4891-905.

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of medicinal chemistry. 2004;47(7):1739-49.

Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, et al. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. Journal of medicinal chemistry. 2006;49(21):6177-96.

Gillmor SA, Craik CS, Fletterick RJ. Structural determinants of specificity in the cysteine protease cruzain. Protein science: a publication of the Protein Society. 1997;6(8):1603.

Greenbaum DC, Mackey Z, Hansell E, Doyle P, Gut J, Caffrey CR, et al. Synthesis and structure-activity relationships of parasiticidal thiosemicarbazone cysteine protease inhibitors against Plasmodium falciparum, Trypanosoma brucei, and Trypanosoma cruzi. Journal of medicinal chemistry. 2004;47(12):3212-9.

Grzonka Z, Jankowska E, Kasprzykowski F, Kasprzykowska R, Lankiewicz L, Wiczk W, et al. Structural studies of cysteine proteases and their inhibitors. Acta Biochimica Polonica. 2000;48(1):1-20.

Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. Journal of medicinal chemistry. 2004;47(7):1750-9.

Hansen N, van Gunsteren WF. Practical aspects of free-energy calculations: a review. Journal of Chemical Theory and Computation. 2014;10(7):2632-47.

Hoare CA, Wallace FG. Developmental stages of trypanosomatid flagellates: a new terminology. 1966.

Homeyer N, Gohlke H. Free energy calculations by the molecular mechanics Poisson−Boltzmann surface area method. Molecular Informatics. 2012;31(2):114-22.

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins: Structure, Function, and Bioinformatics. 2006;65(3):712-25.

Huang L, Brinen LS, Ellman JA. Crystal structures of reversible ketone-based inhibitors of the cysteine protease cruzain. Bioorganic & medicinal chemistry. 2003;11(1):21-9.

Jaishankar P, Hansell E, Zhao D-M, Doyle PS, McKerrow JH, Renslo AR. Potency and selectivity of P2/P3-modified inhibitors of cysteine proteases from trypanosomes. Bioorganic & medicinal chemistry letters. 2008;18(2):624-8.

Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. Journal of Computational Chemistry. 2000;21(2):132-46.

Jorgensen WL. Theoretical studies of medium effects on conformational equilibria. The Journal of Physical Chemistry. 1983;87(26):5304-14.

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of chemical physics. 1983;79(2):926-35.

Jorgensen WL, Tirado-Rives J. Monte Carlo vs molecular dynamics for conformational sampling. The Journal of Physical Chemistry. 1996;100(34):14508-13.

Kaus JW, Pierce LT, Walker RC, McCammon JA. Improving the efficiency of free energy calculations in the AMBER molecular dynamics package. Journal of chemical theory and computation. 2013;9(9):4131-9.

Kerr ID, Lee JH, Farady CJ, Marion R, Rickert M, Sajid M, et al. Vinyl sulfones as antiparasitic agents and a structural basis for drug design. Journal of Biological Chemistry. 2009;284(38):25697-703.

Kerr ID, Wu P, Marion-Tsukamaki R, Mackey ZB, Brinen LS. Crystal structures of TbCatB and rhodesain, potential chemotherapeutic targets and major cysteine proteases of Trypanosoma brucei. PLoS Negl Trop Dis. 2010;4(6):e701.

Kollman P. Free energy calculations: applications to chemical and biochemical phenomena. Chemical reviews. 1993;93(7):2395-417.

Langousis G, Hill KL. Motility and more: the flagellum of Trypanosoma brucei. Nature Reviews Microbiology. 2014;12(7):505-18.

Lauria-Pires L, Braga MS, Vexenat AC, Nitz N, Simoes-Barbosa A, Tinoco DL, et al. Progressive chronic Chagas heart disease ten years after treatment with anti-Trypanosoma cruzi nitroderivatives. The American journal of tropical medicine and hygiene. 2000;63(3):111-8.

Lecaille F, Kaleta J, Brömme D. Human and parasitic papain-like cysteine proteases: their role in physiology and pathology and recent developments in inhibitor design. Chemical reviews. 2002;102(12):4459-88.

Ley V, Andrews NW, Robbins ES, Nussenzweig V. Amastigotes of Trypanosoma cruzi sustain an infective cycle in mammalian cells. The Journal of experimental medicine. 1988;168(2):649-59.

Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins: Structure, Function, and Bioinformatics. 2010;78(8):1950-8.

Loncharich RJ, Brooks BR, Pastor RW. Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N′-methylamide. Biopolymers. 1992;32(5):523-35.

Lonsdale-Eccles JD, Grab DJ. Trypanosome hydrolases and the blood–brain barrier. Trends in parasitology. 2002;18(1):17-9.

López-Velázquez G, Hernández R, López-Villaseñor I, Reyes-Vivas H, Segura-Valdez MdL, Jiménez-García LF. Electron microscopy analysis of the nucleolus of Trypanosoma cruzi. Microscopy and Microanalysis. 2005;11(04):293-9.

Löser R, Schilling K, Dimmig E, Gütschow M. Interaction of papain-like cysteine proteases with dipeptide-derived nitriles. Journal of medicinal chemistry. 2005;48(24):7688-707.

Marion R, Hansell E, Caffrey C, Roush W, Brinen LS. The high resolution crystal structure of rohedsain, the major cathepsin L protease from T. brucei rhodesiense, bound to inhibitor K11002.

Martins AV, Gomes AP, Gomes de Mendonça E, Lopes Rangel Fietto J, Santana LA, de Almeida Oliveira MG, et al. Biology of Trypanosoma cruzi: An update. Infectio. 2012;16(1):45-58.

McCammon JA, Karplus M. The dynamic picture of protein structure. Accounts of Chemical Research. 1983;16(6):187-93.

McGrath ME, Eakin AE, Engel JC, McKerrow JH, Craik CS, Fletterick RJ. The crystal structure of cruzain: a therapeutic target for Chagas' disease. Journal of molecular biology. 1995;247(2):251-9.

Michel J, Foloppe N, Essex JW. Rigorous Free Energy Calculations in Structure-Based Drug Design. Molecular Informatics. 2010;29(8-9):570-8.

Miller III BR, McGee Jr TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA. py: an efficient program for end-state free energy calculations. Journal of Chemical Theory and Computation. 2012;8(9):3314-21.

MOE. Molecular Operating Environment (MOE), 2013.08. 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7: Chemical Computing Group Inc.; 2016.

Molina J, Martins-Filho O, Brener Z, Romanha AJ, Loebenberg D, Urbina JA. Activities of the triazole derivative SCH 56592 (posaconazole) against drug-resistant strains of the protozoan parasiteTrypanosoma (Schizotrypanum) cruzi in immunocompetent and immunosuppressed murine hosts. Antimicrobial agents and chemotherapy. 2000;44(1):150-5.

Mott BT, Ferreira RS, Simeonov A, Jadhav A, Ang KK-H, Leister W, et al. Identification and optimization of inhibitors of trypanosomal cysteine proteases: cruzain, rhodesain, and TbCatB. Journal of medicinal chemistry. 2009;53(1):52-60.

Nägler DK, Ménard R. Family C1 cysteine proteases: biological diversity or redundancy? Biological chemistry. 2003;384(6):837-43.

Ooi C-P, Bastin P. More than meets the eye: understanding Trypanosoma brucei morphology in the tsetse. Frontiers in cellular and infection microbiology. 2013;3.

Otto H-H, Schirmeister T. Cysteine proteases and their inhibitors. Chemical reviews. 1997;97(1):133-72.

Pastor RW, Brooks BR, Szabo A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. Molecular Physics. 1988;65(6):1409-19.

Pearlman DA. Free energy calculations: methods for estimating ligand binding affinities. Free energy calculations in rational drug design Kluwer Academic/Plenum Publishers, New York. 2001:9-35.

Pearlman DA, Rao BG. Free energy calculations: methods and applications. Encyclopedia of computational chemistry. 1998.

Polgar L. Catalytic mechanisms of cysteine peptidases. Handbook of proteolytic enzymes. 2004:1072-9.

Puente XS, Sánchez LM, Overall CM, López-Otín C. Human and mouse proteases: a comparative genomic approach. Nature Reviews Genetics. 2003;4(7):544-58.

Que X, Reed SL. Cysteine proteinases and the pathogenesis of amebiasis. Clinical Microbiology Reviews. 2000;13(2):196-206.

Rapaport DC. The art of molecular dynamics simulation: Cambridge university press; 2004.

Rawlings ND, Barrett AJ. Families of cysteine peptidases. Methods in enzymology. 1993;244:461-86.

Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. Nucleic acids research. 2010;38(suppl 1):D227-D33.

Rodenko B, de Koning HP. Rational Selection of Anti-Microbial Drug Targets: Unique or Conserved? Trypanosomatid Diseases: Molecular Routes to Drug Discovery. 2013:279-96.

Rodrigues CR, Flaherty TM, Springer C, McKerrow JH, Cohen FE. CoMFA and HQSAR of acylhydrazide cruzain inhibitors. Bioorganic & medicinal chemistry letters. 2002;12(11):1537-41.

Ryckaert J-P, Ciccotti G, Berendsen HJ. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics. 1977;23(3):327-41.

Sajid M, McKerrow JH. Cysteine proteases of parasitic organisms. Molecular and biochemical parasitology. 2002;120(1):1-21.

Sajid M, Robertson SA, Brinen LS, McKerrow JH. Cruzain : the path from target validation to the clinic. Advances in experimental medicine and biology. 2011;712:100-15.

Sárkány Z, Szeltner Z, Polgár L. Thiolate-imidazolium ion pair is not an obligatory catalytic entity of cysteine peptidases: the active site of picornain 3C. Biochemistry. 2001;40(35):10601-6.

Scharfstein J, Schmitz V, Morandi V, Capella MM, Lima APC, Morrot A, et al. Host cell invasion by Trypanosoma cruzi is potentiated by activation of bradykinin B2 receptors. The Journal of experimental medicine. 2000;192(9):1289-300.

Schechter I, Berger A. On the size of the active site in proteases. I. Papain Biochem Biophys Res Commun. 1967;27:157-62.

Schilling O, Overall CM. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. Nature biotechnology. 2008;26(6):685-94.

Serveau C, Lalmanach G, Juliano M, Scharfstein J, Juliano L, Gauthier F. Investigation of the substrate specificity of cruzipain, the major cysteine proteinase of Trypanosoma cruzi, through the use of cystatin-derived substrates and inhibitors. Biochem J. 1996;313:951-6.

Shirts MR, Pande VS. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. The Journal of chemical physics. 2005;122(14):144107.

Shokhen M, Khazanov N, Albeck A. Challenging a paradigm: Theoretical calculations of the protonation state of the Cys25-His159 catalytic diad in free papain. Proteins: Structure, Function, and Bioinformatics. 2009;77(4):916-26.

Siklos M, BenAissa M, Thatcher GR. Cysteine proteases as therapeutic targets: does selectivity matter? A systematic review of calpain and cathepsin inhibitors. Acta Pharmaceutica Sinica B. 2015.

Smith HJ, Simons C. Enzymes and their inhibitors: drug development: CRC press; 2004.

Steinbrecher T, Joung I, Case DA. Soft-core potentials in thermodynamic integration: Comparing one-and two-step transformations. Journal of computational chemistry. 2011;32(15):3253-63.

Steverding D. Proteases of Trypanosoma brucei. Trypanosomatid Diseases: Molecular Routes to Drug Discovery. 2013:365-82.

Steverding D, Caffrey CR, Sajid M. Cysteine proteinase inhibitors as therapy for parasitic diseases: advances in inhibitor design. Mini reviews in medicinal chemistry. 2006;6(9):1025-32.

Steverding D, Sexton DW, Wang X, Gehrke SS, Wagner GK, Caffrey CR. Trypanosoma brucei: chemical evidence that cathepsin L is essential for survival and a relevant drug target. International journal for parasitology. 2012;42(5):481-8.

Storer AC, Ménard R. Catalytic mechanism in papain family of cysteine peptidases. Methods in enzymology. 1994;244:486.

Straatsma T, Berendsen H, Stam A. Estimation of statistical errors in molecular simulation calculations. Molecular Physics. 1986;57(1):89-95.

Turk B, Turk D, Turk V. Lysosomal cysteine proteases: more than scavengers. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology. 2000;1477(1):98-111.

Turk D, Gunčar G, Podobnik M, Turk B. Revised definition of substrate binding sites of papain-like cysteine proteases. Biological chemistry. 1998;379(2):137-48.

Turk V, Stoka V, Vasiljeva O, Renko M, Sun T, Turk B, et al. Cysteine cathepsins: from structure, function and regulation to new frontiers. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics. 2012;1824(1):68-88.

Urbina JA, Lira R, Visbal G, Bartrolí J. In vitro antiproliferative effects and mechanism of action of the new triazole derivative UR-9825 against the protozoan parasite Trypanosoma (Schizotrypanum) cruzi. Antimicrobial agents and chemotherapy. 2000;44(9):2498-502.

Urbina JA, Payares G, Molina J, Sanoja C, Liendo A, Lazardi K, et al. Cure of short- and long-term experimental Chagas' disease using D0870. Science. 1996;273(5277):969-71.

Vickerman K. Developmental cycles and biology of pathogenic trypanosomes. British medical bulletin. 1985;41(2):105-14.

Wang J, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. Journal of molecular graphics and modelling. 2006;25(2):247-60.

Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. Journal of computational chemistry. 2004;25(9):1157-74.

WHO WHO. Sustaining the drive to overcome the global impact of neglected tropical diseases: second WHO report on neglected diseases: World Health Organization; 2013.

WHO WHO. Fact Sheet 259: WHO; 2015 [updated May 2015]. Available from: http://www.who.int/mediacentre/factsheets/fs259/.

WHO WHO. Human African trypanosomiasis: update of the methodological framework for clinical trials: report of the first meeting of the Development of New Tools subgroup, Geneva, 24 September 2014 2015.

Wilkinson SR, Taylor MC, Horn D, Kelly JM, Cheeseman I. A mechanism for cross-resistance to nifurtimox and benznidazole in trypanosomes. Proceedings of the National Academy of Sciences. 2008;105(13):5022-7.

Yang PY, Wang M, Li L, Wu H, He CY, Yao SQ. Design, Synthesis and Biological Evaluation of Potent Azadipeptide Nitrile Inhibitors and Activity-Based Probes as Promising Anti-Trypanosoma brucei Agents. Chemistry-A European Journal. 2012;18(21):6528-41.

Zwanzig RW. High-temperature equation of state by a perturbation method. I. nonpolar gases. The Journal of Chemical Physics. 1954;22(8):1420-6.

3 APPENDIX A – DOCK 6.6 INPUTS

```
###Dock 6.6 input for Grid Score scoring
ligand_atom_file                    ligand.mol2
limit_max_ligands                   no
skip_molecule                       no
read_mol_solvation                  no
calculate_rmsd                      yes
use_rmsd_reference_mol              yes
rmsd_reference_filename             ligand-reference.mol2
use_database_filter                 no
orient_ligand                       yes
automated_matching                  yes
receptor_site_file                  selected_spheres.sph
max_orientations                    500
critical_points                     no
chemical_matching                   no
use_ligand_spheres                  no
use_internal_energy                 yes
internal_energy_rep_exp             12
flexible_ligand                     yes
user_specified_anchor               no
limit_max_anchors                   no
min_anchor_size                     40
pruning_use_clustering              yes
pruning_max_orients                 100
pruning_clustering_cutoff           100
pruning_conformer_score_cutoff      25.0
use_clash_overlap                   no
write_growth_tree                   no
bump_filter                         no
score_molecules                     yes
contact_score_primary               no
contact_score_secondary             no
grid_score_primary                  yes
grid_score_secondary                no
grid_score_rep_rad_scale            1
grid_score_vdw_scale                1
grid_score_es_scale                 1
grid_score_grid_prefix              grid
multigrid_score_secondary           no
dock3.5_score_secondary             no
continuous_score_secondary          no
descriptor_score_secondary          no
gbsa_zou_score_secondary            no
gbsa_hawkins_score_secondary        no
SASA_descriptor_score_secondary     no
```

```
amber_score_secondary               no
minimize_ligand                     yes
minimize_anchor                     yes
minimize_flexible_growth            yes
use_advanced_simplex_parameters     no
simplex_max_cycles                  1
simplex_score_converge              0.1
simplex_cycle_converge              1.0
simplex_trans_step                  1.0
simplex_rot_step                    0.1
simplex_tors_step                   10.0
simplex_anchor_max_iterations       500
simplex_grow_max_iterations         500
simplex_grow_tors_premin_iterations 0
simplex_random_seed                 0
simplex_restraint_min               no
atom_model                          all
vdw_defn_file                       ~/dock6/parameters/vdw_AMBER_parm99.defn
flex_defn_file                      ~/dock6/parameters/flex.defn
flex_drive_file                     ~/dock6/parameters/flex_drive.tbl
ligand_outfile_prefix               flex
write_orientations                  no
num_scored_conformers               1
rank_ligands                        no
```

```
###Dock 6.6 input for Amber Score scoring
ligand_atom_file                       scored.amber_score.mol2
limit_max_ligands                      no
skip_molecule                          no
read_mol_solvation                     no
calculate_rmsd                         no
use_database_filter                    no
orient_ligand                          no
use_internal_energy                    no
flexible_ligand                        no
bump_filter                            no
score_molecules                        yes
contact_score_primary                  no
contact_score_secondary                no
grid_score_primary                     no
grid_score_secondary                   no
multigrid_score_primary                no
multigrid_score_secondary              no
dock3.5_score_primary                  no
dock3.5_score_secondary                no
continuous_score_primary               no
```

```
continuous_score_secondary            no
descriptor_score_primary              no
descriptor_score_secondary            no
gbsa_zou_score_primary                no
gbsa_zou_score_secondary              no
gbsa_hawkins_score_primary            no
gbsa_hawkins_score_secondary          no
SASA_descriptor_score_primary         no
SASA_descriptor_score_secondary       no
amber_score_primary                   yes
amber_score_secondary                 no
amber_score_receptor_file_prefix      protein_semH
amber_score_movable_region            nothing
amber_score_gb_model                  5
amber_score_nonbonded_cutoff          18.0
amber_score_temperature               300.0
amber_score_abort_on_unprepped_ligand yes
ligand_outfile_prefix                 output
write_orientations                    no
num_scored_conformers                 1
rank_ligands                          no
```

4 APPENDIX B – AMBER MD INPUTS

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=1000.0,restraintmask="!@H="
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=1000.0,restraintmask="!@H=&:1-216",
/
```

equilibration

```
&cntrl
        ntb=1,ntc=2,ntf=2,ntt=3,gamma_ln=2.0,cut=8.0,
        ntr=1,restraint_wt=1000.0,restraintmask="!@H=&:1-216",
        nstlim=100000,dt=0.001,nmropt=1,
/
&wt TYPE='TEMP0', istep1=0, istep2=100000, value1=100.0,value2=300.0 /
&wt TYPE='END' /
```

pressure equilibration

```
&cntrl
        ntb=2,ntp=1,pres0=1.0,tautp=2.0,
        ntc=2,ntf=2,ntt=3,gamma_ln=2.0,
        tempi=300.0,temp0=300.0,
        ntr=1,restraint_wt=1000.0,restraintmask="!@H=&:1-216",
        nstlim=100000,dt=0.002,
/
```

equilibration

```
&cntrl
        ntb=1,ntc=2,ntf=2,ntt=3,gamma_ln=2.0,cut=8.0,
        ntr=1,restraint_wt=1000.0,restraintmask="!@H=&:1-216",
        nstlim=100000,dt=0.001,nmropt=1,
/
&wt TYPE='TEMP0', istep1=0, istep2=100000, value1=300.0,value2=100.0 /
&wt TYPE='END' /
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=1000.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=500.0,restraintmask="!@H=&:1-216",/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=200.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=100.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=50.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=20.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=10.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=5.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=4.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=3.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=2.0,restraintmask="!@H=&:1-216",/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=1.0,restraintmask="!@H=&:1-216",
/
```

minimization

```
&cntrl
        imin=1,ncyc=500,maxcyc=1000,ntr=1,cut=8.0,
        restraint_wt=0.5,restraintmask="!@H=&:1-216",
/
```

equilibration

```
&cntrl
        ntb=1,ntc=2,ntf=2,ntt=3,gamma_ln=2.0,cut=8.0,
        nstlim=200000,dt=0.002,nmropt=1,
/
&wt TYPE='TEMP0', istep1=0, istep2=200000, value1=100.0,value2=300.0 /
&wt TYPE='END' /
```

pressure equilibration

```
&cntrl
        ntb=2,ntp=1,pres0=1.0,tautp=2.0,
        ntc=2,ntf=2,ntt=3,gamma_ln=2.0,
        tempi=300.0,temp0=300.0,
        nstlim=50000,dt=0.002,
/
```

10ns npt simulation for production

```
&cntrl
        ntb=2
         ntp=1,pres0=1.0,taup=2.0
        iwrap=1
        ioutfm=1
         ntx=5, irest=1
        cut=8.0
        ntc=2,ntf=2
         ntt=3,gamma_ln=2.0
         tempi=300.0,temp0=300.0
        nstlim=5000000,dt=0.002
        ntpr=10000,ntwx=10000
/
```

5 APPENDIX C – AMBER TI INPUTS

Decharge

```
minimization
 &cntrl
        imin = 1, ntmin = 2,
        maxcyc = 1000,
        ntpr = 200, ntwe = 200,
        ntb = 1,
        ntr = 1, restraint_wt = 5.00,
        restraintmask='!:WAT & !@H=',

        icfe = 1, ifsc = 1, clambda = 0.5, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc = 0, crgmask = ':2@F',
 /
 &ewald
 /
```

```
equilibration
&cntrl
        ntb=1,ntc=2,ntf=1,ntt=3,gamma_ln=2.0,cut=8.0,
        nstlim=125000,dt=0.002,nmropt=1,
        ntr = 1, restraint_wt = 5.00,restraintmask='!:WAT & !@H=',
        ioutfm = 1, iwrap = 1,
        ntwe = 25000, ntwx = 25000, ntpr = 25000, ntwr = 62500,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc = 0, crgmask = ':2@F',
/
&wt TYPE='TEMP0', istep1=0, istep2=120000, value1=100.0,value2=300.0 /
&wt TYPE='END' /
```

```
pressure equilibration
&cntrl
        ntb=2,ntp=1,pres0=1.0,tautp=2.0,
        ntc=2,ntf=1,ntt=3,gamma_ln=2.0,
        tempi=300.0,temp0=300.0,
        nstlim=125000,dt=0.002,
        ntr = 1, restraint_wt = 0.50,restraintmask='!:WAT & !@H=',
        ioutfm = 1, iwrap = 1,
        ntwe = 25000, ntwx = 25000, ntpr = 25000, ntwr = 62500,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc = 0, crgmask = ':2@F',
/
```

TI simulation

```
 &cntrl
        imin = 0, nstlim = 250000, irest = 1, ntx = 5, dt = 0.002,
        ntt = 3, temp0 = 300.0, gamma_ln = 2.0, ig = -1,
        ntc = 2, ntf = 1,
        ntb = 2,
        ntp = 1, pres0 = 1.0, taup = 2.0,
        Ioutfm = 1, iwrap = 1,
        ntwe = 2500, ntwx = 10000, ntpr = 10000, ntwr = 10000,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 1,
        ifmbar = 1, bar_intervall = 1000, bar_l_min = 0.0, bar_l_max = 1.0,
        bar_l_incr = 0.1,
        timask1 = ':1', timask2 = ':2',
        ifsc = 0, crgmask = ':2@F',
 /

 &ewald
 /
```

VDW Bonded

minimization

```
 &cntrl
        imin = 1, ntmin = 2,
        maxcyc = 1000,
        ntpr = 200, ntwe = 200,
        ntb = 1,
        ntr = 1, restraint_wt = 5.00,
        restraintmask='!:WAT & !@H=',

        icfe = 1, ifsc = 1, clambda = 0.5, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc=1, scmask1=':1@F', scmask2=':2@C01,H01,H02,H03',
        crgmask=':1@F | :2@C01,H01,H02,H03'
 /
 &ewald
 /
```

equilibration

```
&cntrl
        ntb=1,ntc=2,ntf=1,ntt=3,gamma_ln=2.0,cut=8.0,
        nstlim=125000,dt=0.002,nmropt=1,
        ntr = 1, restraint_wt = 5.00,restraintmask='!:WAT & !@H=',
        ioutfm = 1, iwrap = 1,
        ntwe = 25000, ntwx = 25000, ntpr = 25000, ntwr = 62500,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc=1, scmask1=':1@F', scmask2=':2@C01,H01,H02,H03',
        crgmask=':1@F | :2@C01,H01,H02,H03'
/
&wt TYPE='TEMP0', istep1=0, istep2=120000, value1=100.0,value2=300.0 /
&wt TYPE='END' /

pressure equilibration
&cntrl
        ntb=2,ntp=1,pres0=1.0,tautp=2.0,
        ntc=2,ntf=1,ntt=3,gamma_ln=2.0,
        tempi=300.0,temp0=300.0,
        nstlim=125000,dt=0.002,
        ntr = 1, restraint_wt = 0.50,restraintmask='!:WAT & !@H=',
        ioutfm = 1, iwrap = 1,
        ntwe = 25000, ntwx = 25000, ntpr = 25000, ntwr = 62500,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc=1, scmask1=':1@F', scmask2=':2@C01,H01,H02,H03',
        crgmask=':1@F | :2@C01,H01,H02,H03'
/

TI simulation
 &cntrl
        imin = 0, nstlim = 500000, irest = 1, ntx = 5, dt = 0.002,
        ntt = 3, temp0 = 300.0, gamma_ln = 2.0, ig = -1,
        ntc = 2, ntf = 1,
        ntb = 2,
        ntp = 1, pres0 = 1.0, taup = 2.0,
        ioutfm = 1, iwrap = 1,
        ntwe = 2500, ntwx = 10000, ntpr = 10000, ntwr = 10000,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 1,
        ifmbar = 1, bar_intervall = 1000, bar_l_min = 0.0, bar_l_max = 1.0,
        bar_l_incr = 0.1,
         timask1 = ':1', timask2 = ':2',
        ifsc=1, scmask1=':1@F', scmask2=':2@C01,H01,H02,H03',
        crgmask=':1@F | :2@C01,H01,H02,H03'
 /

 &ewald
 /
```

Recharge

Minimization

```
 &cntrl
        imin = 1, ntmin = 2,
        maxcyc = 1000,
        ntpr = 200, ntwe = 200,
        ntb = 1,
        ntr = 1, restraint_wt = 5.00,
        restraintmask='!:WAT & !@H=',

        icfe = 1, ifsc = 1, clambda = 0.5, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc = 0, crgmask = ':1@C01,H01,H02,H03',
 /
 &ewald
 /
```

equilibration

```
&cntrl
        ntb=1,ntc=2,ntf=1,ntt=3,gamma_ln=2.0,cut=8.0,
        nstlim=125000,dt=0.002,nmropt=1,
        ntr = 1, restraint_wt = 5.00,restraintmask='!:WAT & !@H=',
        ioutfm = 1, iwrap = 1,
        ntwe = 25000, ntwx = 25000, ntpr = 25000, ntwr = 62500,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc = 0, crgmask = ':1@C01,H01,H02,H03',
/
&wt TYPE='TEMP0', istep1=0, istep2=120000, value1=100.0,value2=300.0 /
&wt TYPE='END' /
```

pressure equilibration

```
&cntrl
        ntb=2,ntp=1,pres0=1.0,tautp=2.0,
        ntc=2,ntf=1,ntt=3,gamma_ln=2.0,
        tempi=300.0,temp0=300.0,
        nstlim=125000,dt=0.002,
        ntr = 1, restraint_wt = 0.50,restraintmask='!:WAT & !@H=',
        ioutfm = 1, iwrap = 1,
        ntwe = 25000, ntwx = 25000, ntpr = 25000, ntwr = 62500,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 0,
        timask1 = ':1', timask2 = ':2',
        ifsc = 0, crgmask = ':1@C01,H01,H02,H03',
/
```

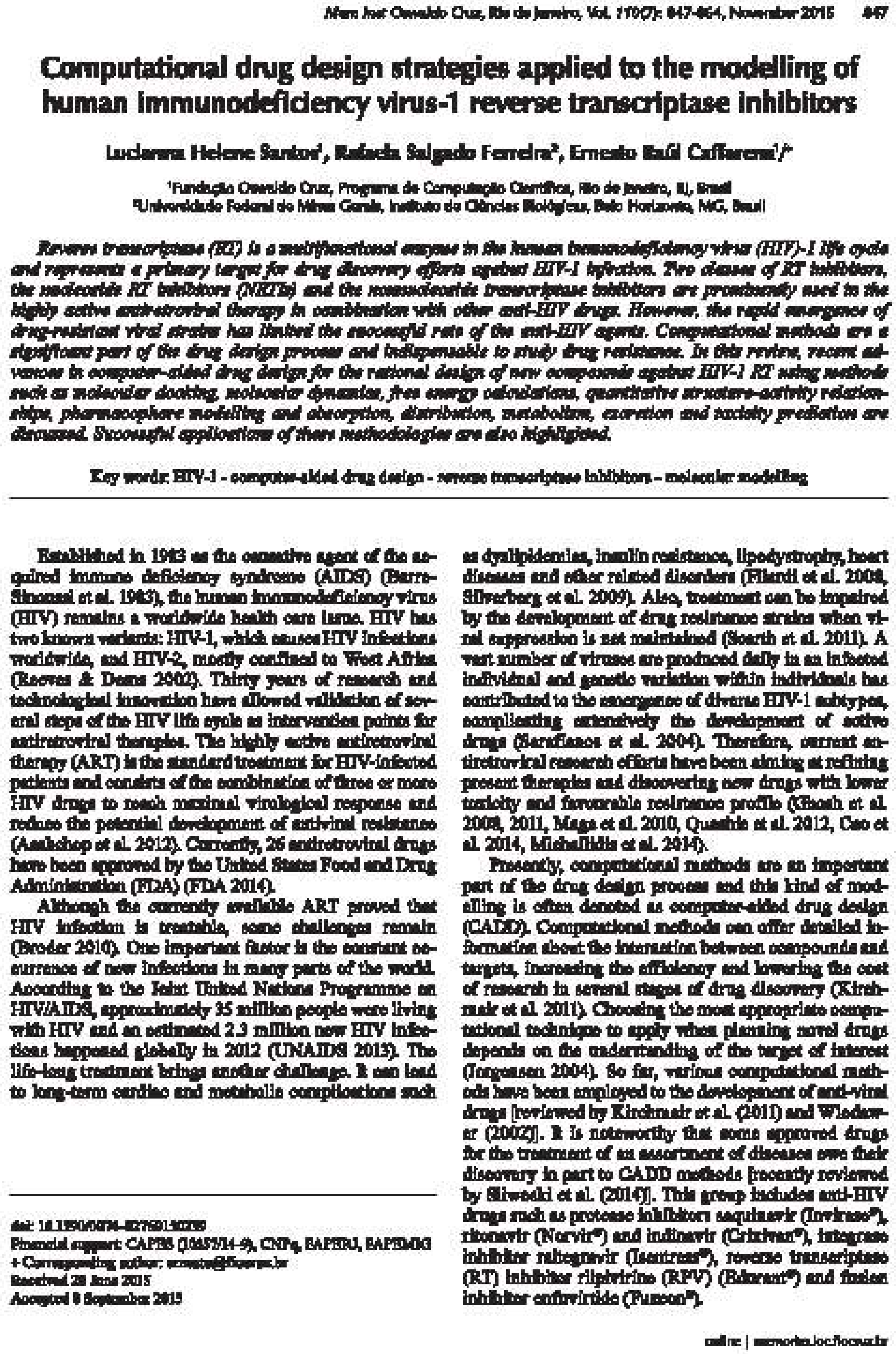
TI simulation

```
&cntrl
        imin = 0, nstlim = 250000, irest = 1, ntx = 5, dt = 0.002,
        ntt = 3, temp0 = 300.0, gamma_ln = 2.0, ig = -1,
        ntc = 2, ntf = 1,
        ntb = 2,
        ntp = 1, pres0 = 1.0, taup = 2.0,
         ioutfm = 1, iwrap = 1,
        ntwe = 2500, ntwx = 10000, ntpr = 10000, ntwr = 10000,

        icfe = 1, clambda = 0.00, scalpha = 0.5, scbeta = 12.0,
        logdvdl = 1,
        ifmbar = 1, bar_intervall = 1000, bar_l_min = 0.0, bar_l_max = 1.0,
        bar_l_incr = 0.1,
        timask1 = ':1', timask2 = ':2',
        ifsc = 0, crgmask = ':1@C01,H01,H02,H03',
/

&ewald
/
```

6 APPENDIX D – PUBLISHED PAPER



Mem Inst Oswaldo Cruz, Rio de Janeiro, Vol. 110(7): 847-864, November 2015 847

Computational drug design strategies applied to the modelling of human immunodeficiency virus-1 reverse transcriptase inhibitors

Lucianna Helene Santos[1], Rafaela Salgado Ferreira[2], Ernesto Raúl Caffarena[1]/+

[1]Fundação Oswaldo Cruz, Programa de Computação Científica, Rio de Janeiro, RJ, Brasil
[2]Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Belo Horizonte, MG, Brasil

Reverse transcriptase (RT) is a multifunctional enzyme in the human immunodeficiency virus (HIV)-1 life cycle and represents a primary target for drug discovery efforts against HIV-1 infection. Two classes of RT inhibitors, the nucleoside RT inhibitors (NRTIs) and the nonnucleoside transcriptase inhibitors are prominently used in the highly active antiretroviral therapy in combination with other anti-HIV drugs. However, the rapid emergence of drug-resistant viral strains has limited the successful rate of the anti-HIV agents. Computational methods are a significant part of the drug design process and indispensable to study drug resistance. In this review, recent advances in computer-aided drug design for the rational design of new compounds against HIV-1 RT using methods such as molecular docking, molecular dynamics, free energy calculations, quantitative structure-activity relationships, pharmacophore modelling and absorption, distribution, metabolism, excretion and toxicity prediction are discussed. Successful applications of these methodologies are also highlighted.

Key words: HIV-1 - computer-aided drug design - reverse transcriptase inhibitors - molecular modelling

Established in 1983 as the causative agent of the acquired immune deficiency syndrome (AIDS) (Barre-Sinoussi et al. 1983), the human immunodeficiency virus (HIV) remains a worldwide health care issue. HIV has two known variants: HIV-1, which causes HIV infections worldwide, and HIV-2, mostly confined to West Africa (Reeves & Doms 2002). Thirty years of research and technological innovation have allowed validation of several steps of the HIV life cycle as intervention points for antiretroviral therapies. The highly active antiretroviral therapy (ART) is the standard treatment for HIV-infected patients and consists of the combination of three or more HIV drugs to reach maximal virological response and reduce the potential development of antiviral resistance (Arts & Hazuda 2012). Currently, 26 antiretroviral drugs have been approved by the United States Food and Drug Administration (FDA) (FDA 2014).

Although the currently available ART proved that HIV infection is treatable, some challenges remain (Broder 2010). One important factor is the constant occurrence of new infections in many parts of the world. According to the Joint United Nations Programme on HIV/AIDS, approximately 35 million people were living with HIV and an estimated 2.3 million new HIV infections happened globally in 2012 (UNAIDS 2013). The life-long treatment brings another challenge. It can lead to long-term cardiac and metabolic complications such as dyslipidemias, insulin resistance, lipodystrophy, heart diseases and other related disorders (Filardi et al. 2008, Silverberg et al. 2009). Also, treatment can be impaired by the development of drug resistance strains when viral suppression is not maintained (Smith et al. 2011). A vast number of viruses are produced daily in an infected individual and genetic variation within individuals has contributed to the emergence of diverse HIV-1 subtypes, complicating extensively the development of active drugs (Sarafianos et al. 2004). Therefore, current antiretroviral research efforts have been aiming at refining present therapies and discovering new drugs with lower toxicity and favourable resistance profile (Ghosh et al. 2008, 2011, Maga et al. 2010, Quashie et al. 2012, Cao et al. 2014, Michailidis et al. 2014).

Presently, computational methods are an important part of the drug design process and this kind of modelling is often denoted as computer-aided drug design (CADD). Computational methods can offer detailed information about the interaction between compounds and targets, increasing the efficiency and lowering the cost of research in several stages of drug discovery (Kirchmair et al. 2011). Choosing the most appropriate computational technique to apply when planning novel drugs depends on the understanding of the target of interest (Jorgensen 2004). So far, various computational methods have been employed to the development of anti-viral drugs [reviewed by Kirchmair et al. (2011) and Wlodawer (2002)]. It is noteworthy that some approved drugs for the treatment of an assortment of diseases owe their discovery in part to CADD methods [recently reviewed by Sliwoski et al. (2014)]. This group includes anti-HIV drugs such as protease inhibitors saquinavir (Invirase®), ritonavir (Norvir®) and indinavir (Crixivan®), integrase inhibitor raltegravir (Isentress®), reverse transcriptase (RT) inhibitor rilpivirine (RPV) (Edurant®) and fusion inhibitor enfuvirtide (Fuzeon®).

doi: 10.1590/0074-02760150239
Financial support: CAPES (10457/14-9), CNPq, FAPERJ, FAPEMIG
+ Corresponding author: ernesto@fiocruz.br
Received 28 June 2015
Accepted 9 September 2015