In silico screening of inhibitors and conformational analysis of HCV NS5B polymerase

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt der

Naturwissenschaftlichen Fakultät I

(Biowissenschaften)

der Martin Luther Universität Halle-Wittenberg

von

Frau Tanaporn Uengwetwanit

geb. 07 September 1981, Thailand

Gutachter/in:

1.Prof.Dr.Wolfgang Sippl, Halle (Saale), Deutschland

2.Prof.Dr.Gabriele Costantino, Parma, Italy

3.Prof.Dr.Gerhard Wolber, Berlin, Deutschland

Datum der Verteidigung: 24.07.2014

Acknowledgements

My warmest gratitude goes to Prof. Wolfgang Sippl, my advisor, whose expertise, enthusiasm, and patience, added considerably to my graduate experience. I have been fortunate to have an advisor who gave me the freedom to explore on my own and at the same time the insightful comments and suggestions.

I am deeply grateful to Dr. Dina Robaa for her suggestions, and provision of thesis proofreading. Appreciation also goes out to for all former or current colleagues in the Medical Chemistry group. In my daily work I have been blessed with a friendly and cheerful group of fellow students. I would like to offer my special thanks to Martin Pippel, Michael Scharfe and Ralf Heinke for all their assistance helped me along the way.

I would like to thank Tobias Hoffmann and Dr. Ralpl Golbik at the Institute for Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg for testing inhibitory activity of HCV NS5B polymerase inhibitors.

I have had the support and encouragement of Dr. Kanin Wichapong who is both a mentor and friend. His advice and comments have been a great help in molecular dynamics simulation.

I would like to acknowledge Dr. Kanthida Kusonmano for insightful and detailed discussion about machine learning.

I also want to special thank Siriphan Manocheewa and Dr. Andy Pang for reviewing the language of my thesis.

I recognize that my study would not have been possible without the financial assistance of Thai Government Science and Technology Scholarship.

During the study at Martin Luther University Halle-Wittenberg, I have been helped and supported by countless people. I would like to thanks them all here. Most importantly, I would like to thank my family for all their love and encouragement.

Contents

Page

List of tables vii
List of figures ix
List of abbreviations xiv
Chapter 1 Introduction 1
1.1 Hepatitis 1
1.2 Hepatitis C Virus (HCV) 1
1.2.1 Genotypes of HCV 1
1.2.2 Genetics and structure of HCV 2
1.2.3 Therapeutic approaches 5
1.3 HCV NS5B polymerase 5
1.3.1 HCV NS5B non-nucleoside inhibitors 8
I) Thumb site I non-nucleoside inhibitors (TS-I NNIs) 10
II) Thumb site II non-nucleoside inhibitors (TS-II NNIs) 11
III) Palm site I non-nucleoside inhibitors (PS-I NNIs) 13
IV) Palm site II non-nucleoside inhibitors (PS-II NNIs) 15
1.4 Virtual screening of HCV NS5B polymerase inhibitors 15
1.5 Aims of the present study 18
Chapter 2 Computational and experimental methods 20
2.1 Protein-ligand binding affinities 20
2.2 Molecular docking 22
2.2.1 Search algorithm 22
2.2.2 Scoring functions 23
2.2.3 Evaluation of docking performance 24
2.3 Molecular Mechanics-Generalized Born / Surface Area (MM-GB/SA) and Molecular Mechanics-Poisson Boltzmann / Surface Area (MM-PB/SA) 26
2.4 Hybrid Quantum Mechanics and Molecular Mechanics Generalized Born Surface Area (QM/MM-GB/SA) 27
2.5 Determining K_d using a fluorescence-based *in vitro* assay 28
2.6 *In vitro* NS5B polymerase activity assay 28

Chapter 3 Optimizing docking/scoring functions for HCV NS5B polymerase inhibitors **30**
3.1 Introduction 30
3.2 Cross-docking study 31
3.3 Evaluation of ensemble docking, flexible side-chain docking and rigid protein docking 32
3.3.1 Methods and datasets 32
3.3.2 Results and discussion 34
3.4 Rescoring of docking poses for the PS-I inhibitors 38
3.4.1 Methods and datasets 38
3.4.2 Results and discussion 39
Chapter 4 Optimizing virtual screening protocols for HCV NS5B polymerase inhibitors **45**
4.1 Introduction 45
4.2 Materials and methods 46
4.2.1 Docking 46
4.2.2 Random Forest (RF) 46
4.2.3 Structural Interaction Fingerprint (SIFt) 48
4.2.4 Two sites docking 49
4.2.5 Dataset 50
4.3 Results and discussion 53
Chapter 5 Virtual screening of HCV NS5B polymerase inhibitors using pharmacophore model and docking **62**
5.1 Introduction 62
5.2 Binding mode analysis of HCV796 63
5.3 Virtual screening based on HCV-796 binding mode 65
5.4 Pharmacophore modeling 68
5.5 Database screening 69
5.6 Experimental results 71
5.7 Binding mode analysis of the pharmacophore hits in genotype 2a HCV NS5B polymerase 73
5.7.1 MD simulations of genotype 2a HCV NS5B polymerase complexed with HCV796 74

5.7.2 MD simulations of the tested pharmacophore hits in complex with the genotype 2a HCV NS5B polymerase 77
5.8 West Nile virus (WNV) NS5 polymerase 82
5.9 Discussion 84
Chapter 6 Analysis of the resistance of Hepatitis C virus NS5B polymerase via docking and molecular dynamics simulation 87
6.1 Introduction 87
6.2 Methods 88
6.3 Results and discussion 89
Chapter 7 Conclusions 95
References 97
List of tables (Appendix) 114
List of figures (Appendix) 115
Appendix 116
Curriculum vitae ..... a
Erklärung ..... e

List of tables

Pages

Table 1. The global distribution of HCV genotypes 2

Table 2. Available crystal structures of HCV NS5B polymerase and non-nucleoside inhibitors available in the RSCB protein data bank. 9

Table 3. Types of scoring functions. 24

Table 4. Quality of crystal structures in TS-II dataset.....33

Table 5. Quality of crystal structures in PS-I dataset. 34

Table 6. Performance after rescoring by MM-GB/SA, QM/MM-GB/SA and other scoring functions. 41

Table 7. Correct pairs of 45 compounds in enrichment study (27 co-crystallized ligands and 18 analogs) ranked by Chem-score, Glide SP, MM-GB/SA and QM/MM-GB/SA. Calculation of correct pairs is described in chapter 2.2.3 Evaluation of docking performance. 42

Table 8. Enrichment factor of rescoring of docked poses of 35 co-crystallized actives. The binding energy of each pose was scored by Glide SP with flexible hydroxyl groups and MM/GB-SA rescoring. 42

Table 9. Predictive performance of RF model-1 on the fivefold cross validation set. 365 PS-I inhibitors and 926 decoys were used in total.....53

Table 10. Predictive performance of RF model-2 on the fivefold cross validation set. 124 potent PS-I and 113 weak PS-I inhibitors were used in total. 54

Table 11. Predictive performance of RF model-1 on the independent validation set. 55

Table 12. Predictive performance of RF model-2 on the independent validation set. 55

Table 13. The number of compounds that passed the selection criteria using six different filtering approaches. 56

Table 14. The number of known PS-I inhibitors (true positives) found in the top 100 ranking by using different approaches on the validation set. 'Hard case' denotes an enrichment study using only the hard case with decoys. Similarly, 'Soft case' denotes an enrichment study using only the soft case with decoys. Note, the decoys dataset was the same in both experiments. 'Whole' represents using the hard case and soft case plus the decoys in the enrichment study. The number of total active compounds in each case is shown in bracket. 57

Table 15. Inhibitory activity of HCV796 in different genotypes [98].....63
Table 16. Inhibitory activity of HCV796 in mutants [97]. The inhibitory activity is represented as fold change calculated as the ratio of the concentration used to a half maximal inhibitory concentration (IC_{50}) against HCV polymerase of the 1b Con1 strain. IC_{50} of HCV796 on 1b Con is 0.04 µM. 63
Table 17. Hydrogen bond summary of the complex of HCV796 and either HCV NS5B polymerase genotype 1b BK1 (3FQK) or Con1 (3FQL) strain. Distance cutoff is 3.00 Å, angle cutoff is 120.00°. 64
Table 18. Hydrogen bond analysis based on 6 ns MD simulation. 67
Table 19. Enrichment study. The study was carried on the dataset of 21 active compounds and 348 decoys. Present of a pharmacophore feature is indicated as 'x', whereas absent of a pharmacophore feature is indicated as '-'.....69
Table 20. Equilibrium dissociation constants (K_d) between inhibitors and HCV polymerase in the absence of RNA. The K_d value of HCV796 (positive control) in the HCV polymerase assay is 27 µM based on the study of Reich [176].....72
Table 21. Differences of the per residue energy (kcal/mol) between the complex of HCV796 and HCV polymerase genotype (GT) 1b Con-1 and genotype 2a. Bars represent values (kcal/mol) on a scale. A blue bar denotes the per-residue energy of genotype (GT) 2a is lower than the per-residue energy of genotype 1b at a given residue position. In vice versa, a red bar denotes the per-residue energy of genotype (GT) 2a is greater than the per-residue energy of genotype 1b at a given residue position. Non-polar solvation energy is relative small and therefore it is not presented here.....76
Table 22. Binding energy of 11 tested pharmacophore hits calculated by the Generalized Born (MM-GB/SA) model using 100 frames from the MD simulation. 77
Table 23. The logarithm of the octanol/water partition coefficient (log P) of 11 tested pharmacophore hits and HCV796 79
Table 24. Inhibition of compound A and compound B against HCV NS5B polymerase containing resistance mutations.....89
Table 25. Mean total per-residue energy (kcal/mol) of the key amino acids in the binding pocket of the wild-type (WT) and mutant enzymes with compound A and compound B. The grey and red background shading represents the energy difference of the residue contributions in the mutants that are between 0.5-1 kcal/mol and greater than 1 kcal/mol relative to the wild-type, respectively. The boxes show the point mutations. 93

List of figures

Pages

Figure 1. Sequence alignment of six genotypes of HCV NS5B polymerases. A representative sequence of each genotype is indicated by genotype's name and GenBank accession number. The color bars above the sequence indicate domains. Green bars indicate finger domain, blue represents palm domain and orange represents thumb domain. Surrounding amino acid residues within 4.5 Å of the palm site II inhibitor HCV796 are indicated by the red crosses on the ruler. 3

Figure 2. HCV genome structure and expression. (A) HCV genome consists of approximately 9600 nucleotides and encodes a polyprotein of ~3000 amino acids. The translated region is flanked by conserved 5' and 3' untranslated regions (UTR). An internal ribosome entry site (IRES) at 5'UTR induces the host ribosomes to initiate the translation of the viral genome. The polyprotein is cleaved by host and viral proteases to generate structural and non-structural proteins. The cleavage sites for ER signal peptidases (black diamond) and virus proteases (arrow down) are indicated. Structural proteins consist of core (C), envelope proteins (E1 and E2) and p7. P7 is a small trans-membrane protein whose putative functions as ion channel [15]. Non-structural (NS) protein 2 is an auto-protease which cleaves itself from NS2/NS3 protein. NS3 serves as both serine protease and helicase/NTPase. NS4A is a cofactor for NS3 protein and is also required for the phosphorylation of NS5A [16]. NS4B induces the rearrangement of intracellular lipid membranes derived from the endoplasmic reticulum to form a structure called membranous web. NS5A is a phosphoprotein of unknown function. NS5B is an RNA dependent RNA polymerase which catalysts RNA synthesis [17, 18] (B) Localization of HCV polyprotein cleavage products. 4

Figure 3. The structure of HCV NS5B polymerases and its allosteric binding sites. The three dimensional structure of NS5B is shown as ribbon colored according to the domains; fingers domain are colored in blue, palm domain is shown in pink and thumb domain is shown in red. Inhibitors binding to the allosteric binding sites are shown in space-filling model. The amino acid in brackets presents resistant mutations occurring in the presence of particular inhibitors. 6

Figure 4. Structural comparison of the closed and open conformation of HCV NS5B. The closed conformation is shown as ribbon and colored according to domains; fingers are

colored in green, palm is colored in blue, thumb is colored in orange and ∧1 loop is colored in cyan. The open conformation is colored in pink. PDB ID: 1YUY and 1YVX were used for the closed and open conformation, respectively. 7
Figure 5. Structure of JTK-109. 10
Figure 6. 2D ligand interaction plot for an indole inhibitor (PDB ID: 2BRL) [38]. 11
Figure 7. Thumb site II inhibitors [62]. 12
Figure 8. 2D ligand interaction plot for the dihydropyrone inhibitor crystallized in 3FRZ [60]. 13
Figure 9. Palm site I inhibitors [55, 62, 71]. 14
Figure 10. 2D ligand interaction plot for a benzothidiazine inhibitor (PDB ID: 3H5U [72]). 14
Figure 11. 2D ligand interaction plot of HCV796, a benzofurancarboxamide derivative (PDB ID: 3FQL). 15
Figure 12. Thermodynamic cycle of protein-ligand binding 26
Figure 13. Performance of different docking/scoring functions on the TS-II dataset (22 crystal structures). Performance was quantified by the correctness of the top ranked pose (denoted as 'top pose') and the pose that was nearest to its native crystal structure within 2.5 Å (i.e. lowest RMSD as 'best pose'). The average number of correct poses and standard deviation bars were calculated from three individual docking runs. 36
Figure 14. Prediction accuracy of different docking/scoring functions on the PS-I dataset (35 crystal structures). Performance was quantified by the correctness of the top ranked pose (denoted as 'top pose') and the pose that was nearest to its native crystal structure within 2.5 Å (i.e. lowest RMSD as 'best pose'). The average number of correct poses and standard deviation bars were calculated from three individual docking runs. 37
Figure 15. Correlation plot between observed free energy ΔG_{obs} ($RTlnIC_{50}$) of 20 co-crystallized inhibitors and predicted enthalpy ΔH_{est} (Left) and predicted binding free energy ΔG_{est} (Right). These 20 co-crystallized inhibitors are subset of 45 compounds in enrichment study (27 co-crystallized ligands and 18 analogs). 7 co-crystallized inhibitors were excluded because of problems calculating the quasi-harmonic entropy. The energy unit is kcal/mol. Predicted energies were calculated by using MM-GB/SA (igb=8) in Amber 12 averaged over 100 snapshots from the last 2ns of MD simulations Entropy was calculated by using quasi-harmonic entropy approximation. 42

Figure 16. Comparative performance of Glide SP scoring function and rescoring docking poses by MM-GB/SA: ROC plot (upper) and percent of actives found at the top 5 % of the screening (lower). The dataset consists of 35 PS-I inhibitors and 859 decoys.....43
Figure 17. Schematic flowchart of SIFt calculation. 49
Figure 18. Boxplot of Glide SP scores of docking 396 PS-I inhibitors, 300 TS-II inhibitors and 326 decoys into two binding sites: palm site I (PS-I) and thumb site I (TS-II). 50
Figure 19. Boxplot of inhibitory profile (pIC_{50} = -log IC_{50}) for different binning clusters of PS-I inhibitors. 52
Figure 20. Aggregate number of known PS-I inhibitors found in top 100 ranking using six filtering approaches and Glide SP without filtering on the validation set. 56
Figure 21. Schematic flowchart of the virtual screening setup to identify novel HCV NS5B polymerase inhibitors. 59
Figure 22. Structures and ChemBridge ID of proposed palm site I inhibitors. 60
Figure 23. Superimposition of the crystal structure of 1Z4U (yellow), with the docked poses of proposed compounds (cyan) ID: 17820853 (Left) and ID: 33772211 (Right). Both hits show a similar shape and interaction profile as the known HCV NS5B inhibitor. 61
Figure 24**.** HCV796, a benzofuran derivative binding to PS-II. 62
Figure 25. Binding mode of HCV796 (balls and sticks) in genotype 1 HCV NS5B polymerase. The Con-1 strain (PDB ID: 3FQL) is shown in pink whereas the BK strain (PDB ID: 3FQK) is shown in blue. 64
Figure 26. Structures and ChemBridge ID of 14 experimental tested compounds (dataset-I) on genotype 2a HCV NS5B polymerase. 66
Figure 27. Overlay of compound C12 and the pharmacophore features of HCV796. The hydrogen bond donor (HD) is represented as green circle and hydrogen bond acceptor (HA) as red circle. 67
Figure 28. Docking poses of compound C3 and C7 at the PS-II of genotype 1b HCV NS5B polymerase. 67
Figure 29. LigandScout pharmacophore model derived from the NS5B-HCV796 structure. The pharmacophore features are represented by LigandScout color codes: hydrogen bond donor: HD (green arrow), hydrogen bond acceptor: HA (red arrow), hydrophobic region: HP (yellow sphere), and excluded volume (black sphere). 69
Figure 30. Flow chart of the virtual screening setup. 70

Figure 31. 2D structures of 11 purchased compounds from ChemBridge, ChemDiv and LifeChemicals which were selected for *in vitro* testing (dataset-II).....71

Figure 32. Relative inhibitory activities of 11 tested compounds. Relative inhibitory activity was calculated by dividing the RNA concentration in the presence of the presumably inhibitor by the RNA concentration in the absence of an inhibitor (Figure A8-Appendix). 100% NS5B polymerase activity was defined as the RNA yield in the control. Error bars represent standard deviation from three replications. 72

Figure 33. Superimposition of palm site II pocket of HCV polymerase genotype 1b Con1 (3FQL; pink) and genotype 2a JFH1 (3I5K; cyan). HCV796 is shown in balls and sticks. 74

Figure 34. Contact preference map for HCV796 in A) the HCV polymerase genotype 1b (3FQL) and B) HCV polymerase genotype 2a (3I5K). The yellow contour denotes the hydrophobic preference and the blue denotes the hydrophilic preference. 75

Figure 35. The docking poses and MD poses of compound T1 and T4. The MD poses were averaged over 100 snapshots from the last 2 ns of the MD simulation. 78

Figure 36. Protein-ligand interaction diagrams of HCV796 with A) genotype 1b [3FQL] and B) genotype 2a HCV NS5B polymerase. 79

Figure 37. Protein-ligand interaction diagrams of compounds T5 and T8 with genotype 2a HCV NS5B polymerase.....80

Figure 38. Protein-ligand interaction diagrams of compounds T9 and T11 with genotype 2a HCV NS5B polymerase. 81

Figure 39. Superimposed structures of PS-II HCV NS5B polymerase (pink) and the homology model of WNV NS5 polymerase (blue). HCV796 is presented in grey balls and sticks. 83

Figure 40. Bound conformation of compound T9 in the PS-II of WNV NS5 polymerase. The structure was averaged over 100 snapshots during the last 2 ns of the MD simulation. No hydrogen bods between inhibitor and NS5 are observed. 83

Figure 41. MD simulation of the WNV NS polymerase-T9 complex. A) Root mean square deviation (Å) of the protein (red) and compound T9 (cyan) B) superimposed structures of the homology model 3I5K (blue) and the average structures over 100 snapshots during last 2 ns (purple). 84

Figure 42. Structural comparison between HCV796 and compound T9. A) Bound conformation of HCV796 and T9. The bound conformation of HCV796 was obtained

from the crystal structure 3FQL. The bound form of compound T9 represents the stable form observed in the MD simulation. B) 2D structure of HCV796 and compound T9. 85
Figure 43. Structures of the two benzimidazole-5-carboxamides. 88
Figure 44. Structure of a co-crystalized ligand NS5B inhibitor (PDB: 2BRL). 88
Figure 45. Bound structures of three inhibitors (balls and sticks)--compound **A** (yellow), compound **B** (orange) and the co-crystallized indole derivative of 2BRL (cyan) -- in the binding site of wild-type HCV NS5B polymerase. 89
Figure 46. 2D representation of the predicted docking pose of compound **A** and compound **B** within thumb site I of the wild-type HCV NS5B polymerase. Hydrophobic amino acids are colored green, polar residues are colored pink. Hydrogen bonds and arene-H interactions are indicated by dashed lines. Ligand and protein solvent exposure is indicated by the blue spheres. 90
Figure 47. Average structures of the inhibitor bound to wild type HCV NS5B polymerase obtained from snapshots of the MD simulations. A) compound **A** and B) compound **B** are shown in ball-and-stick mode. 91

List of abbreviations

2D	=	2 Dimension
3D	=	3 Dimension
K_d	=	Dissociation constant
Ala (A)	=	Alanine
Arg (R)	=	Arginine
Asn (N)	=	Asparagine
Asp (D)	=	Aspartic acid
Cys (C)	=	Cysteine
DAA	=	Direct-acting antiviral agent
EF	=	Enrichment factor
FN	=	False negative
FP	=	False positive
FPR	=	False positive rate
GB	=	Generalized Born
GA	=	Genetic algorithm
GT	=	Genotype
Glu (E)	=	Glutamic acid
Gln (Q)	=	Glutamine
Gly (G)	=	Glycine
HCV	=	Hepatitis C virus
His (H)	=	Histidine
HA	=	Hydrogen bond acceptor
HD	=	Hydrogen bond donor
HP	=	Hydrophobic
Inh	=	Inhibitor(s)
Ile (I)	=	Isoleucine
Leu (L)	=	Leucine
Lys (K)	=	Lysine
Met (M)	=	Methionine

MD	=	Molecular dynamics
MM	=	Molecular mechanics
NNI	=	Non-nucleoside (or nucleotide) inhibitors
NI	=	Nucleoside (or nucleotide) inhibitors
Ntree	=	Number of trees to grow in random forest
Mtry	=	Number of variables in random forest
PS	=	Palm site
Phe (F)	=	Phenylalanine
PB	=	Poisson-Boltzmann
Pro (P)	=	Proline
QM	=	Quantum mechanics
RF	=	Random forest
ROC	=	Receiver operating characteristic
RdRp	=	RNA dependent RNA polymerase
RMSD	=	Root mean square deviation
Ser (S)	=	Serine
SIFt	=	Structural interaction fingerprint
SVM	=	Support vector machine
SA	=	Surface area
Tc	=	Tanimoto coefficient
IC_{50}	=	The concentration of an inhibitor where the response is reduced by half.
DUD	=	The directory of useful decoys
Thr (T)	=	Threonine
TS	=	Thumb site
TN	=	True negative
TP	=	True positive
TPR	=	True positive rate
Trp (W)	=	Tryptophan
Tyr (Y)	=	Tyrosine
FDA	=	US Food and Drug Administration
Val (V)	=	Valine
vdW	=	Van der Waal's

VS	=	Virtual screening
WNV	=	West-Nile virus

Chapter 1 Introduction

1.1 Hepatitis

Hepatitis, or liver inflammation, is caused by several factors; the most common however is viral infection. There are five main types of hepatitis viruses; namely types A, B, C, D and E. All types of hepatitis infections show common symptoms which can include one or more of the following: fever, loss of appetite, extreme fatigue, nausea, vomiting, abdominal pain, jaundice and dark urine [1]. Acute hepatitis resolves completely in six months after infection while chronic infections can last significantly longer.

Hepatitis A (HAV) and hepatitis E virus (HEV) are transmitted via the fecal-oral route; good sanitation is hence the most effective way to combat the disease. The infection is rarely fatal and HAV or HAE patients usually recover with no lasting liver damage [2, 3]. Hepatitis B (HBV), hepatitis C (HCV) and hepatitis D or delta virus (HDV), on the other hand, spread through percutaneous contact with infected blood or body fluid and also through sexual intercourse. HDV requires the help of a hepadnavirus like hepatitis B virus for its own replication; therefore hepatitis D infection cannot occur without HBV. Both hepatitis B and C viruses can cause chronic infections. The potential for developing a chronic infection is 30-50% in HBV-infected children aged between one and five years, and 6-10% in older children and adults [4], while 75-85% of HCV-infected patients develop a chronic infection [5]. Currently, effective vaccines against hepatitis A and B viruses are available, but there is no vaccination against hepatitis C.

1.2 Hepatitis C Virus (HCV)

Developing into a chronic infection is a remarkable feature of HCV infections. It is estimated that approximately 170 million people are infected with HCV [6]. 60-70% will develop chronic liver disease and 5-20% of the chronic live patients will develop cirrhosis [7, 8].

1.2.1 Genotypes of HCV

The HCV RNA genome exhibits a high degree of genetic variability owing to the high replication rate and the lack of a proofreading mechanism during RNA synthesis. Due to the sequence diversity of hepatitis C, at least six genotypes of HCV have been assigned by phylogenetic methods [9]. Genotypes are defined by 31-33% nucleotide differences and subtypes

by 20-25% [10]. Sequence alignment of the six genotypes is shown in Figure 1. Viral mutations occur spontaneously over time and enable the virus population to persist in their hosts. Slightly different genetic variations of a present genotype is referred to as quasispecies [11]. Genetic diversity of HCV is essential for proper association between the genotype and clinical response (Table 1) [12, 13]. Moreover different genotypes have different geographical distribution.

Table 1. The global distribution of HCV genotypes

Genotypes	Region
1a	Mostly in North and South America and also common in Australia
1b	Common in North America, Europe and Japan
2b	Most common genotype 2 in the USA and northern Europe
2c	Most common genotype 2 in western and southern Europe
3a	Common in southern Asia
4a	Highly prevalent in Egypt
4c	Highly prevalent in central Africa
5	Common in south Africa
6	Common in Asia

1.2.2 Genetics and structure of HCV

HCV is a member of genus *Hepacivirus*, family *Flaviviridae* which also includes *Flavivirus* and *Pestivirus* [14]. *Flaviviruses* are recognized human pathogens and include yellow fever virus, dengue fever virus, Japanese encephalitis virus and Tick-borne encephalitis virus. *Pestiviruses* infect cattle, sheep and swine causing grave problems to the agricultural economy.

The HCV genome shares a number of basic structural characteristics with the other members of the *Flaviviridae*. It is a positive single-stranded RNA [(+) ssRNA] surrounded by a hexagonal capsid and an envelope made up of two lipid bilayers, where at least two or more envelope proteins (E) are anchored. The approximately 9.6 kb HCV genome encodes one open reading frame (ORF) which is translated into one polyprotein. The polyprotein is cleaved by cellular and viral proteases into both structural (core, E1, E2) and non-structural components (p7, NS2, NS3, NS4A/B and NS5A/B) (Figure 2).

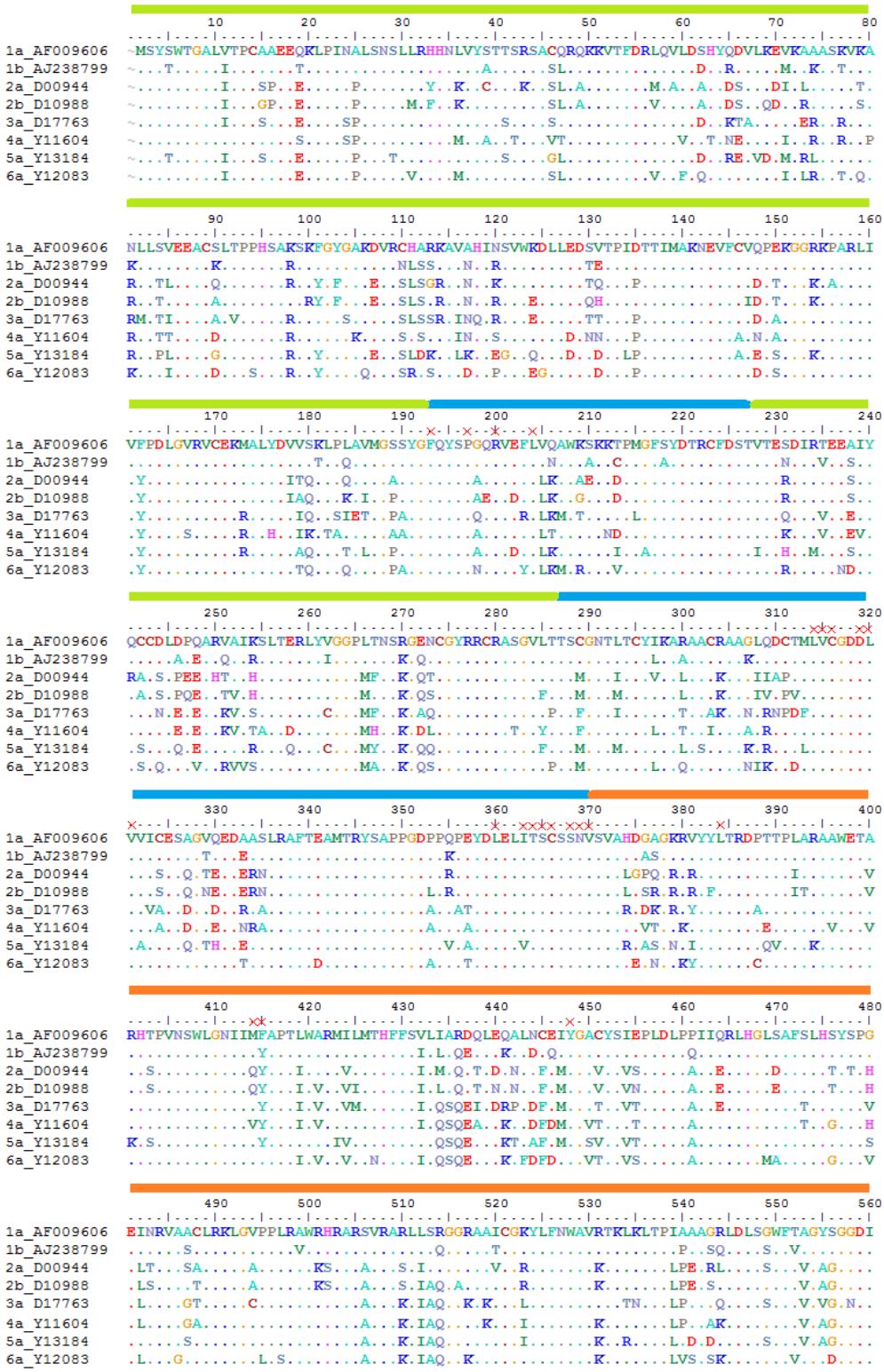


Figure 1. Sequence alignment of six genotypes of HCV NS5B polymerases. A representative sequence of each genotype is indicated by genotype's name and GenBank accession number. The color bars above the sequence indicate domains. Green bars indicate finger domain, blue represents palm domain and orange represents thumb domain. Surrounding amino acid residues within 4.5 Å of the palm site II inhibitor HCV796 are indicated by the red crosses on the ruler.

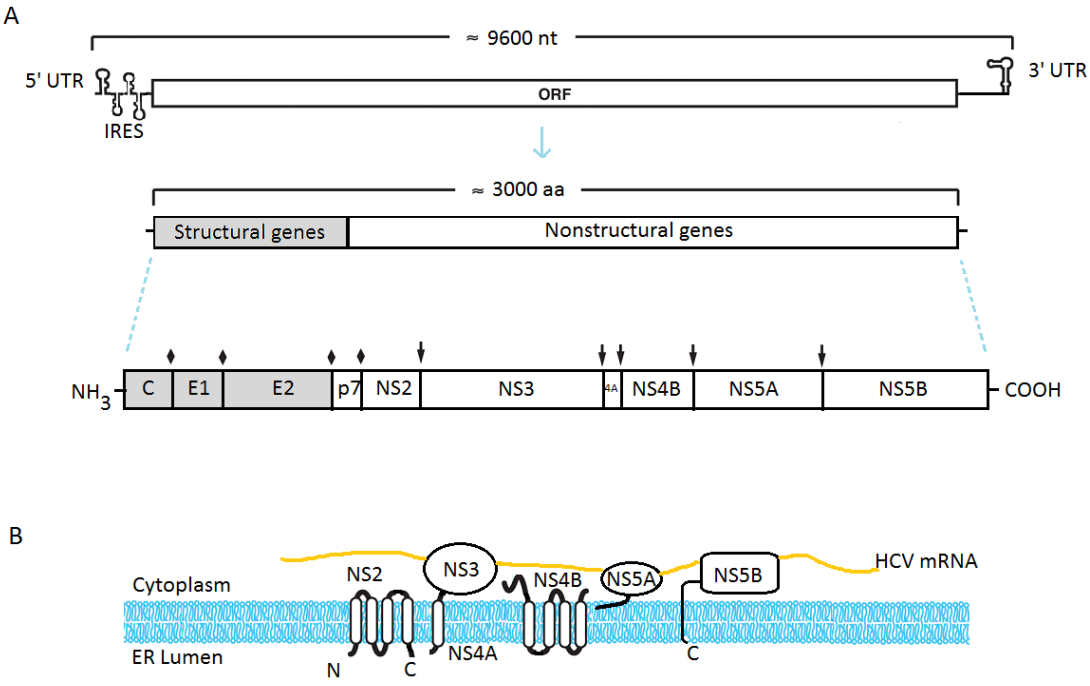


Figure 2. HCV genome structure and expression. (A) HCV genome consists of approximately 9600 nucleotides and encodes a polyprotein of ~3000 amino acids. The translated region is flanked by conserved 5' and 3' untranslated regions (UTR). An internal ribosome entry site (IRES) at 5'UTR induces the host ribosomes to initiate the translation of the viral genome. The polyprotein is cleaved by host and viral proteases to generate structural and non-structural proteins. The cleavage sites for ER signal peptidases (black diamond) and virus proteases (arrow down) are indicated. Structural proteins consist of core (C), envelope proteins (E1 and E2) and p7. P7 is a small trans-membrane protein whose putative functions as ion channel [15]. Non-structural (NS) protein 2 is an auto-protease which cleaves itself from NS2/NS3 protein. NS3 serves as both serine protease and helicase/NTPase. NS4A is a cofactor for NS3 protein and is also required for the phosphorylation of NS5A [16]. NS4B induces the rearrangement of intracellular lipid membranes derived from the endoplasmic reticulum to form a structure called membranous web. NS5A is a phosphoprotein of unknown function. NS5B is an RNA dependent RNA polymerase which catalysts RNA synthesis [17, 18] (B) Localization of HCV polyprotein cleavage products.

1.2.3 Therapeutic approaches

The standard treatment is a combination of pegylated interferon alpha and ribavirin. The primary goal of the treatment is to achieve a sustained virological response (SVR). SVR was defined as an undetectable HCV RNA in the serum six months after therapy was completed [6]. However the US Food and Drug Administration (FDA) now considers an assessment of SVR 12 weeks after the cessation of treatment, to be the primary endpoint [19]. Less than 50% of HCV genotype 1 and genotype 4 infected patients achieved a SVR [20, 21]. Patients infected with HCV genotypes 2, 3, 5 and 6 show better responses to the treatment. HCV genotype 2 is the easiest genotype to treat with current therapy, the success rate is up to 95% [22]. Besides the high variability of the treatment efficiency, a wide array of side effects has been reported: severe depression, hemolytic anemia, renal dysfunction and most commonly flu-like symptoms and fatigue. The HCV protease inhibitors: telaprevir and boceprevir, are the first direct-acting antiviral agents (DAAs) that have received drug approval in 2011 [19]. The triple therapy of pegylated interferon, ribavirin and a protease inhibitor (telaprevir or boceprevir) increases the efficiency in HCV genotype 1 infection up to 75% [23]. This triple regimen represents the new standard of care in HCV genotype 1 infected patients [24]. The protease inhibitors are expensive. The treatment dosing is every 7-9 hours. There have been concerns on drug interaction, poor tolerability and drug resistance with telaprevir and boceprevir. Moreover, the protease inhibitors are effective only in genotype 1 [19, 25]. Recently, there are four new DAAs that have been approved to the market and established new treatment for HCV chronic infection. Since December 2013, an uridine nucleotide analog-- sofosbuvir which inhibits NS5B polymerase, plus ribavirin are approved as first interferon-free therapy for genotype 1-4 infected patients [24]. A macrocyclic NS3/4A protease inhibitor-- simeprevir, and a peptidomimetic linear ketoamide-- faldaprevir together with pegylated interferon alpha are indicated for the treatment of patients with genotype 1 [24]. A NS5A inhibitor-- daclatasvir plus sofosbuvir represent another new regimen for patients infected with genotype 1-3 [26].

1.3 HCV NS5B polymerase

The HCV NS5B protein encodes an RNA dependent RNA polymerase (RdRp) which is responsible for RNA synthesis using an RNA template. NS5B is validated as a potential drug target as it is an essential enzyme for viral RNA replication [27].

The NS5B polymerase has five conserved motives and Gly-Asp-Asp (GDD) sequences as characteristic for polymerases [28]. NS5B has been classified as a tail anchored protein [29]; the 21 amino acids long hydrophobic C-terminus tethers to the endoplasmic reticulum (ER) membrane. These C-terminal residues are pivotal for the *in vivo* localization of the HCV replication complex [30, 31]. However, truncation of the C-terminus does not affect the enzyme's activity *in vitro*. NS5B lacking the 21 C-terminus (ΔC21) has been used in biochemical and structural studies because it does not require the use of heavy detergents for its purification [32].

The structure of NS5B polymerase shows three domains: finger, thumb and palm (Figure 3). The active site is located at the palm domain and is encircled by the finger and thumb domains. The small loop (∧1 and ∧2), connecting the finger and thumb domains [33], in addition to a β-hairpin loop, protruding in the RNA binding channel, both regulate the activity of the enzyme in RNA synthesis. The divalent metal ion is also required for the enzymatic activity.

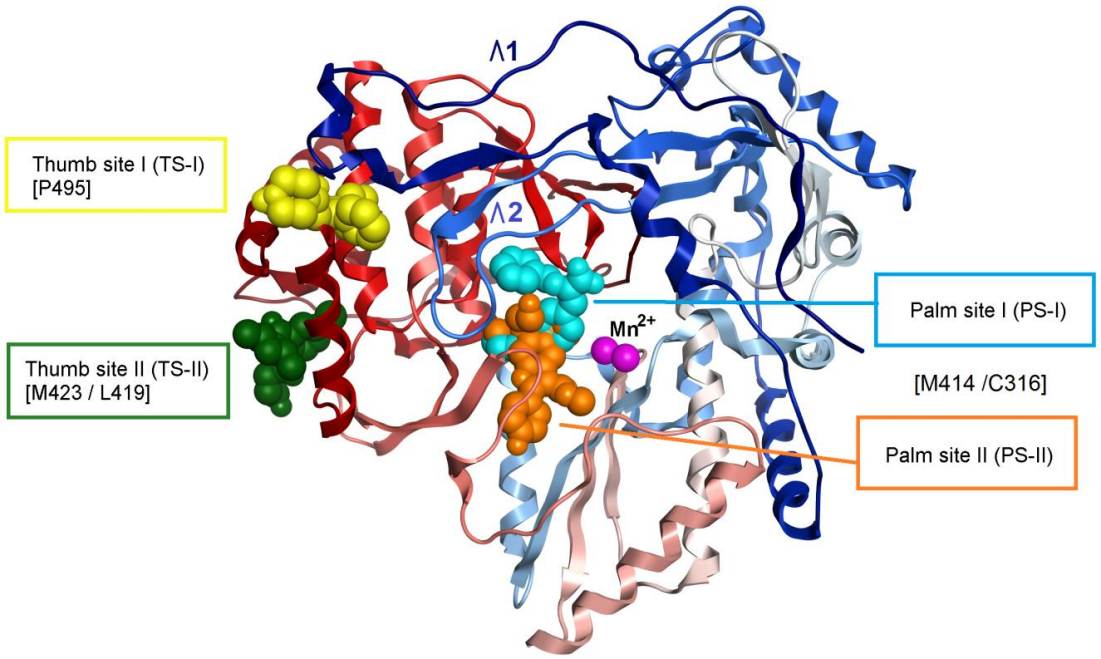


Figure 3. The structure of HCV NS5B polymerases and its allosteric binding sites. The three dimensional structure of NS5B is shown as ribbon colored according to the domains; fingers domain are colored in blue, palm domain is shown in pink and thumb domain is shown in red. Inhibitors binding to the allosteric binding sites are shown in space-filling model. The amino acid in brackets presents resistant mutations occurring in the presence of particular inhibitors.

During RNA replication, HCV NS5B polymerase undergoes several conformational changes. The closed conformation represents the initiation state of the polymerase as it is too narrow to

accommodate the duplex formed by the template and nascent RNA. The polymerase then undergoes transition to the open conformation, which has a larger cavity necessary for the elongation process [34, 35]. The available crystal structures support the hypothesis that inhibitors binding to the thumb subdomain bind only to the closed conformation and thereby inhibit a conformational change that is required for elongation [36, 37]. In the closed conformation, an elongated loop (Δ1 loop, Ile11-Ser46) at the tip of the finger domain protrudes to contact the thumb domain by packing its short alpha helix (helix A) against the alpha helix O (residues 388-401) and Q (residues 418-437) of the thumb domain (Figure 4) [34]. In the open conformation the tip of the fingertip Δ 1 loop, which has an alpha-helical structure in the closed conformation, moves away from the thumb domain and adopts a beta-hairpin-like structure [38].

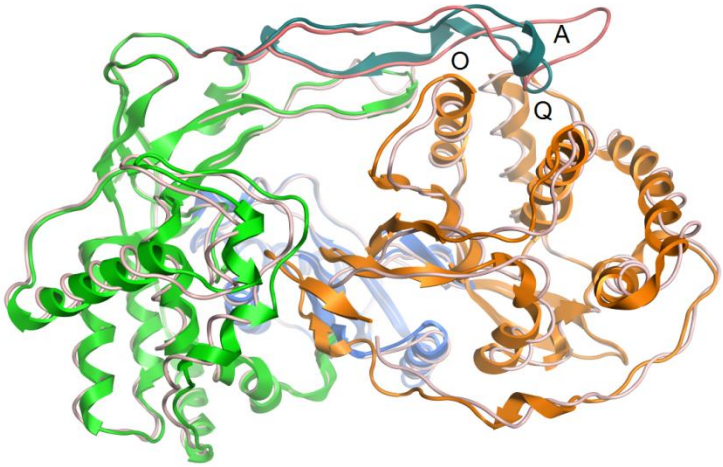


Figure 4. Structural comparison of the closed and open conformation of HCV NS5B. The closed conformation is shown as ribbon and colored according to domains; fingers are colored in green, palm is colored in blue, thumb is colored in orange and ∧1 loop is colored in cyan. The open conformation is colored in pink. PDB ID: 1YUY and 1YVX were used for the closed and open conformation, respectively.

Currently, studies on HCV replication are still restricted to genotypes 1 and 2 due to a robust cell culture system. To propagate in cell culture, genotype 1 undergoes adaptive mutations that enhance RNA replication, but these mutations result in the loss of virus particle production *in vivo* [39, 40]. The first generation of functional HCV replicons were derived from the consensus Con1 cDNA that was isolated from the liver of a patient chronically infected with the genotype 1b [41]. Slight variation in HCV sequence can dramatically alter the replicative ability. There are

two isolation of genotype 1, 1a and 1b [41]. NS5Bs from the BK (HCV-1b) isolate are about 5- to 10-fold more active than those derived from the H77 (HCV-1a) isolate [42]. An entire genomic RNA of JFH1 genotype 2a strain of HCV has the ability to produce infectious HCV particles both *in vitro* and *in vivo* without the requirement of cell culture adaptive mutations. A J6 strain is also a genotype 2a strain but only a chimeric RNA of J6/JFH can replicate efficiently in Huh7 cells [43, 44].

1.3.1 HCV NS5B non-nucleoside inhibitors

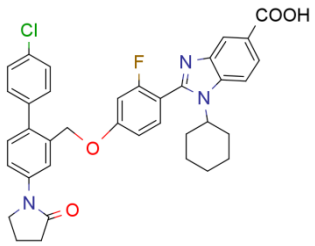
Several inhibitors targeting NS5B have been identified which can be grouped into nucleoside (NIs) and non-nucleoside inhibitors (NNIs). Nucleoside or nucleotide inhibitors mimic the natural polymerase substrate and competitively bind to the active site of the enzyme, which is situated in the palm domain. Nucleoside inhibitors which are incorporated into the nascent HCV RNA chain, cause chain termination. In contrast, non-nucleoside inhibitors are chemically diverse and bind to allosteric sites. At least four allosteric binding sites have been identified as non-nucleoside inhibitors' (NNI) binding site; thumb site I (TS-I or NNI-I), thumb site II (TS-II or NNI-II), palm site I (PS-I or NNI-III) and palm site II (PS-II or NNI-IV) (Figure 3). In this work, the focus was primarily set on non-nucleoside inhibitors. A summary of allosteric sites and co-crystallized NNIs is presented in Table 2 [45].

Table 2. Available crystal structures of HCV NS5B polymerase and non-nucleoside inhibitors available in the RSCB protein data bank.

Site	NNI chemotypes	PDB ID
TS-I (NNI-I)	Indoles	2BRK, 2BRL, 2DXS, 2WCX, 3MWW, 2XWY
TS-II (NNI-II)	Phenylalanines	1NHU, 1NHV
	Dihydropyranones	1OS5, 2HAI, 3FRZ
	Thiophene carboxylic acids	1YVX, 1YVZ, 2GIR, 3MF5, 2D3U, 2D3Z, 2D41
	Thiazolones	2HWH, 2HWI, 2I1R, 2O5D
	Bromophenyl methanones	3CIZ, 3CJ0, 3CJ2, 3CJ3, 3CJ4, 3CJ5
	Benzoisoquinolines-dione	2WHO, 3HVO
	Hexanoic acids	2WRM
	Quinolones	3PHE
	Acrylic acids	1YVF, 1Z4U
	Rhodanines	2AWZ, 2AXO, 2AX1
	Benzothiadiazines	2FVC, 2GIQ, 3HHK, 3BSA, 3BSC, 3CDE, 3BR9, 3E51, 3CO9, 3CVK, 3H2L, 3H98, 3GYN, 3IGV
PS-I (NNI-III)	Benzoisothiazoles dioxide	3D28, 3D5M, 3H5U, 3H5S
	Proline sulfonamides	2GC8
	Acylpyrrolodines	2JC0, 2JC1
	Anthranilic acids	2QE2, 2QE5
	Benzodiazepines	3GOL, 3CSO, 3GNV, 3GNW, 3HKW, 3HKY
PS-II (NNI-IV)	Benzofurans	3FQK, 3FQL

I) Thumb site I non-nucleoside inhibitors (TS-I NNIs)

Thumb site I, also known as finger loop site or non-catalytic GTP binding site, is approximately 30 Å away from the active site [46]. The benzimidazole-5-carboxylic acid derivative is a lead scaffold targeting this pocket and was discovered by Japan Tobacco and Boehringer Ingelheim. Two compounds of this structural family, JKT-003 (structure not disclosed) and JTK-109 [47] (Figure 5), were the first HCV NNIs that have been submitted to clinical trials, which have now been terminated for an undisclosed reason. An indole replacement of the benzimidazole core has been introduced to improve cellular permeability and replicon potency [48-50]. The co-crystal structure of an indole inhibitor in complex with HCV polymerase reveals that the binding of this class of inhibitors induces conformational change in the thumb and finger domains. A small alpha helix A moves 8 Å away from the open GTP binding site and consequently blocks the enzyme in an inactive conformation [38, 51]. Thus benzothiadiazines inhibit the initiation phase of RNA synthesis but have no effect on the elongation phase [52]. Benzothiadiazines are noncompetitive with NTPs or RNA [53]. The binding site of this class of inhibitors is shown in Figure 6. TS-I NNIs show a potent inhibitory activity on the genotype 1a, 1b and 3a but are less effective against genotype 2a [46].



IC_{50} (NS5B) = 0.017 µM; EC_{50} (replicon) = 0.32 µM

Figure 5. Structure of JTK-109.

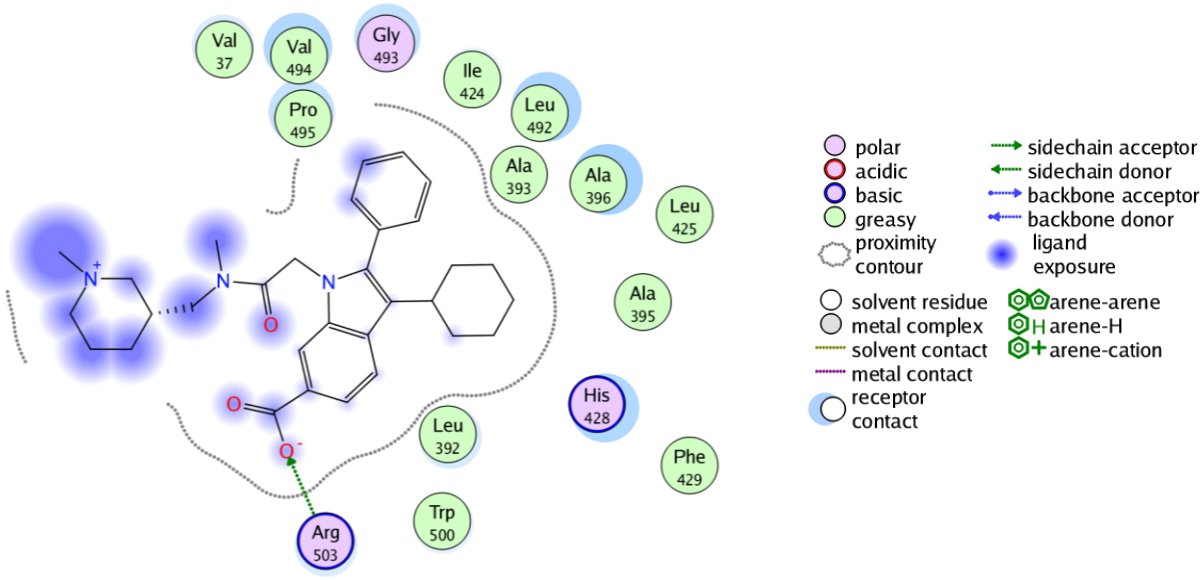
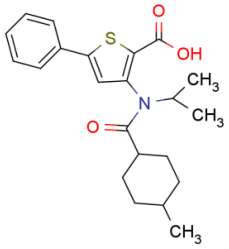


Figure 6. 2D ligand interaction plot for an indole inhibitor (PDB ID: 2BRL) [38].

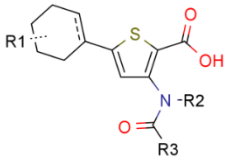
II) Thumb site II non-nucleoside inhibitors (TS-II NNIs)

TS-II is a 30 Å long hydrophobic cleft near the base of the thumb domain [54]. The TS-II is located 35 Å away from the active site and 10-15 Å from both the allosteric GTP binding site and from TS-I [55]. Several scaffolds such as thiophene-based carboxylic acid derivatives, dihydropyranone derivatives, and phenylalanine-based inhibitors, have been reported to bind to TS-II as shown in Table 2 and Figure 7. Even though amino acids in this pocket are quite conserved across different genotypes, the inhibitors are only efficient at genotype 1 [56]. The inhibitors typically bind in a dimple region defined by residues Leu419, Trp528, Tyr477 and Arg422. Hydrogen bond interactions with the backbone amides of Ser-476 and Tyr477 either directly or via a bridging water molecule are a key feature of this pocket [57-60] as shown in Figure 8. TS-II NNIs have been proposed to have similar mechanisms of inhibitions as TS-I NNIs [36, 61].

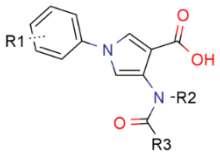
Thiophene and five-membered heterocyclic rings



Shire Biochem
IC_{50} (NS5B) =1.5 µM
EC_{50} (replicon) = 0.3 µM

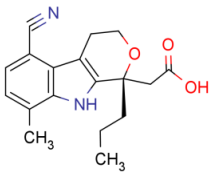


Virochem
IC_{50} (NS5B) < 5 µM

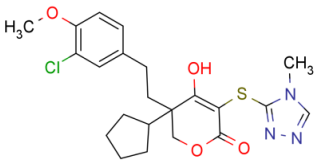


SmithKline Beecham
IC_{50} (NS5B) < 5 µM

Pyranoindole and hydroxydihydropyranones

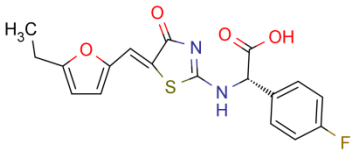


HCV-371
Wyeth/ViroPharm
IC_{50} (NS5B) =1.5 µM
EC_{50} (replicon) = 0.3 µM

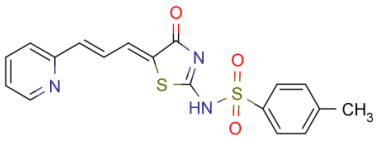


Pfizer
IC_{50} (NS5B) = 0.038 µM
EC_{50} (replicon) > 10 µM

Thiazolones



IC_{50} (NS5B) = 3 µM



IC_{50} (NS5B) = 0.6 µM
EC_{50} (replicon) = 35 µM

Figure 7. Thumb site II inhibitors [62].

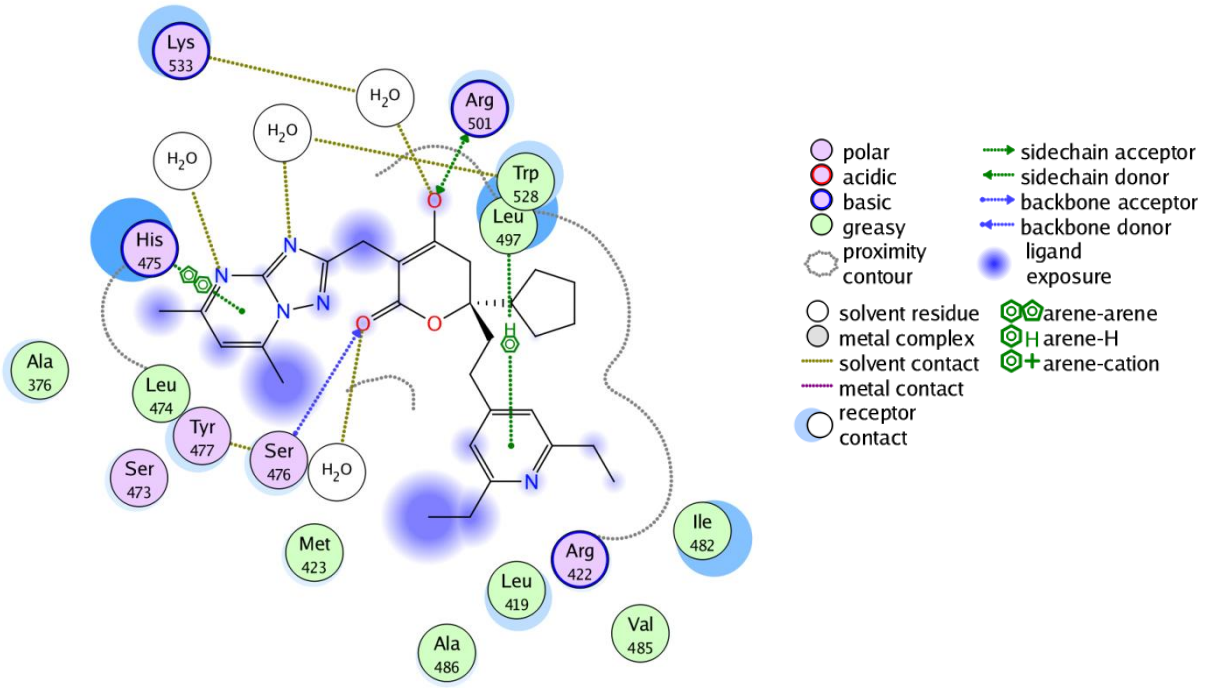
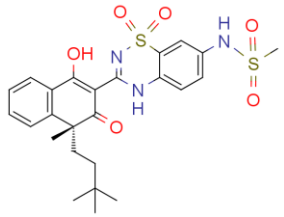


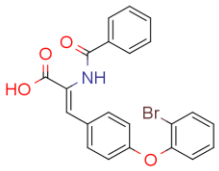
Figure 8. 2D ligand interaction plot for the dihydropyrone inhibitor crystallized in 3FRZ [60].

III) Palm site I non-nucleoside inhibitors (PS-I NNIs)

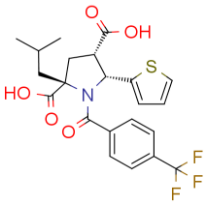
The third allosteric pocket is located at the junction of the thumb and palm domain near the active site. The first PS-I NNI was a benzothiadiazine, which was reported by GlaxoSmithKline in 2001 [53, 63]. Structural diversity has been further disclosed including acylpyrrolidines, rhodanines and isothiazoles (Figure 9). The inhibitors are expected to prevent the initiation of RNA synthesis and an elongation complex [62]. A-848837, a benzothiadizine derivative, demonstrates an excellent inhibitory potency in animals. However A-848837 inhibits only HCV polymerase genotype 1 and resistance emerges during treatment [64]. The observed *in vitro* resistant mutations were C316Y, M414T, Y448H/C, C251R, G554D, S556G or G558R, D559G [65-70]. Notably, the M414T mutation conferred cross-resistance to all NNIs regardless of the binding sites. The mechanism of resistance has not been elucidated in details [55, 56]. An example of the interactions of a PS-I NNI and HCV polymerase is shown in Figure 10.



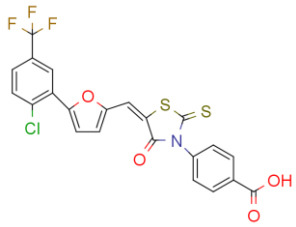
Thiadiazine derivative (A-848837)
IC_{50} (NS5B) = 0.3-0.6 nM)



Acrylic acid derivative
IC_{50} (NS5B) = 100 nM



GlaxoSmithKline inhibitor
IC_{50} (NS5B) = 0.19 µM



Rhodanines

Figure 9. Palm site I inhibitors [55, 62, 71].

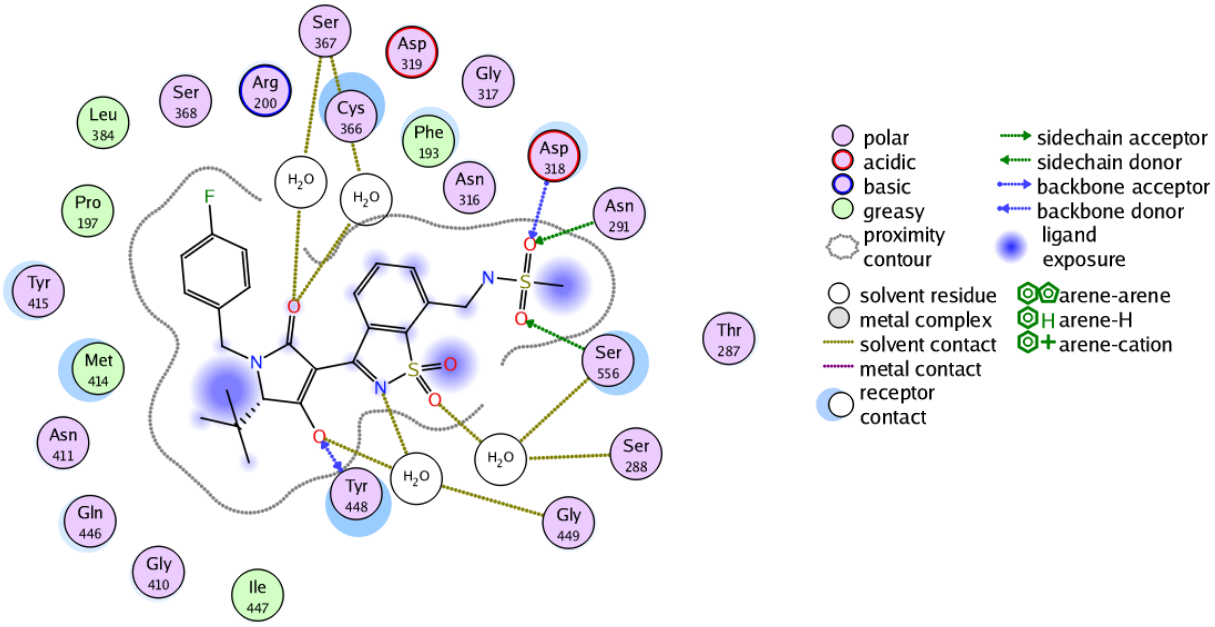


Figure 10. 2D ligand interaction plot for a benzothidiazine inhibitor (PDB ID: 3H5U [72]).

IV) Palm site II non-nucleoside inhibitors (PS-II NNIs)

PS-II is located between the primer grip motif (residues 364-369) and the central beta sheet (residues 214-219, 319-325 and 310-316). It partially overlaps with PS-I, sharing the amino acid residues: Phe193, Met414, Tyr415 and Tyr448 [51]. HCV796 was the first PS-II NNI with a positive clinical trial profile (Figure 11) [73, 74]. HCV796 displays potent and broad spectrum activity. The IC_{50} values range between 0.01-0.57 µM in genotypes 1a/1b, 3 and 4 and 1.7 µM in genotype 2 [75, 76]. In Phase II clinical trial, HCV796 shows an elevated of liver enzymes. Therefore, the study was terminated due to safety concerns.

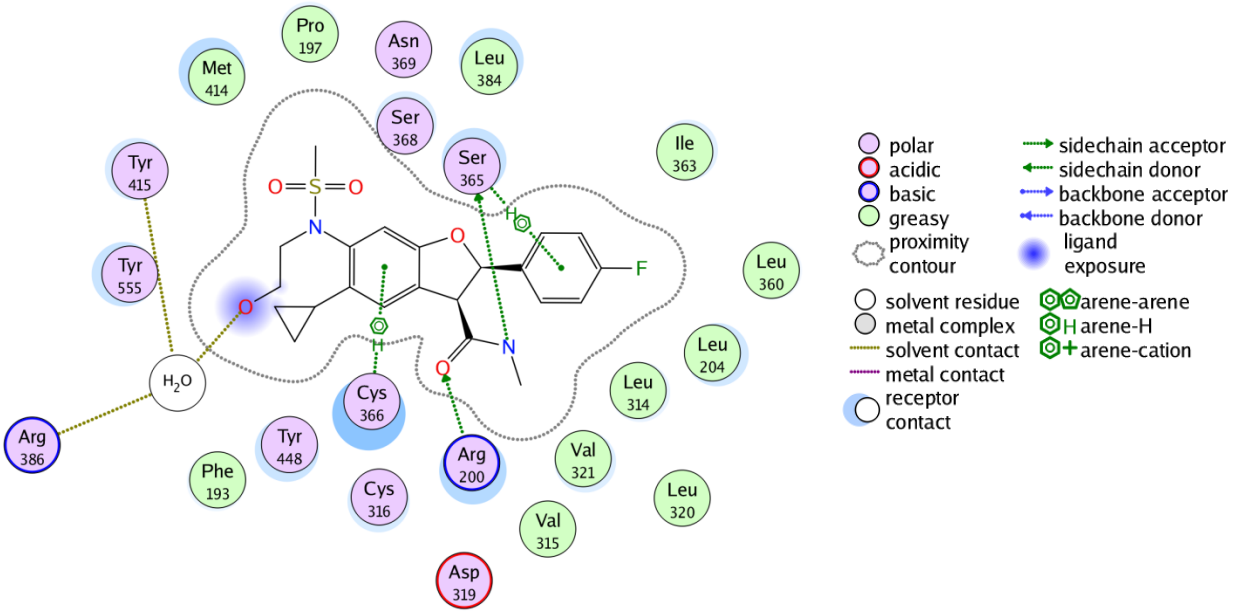


Figure 11. 2D ligand interaction plot of HCV796, a benzofurancarboxamide derivative (PDB ID: 3FQL).

1.4 Virtual screening of HCV NS5B polymerase inhibitors

Virtual screening (VS) aims to identify potential hits from the enormous number of chemical compounds in databases. Various computational methods have been applied as strategies to complement and streamline experimental assays in drug discovery projects. VS workflows generally consist of various stages and methods because each individual method has its own thorny problems and the performance of each method is not consistent [77]. There are plenty of ways to combine computational methods for VS. Each research group has developed and customized virtual screening workflows depending on the purpose and amount of information known about a particular target. Docking and pharmacophore-based VS, which are widely used and integrated in VS workflows, are discussed further in this thesis.

Small molecule docking is commonly used to predict compound-bound conformations, VS for hit identification and binding affinity prediction. Performance of docking varies and depends on the programs and protein targets. Warren et al [78] compared 10 docking programs (Dock4, DockIt, FlexX, Flo+, Fred, Glide, Gold, LigFit, MOE and MVP) against eight proteins including HCV polymerase. In that study, no docking program was able to predict the bound-conformation of the HCV NS5B polymerase inhibitors close to the native structure (within 2 Å) for over 40% of the studied complexes. Whereas, the results for other protein targets showed at least one program achieving an accuracy of pose prediction over 40% and the best performance was over 90%. Similarly, the best VS accuracy for HCV polymerase obtained from MVP and Flo+ programs was moderate. MVP and Flo+ programs gave enrichment factors of 3.6 and 3.4, respectively (maximum enrichment factor was 9.5). The authors stated that the efficiency of docking programs for HCV NS5B polymerase was low because of the size of the search space. In addition, the binding site of HCV polymerase accommodating the template, NTP and the complementary RNA products is extremely large. The docking programs show problems to generate an optimum number of conformations that include the co-crystallized structure.

Nevertheless, individual molecular dockings have been successfully used to obtain novel HCV NS5B inhibitors. Louise-May et al [79] used Glide to screen a customized library of 90,000 lead-like compounds against TS-II of HCV NS5B polymerase. The best scoring pose for each ligand was retained for ranking. The top 1318 compounds which scored below -7.17 were visually inspected, and 50 compounds were experimentally tested. The active compounds showed an IC_{50} between 50 and100 µM. Golub et al [80] used DOCK to screen 120,000 drug-like compounds from Otava Ltd, for new NNIs binding TS-II. A docking score cutoff value of $\leq$ -35 kcal/mol was applied and yielded 41,000 compounds. The binding interaction of known TS-II inhibitors were investigated and used as filtering criteria. Compounds that lacked the key binding interaction were removed. 984 compounds passed the filters and 59 compounds were experimentally investigated. Eight compounds exhibited IC_{50} values between 16 µM and 57 µM. These two studies demonstrate that docking helps to downsize the number of compounds to be screened effeciently. But in order to reduce the number for experimental tests, the final step of protocol relies on the knowledge of known inhibitors and visual inspection instead of considering only docking scores. Commonly, the best docking should obtain 50-60% of the actives in the top 5% of ranked compounds [81].

Water molecules can mediate protein-ligand interactions and play an important role for facilitating tight binding. Treatment of water molecules is thus a key to improve docking performance. Barreca et al [82] analyzed 40 crystal structures of HCV polymerase with PS-I inhibitors and classified the inhibitors into water-mediated and non water-mediated inhibitors. Docking with the conserved water molecules improves the pose prediction in water-mediated inhibitors but not in non water-mediated inhibitors, because water molecules displace ligand binding. The study also took protein flexibility into account. The best performing target structures for the water-mediated and non water-mediated inhibitors (up to five structures), were used for ensemble docking.

FITTED 1.5 [81] is a docking program which was developed by focusing on HCV polymerase. Implementation of protein flexibility and the consensus docking approach significantly improves the accuracy of the program to predict HCV polymerase inhibitors compared to its previous version, FITTED 1.0.

Regarding flexibility, molecular dynamics is a technique allowing to observe dynamic movement of protein and ligand. Molecular dynamics requires a high computational cost, therefore it is generally used to study only small set of protein-ligand complexes [83, 84] instead of being directly used for *in silico* screening.

Pharmacophore-based VS is fundamentally different from docking. A pharmacophore is defined by the IUPAC [85] as "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response". Pharmacophore models can be generated by two ways: ligand-based and structure-based pharmacophore modeling. Ligand-based pharmacophore modeling extracts common chemical features from a set of known ligands, whereas structure-based pharmacophore modeling uses a protein-ligand complex. Once a pharmacophore model is generated, it can be used as a query to search for potential ligands. Efficiency of pharmacophore-based VS depends on model optimization which associates available information of bioactivity. Typically, pharmacophore-based VS has a high false positive rate and a high false negative rate because a single pharmacophore query cannot cover all the protein-ligand interactions [86-88]. A combination of pharmacophore methods with docking is a strategy that has been reported to identify novel HCV polymerase inhibitors [89, 90]. In general, VS and docking of NS5B inhibitors show that active inhibitors can be identified, but usually the hits are only active in the micromolar range. Therese et al [91] employed six pharmacophore models generated from both

structure-based and ligand-based techniques to screen the Asinex database. The hit compounds were further docked using Glide and filtered out based on interactions. This yielded 10 compounds where two compounds showed inhibitory activity with IC_{50} values of 28.8 and 47.3 µM against HCV NS5B polymerase.

Docking and pharmacophore-based VS have their own strength and weakness [92, 93]. Docking provides useful information about ligand bound conformation but it has difficulties in ranking compounds according to their binding affinities [78]. To account for protein flexibility, ensemble docking might be a method to improve docking. Pharmacophore-based VS has the flexibility to adjust a tolerance radius for each pharmacophore feature. Pharmacophore-based VS is more conceptual than similarity based screening, it thus has the potential to find novel active ligands that are structurally different from a reference ligand, in other words, scaffold hopping [94]. Combining different methods which complement each other result in an improved overall performance [95, 96].

1.5 Aims of the present study

An increasing number of successful application of computational methods enables *in silico* approaches to become an integral part of the drug discovery process. This thesis describes efforts in applying computational methods towards the discovery of novel hits against HCV NS5B polymerase. The study focuses on HCV NS5B polymerase which is an enzyme critical for viral lifecycle and is a proven drug target.

There are several computational methods and tools available, but it is well known that the performance of each method varies from target to target. Furthermore limited knowledge of targeted proteins or ligands, and programs available in each work group limit the choice of the protocol. Thus it is worthwhile to evaluate and develop a protocol that is suitable for a particular target of interest in the first step of drug discovery. In addition, the limited efficacy of HCV inhibitors is due to genetic variation in the HCV genome. Various genotypes and mutations associated with the resistance to the different NNIs have been reported [65-70, 97, 98] but the mechanisms of resistance have not been fully elucidated. So it is crucial to use novel computational methods to have a better understanding of such mechanisms, and the results can be useful for the development of novel inhibitors.

Therefore, specific objective of this thesis are:

- To develop a screening protocol to identify novel HCV NS5B polymerase inhibitors.
- To discover novel inhibitors targeting the palm site of HCV NS5B polymerase.
- To understand the impact of resistance on the inhibitory activity via docking and molecular dynamics simulation.

Chapter 2 Computational and experimental methods

2.1 Protein-ligand binding affinities

Drugs function when they interact and bind to a target protein. Characterization of protein-drug interaction requires not only knowledge of the structures of the complexes, but also the free energy contribution of the interaction. The binding process can be classified as either irreversible or reversible. Irreversible binding occurs when a covalent bond is formed between drug and protein. Most drugs establish non-covalent interactions with a target protein, and this reversible interaction can be described by thermodynamic state of the complex formation (shown in Equation 1). Like any other spontaneous process, binding occurs only when it is associated with a negative Gibbs' free energy of binding (ΔG).

$$\Delta G_{binding} = G_{complex} - (G_{protein} - G_{ligand})$$

Equation 1

In Equation 1, $G_{complex}$, $G_{protein}$ and G_{ligand} are the Gibbs free energy of protein-ligand binding, protein and ligand, respectively. The free energy can be related with the chemical equilibrium as in Equation 2 [99]. Under equilibrium condition, the concentrations of the free protein *[P]*, free ligand *[L]*, and the bound ligand *[PL]* are constant. At equilibrium, the association constant K_a is defined as the ratio of the bound ligand to the free protein and protein concentration (Equation 2). The equilibrium dissociation constant K_d is the inverse of K_a.

$$P + L \leftrightarrow PL$$

$$\boldsymbol{K_a = [PL]/[P][L] = 1/K_d}$$

Equation 2

When inhibitor is added to the reaction, the total concretion of the inhibitor that gives 50% inhibition is IC_{50}. There are three types of inhibitor mechanisms: competitive, uncompetitive and noncompetitive inhibitors [100].

For competitive inhibitors:

$$K_i = \frac{IC_{50}}{1+\frac{[L]}{K_m}} \quad , K_i \cong IC_{50} \; if \; [L] \ll K_m$$

For uncompetitive inhibitors:

$$K_i = \frac{IC_{50}}{1+\frac{K_m}{[L]}} \quad , \; K_i \cong IC_{50} \; if \; [L] \gg K_m$$

For noncompetitive inhibitors:

$$K_i = IC_{50} \; when \; [L] = K_m \; or \; [L] \gg K_m \; or \; [L] \ll K_m$$

Where K_m is the concentration of the ligand at the half of the maximal rate of protein-ligand reaction at which all the protein molecules are saturated with ligands.

Equation 3

K_i is the equilibrium dissociation constant for the inhibitor. According to Equation 3, the value of IC_{50} varies depending upon how tightly the ligand binds to a protein and also upon its concentration. However, the IC_{50} value is often used instead of K_i. Because it is easier to determine and IC_{50} values correlate often to K_i [101]. The association constant K_a or the dissociation constant K_d and K_i can be related to the free energy using Equation 4. R (1.986 cal/mol/K or 8.313 J/mol/K) is the gas constant, T is the absolute temperature expressed in Kelvin, and K_a is the association constant in M^{-1} units.

$$\Delta G_{binding} = -RT \ln K_a = RT \ln K_d = RT \ln K_i \cong RT \ln IC_{50} \cong -RTpIC_{50}$$

Equation 4

Computational methods

Because traditional drug discovery is a random trial and error process, it is expensive in cost and time. Computer-aided drug design exploiting state-of-the-art technologies to predict protein-ligand binding affinity has become important for accelerating and economizing drug discovery and development. Docking is one of the popular approaches used to identify potential active compounds [102]. Comparative studies on the performance of numerous docking programs have shown that docking programs give a higher enrichment of the active compounds compare to random screening [78, 103, 104]. However, no single protocol consistently outperforms the others on different protein targets [78, 104-106]. Docking provides sufficient reproducibility of the correct binding poses in a reasonable time but remains unsatisfactory in ranking compounds according to their binding affinity [107]. For this reason, molecular mechanics (MM) and continuum solvation models Poisson-Boltzmann (PB) or generalized Born (GB) surface area

(SA) have been successfully employed as post-docking procedures to re-calculate the relative binding affinity before selecting compounds.

2.2 Molecular docking

Docking programs evaluate feasible binding geometries (often called binding poses) and predict binding affinities. Docking programs are mainly comprised of two operations: search function and scoring function. Both searching and scoring function work tightly together. Search algorithm explores the conformational space and the scoring function gives the estimated binding energy of the predicted pose. The process iterates to find the lowest optimum conformation. The scoring function therefore has to assign the best score to the correct pose of each compound. In principle, it should assign higher scores to the more potent compounds. In the current work, the term docking always refers to flexible ligands and rigid protein docking, unless other specific indication is given.

2.2.1 Search algorithm

The two most common search methods are Genetic algorithm (GA) and Monte Carlo (MC) methods [108]. GA and its modified versions are implemented in many programs such as GOLD [109] , DARWIN [110], PSI-DOCK [111] and AUTODOCK [112]. GA searches for the optimal conformation in a process similar to inheritance patterns in evolution and selection. Properties of the ligand conformation (torsion angle, rotation and translation) are assembled together as genes in chromosomes. These initial assemblies act as parents and produce offspring or poses through genetic operators such as crossover and mutation. The resulting chromosome assembly are evaluated and given a fitness ranking. The resulting offspring that passes the selection of fitness would replace the whole population and become a parent of the subsequent generation. The cycle repeats until a predefined number of generations are reached.

On the other hand, Monte Carlo randomly generates an initial pose and scores it. A new pose is generated by random conformational change, translation and rotation, and this new pose is then scored and compared with the previous pose using a metropolis criterion. The process repeats until the number of desired poses is obtained. ParDOCK [113] and Glamdock [114] are two examples of docking tools that use a Monte Carlo search. Since Monte Carlo methods randomly and independently choose the chemical space on each move, the same position can be sampled again. Modified MC with tabu algorithm can keep track of the generated positions and so can

avoid resampling. Alternately, MC simulated annealing thoroughly searches for similar positions with low energy, and it has been used to efficiently explore the conformational space of ligands [108]. Other search methods, for example particle swarm optimization in PARADOCKS [115], exhaustive search in Glide [116], have also been employed in docking programs.

2.2.2 Scoring functions

Scoring functions are mathematical methods used to predict the strength of protein-ligand interaction as a measure of binding affinity. Different scoring functions have been developed and evaluated, however no scoring function showed good performance for all of the studied target proteins [78].

Scoring functions can be grouped into three broad categories: Force Field (FF) based, empirical based and knowledge based (Table 3). A FF is a mathematical function that returns the energy of a system as a function of the conformation of the system. FF based approaches can be written in terms of potential energy (V) functions of the various structural features as shown in Equation 5 [117].

$$V = V(r) + V(\theta) + V(\phi) + V(nb) + (\text{specific terms})$$

Equation 5

The terms are bond stretching V(r), bond angle bending V(θ), bond torsion V(ϕ), and non-bonded interactions V(nb). The non-bonded interaction generally refers to van der Waals and/or Coulomb force. The specific terms are, for example, out of plane bending, electrostatic interactions and possible hydrogen bonding [117]. The solvation effects are sometimes taken into account using a distance dependent dielectric constant in the Coulombic part. Entropic energy and intra-molecular interactions are completely ignored in a pure FF based approach [118] such as DOCK. However some FF based approaches, such as in Gold-score, are combined with empirical terms to compensate for the missing interactions.

The empirical approach is derived from a set of protein-ligand complexes with known binding affinity. This type of scoring function uses empirically weighted interaction terms such as van der Waals, electrostatic, and solvation energies. The coefficients of each term are obtained using multivariate regression methods to fit a training set of protein-ligand complexes to measure binding constants[118]. Since empirical scoring functions are implicitly calculated, they are often easier to calculate than FF scoring functions[118]. Also, empirical scoring functions are

modeled from a training set, so the accuracy of these scoring functions are limited to complex structures that are similar to the ones used in the training set.

Knowledge based scoring functions are developed from statistical analysis based on observed frequencies of interaction seen in protein-ligand complexes. The distribution of the interaction frequency is translated into energies using the Boltzmann distribution. Similar to empirical scoring functions, only interactions that are part of the training set can be properly accounted for, because knowledge based scoring functions are also derived from training data.

Table 3. Types of scoring functions.

Type	Scoring function
Force field	DOCK [119] , GOLD/Gold-score [120]
Empirical	Glide [116, 121] , PLP [122], Chem-score [123]
Knowledge-based	PMF [124], GOLD/ASP [125]

2.2.3 Evaluation of docking performance

The objectives of docking are to predict the correct placement of small molecules within the protein binding site, and to rank them according to their affinity. The Root Mean Square Deviation (RMSD) is commonly used to evaluate the pose prediction. Compounds with known conformation and orientation, which are generally obtained from crystal structures, are re-docked to their target sites. The standard deviations between the predicted poses and the original conformation are then calculated. In this study, a predicted poses were considered as a correct pose if the RMSD is lower than 2.5 Å.

An enrichment study is used to assess the accuracy of virtual screening. It determines how well a docking program selects the active compounds out of a decoy set compared to random selection. Enrichment factor (EF) compares the relative number of active compounds in the specific top ranked compounds to that in total compounds (Equation 6 and Equation 7). The maximum enrichment is determined by the total number of active compounds and the total number of molecules in the database. For instance, there are 100 active compounds among the total 10,000 molecules in the database, i.e. the achievable maximum is 10,000/100 = 100. If 5% (5

compounds) of active compounds were found among the top 1% (100 compounds) of the database, then the enrichment factor would be fivefold over random (EF = 5) at the 1% of the database [126].

$$EF = \frac{L_{subset}/N_{subset}}{L_{total}/N_{total}}$$

Equation 6

$$EF_{max} = {}^{N_{total}}/_{L_{subset}}$$

Equation 7

Where

L_{subset} is the number of active compounds in the sample subset.

N_{subset} is the total number of compound (active compounds and decoys) in the subset.

L_{total} is the total number of active compounds.

N_{total} is the total number of compounds.

Receiver operating characteristic (ROC) curves is a graph of sensitivity (y-axis) versus specificity (x-axis). Similarly, plotting between the percentage of active compounds retrieved at different top ranking represents the effectiveness of a docking [127].

Both ROC curve and EF indicate on the ability of a docking program to identify active compounds from a large set of the decoys and place them at the top of the hit list, but they do not illustrate the precision in ranking compounds in relation to their binding affinity. The number of correct pairs [128] is a method that can be used to indicate precision in ranking compounds in relation to their binding affinity. To calculate the correct pairs, all pairs between a compound in rank i (C_i) and a compound in a lower rank i+ (C_{i+}) are considered. A correct pair is counted if the IC_{50} value of the compound C_i is one order of magnitude higher in comparison to IC_{50} of the compound C_{i+}. The better effectiveness of docking is thus indicated by a higher number of correct pairs.

2.3 Molecular Mechanics-Generalized Born / Surface Area (MM-GB/SA) and Molecular Mechanics-Poisson Boltzmann / Surface Area (MM-PB/SA)

In MM-GB(PB)/SA, the binding free energy between a protein and a ligand to form a complex (Equation 5) is calculated by using the energy of unphysical process as illustrated in Figure 12 and Equation 9.

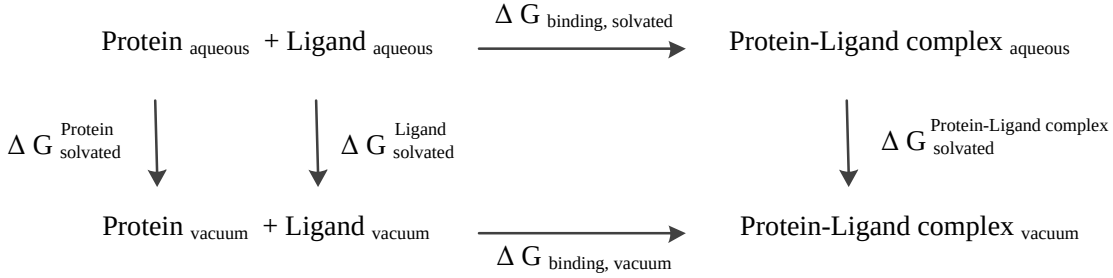


Figure 12. Thermodynamic cycle of protein-ligand binding

$$\Delta G_{binding\ ,solvated} = \Delta G_{binding,\ vacuum} + \Delta G_{solvated}^{P-L\ complex} - (\Delta G_{solvated}^{Protein} + \Delta G_{solvated}^{Ligand})$$

Equation 8

The Gibbs free energy is composed of enthalpy (ΔH) and entropy (-TΔS). The free energy change associated with each term in Equation 8 is estimated according to Equations 9-12.

$$\Delta G = \Delta H - T\Delta S \quad \textbf{Equation 9}$$

$$\Delta H = \Delta E_{MM} + \Delta G_{solvated} \quad \textbf{Equation 10}$$

$$\Delta E_{MM} = \Delta E_{internal} + \Delta E_{vdw} + \Delta E_{electrostatic} + \Delta G_{solvated} \quad \textbf{Equation 11}$$

$$\Delta G_{solvated} = \Delta G_{GB\ or\ PB} + \Delta G_{SASA} \quad \textbf{Equation 12}$$

E_{MM} is the MM energy in vacuum or gas phase. ΔE_{MM} includes $\Delta E_{internal}$ (bond, angle and dihedral energies), electrostatic ($\Delta E_{electrostatic}$) and van der Waals (ΔE_{vdw}) energies. The solvation free energy ($\Delta G_{solvated}$) consists of two components, the polar and nonpolar solvation free energies. The polar solvation energy is calculated by Generalized Born (GB) or Poisson

Boltzmann (PB) in a continuum solvent model, while the nonpolar solvation free energy is estimated by proportion of solvent accessible surface area (SASA). The conformational entropy change $-T\Delta S$ can be computed by normal mode analysis on a set of conformational snapshots taken from MD simulations. When the energy comparison is carried out on similar systems such as ligands binding to the same protein, the entropy change is assumed to be similar. Entropic contributions therefore are normally neglected. Moreover normal mode analysis calculations are computationally expensive and tend to have a large margin of error thus introducing significant uncertainties in the results. Therefore in this study, the energy neglecting the entropy term will be called as binding energy, not binding free energy.

Although MM-GB(PB)/SA is usually applied on the average of ensemble MD trajectories during MD simulation in explicit solvent, it is time consuming. Single energy minimized structure is an alternative and rapid approach that show reasonable estimate of the ligand binding free energies [129, 130], and also it is a good at discriminating the compounds that have difference of IC_{50} values greater or equivalent to 100-1000 times ($\Delta pIC_{50} \geq$ 2-3)[131]. Examples of programs that calculate MM-GB(PB)/SA are Amber [132], GROMACS [133] and Prime MM-GB/SA in the Schrödinger software.

2.4 Hybrid Quantum Mechanics and Molecular Mechanics Generalized Born Surface Area (QM/MM-GB/SA)

Hybrid QM/MM divides the system into two parts: QM and MM regions. The QM region consists of bound ligand and its neighboring protein residues in the binding site. The remainder of the system is computed at the MM level.

In the QM/MM-GB/SA approach, MM energy (ΔE_{MM} in Equation 11) is replaced by QM/MM energy. The energy of the system ($E_{QM/MM}$) is a summation of the energy of the QM subsystem (E_{QM}), the MM subsystem (E_{MM}) and the interaction energy between both subsystems ($E_{QM/MM}$) (Equation 13).

$$E_{QM/MM} = E_{QM} + E_{MM} + E_{QM/MM} \qquad \textbf{Equation 13}$$

E_{MM} is calculated from the MM atom positions using the Amber force field and parameters, whereas E_{QM} is evaluated using semi-empirical Hamiltonians RM1. The electrostatic energy between the QM and MM regions arises from the electric field of the MM region atoms and van der Waal interactions of contact border atoms.

2.5 Determining K_d using a fluorescence-based *in vitro* assay

Referring to equation 2, the concentration of the protein-ligand complex [PL] is given by:

$$[PL] = \frac{[P_0][L]}{K_d + [L]}$$

Where, [P_0] is the concentration of the total protein concentration that was added in the test.

Equation 14

In fluorescence studies, the concentration of the protein-ligand complex ([PL]) is measured by the change in fluorescence, Fl [134].

$$Fl = \frac{F_{max}[L]}{K_d + [L]}$$

Where,

Fl is the relative fluorescence intensity at a given ligand concentration.

F_{max} is the maximum fluorescence intensity at saturation of a given binding site.

L is the free ligand concentration.

Equation 15

The total amount of ligand [L_0] added, is known and the free ligand concentration is not directly measured. In practice, the K_d value can be calculated by using the following non-linear regression equation:

$$Fl = F_{RL}[RL\} = 0.5F_{RL}\{(L_0 + R_0 + K_d) - \sqrt{[(L_0 + R_0 + K_d)^2 - 4R_0L_0]}\}$$

Where, F_{RL} is the fluorescence change per unit concentration of the RL complex.

Equation 16

2.6 *In vitro* NS5B polymerase activity assay

To determine inhibitory activity of compounds, the functional activity of NS5B polymerase is curried out. If a compound can inhibit NS5B polymerase activity, the amount of newly synthesized RNA will deplete in relation to the inhibitory activity. The activity of HCV NS5B polymerase was carried out by Tobias Hoffmann and Dr. Ralpl Golbik at the Institute for

Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg [135]. Details of the procedure are described as following.

Before adding RNA, 30 nM of NS5B and the inhibitor was incubated at room temperature for 15 minutes in a mixture consisting of 4µM of 2X Assay-Buffer, 0.5 µM of 40 mM $MnCl_2$, and 0.4 µL of 100 mM DTT. After pre-incubation, 1 µL of 10 µCi [α-^{32}P]CTP, 0.4 µL of 10 mM CTP, 2 µM radiolabeled NTP-Mix (Nucleoside 5'-triphosphates : 25 mM of each ATP, UTP, and GTP), and 20 nM RNA-template were added. Distilled water was added to the mixture achieving a reaction volume of 40 µL. The mixture was then incubated at 37 degree Celsius for an hour, and it was subsequently diluted to 200 mL. 200 µL of chloroform/phenol solution was added to the mixture and centrifuged at 13,000 rpm for 5 minutes. Aqueous phase containing RNA was transferred to a fresh tube. The RNA was precipitated by adding 472 µL of 100% ethanol, 36 µL of 6M ammonium acetate, and 20 µg tRNA, and incubated for 30 minutes at -20 degree Celsius before 30 minutes of centrifugation at 13,000 rpm. The supernatant was removed and the pellet was washed with 500 µL of 70% ethanol. The residual was centrifuged again at the same speed for 5 minutes and ethanol was discarded. The pellet was dissolved with 40 µL water. Gel electrophoresis was performed with 2 µL of the solution to check the amount and integrity of the RNA. Radioactive signals were detected by autoradiography and were imaged by a Molecular Dynamics Storm 860 scanner. Signal strength relative to a control was determined by the program, ImageQuant.

Chapter 3 Optimizing docking/scoring functions for HCV NS5B polymerase inhibitors

3.1 Introduction

Virtual screening is typically grouped into structure-based and ligand-based approaches. Molecular docking is the predominant method among structure-based approaches and it is used to screen large compound databases against a specific binding site. Docking programs predict the feasible binding geometries (often called binding poses) and then score the poses according to the binding strength. Numerous comparative studies on the performance of docking programs have been published, and they have shown that docking programs often give a higher enrichment of active compounds compared to random screening [78, 103, 104]. However the docking programs remain flawed in ranking compounds according to their binding strength [78, 104, 106, 107]. To improve the performance, a thorough investigation of different docking protocols including rigid docking, flexible protein docking, and ensemble and rescoring approaches was employed in the current work. The study was carried out focusing on thumb site II (TS-II) and palm site I (PS-I) of HCV NS5B polymerase, for which numerous crystal structures of protein-ligand complexes are available.

Usually in docking, the conformation of ligands is fully allowed to move whereas proteins are usually kept rigid to reduce the computational time. The rigid protein docking approach has long been used for most structure-based approaches and was shown to be a successful complement to high-throughput screening. Nonetheless, the approach has its own problems and limitations. One of which is the neglection of protein flexibility. All proteins have inherent adaptation of conformation relevant to their function, so protein flexibility is required to correctly dock ligands [136-138]. Several approaches have been proposed to incorporate protein flexibility in docking, but current computational facilities allow only a limited protein flexibility like in ensemble docking, soft docking [139], side chain flexibility, or induced fit docking [140]. Among the different approaches, ensemble docking is one of the most popular and often outperforms rigid protein docking [141, 142].

Another problem of docking is the performance of the used scoring function. Rescoring is an approach to handle this problem. Molecular Mechanics Generalized Born-Surface Area (MM-GB/SA) and Quantum Mechanics-Molecular Mechanics (QM/MM) are commonly used for

accurately reproducing relative binding energies at least for molecules from congeneric series. Typically the estimated binding free energy is calculated by averaging the values over multiple molecular dynamic snapshots. Since these approaches require expensive computation to run molecular dynamic simulation, it has been proposed that the energy calculated from a single snapshot after minimization is often adequate and suitable for virtual screening purposes [131, 143-145].

3.2 Cross-docking study

Three different docking programs, (Gold 4.1[120], Glide [116] and ParaDockS [115]), with six scoring functions (Gold-score, Chem-score, the Astex Statistical Potential (ASP) in GOLD; p-score and Potential of Mean Force (PMF) in PARADOCKS and Glide standard precision docking (SP)) were evaluated in a cross-docking. Cross-docking refers to docking a ligand into protein structures originally bound with another ligand. The docking program should be able to reproduce the native poses observed in the crystal structures. Moreover comparison of the docking results obtained with different protein structures allows to assess how variations in the protein structures affect the effectiveness of rigid-protein docking. The results demonstrated that there is significant variability in the performance of scoring functions/docking programs based on the used protein structure. As can be seen in the 22 collected crystal structures of the TS-II dataset (Table A1, Figure A1-Appendix) and the 29 crystal structures of the PS-I dataset (Table A2, Figure A2-Appendix), the docking accuracy was found to be variable (Table A3-A5-Appendix). Besides the performance of each scoring function/docking program, the conformation of the binding site in each protein significantly contributes to the accuracy of binding prediction. The TS-II showed several flexible side-chains due to induced-fit adaptation and contained different amino acids according to the genotype (Figure A3-Appendix). Among six docking/scoring functions explored on PS-II pocket, Chem-score and ASP-score performed better compared to P-score and PMF-score (in PARADOCKS). The results of the TS-II dataset with the best performance gave 16 correct poses out of 22, and they were obtained from crystal structure PDB: 3CJ4 and Chem-score as scoring function. On the other hand, the PS-I site shows less mutations and preserves a relatively rigid structure across all of the studied protein structures. In the PS-I dataset, the results obtained from different scoring functions or protein structures were not significantly different compared to the results of the TS-II dataset. Gold-score docking and using the structure PDB: 3GNV gave the best prediction in the PS-I dataset (20 correct poses out of 29 co-crystallized ligands). Furthermore, it is noticed that all

docking/scoring functions failed to predict a correct pose for at least one of the co-crystallized ligands.

The obtained results indicate the limitation of single rigid protein docking in case of the HCV NS5B polymerase. So taking protein flexibility into account should improve the docking accuracy. The scoring functions (Gold-score, Chem-score, ASP-score and Glide SP) along with structural conformations that performed best in the cross-docking study were then further tested on flexible side-chain and ensemble docking. PMF and p-score in PARADOCKS were ignored for further studies because their performance did not surpass other scoring functions in the cross-docking study. As the performance of ensemble docking would depend on the selected conformations in the ensemble [146], different sets of ensemble protein structures were explored. Moreover, rescoring with MM-GB/SA, QM/MM-GB/SA and further scoring function were validated with respect to experimental data.

3.3 Evaluation of ensemble docking, flexible side-chain docking and rigid protein docking

3.3.1 Methods and datasets

Gold-score, Chem-score and ASP-score were used for rigid protein docking and ensemble docking. Flexible side chain docking was carried out using Glide where the hydroxyl groups of residue in the binding pocket were allowed to be flexible. GLIDE can perform ensemble docking by docking a ligand sequentially into all multiple rigid receptor conformations and post-processing the single protein structure results. The docking time is thus proportional to the number of ensemble protein structures. GOLD is developed time-efficiently search algorithm for ensemble docking. Hence in this study, only ensemble docking using GOLD was evaluated. In rigid protein docking and flexible side chain docking, two protein structures that gave the best docking prediction in the cross-docking study: PDB: 3GNV / 3HWK for PS-I dataset and PDB: 3CJ4 / 3HAI for TS-II dataset were selected. These best performing proteins were used to evaluate the effect of protein conformation on the performance of flexible side chain docking, and to compare the performance of individual protein docking with ensemble docking. Dataset of TS-II comprised 22 co-crystallized inhibitors (Table A1, Figure A1-Appendix). The dataset of PS-I contained 35 co-crystallized inhibitors (6 compounds, Figure A4-Appendix, were

additionally added to the 29 compounds used in the cross-docking study, Table A2 and Figure A2-Appendix)).

Ensemble docking was evaluated with three different conformational ensemble groups. Each ensemble group consisted of three protein structures. Three structures are reported to be an adequate number to account for the flexibility of this target and to improve the docking accuracy [146]. Adding more conformations could lead to worse performance [146, 147] as a result of increasing potential artifacts. Ensemble group I (Ensemble–I) consisted of three well-performing conformers from the cross-docking studies (Table A3, A4 and A5-Appendix). Ensemble group II (Ensemble-II) consisted of three crystal structures which show high structure quality (Table 4 and Table 5). Ensemble group I and II were prepared and tested for both PS-I and TS-II dataset. Referring to the TS-II dataset in the cross-docking, there were three clustered conformations. If a representative conformation in each cluster is not included in ensemble docking, docking would have missed potential inhibitors. We therefore investigated an extra ensemble group for the TS-II dataset: Ensemble group III (Ensemble–III). The Ensemble-III consisted of two representative structures from two clustered conformations and lacked one relevant flexible protein conformation.

Ensemble proteins in TS-II dataset

Ensemble-I (well performing conformers): 3CJ4, 1YVX, 2HAI
Ensemble-II (high quality structures): 3CJ2, 2D3Z, 2HAI
Ensemble-III (missing a representative flexible conformer): 3FRZ, 2D3Z, 2HAI

Ensemble proteins in PS-I dataset

Ensemble-I (well performing conformers): 3GNV, 1Z4U, 3CSO
Ensemble-II (high quality structures): 2GIQ, 3HHK, 3HKW

Table 4. Quality of crystal structures in TS-II dataset.

PDB ID	1YVX	2D3Z	2HAI	3CJ2	3CJ4	3FRZ
Resolution (Å)	2	1.8	1.58	1.75	2.07	1.86

Table 5. Quality of crystal structures in PS-I dataset.

PDB ID	1Z4U	2GIQ	3CSO	3GNV	3HHK	3HKW
Resolution (Å)	2.8	1.65	2.71	2.75	1.7	1.55

3.3.2 Results and discussion

Among the ensemble groups, the ensemble docking with ensemble group I (well performing-conformers) and II (high quality structures) outperformed the ensemble group-III (missing a representative flexible conformer) (Figure 13). This supported the observation that the performance of ensemble docking relies on the selection of appropriate protein structures. The success rate of ensemble group I and group II were not significantly different in the TS-II dataset (Figure 13). However, the result from PS-I dataset (Figure 14) showed that using ensemble group I performed better than ensemble group II. In the TS-II dataset, a similar performance of using the ensemble group I and group II was observed due to the fact that both ensemble groups comprised the same protein conformations even though they contained different crystal structures. For instance, the crystal structure 3CJ4 of ensemble group I had the same binding-site conformation as the crystal structure 3CJ2 of ensemble group II. The results implied that the protein structures that performed best in a single protein structure docking should be selected to optimize the accuracy of ensemble docking, because it would ensure that the conformational variability is incorporated in the docking run. The performances of ensemble docking using ASP-score or Chem-score were found to be similar and both scoring functions performed better than Gold-score.

Comparison of rigid protein docking, docking with flexible side-chains, and ensemble docking showed that in all cases of the TS-II dataset (Figure 13) ensemble group I and group II were better than group III. The results of ensemble group III were less accurate than rigid protein docking and docking with flexible side-chains. In contrary, (Figure 14), docking with flexible side-chains in the PS-I dataset was found to be the best. Ensemble group I performed better than rigid protein docking but not better than docking with flexible side-chains. According to section 3.2 (cross-docking study), the TS-II shows a higher conformational variation than the PS-I. Thus using a suitable set of multiple protein conformations in ensemble docking should improve the efficiency in docking the TS-II dataset. Using one rigid protein structure would fail to accurately predict the binding mode of some inhibitors. On the other hand, if the binding site does not adopt

conformational changes upon binding as in the PS-I, docking with flexible side-chains would be a suitable approach. Adding more protein structures might result in increasing the number of false positives.

From the above analysis, we analyzed the results with respect to the top ranked pose. In the next step we tested the top-10 ranked poses of each compound and selected the pose that was nearest to its native crystal structure (i.e. lowest RMSD as 'best pose'). Noticeably, the number of correct poses obtained from the best poses of the ensemble docking group I with Chem-score in the PS-I dataset (69.5%) was significantly higher than the best results obtained from the top ranked poses in Glide with flexible hydroxyl group (55.2%). This suggests that one way to improve the pose prediction is by improving the accuracy of the scoring. Therefore further post-docking processes were examined by using MM-GB/SA, QM/MM and further scoring functions to rescore all calculated docking poses.

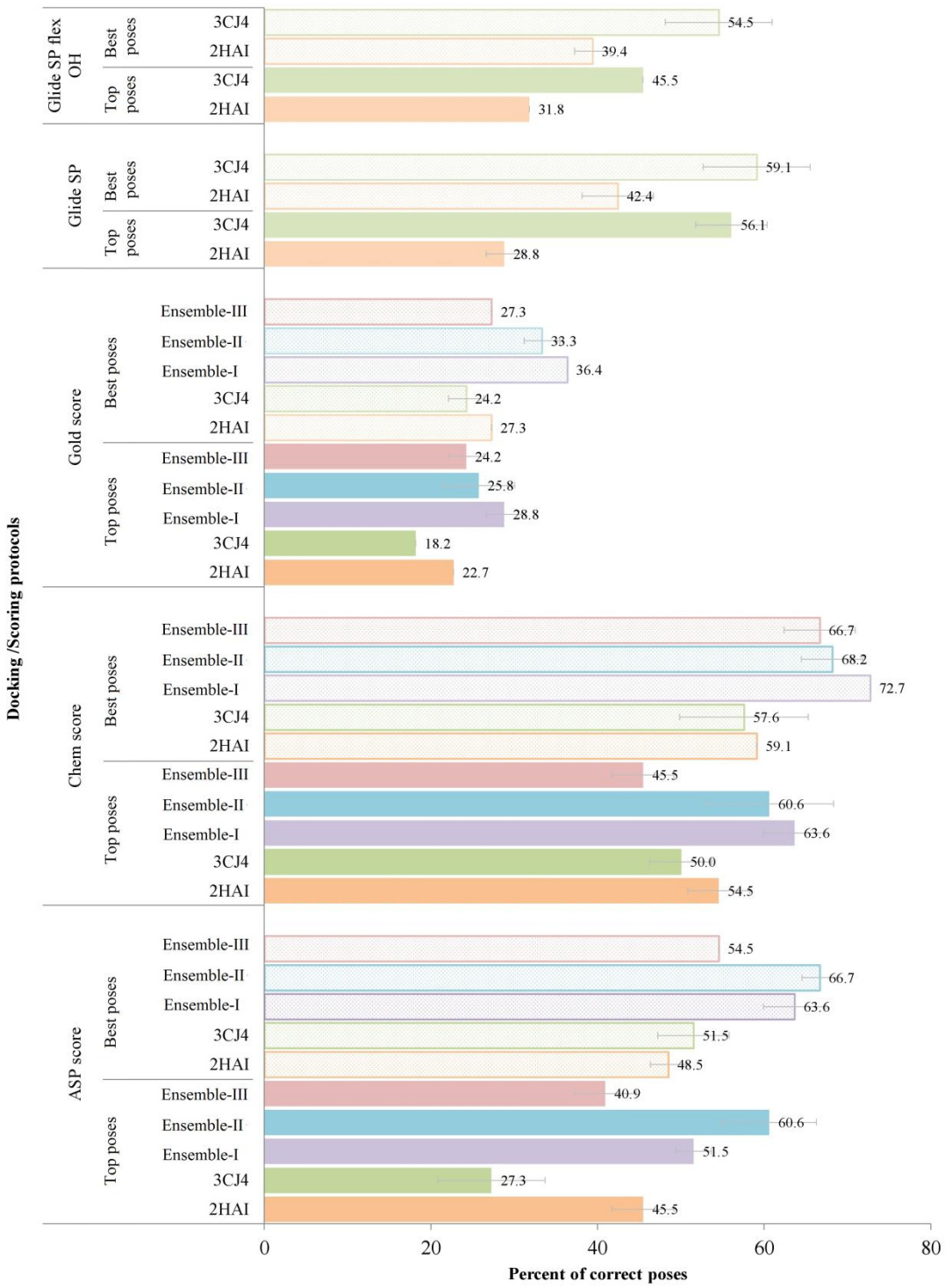


Figure 13. Performance of different docking/scoring functions on the TS-II dataset (22 crystal structures). Performance was quantified by the correctness of the top ranked pose (denoted as 'top pose') and the pose that was nearest to its native crystal structure within 2.5 Å (i.e. lowest RMSD as 'best pose'). The average number of correct poses and standard deviation bars were calculated from three individual docking runs.

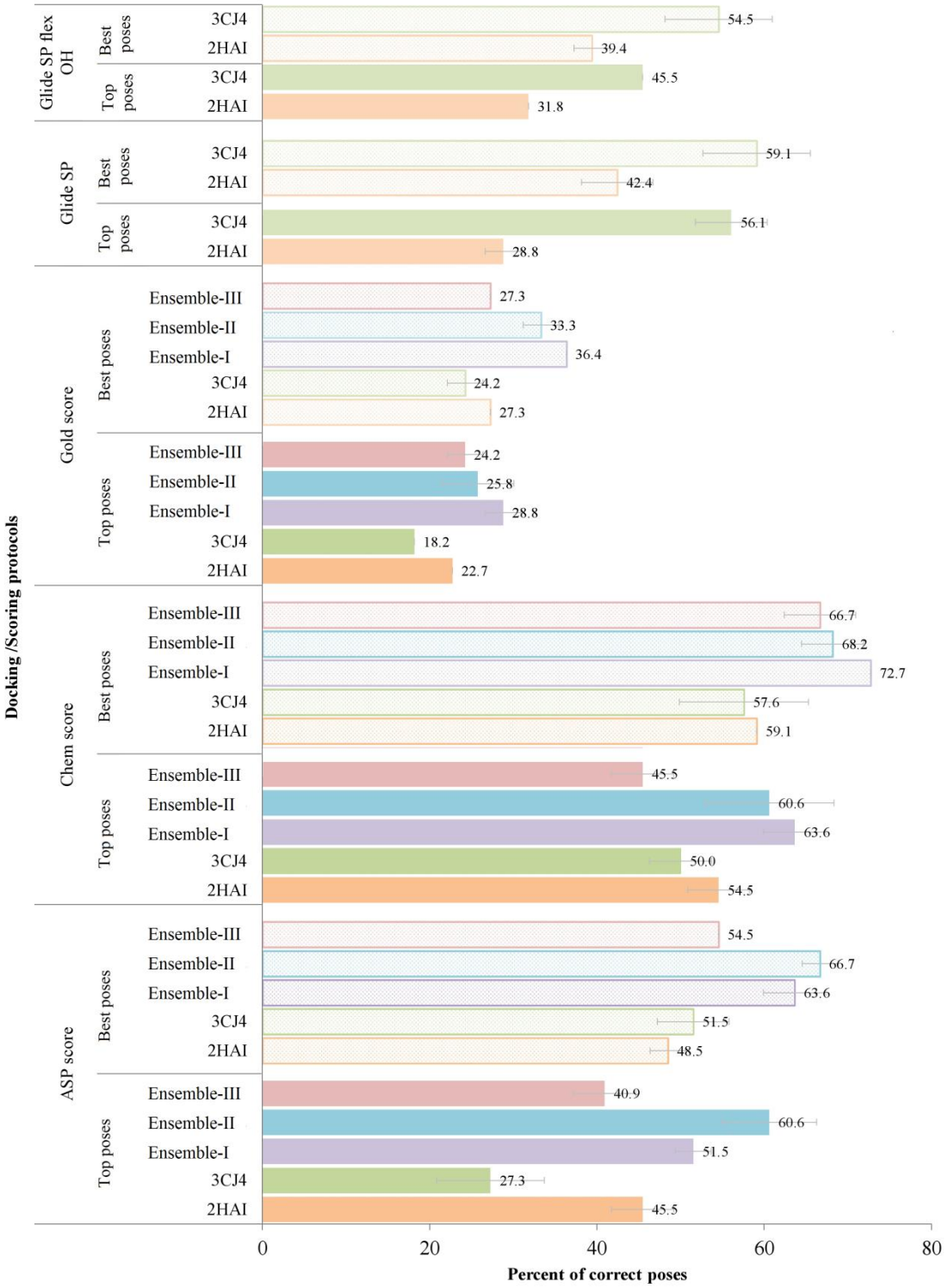


Figure 14. Prediction accuracy of different docking/scoring functions on the PS-I dataset (35 crystal structures). Performance was quantified by the correctness of the top ranked pose (denoted as 'top pose') and the pose that was nearest to its native crystal structure within 2.5 Å (i.e. lowest RMSD as 'best pose'). The average number of correct poses and standard deviation bars were calculated from three individual docking runs.

3.4 Rescoring of docking poses for the PS-I inhibitors

We showed that ensemble docking is helpful for enhancing the docking accuracy for flexible binding sites as the TS-II region. However it had no benefit for the more conserved and rigid PS-I. Therefore we aimed to improve the docking accuracy by using a rescoring method including MM-GB/SA, QM/MM-GB/SA, and other scoring functions [148-150]. Chem-score and Glide SP were used for generating the docking poses since they generated superior results compared to other setups in the previous step. Ten poses per ligand were then passed to rescoring.

3.4.1 Methods and datasets

Molecular Mechanics Generalized Born Surface Area (MM-GB/SA) and Quantum Mechanics- Molecular Mechanics Generalized Born Surface Area (QM/MM-GB/SA) methods

All MM-GB/SA and QM/MM-GB/SA were performed in AMBER 11 [151]. Ligand preparation was carried out by using the GAFF force field and AM1-BCC charge model, and ff99SB force field was applied to protein structures. The starting structure was neutralized with counter ions and solvated in a truncated octahedral box of TIP3P water with 13 Å of water around every atom of the complex. All simulations were performed under periodic boundary conditions. The solvated complex was minimized in two stages in order to remove possible bad contacts. In the first stage, 1000 steps of steepest descent minimization followed by 2000 steps of conjugate gradient minimization were applied with fixed protein. In the second stage, the entire system was minimized without restraints by 1000 steps of steepest descent minimization and conjugate gradient minimization. Then water molecules and counter ions were removed, and the free energy was calculated. MM-GB/SA and QM/MM-GB/SA calculation were performed based on a single minimized structure.

In the hybrid QM/MM-GB/SA method, the ligand and selected residues 4.5 Å around the ligand were defined as the QM region. The remainder of the system was calculated at the MM level. The entropy (ΔS) term was estimated on the number of rotatable bonds of the ligands.

Dataset for ranking evaluation

To evaluate the prediction accuracy, a modified PS-I dataset where the ligands with unknown IC_{50} were replaced by analogs of other co-crystallized ligands with reported inhibitory activity,

was used. The modified dataset contained 45 compounds in total including 27 co-crystallized structures and 18 analogs (Table A6 and Figure A5-Appendix) known as PS-I inhibitors.

Dataset for enrichment study

Decoys

Decoys were added for the enrichment study and were obtained from the Directory of Useful Decoys (DUD). The decoys were selected based on the similarity of their 1D properties to the active compounds [152]. 859 decoys were selected from the DUD collection of decoys of HIV-reverse transcriptase inhibition. Moreover, 106 inactives or weakly actives (IC_{50} values greater than 100 µM for HCV NS5B polymerase) were compiled from the literature. Totally 965 decoys were thus used for the current study.

Active compounds

The 35 PS-I inhibitors were included in the dataset. (Figure A2 and Figure A4-Appendix). This dataset is the same as the palm dataset used in the ensemble docking study (chapter 3.3.1).

3.4.2 Results and discussion

Rescoring by QM/MM-GB/SA (Table 6) performed worse than pure docking. This might be due to the QM/MM boundary settings. The quantum region was set by the ligand and extended to 4.5 Å around the ligand. Each quantum region setting was kept fixed throughout the simulation. Extending the boundary further than this was problematic due to its computational expense. Putting the boundary closes to the ligand might not represent the chemical realism. The results shown here were derived from setting the quantum region at 4.5 Å around the average structures of the superposed ligands, i.e. the amino acid residues defined as QM region were kept the same for each ligand. Since the ligands show different sizes (304 - 613 Da), each system might need a different proper QM/MM partitioning. However, adjusting the setting in each system was not feasible for screening a large compound database. In the study by Jerry et al [153] QM/MM-GB/SA was successful in identifying three correct poses from 20 docked poses of one HCV NS5B complex structure, hence QM/MM-GB/SA might be appropriate for pose prediction. Alternatively, it might be suitable for the comparison of congeneric compounds that have the same chemical scaffold. Another reason for the missing correlation between predicted energy and experimental data might be the entropy. The estimated entropy from the rotatable bonds

could not predict all entropic effects as indicated by the positive values of the estimated binding free energy.

The results (Table 6) showed that rescoring by MM-GB/SA and by Glide SP slightly improved the pose prediction in ensemble docking. However, Glide SP allowing flexible OH groups' rotation was still found to be the best. MM-GB/SA improved the number of correct poses especially in the obtained docking poses from the rigid protein docking using Chem-score. No significant improvement was observed when the docked poses generated from side chain flexibility docking in Glide SP were rescored. Rescoring docked poses of Chem-score docking by Glide SP also showed an improvement but not in the case of rescoring the docked poses of Glide SP by Chem-score. MM-GB/SA seemed to provide a good correlation between scoring values and biological activity as the number of correct pairs after rescoring by MM-GB/SA increased to 456 compared to 309 obtained by side chain flexibility docking in Glide SP. (The number of correct pairs represents the effectiveness of the docking score to rank the compounds. A correct pair is counted when biological activity of the higher ranked compound is greater than of the lower ranked compound at least one magnitude.) However, when only the co-crystallized ligands were considered, Table 7 shows that the number of correct pairs obtained by using MM-GB/SA (76 correct pairs) was lower than using Glide SP (140 correct pairs). Moreover there was a low correlation between predicted binding free energy and experimental data (Figure 15). This suggests that MM-GB/SA is better in selecting poses for this dataset than correctly ranking the compounds. This might be due to the fact that scoring functions are optimized to be general whereas MM-GB/SA is sensitive to differences in chemical structure especially when the entropy is ignored. In addition, the entropy was calculated using the quasi-harmonic entropy approximation in Amber 12. However, it did not improve the correlation with the observed experimental data of the crystal structures compared to the enthalpies as shown in Figure 15. Therefore, for the rest of the work, the binding energy is calculated by means of MM-GB/SA and the term 'binding energy' will refer to the enthalpy omitting the entropy.

The enrichment study applied on 35 active crystal structures and 965 decoys was employed for further evaluation. The ROC curve (Figure 16), i.e. the plotting of the false positive rate (FPR) against the true positive rate (TPR), showed that overall both pure docking by Glide SP with flexible hydroxyl group and MM-GB/SA resulted in a good performance. In the first one percent of the screening (Figure 16 and Table 8), docking by Glide SP with flexible hydroxyl group could retrieve more active compounds than MM-GB/SA; EF1% of Glide SP was 33.33 whereas

it was 23.81 in case of MM-GB/SA. After the first one percent, rescoring by MM-GB/SA gave more active compounds than Glide SP. Considering a typical virtual screening situation where one needs to distinguish a small number of potentially active compounds from a library of several thousand compounds, top 1% to 5% ranked compounds are commonly selected from virtual screening. Thus using Glide SP with flexible hydroxyl group was shown to perform better than MM-GB/SA. Additionally, Glide SP is much less computational time compared to MM-GB/SA rescoring.

Table 6. Performance after rescoring by MM-GB/SA, QM/MM-GB/SA and other scoring functions.

Docking / Scoring protocols		**Rescoring by**			**Best poses in 10 runs**
		MM-GB/SA	**QM/MM-GB/SA**	**Chem-score or SP**	
Rigid docking : (Chem-score)					
No. of correct poses (Max. = 27) [a*]	7	16	1	16	18
Percent of correct poses	25.93	59.26	3.70	59.26	66.67
Average RMSD (Å)	3.99	2.64	6.09	2.84	2.26
SD RMSD [b*] (Å)	1.88	1.74	1.43	1.70	1.45
Correct pairs (Max = 638) [c*]	283	440	137	396	249
Ensemble : (Chem-score)					
No. of correct poses (Max. = 27) [a*]	12	14	3	14	19
Percent of correct poses	44.44	51.85	11.11	51.85	70.37
Average RMSD (Å)	3.51	2.55	5.58	2.92	2.12
SD RMSD [b*] (Å)	2.24	1.65	1.95	2.20	1.50
Correct pairs (Max = 638) [c*]	304	374	143	405	299
Glide SP flexible OH					
No. of correct poses (Max. = 27) [a*]	17	18	9	18	20
Percent of correct poses	62.96	66.67	33.33	66.67	74.07
Average RMSD (Å)	2.38	2.27	3.59	2.41	1.91
SD RMSD [b*] (Å)	1.58	1.44	1.79	1.34	1.22
Correct pairs (Max = 638) [c*]	309	456	202	305	448

[a*] The number of correct poses were considered only for the 27 co-crystallized ligands.
[b*] SD RMSD stands for standard deviation of root mean square deviation
[c*] There are 638 total pairs calculated for 45 compounds. Calculation of correct pairs is described in chapter 2.2.3 Evaluation of docking performance.

Table 7. Correct pairs of 45 compounds in enrichment study (27 co-crystallized ligands and 18 analogs) ranked by Chem-score, Glide SP, MM-GB/SA and QM/MM-GB/SA. Calculation of correct pairs is described in chapter 2.2.3 Evaluation of docking performance.

Compounds	Total pairs	The number of correct pairs			
		Chem-score	Glide SP	MM-GB/SA	QM/MM-GB/SA
27 co-crystallized ligands	168	128	140	76	43
27 co-crystallized ligands and 18 analogs	638	446	481	271	147

Table 8. Enrichment factor of rescoring of docked poses of 35 co-crystallized actives. The binding energy of each pose was scored by Glide SP with flexible hydroxyl groups and MM/GB-SA rescoring.

Rescoring method	Enrichment factor (EF) at		
	1%	3%	5%
Docking poses from Glide SP	33.33	20.63	15.24
Glide SP rescoring of co-crystallized ligands	42.86	28.57	21.90
MM-GB/SA rescoring of docking poses	23.81	22.22	20.00
MM-GB/SA rescoring of co-crystallized ligands	23.81	22.22	20.95
	EF max = 47.62		

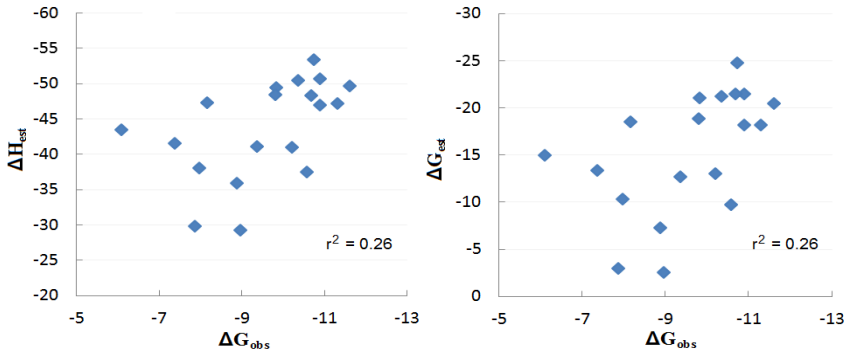


Figure 15. Correlation plot between observed binding free energy ΔG_{obs} ($RT\ln IC_{50}$) of 20 co-crystallized inhibitors and predicted enthalpy ΔH_{est} (Left) and predicted binding free energy ΔG_{est} (Right). These 20 co-crystallized inhibitors are subset of 45 compounds in enrichment study (27 co-crystallized ligands and 18 analogs). 7 co-crystallized inhibitors were excluded because of problems calculating the quasi-harmonic entropy. The energy unit is kcal/mol. Predicted energies were calculated by using MM-GB/SA (igb=8) in Amber 12 averaged over 100 snapshots from the last 2ns of MD simulations Entropy was calculated by using quasi-harmonic entropy approximation.

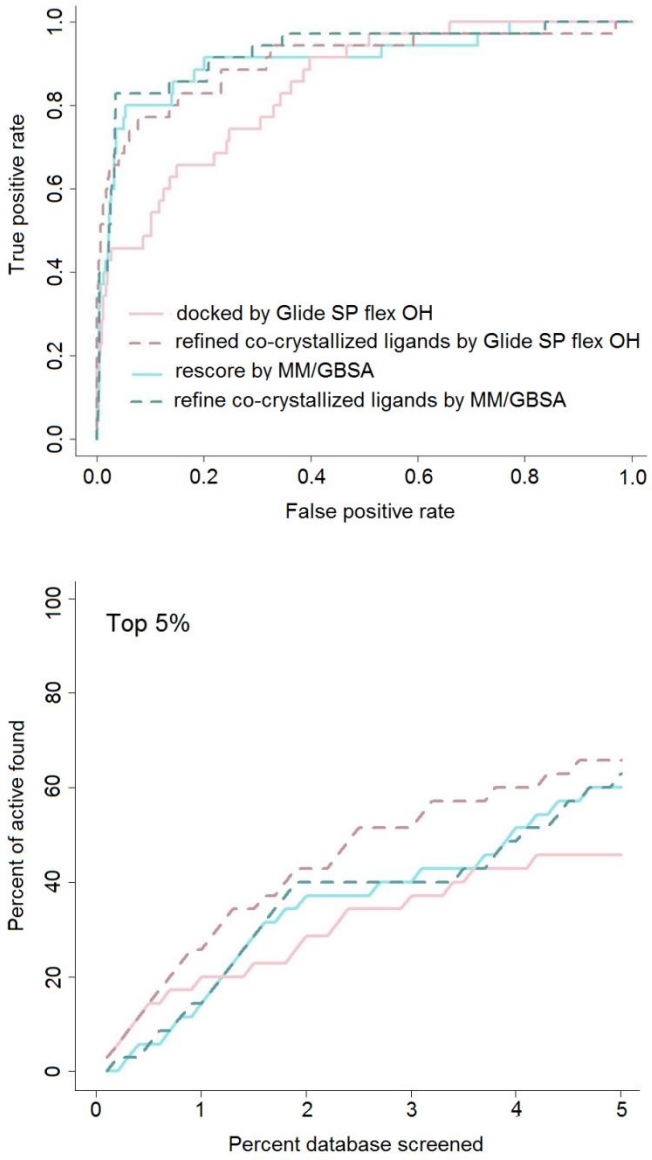


Figure 16. Comparative performance of Glide SP scoring function and rescoring docking poses by MM-GB/SA: ROC plot (upper) and percent of actives found at the top 5 % of the screening (lower). The dataset consists of 35 PS-I inhibitors and 859 decoys.

In summary, an evaluation of rigid protein docking, flexible side chain docking and ensemble docking for HCV NS5B polymerase inhibitors was carried out on two allosteric binding sites; TS- II and PS-I with respect to their ability to reproduce the know protein-ligand binding poses and to correctly predict the binding affinity. The results showed that even with the same protein structure, the most suitable protocol for each binding site was different depending on the properties of the pocket. Multiple docking or ensemble docking using Chem-score in Gold

provided the best accuracy for pose prediction into a flexible target site. However, in the rather rigid PS-I, the proper selection of a single protein docking protocol, Glide SP, performed better than ensemble docking. Moreover, selecting the appropriate protein structure was important for the docking performance of both rigid protein and ensemble docking. Random selection of protein conformers in ensemble docking can lead to worse result than single protein docking. The results suggested that ensemble proteins should be selected from well-performing conformers in single protein docking and they should cover a minimum of representative conformations.

Rescoring by MM-GB/SA improved pose prediction but it did not significantly outperforms solo Glide SP docking with flexible hydroxyl groups in the enrichment study. Besides, MM-GB/SA showed less benefit regarding compound ranking and the computational method is time consuming. In this study, only the PS-I dataset was used to evaluate the rescoring approach because neglecting entropy in binding energy calculation may not be suitable for the flexible TS-II. The entropy mainly reflects two contributions: changes in solvation entropy and changes in conformational entropy. Binding free energy calculation in implicit solvent simulation only takes conformational entropy into account. The conformational entropy is decomposed into three parts, the translational, the rotational and the vibrational entropies. Normally, the entropy can be ignored if the compounds are similar and bind to the same pocket without a significant conformational change. Thus the relative binding energy comparison of the compounds having similar entropy should not be affected. However, in case of TS-II, it is a solvent exposed binding site and a conformational rearrangement occurs upon binding of the inhibitors. Thus, the conformational entropy should not be neglected, and solvent entropy effects might play a key role.

Chapter 4 Optimizing virtual screening protocols for HCV NS5B polymerase inhibitors

4.1 Introduction

In the previous chapter, the results showed that docking gave low hit rates and rescoring by MM-GB/SA, QM/MM or other scoring functions did not significantly improve the results compared to normal docking using Glide SP. In this chapter, an additional filtering step following the docking (post-docking) was therefore assessed to determine whether it could help to enhance the hit rate. The approaches used here include Random Forest (RF) classification, Structural Interaction Fingerprint (SIFt) [154] and a customized docking scheme to filter false positives.

Random forest (RF) is a learning method based on an ensembles of trees [155]. The RF was explored in this work to develop classification models for classifying palm site I (PS-I) inhibitors. 3D molecular descriptors were used to model the relationships between molecular structures, physicochemical properties and biological activities.

Structural Interaction Fingerprint (SIFt) is a one dimensional binary string translated from 3D structural binding information of a protein-ligand complex [154]. This structural interaction profile is used to analyze docking poses by comparing the reference crystal structure's SIFt with the SIFt of the docked poses using Tanimoto coefficient. Instead of a one to one comparison between two molecules, the common interaction patterns from 29 crystal structures were derived and used as references in the current work.

As a potential drug candidate, a compound must bind its target site with high affinity and specificity. In other words, if a compound is docked independently into two binding sites, where one is the target site and the other one a dummy site, the docking program should assign the compound a favorable score for its target binding site and a low score for the dummy binding site. Any compound that possesses high scores in both binding sites should be excluded. (From now on, this approach will be referred to as '*two sites docking*'.)

Three post-docking approaches: RF, SIFt and two sites docking, as well as combinations of different methods were evaluated. The successful methods were further used to identify novel HCV NS5B inhibitors.

4.2 Materials and methods

4.2.1 Docking

The compounds were prepared using the wash function and minimized with the MMFF94x force field in MOE. The protein structure PDB ID: 3GNV [156] was prepared by the protein preparation wizard in MOE. Molecular docking was carried out using Glide SP with flexible hydroxyl groups [116]. The number of poses per ligand included for post-docking minimization was set to 10 while the other parameters were kept at their default values.

4.2.2 Random Forest (RF)

RF is an ensemble of randomized decision trees developed by Breiman [155]. In comparison to other classification methods such as support vector machine (SVM) and neural network, RF yields more comparable and better results [155, 157]. Although SVM has been shown to yield high accuracy in classifying HCV NS5B polymerase inhibitors into active and weakly active palm site inhibitors [158], it has not been studied in the context of virtual screening. RF has two main advantages over SVM. Whereas SVM requires a pre-selection of relevant descriptors, RF does not as the selection is intrinsic to its algorithm [157]. Moreover, RF is a multi-class classification so there is no need to build multiple binary classifiers as needed for SVM. This feature is particularly suitable for HCV NS5B inhibitors, which have more than one binding site and categorized activities.

RF is an ensemble learning method where every tree in the forest is built using random subsets of samples and variables [155]. Each tree is grown by using two-thirds of the training data, which is randomly drawn with replacements. The remaining samples are the out of bag (OOB) set, which calibrates the performance of each tree similar to cross-validation. At each node, the number of variables (Mtry) is randomly chosen to give the best split from among those variables. The Gini index is used as the splitting criterion for each tree. The tree grows until the node is pure and without pruning. Classification of new data is determined by majority votes of the tree in the forest. The error rate is the proportion of times that a case in the OOB set is wrongly classified.

Generally, there are two tuning parameters: the number of trees to grow (Ntree) and the number of variations to select per node (Mtry) that can be optimized in RF. RF prediction follows a

majority vote of the terminal nodes of each tree. The Ntree value therefore has to be sufficiently large to create diversity and to have the desired predictive ability. On the other hand, the choice of the number for Mtry controls the accuracy of individual trees and the diversity of individual trees. Increasing Mtry values is appropriate if the data contains a lot of variables that are weakly predictive. A low value of Mtry is advisable when most variables are highly related to the outcome. In our preliminary study, the Ntree values were varied between 200 and 10,000 in intervals of 200. The results showed that there was no significant change in class errors in all models when the Ntree value was increased. Increasing Ntree values can enable prediction stability but at the cost of computational time. In this study, as the dataset was small, a large Ntree value was chosen (10000). The optimal Mtry values were explored from $\sqrt{v}$, $2\sqrt{v}$, $3\sqrt{v}$, … , $12\sqrt{v}$ and v where v is the total number of variables. The results were similar when varying Mtry. Thus the default Mtry, i.e. the square root of the total number of variables, was selected. Besides tuning these parameters, imbalanced data can result in poor predictive performance of the minority class. Using stratified bootstrap, i.e. a sample with replacement from within each class, is an approach to reducing bias used in RF. However, this cannot solve the problem entirely [159]. Two methods are applied when dealing with imbalanced data sets. One is class weighting; however, this option was not available in the R package. Another option, which was used in this study, is the number of samplings drawn to grow each tree from each class. It was equal to the number of the smallest class to avoid imbalanced sampling between each class.

The randomForest 4.6-2 package implemented in R was used in this study [160]. The run time on an AMD Phenom ™ II x6 1090T GHz and 12 GB memory ranged from one to five minutes.

i) Random forest models

Two RF models were built and evaluated. Model-1 was designed to classify *'binding'* out of *'non-binding or decoys'* while model-2 was aimed to classify compounds into potentially *'potent'* or *'weakly'* active compounds.

ii) Evaluation of the prediction performance

A fivefold cross validation was used to evaluate the actual prediction capability of the RF classifier. The datasets were randomly divided into five equally large subsets, four of which were used for training and the remaining one for evaluating the model. Then the process was iterated by rotating the dataset assignment until each set had been used for both training and testing. For

each iteration, the classification capability was assessed by the prediction's accuracy, sensitivity and specificity.

Sensitivity (or recall) = TP / (TP + FN)

Specificity = TN / (TN + FP)

Precision = TP / (TP + FP)

Where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives and FP is the number of false positives.

iii) Molecular descriptors

3D molecular descriptors were calculated using the Molecular Operating Environment suite (MOE) [161] and PaDEL [162]. Descriptors that had null values across the majority of the samples were eliminated. The descriptors related to chiral centers were excluded since this information is generally not specified in a chemical library. A total of 38 descriptors from MOE and 115 descriptors from PaDEL were used in this study (Table A7 and Table A8-Appendix).

4.2.3 Structural Interaction Fingerprint (SIFt)

Structural Interaction Fingerprint (SIFt) [154] is a useful method for representing and analyzing structurally characterized protein-ligand interactions. SIFts are simply one-dimensional fingerprints that encode the specific interactions that a ligand has with binding site residues. The Interaction Fingerprints were generated by the python script (interaction_fingerprints.py) that is implemented in the maestro program. This SIFt is based on a 9-digit binary interaction pattern that describes physical ligand–protein interactions including H-bond donor, H-bond acceptor, hydrophobic, aromatic, charge, polar, side chain, and backbone. A schematic workflow of SIFt is shown in Figure 17. First, 29 complex crystal structures were clustered into six groups based on MACC key fingerprints of co-crystallized ligands. The cutoff criterion was set to 70% similarity. Then a profile SIFt of these crystal structures was generated to define the conserved interactions of each cluster. These conserved interaction fingerprints were then used as the reference protein-ligand interactions for calculating Tanimoto coefficients (Tc) against each query. The highest Tc for each query was further used for ranking and filtering. To avoid a size dependence of similarity coefficients [163], only the bits of the same 21 amino acids at the binding site were used to calculate Tc. SIFt was evaluated in two ways: (i) ranking by Tc and (ii) filtering by 0.5 Tc and then ranking by Glide SP score.

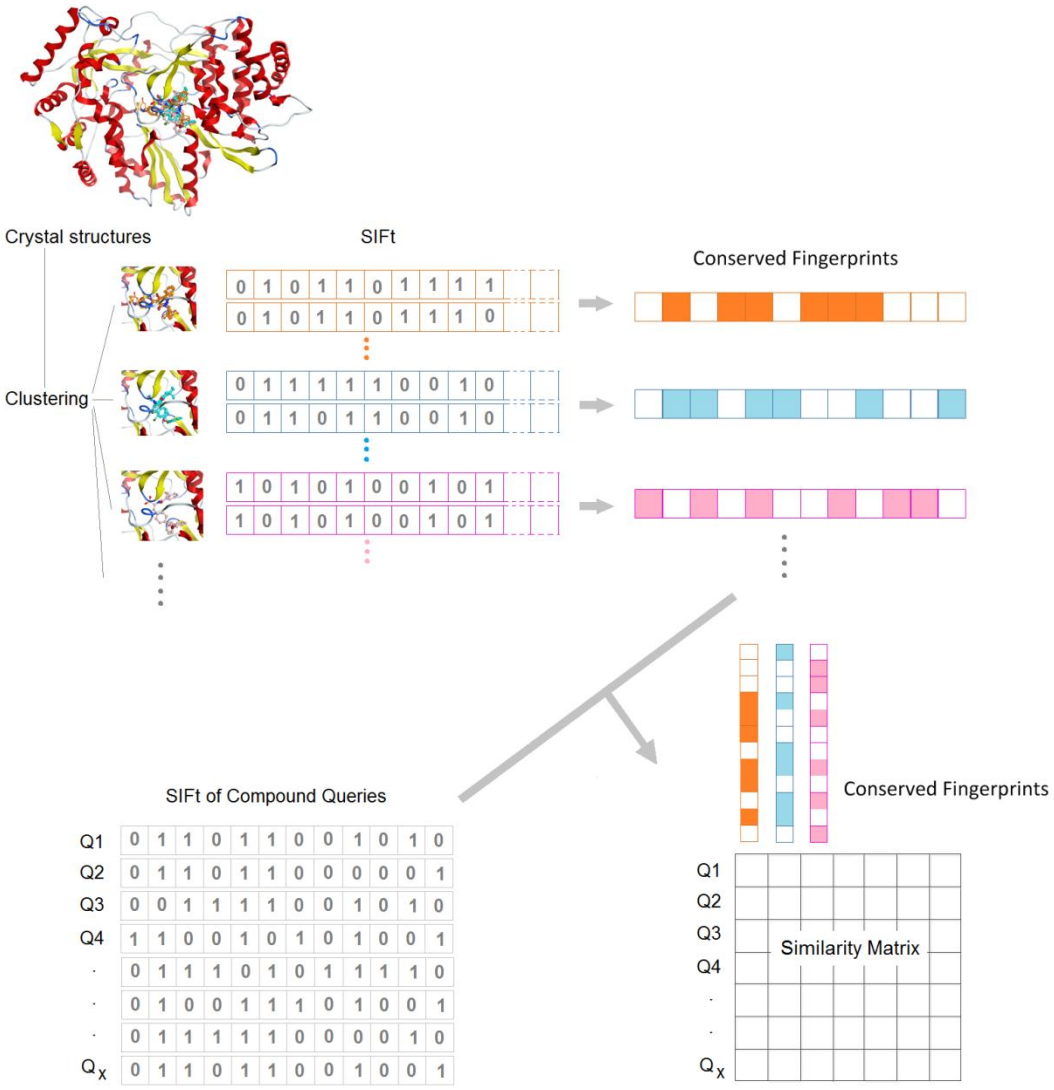


Figure 17. Schematic flowchart of SIFt calculation.

4.2.4 Two sites docking

We hypothesized that a given score of a ligand bound to a target binding site should be higher than that of an unspecific binding site. The target binding site in this study was PS-I of HCV NS5B polymerase and TS-II was used as a dummy binding site. The distribution of Glide SP scores of the training/testing dataset is shown in Figure 18. With Glide SP, lower scores represent better binding. The results showed that the median scores of the PS-I dataset docked in PS-I were lower compared to docking into the dummy binding site (TS-II) and the interquartile ranges did not overlap. However, the gap between the median of TS-II in two different binding sites was small. Based on this result, a score of -6.5 was set as cutoff for filtering docking results.

Potential inhibitors should have a score less than the cutoff in the target binding site (PS-I) and not less than this cutoff in a dummy binding site (TS-II in this study).

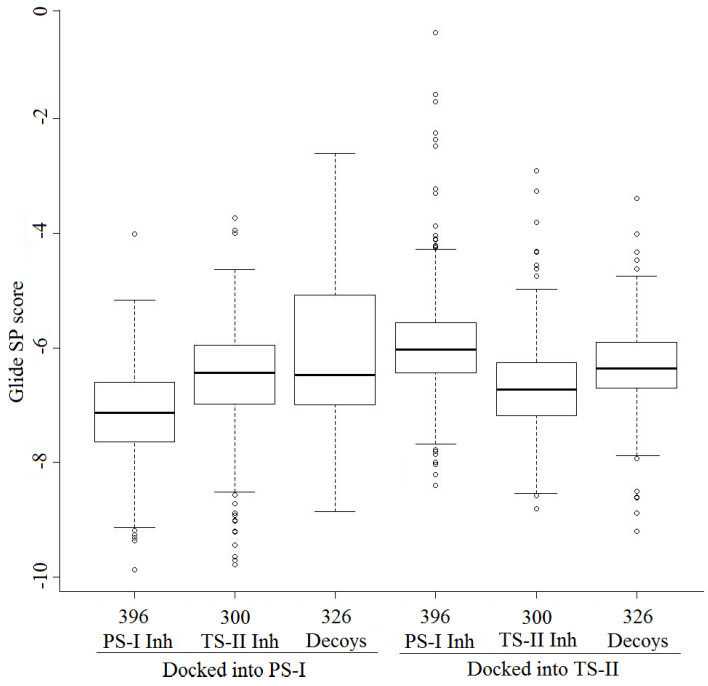


Figure 18. Boxplot of Glide SP scores of docking 396 PS-I inhibitors, 300 TS-II inhibitors and 326 decoys into two binding sites: palm site I (PS-I) and thumb site I (TS-II).

4.2.5 Dataset

Known PS-I inhibitors

The dataset of HCV inhibitors was manually collected from the literature. As aforementioned, there are at least four binding sites in HCV S5B polymerase: TS-I, TS-II, PS-I and PS-II. PS-I was the target binding site in this chapter, whereas TS-II and its inhibitors were used for method evaluation. The IC_{50} values of all the molecules used in this chapter were determined for HCV NS5B genotype 1.

To prepare the correct bound conformation, the docking poses of each compound from Gold and Glide SP were inspected and selected based on their similarity to the binding mode of the co-crystallized ligand with the highest 2D similarity. The 2D similarities (Tc) of each PS-I inhibitor against the 29 co-crystallized ligands (the same palm dataset as used in chapter 3) were calculated by two types of fingerprints in Open Babel [164]: FP2, a path-based fingerprint that indexes small molecule fragments and fingerprint type FP4, which uses a series of SMARTS

queries. Subsequently, only the co-crystallized ligand that had the highest similarity for each query was used as reference. The minimum similarity cutoff was set to 0.5. The cutoff was low compared to general practice where similarities higher than 0.7 are used, because the compounds eventually have to be inspected as to whether the pose binds similarly to the reference structure or not. A total of 496 PS-I inhibitors were selected for this study.

The PS-I dataset was divided into a training/testing set and a validation set. The validation set was used to assess the performance of the filtering method, whereas the training/testing set was used to build the RF models. The PS-I inhibitors were clustered into groups based on their Tanimoto coefficient using binning clustering in ChemMine [165]. At the 0.5 bin cutoff, 9 groups of inhibitors were yielded. However, 4 groups contained only a small number (1, 1, 1 and 9) of ligands so they were combined with the group that had the smallest number of compounds. This resulted in five binning groups as shown in Figure 19. Only the compounds in binning groups A, B and C were used for training/testing the RF model. A total of 396 compounds selected from binning groups A, B and C were included in the training/testing set. The validation set, consisting of ligands from the same binning group as the training/testing set (50 compounds from groups A, B and C), was considered as *'soft case'*, while the other validation set compiled from different binning groups (49 compounds from groups D and E) was referred to as '*hard case'*. Thus the whole validation set included soft case (50 compounds), hard case (49 compounds) and decoys (1693 compounds). The validation set was divided into soft case and hard case in order to evaluate the prediction performance of RF models. As RF is a learning method, if the query compounds are similar to the compounds in the training (soft case), it would be easy for RF to accurately predict the class of the query. However, in real scenarios, virtual screening aims not only to find the active compounds but also new scaffolds. Thus, it would be interesting to evaluate if the RF models are generally able to predict compounds which are significantly dissimilar to the training set (hard case).

Decoys

Decoys were generated by the service of the Database of Useful Decoys (DUD) website[166]. The decoy selection criteria were based on properties matching given ligands. These are six properties including molecular weight, estimated water-octanol partition coefficient, rotatable bonds, hydrogen bond acceptors, hydrogen bond donors, and net charge. A total of 1693 decoys were included in the validation set for enrichment study.

The training/testing set of the RF models

RF model-1, which was generated to classify candidates as 'binding or non-binding', was built with 396 PS-I inhibitors, 326 DUD decoys, 300 TS inhibitors and 300 incorrect poses. The published TS inhibitors were compiled and used as non-binding molecules for PS. Incorrect poses were included in this set because it was intended to yield 'binding' candidates that not only had similar binding to known inhibitors but also had a binding conformation representing a near-native pose. As model-2 was trained to classify inhibitory activity, inhibitory activity was divided into 'potent and weak' based on pIC_{50} values. Potent activity was defined as pIC_{50} (-log IC_{50}) equivalent to or greater than 7, while weak activity had pIC_{50} less than or equivalent to 6. Compounds with pIC_{50} between 6 and 7 were not used for the training model but were only kept in the validation set. Therefore, the training/testing dataset of model-2 included 124 potent PS-I inhibitors and 113 weak PS-I inhibitors.

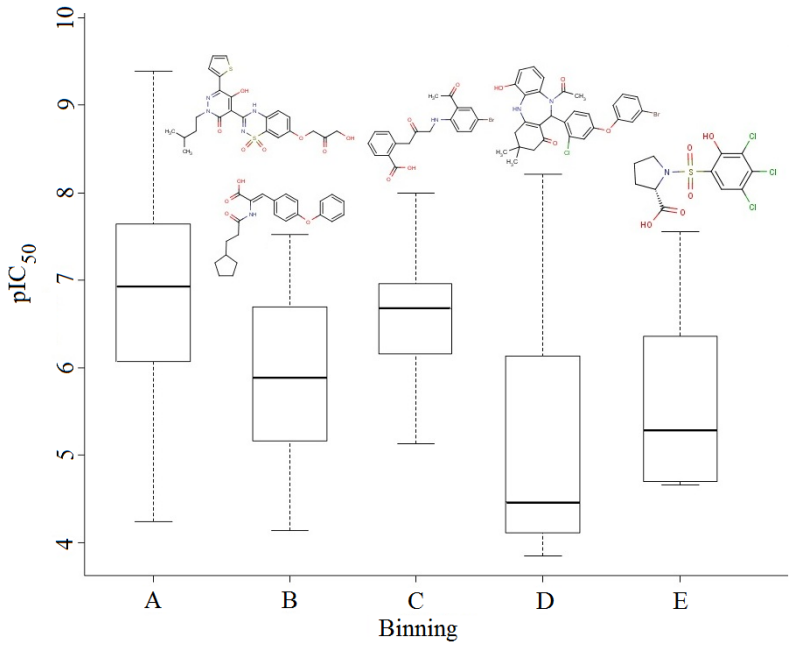


Figure 19. Boxplot of inhibitory profile (pIC_{50} = -log IC_{50}) for different binning clusters of PS-I inhibitors.

4.3 Results and discussion

A main goal of this work was to improve the hit identification of HCV NS5B polymerase inhibitors via a post-docking process. Three filtering methods, RF, SIFt and 'two sites docking', as well as combined methods, were evaluated. In total, there were six filtering models: i) RF model-1, ii) RF model-2, iii) combined model-I and model-II, iv) filtering and ranking by SIFT similarity score (denoted as SIFt), v) filtering by SIFt similarity score and ranking by Glide SP score (denoted as SIFt & SP), and vi) two sites docking.

In order to develop a RF model for classification, two RF models were built and tested by five-fold cross validation. Overall, the RF models (Table 9 and Table 10) -- model-1, which classified binding vs. non-binding and model-2, which classified potent vs. weak inhibitors -- conferred highly accurate classification. RF model-1 yielded an average 94% sensitivity, specificity and precision whereas RF model-2 yielded an average 72% specificity, 84% sensitivity and 80% precision. The results show that the present RF models exhibits satisfactory classification with respect to the prediction of the testing/training sets and can be used for classification.

Table 9. Predictive performance of RF model-1 on the fivefold cross validation set. 365 PS-I inhibitors and 926 decoys were used in total.

Dataset	Actual class	Predicted class		Total	Sensitivity	Specificity	Precision
		Binding	Non-binding				
Fold1	Known PS inhibitors	74	4	78	0.95	0.98	0.96
	Decoys	3	177	180			
Fold2	Known PS inhibitors	71	9	80	0.89	0.99	0.97
	Decoys	2	185	187			
Fold3	Known PS inhibitors	77	3	80	0.96	0.99	0.97
	Decoys	2	183	185			
Fold4	Known PS inhibitors	74	5	79	0.94	0.98	0.96
	Decoys	3	182	185			
Fold5	Known PS inhibitors	75	4	79	0.95	1	1
	Decoys	0	189	189			

Table 10. Predictive performance of RF model-2 on the fivefold cross validation set. 124 potent PS-I and 113 weak PS-I inhibitors were used in total.

Dataset	Actual class	Predicted class		Total	Sensitivity	Specificity	Precision
		Potent actives	Weakly actives				
Fold1	Potent	25	7	32	0.78	0.83	0.86
	Weak	4	19	23			
Fold2	Potent	27	5	32	0.84	0.82	0.87
	Weak	4	18	22			
Fold3	Potent	28	4	32	0.88	0.73	0.82
	Weak	6	16	32			
Fold4	Potent	22	7	29	0.76	0.65	0.73
	Weak	8	15	33			
Fold5	Potent	29	2	31	0.94	0.57	0.74
	Weak	10	13	23			

As RF is a machine learning method, knowledge about active compounds confines the predictive power of the model. In a real-life scenario, the chemical diversity of active compounds is unknown. The number of known actives is usually smaller and limited to some scaffolds. It would be therefore difficult for the trained model to predict unseen data. In this context, when assessing the performance of filtering methods, we used the validation set that compiled active compounds sharing structural similarity with the training set i.e. soft case and dissimilar to the training set i.e. hard case. The results showed as expected that the models were much more accurate in classifying the soft case than the hard case (Table 11 and Table 12). Using the validation set, the sensitivity of model-1 was reduced from 92% for the soft case to 14% for the hard case with model-1, and from 64% to 1% in the case of model-2. But both models retained high specificity. The specificities of model-1 and model-2 were 93% and 88%, respectively.

Nevertheless, in comparison to other filtering methods (Table 13), the RF models help to downsize the dataset to a manageable small subset and result in an improvement of the ranking (Figure 20). The number of active compounds retrieved in the top 100 ranked compounds was

plotted in Figure 20 and the number of PS-I inhibitors in both hard case and soft case are shown in Table 14. The integrated RF models showed much better performance than solo docking, and RF model-1 outperformed the other strategies. RF model-1 placed most of the soft-case PS-I inhibitors (42/50 inhibitors) in the top 100 ranked hits. Since the RF classifiers performed well only on the dataset similar to the training data, this might imply that using fingerprint similarity should give an equivalent result. However, the RF models still have some advantages. RF model-1 was also trained by presumably true bound conformations and incorrect conformations; therefore, it can be used to distinguish correct poses from wrong poses. In the evaluation of the dataset consisting of 250 wrong poses and 50 presumably right poses of soft case inhibitors, RF model-1 showed 92% sensitivity, 75% specificity and 43% precision for 'binding' class prediction.

Table 11. Predictive performance of RF model-1 on the independent validation set.

Dataset	Actual class	Predicted class		Total	Sensitivity	Specificity	Precision
		Binding	Non-binding				
Validation	Hard case	7	42	49	0.14	0.93	0.05
	Soft case	46	4	50	0.92	0.93	0.27
	Decoys	122	1571	1693			

Table 12. Predictive performance of RF model-2 on the independent validation set.

Dataset	Actual class	Predicted class		Total	Sensitivity	Specificity	Precision
		Potent active	Weakly active				
Validation							
	Hard case (49 actives)				0.1	0.88	0.03
	Potent	-	22	22			
	Intermediate	1	7	8			
	Weak	4	15	19			
	Soft case (50 actives)				0.64	0.88	0.14
	Potent	21	4	25			
	Intermediate	9	11	20			
	Weak	2	3	5			
	Decoys	193	1500	1693			

Table 13. The number of compounds that passed the selection criteria using six different filtering approaches.

Group	Absolute number of compounds passing the filter					
	SIFt	**SIFt & SP**	**RF Model-1**	**RF Model-2**	**RF Model-1 & 2**	**Two sites docking**
PS-I inhibitors [99]	94	94	54	37	32	54
Decoys [1693]	823	823	122	193	33	266
Total [1792]	917	917	176	230	65	320

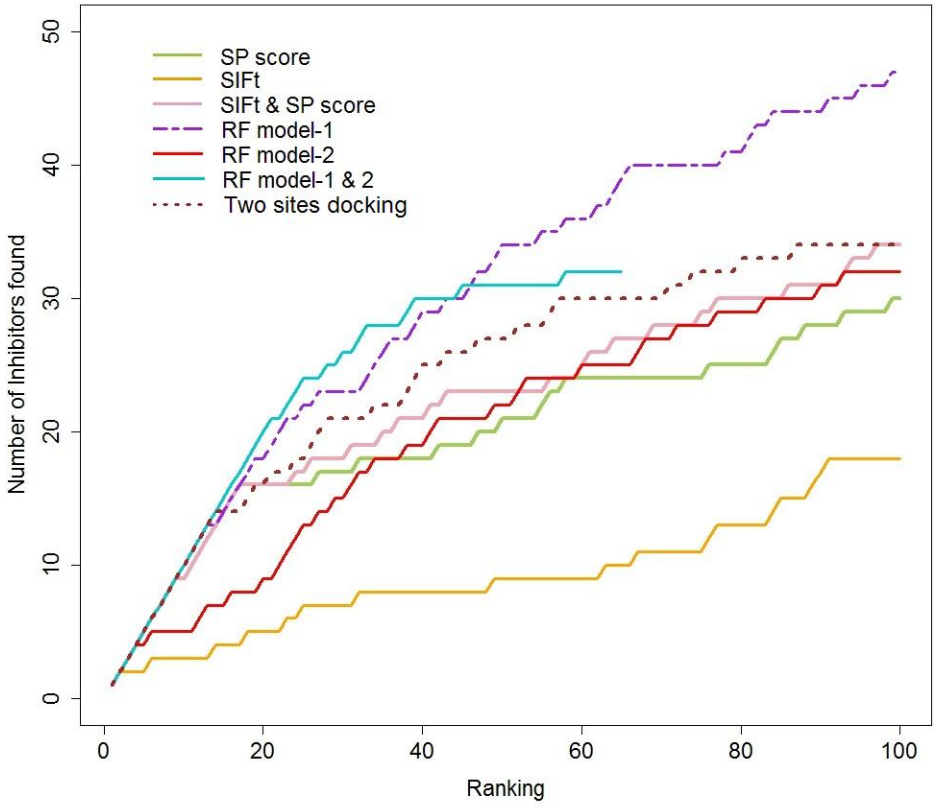


Figure 20. Aggregate number of known PS-I inhibitors found in top 100 ranking using six filtering approaches and Glide SP without filtering on the validation set.

Table 14. The number of known PS-I inhibitors (true positives) found in the top 100 ranking by using different approaches on the validation set. 'Hard case' denotes an enrichment study using only the hard case with decoys. Similarly, 'Soft case' denotes an enrichment study using only the soft case with decoys. Note, the decoys dataset was the same in both experiments. 'Whole' represents using the hard case and soft case plus the decoys in the enrichment study. The number of total active compounds in each case is shown in bracket.

In top ranking	Number of known PS1 inhibitors											
	SP score			SIFt			SIFt & SP score			Two docking sites		
	Hard case [49]	Soft case [50]	Whole [99]	Hard case [49]	Soft case [50]	Whole [99]	Hard case [49]	Soft case [50]	Whole [99]	Hard case [49]	Soft case [50]	Whole [99]
10	9	0	9	1	2	3	9	1	9	10	4	10
20	16	2	16	2	4	5	16	2	16	14	7	16
30	16	3	17	2	5	7	16	4	21	15	10	21
40	16	6	18	2	6	8	16	6	24	17	12	25
50	17	6	21	2	7	9	18	6	27	18	12	27
100	20	13	30	9	11	18	20	13	38	18	17	34

In top ranking	Number of known PS1 inhibitors								
	Random Forest								
	Model-1			Model-2			Model-1 & Model-2		
	Hard case [49]	Soft case [50]	Whole [99]	Hard case [49]	Soft case [50]	Whole [99]	Hard case [49]	Soft case [50]	Whole [99]
10	3	10	10	4	3	5	1	10	10
20	5	18	18	4	8	9	1	20	20
30	5	21	23	4	14	15	1	26	26
40	5	25	29	4	17	19	1	29	30
50	6	30	34	4	20	22	1	30	31
100	7	42	47	5	28	32	1	31	32

RF model-1 exhibited better recall rate than model-2. Model-1 was well trained by structurally diverse compounds including both known inhibitors and decoys, whereas model-2 was trained only by known inhibitors (potent or weakly actives). Model-1, therefore, has the advantage over model-2 in predicting unseen data. The performance of the model-1 and model-2 combination improved as a result of the retrieval of active compounds that were similar to the known inhibitors (Table 14). The RF models failed to recall true actives in the hard case. In applications, where a novel scaffold or high recall rate is prioritized, the RF models will not be suitable. On

the other hand, if a small subset of actives should be identified in VS, the present RF models will be a good choice for filtering compare to the other methods.

Using SIFt and ranking by Glide SP score was performing better than using SIFt and ranking by similarity scores because the magnitude of similarity is not responsible for differing activities. For example, a small difference between analogous compounds can result in a loss or gain of activity but they might have the same fingerprints. Furthermore, the Tanimoto coefficient is a global similarity measure; compounds that have the same similarity score might have different binding modes and different activities. Hence, the docking score that estimates the relative binding affinity should perform better in ranking compounds than their similarity score. Nevertheless, the results showed that using SIFt and ranking by Glide SP was slightly better in terms of hit identification compared to simple docking.

Overall two sites docking achieved better enrichment than the other methods except RF model-1 and a combination between RF model-1 and RF model-2 (Figure 20). When both the soft case and hard case were taken into consideration, two sites docking was the best filtering approach (Table 14). In the top 100 ranked compounds of the hard case, two sites docking recalled more active compounds than the RF models. And in the top 100 ranked compounds of the soft case, two sites docking could enrich the actives better than both SIFt methods and docking alone. The ranking score in docking might not well correlate with the binding affinity, but our enrichment study showed that the docking score gives acceptable results in retrieving true active compounds.

This study was set up to explore three post-docking strategies and identified the most suitable method to screen for novel HCV NS5B polymerase inhibitors. RF enriched true active compounds better than the other methods. However, RF gives the highest hit rate of already known scaffolds. Thus, if a particular scaffold is of interest, RF represents the best method of choice. On the other hand, finding interesting new scaffolds may be preferable over the total hit rate. In this case, two sites docking is suited for an efficient identification. Moreover, two sites docking does not require additional time to prepare a predictive model. Based on the results, combining RF and two sites docking would improve overall hits identification. A proposed virtual screening setup is shown in Figure 21. The screening is first done based on two sites docking. The cut off criteria was strictly set to a score of -7.5. In cases where the database size is too large, the number of compounds passing this cutoff might still be very large; hence only 1–10% of the top ranked compounds will be passed on to docking into the dummy site (TS-II).

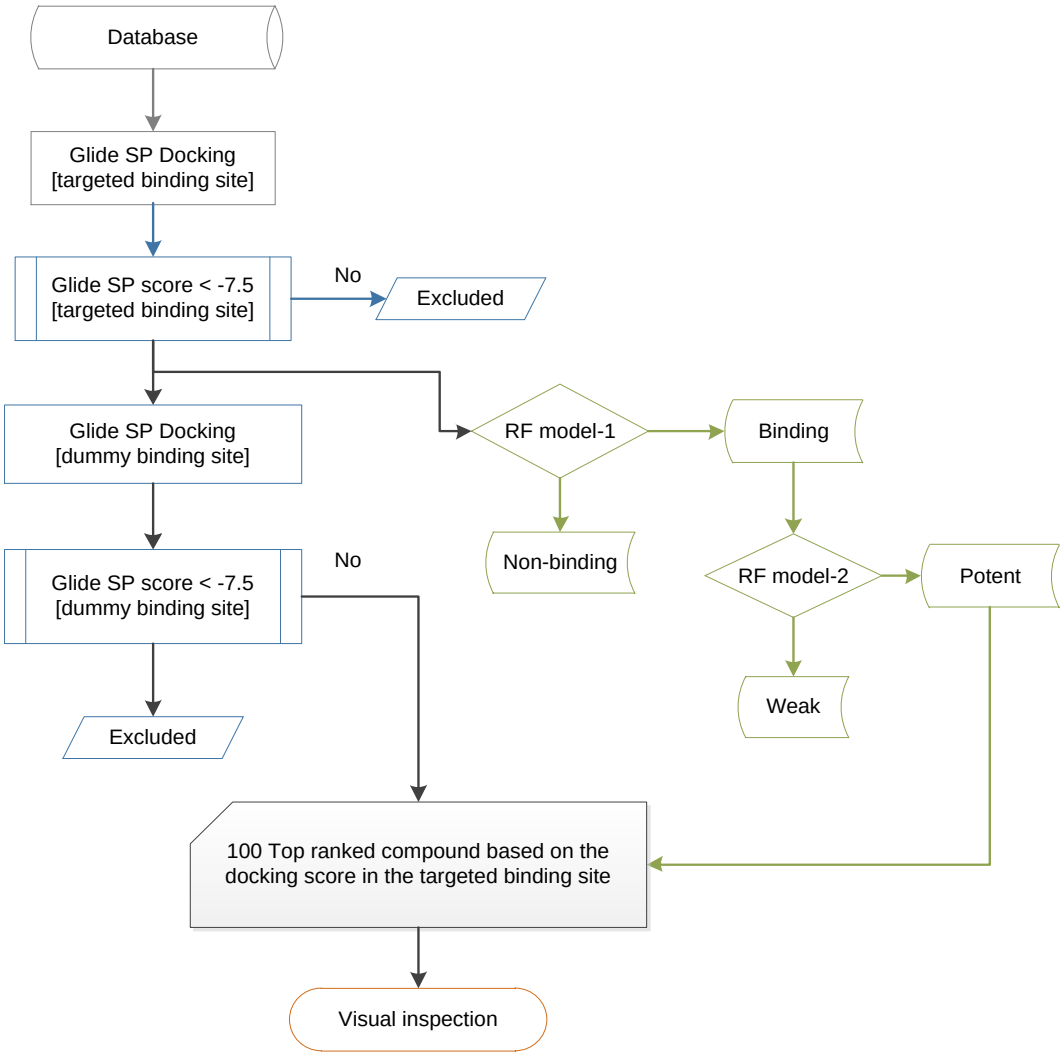


Figure 21. Schematic flowchart of the virtual screening setup to identify novel HCV NS5B polymerase inhibitors.

The aim of this study was not only to develop a protocol but to utilize it for screening of novel inhibitors. As an example we applied this VS setup for screen the ChemBridge database consisting of 392,589 compounds. Docking scores of 19,983 compounds passed the cutoff criteria (< 7.5) at the target site (PS-I) and 19,034 compounds finally passed through dummy site's criteria. There were 84 binning clusters using similarity cutoff 0.6 in ChemMine. All compounds were visually inspected, finally we proposed 28 compounds as PS-I inhibitors (Figure 22). Even though the compounds have not been experimentally tested, their binding modes resemble those of a known inhibitor (PDB: 1Z4U) as shown in Figure 23. The IC_{50} of the inhibitor co-crystallized in1Z4U is 0.07 µM.

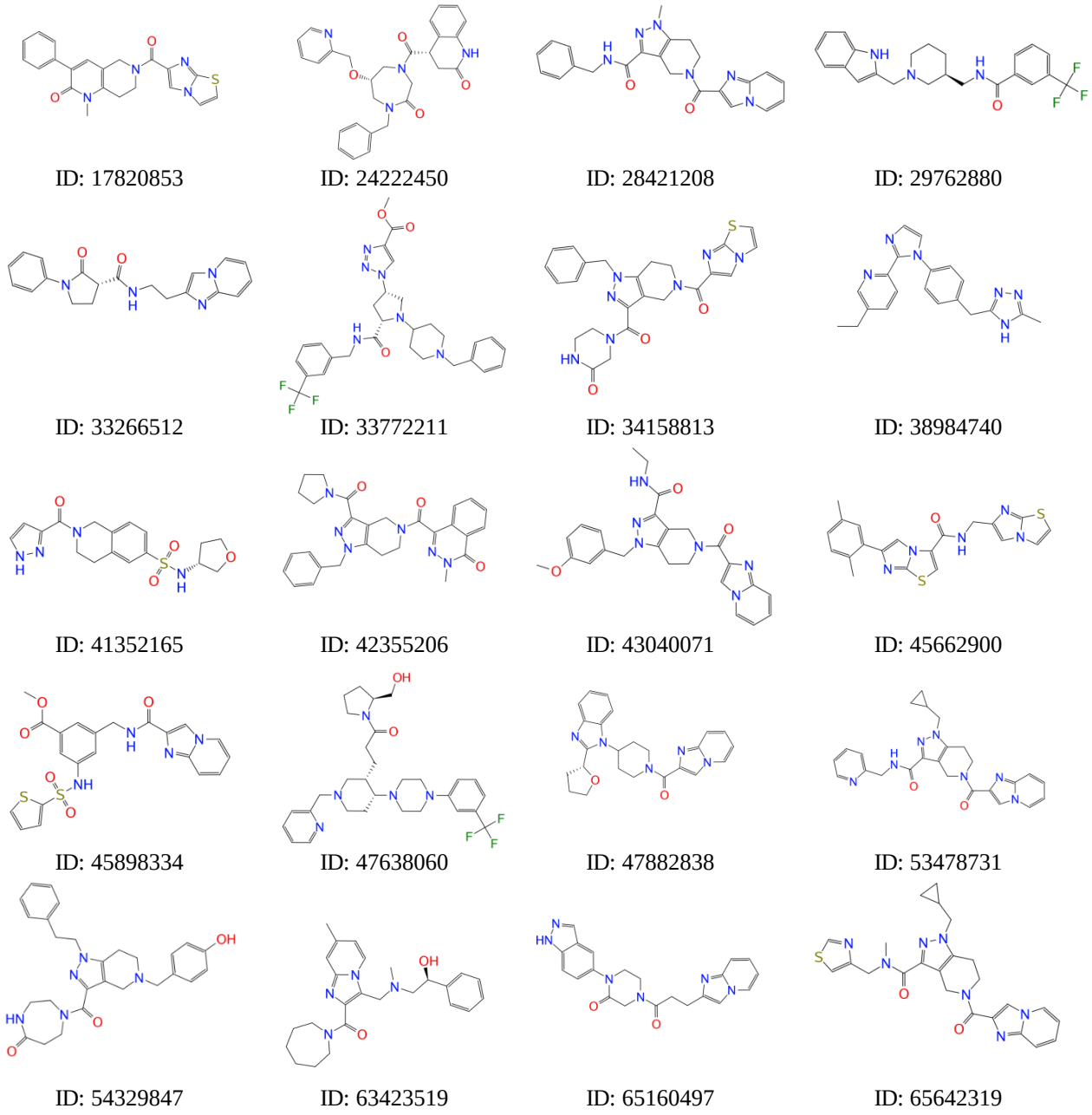
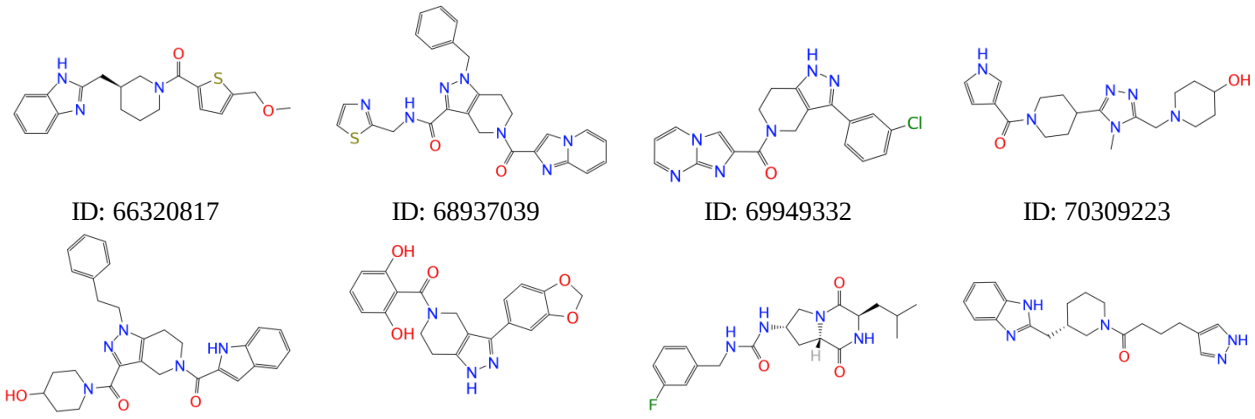


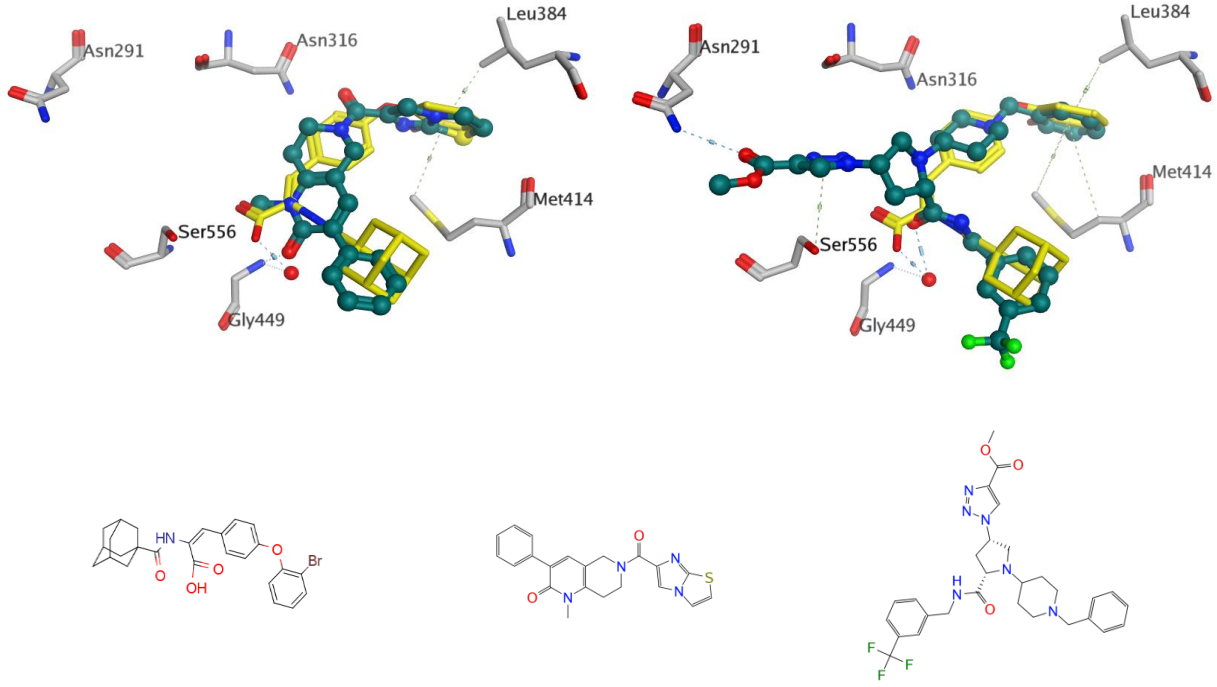
Figure 22. Structures and ChemBridge ID of proposed palm site I inhibitors.



ID: 66320817 ID: 68937039 ID: 69949332 ID: 70309223

ID: 74187007 ID: 74985664 ID: 75340320 ID: 75424279

Figure 23 (Continue). Structures and ChemBridge ID of proposed palm site I inhibitors.



PDB ID: 1Z4U ID: 17820853 ID: 33772211

Figure 23. Superimposition of the crystal structure of 1Z4U (yellow), with the docked poses of proposed compounds (cyan) ID: 17820853 (Left) and ID: 33772211 (Right). Both hits show a similar shape and interaction profile as the known HCV NS5B inhibitor.

Chapter 5 Virtual screening of HCV NS5B polymerase inhibitors using pharmacophore model and docking

5.1 Introduction

Among NNIs, the inhibitor HCV796 (Figure 24) displayed antiviral activity across multiple HCV genotypes in clinical trials [73, 167]. HCV796 exhibits median inhibitory concentration (IC_{50}) of 0.01 to 0.14 µM for genotype 1 [75] with a half maximal effective concentration (EC_{50}) of 4-25 nM against genotype 2 replicons (Table 15 and Table 16). Although HCV796 was terminated in the clinical trial phase II because of hepatocellular toxicity, screening for compounds, which interacts with HCV NS5B polymerase in a similar way as HCV796, represents a promising approach. Pharmacophore-based virtual screening and molecular docking approaches were used in the current work to identifying new inhibitors. Two series of inhibitors (dataset-I: 14 compounds and dataset-II: 11 compounds) were identified by VS and their inhibitory effect was tested experimentally *in vitro* assay.

First, virtual screening was done based on the available crystal structure of genotype 1 HCV NS5B polymerase. Due to availability experiment, *in vitro* testing was carried out on genotype 2a HCV NS5B polymerase. As HCV796 is effective in both genotypes and its binding site is quite conserved (Figure 1), the first underlying assumption was that the genetic variation between genotype 1b and 2a does not change the binding mode i.e. the pharmacophore model and docking should give similar results regardless of the genotype variation. Nevertheless this genetic variation impacts binding affinity. MD simulation and binding energy calculation were thus employed to understanding the binding affinity for genotype 1b and genotype 2a.

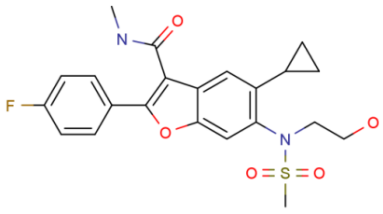


Figure 24**.** HCV796, a benzofuran derivative binding to PS-II.

Table 15. Inhibitory activity of HCV796 in different genotypes [98].

Inhibitor	EC_{50} (µM)		
	Genotype 1b (Con1)	**Genotype 1a (H77)**	**Genotype 2a (J6/JFH)**
HCV796	0.007	0.004	0.25

Table 16. Inhibitory activity of HCV796 in mutants [97]. The inhibitory activity is represented as fold change calculated as the ratio of the concentration used to a half maximal inhibitory concentration (IC_{50}) against HCV polymerase of the 1b Con1 strain. IC_{50} of HCV796 on 1b Con is 0.04 µM.

Mutation	Remark	Fold shift over
		1b Con1 (IC_{50} 0.04 µM)
M414T	Palm site I	1.5x
G554D		1x
C316N	Palm site II	3x
C316Y		50x
P495L	Thumb site I	1x
M423T	Thumb site II	1x
1b-BK	Genotype 1b	6x
1a-H77	Genotype 1a	1x

5.2 Binding mode analysis of HCV796

There are two crystal structures of HCV796 bound to PS- II of both BK (PDB ID: 3FQK) and Con-1 strain (PDB ID: 3FQL) of NS5B genotype 1b. Both structures show high structural similarity and sequence identity (96% identity). As illustrated in Figure 25, both crystal structures show two hydrogen bonds between the inhibitor and, Ser365 and Arg200 but there is a difference in the orientation of Arg200. (i) Ser365 forms a hydrogen bond between its side chain hydroxyl group and the amide nitrogen of HCV796. (ii) Arg200 interacts with the sulfonyl moiety of HCV796 in 3FQK and with the amide oxygen of HCV796 in 3FQL. MD studies carried out for these crystal structures showed that the hydrogen bond interaction of the sulfonyl moiety with Arg200 was the most preserved throughout the simulation (Table 17). The rotamer

of Arg200 in 3FQL was therefore edited similar to that in 3FQK and used with both crystal structures for docking studies. Based on structural comparison between apo-structure and HCV796 bound structure, the side chain of Arg200 appears to be flexible and controls the access to PS-II. Similar observation was made in the comparison between the bound structure of HCV796 and other allosteric site inhibitors. Thus it is conceivable that the interaction with Arg200 is essential for the inhibitory effect of HCV796.

The hydrogen bond with Asn316 was found in the 3FQK structure, but was not observed in the MD simulation of 3FQL, in which Asn is substituted by Cys316. However the affinity of HCV796 in the BK strain (3FQK) is lower than in the Con-1 strain (3FQL). Hence, the interaction between Asn316 and HCV796 in 3FQK appears to be less energetically favorable compared to the Cys316-HCV796 interaction found in 3FQL. The decrease of inhibition is due to a steric clash [168, 169]. The mutation C316N represents a natural polymorphism of HCV NS5B 1b subtype.

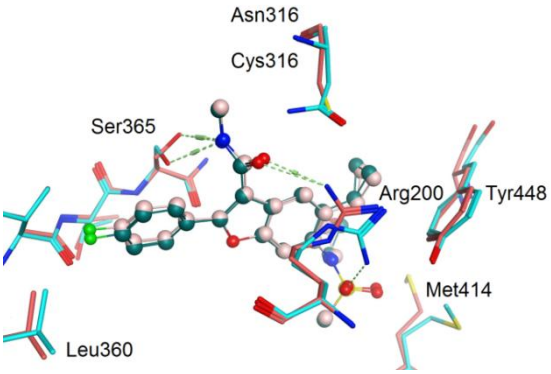
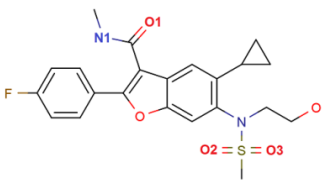


Figure 25. Binding mode of HCV796 (balls and sticks) in genotype 1 HCV NS5B polymerase. The Con-1 strain (PDB ID: 3FQL) is shown in pink whereas the BK strain (PDB ID: 3FQK) is shown in blue.

Table 17. Hydrogen bond summary of the complex of HCV796 and either HCV NS5B polymerase genotype 1b BK1 (3FQK) or Con1 (3FQL) strain. Distance cutoff is 3.00 Å, angle cutoff is 120.00°.



Complex	Hydrogen bond between		%Occupied
	Ligand	Protein	
BK1	O2	Arg200	80.98
PDB: 3FQK	O1	Asn316	67.5
	N1	Ser365	40.4
Con1	O2	Arg200	85.62
PDB: 3FQL	N1	Ser365	40.33

5.3 Virtual screening based on HCV-796 binding mode

A first series of 14 compounds was selected based on top ranked (Gold-score) docking poses (Figure 26 and Figure A6-Appendix). The HCV NS5B polymerase activity was tested by Tobias Hoffmann and Dr. Ralpl Golbik at the Institute for Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg [135]. The experimental results showed that compounds C4 – C8 and C13 had relatively weak inhibition compared to the control while the other compounds did not show any inhibitory activity. According to MD simulations, only three weakly active compounds: C6 and C7, and one inactive compound C12 showed hydrogen bonds with Arg200 and Ser365 similar to HCV796 (Table 18). Interestingly, C12 showed the same hydrogen bond patterns and its structure also has two rings similar to HCV796. The superimposed compound C12 with HCV796 pharmacophore features showed that C12 lacks hydrogen acceptors atom that interacts with Arg200 (Figure 27). This result supports the hypothesis that the interaction with Arg200 is a key factor for strong binding.

Compound C3, C6 and C7 share the same scaffold (p-sulfonyl-acetanilide) but only C6 and C7 show weak activity. Compound C6 and C7 also make similar hydrogen bond interactions with Arg200 and Ser365 as HCV-796 (Table 18). Among the three compounds, the complex of C7-HCV polymerase showed the most stable conformation during MD simulation and longest hydrogen bond interactions with Arg200 and Ser365 (Figure A7-Appendix). The extended part of compound C7 (2,2,4-trimethyl-1,2-dihydroquinolin-6-ol) is more hydrophobic compared to the extended part of C3 (2-phenylethan-1-amine) (Figure 28). This implies that the hydrophobicity might affects the binding.

To increase the hit rate and to identify further potent inhibitors, a pharmacophore based VS was conducted in the next step.

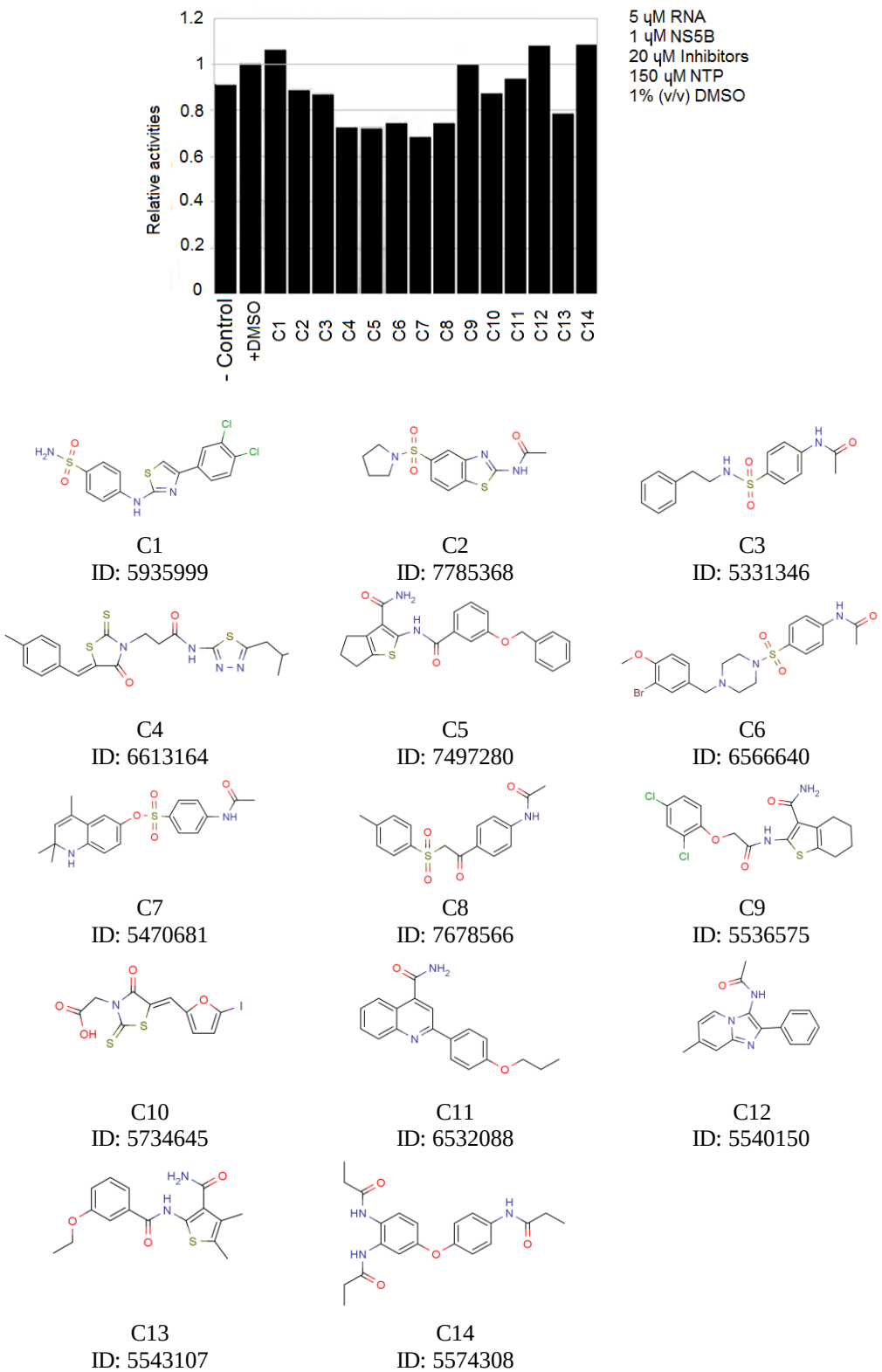


Figure 26. Structures and ChemBridge ID of 14 experimental tested compounds (dataset-I) on genotype 2a HCV NS5B polymerase.

Table 18. Hydrogen bond analysis based on 6 ns MD simulation.

Compounds	Hydrogen bond between		%Occupied	Distance (±SD)
	Ligand	Protein		
C6	O@amide	NH1@Arg200	43.43	2.82 (0.09)
	N@amide	OG@Ser365	8.18	2.88 (0.07)
C7	O@sulforyl	NH2@Arg200	42.75	2.84 (0.09)
	N@amide	OG@Ser365	59.23	2.89 (0.07)
C12	O@amide	NH1@Arg200	77.05	2.83 (0.09)
	N@amide	OG@Ser365	30	2.91 (0.07)

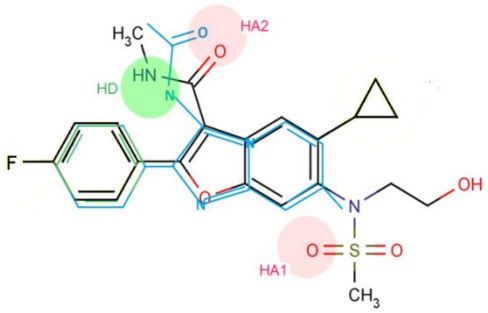


Figure 27. Overlay of compound C12 and the pharmacophore features of HCV796. The hydrogen bond donor (HD) is represented as green circle and hydrogen bond acceptor (HA) as red circle.

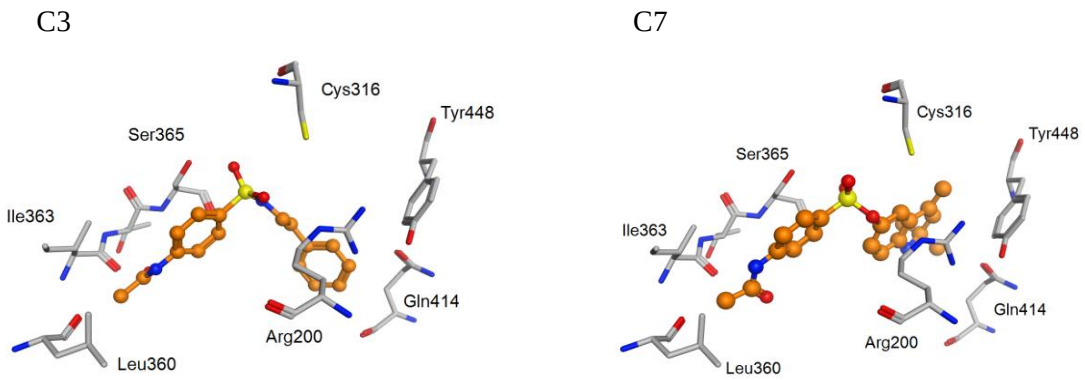


Figure 28. Docking poses of compound C3 and C7 at the PS-II of genotype 1b HCV NS5B polymerase.

5.4 Pharmacophore modeling

Based on the study described above, the pharmacophore models were initially derived from the crystal structures of HCV polymerase complexed with HCV796 (3FQK and 3FQL) using the software Ligandscout [170]. The initial pharmacophore models were merged which composed of five features; two hydrophobic features (HP), two hydrogen bond acceptors (HA) and one hydrogen bond donor (HD) as shown in Figure 29. In order to build a suitable pharmacophore model, an enrichment study generally requires a large number of diverse structures for the test set, with an activity range of at least 3.5 orders of magnitude. Due to the number of known PS-II inhibitors is a few, an enrichment study of the pharmacophore hypotheses was carried on the dataset consisting of 21 active compounds and 348 decoys. The 21 actives were compiled from the in-house dataset-I (6 compounds: C4-C8 and C13), and a study of Kim et al [171] (15 compounds: Table A9-Appendix).

Based on the enrichment study (Table 19), the pharmacophore model which composes of 2 features: HD and HA1 (model 1) was selected to avoid extreme restrictions imposed on the pharmacophore screening. These features: HD and HA1 which were expected from the hydrogen bond interaction were set as mandatory features. According to the published mutations, the decrease of inhibition was due to the mutation of the residues 314, 316, 363, 365, 368, 414 (Table 16) [172-175]. These amino acids interact with HCV796. Amino acids Leu314 and Leu316 are located within the active site loop. Amino acids 363-368 are located in the serine rich loop and Met414 is found in the alpha-helix M. It implies that the pharmacophore features at these regions would be important. Hence, the remainders (HP and HA2) were set as optional features for the virtual screening. Exclusion volumes, which represent the forbidden area for a ligand, were also included in the pharmacophore model.

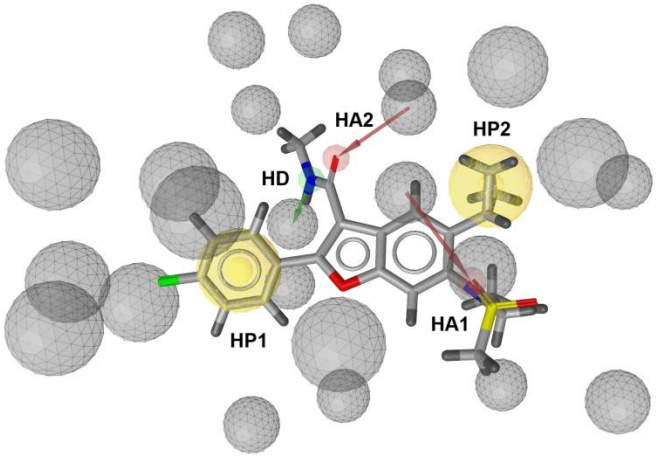


Figure 29. LigandScout pharmacophore model derived from the NS5B-HCV796 structure. The pharmacophore features are represented by LigandScout color codes: hydrogen bond donor: HD (green arrow), hydrogen bond acceptor: HA (red arrow), hydrophobic region: HP (yellow sphere), and excluded volume (black sphere).

Table 19. Enrichment study. The study was carried on the dataset of 21 active compounds and 348 decoys. Present of a pharmacophore feature is indicated as 'x', whereas absent of a pharmacophore feature is indicated as '-'.

Model	Features					Hits		
	HD	HA1	HA2	HP1	HP2	Total	Actives	Inactives
1	x	x	-	-	-	151	14	137
2	x	x	x	-	-	116	14	102
3	x	x	-	x	-	101	2	99
4	x	x	-	-	x	114	3	111
5	x	x	x	x	-	38	1	37
6	x	x	x	-	x	38	1	37
7	x	x	-	x	x	35	1	34
8	x	x	x	x	x	3	1	2

5.5 Database screening

The flowchart of the virtual screening is shown in Figure 30. First, the generated structure-based pharmacophore was used to screen compound databases. Then, two different docking programs: GOLD 4.1 [120] (Gold-score), and Glide SP [116] were applied to dock the hit compounds from the pharmacophore screening. Referring to the evaluation of docking/scoring functions as studied in Chapter 3, Glide SP was found to be the best scoring function for PS- I. However, Gold-score was also used because it was shown to be able in reproducing the native crystal

structure (Table A10-Appendix). For the protein structure 3FQL and two rotamers of Arg200 were used for the docking. The docking poses were further filtered by the designated pharmacophore to ensure that the compounds should fulfill the binding hypotheses. The consensus results and the top 100 ranked compounds from the ChemBridge, ChemDiv and LifeChemical databases were selected and the docking poses were visually analyzed. Among the top-ranked solutions 11 compounds (dataset-II) were available for purchasing (Figure 31). All compounds were tested *in vitro* using the assay described in 2.6.

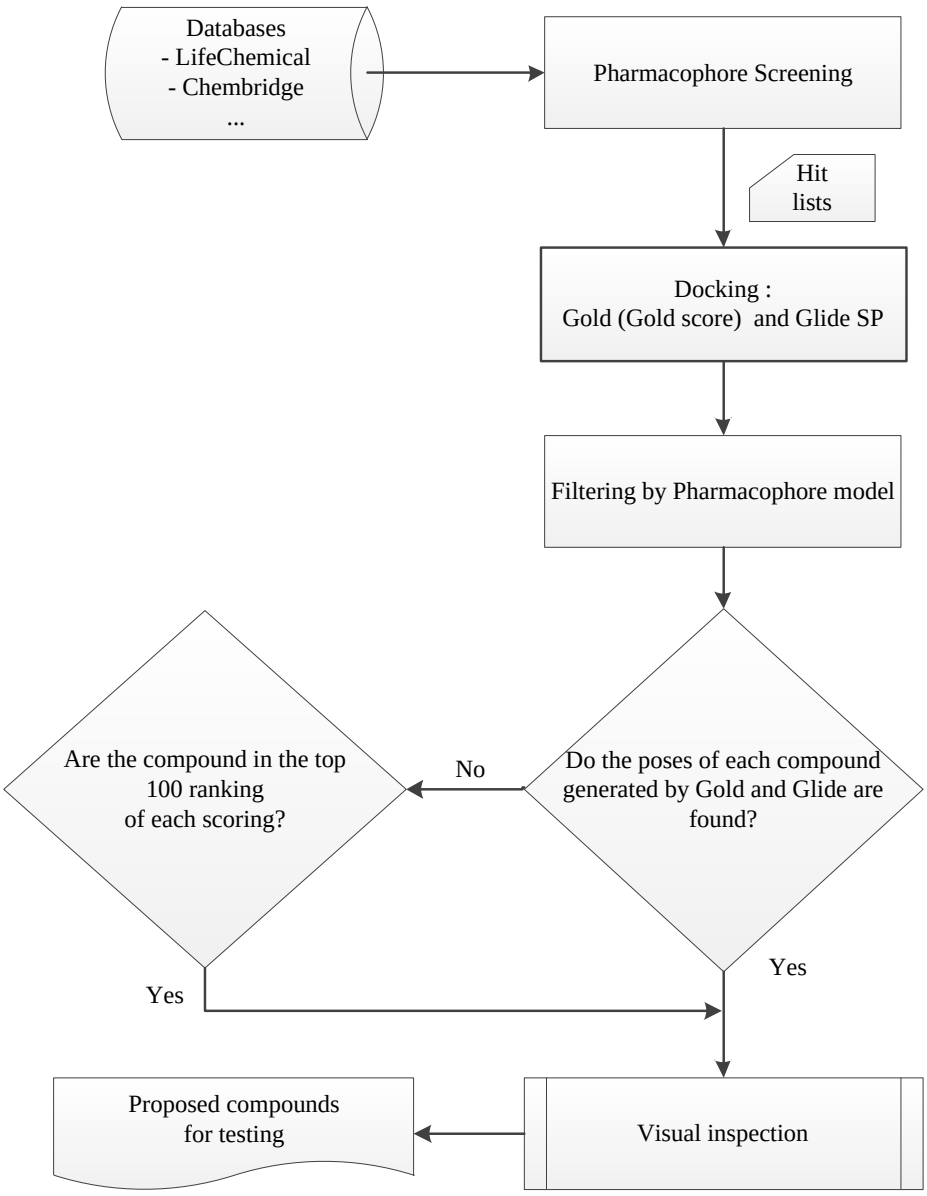
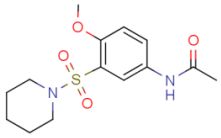
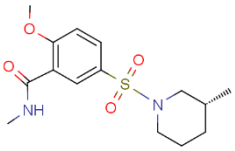


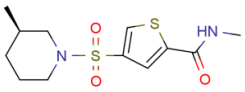
Figure 30. Flow chart of the virtual screening setup.



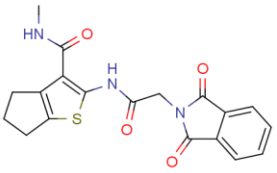
T1 :
9061595 (ChemBridge)



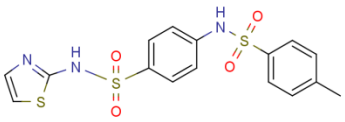
T2 :
9136363 (ChemBridge)



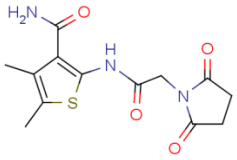
T3 :
9151960 (ChemBridge)



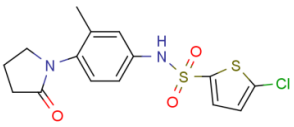
T4 :
G839-0062 (ChemDiv)



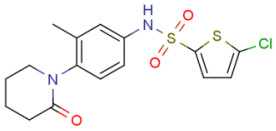
T5 :
F1443-5254 (LifeChemicals)



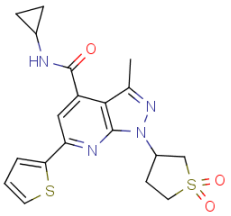
T6 :
F2590-0217 (LifeChemicals)



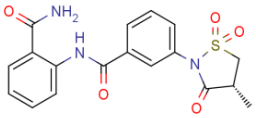
T7 :
F27783-1344 (LifeChemicals)



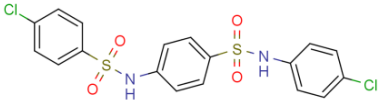
T8 :
F2784-1340 (LifeChemicals)



T9 :
F5121-0816 (LifeChemicals)



T10 :
F5296-0710 (LifeChemicals)



T11 :
6810740 (ChemBridge)

Figure 31. 2D structures of 11 purchased compounds from ChemBridge, ChemDiv and LifeChemicals which were selected for *in vitro* testing (dataset-II).

5.6 Experimental results

Binding and activity assays were determined by Tobias Hoffmann and Dr. Ralpl Golbik at the Institute for Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg [135]. The equilibrium dissociation constant (K_d) and relative inhibitory activity of 11 putative inhibitors (dataset-II) were determined as shown in Table 20 and Figure 32. Besides the binding and inhibition assay on HCV NS5B polymerase, inhibition of NS5 of the West-Nile virus (WNV), which is in the same family (*Flaviviridae*) as HCV, was also examined.

Table 20. Equilibrium dissociation constants (K_d) between inhibitors and HCV polymerase in the absence of RNA. The K_d value of HCV796 (positive control) in the HCV polymerase assay is 27 µM based on the study of Reich [176].

Inhibitor	Kd for NS5B polymerase (µM)	Kd for WNV NS5 (µM)
T1	5.0	7.0
T2	2.5	25.0
T3	1.0	NA
T4	2.5	50.0
T5	17.0	50.0
T6	6.0	11.0
T7	8.0	NA
T8	30.0	60.0
T9	30.0	25.0
T10	60.0	48.0
T11	27.0	35.0

*NA = No binding

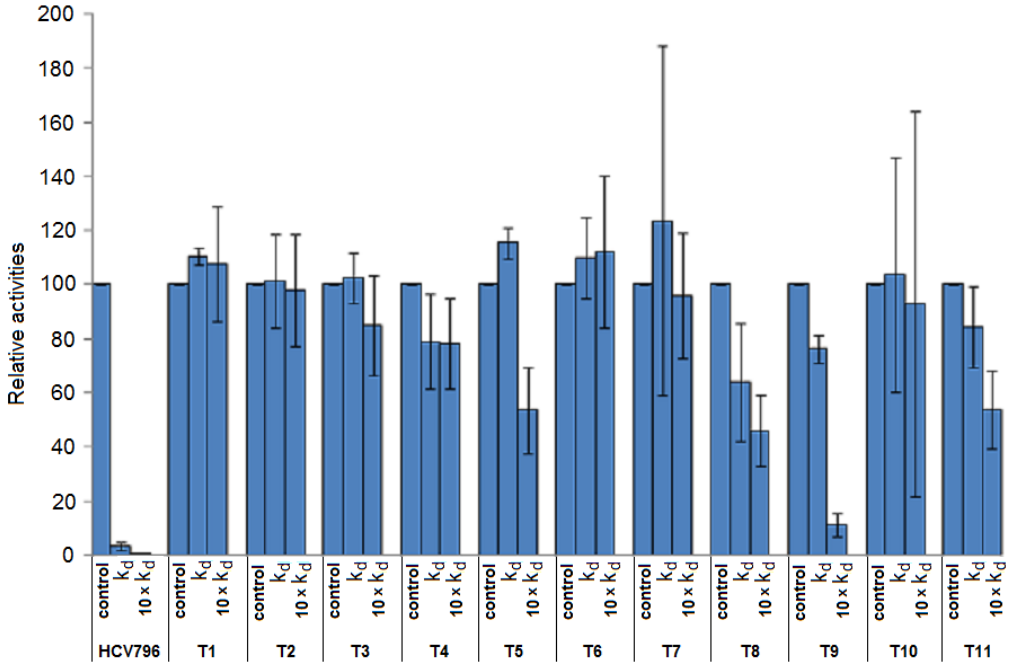


Figure 32. Relative inhibitory activities of 11 tested compounds. Relative inhibitory activity was calculated by dividing the RNA concentration in the presence of the presumably inhibitor by the RNA concentration in the absence of an inhibitor (Figure A8-Appendix). 100% NS5B polymerase activity was defined as the RNA yield in the control. Error bars represent standard deviation from three replications.

In theory, a lower K_d value represents a higher affinity of the inhibitor for its target. However the results of K_d presented here were determined in the absence of the substrate i.e. RNA and nucleotide triphosphate (NTP). Thus the compounds which have a small K_d value, might not tightly bound the target protein when compared to the substrate binding. We observed that compound T2-T4 show low K_d values ($\leq$ 2.5 µM), but showed no inhibition. While compound T5, T8, T9 and T11 have much higher K_d values, but showed inhibition activity at ten times the concentration of their K_d values (Table 20 and Figure 32). The K_D value of HCV796 (positive control) in the polymerase assay is 27 µM based on the study of Reich [176].

The docking poses of the selected pharmacophore hits tested in genotype 1b (PDB ID: 3FQL) share similar binding interaction as HCV796 (Figure A9-Appendix). However, only 4 compounds -- T5, T8, T9 and T11—were found to have an inhibition activity, albeit weakly. This result might due to the genotype difference between VS screening and experiment as aforementioned. To better understand the obtained results, MD simulation of NS5B genotype 2 and the pharmacophore hits was carried out.

5.7 Binding mode analysis of the pharmacophore hits in genotype 2a HCV NS5B polymerase

The amino acid residues surrounding HCV796 genotypes 1b isolate Con1 (PDB ID: 3FQL) and genotype 2a isolate JFH1 (PDB ID: 3I5K) are highly similar as shown in Figure 33. Only one residue at the position 414 is different (Met414 in genotype 1b and Gln414 in genotype 2a). Nevertheless this might significantly affect the binding affinity as it has been reported that the mutation M414T causes an 1.5 fold change in the inhibition (Table 16). The EC_{50} HCV796 for genotype 2a is 35 fold genotype 1b (Table 15).

Since the crystal structure of genotype 2a (PDB ID: 3I5K) represents an apo-structure, it cannot directly be used for docking studies. Hence, the amino acid sequence of PDB ID: 3I5K was used to build a model based on the coordinates of PDB ID: 3FQL in order to represent the HCV796 bound conformation. The homology model of genotype 2a was prepared by using SWISS-model [177]. Furthermore, the structure of three different rotamers of Arg200 were generated and used in an ensemble docking in Gold (Gold-score). The top ranked poses were then subjected to MD simulation.

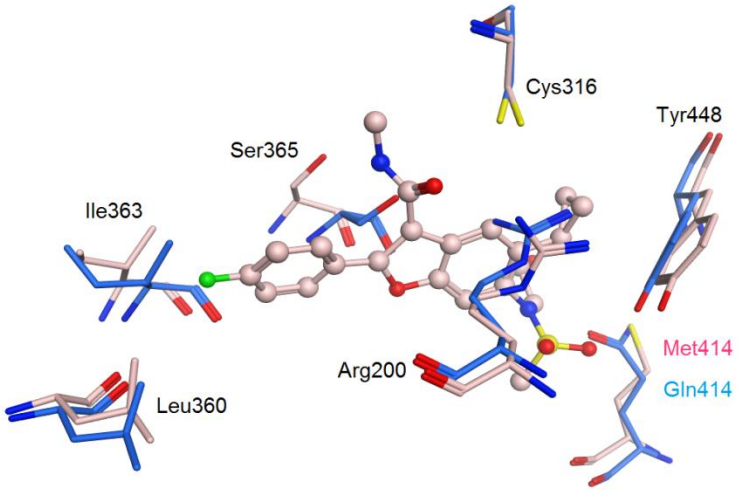


Figure 33. Superimposition of palm site II pocket of HCV polymerase genotype 1b Con1 (3FQL; pink) and genotype 2a JFH1 (3I5K; cyan). HCV796 is shown in balls and sticks.

5.7.1 MD simulations of genotype 2a HCV NS5B polymerase complexed with HCV796

The stability of the complex in the MD simulation was assessed by RMSD plots of the backbone atoms. The RMSD and interaction energies for HCV796 in both genotypes exhibit similar behavior. The binding energy in genotype 1b is -60.66 (± 3.55) kcal/mol, and -62.28 (± 4.18) kcal/mol in genotype 2a. In fact, HCV796 exhibits lower affinity genotype 2a (Table 15 and Table 16). Two common hydrogen bonds are also present in both complexes as mentioned in chapter 5.2 (Binding mode analysis of HCV796). One is a H-bond between the amide nitrogen of HCV796 and the side chain of Ser365 (41% occupancy). Another hydrogen bond is observed between the sulfonyl moiety of HCV796 and Arg200 exhibiting over 90% occupancy. To understand the binding of HCV796 in genotype 1b and 2a, energy decomposition per residue was employed to obtain a quantitative contribution of the individual forces governing the binding affinity.

The energy contribution of each amino acid residue surrounding HCV796 (4.5 Å) was computed using per-residue free energy decomposition and generalized Born solvation model implemented in the MMPBSA.py script of AMBER12. The per-residue free energy decomposition includes five terms: compose internal energy, vdW, electrostatic energy, polar solvation and non-polar solvation. All energy components were calculated using 100 snapshots taken from the 4 ns MD trajectory. Differences in the per-residue energy between the complexes of HCV polymerase

genotype 1b and genotype 2a were computed as shown in Table 21. The binding energies of both complexes from the two genotypes were similar but the energy component for each residue varied significantly. It is noticeable that the electrostatic interaction is a major energetic contribution to the binding, especially at the residues 316 and 414. At the residue 316, both genotypes have a cysteine but a subtle conformational change affects the binding affinity. At residue 414 in genotype 1b a methionine is found, whereas a glutamine is found in genotype 2a. Hence the substitution of a hydrophobic amino acid, Met, with a polar and uncharged amino acid, Gln, changes not only the physiochemical property of the binding pocket but also the shape and size (Figure 34). This result is consistent with single mutation analysis. Mutation M414T causes a 1.5 fold change in IC_{50} of HCV796 for NS5B Con1. Mutations C316N and C316Y show 3 and 50 fold changes, respectively in IC_{50} of HCV796 for NS5B Con1 (Table 16). These data supports that residue 316 and 414 significantly contribute to the binding affinity of HCV796.

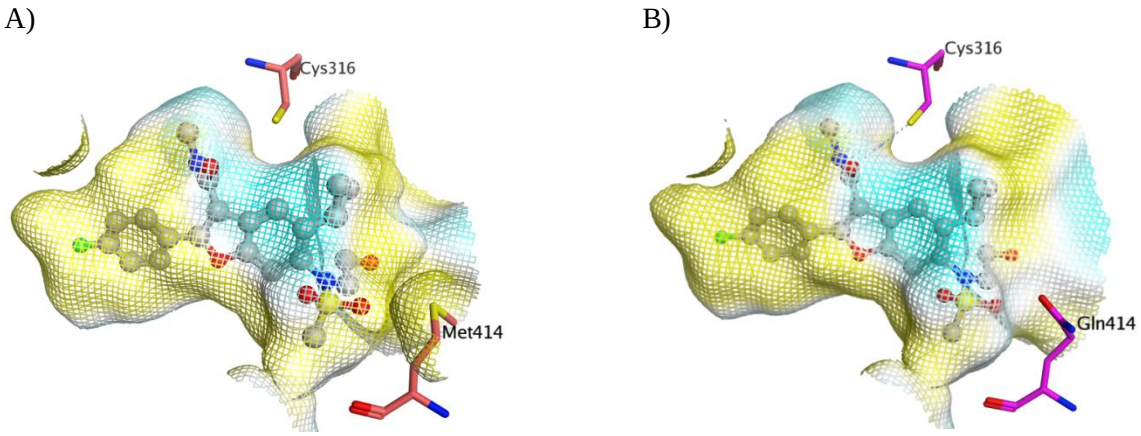


Figure 34. Contact preference map for HCV796 in A) the HCV polymerase genotype 1b (3FQL) and B) HCV polymerase genotype 2a (3I5K). The yellow contour denotes the hydrophobic preference and the blue denotes the hydrophilic preference.

Table 21. Differences of the per residue energy (kcal/mol) between the complex of HCV796 and HCV polymerase genotype (GT) 1b Con-1 and genotype 2a. Bars represent values (kcal/mol) on a scale. A blue bar denotes the per-residue energy of genotype (GT) 2a is lower than the per-residue energy of genotype 1b at a given residue position. In vice versa, a red bar denotes the per-residue energy of genotype (GT) 2a is greater than the per-residue energy of genotype 1b at a given residue position. Non-polar solvation energy is relative small and therefore it is not presented here.

Position	Sequence		Different energy between GT1b and 2a (Δ = GT1b -2a)			
	1b	2a	Internal	vdW	Electrostatic	Polar Solvation
193	F	.	0.31	-0.97	-1.75	0.88
197	P	.	0.04	-0.07	2.00	-0.12
200	R	.	-1.45	-0.19	-10.00	10.24
204	L	.	0.73	-0.90	-10.50	1.13
314	L	.	0.61	-0.14	-0.42	0.71
315	V	.	1.09	0.09	-1.69	0.51
316	C	.	8.88	-0.77	-61.37	-0.33
319	D	.	-0.46	-1.19	8.35	-5.97
320	L	.	0.18	-0.17	-0.49	0.23
321	V	.	-0.19	-0.12	0.01	-0.05
360	L	.	-0.29	-0.40	1.31	-1.23
363	I	.	0.47	-0.16	0.20	0.03
364	T	.	-0.14	-0.13	0.65	-0.04
365	S	.	-0.39	-0.30	4.59	-2.88
366	C	.	-0.01	-0.18	0.28	-0.38
367	S	.	0.11	0.05	-2.59	2.27
368	S	.	0.14	0.08	-0.42	0.56
369	N	.	-0.06	-0.21	0.47	-0.56
370	V	.	0.73	-0.49	-0.06	0.18
384	L	.	-0.17	-0.36	0.23	-0.32
414	M	Q	-6.70	-0.96	67.12	2.17
415	Y	Y	-0.04	0.39	-0.65	0.20
448	Y	.	-0.43	0.54	-0.70	-1.31
555	Y	A	6.75	-4.22	-20.10	1.26
Minimum different value			-6.7	-4.2	-61.4	-6.0
Maximum different value			8.9	0.5	67.1	10.2

5.7.2 MD simulations of the tested pharmacophore hits in complex with the genotype 2a HCV NS5B polymerase

The binding energies were calculated using MM-GB/SA and taking 100 snapshots for the last 2 ns of a 10 ns simulation. The result shows that the binding energies cannot clearly discriminate between active and inactive compounds (Table 22). (RMSD plots are shown in Figure A10-Appendix). MD simulations of the 11 tested pharmacophore hits show that the energetically preferred bound conformations were different from the docking poses. For illustration purpose, the docking poses and simulation poses of compound T3, which showed the lowest K_d value, and compound T4, whose structures share some similarity with HCV796 are shown in Figure 35. HCV796 consists of a N-methyl-1-benzofuran-3-carboxamide and T4 has a N-methyl-cyclopenta-thiophene-3-carboxamide. (Poses of the other compounds are shown in Figure A11-Appendix.) Moreover, the poses showed that the inhibitors lost the preferable hydrogen bond interactions with Ser365 and Arg200. These results might indicate why they showed no strong inhibition. However, an absence of hydrogen bond interactions was also found in T5, T8, T9 and T11, which showed an inhibition in the replicon assay.

Table 22. Binding energy of 11 tested pharmacophore hits calculated by the Generalized Born (MM-GB/SA) model using 100 frames from the MD simulation.

Compounds	MM-GB/SA (kcal/mol)		
	Average	Std. Dev.	Std. Err. of Mean
T1	-38.33	2.41	0.24
T2	-43.51	2.18	0.21
T3	-30.90	2.03	0.20
T4	-50.41	2.84	0.28
T5	-52.84	2.62	0.26
T6	-40.28	2.12	0.21
T7	-44.57	2.26	0.22
T8	-50.52	2.59	0.25
T9	-54.33	2.50	0.25
T10	-52.99	1.88	0.54
T11	-42.91	2.45	0.24

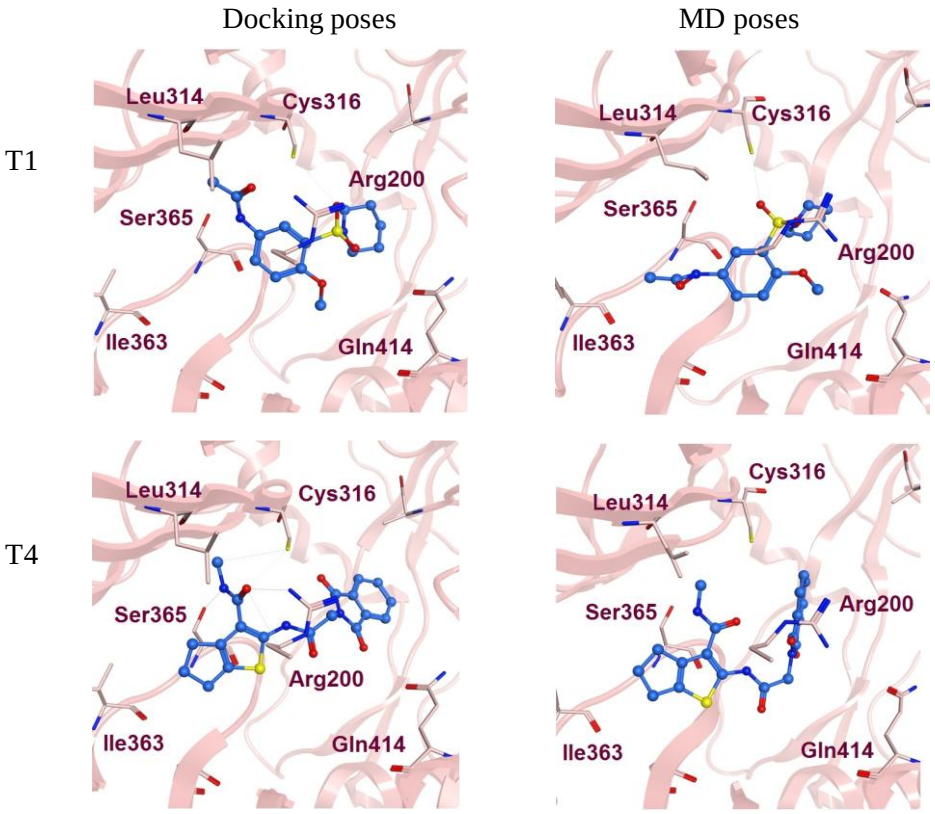


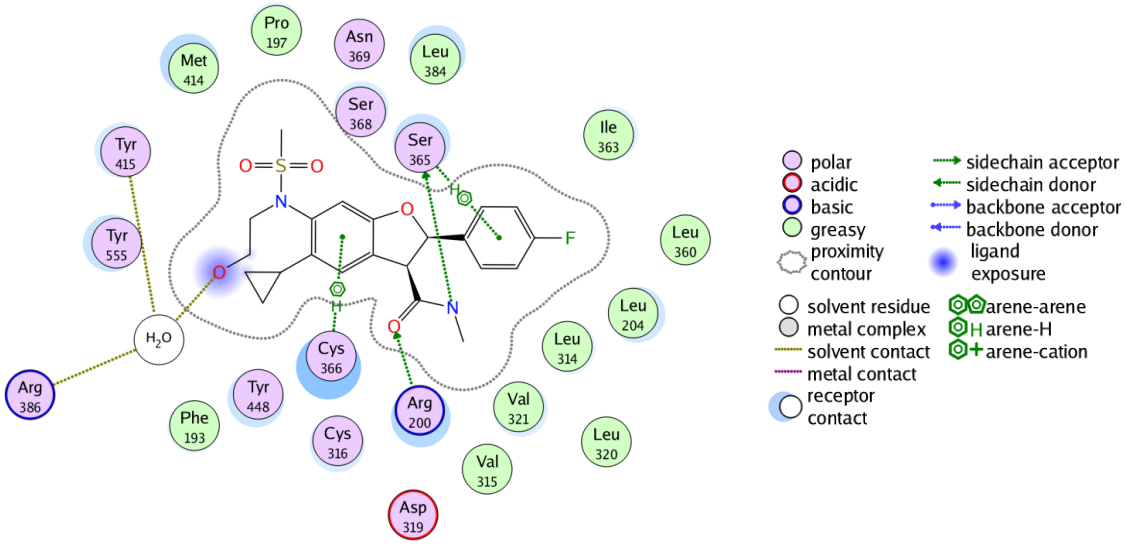
Figure 35. The docking poses and MD poses of compound T1 and T4. The MD poses were averaged over 100 snapshots from the last 2 ns of the MD simulation.

With respect to the aforementioned MD simulation of HCV796 and HCV polymerase genotype 2a, the electrostatic interaction might stabilize the binding of these compounds. The 2D protein-ligand interaction diagrams (Figure 36, Figure 37 and Figure 38) show that Cys316 and Gln414 are close to the ligand with which they can interact. Besides, there are several π-interactions found for compound T5, T9 and T10. Due to the fact that palm site II is a deep and hydrophobic pocket, the binding affinity might elicit through hydrophobic interactions. Hydrophobic interactions occur due to the close proximity between non-polar amino acid side chains of the protein and lipophilic groups of the ligand [178]. Furthermore, the logarithm of the octanol/water partition coefficient (logP) which is used to quantify hydrophobicity (Table 23) show that the putative inhibitors (T5, T8, T9 and T11) have comparatively higher logP values than the inactive compounds. Thus the results support the notion that the hydrophobic interaction is a major contribution to the binding affinity.

Table 23. The logarithm of the octanol/water partition coefficient (log P) of 11 tested pharmacophore hits and HCV796

Compound	Log P	Compound	Log P
HCV796	1.59	T6	-0.66
T1	0.91	T7	3.2
T2	1.27	T8	3.64
T3	0.85	T9	1.56
T4	1.5	T10	0.86
T5	2.68	T11	4.59

A)



B)

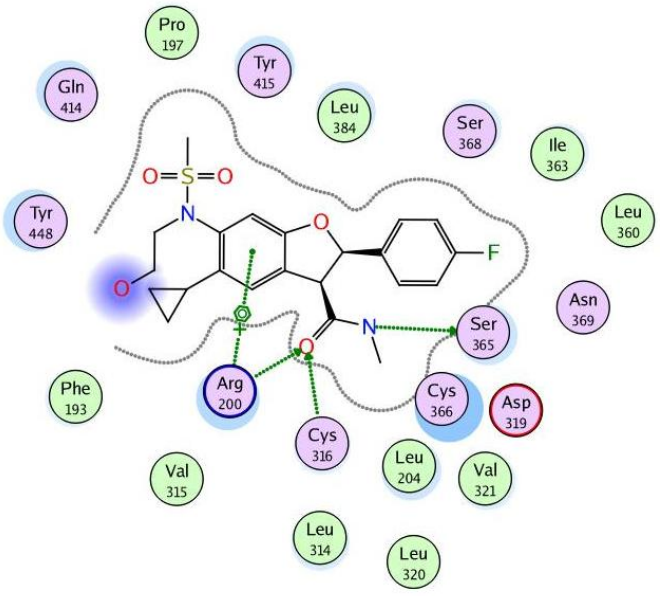


Figure 36. Protein-ligand interaction diagrams of HCV796 with A) genotype 1b [3FQL] and B) genotype 2a HCV NS5B polymerase.

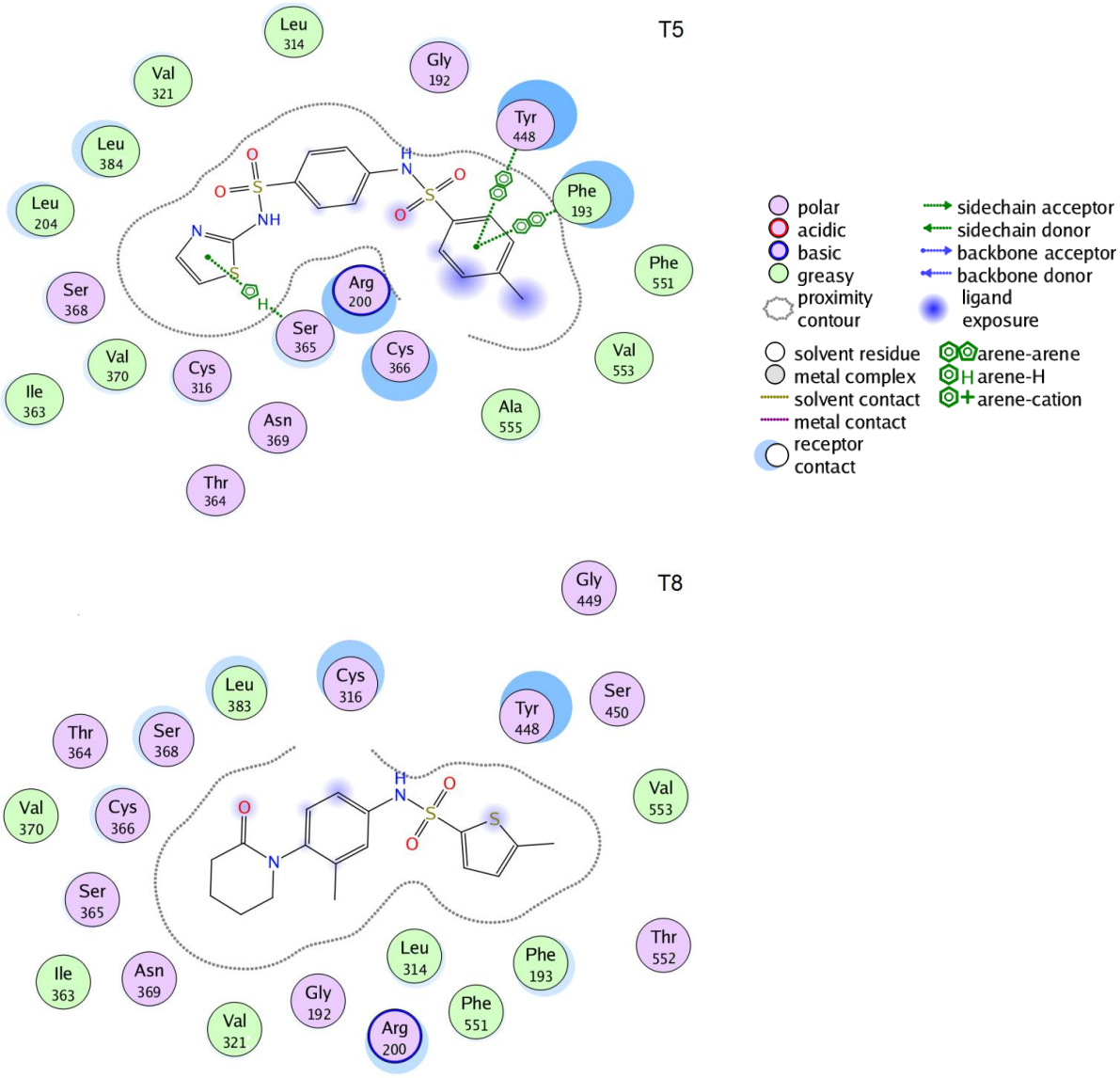


Figure 37. Protein-ligand interaction diagrams of compounds T5 and T8 with genotype 2a HCV NS5B polymerase.

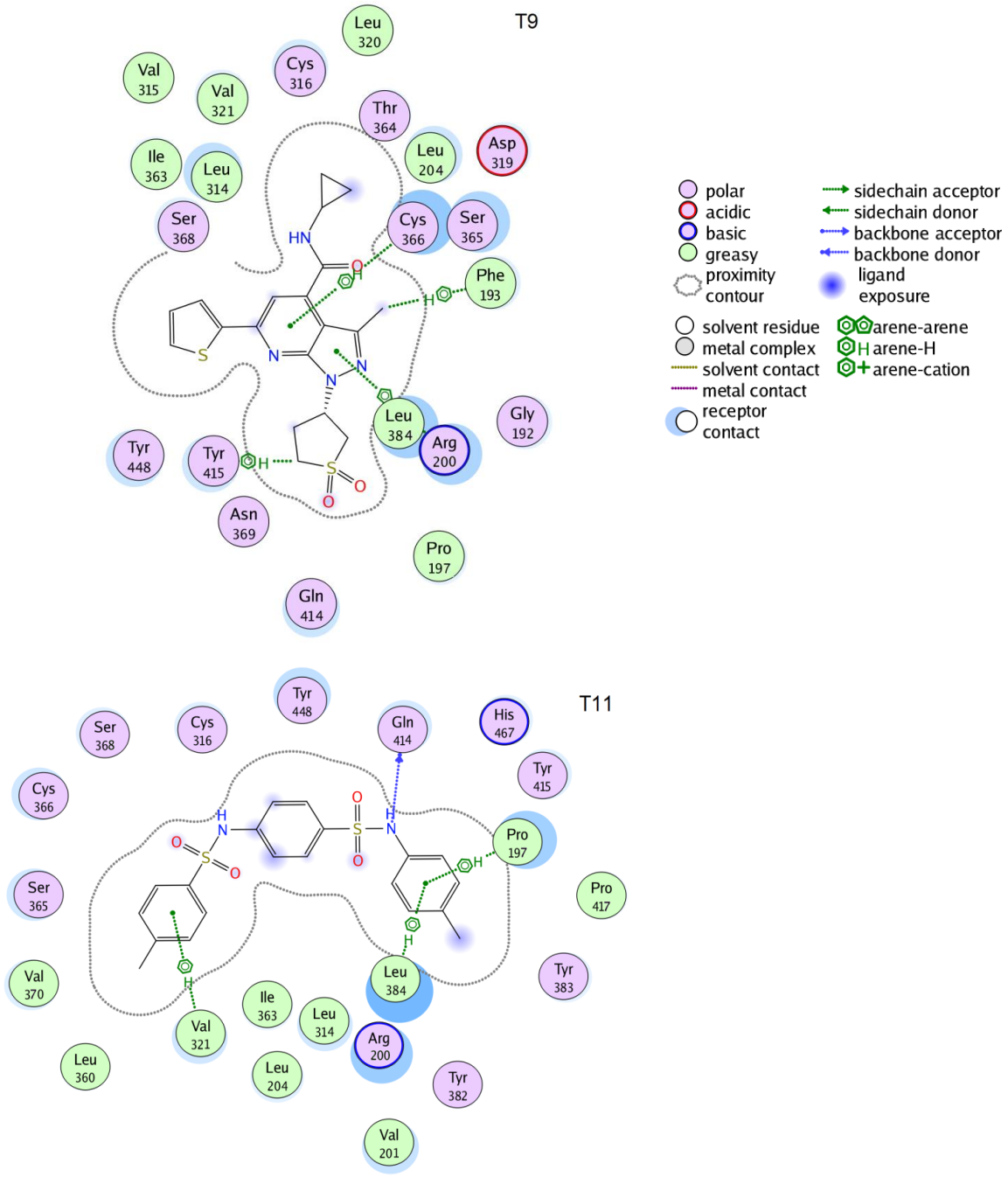


Figure 38. Protein-ligand interaction diagrams of compounds T9 and T11 with genotype 2a HCV NS5B polymerase.

5.8 West Nile virus (WNV) NS5 polymerase

Flavivirus NS5 consists of an N-terminal methyltransferase (MTase) and a C-terminal RdRp domain [179]. The architecture of the palm domain among all RNA dependent RNA polymerases (RdRp) is highly conserved [180], however they show very low sequence similarity. Sequence identity between HCV NS5B polymerase (3FQL) and WNV NS5 polymerase (2HFZ) is approximately 14%. A generated homology model of WNV NS5 polymerase (2HFZ) aligned with 3FQL is shown Figure 39. The PS-II pocket of WNV NS5 is smaller than the pocket of HCV NS5B. Key amino acids involved in binding interactions are different. For instance, Arg200, Cys316 and Met414 of HCV NS5B are in the same position as Phe194, Trp360 and Trp408 of WNV NS5. HCV NS5B inhibitors could not form hydrogen bonds with WNV NS5 and are not able to show the same binding mode as with HCV NS5B (Figure 40). It was therefore not surprising that the 11 tested compounds (dataset-II) (Table 20) which were selected based on the structure of HCV NS5B polymerase, showed K_d values for WNV NS5 polymerase higher than for HCV NS5B polymerase.

We employed an MD simulation to check the stability of the complexes of the homology model of WNV NS5 polymerase and the docked compounds. We observed that the model is not stable indicating that the homology model is not correct in all parts (Figure 41). It can be suggested, that the model is not accurate enough for docking studies because generally, 30% sequence identity is considered a threshold for successful homology modeling [181]. Due to the low sequence identity, it must be stated that the current model is not suitable for docking studies.

To the best our knowledge, there are no compounds known that inhibit both HCV and WNV polymerases. In addition, there is also no structure of an inhibitor bound to WNS NS5 polymerase available. Even though the overall structure of WNV and HCV are similar but the sequence similarity is rather low. Amino acid residues in the PS-II pocket are different. Thus if compounds can bind both HCV and WNV polymerase, their binding modes are likely to be different. Further studies and experimental data are necessary to identify potent WNV polymerase inhibitors.

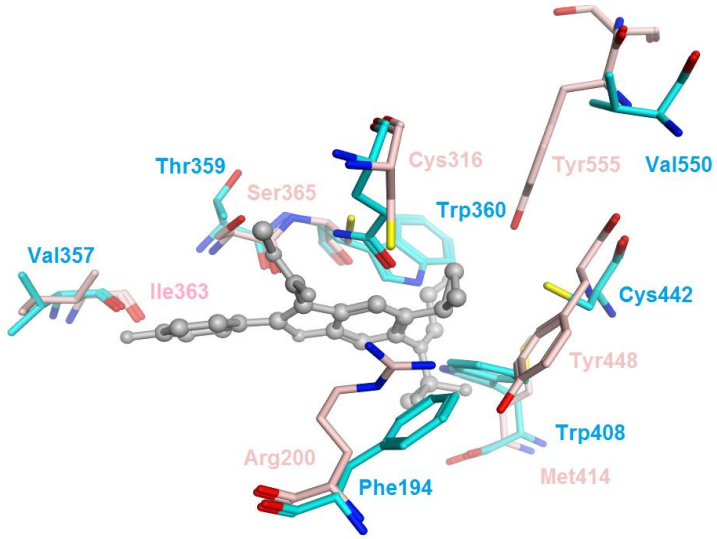


Figure 39. Superimposed structures of PS-II HCV NS5B polymerase (pink) and the homology model of WNV NS5 polymerase (blue). HCV796 is presented in grey balls and sticks.

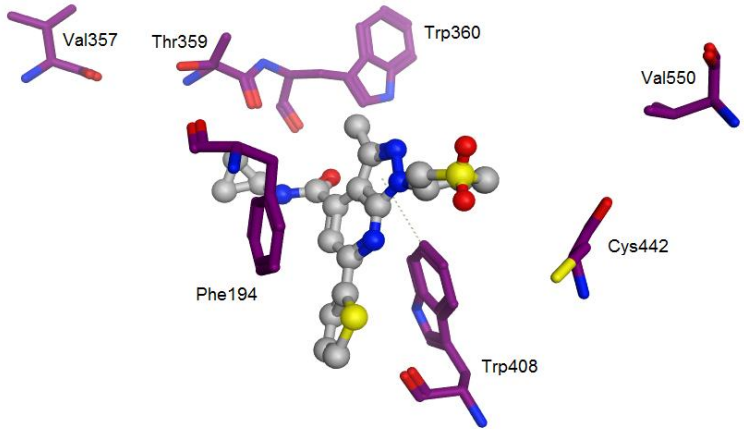


Figure 40. Bound conformation of compound T9 in the PS-II of WNV NS5 polymerase. The structure was averaged over 100 snapshots during the last 2 ns of the MD simulation. No hydrogen bods between inhibitor and NS5 are observed.

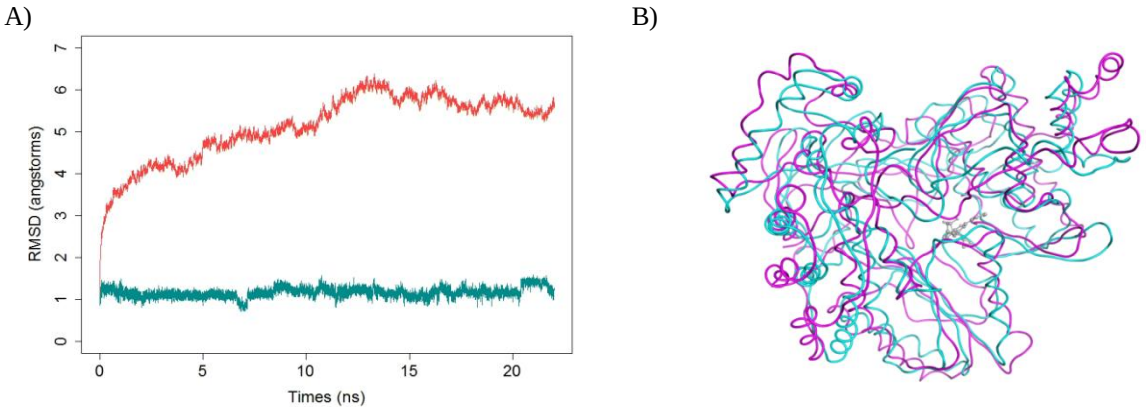


Figure 41. MD simulation of the WNV NS polymerase-T9 complex. A) Root mean square deviation (Å) of the protein (red) and compound T9 (cyan) B) superimposed structures of the homology model 3I5K (blue) and the average structures over 100 snapshots during last 2 ns (purple).

5.9 Discussion

Overall, the derived results suggest that two hydrogen bonds between the inhibitor and both Ser365 and Arg200 are essential for inhibiting HCV NS5B polymerase, and hence, the antiviral activity. The lack of the hydrogen bonds could potentially be compensated, at least partially, by favorable electrostatic and hydrophobic interactions as found for the weakly active compounds from the VS. However, given the current results, it is still unclear whether these weakly active compounds can actually bind to PS-II. Without the two hydrogen bonds, the compounds might not be able to induce the Arg200 rearrangement and access the binding pocket. In the experimental assay, these compounds show favorable binding affinity for the HCV NS5B polymerase i.e. have small K_d values, but exhibit no strong anti-HCV activity. Preliminary data showed that the IC_{50} values of T8, T9 and T11 are in the range of the K_d values. However, the compounds were not able to inhibit the HCV NS5B polymerase to 100%. As the presence of RNA and NTPs in the functional assay might affect the binding of these compounds, it can be suggested that either RNA and/or NTPs might block their binding or induces structural changes. It has been reported that the binding of RNA to NS5B polymerase does not alter the structure of the PS-II and does not affect the initial binding of HCV796 to the polymerase [169]. Thus, if their loss of activity is because of RNA blocking, it implies that the compounds should bind at the same site as RNA (catalytic site) or nearby this area such as PS-I. Long-time MD simulations might be interesting for further studies to identify additional binding sites. In a recent study, Cang and group performed microsecond MD simulation of β_2 adrenergic receptor complexed

with cholesterol. Three cholesterol-binding sites which are observed in the crystal structure are identified from the MD simulation [182]. Further experimental tests, such as a single point mutations or crystallization, are required to verify the putative binding site of the novel inhibitors.

Among the 11 tested pharmacophore hits, compound T9 was found to be the most potent PS-II inhibitor. Compound T9 exhibits anti-HCV activity equivalent to HCV796 at the concentration of 10 times its K_d value. Despite its moderate activity, the T9-HCV NS5B polymerase structural information is useful for further studies. The structure of compound T9 is relatively similar to HCV-79 (Figure 42). For example, the pyrazolopyridine in compound T9 is structurally related to the benzofuran in HCV796, and the thiophene in compound T9 is similar to the fluorobenzene in HCV796. Only a cyclopropane, which is found in HCV796, is missing in T9.

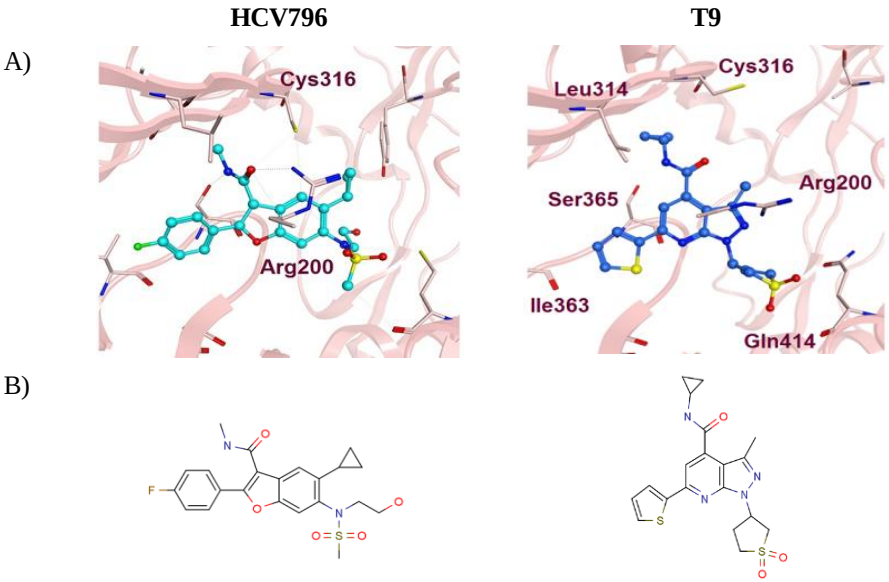


Figure 42. Structural comparison between HCV796 and compound T9. A) Bound conformation of HCV796 and T9. The bound conformation of HCV796 was obtained from the crystal structure 3FQL. The bound form of compound T9 represents the stable form observed in the MD simulation. B) 2D structure of HCV796 and compound T9.

The cyclopropane in HCV796 is located at the hydrophobic area in PS-I. Referring to our pharmacophore model (Figure 29), this cyclopropane represents HP2. Thus if a hydrophobic group is added to compound T9 at the same position as the cyclopropane of HCV796, it should

improve the affinity and activity. Structural similarity search was carried out using the ZINC database (access date: January 15, 2013) [183] based on the structure of compound T9 using 50% molecular identity. We found 5022 matching compounds. However, in most of the compounds the cyclopropane or thiophene of compound T9 is substituted by other functional groups. There is no compound that exactly shows a hydrophobic group located at the pyrazole ring as we expected.

As the generated pharmacophore model possesses this hydrophobic feature (HP2) as an optional feature in the screening, the compounds that have this feature should be included in the hit entries of the pharmacophore screening. We docked these compounds into the PS-II of HCV polymerase genotype 2a (modified 3I5K structure as stated in the MD simulation study). Then the compounds were filtered using the pharmacophore model (HD, HA1, HP1 and HP2). Unfortunately, we could not find any compound that shows all these features. Thus, only chemical derivatives could be used to end up with the desired compounds.

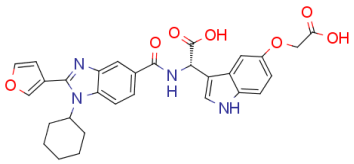
Regarding the results of the energy contribution per-residue, the electrostatic energy, especially at position 316 and 414 was found to be important for binding affinity. Using electrostatic similarity search such as EON (a commercial program for an electrostatic comparison) in the post-processing of docking results might help to identify novel inhibitors. Besides computational methods, experimental testing in HCV polymerase genotype 1 might be interesting to validate the data, especially for compound T9. Moreover, we still lack any knowledge of known PS-II inhibitors and potential binding affinities in different genotypes. This information could provide useful insight into the binding mechanism and can be used to improve the virtual screening.

Chapter 6 Analysis of the resistance of Hepatitis C virus NS5B polymerase via docking and molecular dynamics simulation

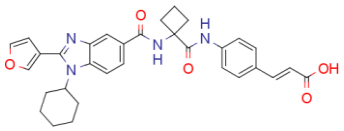
6.1 Introduction

Due to the rapid replication rate along with the lack of a proofreading mechanism in the viral polymerase, viral mutations occur spontaneously over time. This mutation results in numerous distinct genotypes [13] and quasispecies [11] as well as cause escape mutants from antiviral agents. The variants show a decreased binding affinity to inhibitors, while retaining the enzyme's activity necessary for viral replication. Several point mutations in NS5B polymerase have been reported which give rise to different levels of resistance [52]. Thus, it is desirable to account for resistant mutations in the early stage of drug design which might help to avoid therapy failure.

In order to improve the inhibitory potency of the compounds, information on inhibitor-enzyme interaction would be valuable. The present study was carried out on the reported point mutations Pro495, Pro496 and Val499 of HCV NS5B polymerase which show reduced affinity towards two benzimidazole-5-carboxamide inhibitors, namely compound **A** and compound **B** (Figure 43 and Table 24) [184]. These derivatives represent a promising group of inhibitors since they possess comparable potencies against cross-genotypes (genotype 1,3,4,5 and 6) [47] [185]. They bind at thumb site I and induce a conformational change resulting in an inactive protein conformation (closed form) [51]. Mutation of Pro495 showed a pronounced reduction in the affinities of both compounds; in fact the activity of the compounds against the mutated enzyme was almost abolished. Meanwhile Pro496 and Val499 mutations also resulted in a decrease in the activity of the compounds, however to a much lesser extent than observed for the Pro495 mutation. Docking and molecular dynamics (MD) simulation were applied to elucidate the probable binding mode of the inhibitors. Then a combined Molecular Mechanics-Generalized Born Surface Area (MM-GB/SA) calculation was employed on wild-type and mutant enzymes in complex with the inhibitors to examine the molecular interactions in detail.



compound **A**



compound **B**

Figure 43. Structures of the two benzimidazole-5-carboxamides.

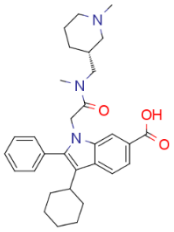


Figure 44. Structure of a co-crystalized ligand NS5B inhibitor (PDB: 2BRL).

6.2 Methods

As the crystal structures of NS5B with compound **A** and compound **B** are not solved, the crystal structure of HCV polymerase genotype 1b bound to an indole derivative (Figure 44), (PDB ID: 2BRL[38]), which shows structural similarity to the studied compounds, was utilized to prepare the protein-inhibitor complexes. The benzimidazole derivatives bind to the thumb domain, which normally interacts with a finger loop and modulates the protein conformation. This finger loop is highly flexible, and is not completely resolved in the chosen crystal structure [186]. The missing residues of the finger loop (∧1 loop) were therefore modeled taking the NS5B crystal structure 2FVC [187] as template. The generated model was subsequently minimized and equilibrated using the Amber 12 package [151]. The model served as structure of the wild type (WT) enzyme. To investigate how mutations affect the binding, models of P495S/L/A, P496A/S and V499A mutants were built using the prepared WT structure as initial configuration. The single point mutations were prepared by using the Molecular Operating Environment (MOE) 2012.10 [161]. The structures of the studied inhibitors were prepared by MOE before docking into the prepared HCV polymerase structures using GLIDE SP [116]. The docked poses were further investigated by MD simulation and energy decomposition using Amber 12.

Table 24. Inhibition of compound A and compound B against HCV NS5B polymerase containing resistance mutations

Enzyme	IC_{50} (µM)	
	compound **A**	compound **B**
Wild type 1b	0.15 (± 0.042)	0.27 (±0.05)
P495S	>25	>32
P495L	>25	>25
P495A	>25	10.6 (±3.2)
P496A	0.52 (±0.16)	2.6 (±0.4)
P496S	0.53 (±0.22)	3.8 (±2.1)
V499A	0.41 (±0.12)	0.63 (±0.15)

6.3 Results and discussion

The docking results showed that the central scaffold of the studied inhibitors, including the benzimidazole, the furan and the cyclohexyl rings, fills the thumb site I pocket and forms mainly hydrophobic interactions with the enzyme (Figure 45). In addition, a hydrogen bond between the amide carbonyl group of the inhibitors and Arg503 is observed. The groups attached to the N-carboxamide moiety of the inhibitors are completely solvent exposed. Arene-H interaction of the benzimidazole ring with Pro495 was also detected in the docking poses of both compounds (Figure 46). Poses obtained from docking into both wild-type (WT) and mutant proteins were similar. Visual inspection of these poses suggested that the lack of the arene interaction between Pro495 and the inhibitor in the Pro495 mutation might affect the inhibitor affinity.

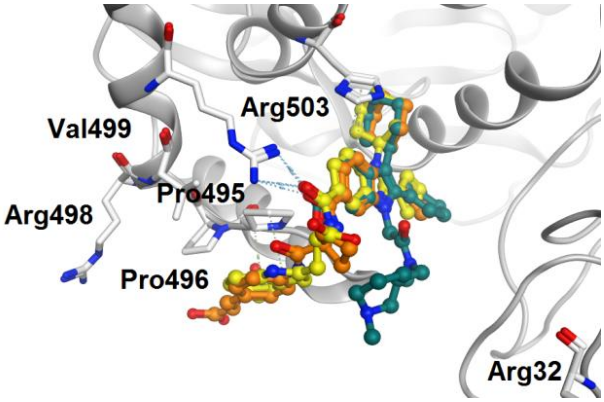


Figure 45. Bound structures of three inhibitors (balls and sticks)--compound **A** (yellow), compound **B** (orange) and the co-crystallized indole derivative of 2BRL (cyan) -- in the binding site of wild-type HCV NS5B polymerase.

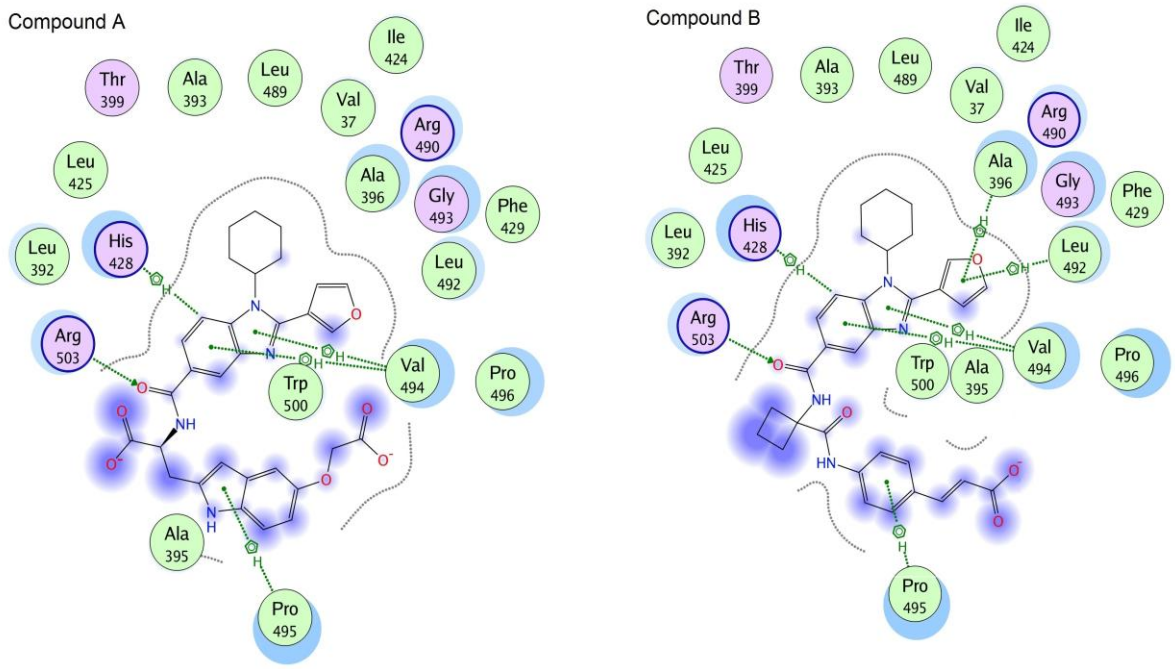


Figure 46. 2D representation of the predicted docking pose of compound **A** and compound **B** within thumb site I of the wild-type HCV NS5B polymerase. Hydrophobic amino acids are colored green, polar residues are colored pink. Hydrogen bonds and arene-H interactions are indicated by dashed lines. Ligand and protein solvent exposure is indicated by the blue spheres.

For Pro496 and Val499 no direct interaction with the inhibitors were observed, which might explain the similar inhibitor activities for Pro496 and Val499 mutants. To get an insight into the energetics of the mutation-induced change in the binding affinities of benzimidazole-inhibitors, we analyzed the binding free energy for wild-type and mutants. This was done by performing MD simulation using the AMBER 12 package [151]. The alpha carbon Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuations (RMSF) of the backbone atoms compared to the starting structures of all simulated systems were monitored (Figure A12 and A13-Appendix). The RMSF of the backbone atoms showed that the mutation does not perturb the overall enzyme structure. However a local geometry change affects the inhibitor binding as shown by the average structure taken from the snapshots of the MD simulations (Figure 47). These showed that the binding mode of the compounds in the wild-type enzyme was slightly changed during MD compared to the obtained docking poses. The finger loop moved closer to the ligand, which enabled Arg32 to form interactions with the ligand as shown in Figure 6. Moreover compound **B** showed an additional hydrogen bond with Arg498 (Figure 47).

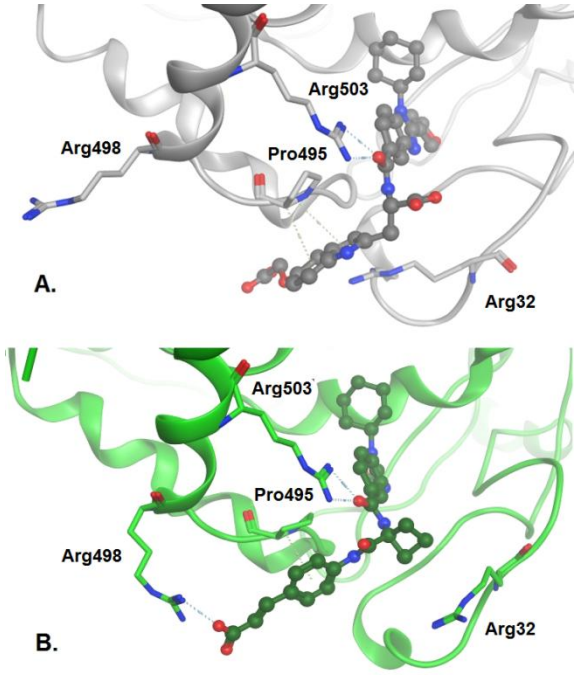


Figure 47. Average structures of the inhibitor bound to wild type HCV NS5B polymerase obtained from snapshots of the MD simulations. A) compound **A** and B) compound **B** are shown in ball-and-stick mode.

Analysis of the hydrogen bond occupancy throughout the MD run revealed that the mutation of Pro495 with serine, leucine or alanine disrupted the stability of the hydrogen bond formed between the adjacent residues Arg503 andArg498 with the inhibitors, as can be seen from snapshots taken from the MD simulation. For instance, the percentage occupation of the hydrogen bond between compound **A** and Arg503 was reduced from 84% (WT) to 41-43 % in the P495S/A/L mutants. While the hydrogen bond between compound **B** and Arg503 in WT-complex exhibited 48% occupancy, and was reduced to 18% in the P495L mutant, 12% in P495A and to 0% in the P495S mutant.

The mutations P496A and P496S mainly affected the stability of the hydrogen bond formed between compound **B** and Arg498, with no effect on the hydrogen bond occupancy between the inhibitors and Arg503. The hydrogen bond formed between Arg498 and compound **B** showed a noticeable reduction from 65% occupancy to 24% in the P496A mutant and 53% in the P496S mutants. Meanwhile, the occupancy of hydrogen bonding between compound **B** and Arg503 were 49% and 51% hydrogen bond occupancy in the P496A and P496S mutants, respectively, compared to 48% in the WT. This might explain why the binding of compound **A**, which did not

form a hydrogen bond with Arg498, was much less affected by the mutation at residue 496 than compound **B.**

Subsequently, 100 snapshots for each complex were extracted from the last 2 ns of the MD simulation for binding energy calculation. The calculated binding energies using the MM-GB/SA method were not able to completely explain the differences between wild-type and mutants (Table A11-Appendix). We then performed a per-residue decomposition of the binding free energy to analyses the contribution of the individual residues. The per-residue energy decomposition includes internal energy, van der Waals (vdW), electrostatic energy, polar solvation and non-polar solvation. All energy components were calculated using 100 snapshots taken from the last 2 ns of the MD trajectory. The calculated per-residue energy decompositions indicated that electrostatic and polar solvation energies contribute highly to the binding affinities (Table A12-Appendix). The vdW contribution for all the mutants changed relative to the WT from -3 to 2 kcal/mol, whereas the difference of per-residue contributions were between -16 to 29 kcal/mol for the electrostatic energy and -23 to 15 kcal/mol for the polar solvation energy. Internal energies and non-polar solvation energies were relatively small compared to other energy components.

We observed that the energy difference of each residue contribution for the mutants relative to the WT did not occur only at the mutation point but also the neighboring residues. Table 25 shows that the key contributors to the change in binding affinity between the WT and the mutants were residues 32, 495, 498 and 503. The dominant energy changes were at residue 495 in compound **A** complexes, and residues 495 and 498 in compound **B** complexes. This implies that the hydrogen bonding with Arg498 is important for the inhibitory activity of compound **B**.

By comparing the per-residue contribution profile of compound **A** complexes at the point mutations, only the mutations P495S/L/A caused a noticeable difference in the total energy contribution relative to the WT. The contribution to the binding energy was generally less favorable when compared to the WT. Meanwhile P496A/S and V499A mutations showed only a small change. This result is in agreement with the experimental data where only the mutations P495S/L/A cause a loss in the inhibitory activity of the compounds (Table 24). The weaker activity of compound B against P496A/S mutants compared to compound A, could however not be explained.

Table 25. Mean total per-residue energy (kcal/mol) of the key amino acids in the binding pocket of the wild-type (WT) and mutant enzymes with compound A and compound B. The grey and red background shading represents the energy difference of the residue contributions in the mutants that are between 0.5-1 kcal/mol and greater than 1 kcal/mol relative to the wild-type, respectively. The boxes show the point mutations.

Position	Mean total per-residue energy (kcal/mol)													
	Compound **A**							Compound **B**						
	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A
32	-0.23	-3.11	-0.07	-2.31	-0.13	-0.15	-4.05	-0.43	0.00	-0.36	-0.45	-0.08	-0.16	0.05
396	-1.36	-1.23	-1.31	-1.61	-1.13	-1.32	-1.56	-1.51	-1.37	-1.61	-1.74	-1.12	-1.14	-1.35
428	-0.07	0.03	0.15	0.12	0.28	0.07	-0.20	-0.13	0.19	-0.07	-0.91	0.10	0.26	-0.12
492	-0.38	-0.21	-1.05	-0.45	-1.85	-0.76	-0.05	-0.92	-1.08	-0.98	-0.93	-0.52	-0.96	-1.05
493	-0.36	-0.61	-0.09	0.33	-0.54	-0.25	-0.50	-0.12	-0.33	0.00	-0.05	-0.06	-0.28	-0.34
494	-1.71	-2.31	-1.85	-0.78	-3.23	-4.69	-2.62	-2.34	-2.32	-1.97	-2.30	-2.25	-1.53	-1.98
495	-3.56	-2.02	-2.30	-1.42	-3.11	-3.36	-4.18	-3.31	-1.22	-3.40	-1.71	-3.01	-3.13	-2.98
496	-1.04	-0.74	-0.32	-0.27	-0.02	-0.45	-0.87	-1.31	-1.26	-1.11	-1.61	-0.36	-2.22	-1.53
497	-0.02	-0.02	-0.01	-0.02	0.01	-0.02	-0.03	-0.05	-0.06	-0.06	-0.06	-0.06	-0.08	-0.04
498	-0.04	-0.04	-0.02	-0.04	-0.04	-0.02	-0.03	-3.44	-3.77	-4.04	-3.70	-4.33	-4.44	-4.53
499	-0.27	-0.30	-0.05	-0.08	-0.13	-0.14	-0.07	-0.99	-1.03	-1.01	-1.05	-1.33	-1.29	-0.79
500	-0.96	-1.28	-0.75	-0.74	-1.26	-0.94	-1.10	-1.24	-1.11	-1.16	-1.19	-1.53	-1.15	-1.49
501	-0.04	-0.05	-0.03	-0.02	-0.04	-0.03	-0.03	-0.08	-0.06	-0.06	-0.07	-0.11	-0.08	-0.08
502	0.01	0.00	0.01	0.01	-0.01	0.00	0.00	0.00	0.00	0.01	0.01	-0.02	0.00	-0.02
503	-1.50	-3.16	-0.40	-0.05	-2.50	-2.09	-0.87	0.03	0.09	-0.02	-0.20	-0.12	-0.82	0.13

In this study, we employed docking and MD simulation to understand the mechanism of resistance of NSB-polymerase mutants for two benzimidazole inhibitors. MD simulations consider also the protein flexibility and provide a more realistic picture of the binding process than docking. MM-GB/SA data that are computed with neglecting entropy has so far

demonstrated an improvement in the predictive aspect [143, 148], but in the present study it was not successful in distinguishing between the binding to wild-type and mutant proteins.

It is possible that the entropic contribution could improve the activity prediction for ranking compounds according to their activity profile. However, the identification of key residues that contribute to the inhibitor binding is crucial for designing new compounds. So instead of calculating the entropy term which is computationally expensive, we employed per-residue energy decomposition, structural comparison of the WT and the mutant complexes, and hydrogen bond analysis to provide an insight into the drug resistance mechanism.

The energy difference of each residue contribution for the mutants relative to the WT reveals the key residues responsible for the inhibitory activity. Besides the point mutation at Pro495, Pro496 and Val499, the neighboring residues Arg32, Arg498 and Arg503 also showed larger differences compared to the WT. This suggests that if mutation or polymorphism in different genotypes occurs at residue 32, 498 or 503, the activity of both compound **A** and compound **B** will be affected. However, the chance that the virus escapes by developing a point mutation at these residues might be low. Arg32 from the finger loop and Pro495, Pro496, Val499 and Arg503 from the thumb domain are namely part of the allosteric GTP binding pocket [33], which regulates the inter-domain interaction during the conformational change of the enzyme necessary for HCV RNA replication *in vivo* but not *in vitro* [188]. Mutations of the residues defining this allosteric GTP binding site cause a significantly lower or even a complete loss of HCV RNA replication [188].

Drug resistance and genotype variation represents a major obstacle to successful drug design. Free energy decomposition analysis has been showed to be effective method to reveal atomic-level understanding of resistance mechanism [189, 190]. We showed that using per-residue energy contribution can help to elucidate the residues that make a major contribution to the inhibitor binding and to predict the outcome of the inhibitors.

Chapter 7 Conclusions

The different approaches and methods used in the current work were applied to develop appropriate virtual screening protocols in order to identify novel inhibitors of HCV NS5B polymerase. Additionally, molecular docking, MD simulations and energy decomposition studies were carried out to assess the effect of mutations on the inhibitory potency of some selected inhibitors.

Molecular docking was employed as main method for VS. The herein applied docking procedures gave satisfactory results in predicting the bioactive conformation of the studied compounds. Docking was able to predict the experimentally observed poses of about 60% of the compounds. However, all employed docking procedures failed to correctly rank the studied compounds according to their biological activity; a problem which is often encountered with docking. Using docking constraints might improve the pose prediction, but it would confine the VS to a specific binding mode, which might limit the structural diversity of the hits. Hence, the combination of docking methods with other techniques, such as machine-learning methods, was explored to improve the hit identification.

The best performance in pose prediction obtained during the course of this work was 69%, which was achieved by using Gold-score and the crystal structure 3GNV. Docking was carried out without considering water molecules, since some of the conserved water molecules could be displaced by the inhibitors, as seen in some crystal structures. However, a recent study by Barreca et al, [82] showed that considering water molecules in docking might improve the docking performance (both pose-prediction and ranking) for PS-I inhibitors, especially those which undergo water-mediated interactions with the protein. A direct comparison of the herein obtained results with the results of Barreca et al is not possible, since the number of the studied inhibitors is different (40 in the study of Barreca et al and 29 in the present work). Despite the satisfactory results, which were obtained in this work, it would be interesting to further study the effect of water molecules by using different scoring functions for example Gold-score, which yielded the best performance in the present work.

In a trial to improve the ability of the procedure to discriminate between active and inactive compounds, a combination of docking with several methods was investigated. Binding free energy calculations using MM-GBSA failed to improve the ranking and no correlation could be

obtained between the experimental activity and the calculated binding free energy, even when the entropic term was approximated.

A machine learning method was also explored. RF showed the best results and was able to improve the ability to discriminate between actives and inactives. The RF models which were developed for post-docking filtration of docking poses, showed promising results in identifying active compounds.

Two sites docking, i.e. parallel docking in PS-I and TS-II sites, also gave positive results, which were comparable to RF. Hence a procedure which combines the selected RF model with two sites docking was suggested to improve the hit identification. Consequently, the ChemBridge database was screened using the suggested protocol and a total of 28 compounds were suggested as potential actives. Even though the proposed compounds have not yet been tested to ensure the success of the protocol, the binding modes of the proposed compounds are similar to known inhibitors, which imply the efficiency of this protocol. In the two sites docking study, PS-I was chosen as target binding site and TS-II as dummy site. However, the top ranked compounds in TS-II, which are considered as false positives of PS-I, might possibly be putative TS-II inhibitors. This method thus can give potential inhibitors targeting both binding sites at the same time.

Another part of the current work was to explain the effect of resistance mutations on the inhibitory potency of two benzimidazole inhibitors. MM-GBSA calculations failed to show any difference between binding to wild-type and mutant enzymes. However, using per-residue decomposition energy helped to define the key amino acid residues involved in protein-ligand interaction. It also was useful in explaining the binding affinity difference that occurs in mutations and genotype variations. This information would be useful for further design of inhibitors that can overcome resistance or are potent across different genotypes.

References

1. Centers for Disease Control and Prevention. *Viral Hepatitis*. 2011 May 12, 2011 [cited 2011 June 1]; Available from: www.cdc.gov/hepatitis

2. World Health Organization. *Hepatitis E Fact sheet N°280*. January 2005 [cited 2011 October 12]; Available from: http://www.who.int/mediacentre/factsheets/fs328/en/index.html.

3. World Health Organization. *Hepatitis A Fact sheet N°328*. May 2008 [cited 2011 October 12]; Available from: http://www.who.int/mediacentre/factsheets/fs328/en/index.html.

4. Organization, W.H. *Hepatitis B Fact sheet N°204*. 2013 July 2013 [cited 2013 September 20]; Available from: http://www.who.int/mediacentre/factsheets/fs204/en/.

5. World Health Organization. *Hepatitis C*. June 1, 2011 [cited 2013 September 20]; Available from: http://www.who.int/entity/csr/disease/hepatitis/Hepc.pdf.

6. Marcellin, P. and T. Asselah, *Viral hepatitis: impressive advances but still a long way to eradication of the disease.* Liver international : official journal of the International Association for the Study of the Liver, 2014. **34 Suppl 1**: p. 1-3.

7. Afdhal, N.H., *The natural history of hepatitis C.* Seminars in liver disease, 2004. **24 Suppl 2**: p. 3-8.

8. Casey, L.C. and W.M. Lee, *Hepatitis C virus therapy update 2013.* Current opinion in gastroenterology, 2013. **29**(3): p. 243-9.

9. Simmonds, P., et al., *Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes.* Hepatology, 2005. **42**(4): p. 962-73.

10. Kuiken, C. and P. Simmonds, *Nomenclature and numbering of the hepatitis C virus.* Methods in molecular biology, 2009. **510**: p. 33-53.

11. Bukh, J., R.H. Miller, and R.H. Purcell, *Genetic heterogeneity of hepatitis C virus: quasispecies and genotypes.* Seminars in liver disease, 1995. **15**(1): p. 41-63.

12. NSW, H. *Hepatitis C factsheets Genotypes*. 2010 [cited 2011 October 7]; Available from: www.hep.org.au/documents/factsheets/Genotypes2010.pdf.

13. World Health Organization. *key Global distribution of HCV genotypes*. 2009 [cited 2011 October 07]; Available from: www.who.int/vaccine_research/documents/ViralCancer7.pdf.

14. Robertson, B., et al., *Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. International Committee on Virus Taxonomy.* Archives of virology, 1998. **143**(12): p. 2493-503.

15. Clarke, D., et al., *Evidence for the formation of a heptameric ion channel complex by the hepatitis C virus p7 protein in vitro.* The Journal of biological chemistry, 2006. **281**(48): p. 37057-68.

16. Yamanaka, T., M. Uchida, and T. Doi, *Innate form of HCV core protein plays an important role in the localization and the function of HCV core protein.* Biochemical and biophysical research communications, 2002. **294**(3): p. 521-7.

17. McGarvey, M.J. and M. Houghton, *Structure and Molecular Virology*, in *Viral Hepatitis*, H. Thomas, S. Lemon, and A. Zuckerman, Editors. 2007, Blackwell Publishing Ltd: Oxford, UK.

18. Lange, C.M., C. Sarrazin, and S. Zeuzem, *Review article: specifically targeted anti-viral therapy for hepatitis C - a new era in therapy.* Alimentary pharmacology & therapeutics, 2010. **32**(1): p. 14-28.

19. Kiser, J.J. and C. Flexner, *Direct-acting antiviral agents for hepatitis C virus infection.* Annual review of pharmacology and toxicology, 2013. **53**: p. 427-49.

20. *EASL Clinical Practice Guidelines: management of hepatitis C virus infection.* Journal of hepatology, 2011. **55**(2): p. 245-64.

21. Antaki, N., et al., *The neglected hepatitis C virus genotypes 4, 5 and 6: an international consensus report.* Journal of the International Association for the Study of the Liver, 2010. **30**(3): p. 342-55.

22. Marciano, S. and A.C. Gadano, *How to optimize current treatment of genotype 2 hepatitis C virus infection.* Liver international : official journal of the International Association for the Study of the Liver, 2014. **34 Suppl 1**: p. 13-7.

23. Dugum, M. and R. O'Shea, *Hepatitis C virus: Here comes all-oral treatment.* Cleveland Clinic journal of medicine, 2014. **81**(3): p. 159-72.

24. Schneider, M.D. and C. Sarrazin, *Antiviral therapy of hepatitis C in 2014: Do we need resistance testing?* Antiviral research, 2014. **105C**: p. 64-71.

25. Serfaty, L., *Is there still a role for PEG IFN+RBV therapy in patients with HCV genotype 1?* Liver international : official journal of the International Association for the Study of the Liver, 2014. **34 Suppl 1**: p. 11-2.

26. Sulkowski, M.S., et al., *Daclatasvir plus sofosbuvir for previously treated or untreated chronic HCV infection.* The New England journal of medicine, 2014. **370**(3): p. 211-21.

27. Lohmann, V., et al., *Biochemical properties of hepatitis C virus NS5B RNA-dependent RNA polymerase and identification of amino acid sequence motifs essential for enzymatic activity.* Journal of virology, 1997. **71**(11): p. 8416-28.

28. Hagedorn, C.H., E.H. van Beers, and C. De Staercke, *Hepatitis C virus RNA-dependent RNA polymerase (NS5B polymerase).* Current topics in microbiology and immunology, 2000. **242**: p. 225-60.

29. Kutay, U., E. Hartmann, and T.A. Rapoport, *A class of membrane proteins with a C-terminal anchor.* Trends in cell biology, 1993. **3**(3): p. 72-5.

30. Yamashita, T., et al., *RNA-dependent RNA polymerase activity of the soluble recombinant hepatitis C virus NS5B protein truncated at the C-terminal region.* The Journal of biological chemistry, 1998. **273**(25): p. 15479-86.

31. Lee, K.J., et al., *The C-terminal transmembrane domain of hepatitis C virus (HCV) RNA polymerase is essential for HCV replication in vivo.* Journal of virology, 2004. **78**(7): p. 3797-802.

32. Uchiyama, Y., et al., *Measurement of HCV RdRp activity with C-terminal 21 aa truncated NS5b protein: optimization of assay conditions.* Hepatology research : the official journal of the Japan Society of Hepatology, 2002. **23**(2): p. 90-97.

33. Bressanelli, S., et al., *Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides.* Journal of virology, 2002. **76**(7): p. 3482-92.

34. Chinnaswamy, S., et al., *A locking mechanism regulates RNA synthesis and host protein interaction by the hepatitis C virus polymerase.* The Journal of biological chemistry, 2008. **283**(29): p. 20535-46.

35. Simister, P., et al., *Structural and functional analysis of hepatitis C virus strain JFH1 polymerase.* J Virol, 2009. **83**(22): p. 11926-39.

36. Biswal, B.K., et al., *Crystal structures of the RNA-dependent RNA polymerase genotype 2a of hepatitis C virus reveal two conformations and suggest mechanisms of inhibition by non-nucleoside inhibitors.* The Journal of biological chemistry, 2005. **280**(18): p. 18202-10.

37. Yi, G., et al., *Biochemical study of the comparative inhibition of hepatitis C virus RNA polymerase by VX-222 and filibuvir.* Antimicrobial agents and chemotherapy, 2012. **56**(2): p. 830-7.

38. Di Marco, S., et al., *Interdomain communication in hepatitis C virus polymerase abolished by small molecule inhibitors bound to a novel allosteric site.* The Journal of biological chemistry, 2005. **280**(33): p. 29765-70.

39. Blight, K.J., A.A. Kolykhalov, and C.M. Rice, *Efficient initiation of HCV RNA replication in cell culture.* Science, 2000. **290**(5498): p. 1972-4.

40. Bukh, J., et al., *Mutations that permit efficient replication of hepatitis C virus RNA in Huh-7 cells prevent productive replication in chimpanzees.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(22): p. 14416-21.

41. *Hepatitis C Viruses: Genomes and Molecular Biology*, S.-L. Tan, Editor 2006, Horizon Bioscience: Norfolk (UK).

42. Ferrari, E., et al., *Characterization of soluble hepatitis C virus RNA-dependent RNA polymerase expressed in Escherichia coli.* Journal of virology, 1999. **73**(2): p. 1649-54.

43. Kato, T., et al., *Efficient replication of the genotype 2a hepatitis C virus subgenomic replicon.* Gastroenterology, 2003. **125**(6): p. 1808-17.

44. Wakita, T., et al., *Production of infectious hepatitis C virus in tissue culture from a cloned viral genome.* Nature medicine, 2005. **11**(7): p. 791-6.

45. Barreca, M.L., et al., *Allosteric inhibition of the hepatitis C virus NS5B polymerase: in silico strategies for drug discovery and development.* Future medicinal chemistry, 2011. **3**(8): p. 1027-55.

46. Beaulieu, P.L., *Finger loop inhibitors of the HCV NS5B polymerase: discovery and prospects for new HCV therapy.* Current opinion in drug discovery & development, 2006. **9**(5): p. 618-26.

47. Hirashima, S., et al., *Benzimidazole derivatives bearing substituted biphenyls as hepatitis C virus NS5B RNA-dependent RNA polymerase inhibitors: structure-activity relationship studies and identification of a potent and highly selective inhibitor JTK-109.* Journal of medicinal chemistry, 2006. **49**(15): p. 4721-36.

48. Harper, S., et al., *Potent inhibitors of subgenomic hepatitis C virus RNA replication through optimization of indole-N-acetamide allosteric inhibitors of the viral NS5B polymerase.* Journal of medicinal chemistry, 2005. **48**(14): p. 4547-57.

49. Vendeville, S., et al., *Finger loop inhibitors of the HCV NS5b polymerase. Part II. Optimization of tetracyclic indole-based macrocycle leading to the discovery of TMC647055.* Bioorganic & medicinal chemistry letters, 2012. **22**(13): p. 4437-43.

50. Beaulieu, P.L., et al., *Improved replicon cellular activity of non-nucleoside allosteric inhibitors of HCV NS5B polymerase: from benzimidazole to indole scaffolds.* Bioorganic & medicinal chemistry letters, 2006. **16**(19): p. 4987-93.

51. Li, H. and S.T. Shi, *Non-nucleoside inhibitors of hepatitis C virus polymerase: current progress and future challenges.* Future medicinal chemistry, 2010. **2**(1): p. 121-41.

52. McKercher, G., et al., *Specific inhibitors of HCV polymerase identified using an NS5B with lower affinity for template/primer substrate.* Nucleic acids research, 2004. **32**(2): p. 422-31.

53. Dhanak, D., et al., *Identification and biological characterization of heterocyclic inhibitors of the hepatitis C virus RNA-dependent RNA polymerase.* The Journal of biological chemistry, 2002. **277**(41): p. 38322-7.

54. Koch, U. and F. Narjes, *Recent progress in the development of inhibitors of the hepatitis C virus RNA-dependent RNA polymerase.* Current topics in medicinal chemistry, 2007. **7**(13): p. 1302-29.

55. De Francesco, R. and A. Carfi, *Advances in the development of new therapeutic agents targeting the NS3-4A serine protease or the NS5B RNA-dependent RNA polymerase of the hepatitis C virus.* Advanced drug delivery reviews, 2007. **59**(12): p. 1242-62.

56. Ludmerer, S.W., et al., *Replication fitness and NS5B drug sensitivity of diverse hepatitis C virus isolates characterized by using a transient replication assay.* Antimicrobial agents and chemotherapy, 2005. **49**(5): p. 2059-69.

57. Li, H., et al., *Identification and structure-based optimization of novel dihydropyrones as potent HCV RNA polymerase inhibitors.* Bioorg Med Chem Lett. , 2006. **16**(18): p. 4834-8.

58. Gopalsamy, A., et al., *Discovery of pyrano[3,4-b]indoles as potent and selective HCV NS5B polymerase inhibitors.* Journal of medicinal chemistry, 2004. **47**(26): p. 6603-8.

59. Wang, M., et al., *Non-nucleoside analogue inhibitors bind to an allosteric site on HCV NS5B polymerase. Crystal structures and mechanism of inhibition.* The Journal of biological chemistry, 2003. **278**(11): p. 9489-95.

60. Li, H., et al., *Discovery of (R)-6-cyclopentyl-6-(2-(2,6-diethylpyridin-4-yl)ethyl)-3-((5,7-dimethyl-[1,2,4]triazolo[1,5-a]pyrimidin-2-yl)methyl)-4-hydroxy-5,6-dihydropyran-2-one (PF-00868554) as a potent and orally available hepatitis C virus polymerase inhibitor.* J.Med.Chem, 2009. **52**: p. 1255-1258.

61. Biswal, B.K., et al., *Non-nucleoside inhibitors binding to hepatitis C virus NS5B polymerase reveal a novel mechanism of inhibition.* Journal of molecular biology, 2006. **361**(1): p. 33-45.

62. Beaulieu, P.L., *Non-nucleoside inhibitors of the HCV NS5B polymerase: progress in the discovery and development of novel agents for the treatment of HCV infections.* Current opinion in investigational drugs, 2007. **8**(8): p. 614-34.

63. Beaulieu, P.L. and Y.S. Tsantrizos, *Inhibitors of the HCV NS5B polymerase: new hope for the treatment of hepatitis C infections.* Current opinion in investigational drugs, 2004. **5**(8): p. 838-50.

64. Molla, A., et al. *Characterization of Pharmacokinetic/Pharmacodynamic Parameters for the Novel HCV Polymerase Inhibitor A-848837*. in *EASL 42nd Meeting of the European Association for the Study of Liver Diseases*. 2007. Barcelona, Spain.

65. Tomei, L., et al., *Characterization of the inhibition of hepatitis C virus RNA replication by nonnucleosides.* Journal of virology, 2004. **78**(2): p. 938-46.

66. Tomei, L., et al., *HCV antiviral resistance: the impact of in vitro studies on the development of antiviral agents targeting the viral NS5B polymerase.* Antiviral chemistry & chemotherapy, 2005. **16**(4): p. 225-45.

67. Nguyen, T.T., et al., *Resistance profile of a hepatitis C virus RNA-dependent RNA polymerase benzothiadiazine inhibitor.* Antimicrobial agents and chemotherapy, 2003. **47**(11): p. 3525-30.

68. Lu, L., et al., *Identification and characterization of mutations conferring resistance to an HCV RNA-dependent RNA polymerase inhibitor in vitro.* Antiviral research, 2007. **76**(1): p. 93-7.

69. Mo, H., et al., *Mutations conferring resistance to a hepatitis C virus (HCV) RNA-dependent RNA polymerase inhibitor alone or in combination with an HCV serine protease inhibitor in vitro.* Antimicrobial agents and chemotherapy, 2005. **49**(10): p. 4305-14.

70. Chen, C.M., et al., *Activity of a potent hepatitis C virus polymerase inhibitor in the chimpanzee model.* Antimicrobial agents and chemotherapy, 2007. **51**(12): p. 4290-6.

71. Lesburg, C.A., R. Radfar, and P.C. Weber, *Recent advances in the analysis of HCV NS5B RNA-dependent RNA polymerase.* Current opinion in investigational drugs, 2000. **1**(3): p. 289-96.

72. de Vicente, J., et al., *Non-nucleoside inhibitors of HCV polymerase NS5B. Part 4: structure-based design, synthesis, and biological evaluation of benzo[d]isothiazole-1,1-dioxides.* Bioorganic & medicinal chemistry letters, 2009. **19**(19): p. 5652-6.

73. Levin, J., *HCV-796 Displays Potent Antiviral Activity in Replicon and in Chimeric Mice Infected with Hepatitis C Virus (HCV)* in *46th Annual ICAAC Interscience Conference on Antimicrobial Agents and Chemotherapy*2006: San Francosco.

74. Levin, J., *HCV-796, new HCV non-nucleoside- 14 day study : Antiviral Activity of the Non-Nucleoside Polymerase Inhibitor, HCV-796, in patients with chronic HCV: preliminary results from a randomized, double-blind, placebo-controlled, ascending multiple dose study*, in *Digestive disease week*2006: Los Angeles.

75. Kneteman, N.M., et al., *HCV796: A selective nonstructural protein 5B polymerase inhibitor with potent anti-hepatitis C virus activity in vitro, in mice with chimeric human livers, and in humans infected with hepatitis C virus.* Hepatology, 2009. **49**(3): p. 745-52.

76. Beaulieu, P.L., *Successful Strategies for the Discovery of Antiviral Drugs*, in *Drug discovery*, M.C. Desai and N.A. Meanwell, Editors. 2013, The Royal Society of Chemistry. p. 279.

77. Drwal, M.N. and R. Griffith, *Combination of ligand- and structure-based methods in virtual screening.* Drug discovery today. Technologies, 2013. **10**(3): p. e395-401.

78. Warren, G.L., et al., *A critical assessment of docking programs and scoring functions.* Journal of medicinal chemistry, 2006. **49**(20): p. 5912-31.

79. Louise-May, S., et al., *Discovery of novel dialkyl substituted thiophene inhibitors of HCV by in silico screening of the NS5B RdRp.* Bioorganic & medicinal chemistry letters, 2007. **17**(14): p. 3905-9.

80. Golub, A.G., et al., *Discovery of new scaffolds for rational design of HCV NS5B polymerase inhibitors.* European journal of medicinal chemistry, 2012. **58**: p. 258-64.

81. Corbeil, C.R., P. Englebienne, and N. Moitessier, *Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0.* Journal of chemical information and modeling, 2007. **47**(2): p. 435-49.

82. Barreca, M.L., et al., *Accounting for Target Flexibility and Water Molecules by Docking to Ensembles of Target Structures: The HCV NS5B Palm Site I Inhibitors Case Study.* Journal of chemical information and modeling, 2014. **54**(2): p. 481-97.

83. Li, T., M. Froeyen, and P. Herdewijn, *Insight into ligand selectivity in HCV NS5B polymerase: molecular dynamics simulations, free energy decomposition and docking.* Journal of molecular modeling, 2010. **16**(1): p. 49-59.

84. Yu, H., et al., *Combined 3D-QSAR, molecular docking, molecular dynamics simulation, and binding free energy calculation studies on the 5-hydroxy-2H-pyridazin-3-one derivatives as HCV NS5B polymerase inhibitors.* Chemical biology & drug design, 2014. **83**(1): p. 89-105.

85. Wermuth, C.G., et al., *Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998).* Pure and Applied Chemistry, 1998. **70**: p. 1129-1143.

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

87. Sun, H., *Pharmacophore-based virtual screening.* Current medicinal chemistry, 2008. **15**(10): p. 1018-24.

88. Kim, J., et al., *Identification of novel HCV RNA-dependent RNA polymerase inhibitors using pharmacophore-guided virtual screening.* Chemical biology & drug design, 2008. **72**(6): p. 585-91.

89. Tian, S., et al., *Development and evaluation of an integrated virtual screening strategy by combining molecular docking and pharmacophore searching based on multiple protein structures.* Journal of chemical information and modeling, 2013. **53**(10): p. 2743-56.

90. Mahmoud, A.H., et al., *A highly selective structure-based virtual screening model of Palm I allosteric inhibitors of HCV Ns5b polymerase enzyme and its application in the discovery and optimization of new analogues.* European journal of medicinal chemistry, 2012. **57**: p. 468-82.

91. Therese, P.J., et al., *Multiple e-Pharmacophore Modeling, 3D-QSAR, and High-Throughput Virtual Screening of Hepatitis C Virus NS5B Polymerase Inhibitors.* Journal of chemical information and modeling, 2014. **54**(2): p. 539-52.

92. Dror, O., et al., *Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications to drug design.* Current medicinal chemistry, 2004. **11**(1): p. 71-90.

93. Plewczynski, D., et al., *Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database.* Journal of computational chemistry, 2011. **32**(4): p. 742-55.

94. Schneider, G., et al., *"Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening.* Angewandte Chemie, 1999. **38**(19): p. 2894-2896.

95. Peach, M.L. and M.C. Nicklaus, *Combining docking with pharmacophore filtering for improved virtual screening.* Journal of cheminformatics, 2009. **1**(1): p. 6.

96. Xu, Z., et al., *Combining pharmacophore, docking and substructure search approaches to identify and optimize novel B-RafV600E inhibitors.* Bioorganic & medicinal chemistry letters, 2012. **22**(17): p. 5428-37.

97. Cheng, C.C., et al., *Pyridine Carboxamides: Potent Palm Site Inhibitors of HCV NS5B Polymerase.* ACS Medicinal Chemistry Letters 2010. **1**(9): p. 466-471.

98. Lam, A.M., et al. *Selection and characterization of hepatitis C virus replicons using combination of NS3 protease and NS5B non-nucleoside inhibitors or combination of NS5B nucleoside inhibitors* in *International HIV and Hepatitis virus drug resistance workshop and curative strategies*. 2010. Dubrovnik, Croatia.

99. Yan, Z. and G. Caldwell, *Optimization in drug discovery : in vitro methods*2004, Totowa, N.J. ; [Great Britain]: Humana Press.

100. Cer, R.Z., et al., *IC50-to-Ki: a web-based tool for converting IC50 to Ki values for inhibitors of enzyme activity and ligand binding.* Nucleic acids research, 2009. **37**(Web Server issue): p. W441-5.

101. Sussman, J. and P. Spadon, *From molecules to medicines : structure of biological macromolecules and its relevance in combating new diseases and bioterrorism*2009, Dordrecht: Springer.

102. Suenaga, A., et al., *An efficient computational method for calculating ligand binding affinities.* PloS one, 2012. **7**(8): p. e42846.

103. Stahl, M. and M. Rarey, *Detailed analysis of scoring functions for virtual screening.* Journal of medicinal chemistry, 2001. **44**(7): p. 1035-42.

104. Moitessier, N., et al., *Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go.* British journal of pharmacology, 2008. **153 Suppl 1**: p. S7-26.

105. Nissink, J.W., et al., *A new test set for validating predictions of protein-ligand interaction.* Proteins, 2002. **49**(4): p. 457-71.

106. Cheng, T., et al., *Comparative assessment of scoring functions on a diverse test set.* Journal of chemical information and modeling, 2009. **49**(4): p. 1079-93.

107. Pearlman, D.A. and P.S. Charifson, *Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system.* Journal of medicinal chemistry, 2001. **44**(21): p. 3417-23.

108. Boyd, D.B., K.B. Lipkowitz, and John Wiley & Sons., *Reviews in computational chemistry. Vol. 17*, 2001, Wiley-VCH: New York.

109. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking.* Journal of molecular biology, 1997. **267**(3): p. 727-48.

110. Taylor, J.S. and R.M. Burnett, *DARWIN: a program for docking flexible molecules.* Proteins, 2000. **41**(2): p. 173-91.

111. Pei, J., et al., *PSI-DOCK: towards highly efficient and accurate flexible ligand docking.* Proteins, 2006. **62**(4): p. 934-46.

112. Osterberg, F., et al., *Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock.* Proteins, 2002. **46**(1): p. 34-40.

113. Gupta, A., et al., *ParDOCK: an all atom energy based Monte Carlo docking protocol for protein-ligand complexes.* Protein and peptide letters, 2007. **14**(7): p. 632-46.

114. Tietze, S. and J. Apostolakis, *GlamDock: development and validation of a new docking tool on several thousand protein-ligand complexes.* Journal of chemical information and modeling, 2007. **47**(4): p. 1657-72.

115. Meier, R., et al., *ParaDockS: a framework for molecular docking with population-based metaheuristics.* Journal of chemical information and modeling, 2010. **50**(5): p. 879-89.

116. Friesner, R.A., 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): p. 1739-49.

117. Doucet, J.-P. and J. Weber, *Computer-aided molecular design : theory and applications*1996, London: Academic. xix, 487 p., 8 p. of plates.

118. Xu, Y., D. Xu, and J. Liang, *Computational methods for protein structure prediction and modeling*. Biological and medical physics, biomedical engineering2007, New York, N.Y.: Springer.

119. Meng, E.C., B.K. Shoichet, and I.D. Kuntz, *Automated docking with grid-based energy evaluation.* Journal of Computational Chemistry, 1992. **13**(4): p. 505-524.

120. Verdonk, M.L., et al., *Improved protein-ligand docking using GOLD.* Proteins, 2003. **52**(4): p. 609-23.

121. Halgren, T.A., 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): p. 1750-9.

122. Verkhivker, G.M., et al., *Towards understanding the mechanisms of molecular recognition by computer simulations of ligand-protein interactions.* Journal of molecular recognition : JMR, 1999. **12**(6): p. 371-89.

123. Eldridge, M.D., et al., *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes.* Journal of computer-aided molecular design, 1997. **11**(5): p. 425-45.

124. Muegge, I. and Y.C. Martin, *A general and fast scoring function for protein-ligand interactions: a simplified potential approach.* Journal of medicinal chemistry, 1999. **42**(5): p. 791-804.

125. Mooij, W.T. and M.L. Verdonk, *General and targeted statistical potentials for protein-ligand interactions.* Proteins, 2005. **61**(2): p. 272-87.

126. Li, H., et al., *An effective docking strategy for virtual screening based on multi-objective optimization algorithm.* BMC bioinformatics, 2009. **10**: p. 58.

127. Wei, B.Q., et al., *A model binding site for testing scoring functions in molecular docking.* Journal of molecular biology, 2002. **322**(2): p. 339-55.

128. Rapp, C., et al., *A molecular mechanics approach to modeling protein-ligand interactions: relative binding affinities in congeneric series.* Journal of chemical information and modeling, 2011. **51**(9): p. 2082-9.

129. Okimoto, N., et al., *High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations.* PLoS computational biology, 2009. **5**(10): p. e1000528.

130. Rastelli, G., et al., *Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA.* Journal of computational chemistry, 2010. **31**(4): p. 797-810.

131. Kuhn, B., et al., *Validation and use of the MM-PBSA approach for drug discovery.* Journal of medicinal chemistry, 2005. **48**(12): p. 4040-8.

132. Case, D.A., et al., *The Amber biomolecular simulation programs.* Journal of computational chemistry, 2005. **26**(16): p. 1668-88.

133. Pronk, S., et al., *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.* Bioinformatics, 2013. **29**(7): p. 845-854.

134. Lajtha, A., et al., *Practical neurochemistry methods*. 3rd ed. ed2007, New York: Springer.

135. Hoffmann, T., *Utersuchung von putativen Inhibitoren an viralen Polymerasen*, in *Institut für Biochemie und Biotechnologie* 2013, Martin-Luther-Universität Halle-Wittenberg Halle.

136. Carlson, H.A., *Protein flexibility and drug design: how to hit a moving target.* Current opinion in chemical biology, 2002. **6**(4): p. 447-52.

137. Totrov, M. and R. Abagyan, *Flexible ligand docking to multiple receptor conformations: a practical alternative.* Current opinion in structural biology, 2008. **18**(2): p. 178-84.

138. Cozzini, P., et al., *Target flexibility: an emerging consideration in drug discovery and design.* Journal of medicinal chemistry, 2008. **51**(20): p. 6237-55.

139. Ferrari, A.M., et al., *Soft docking and multiple receptor conformations in virtual screening.* Journal of medicinal chemistry, 2004. **47**(21): p. 5076-84.

140. Sherman, W., et al., *Novel procedure for modeling ligand/receptor induced fit effects.* Journal of medicinal chemistry, 2006. **49**(2): p. 534-53.

141. Barril, X. and S.D. Morley, *Unveiling the full potential of flexible receptor docking using multiple crystallographic structures.* Journal of medicinal chemistry, 2005. **48**(13): p. 4432-43.

142. Huang, S.Y. and X. Zou, *Efficient molecular docking of NMR structures: application to HIV-1 protease.* Protein science : a publication of the Protein Society, 2007. **16**(1): p. 43-51.

143. Guimaraes, C.R. and M. Cardozo, *MM-GB/SA rescoring of docking poses in structure-based lead optimization.* Journal of chemical information and modeling, 2008. **48**(5): p. 958-70.

144. Thompson, D.C., C. Humblet, and D. Joseph-McCarthy, *Investigation of MM-PBSA rescoring of docking poses.* Journal of chemical information and modeling, 2008. **48**(5): p. 1081-91.

145. Rastelli, G., et al., *Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening.* Chemical biology & drug design, 2009. **73**(3): p. 283-6.

146. Rueda, M., G. Bottegoni, and R. Abagyan, *Recipes for the selection of experimental protein conformations for virtual screening.* Journal of chemical information and modeling, 2010. **50**(1): p. 186-93.

147. Polgar, T. and G.M. Keseru, *Ensemble docking into flexible active sites. Critical evaluation of FlexE against JNK-3 and beta-secretase.* Journal of chemical information and modeling, 2006. **46**(4): p. 1795-805.

148. Hou, T., et al., *Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations.* Journal of chemical information and modeling, 2011. **51**(1): p. 69-82.

149. Gleeson, M.P. and D. Gleeson, *QM/MM as a tool in fragment based drug discovery. A cross-docking, rescoring study of kinase inhibitors.* Journal of chemical information and modeling, 2009. **49**(6): p. 1437-48.

150. Khandelwal, A., et al., *A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands.* Journal of medicinal chemistry, 2005. **48**(17): p. 5437-47.

151. Case, D.A., et al., *AMBER 12.* University of California, San Francisco, 2012.

152. Verdonk, M.L., et al., *Virtual screening using protein-ligand docking: avoiding artificial enrichment.* Journal of Chemical Information and Computer Sciences, 2004. **44**(3): p. 793-806.

153. Parks, J.M., et al., *Hepatitis C virus NS5B polymerase: QM/MM calculations show the important role of the internal energy in ligand binding.* The journal of physical chemistry. B, 2008. **112**(10): p. 3168-76.

154. Deng, Z., C. Chuaqui, and J. Singh, *Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions.* Journal of medicinal chemistry, 2004. **47**(2): p. 337-44.

155. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**: p. 5-32.

156. Vandyck, K., et al., *Structure-based design of a benzodiazepine scaffold yields a potent allosteric inhibitor of hepatitis C NS5B RNA polymerase.* Journal of medicinal chemistry, 2009. **52**(14): p. 4099-102.

157. Svetnik, V., et al., *Random forest: a classification and regression tool for compound classification and QSAR modeling.* Journal of Chemical Information and Computer Sciences, 2003. **43**(6): p. 1947-58.

158. Wang, M., et al., *Classification of HCV NS5B Polymerase Inhibitors Using Support Vector Machine.* International journal of molecular sciences, 2012. **13**(4): p. 4033-47.

159. Chen, C., A. Liaw, and L. Breiman, *Using Random Forest to Learn Imbalanced Data*, in *Statistics Technical Reports*2004, University of California.

160. Liaw, A. and M. Wiener, *Classification and Regression by randomForest.* R News, 2002. **2**(3): p. 18-22.

161. *Molecular Operating Environment (MOE), 2012.10*, 2012, Chemical Computing Group Inc.: Montreal, QC, Canada.

162. Yap, C.W., *PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints.* Journal of computational chemistry, 2011. **32**(7): p. 1466-74.

163. Bender, A. and R.C. Glen, *Molecular similarity: a key technique in molecular informatics.* Organic & biomolecular chemistry, 2004. **2**(22): p. 3204-18.

164. O'Boyle, N.M., et al., *Open Babel: An open chemical toolbox.* Journal of cheminformatics, 2011. **3**: p. 33.

165. Backman, T.W., Y. Cao, and T. Girke, *ChemMine tools: an online service for analyzing and clustering small molecules.* Nucleic acids research, 2011. **39**(Web Server issue): p. W486-91.

166. Mysinger, M.M., et al., *Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking.* Journal of medicinal chemistry, 2012. **55**(14): p. 6582-94.

167. Levin, J., *Antiviral Activity of the Non-Nucleoside Polymerase Inhibitor, HCV-796, in patients with chronic HCV: preliminary results from a randomized, double-blind, placebo-controlled, ascending multiple dose study*, in *Digestive Disease Week*2006: Los Angeles.

168. Robinson, M., et al., *Preexisting drug-resistance mutations reveal unique barriers to resistance for distinct antivirals.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(25): p. 10290-5.

169. Hang, J.Q., et al., *Slow binding inhibition and mechanism of resistance of non-nucleoside polymerase inhibitors of hepatitis C virus.* The Journal of biological chemistry, 2009. **284**(23): p. 15517-29.

170. Wolber, G. and T. Langer, *LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters.* Journal of chemical information and modeling, 2005. **45**(1): p. 160-9.

171. Kim, N.D., et al., *Discovery of novel HCV polymerase inhibitors using pharmacophore-based virtual screening.* Bioorganic & medicinal chemistry letters, 2011. **21**(11): p. 3329-34.

172. Keeffe, E.B., *Future treatment of chronic hepatitis C.* Antiviral therapy, 2007. **12**(7): p. 1015-25.

173. Manns, M.P., et al., *The way forward in HCV treatment--finding the right path.* Nature reviews. Drug discovery, 2007. **6**(12): p. 991-1000.

174. McCown, M.F., et al., *The hepatitis C virus replicon presents a higher barrier to resistance to nucleoside analogs than to nonnucleoside polymerase or protease inhibitors.* Antimicrobial agents and chemotherapy, 2008. **52**(5): p. 1604-12.

175. Howe, A.Y., et al., *Molecular mechanism of hepatitis C virus replicon variants with reduced susceptibility to a benzofuran inhibitor, HCV-796.* Antimicrobial agents and chemotherapy, 2008. **52**(9): p. 3327-38.

176. Reich, S., et al., *Mechanisms of activity and inhibition of the hepatitis C virus RNA-dependent RNA polymerase.* The Journal of biological chemistry, 2010. **285**(18): p. 13685-93.

177. Bordoli, L., et al., *Protein structure homology modeling using SWISS-MODEL workspace.* Nature protocols, 2009. **4**(1): p. 1-13.

178. Klebe, G., *Drug design Methodology, concepts, and mode-of-action*. Vol. 1. 2013, Marburg: Springer.

179. Nomaguchi, M., et al., *Requirements for West Nile virus (-)- and (+)-strand subgenomic RNA synthesis in vitro by the viral RNA-dependent RNA polymerase expressed in Escherichia coli*. The Journal of biological chemistry, 2004. **279**(13): p. 12141-51.

180. Malet, H., et al., *Crystal structure of the RNA polymerase domain of the West Nile virus non-structural protein 5.* The Journal of biological chemistry, 2007. **282**(14): p. 10678-89.

181. Xiang, Z., *Advances in homology protein structure modeling.* Current protein & peptide science, 2006. **7**(3): p. 217-27.

182. Cang, X., et al., *Mapping the functional binding sites of cholesterol in beta2-adrenergic receptor by long-time molecular dynamics simulations.* The journal of physical chemistry. B, 2013. **117**(4): p. 1085-94.

183. Irwin, J.J., et al., *ZINC: a free tool to discover chemistry for biology.* Journal of chemical information and modeling, 2012. **52**(7): p. 1757-68.

184. Kukolj, G., et al., *Binding site characterization and resistance to a class of non-nucleoside inhibitors of the hepatitis C virus NS5B polymerase.* The Journal of biological chemistry, 2005. **280**(47): p. 39260-7.

185. Desai, M.C.e.o.c. and N.A.e.o.c. Meanwell, *Successful strategies for the discovery of antiviral drugs.*

186. Rigat, K., et al., *Ligand-induced changes in hepatitis C virus NS5B polymerase structure.* Antiviral research, 2010. **88**(2): p. 197-206.

187. Tedesco, R., et al., *3-(1,1-dioxo-2H-(1,2,4)-benzothiadiazin-3-yl)-4-hydroxy-2(1H)-quinolinones, potent inhibitors of hepatitis C virus RNA-dependent RNA polymerase.* Journal of medicinal chemistry, 2006. **49**(3): p. 971-83.

188. Cai, Z., et al., *Mutagenesis analysis of the rGTP-specific binding site of hepatitis C virus RNA-dependent RNA polymerase.* Journal of virology, 2005. **79**(18): p. 11607-17.

189. Jiao, P., et al., *Understanding the drug resistance mechanism of hepatitis C virus NS5B to PF-00868554 due to mutations of the 423 site: a computational study.* Molecular bioSystems, 2014. **10**(4): p. 767-77.

190. Xue, W., et al., *Molecular modeling and residue interaction network studies on the mechanism of binding and resistance of the HCV NS5B polymerase mutants to VX-222 and ANA598.* Antiviral research, 2014. **104**: p. 40-51.

191. Love, R.A., et al., *Crystallographic identification of a noncompetitive inhibitor binding site on the hepatitis C virus NS5B RNA polymerase enzyme.* Journal of virology, 2003. **77**(13): p. 7575-81.

192. Le Pogam, S., et al., *Selection and characterization of replicon variants dually resistant to thumb- and palm-binding nonnucleoside polymerase inhibitors of the hepatitis C virus.* Journal of virology, 2006. **80**(12): p. 6146-54.

193. Yan, S., et al., *Structure-based design of a novel thiazolone scaffold as HCV NS5B polymerase allosteric inhibitors.* Bioorg.Med.Chem.Lett., 2006. **16**: p. 5888-5891.

194. Yan, S., et al., *Novel thiazolones as HCV NS5B polymerase allosteric inhibitors: Further designs, SAR, and X-ray complex structure.* Bioorg.Med.Chem.Lett., 2007. **17**: p. 63-67.

195. Yan, S., et al., *Thiazolone-acylsulfonamides as novel HCV NS5B polymerase allosteric inhibitors: convergence of structure-based drug design and X-ray crystallographic study.* Bioorg.Med.Chem.Lett., 2007. **17**: p. 1991-1995.

196. Ontaria, J.M., et al., *Identification and biological evaluation of a series of 1H-benzo[de]isoquinoline-1,3(2H)-diones as hepatitis C virus NS5B polymerase inhibitors.* J.Med.Chem, 2009. **52**: p. 5217-5227.

197. Antonysamy, S.S., et al., *Fragment-based discovery of hepatitis C virus NS5b RNA polymerase inhibitors.* Bioorganic & medicinal chemistry letters, 2008. **18**(9): p. 2990-5.

198. Pfefferkorn, J.A., et al., *Inhibitors of HCV NS5B polymerase. Part 1: Evaluation of the southern region of (2Z)-2-(benzoylamino)-3-(5-phenyl-2-furyl)acrylic acid.* Bioorganic & medicinal chemistry letters, 2005. **15**(10): p. 2481-6.

199. Pfefferkorn, J.A., et al., *Inhibitors of HCV NS5B polymerase. Part 2: Evaluation of the northern region of (2Z)-2-benzoylamino-3-(4-phenoxy-phenyl)-acrylic acid.* Bioorganic & medicinal chemistry letters, 2005. **15**(11): p. 2812-8.

200. Gopalsamy, A., et al., *Discovery of proline sulfonamides as potent and selective hepatitis C virus NS5b polymerase inhibitors. Evidence for a new NS5b polymerase binding site.* J.Med.Chem, 2006. **49**: p. 3052-3055.

201. Slater, M.J., et al., *Optimization of novel acyl pyrrolidine inhibitors of hepatitis C virus RNA-dependent RNA polymerase leading to a development candidate.* J.Med.Chem, 2007. **50**.

202. Nittoli, T., et al., *Identification of anthranilic acid derivatives as a novel class of allosteric inhibitors of hepatitis C NS5B polymerase.* J.Med.Chem., 2007. **50**: p. 2108-2116.

203. Zhou, Y., et al., *Novel HCV NS5B polymerase inhibitors derived from 4-(1',1'-dioxo-1',4'-dihydro-1'lambda6-benzo[1',2',4']thiadiazin-3'-yl)-5-hydroxy-2H-pyridazin-3-ones. Part 1: exploration of 7'-substitution of benzothiadiazine.* Bioorg.Med.Chem.Lett., 2008. **18**: p. 1413-1418.

204. Ruebsam, F., et al., *Pyrrolo[1,2-b]pyridazin-2-ones as potent inhibitors of HCV NS5B polymerase.* Bioorg.Med.Chem.Lett., 2008. **18**: p. 3616-3621.

205. Nyanguile, O., et al., *1,5-benzodiazepines, a novel class of hepatitis C virus polymerase nonnucleoside inhibitors.* Antimicrobial agents and chemotherapy, 2008. **52**(12): p. 4420-31.

206. Ruebsam, F., et al., *Hexahydro-pyrrolo- and hexahydro-1H-pyrido[1,2-b]pyridazin-2-ones as potent inhibitors of HCV NS5B polymerase.* Bioorg.Med.Chem.Lett., 2008. **18**: p. 5002-5005.

207. Ellis, D.A., et al., *4-(1,1-Dioxo-1,4-dihydro-1lambda6-benzo[1,4]thiazin-3-yl)-5-hydroxy-2H-pyridazin-3-ones as potent inhibitors of HCV NS5B polymerase.* Bioorg.Med.Chem.Lett. , 2008. **18**: p. 4628-4632.

208. Kim, S.H., et al., *Structure-based design, synthesis, and biological evaluation of 1,1-dioxoisothiazole and benzo[b]thiophene-1,1-dioxide derivatives as novel inhibitors of hepatitis C virus NS5B polymerase.* Bioorganic & medicinal chemistry letters, 2008. **18**(14): p. 4181-5.

209. de Vicente, J., et al., *Non-nucleoside inhibitors of HCV polymerase NS5B. Part 2: Synthesis and structure-activity relationships of benzothiazine-substituted quinolinediones.* Bioorganic & medicinal chemistry letters, 2009. **19**(13): p. 3642-6.

210. Ellis, D.A., et al., *5,5'- and 6,6'-dialkyl-5,6-dihydro-1H-pyridin-2-ones as potent inhibitors of HCV NS5B polymerase.* Bioorg.Med.Chem.Lett., 2009. **19**: p. 6047-6052.

211. Ruebsam, F., et al., *Discovery of tricyclic 5,6-dihydro-1H-pyridin-2-ones as novel, potent, and orally bioavailable inhibitors of HCV NS5B polymerase.* Bioorg.Med.Chem.Lett., 2009. **19**: p. 6404-6412.

212. de Vicente, J., et al., *Non-nucleoside inhibitors of HCV polymerase NS5B. Part 3: synthesis and optimization studies of benzothiazine-substituted tetramic acids.* Bioorganic & medicinal chemistry letters, 2009. **19**(19): p. 5648-51.

213. Wang, G., et al., *HCV NS5B polymerase inhibitors 2: Synthesis and in vitro activity of (1,1-dioxo-2H-[1,2,4]benzothiadiazin-3-yl) azolo[1,5-a]pyridine and azolo[1,5-a]pyrimidine derivatives.* Bioorganic & medicinal chemistry letters, 2009. **19**(15): p. 4480-3.

214. Shaw, A.N., et al., *Substituted benzothiadizine inhibitors of Hepatitis C virus polymerase.* Bioorganic & medicinal chemistry letters, 2009. **19**(15): p. 4350-3.

215. Gopalsamy, A., et al., *Discovery of proline sulfonamides as potent and selective hepatitis C virus NS5b polymerase inhibitors. Evidence for a new NS5b polymerase binding site.* Journal of medicinal chemistry, 2006. **49**(11): p. 3052-5.

216. Slater, M.J., et al., *Optimization of novel acyl pyrrolidine inhibitors of hepatitis C virus RNA-dependent RNA polymerase leading to a development candidate.* Journal of medicinal chemistry, 2007. **50**(5): p. 897-900.

217. Nittoli, T., et al., *Identification of anthranilic acid derivatives as a novel class of allosteric inhibitors of hepatitis C NS5B polymerase.* Journal of medicinal chemistry, 2007. **50**(9): p. 2108-16.

218. Zhou, Y., et al., *Novel HCV NS5B polymerase inhibitors derived from 4-(1',1'-dioxo-1',4'-dihydro-1'lambda6-benzo[1',2',4']thiadiazin-3'-yl)-5-hydroxy -2H-pyridazin-3-ones. Part 1: exploration of 7'-substitution of benzothiadiazine.* Bioorganic & medicinal chemistry letters, 2008. **18**(4): p. 1413-8.

219. Ruebsam, F., et al., *Pyrrolo[1,2-b]pyridazin-2-ones as potent inhibitors of HCV NS5B polymerase.* Bioorganic & medicinal chemistry letters, 2008. **18**(12): p. 3616-21.

220. Ruebsam, F., et al., *Hexahydro-pyrrolo- and hexahydro-1H-pyrido[1,2-b]pyridazin-2-ones as potent inhibitors of HCV NS5B polymerase.* Bioorganic & medicinal chemistry letters, 2008. **18**(18): p. 5002-5.

221. Ellis, D.A., et al., *4-(1,1-Dioxo-1,4-dihydro-1lambda6-benzo[1,4]thiazin-3-yl)-5-hydroxy-2H-pyridazin- 3-ones as potent inhibitors of HCV NS5B polymerase.* Bioorganic & medicinal chemistry letters, 2008. **18**(16): p. 4628-32.

222. Dragovich, P.S., et al., *Novel HCV NS5B polymerase inhibitors derived from 4-(1',1'-dioxo-1',4'-dihydro-1'lambda(6)-benzo[1',2',4']thiadiazin-3'-yl)-5-hydro xy-2H-pyridazin-3-ones. Part 5: Exploration of pyridazinones containing 6-amino-substituents.* Bioorganic & medicinal chemistry letters, 2008. **18**(20): p. 5635-9.

223. McGowan, D., et al., *1,5-Benzodiazepine inhibitors of HCV NS5B polymerase.* Bioorganic & medicinal chemistry letters, 2009. **19**(9): p. 2492-6.

224. Ellis, D.A., et al., *5,5'- and 6,6'-dialkyl-5,6-dihydro-1H-pyridin-2-ones as potent inhibitors of HCV NS5B polymerase.* Bioorganic & medicinal chemistry letters, 2009. **19**(21): p. 6047-52.

List of tables (Appendix)

Pages

Table A1. Details of 22 crystal structure of thumb site II dataset……. 116
Table A2. Details of 29 crystal structures of palm site I dataset……...….... 118
Table A3. Cross-docking results using Chem-score for 22 crystal structures of the thumb dataset (upper) and 29 crystal structures of the palm dataset (lower). The tables display the root mean square deviation (RMSD) and the correct predicted poses are labeled in blue shade……. 122
Table A4. Cross-docking results of the thumb dataset (22 crystal structures) using 6 scoring functions; Gold-score, Chem-score, ASP-score, P-score, PMF and GLIDE SP. Average root mean square deviation (RMSD) and the numbers of correct poses are presented according to their protein cluster group. A color bar in the 'total' column denotes the percentage of correct pose prediction……..….. 123
Table A5. Cross-docking results of the palm dataset (29 crystal structures) using 6 scoring functions; Gold-score, Chem-score, ASP-score, P-score, PMF and GLIDE SP. A color bar denotes the percentage of correct pose prediction……... 124
Table A6. 45 ligands used in the rescoring study……. 125
Table A7. 3D Padel descriptors used in this study…… 129
Table A8. MOE descriptors used in this study……... 130
…...
Table A9. 30 compounds from the study of Kim et al [171]…….. 134
Table A10. Root mean square deviation (RMSD) of re-docking HCV-796 into 3FQK and 3FQL structure……...…… 136
Table A11. Inhibitory profiles and estimated binding free energy calculations (GBTOT) of complex between inhibitors and HCV polymerases (± standard deviation) calculated from MM-GB/SA. All energies are in kcal/mol. Predicted binding free energies were averaged over 100 snapshots during last 2 ns of MD simulation…….…… 145
Table A12. Per-residue energy (kcal/mol) of the key amino acids in the binding pocket of the wild-type (WT) and mutant enzymes with compound A and compound B. The grey and red background shading represents the energy difference of the residue contributions in the mutants that are between 0.5-1 kcal/mol and greater than 1 kcal/mol relative to the wild-type, respectively. The boxes show the mutation points…… 146

List of figures (Appendix)

Pages

Figure A1. 22 co-crystallized ligands of thumb site II dataset used in cross-docking study…….. 117
Figure A2. 29 co-crystallized ligand of palm site I dataset used in cross-docking study. 119
Figure A3. Superimposition of X-ray structures PDB ID: 2HAI (Pink), 1YVX (Yellow) and 3CJ4 (Cyan) representing group 1, 2 and 3, respectively. The inhibitors of each group are shown by balls and sticks in the same color as their protein structures…….…….. 121
Figure A4. Additional co-crystallized ligands of palm site I dataset used in ensemble study…….…… 125
Figure A5. 45 ligand structures used in rescoring study…… 127
Figure A6. Docked poses of 14 tested compounds (Dataset-I) and their Gold-scores … 131
Figure A7. Root Mean Square Deviations (RMSD) plots of the complexes and ligands during MD simulation…….**.....** 133
Figure A8. The autoradiography of the HCV NS5B polymerase assays with the inhibitors T1-T11 (Dataset-II). Each compound was tested independently three times…….. 136
Figure A9. Docked poses of 11 tested compounds (Dataset-II) in structure of HCV polymerase genotype 1b Con1 (PDB ID: 3FQL) and their Glide SP scores……**...** 137
Figure A10. RMSD plots of the complexes (red) and 11 tested compounds (teal) during MD simulation…….…….. 139
Figure A11. Bound conformations of 11 tested compounds generated from average 100 snapshots of MD simulations…….……**....** 141
Figure A12. Root Mean Square Deviations (RMSD) of the HCV polymerase (red) complexed with inhibitor (teal) during 12 ns simulation…….…. 143
Figure A13. Comparison of the flexibility of HCV-NS5B-inhibitor complexes. Compound A and compound B bound to wild-type HCV NS5B polymerase are shown in the upper part, whereas binding to the P495A mutant is shown in the lower part…….. 145

Appendix

Table A1. Details of 22 crystal structure of thumb site II dataset.

PDB ID	Resolution	Ligand ID	Chain Length	Reference
1OS5	2.2	NH1	576	[191]
2D3U	2.0	CCT	570	[61]
2D3Z	1.8	FIH	570	[61]
2D41	2.1	SNH	570	[61]
2GIR	1.9	NN3	568	[192]
2HAI	1.58	CME	576	[57]
2HWH	2.3	RNA	576	[193]
2HWI	2.0	VRX	576	[193]
2I1R	2.2	VXR	576	[194]
2O5D	2.2	VR1	576	[195]
3FRZ	1.86	AG0	576	[60]
1NHU	2.0	153	578	[59]
1NHV	2.9	154	578	[59]
2WHO	2.0	VGI	536	[196]
3CIZ	1.87	SX1	576	[197]
3CJ0	1.9	SX2	576	
3CJ2	1.75	SX3	576	
3CJ3	1.87	SX4	576	
3CJ4	2.07	SX5	576	
3CJ5	1.92	SX6	576	
1YVX	2.0	IPC	570	[36]
1YVZ	2.2	JPC	570	

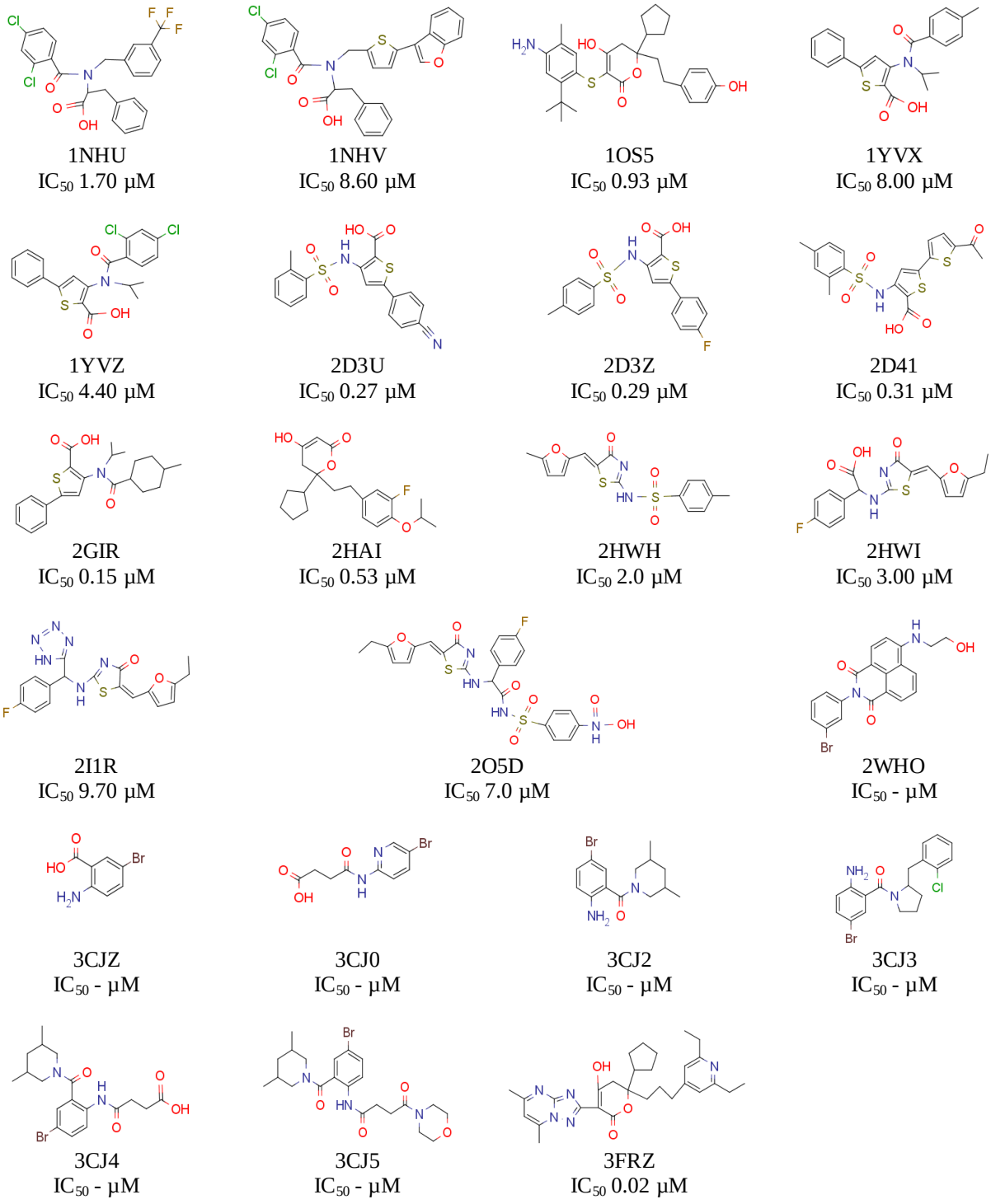


Figure A1. 22 co-crystalized ligands of thumb site II dataset used in cross-docking study.

Table A2. Details of 29 crystal structures of palm site I dataset.

PDB ID	Resolution	Ligand ID	Chain Length	References
1YVF	2.5	PH7	577	[198]
1Z4U	2.8	PH9	577	[199]
2GC8	2.2	885	578	[200]
2GIQ	1.65	NN2	568	[192]
2JC0	2.2	699	570	[201]
2JC1	2.0	698	570	
2QE5	2.6	617	578	[202]
3BR9	2.3	DEY	578	[203]
3BSA	2.3	1PD	578	
3BSC	2.65	2PD	578	
3CDE	2.1	N3H	578	
3CO9	2.1	3MS	578	[204]
3CSO	2.71	XNI	581	[205]
3CVK	2.31	N34	578	[206]
3CWJ	2.4	321	578	[207]
3D28	2.3	B34	578	[208]
3D5M	2.2	4MS	578	
3E51	1.9	N35	578	[203]
3G86	2.2	T18	576	[209]
3GNV	2.75	XNZ	581	[156]
3GNW	2.39	XNC	581	
3GYN	2.15	B42	578	[210]
3H2L	1.9	YAK	578	[211]
3H59	2.1	H59	576	[212]
3H5S	2.0	H5S	576	[72]
3H5U	1.95	H5U	576	
3H98	1.9	B5P	576	[213]
3HHK	1.7	77Z	563	[214]
3IGV	2.6	B80	578	[210]

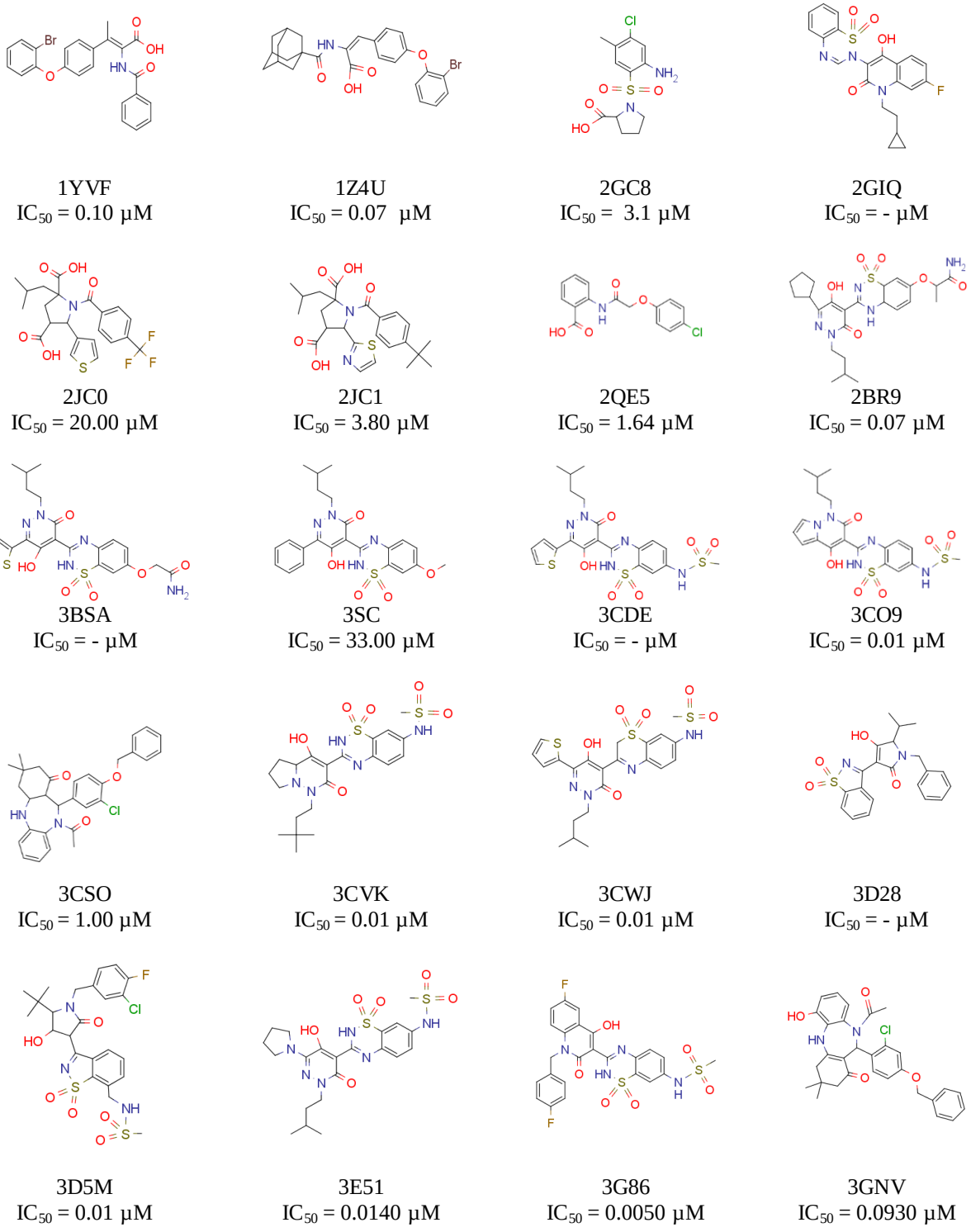
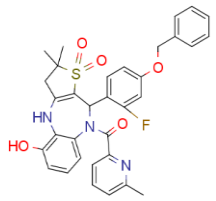
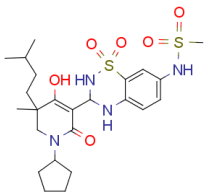


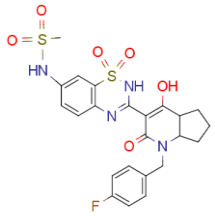
Figure A2. 29 co-crystallized ligands of palm site I dataset used in cross-docking study.



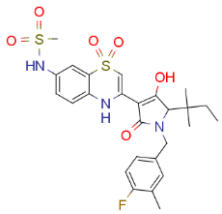
3GNW
IC_{50} = 0.0260 µM



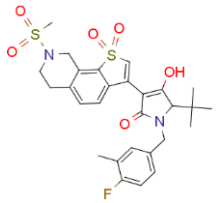
3GYN
IC_{50} = 0.0470 µM



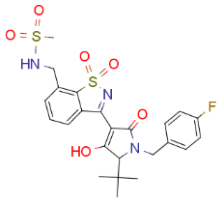
3H2L
IC_{50} = - µM



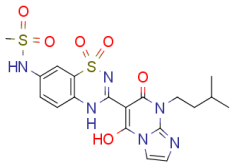
3H59
IC_{50} = 0.0130 µM



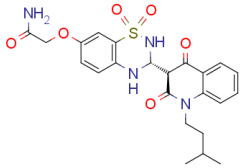
3H5S
IC_{50} = 0.0050 µM



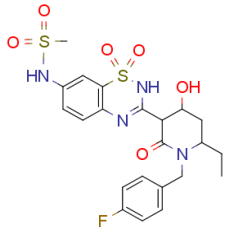
3H5U
IC_{50} = 0.0030 µM



3H98
IC_{50} = 0.0050 µM

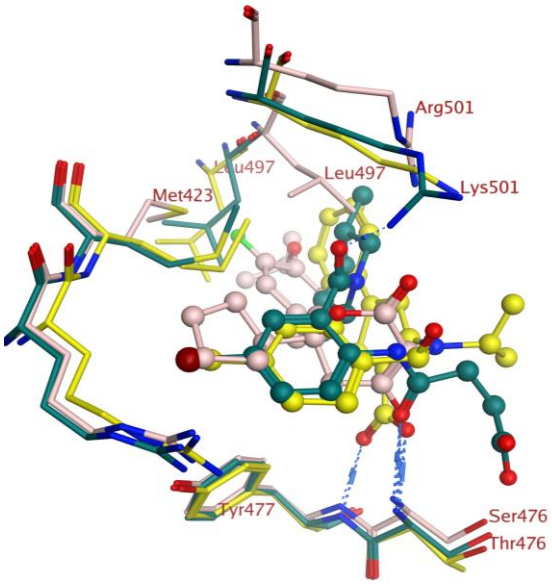


3HHK
IC_{50} = - µM



3IGV
IC_{50} = 0.0100 µM

Figure A2 (Continued). 29 co-crystallized ligands of palm site I dataset used in cross-docking study.



	Residual difference			
Group 1	Arg501	Leu419	Ile482	Ser476
Group 2	Lys501	Ile419	Leu482	Thr476
Group 3	Arg501	Leu419	Ile482	Ser476

Figure A3. Superimposition of X-ray structures PDB ID: 2HAI (Pink), 1YVX (Yellow) and 3CJ4 (Cyan) representing group 1, 2 and 3, respectively. The inhibitors of each group are shown by balls and sticks in the same color as their protein structures.

Table A3. Cross-docking results using Chem-score for 22 crystal structures of the thumb dataset (upper) and 29 crystal structures of the palm dataset (lower). The tables display the root mean square deviation (RMSD) and the correct predicted poses are labeled in blue shade.

CHEM score		Ligands																						Avg. RMSD
		1OS5	2D3U	2D3Z	2D41	2GIR	2HAI	2HWH	2HWI	2I1R	2O5D	3FRZ	1NHU	1NHV	2WHO	3CIZ	3CJ0	3CJ2	3CJ3	3CJ4	3CJ5	1YVX	1YVZ	
Proteins	1OS5	5.3	0.4	1.1	0.8	0.7	3.1	1.8	6.5	1.2	4.3	3.7	6.4	8.5	8.1	4.5	4.7	5.7	5.5	5.1	6.6	2.6	2.4	4.0
	2D3U	7.8	0.6	0.7	0.8	0.5	4.0	1.5	1.0	1.2	10.0	2.2	2.5	8.8	7.8	2.8	1.8	3.8	5.4	5.4	5.2	2.3	2.0	3.5
	2D3Z	4.2	0.7	1.0	1.0	0.9	3.3	6.6	1.1	1.2	2.9	2.9	2.5	8.1	7.8	2.7	2.0	3.7	5.4	5.7	5.5	2.5	2.2	3.3
	2D41	3.1	0.7	1.3	0.7	0.8	3.1	1.6	1.2	1.3	1.6	2.1	5.1	7.9	9.6	2.7	2.9	3.0	5.3	2.1	6.0	2.0	1.7	3.0
	2GIR	7.9	0.6	0.7	0.8	0.7	3.6	1.5	1.4	1.2	2.7	6.0	2.7	8.9	7.8	2.7	1.8	3.9	5.4	5.2	5.2	2.5	2.1	3.4
	2HAI	0.9	0.6	0.9	0.8	0.6	1.6	1.9	5.1	1.4	1.3	2.8	2.5	8.0	8.1	4.2	4.2	3.6	5.5	4.0	4.9	2.1	2.5	3.1
	2HWH	0.9	0.9	1.2	1.1	0.8	3.0	1.2	2.7	1.4	1.3	2.6	3.7	8.0	3.7	2.6	2.4	3.0	3.6	3.9	6.3	2.5	2.6	2.7
	2HWI	0.8	1.0	1.2	0.9	0.9	3.4	1.2	0.9	1.1	1.5	2.9	4.7	7.9	7.9	2.6	5.0	3.0	5.3	4.2	4.7	2.6	2.7	3.0
	2I1R	4.6	0.5	1.0	1.0	0.9	3.2	1.2	0.9	1.3	1.5	2.6	4.7	8.0	7.9	2.6	2.5	3.1	3.7	4.0	5.4	2.2	2.5	3.0
	2O5D	3.8	0.7	1.0	1.0	0.9	3.4	1.7	0.8	1.2	1.3	2.6	5.6	7.9	7.8	2.6	4.9	3.6	3.5	4.1	5.3	2.4	2.6	3.1
	3FRZ	1.2	0.8	1.1	0.8	0.8	1.4	1.9	5.4	1.7	4.5	2.3	3.9	7.9	8.1	2.7	2.4	3.6	5.6	1.9	5.3	2.4	2.7	3.1
	1NHU	2.9	2.2	2.6	2.2	1.6	2.7	2.0	3.4	4.1	4.2	4.1	5.6	7.2	8.0	2.3	4.1	0.7	0.7	3.2	3.9	2.1	2.1	3.3
	1NHV	3.4	2.7	2.5	2.5	2.2	2.7	3.6	4.0	3.8	6.0	1.7	5.6	7.3	8.1	3.6	1.2	1.1	4.9	1.0	1.5	1.2	1.3	3.3
	2WHO	1.6	1.7	2.3	2.3	1.5	3.1	1.7	4.0	4.3	6.3	2.6	6.4	8.4	8.2	2.4	1.6	3.9	1.0	0.9	1.5	1.0	1.2	3.1
	3CIZ	1.3	9.3	2.6	9.4	1.9	3.4	3.4	2.9	3.9	3.7	2.9	7.1	9.1	8.5	2.9	5.8	1.3	0.7	1.7	4.0	1.2	2.7	4.1
	3CJ0	1.8	10.0	2.6	9.8	1.8	3.4	3.1	3.2	2.3	3.8	8.1	6.5	9.8	4.7	2.8	0.7	3.7	3.1	3.4	3.8	2.9	3.2	4.3
	3CJ2	3.1	1.7	2.5	2.3	2.3	3.2	1.8	3.6	2.4	4.0	2.2	5.1	7.1	1.8	3.1	4.9	1.0	0.7	0.9	3.8	1.7	5.9	3.0
	3CJ3	8.0	9.2	3.0	9.6	5.5	2.9	3.8	4.0	1.6	6.1	2.8	5.2	7.4	8.6	2.1	5.8	3.5	0.7	4.2	4.4	3.0	5.7	4.9
	3CJ4	2.3	2.1	2.3	2.5	1.7	3.8	3.6	3.0	1.8	3.9	2.3	2.1	8.6	1.4	3.2	0.8	0.7	1.6	0.6	1.9	1.8	1.6	2.4
	3CJ5	1.6	2.0	2.5	2.5	2.2	3.4	3.3	3.5	2.4	10.8	2.7	7.1	7.4	1.4	3.1	5.9	0.8	0.8	1.1	3.7	1.5	1.3	3.2
	1YVX	2.7	2.5	1.8	2.0	1.9	2.9	3.2	7.6	1.7	10.2	2.7	4.7	7.6	8.6	2.7	3.6	1.5	2.8	4.1	4.5	1.0	1.0	3.7
	1YVZ	2.7	1.9	1.8	1.6	1.7	3.7	3.5	7.1	1.5	4.2	2.4	5.7	8.0	8.6	2.8	3.5	3.8	2.0	4.1	3.0	1.9	0.6	3.5

RMSD < 1 < 2 < 2.5

CHEM score		Ligands																													Avg RMSD
		1YVF	1Z4U	2GC8	2JC0	2JC1	2QE5	2GIQ	3BR9	3BSA	3BSC	3CDE	3CO9	3CVK	3E51	3GYN	3H2L	3H98	3IGV	3CWJ	3G86	3H59	3CSO	3GNV	3GNW	3HHK	3D28	3D5M	3H5S	3H5U	
Proteins	1YVF	0.8	0.6	3.6	5.0	4.9	7.2	6.8	4.4	4.3	5.1	2.1	2.4	1.3	4.2	4.5	0.8	1.9	1.4	4.5	4.2	0.6	6.7	6.7	7.0	7.7	0.6	0.8	1.6	0.6	3.5
	1Z4U	0.7	0.9	3.3	5.2	1.8	2.5	7.3	3.7	4.4	5.0	2.5	1.9	1.9	4.3	4.4	1.9	2.2	1.9	4.2	4.2	1.4	1.9	1.8	2.7	5.6	0.9	1.5	1.8	1.3	2.9
	2GC8	2.2	6.4	3.1	1.6	1.2	2.4	6.7	4.2	4.4	4.6	7.9	2.4	1.8	4.4	4.5	1.6	2.1	5.1	4.5	4.3	5.7	6.9	6.9	4.7	2.3	4.1	5.4	2.0	5.5	4.1
	2JC0	5.3	6.2	3.0	1.2	1.5	2.5	7.2	4.4	4.0	4.4	6.3	1.6	2.1	5.0	1.9	1.0	1.8	7.0	4.3	4.3	1.9	6.9	7.0	4.9	3.7	5.9	4.6	2.0	6.9	4.1
	2JC1	5.0	5.3	3.1	1.7	1.1	6.1	2.3	4.7	2.2	4.5	3.3	2.0	1.9	4.5	4.4	2.0	2.0	1.6	4.3	4.6	1.4	7.1	6.7	4.5	1.6	1.1	1.7	1.5	1.0	3.2
	2QE5	7.4	5.7	3.6	4.9	1.7	3.9	7.2	4.3	4.4	4.5	3.0	1.0	1.5	9.1	4.2	1.2	2.5	1.9	4.3	4.2	1.8	7.1	7.1	5.8	3.4	5.6	1.9	1.5	5.1	4.1
	2GIQ	7.8	5.7	2.6	1.2	1.3	5.1	2.0	4.1	4.4	4.6	7.9	1.3	2.0	4.3	1.2	1.0	1.2	2.0	4.3	4.3	1.9	6.8	6.8	5.1	2.1	0.8	2.1	1.7	1.0	3.3
	3BR9	5.0	5.7	2.8	1.4	1.3	4.3	7.1	4.3	4.7	4.7	7.6	1.3	2.0	4.2	4.2	1.7	2.5	5.8	4.3	4.6	1.8	5.1	4.8	5.1	3.5	4.0	5.6	1.9	0.9	3.9
	3BSA	5.0	5.6	3.5	4.8	1.5	3.9	7.4	4.1	4.4	4.7	7.8	3.7	1.8	4.6	1.0	1.6	2.4	1.8	4.2	4.8	1.9	7.5	6.7	6.0	3.4	4.4	1.7	1.8	1.1	3.9
	3BSC	4.7	5.6	3.2	4.9	0.8	2.8	6.3	4.2	4.5	4.8	7.7	3.7	1.8	5.2	1.1	1.5	1.7	1.5	4.9	4.6	1.9	7.3	5.1	5.9	3.6	0.9	1.6	1.9	1.0	3.6
	3CDE	5.0	5.6	3.2	1.2	1.0	3.2	7.0	4.5	4.6	4.6	7.6	1.2	2.0	4.5	4.4	1.1	2.4	1.8	4.2	4.3	4.7	7.5	6.8	4.5	3.3	3.9	5.3	2.0	0.9	3.9
	3CO9	5.1	5.6	3.0	1.4	1.2	4.0	4.0	4.2	9.8	4.6	7.8	1.3	2.1	4.2	4.2	1.9	7.8	1.9	4.1	4.3	1.7	3.5	7.0	8.7	8.7	0.7	1.7	2.0	1.3	4.1
	3CVK	4.7	5.4	2.8	2.2	0.8	4.0	7.1	4.3	4.6	4.4	7.6	1.4	1.7	9.1	4.4	0.8	9.1	1.9	4.0	4.4	2.0	3.4	6.9	5.9	1.9	2.6	1.2	1.3	0.9	3.8
	3E51	7.4	5.7	3.2	1.0	1.1	3.0	7.2	4.1	4.7	4.6	7.7	2.1	1.9	4.3	4.5	2.0	2.2	1.6	4.3	4.2	1.6	7.0	6.9	3.8	3.3	0.8	1.6	1.8	1.3	3.6
	3GYN	6.7	5.7	3.2	1.6	1.2	2.8	7.1	4.3	8.4	6.9	7.8	8.7	8.1	8.6	8.3	1.9	8.3	6.0	8.9	8.3	8.5	7.0	4.8	4.8	8.1	1.0	1.7	1.6	1.3	5.6
	3H2L	4.8	5.5	2.9	4.9	0.8	5.0	7.1	4.6	4.9	4.8	2.3	1.2	2.0	9.3	4.3	1.9	2.3	1.9	4.2	4.3	1.1	6.8	6.7	6.0	3.3	0.7	1.1	1.2	0.7	3.7
	3H98	5.0	5.8	2.8	4.9	1.2	4.0	4.2	4.5	4.4	4.7	7.8	1.3	1.7	4.7	4.3	1.1	3.5	1.8	4.3	4.3	5.7	5.8	5.1	5.0	4.1	0.9	1.6	1.5	4.7	3.8
	3IGV	4.7	5.3	2.8	5.0	1.1	4.0	7.2	4.6	4.6	4.9	7.7	1.2	2.0	4.3	4.5	1.9	2.2	1.8	4.3	4.3	1.6	3.6	7.0	6.1	3.2	2.5	1.2	1.6	1.0	3.7
	3CWJ	4.9	5.5	2.7	5.7	1.8	4.5	7.2	4.1	4.5	4.3	3.7	1.3	1.9	4.2	4.5	1.9	3.1	5.2	4.6	4.3	5.9	6.0	7.0	4.6	1.2	3.9	4.9	1.9	4.9	4.2
	3G86	4.7	5.4	3.5	4.9	1.2	4.0	7.1	4.0	4.5	4.9	2.3	1.5	2.0	4.2	4.0	2.0	1.4	1.9	3.9	4.3	0.8	3.6	6.7	2.8	3.3	2.5	1.9	1.2	1.1	3.3
	3H59	4.8	5.3	3.0	1.5	5.7	5.2	4.1	4.3	4.4	5.4	2.5	2.3	2.0	2.2	4.5	1.9	3.9	1.8	4.0	1.9	1.0	3.6	6.5	6.9	2.3	0.6	1.0	1.5	1.3	3.3
	3CSO	5.0	0.4	2.8	4.9	1.1	3.4	4.1	4.6	4.4	4.7	7.3	1.5	5.4	3.9	6.7	1.4	1.5	5.6	4.3	4.1	6.2	1.7	1.9	6.2	5.9	5.0	6.0	1.2	6.2	4.1
	3GNV	5.1	5.5	3.6	5.0	1.0	3.9	6.3	3.7	1.9	1.4	6.6	0.9	5.9	4.3	6.9	1.9	6.2	5.5	3.9	4.3	6.1	2.2	1.5	6.2	6.1	4.3	1.1	1.3	5.7	4.1
	3GNW	5.8	5.5	3.3	5.8	1.1	2.0	6.3	4.6	4.7	4.8	2.8	2.1	2.0	4.3	4.5	2.4	6.2	2.0	4.6	4.3	5.8	5.4	1.3	0.6	6.4	4.4	1.5	1.6	5.9	3.9
	3HHK	5.2	5.7	3.3	5.2	1.4	3.9	7.0	4.2	1.3	5.1	2.4	1.0	1.8	4.3	4.4	1.0	1.5	2.3	3.8	4.2	4.8	6.4	6.8	4.9	3.2	0.6	2.0	2.0	1.2	3.5
	3D28	5.8	6.1	2.9	2.1	1.3	5.4	6.9	4.5	4.6	4.4	2.4	2.4	1.4	4.0	4.3	0.8	2.3	1.5	4.3	4.7	1.0	7.1	6.8	2.4	2.3	5.6	0.9	0.8	1.0	3.4
	3D5M	4.9	5.6	2.8	2.0	1.1	4.1	7.2	4.6	4.9	5.0	2.6	1.3	1.7	4.7	4.5	1.9	9.7	1.9	4.7	4.4	1.6	3.7	6.6	6.0	2.3	2.5	1.3	1.3	0.9	3.7
	3H5S	4.9	5.6	3.0	5.4	5.7	5.2	4.8	4.3	4.9	4.9	2.2	1.2	1.9	3.9	4.2	0.9	1.3	1.9	4.0	4.4	1.7	6.7	6.7	2.3	3.2	2.6	1.0	1.2	0.8	3.5
	3H5U	5.0	5.5	2.9	1.5	5.6	4.1	6.6	4.6	4.3	5.1	7.9	1.3	1.4	4.4	4.1	1.2	2.4	2.1	4.0	4.3	1.3	3.9	6.5	6.2	8.0	0.6	0.7	1.2	1.2	3.7

RMSD < 1 < 2 < 2.5

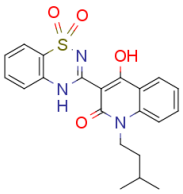
Table A4. Cross-docking results of the thumb dataset (22 crystal structures) using 6 scoring functions; Gold-score, Chem-score, ASP-score, P-score, PMF and GLIDE SP. Average root mean square deviation (RMSD) and the numbers of correct poses are presented according to their protein cluster group. A color bar in the 'total' column denotes the percentage of correct pose prediction.

Structure (PDB ID)		Gold score							Chem score							ASP score						
		Avg RMSD			No. of corrected poses				Avg RMSD			No. of corrected poses				Avg RMSD			No. of corrected poses			
		Gr.1	Gr.2	Gr.3	Gr.1	Gr.2	Gr.3	Total	Gr.1	Gr.2	Gr.3	Gr.1	Gr.2	Gr.3	Total	Gr.1	Gr.2	Gr.3	Gr.1	Gr.2	Gr.3	Total
Group 1 (11 cpxs.)	1OS5	6.9	7.2	6.9	2	0	0	2	2.6	2.5	6.1	6	1	0	7	2.7	2.4	5.8	7	1	0	8
	2D3U	5.9	4.6	6.8	4	1	0	5	2.7	2.1	4.8	8	2	1	11	2.0	2.1	6.3	8	2	0	10
	2D3Z	6.0	7.8	6.6	3	0	0	3	2.3	2.3	4.8	6	2	1	9	1.7	2.1	6.3	8	2	0	10
	2D41	7.0	7.7	6.6	2	0	0	2	1.6	1.9	4.9	9	2	1	12	1.5	2.0	6.2	9	2	0	11
	2GIR	6.9	2.3	7.0	2	1	0	3	2.5	2.3	4.8	7	2	1	10	2.0	2.1	6.4	7	2	0	9
	2HAI	6.1	7.6	6.3	4	0	0	4	1.6	2.3	5.0	9	2	0	11	1.7	2.3	6.3	9	1	0	10
	2HWH	6.3	7.3	5.9	3	0	0	3	1.5	2.6	4.1	8	0	1	9	1.7	2.4	5.9	9	2	0	11
	2HWI	5.3	5.0	5.5	4	0	0	4	1.4	2.6	5.0	9	0	0	9	1.9	2.5	5.7	8	2	0	10
	2I1R	5.8	7.9	6.3	4	0	0	4	1.7	2.4	4.6	8	2	1	11	2.5	2.4	6.0	7	2	0	9
	2O5D	6.2	8.0	6.2	4	0	0	4	1.7	2.5	5.0	8	1	0	9	1.8	2.4	6.1	9	1	0	10
	3FRZ	5.3	8.2	6.2	5	0	0	5	2.0	2.5	4.6	9	1	2	12	2.0	2.5	5.8	8	2	0	10
Group 2 (2 cpxs.)	1YVX	7.5	1.1	6.2	2	2	0	4	3.6	1.0	4.5	4	2	1	7	2.8	1.1	5.5	6	2	1	9
	1YVZ	7.9	0.8	7.3	2	2	0	4	2.9	1.3	4.6	6	2	1	9	2.8	1.2	4.2	7	2	4	13
Group 3 (9 cpxs.)	1NHU	6.4	2.3	4.8	1	1	2	4	2.9	2.1	4.0	4	2	3	9	4.6	2.0	4.6	2	2	4	8
	1NHV	8.4	8.5	6.9	0	0	0	0	3.2	1.3	3.8	3	2	4	9	6.1	1.9	4.7	3	2	3	8
	2WHO	5.9	1.6	5.1	2	2	3	7	2.9	1.1	3.8	6	2	5	13	4.4	1.7	3.3	3	2	5	10
	3CIZ	7.8	1.9	6.7	2	2	1	5	4.1	2.0	4.6	2	1	3	6	6.4	2.1	4.4	1	2	4	7
	3CJ0	8.3	8.6	8.2	1	0	0	1	4.5	3.1	4.3	3	0	1	4	6.1	5.0	6.4	3	0	0	3
	3CJ2	7.6	1.9	7.3	0	1	0	1	2.7	3.8	3.2	7	1	4	12	4.7	1.9	4.6	2	2	3	7
	3CJ3	9.1	3.6	6.8	0	0	0	0	5.1	4.3	4.7	1	0	2	3	5.3	2.1	6.3	2	2	1	5
	3CJ4	7.6	2.1	4.1	0	2	2	4	2.7	1.7	2.3	7	2	7	16	5.4	1.7	4.8	2	2	3	7
	3CJ5	7.0	1.8	5.1	2	2	2	6	3.4	1.4	3.5	5	2	4	11	4.0	2.1	4.5	3	2	3	8

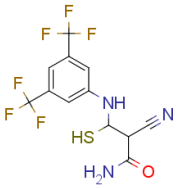
Structure (PDB ID)		P-score							PMF							Glide SP						
		Avg RMSD			No. of corrected poses				Avg RMSD			No. of corrected poses				Avg RMSD			No. of corrected poses			
		Gr.1	Gr.2	Gr.3	Gr.1	Gr.2	Gr.3	Total	Gr.1	Gr.2	Gr.3	Gr.1	Gr.2	Gr.3	Total	Gr.1	Gr.2	Gr.3	Gr.1	Gr.2	Gr.3	Total
Group 1 (11 cpxs.)	1OS5	3.8	3.4	6.2	4	1	0	5	5.2	2.5	5.8	3	1	0	4	4.5	2.2	6.2	6	2	0	8
	2D3U	2.4	3.3	5.7	6	1	0	7	3.5	2.2	5.5	4	1	0	5	3.6	3.5	5.1	5	1	3	9
	2D3Z	3.4	3.4	5.9	5	1	0	6	2.9	1.9	5.7	6	2	0	8	3.3	2.6	5.3	6	1	0	7
	2D41	2.8	3.1	6.0	5	1	0	6	3.6	3.7	5.8	5	1	0	6	4.2	7.1	5.7	5	0	0	5
	2GIR	3.2	3.3	6.4	5	1	0	6	3.2	4.8	5.8	5	1	0	6	3.3	2.1	6.0	5	2	0	7
	2HAI	3.2	3.2	5.9	4	1	0	5	4.8	4.1	6.1	3	1	0	4	4.4	4.9	4.8	4	1	0	5
	2HWH	2.3	3.1	5.9	6	1	0	7	3.2	6.2	5.8	5	0	0	5	5.9	4.7	5.9	4	1	0	5
	2HWI	2.9	3.2	6.1	5	1	0	6	3.3	2.2	5.5	5	2	0	7	4.6	1.9	5.6	5	2	0	7
	2I1R	2.8	3.1	6.0	5	1	0	6	3.4	1.9	6.2	4	2	0	6	4.3	2.4	5.9	4	1	1	6
	2O5D	2.6	3.2	6.4	5	1	0	6	4.7	4.0	6.1	2	1	0	3	4.6	2.6	5.8	3	1	0	4
	3FRZ	3.3	3.3	5.8	5	1	0	6	4.0	3.6	6.2	4	0	0	4	2.6	1.9	4.7	6	2	0	8
Group 2 (2 cpxs.)	1YVX	3.2	3.0	5.5	4	1	0	5	4.0	1.8	5.5	4	2	0	6	5.3	1.0	4.0	5	2	0	7
	1YVZ	3.1	2.8	5.8	3	1	1	5	4.0	1.8	5.2	4	2	1	7	4.8	0.5	4.8	4	2	0	6
Group 3 (9 cpxs.)	1NHU	4.2	3.4	5.6	0	0	0	0	5.1	2.6	4.1	0	1	2	3	4.3	1.5	4.9	3	2	2	7
	1NHV	4.0	3.2	5.5	3	1	0	4	4.2	5.5	5.2	1	0	0	1	6.1	1.1	5.3	1	2	1	4
	2WHO	3.7	3.2	5.5	0	1	0	1	5.8	1.7	4.7	0	2	2	4	5.4	2.9	3.5	0	0	3	3
	3CIZ	3.5	3.6	5.6	3	0	0	3	5.1	2.3	4.7	0	2	2	4	5.2	1.6	3.6	2	2	3	7
	3CJ0	3.6	2.1	6.6	2	1	0	3	4.4	1.8	5.6	1	2	0	3	8.6	1.2	6.5	1	2	1	4
	3CJ2	3.8	3.2	4.8	1	1	0	2	4.3	2.7	4.5	0	1	2	3	5.6	1.3	3.4	1	2	2	5
	3CJ3	3.9	3.4	5.8	1	0	0	1	5.3	2.6	5.1	1	1	1	3	6.7	1.4	4.4	0	2	3	5
	3CJ4	3.7	3.1	5.7	1	1	0	2	4.2	2.7	4.5	2	1	2	5	5.5	1.3	4.0	3	2	3	8
	3CJ5	3.7	3.1	5.4	2	1	0	3	4.6	3.0	4.4	0	0	2	2	5.6	1.5	3.4	1	2	4	7

Table A5. Cross-docking results of the palm dataset (29 crystal structures) using 6 scoring functions; Gold-score, Chem-score, ASP-score, P-score, PMF and GLIDE SP. A color bar denotes the percentage of correct pose prediction.

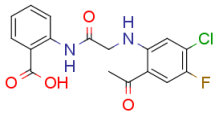
Proteins (PDB ID)	Gold score		Chem score		ASP score		PMF score		p-score		Glide SP	
	Avg RMSD	No. of correct poses	Avg RMSD	No. of correct poses	Avg RMSD	No. of correct poses	Avg RMSD	No. of correct poses	Avg RMSD	No. of correct poses	Avg RMSD	No. of correct poses
1YVF	4.6	4	3.5	13	4.3	5	3.7	13	3.3	13	3.8	13
1Z4U	4.6	7	2.9	17	3.4	10	4.0	8	3.0	13	2.7	17
2GC8	3.9	11	4.1	10	4.1	7	4.3	7	3.8	6	3.8	8
2GIQ	4.1	6	3.3	15	3.6	8	3.9	7	3.5	10	3.6	10
2JC0	3.6	11	4.1	10	4.2	6	4.8	4	4.1	5	4.2	6
2JC1	3.6	11	3.2	15	3.3	9	3.8	12	3.4	11	3.3	14
2QE5	4.6	7	4.1	9	4.1	7	4.1	5	4.0	9	4.2	10
3BR9	3.9	10	3.9	9	4.0	9	4.1	4	4.0	8	4.0	9
3BSA	3.9	11	3.9	10	4.1	9	3.5	10	3.8	9	3.7	8
3BSC	3.8	11	3.6	11	4.1	7	4.4	8	3.3	12	4.3	7
3CDE	4.6	8	3.9	9	3.8	8	3.7	6	3.9	8	3.8	9
3CO9	3.9	10	4.1	11	3.6	9	4.6	6	3.5	11	3.6	10
3CS0	2.6	18	4.1	8	3.4	10	3.7	15	2.4	17	3.0	12
3CVK	3.3	12	3.8	11	3.2	10	3.4	12	3.1	11	3.1	12
3CWJ	4.3	9	4.2	6	3.5	9	3.8	8	3.5	12	3.4	12
3D28	3.0	14	3.4	14	3.7	7	3.4	11	3.3	9	3.6	10
3D5M	3.5	11	3.7	11	3.1	11	4.2	9	3.1	12	2.8	12
3E51	4.0	9	3.6	12	3.7	9	3.8	7	3.9	8	4.0	9
3G86	4.1	11	3.3	11	3.4	9	3.4	13	3.1	11	2.9	13
3GNV	2.2	20	4.1	9	3.2	10	3.4	14	2.7	16	2.8	14
3GNW	2.9	14	3.9	10	3.6	9	4.8	7	3.1	13	3.4	13
3GYN	3.9	8	5.6	7	3.7	7	4.7	8	4.1	8	3.6	12
3H2L	3.7	11	3.7	12	3.2	10	3.7	12	2.9	14	3.2	12
3H59	3.6	11	3.3	13	3.2	10	3.7	12	2.9	11	2.8	15
3H5S	3.2	14	3.5	11	3.3	8	3.5	12	2.9	13	3.0	10
3H5U	3.3	13	3.7	11	3.4	9	3.9	10	3.1	11	3.1	15
3H98	4.4	9	3.8	8	3.6	9	3.6	9	4.0	6	3.9	7
3HHK	4.5	8	3.5	12	3.6	9	4.0	8	3.5	9	3.6	8
3IGV	3.2	13	3.7	11	3.4	9	3.7	11	2.9	12	3.1	12



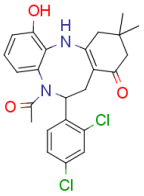
2FVC
IC_{50} = - µM



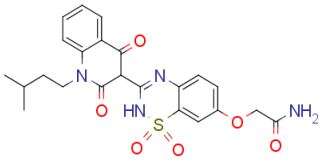
2IJN
IC_{50} = 5.900 µM



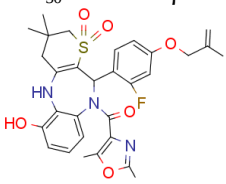
2QE2
IC_{50} = 0.0170 µM



3GOL
IC_{50} = 0.0810 µM



3HHK
IC_{50} = - µM



3HKW
IC_{50} = - µM

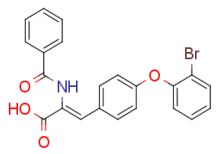
Figure A4. Additional co-crystallized ligands of palm site I dataset used in ensemble study.

Table A6. 45 ligands used in the rescoring study.

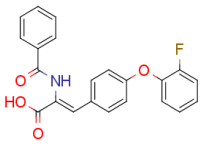
Name	IC_{50} (µM)	No. in source	Source
1YVF	0.100	59	[198]
1YVF-A1	6.900	63	[198]
1YVF-A2	48.000	68	[198]
1YVF-I1	Inactive	72	[198]
1YVF-I2	Inactive	55	[198]
1YVF-I3	Inactive	53	[198]
1Z4U	0.070	49	[198]
1Z4U-A1	2.800	46	[198]
1Z4U-A2	0.850	50	[198]
1Z4U-I1	Inactive	66	[198]
1Z4U-I2	Inactive	67	[198]
2FVC	0.032	2	[187]
2FVC-A1	0.340	72	[187]
2FVC-A2	0.105	74	[187]
2FVC-A3	10.000	105	[187]
2FVC-A4	10.000	104	[187]
2GC8	3.100	6	[215]
2GC8-A1	0.080	25	[215]
2GC8-A2	0.770	21	[215]

Table A6 (Continued). 45 ligands used in the rescoring study.

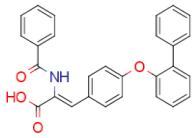
Name	IC_{50} (µM)	No. in source	Source
2GIQ	0.260		[192]
2JC0	20.000	1a	[216]
2JC1	3.800	7a	[216]
2QE2	0.017	14i	[217]
2QE5	1.640	3a	[217]
3BR9	0.070	8d	[218]
3BR9-A1	1.300	8b	[218]
3BR9-A2	0.340	8c	[218]
3BR9-A3	7.200	8r	[218]
3BSC	33.000	3	[218]
3CO9	0.010	3c	[219]
3CSO	1.000	4a	[205]
3CVK	0.010	4c	[220]
3CWJ	0.010	3a	[221]
3D5M	0.010	34	[208]
3E51	0.014	2e	[222]
3G86	0.005	18	[209]
3GNV	0.093	(R)-1b	[156]
3GNW	0.026	(S)-4c	[156]
3GOL	0.081	(R)-11d	[223]
3GYN	0.047	49	[224]
3H59	0.013	20	[212]
3H5S	0.005	16	[72]
3H5U	0.003	1b or 5	[72]
3H98	0.005	4a	[72]
3IGV	0.010	29	[224]



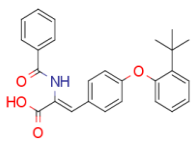
1YVF
IC_{50} = 0.10 µM



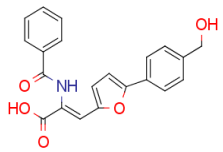
1YVF-A1
IC_{50} = 6.90 µM



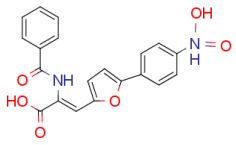
1YVF-A2
IC_{50} = 48.00 µM



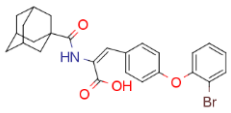
1YVF-I1
IC_{50} = 100.00 µM



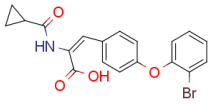
1YVF-I2
IC_{50} = 100.00 µM



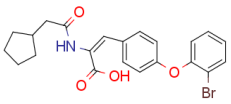
1YVF-I3
IC_{50} = 100.00 µM



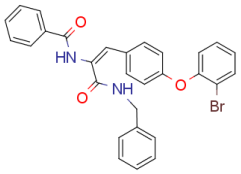
1Z4U
IC_{50} = 0.07 µM



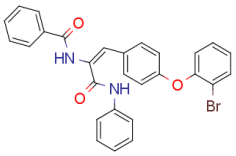
1Z4U-A1
IC_{50} = 2.80 µM



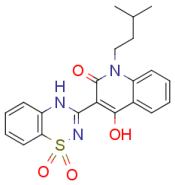
1Z4U-A2
IC_{50} = 0.85 µM



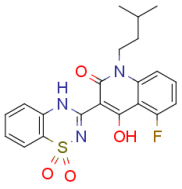
1Z4U-I1
IC_{50} = 100.00 µM



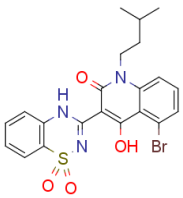
1Z4U-I2
IC_{50} = 100 µM



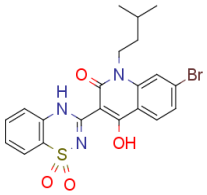
2FVC
IC_{50} = 0.03 µM



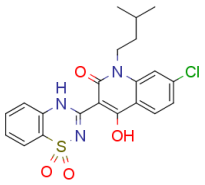
2FVC-A1
IC_{50} = 0.34 µM



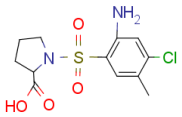
2FVC-A2
IC_{50} = 0.11 µM



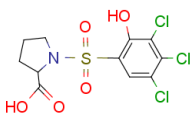
2FVC-A3
IC_{50} = 10.00 µM



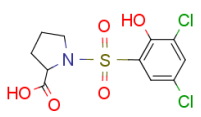
2FVC-A4
IC_{50} = 10.00 µM



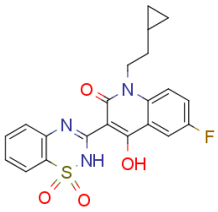
2GC8
IC_{50} = 3.10 µM



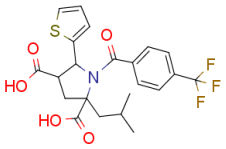
2GC8-A1
IC_{50} = 0.08 µM



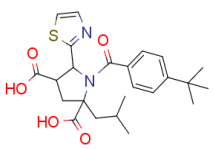
2GC8-A2
IC_{50} = 0.77 µM



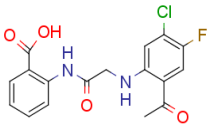
2GIQ
IC_{50} = 0.26 µM



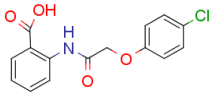
2JC0
IC_{50} = 20.00 µM



2JC1
IC_{50} = 3.80 µM

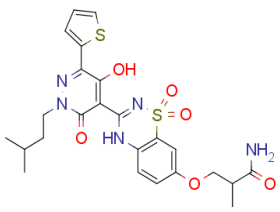


2QE2
IC_{50} = 0.02 µM

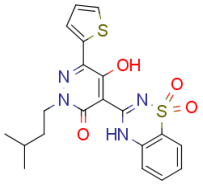


2QE5
IC_{50} = 1.64 µM

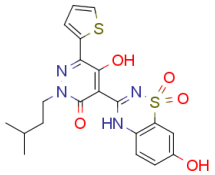
Figure A5. 45 ligand structures used in rescoring study.



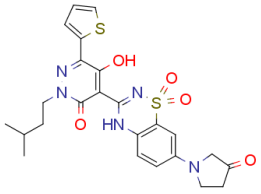
3BR9
IC_{50} = 0.07 µM



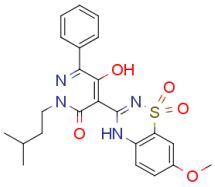
3BR9-A1
IC_{50} = 1.30 µM



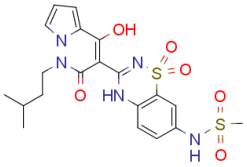
3BR9-A2
IC_{50} = 0.34 µM



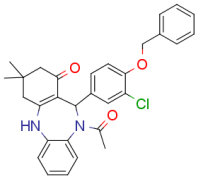
3BR9-A3
IC_{50} = 7.20 µM



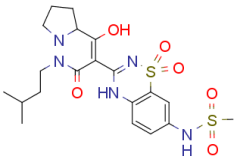
3BSC
IC_{50} = 33.00 µM



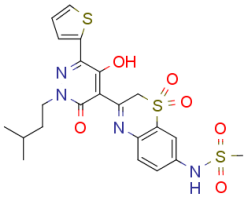
3CO9
IC_{50} = 0.01 µM



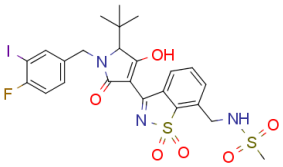
3CS0
IC_{50} = 1.00 µM



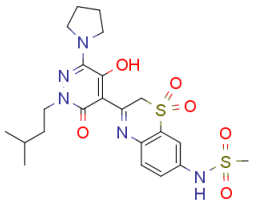
3CVK
IC_{50} = 0.01 µM



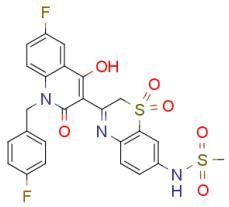
3CWJ
IC_{50} = 0.0100 µM



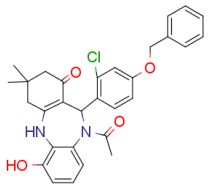
3D5M
IC_{50} = 0.0100 µM



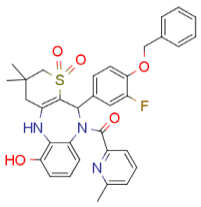
3E51
IC_{50} = 0.0140 µM



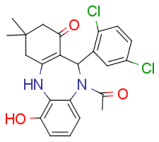
3G86
IC_{50} = 0.0050 µM



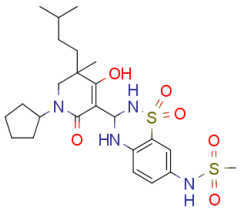
3GNV
IC_{50} = 0.0930 µM



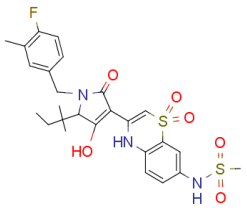
3GNW
IC_{50} = 0.0260 µM



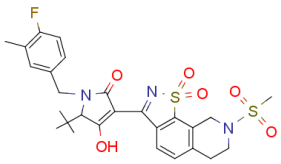
3GOL
IC_{50} = 0.0810 µM



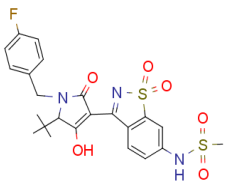
3GYN
IC_{50} = 0.0470 µM



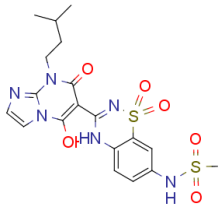
3H59
IC_{50} = 0.0130 µM



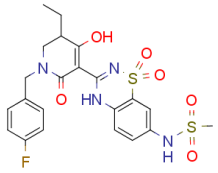
3H5S
IC_{50} = 0.0050 µM



3H5U
IC_{50} = 0.0030 µM



3H98
IC_{50} = 0.00050 µM



3IGV
IC_{50} = 0.0100 µM

Figure A5 (Continued). 45 ligand structures used in rescoring study.

Table A7. 3D Padel descriptors used in this study.

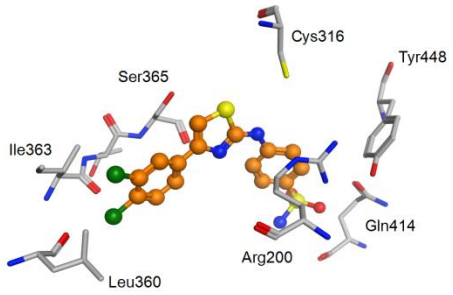
Descriptor type	Number	Descriptor
Charged partial surface area	29	PPSA-1, PPSA-2, PPSA-3, PNSA-1, PNSA-2, PNSA-3, DPSA-1, DPSA-2, DPSA-3, FPSA-1, FPSA-2, FPSA-3, FNSA-1, FNSA-2, FNSA-3, WPSA-1, WPSA-2, WPSA-3, WNSA-1, WNSA-2, WNSA-3, RPCG, RNCG, RPCS, RNCS, THSA, TPSA, RHSA, RPSA
Gravitational index	9	GRAV-1, GRAV-2, GRAV-3, GRAVH-1, GRAVH-2, GRAVH-3, GRAV-4, GRAV-5, GRAV-6
Length over breadth	2	LOBMAX, LOBMIN
Moment of inertia	7	MOMI-X, MOMI-Y, MOMI-Z, MOMI-XY, MOMI-XZ, MOMI-YZ, MOMI-R
Petitjean shape index	3	geomRadius, geomDiameter, geomShape
WHIM	65	WA.eneg, WA.mass, WA.polar, WA.unity, WA.volume, WD.eneg, WD.mass, WD.polar, WD.unity, WD.volume, Weta1.eneg, Weta1.mass, Weta1.polar, Weta1.unity, Weta1.volume, Weta2.eneg, Weta2.mass, Weta2.polar, Weta2.unity, Weta2.volume, Weta3.eneg, Weta3.mass, Weta3.polar, Weta3.unity, Weta3.volume, WK.eneg, WK.mass, WK.polar, WK.unity, WK.volume, Wlambda1.eneg, Wlambda1.mass, Wlambda1.polar, Wlambda1.unity, Wlambda1.volume, Wlambda2.eneg, Wlambda2.mass, Wlambda2.polar, Wlambda2.unity, Wlambda2.volume, Wlambda3.eneg, Wlambda3.mass, Wlambda3.polar, Wlambda3.unity, Wlambda3.volume, Wnu1.eneg, Wnu1.mass, Wnu1.polar, Wnu1.unity, Wnu1.volume, Wnu2.eneg, Wnu2.mass, Wnu2.polar, Wnu2.unity, Wnu2.volume, WT.eneg, WT.mass, WT.polar, WT.unity, WT.volume, WV.eneg, WV.mass, WV.polar, WV.unity, WV.volume,

Table A8. MOE descriptors used in this study.

Code	Description
ASA	Water accessible surface area
ASA_H	Total hydrophobic surface area
E	Potential energy
E_ang	Angle bend energy
E_ele	Electrostatic energy
E_nb	Non-bonded energy
E_oop	Out of plane energy
E_sol	Solvation energy
E_stb	Stretch bend energy
E_str	Bond stretch energy
E_strain	E minus energy of the local minimum
E_tor	Torsion energy
E_vdw	Van der Waals energy
PM3_Eele	Electronic energy (kcal/mol)
PM3_HF	Heat of formation (kcal)
PM3_HOMO	HOMO energy (eV)
PM3_IP	Ionization potential (kcal/mol)
PM3_LUMO	LUMO energy (kcal/mol)
pmi	Principal moment of inertia
vol	Van der Waals volume
VSA	Van der Waals surface area
vsurf_A	Ampiphilic moment
vsurf_CP	Critical packing parameter
vsurf_CW1	Capacity factor at -0.2
vsurf_D1	Hydrophobic volume at -0.2
vsurf_DD12	vsurf_EDmin1, vsurf_EDmin2 distance
vsurf_EDmin1	Lowest hydrophobic energy
vsurf_EWmin1	Lowest hydrophilic energy
vsurf_G	Surface globularity
vsurf_HB1	H-bond donor capacity at -0.2
vsurf_HL1	First hydrophilic-lipophilic balance
vsurf_ID1	Hydrophobic integy momet at -0.2
vsurf_IW1	Hydrophilic integy moment at -0.2
vsurf_R	Surface rugosity
vsurf_S	Interaction field area
vsurf_V	Interaction field volume
vsurf_W1	Hydrophilic volume at -0.2
vsurf_Wp1	Polar volume at -0.2

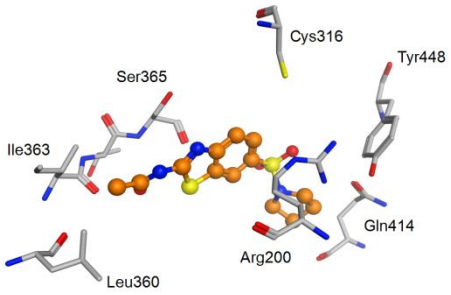
C1:

Gold-score = 64.27



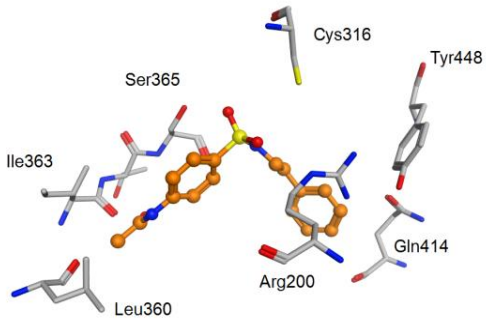
C2:

Gold-score = 60.70



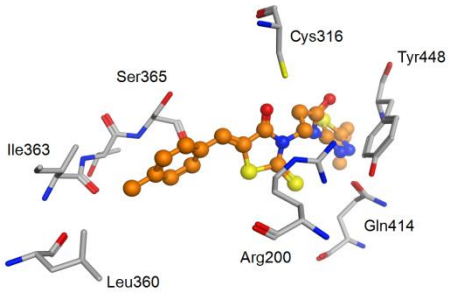
C3

: Gold-score = 52.84



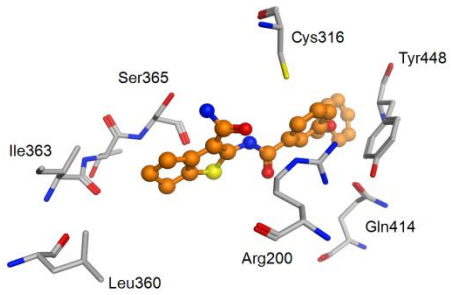
C4

: Gold-score = 60.19



C5

: Gold-score = 58.07



C6

: Gold-score = 56.78

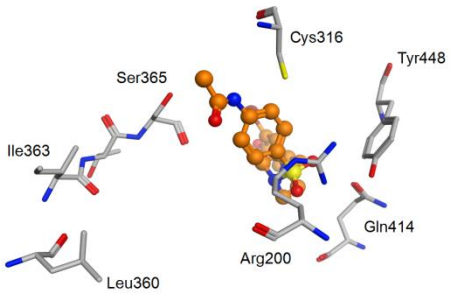
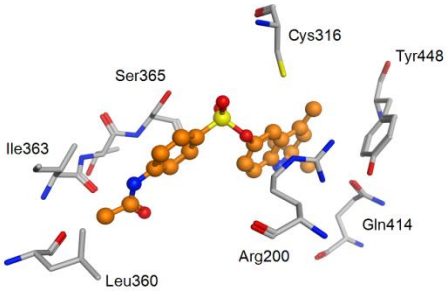
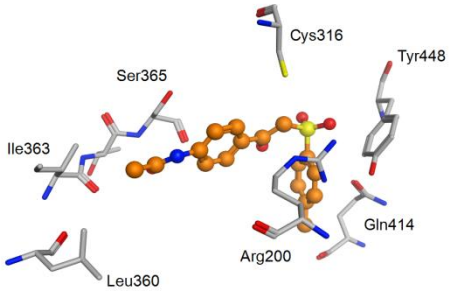


Figure A6. Docked poses of 14 tested compounds (Dataset-I) and their Gold-scores.

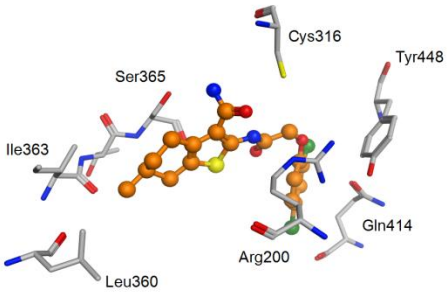
C7
: Gold-score = 54.12



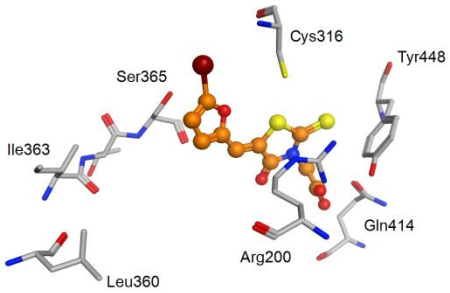
C8
: Gold-score = 53.35



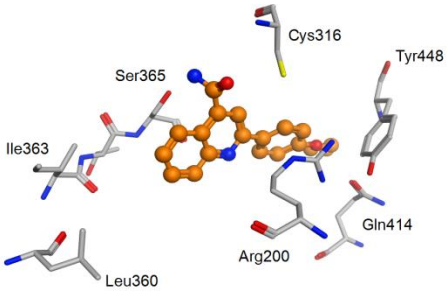
C9
: Gold-score = 52.84



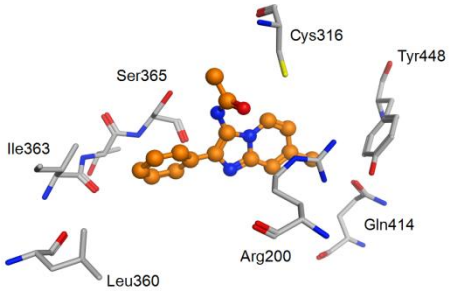
C10
: Gold-score = 52.34



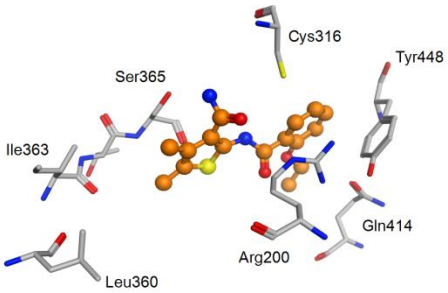
C11
: Gold-score = 51.50



C12
: Gold-score = 51.16



C13
: Gold-score = 50.84



C14
: Gold-score = 54.00

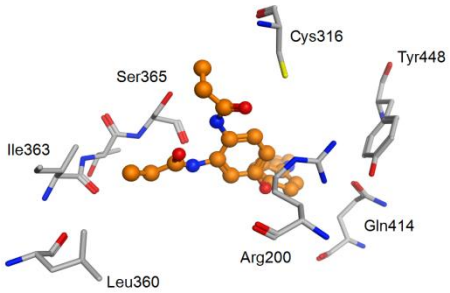
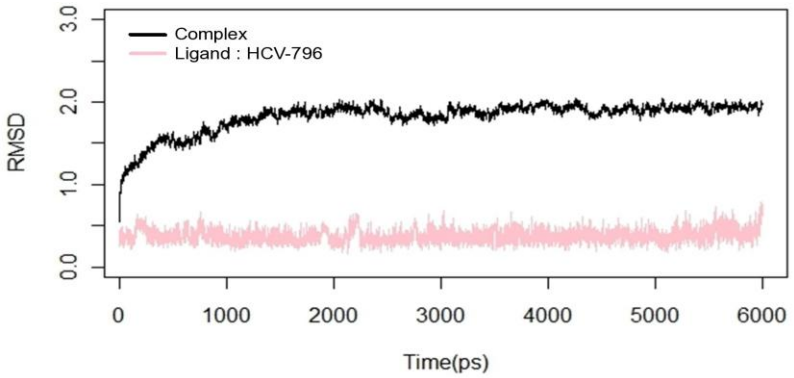
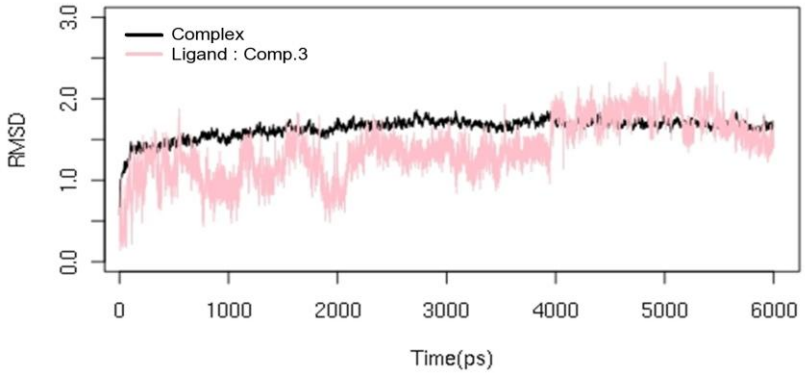


Figure A6 (Continued). Docked poses of 14 tested compounds (Dataset-I).

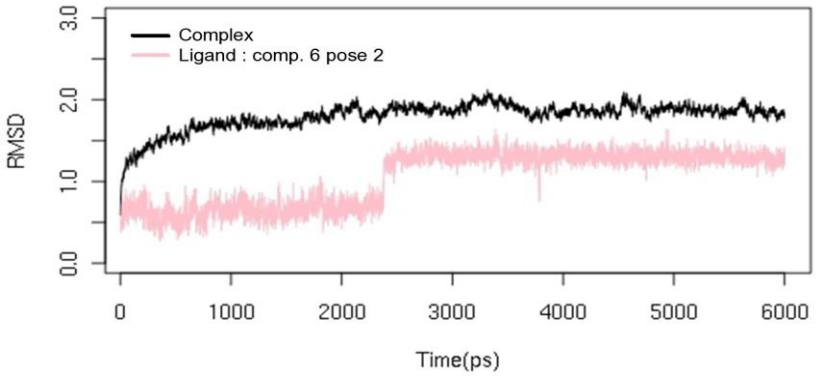
HCV796



C3



C6



C7

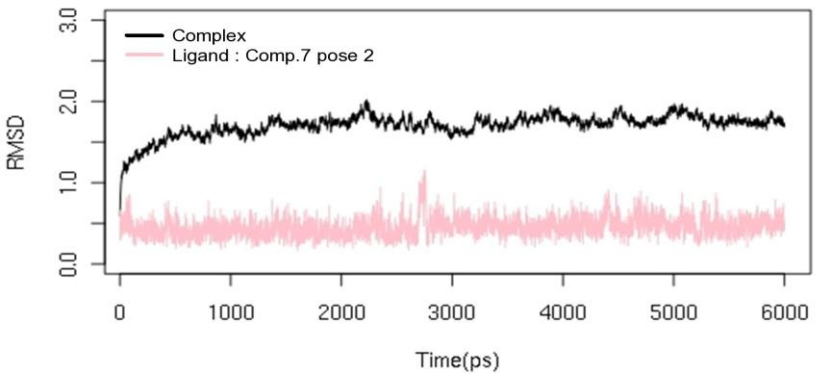
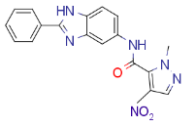
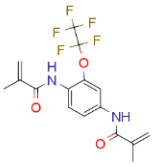
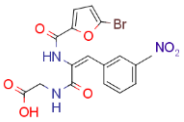
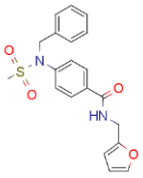
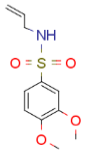
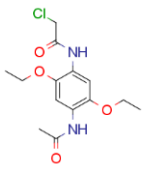
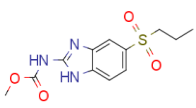


Figure A7. Root Mean Square Deviations (RMSD) plots of the complexes and ligands during MD simulation.

Table A9. 30 compounds from the study of Kim et al [171].

Source index	Structures	RdRp activity (@20 μM)
1		54%
2		48%
3		-
4		8%
5		-
6		-
7		9%
8		-

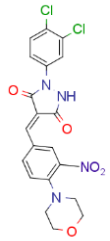
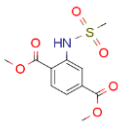
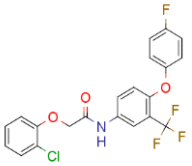
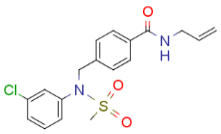
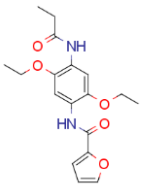
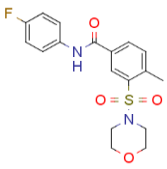
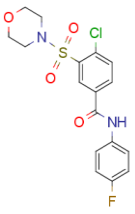
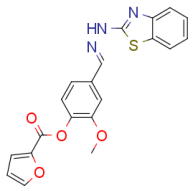
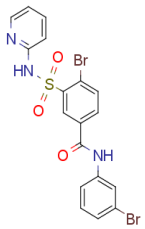
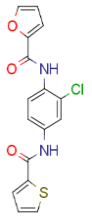
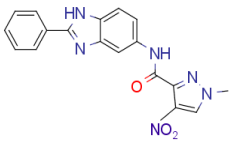
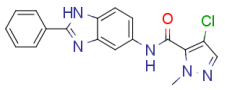
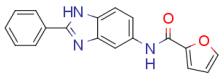
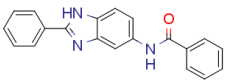
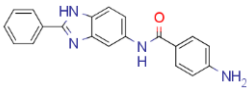
Source index	Structures	RdRp activity (@20 μM)
9		13%
10		-
11		-
12		-
13		-
14		-
15		-

Table A9 (Continued). 30 compounds from the study of Kim et al [171].

Source index	Structures	RdRp activity (@20 μM)
16		-
17		-
18		-
19		14%
20		37.7%
21		-
22		20%
23		19.7%

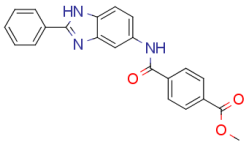
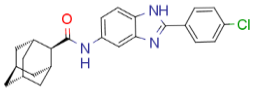
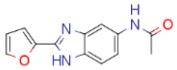
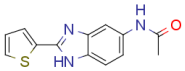
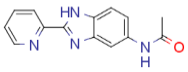
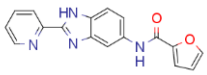
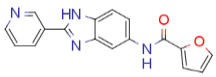
Source index	Structures	RdRp activity (@20 μM)
24		90%
25		-
26		42%
27		18.5%
28		58.6
29		42.4%
30		49.9%

Table A10. Root mean square deviation (RMSD) of re-docking HCV-796 into 3FQK and 3FQL structure.

Docking/Scoring	RMSD (STD) Å	
	3FQL	3FQK
Gold-score	1.89 (0.15)	1.17 (0.01)
Chem-score	2.03 (2.00)	3.79 (3.32)
ASP-score	8.39 (0.55)	5.87 (3.05)
Glide SP	0.96 (0.17)	0.55 (0.01)
P score	1.17 (0.37)	1.43 (0.65)

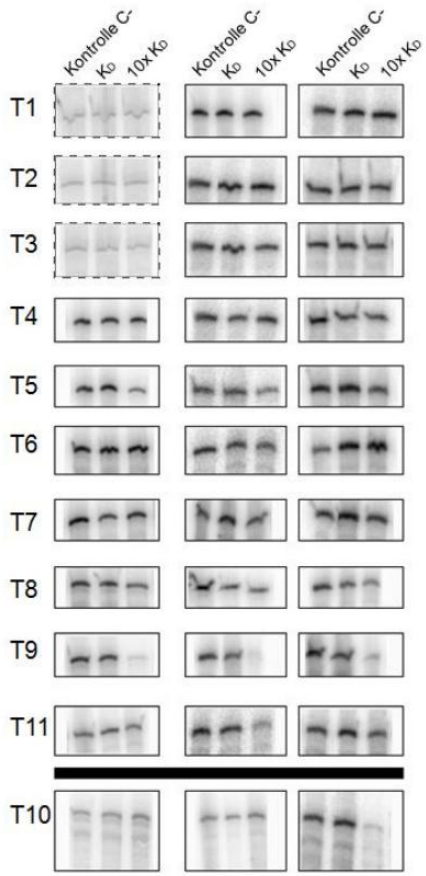
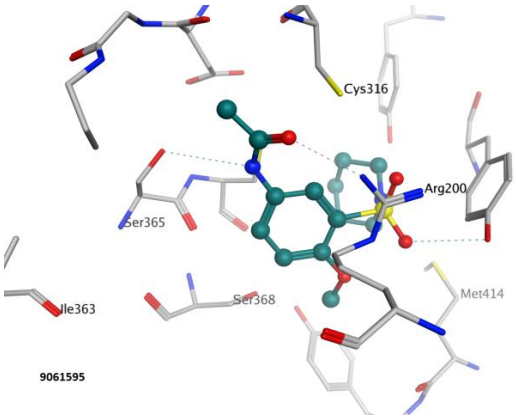


Figure A8. The autoradiography of the HCV NS5B polymerase assays with the inhibitors T1-T11 (dataset-II). Each compound was tested independently three times.

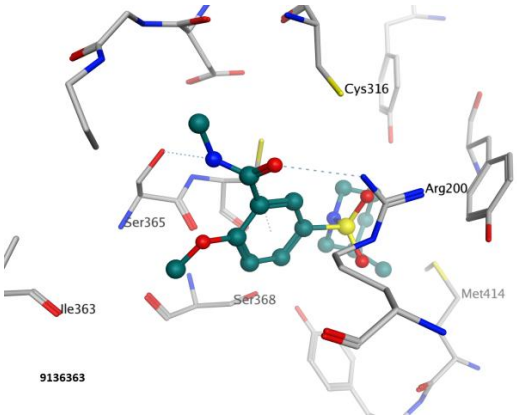
T1

: Glide SP score = -6.81



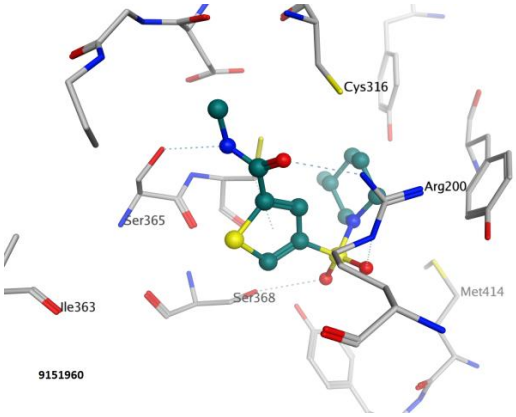
T2

: Glide SP score = -7.88



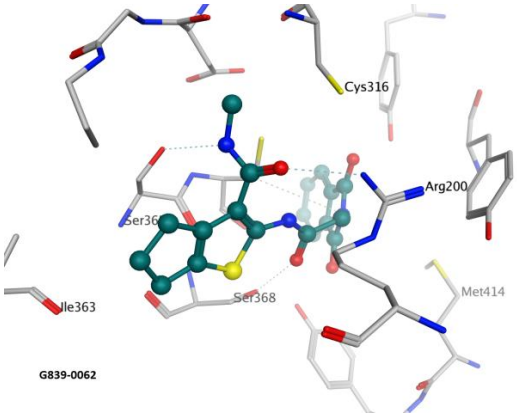
T3

: Glide SP score = -6.81



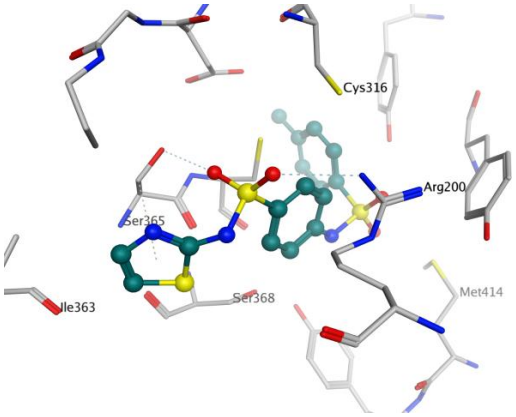
T4

: Glide SP score = -7.19



T5

: Glide SP score = -8.16



T6

: Glide SP score = -7.21

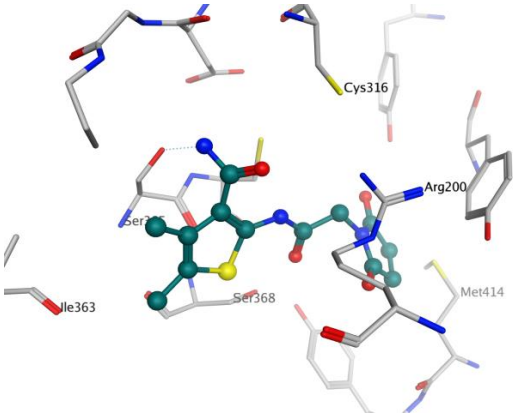
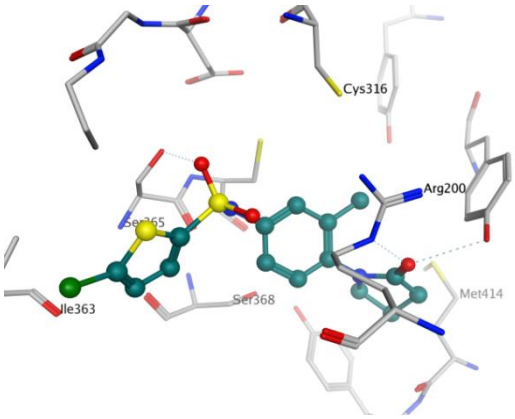


Figure A9. Docked poses of 11 tested compounds (Dataset-II) in structure of HCV polymerase genotype 1b Con1 (PDB ID: 3FQL) and their Glide SP scores.

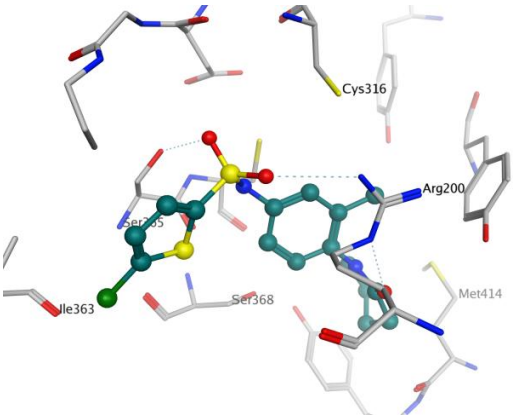
T7

: Glide SP score = -8.56



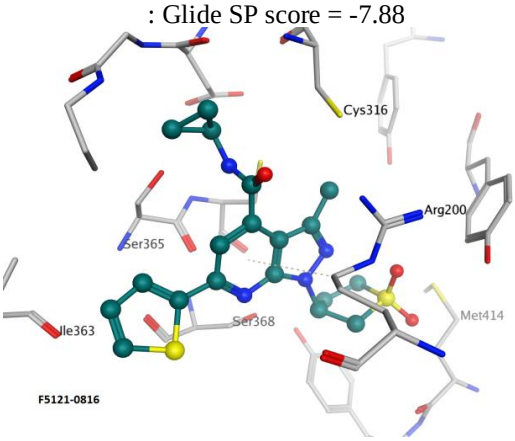
T8

: Glide SP score = -7.98



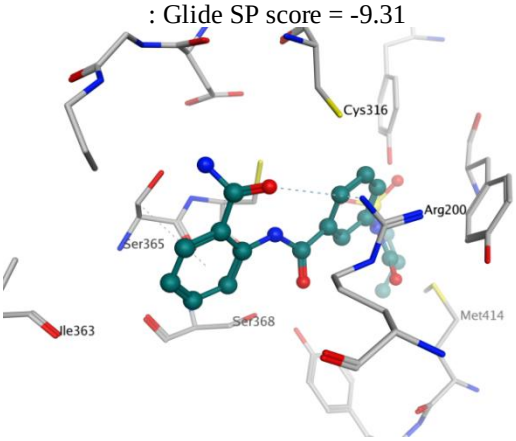
T9

: Glide SP score = -7.88



T10

: Glide SP score = -9.31



T11

: Glide SP score = -7.43

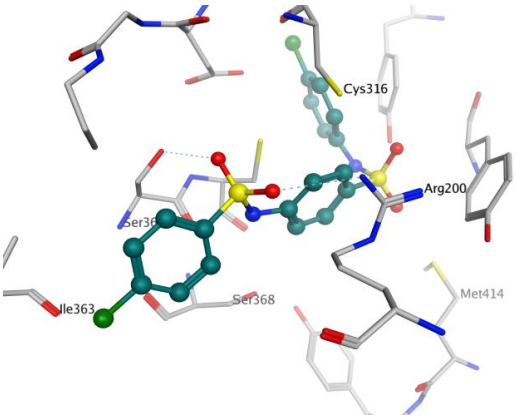


Figure A9 (Continued). Docked poses of 11 tested compounds (Dataset-II) in structure of HCV polymerase genotype 1b Con1 (PDB ID: 3FQL) and their Glide SP scores.

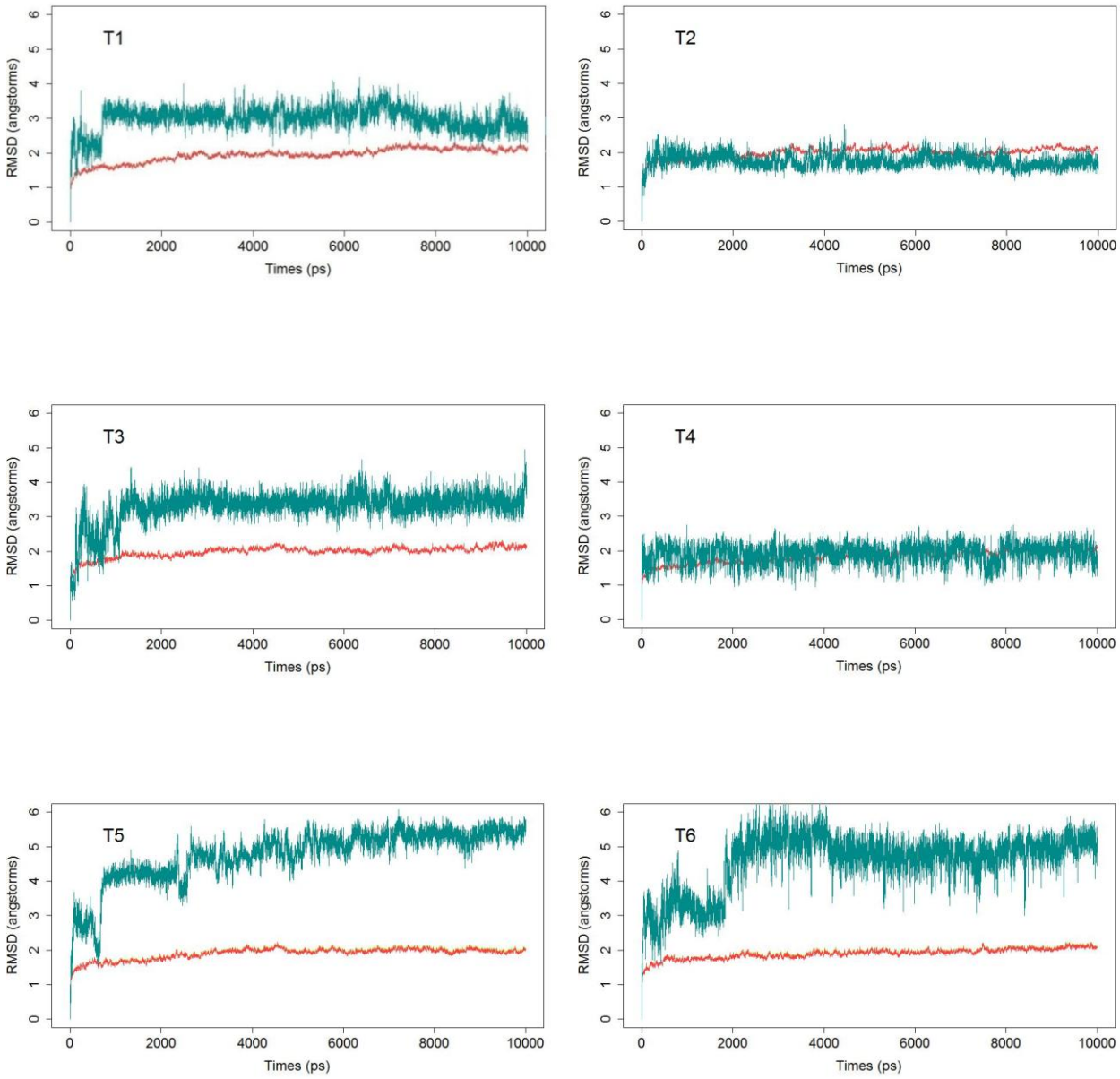


Figure A10. RMSD plots of the complexes (red) and 11 tested compounds (teal) during MD simulation.

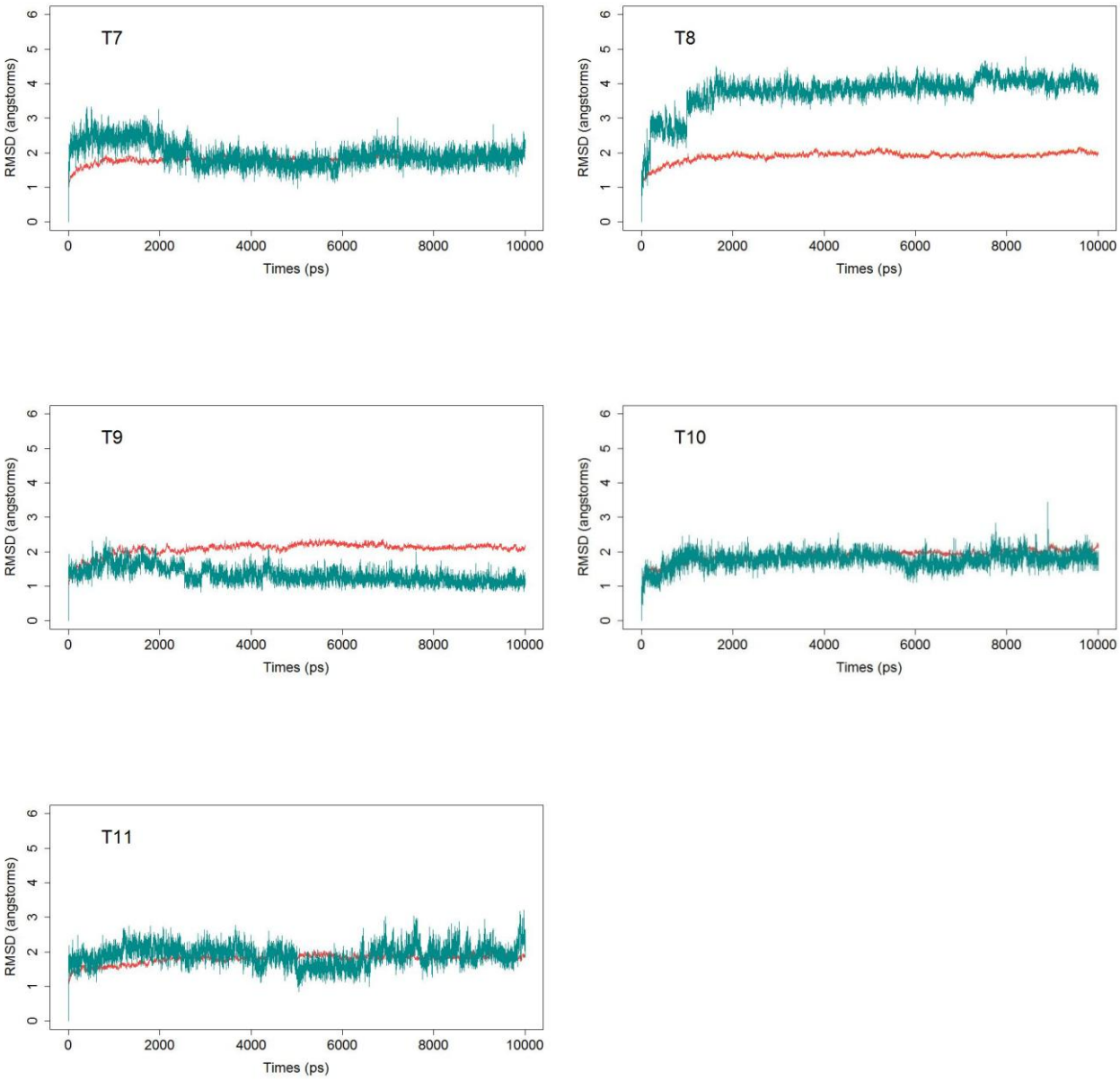


Figure A10 (Continued). RMSD plots of the complexes (red) and 11 tested compounds (teal) during MD simulation.

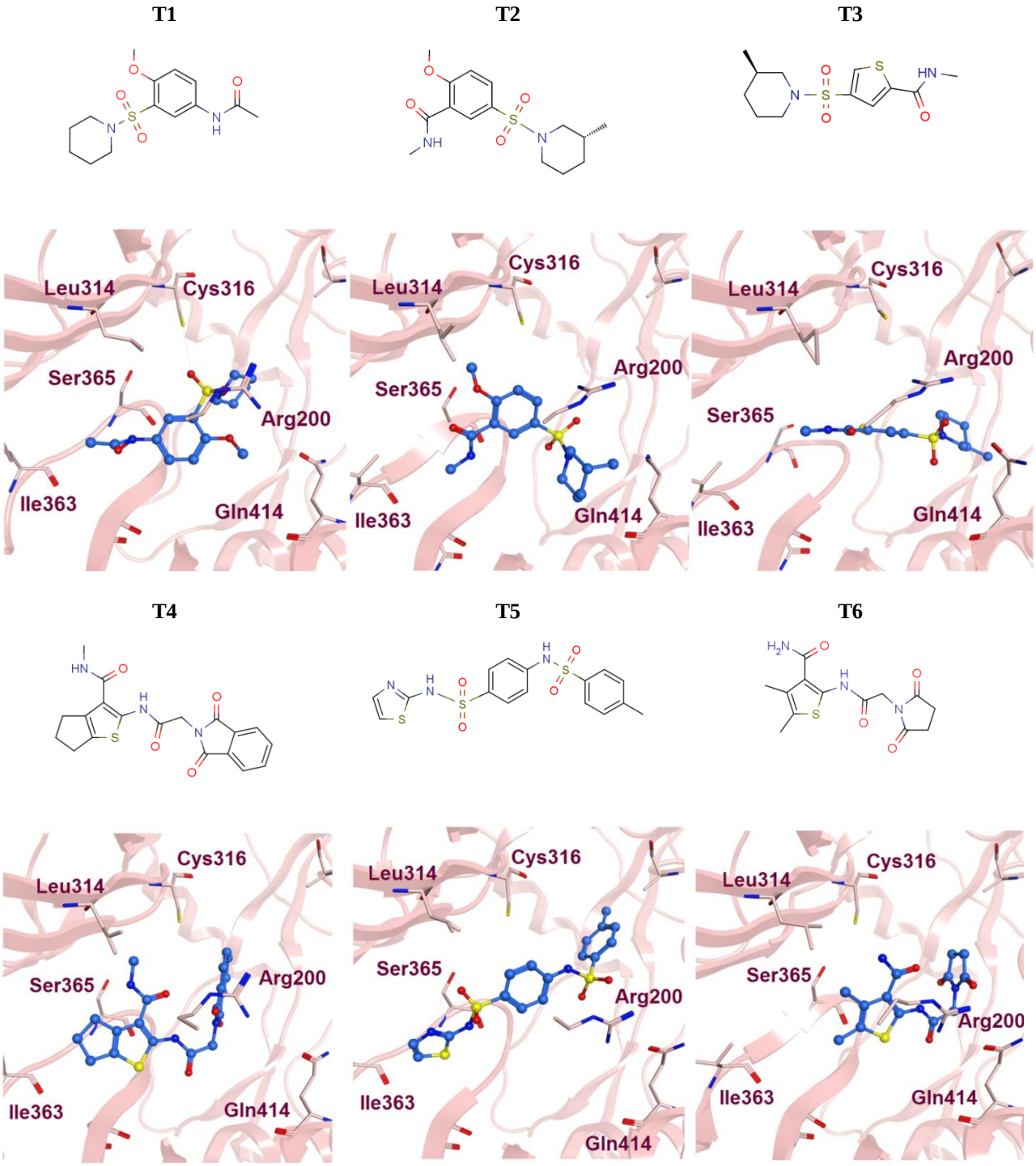


Figure A11. Bound conformations of 11 tested compounds generated from average 100 snapshots of MD simulations.

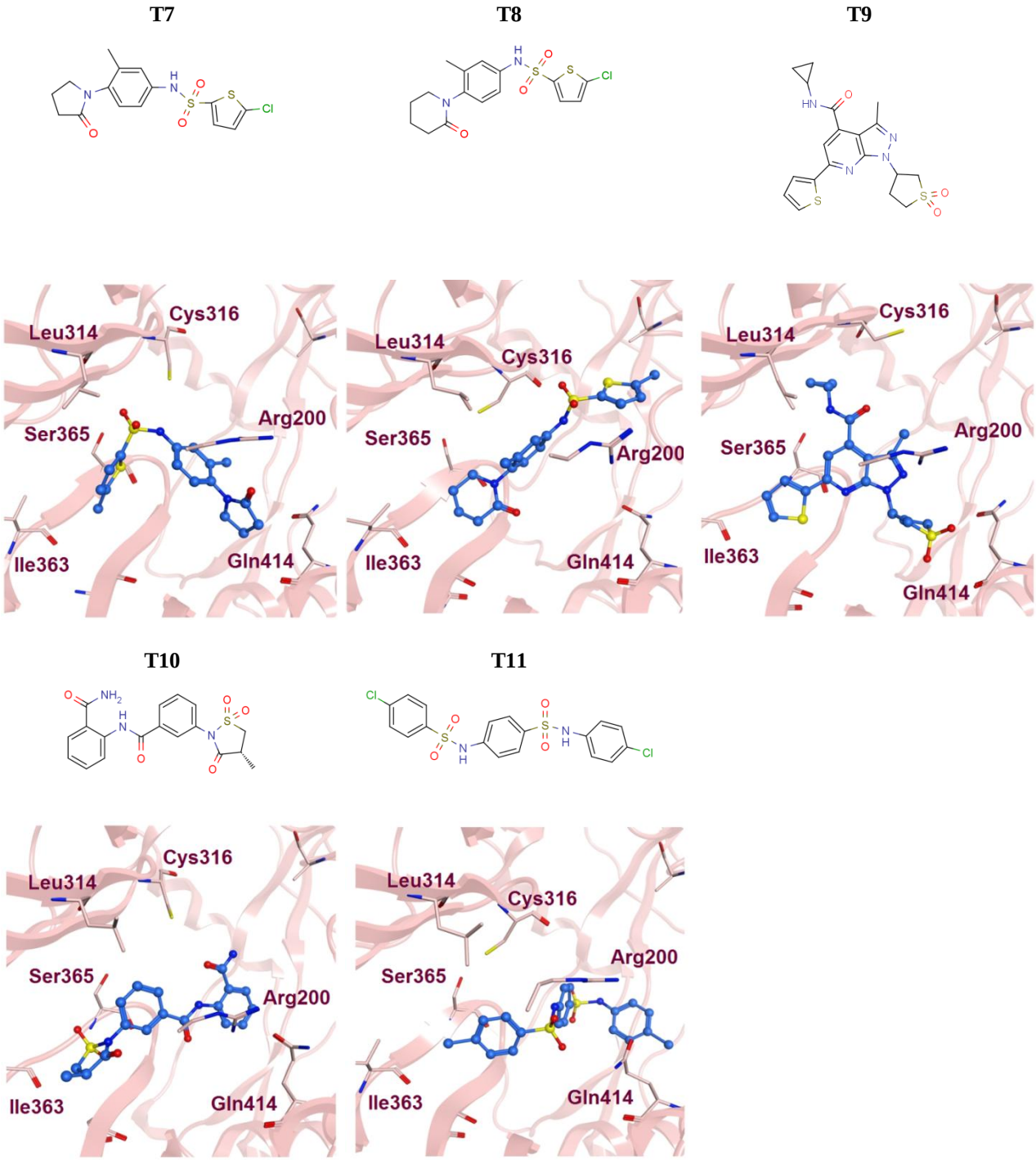


Figure A11 (Continued). Bound conformations of 11 tested compounds generated from average 100 snapshots of MD simulations.

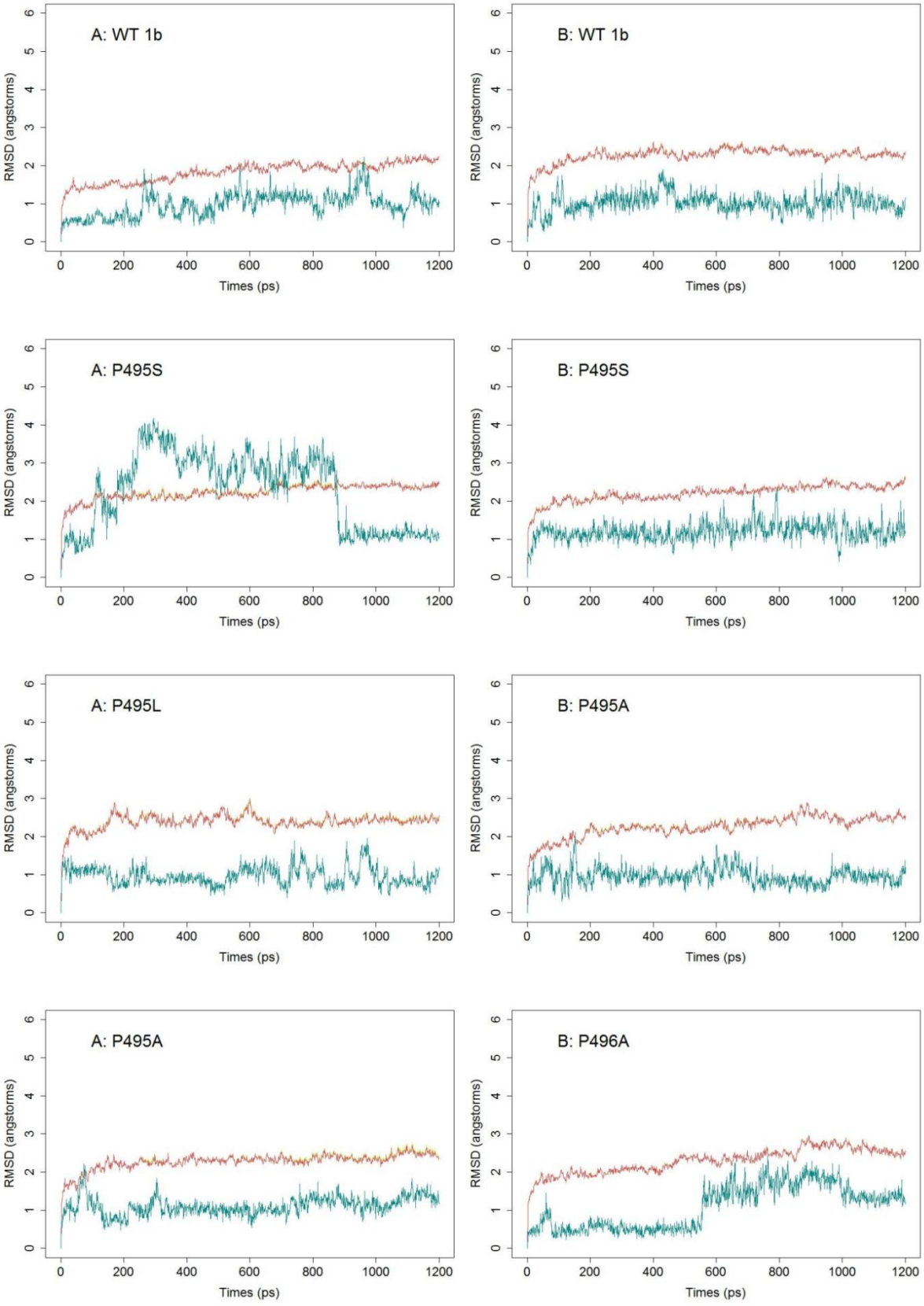


Figure A12. Root Mean Square Deviations (RMSD) of the HCV polymerase (red) complexed with inhibitor (teal) during 12 ns simulation.

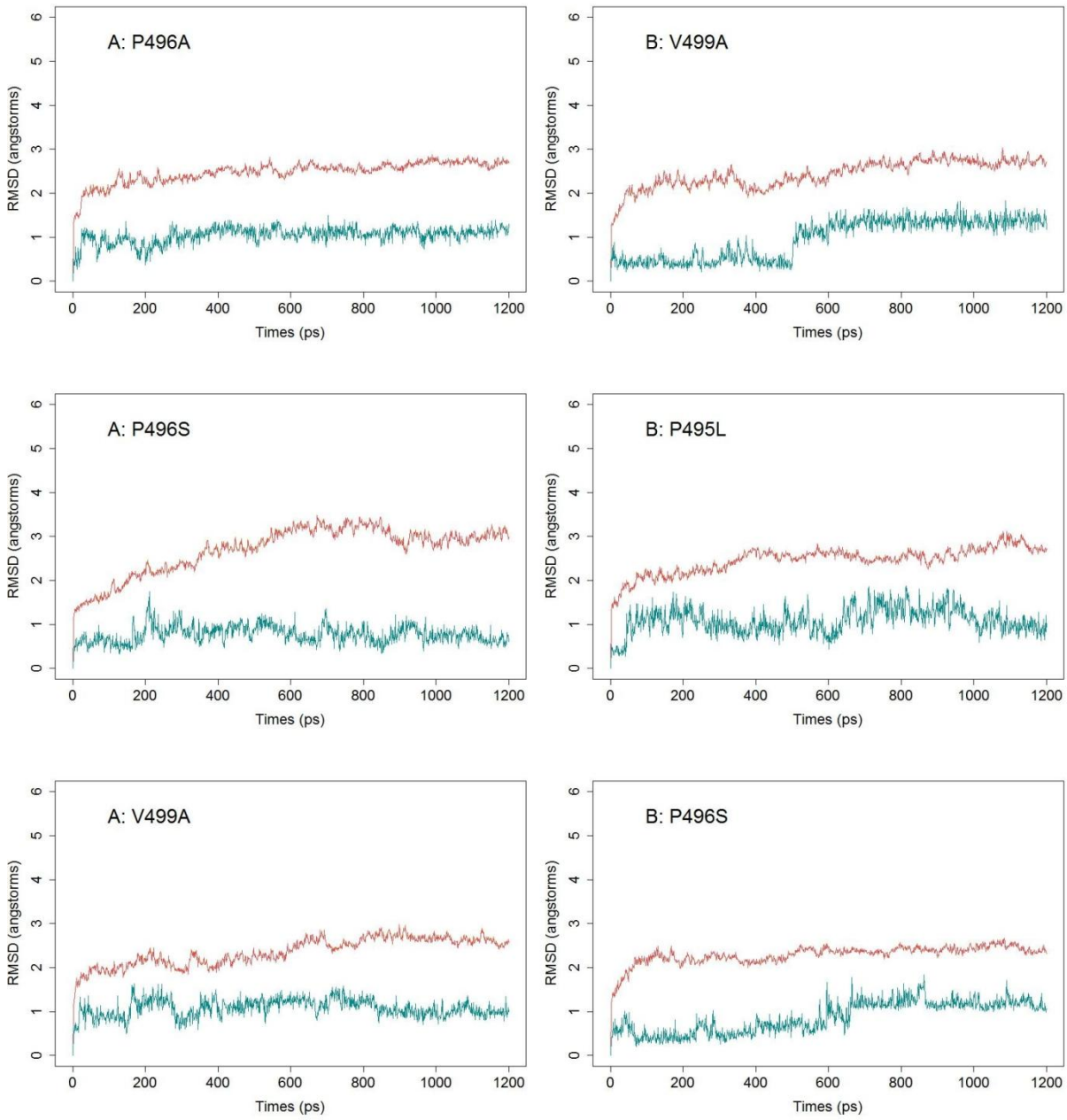


Figure A12 (Continued). Root Mean Square Deviations (RMSD) of the HCV polymerase (red) complexed with inhibitor (teal) during 12 ns simulation.

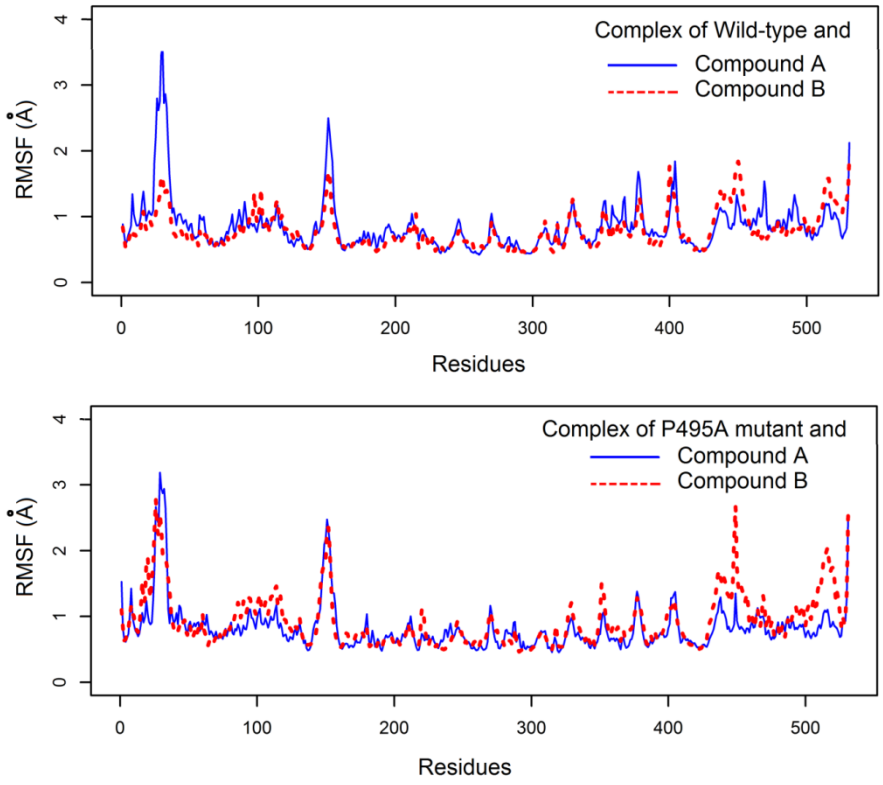


Figure A13. Comparison of the flexibility of HCV-NS5B-inhibitor complexes. Compound **A** and compound **B** bound to wild-type HCV NS5B polymerase are shown in the upper part, whereas binding to the P495A mutant is shown in the lower part.

Table A11. Inhibitory profiles and estimated binding free energy calculations (GBTOT) of complex between inhibitors and HCV polymerases (± standard deviation) calculated from MM-GB/SA. All energies are in kcal/mol. Predicted binding free energies were averaged over 100 snapshots during last 2 ns of MD simulation.

Structure	**Compound A**		**Compound** B	
	IC_{50} (µM)	GBTOT (kcal/mol)	IC_{50} (µM)	GBTOT (kcal/mol)
WT- 1b	0.15 (± 0.042)	-56.46 (±2.6)	0.27 (±0.05)	-52.22 (±2.5)
P495S	>25	-47.76 (±3.4)	>32	-48.32 (±3.0)
P495L	>25	-47.66 (±2.9)	>25	-48.53 (±2.7)
P495A	>25	-54.51 (±2.9)	10.6 (±3.2)	-45.03 (±2.3)
P496A	0.52 (±0.16)	-57.52 (±3.6)	2.6 (±0.4)	-49.20 (±2.6)
P496S	0.53 (±0.22)	-53.58 (±2.5)	3.8 (±2.1)	-49.30 (±2.6)
V499A	0.41 (±0.12)	-48.61 (±2.5)	0.63 (±0.15)	-55.07 (±2.6)

Table A12. Per-residue energy (kcal/mol) of the key amino acids in the binding pocket of the wild-type (WT) and mutant enzymes with compound A and compound B. The grey and red background shading represents the energy difference of the residue contributions in the mutants that are between 0.5-1 kcal/mol and greater than 1 kcal/mol relative to the wild-type, respectively. The boxes show the mutation points.

Residue	van der Waals (kcal/mol)													
	Compound **A**							Compound **B**						
	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A
32	-1.8	-0.7	-0.1	-3.9	-0.1	-0.2	0.8	-1.2	0.0	-1.4	-0.6	-0.2	-0.4	-0.1
396	-1.4	-1.6	-1.5	-1.6	-1.1	-1.4	-1.5	-1.4	-1.3	-1.4	-1.6	-1.0	-1.1	-1.2
428	-2.4	-1.8	-2.3	-2.5	-1.9	-2.2	-2.4	-2.2	-1.4	-2.1	-2.2	-2.0	-1.7	-2.0
492	-1.8	-0.3	-1.7	-1.4	-2.0	-0.9	-1.4	-1.1	-1.6	-1.4	-1.2	-0.9	-1.4	-1.2
493	-1.0	-1.1	-0.6	-1.2	-1.7	-1.0	-1.2	-0.3	-0.6	-0.3	-0.3	-0.7	-0.5	-0.7
494	-2.6	-3.3	-2.6	-2.2	-3.8	-3.1	-3.3	-2.5	-2.7	-2.2	-2.7	-2.3	-2.1	-2.3
495	**-3.3**	-2.2	-2.2	-1.4	-3.4	-3.6	-3.8	**-3.0**	-1.8	-2.9	-1.7	-2.9	-3.1	-2.9
496	**-1.0**	-0.7	-0.3	-0.3	-0.2	-0.4	-0.8	**-1.1**	-1.0	-0.9	-1.3	-0.4	-0.4	-1.4
497	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	-0.1	-0.1	-0.1	-0.1
498	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.1	1.0	0.3	0.5	0.4	0.5
499	**-0.3**	-0.3	-0.1	-0.1	-0.2	-0.2	-0.1	**-0.9**	-0.9	-0.9	-0.9	-1.2	-1.1	-0.7
500	-1.0	-1.4	-0.9	-0.9	-1.2	-1.0	-1.1	-1.2	-1.2	-1.2	-1.2	-1.5	-1.3	-1.5
501	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
502	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	-0.1
503	-0.1	0.7	-0.4	-0.8	0.2	0.1	-0.2	-0.9	-0.4	-0.8	-0.4	-0.7	-0.7	-0.9

Table A12 (Continued). Per-residue energy (kcal/mol) of the key amino acids in the binding pocket of the wild-type (WT) and mutant enzymes with compound A and compound B. The grey and red background shading represents the energy difference of the residue contributions in the mutants that are between 0.5-1 kcal/mol and greater than 1 kcal/mol relative to the wild-type, respectively. The boxes show the mutation points.

Residue	Electrostatic (kcal/mol)													
	Compound **A**							Compound **B**						
	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A
32	-	-47.0	-11.8	-20.1	-16.3	-17.9	-53.8	-	-5.7	-18.3	-10.1	-8.5	-8.6	-8.9
396	0.5	-0.5	0.7	0.9	0.4	0.7	0.5	0.0	0.0	0.0	-0.2	-0.2	-0.1	-0.1
428	0.2	0.5	-0.6	1.2	1.1	0.0	-0.2	0.1	0.3	0.1	0.0	0.3	0.3	0.0
492	1.7	0.2	1.4	0.7	0.4	0.7	1.2	-0.4	0.1	-0.5	-0.3	-0.7	-0.3	-0.5
493	0.6	-0.3	0.0	0.8	0.7	0.5	0.6	-0.4	-0.8	-0.3	-0.4	0.0	-0.3	-1.2
494	-2.4	-2.0	-1.5	-0.3	-2.4	-5.8	-1.8	0.6	0.6	0.4	0.7	0.0	0.5	0.3
495	**-2.7**	-5.7	-1.1	-1.5	-1.2	-1.3	-2.6	**-3.0**	-4.3	-2.8	-2.9	-2.5	-3.0	-2.6
496	**0.7**	1.5	0.8	0.5	1.2	0.5	0.8	**-1.1**	-0.8	-0.7	-1.4	-0.4	-6.9	-1.1
497	0.4	0.5	0.3	0.2	0.5	0.3	0.3	0.0	0.0	0.1	-0.1	-0.1	-0.2	-0.1
498	-	-10.5	-9.9	-10.0	-12.0	-9.3	-9.8	-	-55.6	-55.7	-52.2	-55.1	-55.3	-54.5
499	**-0.5**	-0.7	-0.2	-0.1	-0.4	-0.1	-0.4	**-1.3**	-1.1	-1.3	-1.3	-1.1	-1.6	-1.3
500	0.4	0.1	0.0	0.2	0.0	0.6	-0.1	-0.8	-0.8	-0.6	-0.9	-0.6	-0.8	-0.6
501	-9.3	-9.9	-8.0	-7.9	-9.3	-8.6	-8.4	-	-11.7	-12.0	-12.2	-16.7	-13.6	-12.8
502	-0.1	-0.2	0.1	-0.3	-0.8	-0.2	-0.7	-0.9	-0.8	-0.4	-0.6	-1.2	-0.7	-1.1
503	-	-52.4	-26.1	-24.9	-47.5	-43.9	-37.0	-	-13.7	-22.4	-12.7	-28.2	-27.8	-26.2

Table A12 (Continued). Per-residue energy (kcal/mol) of the key amino acids in the binding pocket of the wild-type (WT) and mutant enzymes with compound A and compound B. The grey and red background shading represents the energy difference of the residue contributions in the mutants that are between 0.5-1 kcal/mol and greater than 1 kcal/mol relative to the wild-type, respectively. The boxes show the mutation points.

Residue	Polar Solvation (kcal/mol)													
	Compound **A**							Compound **B**						
	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A
32	26.	45.1	11.9	22.2	16.4	17.9	49.3	12.	5.7	19.7	10.4	8.7	9.0	9.2
39	-0.3	0.9	-0.4	-0.7	-0.3	-0.4	-0.4	0.1	0.0	0.0	0.2	0.1	0.1	0.0
42	2.4	1.5	3.4	1.7	1.3	2.5	2.7	2.1	1.5	2.2	1.5	2.0	1.8	2.1
49	-0.2	-0.1	-0.6	0.3	-0.2	-0.4	0.2	0.7	0.6	1.0	0.6	1.2	0.8	0.7
49	0.1	1.0	0.6	0.9	0.5	0.3	0.1	0.6	1.1	0.6	0.6	0.7	0.5	1.6
49	3.5	3.3	2.6	1.9	3.2	4.6	2.7	-0.2	0.1	0.2	-0.1	0.2	0.2	0.1
49	**2.8**	6.1	1.3	1.7	1.9	1.9	2.6	**3.0**	5.2	2.7	3.1	2.7	3.3	2.7
49	**-0.7**	-1.4	-0.8	-0.4	-1.0	-0.5	-0.7	**1.1**	0.7	0.6	1.2	0.5	5.1	1.1
49	-0.4	-0.4	-0.3	-0.2	-0.4	-0.3	-0.3	0.0	0.0	-0.1	0.1	0.1	0.2	0.1
49	11.	10.5	9.9	10.0	12.0	9.3	9.8	48.	51.0	50.9	48.4	50.5	50.6	49.8
49	**0.5**	0.7	0.2	0.1	0.4	0.1	0.4	**1.3**	1.2	1.3	1.3	1.2	1.5	1.3
50	-0.4	0.2	0.2	0.2	0.0	-0.5	0.1	0.8	1.0	0.7	1.1	0.7	1.0	0.7
50	9.2	9.9	8.0	7.9	9.2	8.6	8.4	12.	11.6	12.0	12.1	16.6	13.5	12.7
50	0.1	0.3	0.0	0.3	0.8	0.2	0.7	0.9	0.9	0.5	0.7	1.3	0.8	1.2
50	39.	48.7	26.3	25.9	45.1	42.0	36.6	27.	14.2	23.4	12.9	29.2	28.0	27.6

Table A12 (Continued). Per-residue energy (kcal/mol) of the key amino acids in the binding pocket of the wild-type (WT) and mutant enzymes with compound A and compound B. The grey and red background shading represents the energy difference of the residue contributions in the mutants that are between 0.5-1 kcal/mol and greater than 1 kcal/mol relative to the wild-type, respectively. The boxes show the mutation points.

Residue	Non-Polar Solvation (kcal/mol)													
	Compound **A**							Compound **B**						
	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A	WT	P495 S	P495 L	P495 A	P496 A	P496 S	V499 A
32	-0.4	-0.4	0.0	-0.6	0.0	0.0	-0.4	-0.4	0.0	-0.3	-0.1	-0.1	-0.1	0.0
39	-0.1	-0.1	-0.2	-0.2	-0.2	-0.2	-0.2	-0.1	-0.2	-0.2	-0.2	-0.1	-0.1	-0.1
42	-0.2	-0.2	-0.3	-0.3	-0.3	-0.3	-0.2	-0.2	-0.1	-0.2	-0.2	-0.2	-0.2	-0.2
49	-0.1	0.0	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1
49	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	0.0	-0.1	0.0	0.0	-0.1	0.0	0.0
49	-0.2	-0.3	-0.4	-0.3	-0.3	-0.4	-0.2	-0.2	-0.3	-0.3	-0.2	-0.2	-0.2	-0.1
49	-0.3	-0.3	-0.4	-0.2	-0.4	-0.4	-0.3	-0.3	-0.3	-0.4	-0.3	-0.4	-0.3	-0.3
49	-0.1	-0.1	0.0	0.0	0.0	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	0.0	-0.1	-0.2
49	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.2	-0.3	-0.2	-0.2	-0.3	-0.2	-0.2
49	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.2	-0.1	-0.2	-0.2	-0.2	-0.1
50	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50	-0.3	-0.3	-0.2	-0.3	-0.3	-0.3	-0.2	-0.3	-0.1	-0.2	-0.1	-0.3	-0.3	-0.3

Curriculum vitae

Name	Tanaporn Uengwetwanit
Date of Birth	September 7, 1981
Nationality	Thai
E-mail	tanaporn.ueng@gmail.com

Education

August 2010 - 2014	Ph.D. student under supervision of Prof.Dr.Wolfgang Sippl, Institute of Pharmaceutical Chemistry, Martin Luther University Halle-Wittenberg, Germany
May 2005 – October 2007	Master of Science (Bioinformatics), King Mongkut's University of Technology Thonburi, Thailand
May2000 – January, 2005	Bachelor of Pharmacy Silpakorn University, Thailand (Second Class Honours)

Internship

April – August, 2007	The Centre of Applied Genomics (TCAG), Toronto, Canada
October – December, 2004	The Center of Vaccine and Development, Mahidol University, Thailand.

Working Experience

2008 – 2009	Research assistant, the National Center for Genetic Engineering and Biotechnology, Thailand

Scholarship and awards

August 2010 - 2014	Thai Government Science and Technology Scholarship
May 2005 – October 2007	Scholarship in the program of Master's Degree in Bioinformatics, King Mongkut's University of Technology Thonburi, Thailand

Workshops and Trainings

- New approaches in drug design & discovery 'the aspect of time in drug design', Marburg, Germany (March 24-27, 2014)
- 28th Molecular modeling workshop, Erlangen, Germany (March 17-19, 2014)
- Innovative approaches to computational drug discovery, CECAM-HQ-EPFL, Lausanne, Switzerland (October 1-4,2013)

- Tutorial for the AMBER set of modeling tools, CECAM-HQ-EPFL, Lausanne, Switzerland (October 8-12, 2012)
- 6th Summer school medicinal chemistry, University of Regensburg (September 26 - 28, 2012)
- Vienna summer school: drug design, a European pharmacoinformatics initiative (September 11-16, 2011)
- 25th Molecular Modeling Workshop, Erlangen, Germany (April 4-6, 2011)
- Computational biology : from (meta)genomes to phenotype and environment, EMBL (August 16-22, 2009)
- Comparative microbial genomics workshop, The Center for Biological Analysis, Technical University of Denmark, (June 2-6, 2008)
- System biology: post-genomic integrative analyses, Technical University of Denmark, (October 16-19, 2006)

Publications

- Uengwetwanit T., Robaa D. and Sippl W., *Analysis of the resistance of hepatitis C virus NS5B polymerase inhibitors via docking and molecular dynamics simulation*, 2014 (submitted)
- Tewes B., Schepmann D., Robaa D., Uengwetwanit T., Fröhlich R., Sippl W, Wünsch B, *Enantiomerically Pure Ifenprodil Analogues*, 2014 (submitted)
- Nguyen, N. H., Maruset, L., Uengwetwanit, T., Mhuantong, W., Harnpicharnchai, P., Champreda, V., Tanapongpipat, S., Jirajaroenrat, K., Rakshit, S. K., Eurwilaichitr, L., Pongpattanakitshote, S., *Identification and characterization of a cellulase-encoding gene from the buffalo rumen metagenomic library*. Biosci Biotechnol Biochem, 2012, 76 (6), 1075-84.
- Nimchua, T., Thongaram, T., Uengwetwanit, T., Pongpattanakitshote, S., Eurwilaichitr, L., *Metagenomic analysis of novel lignocellulose-degrading enzymes from higher termite guts inhabiting microbes*. J Microbiol Biotechnol 2012, 22 (4), 462-9.
- Kanokratana, P., Uengwetwanit, T., Rattanachomsri, U., Bunterngsook, B., Nimchua, T., Tangphatsornruang, S., Plengvidhya, V., Champreda, V., Eurwilaichitr, L., *Insights into the phylogeny and metabolic potential of a primary tropical peat swamp forest microbial community by metagenomic analysis*. Microb Ecol, 2011, 61 (3), 518-28.

- Chantasingh, D., Kitikhun, S., Eurwilaichitr, L., Uengwetwanit, T., and Pootanakit, K., *Functional expression in Beauveria bassiana of a chitinase gene from Ophiocordyceps unilateralis , an ant-pathogenic fungus*, Biocontrol Science and Technology, 2011, 21 (6).
- Bunterngsook, B.; Kanokratana, P.; Thongaram, T.; Tanapongpipat, S.; Uengwetwanit, T.; Rachdawong, S.; Vichitsoonthonkul, T.; Eurwilaichitr, L., *Identification and characterization of lipolytic enzymes from a peat-swamp forest soil metagenome*. Biosci Biotechnol Biochem, 2010, 74 (9), 1848-54.

Posters and oral presentations

- Uengwetwanit T. and Sippl W, *Filtering strategies for improved virtual screening of HCV NS5B polymerase inhibitors*, 28th Molecular Modeling workshop, 2014, Erlangen, Germany
- Uengwetwanit T. and Sippl W., *Evaluating molecular dynamics simulation for predicting binding affinity changes upon single point mutation: a case study on non-nucleoside HCV NS5B inhibitors*, Drug design 2013: Fragment and ligand based drug design, 2013, Oxford, UK
- Uengwetwanit T. and Sippl W., *Improving virtual screening hit identification of HCV NS5B inhibitors by random forest classification*, Drug discovery and selection RICT 2013 : When chemical biology meets drug design, 2013, Nice, France
- Uengwetwanit T. and Sippl W., *Optimized Docking/Scoring Protocols for HCV NS5B Polymerase Inhibitors through Receptor Flexibility and Rescoring*, 6th Summer school Medicial Chemistry, 2012, Regensburg, Germany
- Uengwatwanit T., Jongjaroenprasert, W., Meechai. A., Ongphiphadhanakul, B., Lertbantanawong, J. and Chan J.H. *An Analytical Process for Screening Susceptibility Genes of Type 2 Diabetes Mellitus Using Pooled DNA Microarray Data*, IAENG International Conference on Bioinformatics, Hong Kong, 2008, 195-199
- Uengwetwanit T., Jongjaroenprasert, W., Meechai. A. and Chan, J.H.,*The effect of preprocessing on the results of pooled DNA SNP microarray analysis*, The 18th Annual meeting of the Thai Social for Biotechnology, 2006, Bangkok

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertationsschrift selbständig und ohne fremde Hilfe angefertigt, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus ihnen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Die Arbeit wurde ausschließlich der Mathematisch-Naturwissenschaftlichen Fakultät der Martin Luther Universität Halle-Wittenberg vorgelegt und an keiner anderen Universität oder Hochschule weder im In- und Ausland zur Erlangung des Doktorgrades eingereicht

Halle (saale), den