

**Computational Approaches for Investigating Protein-Ligand
Interactions - Towards an in-depth Understanding of the Dengue
Virus Methyltransferase**

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät der
Universität Basel

von

Tobias Benjamin Schmidt

aus

Basel (BS), Schweiz

Basel, 2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Torsten Schwede

Prof. Dr. Markus Meuwly

Basel, den 13. November 2012

Prof. Dr. Jörg Schibler
Dekan






Attribution-Noncommercial-No Derivative Works 2.5 Switzerland

You are free:

to share — to copy, distribute and transmit the work

Under the following conditions:

-  **ATTRIBUTION.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
-  **NONCOMMERCIAL.** You may not use this work for commercial purposes.
-  **NO DERIVATIVE WORKS.** You may not alter, transform, or build upon this work.

With the understanding that:

WAIVER. Any of the above conditions can be waived if you get permission from the copyright holder.

PUBLIC DOMAIN. Where the work or any of its elements is in the public domain under applicable law, that status is in no way affected by the license.

OTHER RIGHTS. In no way are any of the following rights affected by the license:

- Your fair dealing or fair use rights, or other applicable copyright exceptions and limitations;
- The author's moral rights;
- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

This is a human-readable summary of the Legal Code (the full license) available at: <http://creativecommons.org/licenses/by-nc-nd/2.5/ch/legalcode.de>

SOURCE: <http://creativecommons.org/licenses/by-nc-nd/2.5/ch/deed.en>

Für Nicole, Anina und Joel

Abstract

Interactions between proteins and their ligands play crucial roles in many biological processes, such as metabolism, signaling, transport, regulation or molecular recognition. Understanding the molecular basis of protein-ligand interactions is thus of great interest, not only for modeling complex biological systems but also for applications in drug discovery. However, structural details for most of these interactions have not been characterized experimentally. Therefore, computational methods have become increasingly important for investigating biological systems at an atomistic level.

This work aims at a better understanding of the molecular basis of disease related viral methyltransferases, their interactions with small molecules and the catalytic mechanism, which may on the long perspective help to develop a treatment against neglected tropical diseases. Furthermore, we aim to advance the current methods for the computational prediction of a protein's molecular function and its biological role in the cell. In addition, we aim to complement currently available computational strategies for estimating protein ligand interaction energies.

Dengue fever is a rapidly emerging, still neglected tropical disease which causes significant mortality and morbidity in humans. For the discovery of novel classes of compounds inhibiting dengue virus methyltransferase, a combination of structure-based virtual screening and enzymatic inhibition assays is employed. From the shortlist of 263 candidates selected by virtual screening, ten compounds are found to specifically inhibit the target enzyme with IC_{50} values in the low μM range. Promising compounds are selected for further experimental characterization and the inhibitory activity of the two most active compounds is confirmed.

For obtaining a better understanding of the molecular basis of the target enzyme's function, molecular dynamics simulations and mixed quantum mechanics/molecular mechanics calculations are employed to investigate the mechanisms of the enzymatically catalyzed reaction at an atomistic level. Based on a structural model of the target protein in complex with its RNA substrate, the impact of mutations on ligand binding, geometric arrangements and reaction energy barriers are evaluated computationally. In addition, for a detailed characterization of the underlying chemical reactions, ab initio electronic structure calculations are performed on model systems approximating the biological structure.

The reliable prediction of ligand binding sites is crucial for characterizing proteins with unknown function. Therefore, the use of computational predictions of protein function and ligand binding sites for proteins without experimental structures are assessed in a blind and objective way. Limitations in the current prediction methods are analyzed and suggestions for

a more reliable evaluation are given. Following those suggestions, an extended and fully automated assessment is implemented in the Continuous Automated Model EvaluatiOn (CAMEO) framework.

Computational identification of protein-ligand interactions can greatly facilitate the drug discovery process. Thus, we establish a straightforward, rapid scoring function that aims to identify the best poses out of an ensemble of pre-docked poses, by quantifying the degree of burial and the electrostatic interactions of the ligand in a binding site. The scoring function is evaluated on a set of high quality protein-ligand complex structures, where the results show promisingly high retrieval rates for selecting the best poses from a pool of decoy poses.

Finally, a novel human-computer interface device is described which facilitates the interaction with the computational representation of complex biological systems by employing natural and intuitive movements.

Contents

1	Introduction	1
1.1	Protein-Ligand Interactions	1
1.2	Estimation of Protein-Ligand Interactions	3
1.2.1	Prediction of Ligand Binding Sites	3
1.2.2	Protein-Ligand Docking and Virtual Screening	3
1.2.3	Estimation of Protein-Ligand Binding Affinities	4
1.2.4	Estimation of Reaction Energy Barriers	4
1.3	Flavivirus	6
1.3.1	Dengue Fever	6
1.3.2	Dengue Virus	6
1.3.3	NS5 Methyltransferase	8
1.4	Objectives	11
2	Identification and Validation of Novel Dengue Methyltransferase Inhibitors	13
2.1	Screening For Novel Inhibitors	15
2.2	Prediction of Non-Specific Inhibitors	29
2.2.1	Introduction	29
2.2.2	Materials and Method	29
2.2.3	Results and Discussion	30
2.2.4	Conclusion	32
2.3	Experimental Characterization of Novel Inhibitors	33
2.3.1	Introduction	33
2.3.2	Materials and Method	33
2.3.3	Results and Discussion	39
2.3.4	Conclusion	41
3	Computational Analysis of the Methyltransferase Reaction	45
3.1	Introduction	45
3.2	Modeling of the Protein-RNA Complex	47
3.2.1	Method	47
3.2.2	Validation of the Structural Model	48
3.2.3	Ligand-Induced Structural Rearrangements	49

3.3	Methylation of Guanosine N7 and Adenosine 2'O in Model Systems	52
3.3.1	Method	52
3.3.2	Geometry	52
3.3.3	Energy Profiles	55
3.3.4	Energy Landscapes	56
3.3.5	Point Charges	57
3.3.6	Two Step Reaction	58
3.4	Impact of Single Point Mutations	61
3.4.1	Materials and Methods	61
3.4.2	Results of Computational Alanine Scanning	66
3.4.3	Summary of Computational Alanine Scanning	72
3.4.4	Experimental and Computational Analysis of Selected Mutants	72
3.4.5	Conclusion	76
3.5	RNA Sequence Specificity	78
3.5.1	Method	79
3.5.2	Results and Discussion	80
3.5.3	Conclusion	83
3.6	Conclusion	85
4	Ligand Binding Site Prediction	87
4.1	Introduction	87
4.1.1	Critical Assessment of Protein Structure Prediction	87
4.2	Assessment of Ligand Binding Site Prediction in CASP9	88
4.3	CAMEO Ligand Binding	100
4.3.1	Introduction	100
4.3.2	CAMEO Workflow	101
4.3.3	Prediction Targets	102
4.3.4	Ligand Annotation	102
4.3.5	Ligand Classification Scheme	103
4.3.6	Ligand Categorization	104
4.3.7	Assessment	104
4.3.8	Scoring	106
4.3.9	Baseline Servers	107
4.3.10	Prediction Format	108
4.3.11	Results and Discussion	110
4.3.12	Conclusion	114
4.4	Geometry Based Ligand Binding Site Prediction	116
4.4.1	Introduction	116
4.4.2	Method	116
4.4.3	Results and Discussion	117
4.4.4	Conclusion	120

5	BEscore: a Novel Method for Rapid Scoring of Protein-Ligand Complexes	121
5.1	Introduction	121
5.2	Method	122
5.2.1	Shape Term (Degree of Burial)	122
5.2.2	Electrostatic Term	124
5.2.3	BEscore	125
5.3	Sets of Receptor-Ligand Complexes	125
5.3.1	Thrombin Set	125
5.3.2	Astex Diverse Set	126
5.3.3	S3DB	126
5.4	Validation	127
5.4.1	Shape Term	127
5.4.2	Electrostatic Term	131
5.4.3	Summary of Individual Terms	132
5.4.4	Comparison to Van der Waals Interaction Energies	133
5.4.5	BEscore	134
5.5	Analysis of Surface Point Distribution	135
5.6	Results	137
5.6.1	Astex Diverse Set	137
5.6.2	Thrombin Set	138
5.6.3	S3DB	139
5.7	Comparison to X-Score and Glide SP	140
5.8	Combining with X-Score and Glide SP	140
5.9	Discussion	140
6	Design and Evaluation of a Novel, Intuitive Human-Computer Interface Device	143
6.1	Introduction	143
6.2	Interface Device Design	145
6.2.1	Hardware Architecture	146
6.2.2	Software Architecture	148
6.3	User Experience	149
6.4	Conclusion	150
7	Summary and Outlook	153
	Acknowledgments	155
	References	157
A	Appendix	173
A.1	Dengue	173
A.1.1	Identification and Validation of Novel Dengue Methyltransferase Inhibitors	173

A.1.2	Computational Analysis of the Methyltransferase Reaction	174
A.2	CAMEO Ligand Binding	175
A.2.1	CAMEO Ligand Binding Format Examples	175
A.3	BEscore	176
A.4	Human-Computer Interface Schematics	182

Chapter 1

Introduction

1.1 Protein-Ligand Interactions

Proteins are biological macromolecules that play a central role in all living cells. They are involved in virtually every physiological process like metabolism, catalysis, signal transduction, cell cycle and transport and they perform structural and mechanical functions such as in the cytoskeleton and in muscles.

In most of these processes interactions between proteins and their ligands play crucial roles. Most of these interactions are unspecific and transient in nature (e.g. interactions with water and ions), some are persistent and may play a structural or functional role (e.g. certain metal ions) and others might be transient but nevertheless highly specific, often resulting in essential changes of the protein or the ligand (e.g. enzyme-substrate complexes or receptor-ligand complexes). Understanding the molecular basis of protein-ligand interactions is thus of great interest, not only for understanding complex biological systems but also for clinical applications.

Although protein-ligand interactions are crucial for the function of a protein, in many cases they are unknown. Despite the kind of ligands interacting with a protein is often known from biochemical analyses, elucidating the structural details of these interactions requires elaborate and time-consuming studies by X-ray crystallography or nuclear magnetic resonance spectroscopy. Therefore, computational methods have become increasingly important to investigate biological systems at an atomistic level. Today, as examples, in-silico approaches facilitate the functional characterization of proteins,¹ allow the identification of possible interactions with small molecules based on three-dimensional protein structures² or help to investigate the molecular mechanism of enzymatic catalysis.³

Ultimately, the function of a protein is determined by its three-dimensional structure, which in turn is governed by its amino acid sequence. Therefore, in cases where no experimentally determined three-dimensional structures are available, comparative modeling techniques can provide insights. These methods rely on the observation that the three-dimensional structure of a protein family is robust against sequence changes.⁴ This allows to build structural models based on similarity to proteins with known structure.

Studies of the interactions of small chemical molecules with the binding site of a disease

related protein can help to develop specific inhibitors with applications both as research tools for probing the effect of inhibition in a protein network as well as early lead compounds for developing new drugs.

1.2 Estimation of Protein-Ligand Interactions

Computational approaches have been developed for a broad variety of applications in molecular biology, ranging from the identification of ligand binding sites, through estimation of protein-ligand interaction energies to detailed descriptions of the electronic structure of catalyzed reactions.

1.2.1 Prediction of Ligand Binding Sites

The number of protein structures with unknown biological function is steadily increasing. To bridge this rapidly growing gap between known sequences and unknown function, numerous computational and experimental techniques have been developed to help identifying the structure-function relationship.^{5,1,6}

Among these methods, computational approaches for determining the precise location of ligand binding sites and protein residues involved in ligand interaction, directly from a protein's sequence, is of high relevance for life science research. Various approaches for the prediction of ligand-binding sites have been proposed,⁷ based on sequence conservation,^{8,9,10,11,12,13} geometric criteria of the protein surface^{14,15,16,17,18} or homology transfer from known structures.^{19,20,21,22}

Recently, methods based on homology transfer have been shown to exhibit excellent results in a blind assessment of prediction methods.^{23,24,25} These methods follow a general scheme: starting from an input sequence, a three dimensional structure is build based on homology modeling techniques. With this model, a database of protein structures with bound ligands is queried to identify proteins with similar structure. Superimposing these structures onto the query structure aligns the bound ligands onto the query and allows to identify contacting protein residues which form the binding site.²² A number of variations to this scheme have been implemented including residue conservation,²² constrained ligand docking²⁶ or local functional site identification.²⁷

Despite the good results of these methods, they are limited to cases where homologous proteins with known ligands are detectable, which is not commonly the case. When homologue structures are available, but their binding sites are unknown, geometric methods, trying to identify the deepest clefts on the protein surface, yield good results. Where no homologue structure of the query protein is detectable, only methods based on sequence conservation are applicable.

1.2.2 Protein-Ligand Docking and Virtual Screening

Computational methods for docking small molecules into the binding sites of biological macromolecules and for scoring their potential interactions with the protein are widely used in drug discovery for hit identification and lead optimization.²⁸ They often help to identify possible drug candidates from a large library of available chemical compounds.

Based on the three dimensional structure of the target protein, docking programs try to predict the best fit of a ligand into the binding pocket. For this, chemical compounds are computationally placed into the target protein binding site and their interaction energy is estimated. Current algorithms used in virtual screening make a number of approximations in order to achieve reasonable computational speeds necessary for screening of large compound libraries. These approximations reduce the numbers of degrees of freedom which are explicitly treated, of which some are replaced with implicit degrees of freedom. This includes constraining of protein motions, implicit treatment of solvent molecules or evaluation of interaction energies based on molecular mechanics force fields or empirical scoring functions.²⁹

In general there are two aims of docking studies: First, the accurate modeling of the binding pose and second, the correct prediction of binding free energies.³⁰

For the sampling of the small molecule's conformational space, a number of algorithms are employed, based on genetic algorithms, incremental build strategies or Monte Carlo sampling. It has been found that most methods for virtual screening work reasonably well in reproducing a close-to-native orientation of the ligand if properly configured and applied to well-behaved systems. However, their ability to predict binding free energies is often very limited.^{31,28,30}

1.2.3 Estimation of Protein-Ligand Binding Affinities

Various approaches have been developed for a more accurate estimation of protein-ligand binding free energies. These calculations often employ molecular dynamics simulations or Monte Carlo sampling of a full system in explicit solvent. However, such methods are very time consuming and are thus only applicable to a small number of compounds.

Most accurate results are obtained with free energy pathway methods, like free energy perturbation (FEP), which sample the whole path from initial to final state. However, these methods are computationally too costly to be routinely applied in a drug discovery process.²⁹ Therefore, numerous approaches have been developed which try to obtain similarly accurate results at lower computational costs. Generally, those methods are end-point methods which consider only the initial and final state. These methods include MM-PBSA / MM-GBSA³² and linear interaction energy (LIE)³³ which are nowadays commonly applied to study the interaction of ligands with biological macromolecules.

1.2.4 Estimation of Reaction Energy Barriers

Mixed Quantum-Mechanical/Molecular-Mechanics Calculations

For describing chemical reactions, quantum-mechanical (QM) methods are often required. However, the application of such methods are limited to systems with a few hundred atoms. On the other hand, even small biological systems, contain orders of magnitude more atoms and are thus incompatible with a full QM treatment. Therefore, mixed quantum-mechanical/molecular-mechanics (QM/MM) approaches have become the method of choice for modeling reactions in biological systems.³

These methods use a QM treatment of the chemically active region, whereas the surrounding is modeled as molecular-mechanics (MM). Combining these methods allows to simulate complex biological systems with good accuracies and reasonable computational costs. These methods can give detailed insights into enzyme catalyzed reactions and other electronic processes, like charge transfer or electronic excitation.³⁴

Potential of Mean Force

The free energy changes as a function of an inter- or intramolecular coordinate is of high relevancy for the computational investigation of physically relevant processes like chemical reactions, ligand migration or conformational changes. The free energy surface along a reaction coordinate is called potential of mean force (PMF). The highest energy point on a PMF is of particular interest, since it corresponds to the transition state of the process, from which kinetic quantities, like rate constants, can be computed. The PMF considers not only the interaction between the solute particles but also incorporates solvent effects if the system is in solution.

Although the PMF is of high relevancy, it is difficult to obtain for complex systems like solvated macromolecules which have many minimum energy conformations. Unfortunately, standard unrestrained molecular dynamics simulations do not adequately sample high energy regions of phase space which contribute significantly to the free energy and thus, yield inaccurate values for the PMF.

One method to overcome these sampling problems is umbrella sampling. In umbrella sampling, the potential energy function is modified in order to adequately sample high energy regions. Bias potentials are placed along a reaction coordinate in order to drive the system from one state to another. The steps along the path are covered by subsequent umbrella windows. In each window, an MD simulation is performed from which the change in free energy can be computed. Subsequently, all windows are combined using the weighted histogram analysis method in order to obtain the free energy profile along the reaction coordinate.³⁵

1.3 Flavivirus

Flaviviruses are small, enveloped RNA viruses, belonging to the Flaviviridae family, together with Hepaciviruses and Pestiviruses.³⁶ The Flavivirus genus contains numerous recognized viral species, which are predominantly transmitted by arthropod vectors, mainly *Aedes* mosquitoes. There are 40 known flaviviruses capable of causing diseases in humans.³⁷ Of those some are medically important pathogens causing significant mortality and morbidity in humans. This includes all four serotypes of dengue virus (DENV1-4), Japanese encephalitis virus (JEV), tick-borne encephalitis virus (TBEV), West Nile virus (WNV) and yellow fever virus (YFV).³⁸ Although vaccines are available for YFV, JEV and TBE, none have been developed for other flaviviral diseases. Currently, there are no specific antiviral drug treatments available against flaviviruses, and disease control is often limited to vector control.

1.3.1 Dengue Fever

Dengue fever (DF), which is caused by all four dengue virus serotypes, is among the most important emerging diseases. In the last 25 years, a dramatic global expansion of DF and the more severe and potentially lethal form of the disease, dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS), has occurred.³⁷ Nowadays, dengue is predominantly prevalent in all tropical regions with annually 50-100 million cases of DF, 500'000 cases of DHF/DSS and around 20'000 death worldwide. For dengue virus, four closely related serotypes have been isolated, where each serotype is sufficiently different, that no cross-protection can occur. Furthermore, sequential infection with different DENV serotypes in long intervals can produce unusually severe disease.³⁹

Vaccine development for DENV has been a challenge for decades, mainly due to the inability of vaccines to protect simultaneously against all four distinct serotypes.⁴⁰ In the absence of vaccines, specific drug treatments are needed, but none were developed so far.

1.3.2 Dengue Virus

Like all flaviviruses, dengue virus is a enveloped, single stranded, positive sense RNA virus. The genome is packaged by viral capsid protein (C) in a host-derived lipid bilayer, into which 180 copies of the envelope protein E in complex with the membrane protein (M) are embedded. This results in a smooth and spherical virion with a diameter of 50 *nm* (Figure 1.1).⁴¹

The single stranded, 11 kb positive sense RNA genome has a single long open reading frame which is flanked by 5'- and 3'-untranslated regions (UTR), which have secondary structure that is essential for the initiation of translation and for replication. The 5' end of the genome has a type 1 cap, whereas the 3' end lacks a poly-A tail.⁴² In the host, the viral RNA is translated into a polyprotein which is cleaved by both host and viral proteases into three structural and seven non-structural (NS) proteins (Figure 1.2).⁴³

The non-structural proteins are involved in viral RNA replication. The best characterized proteins are NS3 and NS5. NS3 has three distinct activities: serine protease in complex with

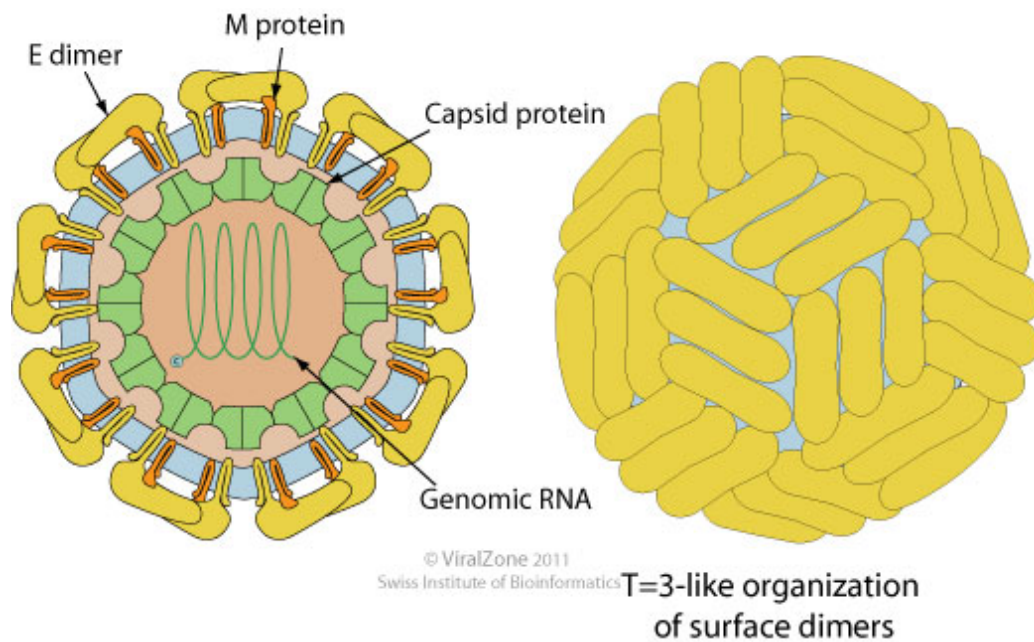


Figure 1.1: Depiction of the flavivirus virion. Source: ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics

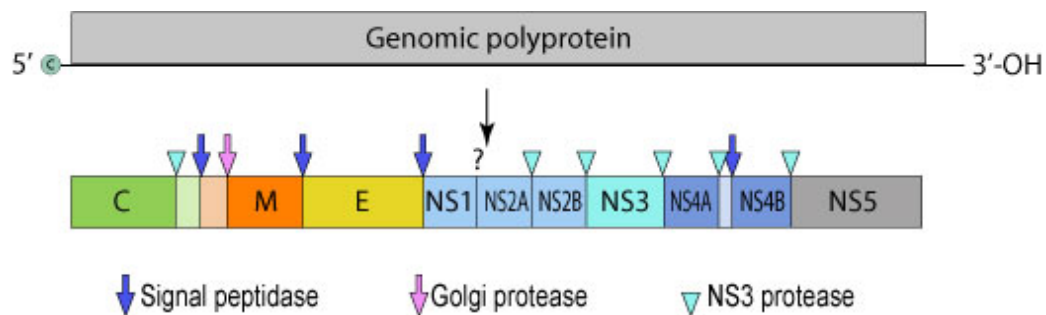


Figure 1.2: Flavivirus genome (top) and polyprotein (bottom) with cleavage sites. (Source: ViralZone www.expasy.ch/viralzone, Swiss Institute of Bioinformatics)

NS2B, required for polyprotein processing; helicase/NTPase activity, required for unwinding double stranded replicative form of RNA; RNA triphosphatase, required for capping of nascent viral RNA. NS5 has three enzymatic functions: S-adenosyl-L-methionine (SAM) dependent methyltransferase (MTase)^{44,45} and guanylyltransferase⁴⁶ required for maturation of the RNA cap; RNA-dependent RNA polymerase (RdRp) required for RNA replication. NS1 is required for flaviviral replication and presumably involved in negative-strand synthesis. NS2A is a trans-membrane protein involved in membrane generation during virus assembly. NS4A is a membrane protein involved in the formation of the viral replication complex. NS4B inhibits type I interferon response of host cells.⁴⁷

A type 1 cap structure is found at the 5'-end of both viral and cellular eukaryotic RNA.⁴⁸ It is essential for viral replication, since it ensures RNA stability by protecting against RNases and it enhances recognition by the ribosomes.^{49,50} The capping process results from four chemical reactions, catalyzed by viral enzymes (Figure 1.3). Starting from the unaltered 5'-

end, consisting of the final nucleotide to which a triphosphate is attached at the 5' position, an RNA triphosphatase (presumably NS3) removes the terminal phosphate group. In the following step, a guanylyltransferase (NS5) transfers one molecule of guanosine monophosphate to the 5'-diphosphate RNA. Finally, the terminal guanosine moiety is methylated in the N7-position by a methyltransferase (NS5), which leads to a cap 0 structure. In addition, a methyltransferase (NS5) further methylates the 2'-hydroxy group of the first RNA nucleotide which leads to the cap 1 structure.⁴⁴ For all flaviviruses, a cap 1 structure with the form ${}^{7Me}GpppA_{2'OMe}G$ -RNA is always present in mature viral RNA, where the first two nucleotides (A,G) are strictly conserved among all flaviviruses.⁴⁸

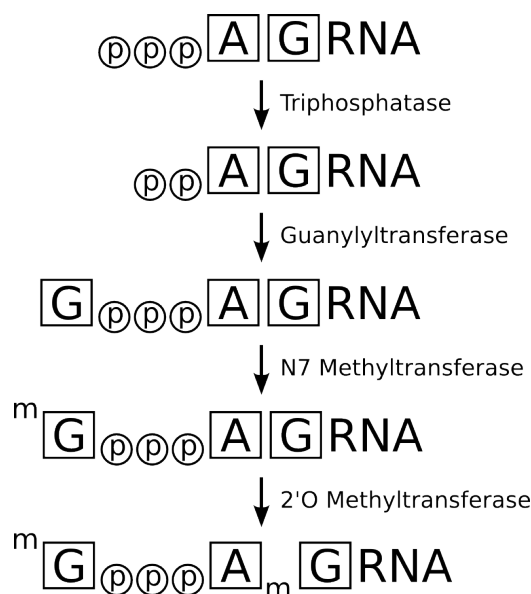


Figure 1.3: Schematic overview of the processes involved in RNA capping.

1.3.3 NS5 Methyltransferase

One of the viral enzymes involved in the capping process, is the NS5 methyltransferase (MTase) which is located at the N-terminal domain of the NS5 protein. This enzyme shares a common fold with many SAM dependent methyltransferases although sequence identity within this family is very low (10-15%).^{44, 51}

Twelve X-ray crystal structures of the dengue MTase domain complexed with S-adenosyl-L-homocysteine (SAH), ribavirin triphosphate (RTP), as well as a variety of RNA cap analogues, have been published. So far, no full length NS5 crystal structure, consisting of the N-terminal MTase and the C-terminal RdRp domains, has been solved.

The enzyme has two specific binding sites where ligands have been co-crystallized: The position of SAH indicates the binding of the methyl donor, SAM. RNA cap analogues bind to a shallow second pocket. The two binding sites are connected by a common Y-shaped positively charged cleft, which suggests the placement of capped RNA along the cleft, positioning the first RNA nucleotide close to SAM, compatible with 2'O-methylation (Figure 1.4).^{44, 52, 53}

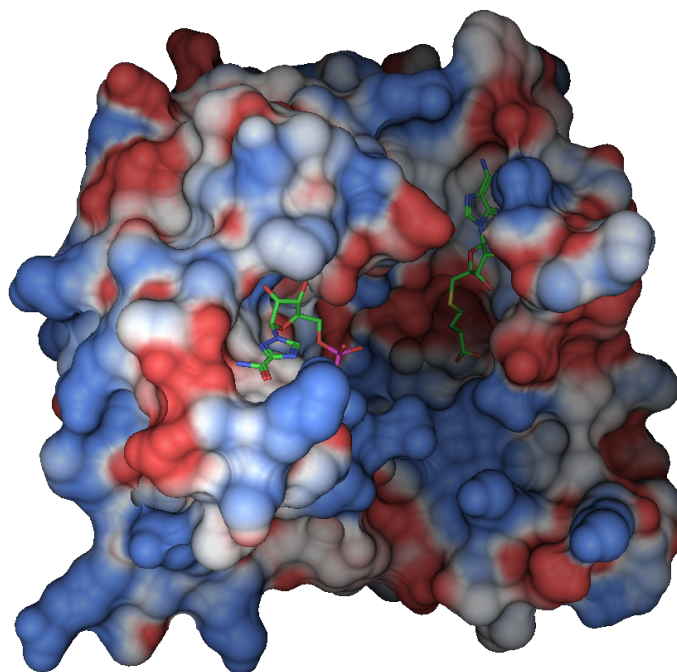


Figure 1.4: Structure of the dengue virus methyltransferase domain.

Although competitive inhibitors are known for both sites of the dengue MTase, the medicinal chemistry of DENV is still in its very early stage.⁵⁴ SAH, as well as sinefungin and dehydrosinefungin have been characterized as efficient sub-micromolar competitive inhibitors of the MTase. The structural similarity to SAM strongly suggests their interaction with the SAM pocket.⁵⁵ A virtual screening campaign, identified a further inhibitor based on structural similarity to SAM, which inhibits MTase activity in the medium-micromolar range.⁵⁶ Furthermore, RTP has been found to inhibit dengue MTase, but shows only weak activity.⁵² An additional inhibitor with activity in the low micromolar range was found, which is expected to bind to the RNA cleft.⁵⁷ Recently, Lim et al. have developed a small molecular inhibitor based on SAM analogs, which selectively blocks DENV MTase.⁵⁸ In addition, using high throughput screening, Stahla-Beek et al have discovered the first inhibitor of the enzyme's guanylyltransferase activity.⁵⁹

The NS5 MTase catalyzes both the guanine N7 and the ribose 2'O methylation, generating sequentially $\text{GpppA-RNA} \rightarrow {}^7\text{MeGpppA-RNA} \rightarrow {}^7\text{MeGpppA}_{2'\text{OMe}}\text{-RNA}$ (Figure 1.5).^{44, 60, 50} For both reactions in flaviviruses, no mechanisms at an atomistic level is known and no structure with a short capped RNA in a conformation suitable for methyltransfer has been solved.

Sequence alignment revealed that the four residues Lys61, Asp146, Lys181 and Glu217 are conserved among many MTases. From biochemical and mutagenesis studies, it has been shown that those four residues are critical for the functioning of the methyltransfer reactions and thus the replication of the virus itself. However, different dependencies on the residues within this motive were found for the N7 and the 2'O methylation reaction, which suggests different underlying mechanisms.⁶⁰ In addition, further mutagenesis studies identified two distinct sets of amino acids on the enzyme's surface required for the N7 and the 2'O methylation, which

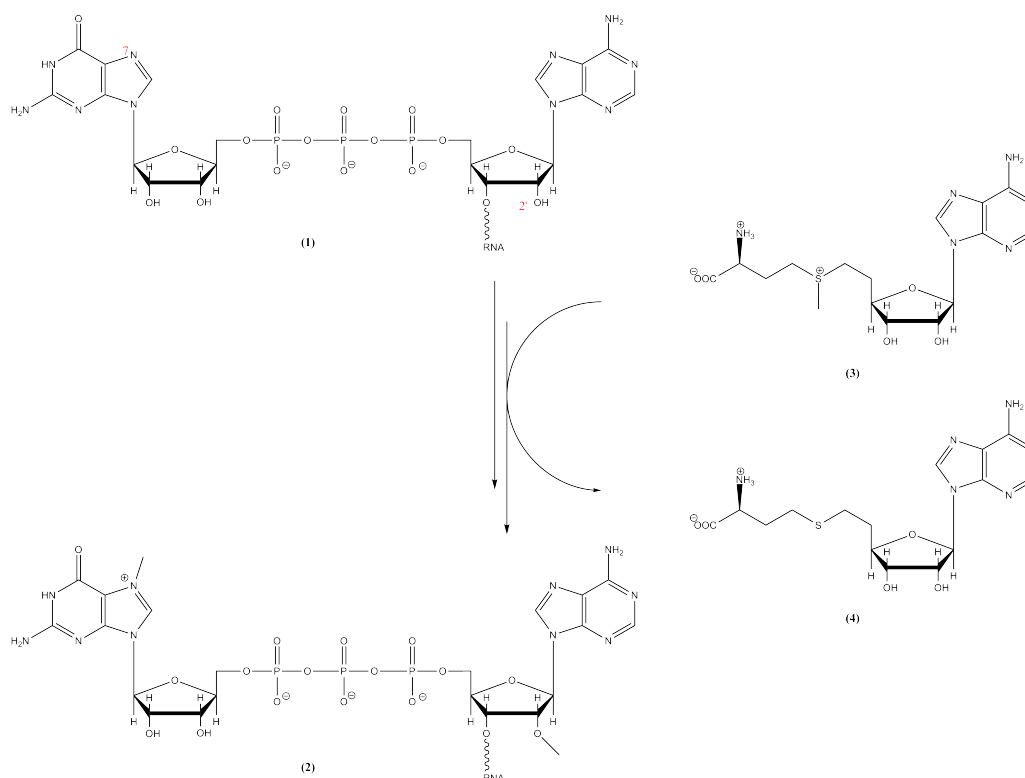


Figure 1.5: Overview of the two methyltransfer reactions catalyzed by the NS5 MTase. Reactants: unmethylated RNA (1), S-adenosyl-L-methionine (3). Products: doubly methylated RNA (2), S-adenosyl-L-homocysteine (4).

suggests that the RNA adopts two different binding modes.⁶¹ In addition, for the N7 reaction, it has been found that it can only take place on RNA templates comprising at least 74 nucleotides of the viral 5' UTR sequence.⁵⁰

From the structures of vaccinia virus VP39 2'O-MTase and mutagenesis studies of RrmJ MTase, a mechanism for the 2'O methylation has been suggested for those enzymes.^{62, 63} It was proposed that the methyltransfer from SAM to the 2'-hydroxy group of the RNA ribose moiety proceeds as a nucleophilic S_N2 type reaction and that it is catalyzed by the conserved residues in the Lys61-Asp146-Lys181-Glu217 tetrad, which mediates deprotonation of the 2'-hydroxy group.^{64, 65} From the structure of the distantly related Ecm1 N7-MTase, an in-line mechanism with no direct contact of the protein was suggested for the N7-methylation. There, the catalysis seems to be achieved through close proximity of the guanosine N7-atom and the SAM methyl group.⁶⁶

1.4 Objectives

This work aims at a better understanding of the molecular basis of disease related viral methyltransferases, their interactions with small molecules and the catalytic mechanism, which may on the long perspective help to develop a treatment against neglected tropical diseases. Furthermore, we aim to advance the current methods for the computational prediction of a protein's molecular function and its biological role in the cell. In addition, we aim to complement currently available computational strategies for estimating protein ligand interaction energies.

This thesis is organized as follows: First, the results from our study on the identification of novel dengue virus methyltransferases are given, followed by a description of the further experimental characterization of the inhibitory effects of selected promising compounds. Second, the computational and experimental analysis of the mechanism of action of the dengue virus methyltransferase is described. Third, insights of the assessment of ligand binding site prediction methods are presented, indicating current limitations in prediction methods and their assessment. In addition, the subsequent implementation of these suggestions is described. Fourth, the development of a rapid scoring function for identifying the correct pose of a ligand bound to a protein is presented. Finally, a novel human-computer interface device is described.

Chapter 2

Identification and Validation of Novel Dengue Methyltransferase Inhibitors

The search for lead compounds which inhibit the dengue methyltransferase can be significantly facilitated by using computational methods. We are using high-throughput structure-based virtual screening of a library of over five million purchasable compounds to reduce the library to a list of a few hundred candidates which are tested in vitro. Subsequently, the experimental binding affinities can be used to increase the accuracy of our predictions and to select further compounds for experimental verification.

The focus of our study is to obtain a better understanding of the molecular properties of viral methyltransferase active sites and their interactions with small molecules, which will guide our search for novel lead compounds against neglected tropical diseases.

For the discovery of novel classes of compounds inhibiting dengue MTase, a combination of large-scale structure-based virtual screening and enzymatic inhibition assays was employed. The virtual screening approach was based on a multi-stage docking strategy and was applied to a library of over five million commercially available compounds. The funnel like strategy included multiple Glide⁶⁷ docking steps of selected subsets with increasing accuracy as well as a refinement and a selection step. Promising compounds were subsequently assayed in vitro using a scintillation proximity assay⁵⁵ at the Novartis Institute for Tropical Diseases (NITD) in Singapore.

Additionally, ligand based virtual screening methods were applied to retrieve additional active compounds. Thus, we constructed a pharmacophore model, based on experimental data obtained by this study and common receptor-ligand interactions predicted by our docking calculations. This model was used to obtain further candidates from the initial compound library which were subsequently assayed in vitro.

From the list of 263 candidates which were assayed experimentally, ten compounds were found to specifically inhibit dengue MTase with IC₅₀ values in the low μM range. Due to the broad setup of the initial library, those active compounds represent a set of diverse chemotypes and predicted binding modes, leading to a variety of different starting points for further drug discovery efforts.

During compound screening, numerous false positive hits were encountered, which non-

specifically inhibit the dengue MTase. Thus, we have tested computational methods to predict non-specific inhibition due to compound aggregation. Those methods use supervised machine learning techniques to classify specific from non-specific inhibitors based on calculated physicochemical properties. While our trained classifier performs well within one dataset, misclassification rates are significantly increased when applied to a completely new set of compounds. Our results suggest that prediction of aggregation behavior is not transferable between assay conditions or biological targets. Thus, such classifiers cannot be used to eliminate predicted non-specific compounds prior to in vitro assays.

To further characterize the active compounds and to validate their specific interaction with the dengue MTase, additional experimental assays were performed. Of the ten active compounds identified through our virtual screening approach, five promising compounds were selected for further follow-up experimental assays to confirm their specific inhibition of the MTase and to distinguish between inhibitory activity of the 2'O and the N7 MTase function. Thereby, the inhibitory activity of the two most active compounds was confirmed.

In addition, we have developed an isothermal titration calorimetry (ITC) assay in order to measure binding constants of the selected active compounds. Due to solubility issues only a subset of three compounds was assayed in the ITC experiment, however, their binding to the dengue MTase could not be confirmed.

For obtaining a better understanding of the important interactions governing protein-ligand binding, for the validation of predicted binding modes as well as for future structure-based compound optimizations, we started efforts to obtain X-ray crystal structures of the inhibitors bound to the MTase.

2.1 Screening For Novel Inhibitors

In the following, a published manuscript is included:

“Novel Inhibitors of Dengue Virus Methyltransferase: Discovery by in Vitro-Driven Virtual Screening on a Desktop Computer Grid”

My contributions to this joint work were the following:

In-depth analysis of high-throughput docking results

Development of pharmacophore hypotheses based on experimentally validated high-throughput docking hits

Pharmacophore based screening for additional compounds and subsequent rescoring of obtained hits

Retrospective analysis of refinement and rescoring procedures

Modeling of compound aggregation behaviour

Novel Inhibitors of Dengue Virus Methyltransferase: Discovery by in Vitro-Driven Virtual Screening on a Desktop Computer Grid

Michael Podvinec,[†] Siew Pheng Lim,[‡] Tobias Schmidt,[†] Marco Scarsi,[†] Daying Wen,[‡] Louis-Sebastian Sonntag,[‡] Paul Sanschagrin,[§] Peter S. Shenkin,[§] and Torsten Schwede^{*†}

[†]Swiss Institute of Bioinformatics and Biozentrum, University of Basel, Klingelbergstrasse 50, CH-4056 Basel, Switzerland, [‡]Novartis Institute for Tropical Diseases, 10 Biopolis Road, Chromos #05-01, 138670 Singapore, and [§]Schrödinger LLC, 120 West 45th Street, 29th Floor, New York, New York 10036-4041

Received June 1, 2009

Dengue fever is a viral disease that affects 50–100 million people annually and is one of the most important emerging infectious diseases in many areas of the world. Currently, neither specific drugs nor vaccines are available. Here, we report on the discovery of new inhibitors of the viral NS5 RNA methyltransferase, a promising flavivirus drug target. We have used a multistage molecular docking approach to screen a library of more than 5 million commercially available compounds against the two binding sites of this enzyme. In 263 compounds chosen for experimental verification, we found 10 inhibitors with IC₅₀ values of < 100 μM, of which four exhibited IC₅₀ values of < 10 μM in in vitro assays. The initial hit list also contained 25 nonspecific aggregators. We discuss why this likely occurred for this particular target. We also describe our attempts to use aggregation prediction to further guide the study, following this finding.

Introduction

Dengue fever is a viral disease that is transmitted between human hosts by *Aedes* mosquitoes, particularly *Aedes aegyptii*. In 1997, 20 million cases of dengue fever were estimated to occur annually.^{1,2} Partially because of increased urbanization and failure to effectively control the spread of the insect vector, more recent estimates suggest this number has risen to 50–100 million, and dengue fever is now seen as one of the most important emerging infectious diseases in many areas of the world.^{3–5} Mild cases of dengue fever result in severe flulike symptoms, including fever, headache, and myalgia, but more severe cases can progress into a hemorrhagic fever and shock syndrome with considerable lethality.⁶ Current treatment practice is nonspecific and symptomatic with a regimen of analgesics and fluid replacement, as neither specific drugs nor vaccines are available.¹

Dengue virus is a plus-strand RNA virus belonging to the *Flavivirus* genus of the *Flaviviridae* family. Four serotypes have been isolated (DENV1–DENV4), and exposure to each of the serotypes conveys only partial immunity. Moreover, the presence of heterologous antibodies against a serotype other than the present infection may precipitate the more severe forms of dengue fever in patients.⁷ In the absence of efficient and cost-effective vaccines, the development of inhibitors of viral or cellular enzyme targets as antiviral therapeutic agents is of particular interest.

The dengue genome, a single RNA strand 10.7 kb in length, is translated into a single polyprotein and later cleaved by viral

and cellular proteases into 10 mature proteins. Three of the proteins have a structural role (C, prM, and E). In addition, seven nonstructural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) are formed.⁸

Of the latter, NS3 and NS5 are the best understood to date, and both enzymes exhibit multiple domains and functions.^{9,10} NS5 is the largest (900 amino acids) and most conserved protein in the dengue genome (67% sequence identity among serotypes 1–4).⁸ It contains the RNA methyltransferase (MTase)^a domain, as well as the RNA-dependent RNA polymerase necessary for virus replication. In this study, we focus on the discovery of compounds inhibiting the NS5 MTase, which has been proposed as a promising drug target against flaviviruses by us and others.^{11–13}

The 5' end of the dengue genome contains a type 1 cap structure, followed by the nucleotides AG, which are conserved in all flaviviruses.¹⁴ Appropriate capping of cellular and viral RNA is known to increase translation efficiency as well as RNA half-life.^{15,16} Host RNA is transcribed in the nucleus and processed by the cellular capping machinery. Dengue virus replication, however, occurs at the membrane of the endoplasmic reticulum; hence, a viral MTase is required for capping of the nascent viral RNA. Of the four steps necessary in *Flavivirus* cap formation, the final two methylation reactions are catalyzed

^aAbbreviations: ATA, aurintricarboxylic acid; CF-I, cell-based flavivirus immunodetection; Cpd, compound; DENV, dengue virus; FN, false negative; FP, false positive; MTase, methyltransferase; NCI DTP, National Cancer Institute Developmental Therapeutics Program; MM-GBSA, molecular mechanics-generalized Born surface area; NS, nonstructural protein; RF, random forest; rmsd, root-mean-square distance; RTP, ribavirin triphosphate; SAH, *S*-adenosyl-L-homocysteine; SAM, *S*-adenosyl-L-methionine; SPA, scintillation proximity assay; TN, true negative; TP, true positive.

*To whom correspondence should be addressed: Swiss Institute of Bioinformatics and Biozentrum, University of Basel, Klingelbergstrasse 50, CH-4056 Basel, Switzerland. Telephone: +41 61 267 15 81. Fax: +41 61 267 15 84. E-mail: torsten.schwede@unibas.ch.

by NS5 MTase with *S*-adenosyl-L-methionine (SAM) as the methyl donor, generating *S*-adenosyl-L-homocysteine (SAH) as a byproduct.^{17,18} The cap guanine is methylated at the N7 position, resulting in a type 0 cap structure. Subsequently, the first RNA base, adenosine, is methylated at the 2'-OH group of the ribose, resulting in the formation of a type 1 cap structure.

The three-dimensional (3D) structure of the dengue NS5 MTase domain was the first *Flavivirus* MTase structure to be determined by X-ray crystallography.¹⁸ Structures of the MTase complexed with *S*-adenosyl-L-homocysteine (SAH), the nonhydrolyzable GTP analogue GDPMP, ribavirin triphosphate (RTP), and a variety of RNA cap analogues (GpppA, GpppG, ⁷MeGpppA, ⁷MeGpppG, and ⁷MeGpppG_{2'OMe}) have been published.^{17–20}

Dengue MTase has an overall globular fold and shares a common fold with many SAM-dependent MTases, consisting of a seven-stranded β -sheet enclosed by four α -helices (subdomain 2).²¹ This domain is surrounded by subdomain 1, an N-terminal extension of a helix–turn–helix motif, followed by a β -strand, an α -helix, and subdomain 3, a C-terminal extension consisting of an α -helix and two β -strands, spatially located between subdomain 2 and the N-terminal extension.¹⁸ The enzyme has two specific binding sites where ligands have been cocrystallized (cf. Figure 1). The position of SAH indicates the binding site of the methyl donor, SAM. RNA cap analogues bind to a shallow second pocket formed between subdomains 1 and 2 (cf. Figure 1A). The two binding sites are connected by a common Y-shaped cleft, which suggests the placement of capped RNA along the cleft, positioning the first RNA nucleotide close to SAM, compatible with 2'-O-methylation. These positions are in accordance with observed positions of the RNA and cofactor in a complex structure of vaccinia virus VP39 MTase.²²

Here, we present the results of our efforts to find novel classes of compounds inhibiting dengue MTase, potentially blocking viral replication. We have used a combination of large-scale structure-based computational analysis and enzyme inhibition assays. On the basis of structural analysis of dengue MTase, separate binding sites for RTP and SAM were targeted. For both sites, competitive inhibitors are known: SAH, sinefungin, and dehydrosinefungin have been characterized as efficient submicromolar competitive inhibitors of this MTase, and structural similarity to SAM strongly suggests their interaction with the SAM pocket.²³ Furthermore, two inhibitors of dengue MTase were published concomitant to this work. An inhibitor ($IC_{50} = 60.5 \mu M$) has been found by Luzhkov et al. based on structural similarity to SAM,¹³ and a docking study by Milani et al. has found aurintricarboxylic acid (ATA) to be a low-micromolar inhibitor of dengue MTase ($IC_{50} = 2.3 \mu M$).²⁴ On the basis of the specific structural interactions of RTP ($IC_{50} = 101 \mu M$)¹⁹ and nucleotide or cap analogues with the RNA cap binding site, we consider this site a valid second target for inhibitors.

Our virtual screening approach was based on initial high-throughput docking calculations performed on a library of more than 5 million commercially available compounds. Using a personal computer (PC) grid to harness the idle computing power of our university's PCs, we were able to perform these calculations without prior focusing of the compound library. After the compounds had been docked, compound poses were refined, and promising candidates were assayed in vitro. Insights from these assays combined with pharmacophoric searches based on the predicted binding

mode of actives were then used to select further compounds for follow-up testing. In the following, we will discuss our combined screening study, as well as the results obtained computationally and in vitro.

Materials and Methods

Chemical Compounds. All compounds in the docking database were associated with purchasing information, and compounds selected for inhibition assays were obtained from a variety of vendors. Compounds **1–9** (Table 1) were obtained from the NCI DTP Open Chemical Repository (<http://dtp.nci.nih.gov>) with the following compound codes: NSC12451, NSC15765, NSC26899, NSC49419, NSC54771, NSC84407, NSC91788, NSC14778, and NSC140047, respectively. Compounds **10–12**, **14**, **15**, **17**, **18**, **20**, **21**, **27**, and **33** were obtained from ChemBridge Corp. (San Diego, CA) (codes 5654575, 6490771, 7018889, 7936171, 7208655, 7746191, 7778100, 5219400, 7364286, 5255882, and 5917902, respectively). Compounds **13**, **24**, **26**, and **35** were from Enamine Ltd. (Kiev, Ukraine) (codes T0520-2463, T0511-8111, T5237786, and T5285909, respectively). Compounds **16**, **19**, **22**, and **23** were from InterBioScreen (Moscow, Russia) (codes STOCK1N-55803, STOCK2S-36613, STOCK3S-13122, and STOCK5S-06910, respectively). Compounds **25** and **28** were from InterChim (Montluçon, France) (codes STOCK1N-17364 and UZI/9041345, respectively). Compounds **29** and **30** were from Aurora Fine Chemicals (San Diego, CA) (codes Kenb-0135169 and Kina-0056391, respectively). Compound **31** was from Ambinter SARL (Paris, France) (code PHAR058572). Compound **32** was from TimTec LLC (Newark, DE) (code ST057026), and compound **35** was from Life Chemicals (Burlington, ON) (code F0777-1485).

Molecular Modeling. (i) **Analysis of Dengue Methyltransferase Mutations and Structural Variability.** For structural studies and for docking, an X-ray crystallographic structure of DENV2 MTase with bound SAH and RTP was used [Protein Data Bank (PDB) entry 1R6A]. To assess the conservation of protein residues, we extracted dengue MTase sequences from a database of all dengue sequences in UniProtKB release 14.0 using a blastp search with the sequence of PDB entry 1R6A as a query.^{25,26} From the retrieved set of sequences, redundant sequences were removed, and 127 unique sequences were aligned using ClustalW with standard parameter settings.²⁷ Next, identity histogram values (I_p) were calculated at each position, where $I_p = (M - 1)/(N - 1)$, with p being the position in the alignment, M the number of prevalent residues in row p , and N the total number of residues in row p . Finally, residues were colored by identity histogram values in the Chimera software package.²⁸

To study the structural variability seen in dengue MTase crystal structures, we obtained all available X-ray structures from the PDB²⁰ and optimally superposed their backbone atoms to the reference structure (PDB entry 1R6A). The average per-residue root-mean-square distances (rmsd) between the 1R6A structure and all other structures were calculated using VMD version 1.8.6²⁹ and colored accordingly.

(ii) **Library of Purchasable Chemical Compounds.** The compound library for screening was collected as follows. The all-purchasable subset of the ZINC V5 database, comprising ~2.7 million molecules from a variety of vendors, was obtained from <http://zinc5.docking.org/>.³⁰ To this collection were added 2.4 million nonredundant compounds from the Schrödinger in-house CACDB database of commercially available compounds. Ligands were prepared for docking using the *LigPrep* process (Schrödinger Suite 2007, Schrödinger LLC, NY). Briefly, the procedure was as follows. Ligands were desalted, neutralized, and parametrized using the OPLS 2005 force field. Next, tautomers and ionization states expected to occur in the pH range of 5.0–9.0 were generated using *ionizer* (Schrödinger Suite 2007). Wherever the stereochemistry of chiral centers

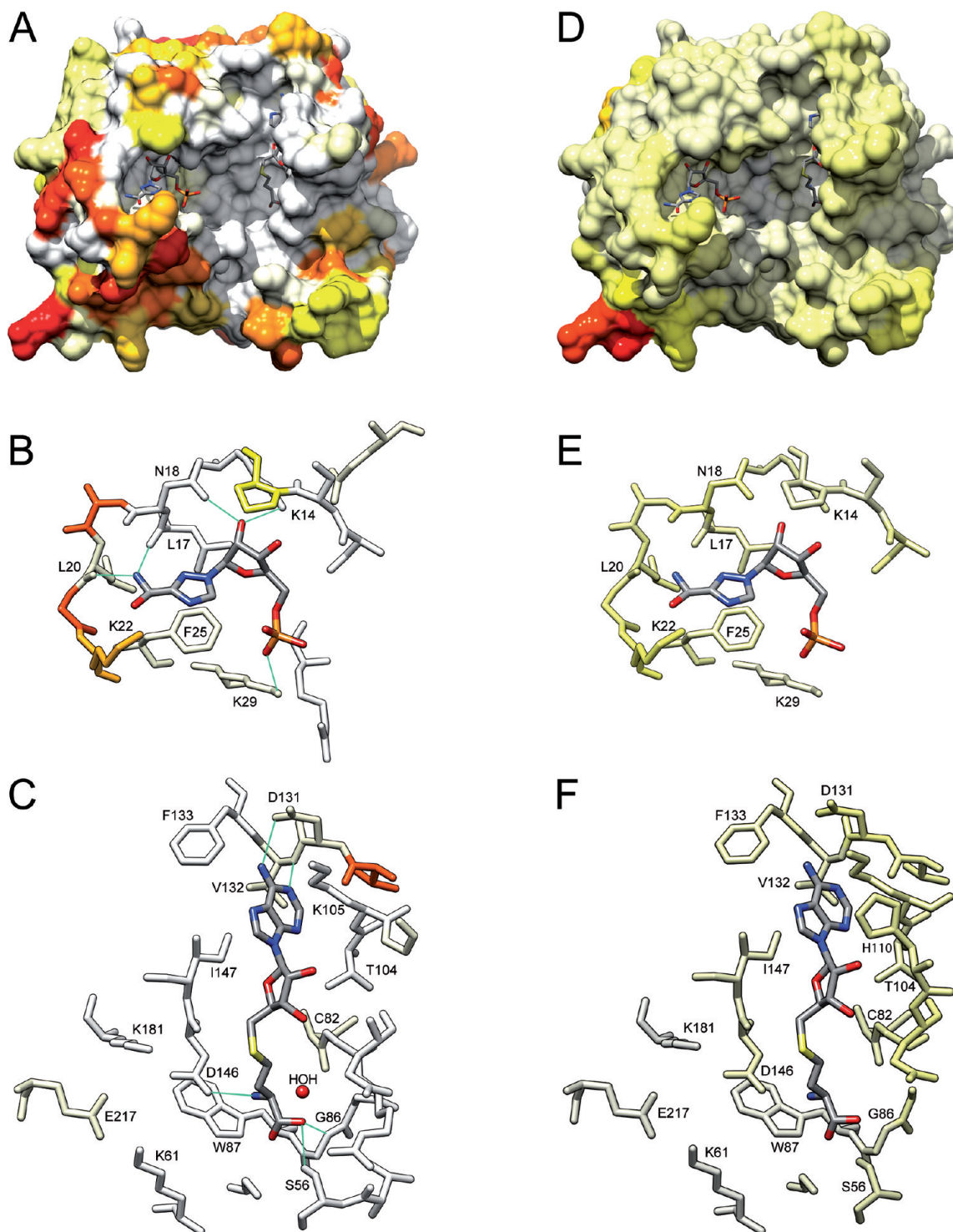


Figure 1. Sequence and structural conservation of DENV2 MTase. (A–C) Sequence conservation of dengue MTase. Sequence conservation is expressed as the identity histogram (I) of an alignment of 127 nonredundant dengue MTase sequences retrieved from UniProtKB. (A) Overall structure of DENV2 MTase in complex with RTP (left) and SAH (right). Ligands are displayed as element-colored licorice sticks. Surface gradient: from light gray ($I = 1$) to yellow ($I = 0.947$, i.e., 95% identical residues) to red ($I = 0.323$, i.e., 33% identical residues). (B) RNA cap binding site. Residues surrounding the inhibitor RTP are shown as licorice sticks, colored by degree of conservation as in panel A. RTP is shown in element-colored sticks (only one of three phosphate groups is shown). Residues undergoing key interactions with the ligand are labeled, and hydrogen bonds are depicted as cyan lines. (C) SAM binding site. Residues surrounding the reaction byproduct SAH are shown as licorice sticks, colored by degree of conservation as in panel A. SAH is shown in element-colored sticks. Residues undergoing key interactions with the ligand are labeled, and hydrogen bonds are depicted as cyan lines. (D–F) Structural variation calculated as a per-residue root-mean-square distance (rmsd) between the displayed structure and all other available DENV2 MTase crystal structures with and without bound ligands. (D) Overall MTase structure. The average rmsd is expressed as a color gradient: from light gray (rmsd = 0.0 Å) to yellow (rmsd = 1.0 Å) to red (rmsd \geq 2.0 Å). (E) RNA cap binding site with residues close to the inhibitor RTP, colored as in panel D. (F) SAM binding site with residues close to the reaction byproduct SAH, colored according to the rmsd as in panel D. Surfaces were calculated using the MSMS package.⁵⁷

Table 1. Predicted Binding Pocket and Measured Inhibition of Docked Compounds

Cpd	ID ^a	binding pocket	IC ₅₀ (μM) (Hill coefficient)	IC ₅₀ (μM)		activity retained	EC ₅₀ (μM)	CC ₅₀ (μM)
				with 0.1% TX100 (μM)	(Hill coefficient)			
1	NSC12451	SAM	29.9 (n.d.)	> 100 (n.d.)				
2	NSC15765	SAM	14.3 (1.9)	43.4 (2.3)	yes	> 100	> 100	
3	NSC26899	SAM	25.3 (3.1)	> 100 (n.d.)				
4	NSC49419	SAM	27.59 (2.5)	> 100 (n.d.)				
5	NSC54771	SAM	27.53 (2.9)	> 100 (n.d.)				
6	NSC84407	SAM	31.43 (2.3)	> 100 (n.d.)				
7	NSC91788	SAM	29.03 (1.4)	> 100 (n.d.)				
8	NSC14778	SAM	1.52 (3.1)	9.46 (2.5)	yes	> 100	> 100	
9	NSC140047	SAM	8.78 (1.9)	4.47 (2.2)	yes	> 100	> 100	
10	ZINC 02911543	RNA cap	7.56 (1.5)	7.14 (1.4)	yes	> 100	> 100	
11	ZINC 01174529	RNA cap	6.83 (2.9)	> 100 (n.d.)				
12	ZINC 03461039	RNA cap	7.11 (2)	> 100 (n.d.)				
13	ZINC 03287966	RNA cap	8.81 (2.3)	> 100 (n.d.)				
14	ZINC 01078518	RNA cap	9.28 (2.4)	64.2 (4.4)	yes	12	22.7	
15	ZINC 01138375	RNA cap	11.35 (3.2)	> 100 (n.d.)				
16	ZINC 02129857	RNA cap	11.92 (1.9)	> 100 (n.d.)				
17	ZINC 01112283	RNA cap	13.16 (2.5)	> 100 (n.d.)				
18	ZINC 02849675	RNA cap	17.64 (2.8)	> 100 (n.d.)				
19	ZINC 00632055	RNA cap	20.32 (1.3)	> 100 (n.d.)				
20	ZINC 01467812	RNA cap	37.46 (1.5)	> 100 (n.d.)				
21	ZINC 02826899	SAM	2.91 (2.7)	> 25 (n.d.)				
22	ZINC 01878835	SAM	4.29 (4.1)	> 25 (n.d.)				
23	ZINC 01758620	SAM	9.62 (2.1)	> 100 (n.d.)				
24	ZINC 00633950	SAM	12.84 (1.7)	> 100 (n.d.)				
25	CACDB 1751080	SAM	16.87 (2.2)	79.8 (0.9)	yes	10.9	30.7	
26	ZINC 02642996	SAM	16.09 (1.6)	> 100 (n.d.)				
27	ZINC 01226983	SAM	21.11 (2)	> 100 (n.d.)				
28	ZINC 02750651	RNA cap	2.81 (1.6)	19.55 (1.3)	yes	> 100	> 100	
29	CACDB964942	RNA cap	13.50 (1.7)	87.1 (2.3)	yes	50.0	75.1	
30	CACDB1563494	SAM	9.84 (2.1)	> 100 (n.d.)				
31	ZINC 01832826	RNA cap	4.42 (1.8)	44.5 (4.1)	yes	> 100	> 100	
32	ZINC 01078734	RNA cap	12.39 (1.8)	> 100 (n.d.)				
33	ZINC01196449	RNA cap	7.99 (1.3)	> 100 (n.d.)				
34	ZINC02379945	RNA cap	14.50 (1.4)	> 100 (n.d.)				
35	ZINC03369470	RNA cap	4.80 (2.1)	4.91 (1.6)	yes	> 100	> 100	

^aCompound structures are depicted in Table S4 of the Supporting Information. n.d. = not determined.

was not specified, a maximum of two chiral centers was expanded into stereoisomers. Up to two low-energy conformations were produced for ligands with flexible ring systems. Ligand structures were minimized in implicit solvent using *bmin* (Schrödinger Suite 2007). The final library consisted of 5428096 structures.

(iii) Protein Preparation. In preparation for virtual screening, the enzyme structure from PDB entry 1R6A was modified as follows. All sulfate ions and water molecules found in the crystal structure were removed with the exception of HOH11, a structural water molecule found close to SAH. Moreover, the pose of the flexible Lys22 residue was replaced with an alternate rotamer, opening up the front of the RNA cap site to potentially accommodate larger ligands.

(iv) Ligand Docking and Compound Selection Procedures. Virtual screening and docking were performed using Glide version 4.5 (Schrödinger Suite 2007) using default docking parameter settings. A set of docking grids was generated independently for the RNA cap site and the SAM binding site using the default parameters. For the SAM site, the ligand's ability to form a hydrogen bond to the backbone N of Val132 (as observed with SAH) was required as a docking constraint. Next, a "funnel" strategy was employed for virtual screening. Initially, all compounds were docked using Glide in HTVS (High-throughput Virtual Screening) mode. After this rapid screening, the following compounds were selected for the next round. (1) All compounds ranked in the top 10% by GlideScore were picked. (2) All isomers (enantiomers, tautomers, and ring conformers) or alternate protonation states of compounds selected

under 1 were chosen. (3) All docked poses forming a hydrogen bond to the Val132 backbone nitrogen were selected using a relaxed distance criterion of 3 Å. In the next round, these compounds were docked into the respective binding sites, using the Glide SP (Standard Precision) protocol. From this stage, compounds were selected as follows. (1) The top 10% of the compounds for each binding site by GlideScore were chosen. (2) Isomers of compounds selected in step 1 were included if found in the top 20% of compounds. These compounds were finally docked using the Glide XP (Extended Precision) procedure, and the 4000 top-ranked molecules from each binding site were selected for further refinement. Details on the number of compounds selected in each step are given in Results.

Following docking, selected compounds were passed through further refinement steps. (1) Additional input conformations for each selected compound were generated by reconstructing the geometry of each of the hit compounds and minimizing in implicit solvent or vacuum using the OPLS-AA or MMFF94 force fields^{31,32} using MacroModel (Schrödinger Suite 2007). Alternate conformations were docked using Glide XP, and only the best-scoring pose was retained. The rationale for this enhanced sampling procedure was to ensure that found poses and scores are not influenced by subtle biases in the starting conformations of compounds induced by the force field. (2) We next applied a correction term to the docking score to account for internal ligand strain. The ligand strain correction term was calculated by optimizing the docked pose of the free ligand resulting from step 1 with torsion angle restraints and then

without such restraints. A portion of the difference in minimized energies (25% of the strain energy in excess of 4 kcal/mol) was used to calculate the correction term, which is applied to the original Glide docking score. (3) We then applied the Prime MM-GBSA rescoring protocol (Schrödinger Suite 2007). This procedure estimates the ligand binding free energy by performing minimization of the receptor–ligand complex, or the receptor and ligand alone. After refinement, a sorted list of the top compounds for each binding site was generated using the best-scoring Glide XP score of the multiple conformations, the strain-corrected XP score, and the MM-GBSA binding free energy, and three ranks were assigned to each compound, along with a consensus rank, calculated as the mean of the individual ranks. In summary, the refinement procedure first resampled ligand conformations and then provided the Glide XP score and two additional scores (strain-corrected Glide XP scores and MM-GBSA binding free energy estimates), as well as a consensus score for the selection of ligands for experimental verification.

For retrospective analysis, initial (nondocked) conformations of all 263 compounds tested in assays (regardless of activity) were collected and docked into both binding sites using Glide XP. Next, refinement and rescoring procedures as described above were applied to all resulting poses. To allow comparisons between the binding sites, all compound scores were converted into Z scores. The better-scoring pose of the two binding sites was then used in the generation of enrichment plots. With two exceptions, all the different scoring schemes employed in this process selected the same binding pocket as predicted in our previous docking attempts. We predicted compound **8** to dock into the SAM site, but as it does not form the H-bond to Val132, its pose in the RNA cap site was chosen. This bond was not required in the initial screening of the NCI library. For compound **31**, only the Prime MM-GBSA approach favored a pose in the SAM site over the RNA cap site, which was predicted by all other scores along with our previous predictions.

(v) **Pharmacophore Searches.** Pharmacophore generation and database searching were performed using Phase version 2.5 (Schrödinger LLC). Three different 3D pharmacophore hypotheses were generated on the basis of (A) a cluster consisting of five compounds [**10**, **13**, **17**, **20**, and **32** (Table 1)], including the confirmed hit compound **10**, (B) compound **10** alone, and (C) the cocrystallized MTase ligands ribavirin monophosphate and guanosine monophosphate. Pharmacophoric features were chosen so that they resembled the interactions between the ligands and the protein predicted by docking (A and B) or present in the cocrystallized structures³³ (C).

To allow flexible pharmacophore matching, a conformational search was performed on all compounds of the library of purchasable chemical compounds described above. For subsequent filtering of the full compound library, Phase default parameters were applied and all features defined in the pharmacophore hypotheses were required to match.

All compounds returned by the pharmacophore searches were subsequently docked to the RNA cap binding site and scored using Glide XP version 4.5 (Schrödinger LLC). The best-scoring pose for each compound was saved for further evaluation. As pharmacophore (A) resulted in a large number of hits, results were clustered by similarity, using MACCS structural keys fingerprints, a Tanimoto metric, and a degree of similarity of 60% using MOE 2007.09 (Chemical Computing Group, Montreal, QC). A diverse result set was obtained by picking the representative with the best GlideScore from each cluster.

Aggregation Prediction. (i) Test Sets. Three test sets were assembled to evaluate different predictors of compound aggregation. Briefly, the DenV test set consists of the molecules tested in this study, and the Med and AmpC test sets are taken from previously published studies on aggregation behavior.^{34–36} See the Supporting Information for details.

(ii) Decision Tree Aggregation Prediction. We assembled a decision tree similar to that which Seidler et al. derived using recursive partitioning.³⁷ See the Supporting Information for details.

(iii) Random Forest Modeling of Aggregation Behavior. We applied the random forest method³⁸ to model aggregation in a manner similar to the approach published by Feng et al.³⁵ See the Supporting Information for details.

Computational Infrastructure. While visual analysis of protein and small-molecule structures, as well as analysis of physico-chemical properties, was performed on standard Linux workstations, the preparation and filtering of the library for docking were performed on a Beowulf-type Linux cluster. Still, these resources were not sufficient to allow us to execute the planned large-scale molecular docking campaigns against the MTase. In the early stages of the project, we therefore built up a grid computing infrastructure at that time consisting of approximately 300 desktop PCs running the Windows 2000 or Windows XP operating system located in our institution's laboratories and classrooms, as well as some laboratory computers from another academic institution. These computers were tasked with docking or pharmacophore search jobs using Univa UD GridMP version 5.3. Schrödinger Suite 2007 supported this resource management system natively. As these types of computations are embarrassingly parallel, this resource's ready availability allowed us to screen large libraries for suitable compounds within reasonable amounts of time.

In Vitro Assays. (i) Methyltransferase Activity Assay. Unless otherwise stated, all compounds were first tested at a single maximum concentration of 25 or 100 μ M followed by IC₅₀ determinations with 2-fold serial dilutions starting from 25 or 100 μ M following a previously described protocol.²³ In brief, inhibitors were assayed in a 96-well white opaque plates (Corning Costar, Lowell, MA) in 50 mM Tris-HCl (pH 7.0), 10 mM KCl, 2 mM MgCl₂, 2 mM MnCl₂, 0.05% (v/v) CHAPS, 2 mM DTT, and 5 units of RNasin inhibitor (Promega, Madison, WI). Typically, 25 nM DENV2 MTase enzyme and 40 nM biotinylated RNA substrate were preincubated with the test compounds at room temperature for 20 min, and the reaction was initiated by addition of 0.56 μ M [*methyl*-³H]AdoMet (72 Ci/mmol) (Amersham Biosciences, Piscataway, NJ). Under these conditions, the inhibitory effect of the reaction end product is negligible, as the amount of SAH produced during the reaction time (5–10 nM SAH produced in 20 min) is too small to have a significant impact on enzyme activity, as shown in Figures 3 and 4 of Lim et al.²³ To detect aggregators, the assay was varied as follows. Detergent sensitivity experiments were performed via addition of 0.01 or 0.1% Triton X-100 to the reaction mix.^{39,40} For spin-down experiments, compound solutions were centrifuged for 15 min at 14000 rpm and room temperature before addition. For IC₅₀ shift assays, 8 or 80 nM DENV2 MTase was used to reach a 10-fold difference in enzyme concentration. All other conditions were kept the same.

Reactions were stopped with buffer containing 100 mM Tris-HCl (pH 7), 50 mM EDTA, 300 mM NaCl, 8 mg/mL streptavidin SPA beads (Amersham Biosciences), and 125 μ M cold *S*-adenosyl-L-methionine. Plates were read in a Trilux microbeta counter (Perkin-Elmer, Boston, MA) with a counting time of 1 min/well. All data points were measured in duplicate wells. IC₅₀ curves were plotted with average counts per minute against the log of compound concentration. The standard deviation was calculated by the nonbiased $n - 1$ method, where standard deviation = $\{[n\sum x^2 - (\sum x)^2]/[n(n - 1)]\}^{1/2}$. Nonlinear regression (curve fit) and the equation for the sigmoidal dose response (variable slope) from GraphPad Prism version 3.02 (GraphPad Prism, Inc., San Diego, CA) were used to interpolate values for IC₅₀. The equation is as follows:

$$Y = \text{bottom} + (\text{top} - \text{bottom}) / [1 + 10^{(\log \text{IC}_{50} - X) \times \text{Hill slope}}]$$

where X is the logarithm of concentration and Y is the response. Y starts at bottom and goes to top with a sigmoid shape. This is identical to the "four-parameter logistic equation".^{41,42}

(ii) **Cell-Based *Flavivirus* Immunodetection (CF-I) Assays and Cytotoxicity Assay.** The ability of compounds to inhibit dengue replication in a cell-based system (EC_{50}) as well as the cytotoxicity of test compounds (CC_{50}) was assayed as previously described.⁴³

Results

Structural Analysis of DENV2 Methyltransferase. Viruses are known to benefit from short generation times and fast evolution rates in escaping host defense reactions as well as in developing resistance against therapeutic molecules.^{44,45} Inhibitors must remain efficient against different serotypes and common mutations to qualify for further development. Figure 1A shows the conservation of individual MTase residues among peptide sequences from clinical isolates deposited in the UniProt database,²⁵ mapped onto the surface of DENV2 MTase in complex with RTP and SAH. For each residue, surface color is determined by the identity histogram value of an alignment of 127 unique DENV MTase sequences, shown in a spectrum from light gray (100% conservation) through yellow (95% conservation) to red (33% conservation). The Y-shaped central cavity is clearly visible, with the RNA cap located at the left-hand pocket (occupied by RTP in the structure shown) and SAM located to the right in place of SAH. As shown in Figure 1B, the RNA cap binding site is rather shallow and open. The aromatic ring of Phe25 undergoes π -stacking interaction with aromatic ring systems of the ligand. Hydrogen bonds can be formed from the ribose moiety to backbone oxygens of Asn18 and Lys14. The backbone oxygens of Leu20 and Leu17 likewise accept H-bonds, stabilizing the “front end” of ribavirin or the cap guanine. Moreover, electrostatic interactions between the phosphate groups of RTP and Lys29, Ser150, and Ser214 further stabilize ligand binding. The binding pocket of SAM (Figure 1C) is considerably more closed than the cap binding site. Important interactions are hydrogen bonds at both ends of the elongated SAM molecule, with Asp131 and Val132 fixing the adenine moiety and Gly86, Ser56, and Asp146 fixing the amino acid moiety at the opposite end. The elongated binding pocket is lined with predominantly apolar residues. Figure 1C also shows the catalytic tetrad comprised of Lys61, Asp146, Lys181, and Glu217 essential for RNA 2'-O-methylation.^{17,46}

Most current docking algorithms treat the protein as a rigid structure. It is therefore of great importance to investigate and account for possible structural rearrangements between the unliganded and liganded states. Figure 1D shows the variations present in the available X-ray structures of the dengue MTase with the average per-residue root-mean-square distance (rmsd) mapped onto the surface of dengue MTase in complex with RTP and SAH. The variability of each residue is encoded by a color gradient from light gray (0 Å rmsd) to yellow (1 Å rmsd) to red (≥ 2 Å rmsd). Overall, the structural variability is small and mainly located on the outer, solvent-exposed surface of the protein. The RNA cap binding site (Figure 1E) and the SAM binding site (Figure 1F) show very little structural variability.

The small number of mutations observed in the MTase active sites and the relative rigidity of the enzyme's active sites, as witnessed by structural comparisons, make the MTase binding sites evolutionary and structurally stable targets for rigid protein molecular docking approaches.

Validation of the Docking and Assay Pipeline. In light of the small number of known inhibitors of the MTase, we

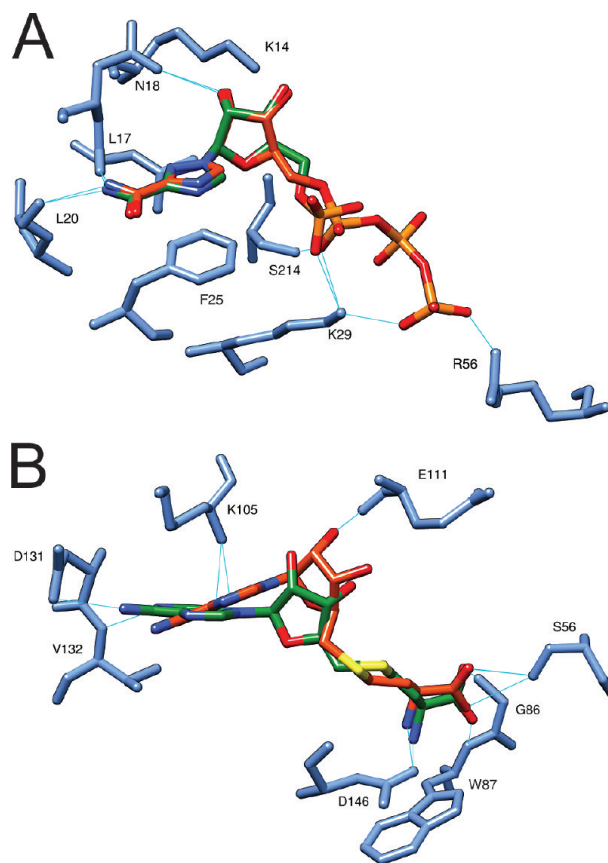


Figure 2. Redocking of RTP and SAH. (A) Crystallized pose of RTP shown with green carbons. The best redocked pose of RTP is shown with orange carbons. Surrounding residues undergoing important interactions are colored light blue. (B) Crystallized pose of SAH shown with green carbons. The best redocked pose of SAH is shown with orange carbons. Main interacting residues are colored light blue.

favoured structure-based over ligand-based computational approaches for the identification of promising inhibitory compounds. As a first step in validating our approach, we therefore redocked the cocrystallized ligands RTP and SAH into their respective binding pockets. These calculations were performed using Glide 4.5 in XP mode. The best-scored resulting poses were found to generally reproduce the binding mode of the crystallized ligands with heavy atom rmsd values of 0.74 and 1.33 for RTP and SAH, respectively (Figure 2).

Despite the fact that the RNA cap site is rather shallow, the redocked pose of RTP closely resembles that of the crystallized compound. Notably, the terminal nitrogen of Lys14 was designated as protonated in our protein preparation procedure, inducing a slight difference in the position of the ribose 3'-hydroxyl group with regard to the experimental structure. As the original data contain coordinates for only ribavirin monophosphate modeled into the $F_o - F_c$ and $2F_o - F_c$ electron density maps,¹⁹ the positions of the RTP β - and γ -phosphates were predicted, with the γ -phosphate stabilized by hydrogen bonds to Arg57 and Lys29. Redocking of SAH resulted in a pose similar to that of the SAH modeled into the electron density maps with a few differences. As a structural water molecule is absent, the sugar 2'-oxygen forms a direct H-bond to the backbone nitrogen of Glu111 rather than a water-mediated interaction as in the

experimental structure. This leads to a tilt in the plane of the adenine and a loss of the hydrogen bonds to Val132 and Asp131. Finally, the amino acid moiety of the SAH molecule is well anchored to Asp146, Trp87, Gly86, and Ser56 in a less strained conformation than that observed in the published structure. Comparable poses were also obtained for further known ligands of dengue MTase, GDPMP, SAM, and sinefungin (data not shown).

Our first effort was to screen a small set of 127000 compounds from the National Cancer Institute Developmental Therapeutics Program (NCI DTP).⁴⁷ Separate docking calculations were set up targeting either the RNA cap or SAM binding site. This study was conducted before the development of the protocol described in Materials and Methods and used a somewhat different procedure. Glide 4.5 was used in HTVS mode as a rapid first pass, mainly to eliminate compounds unlikely to fit to the binding pockets. Resulting poses were ranked by GlideScore, and the top 40% of hits against each binding site (79888 poses for the RNA cap site and 79946 poses for the SAM site) were extracted and redocked using Glide in SP mode. From the SP results, 100 compounds per site were chosen, on the basis of their ranked GlideScore and visual inspection for plausibility. Of this compound list, 40 were randomly selected for in vitro testing for dengue MTase inhibition. Of 36 compounds obtained from the NCI, nine compounds (Table 1, compounds 1–9) were found to be inhibitors of the MTase in vitro and will be further discussed below. With the workflow from computation to inhibition data in place, we now were ready to screen the large library of commercially available compounds previously compiled.

High-Throughput Docking. Both binding sites were next targeted in molecular docking campaigns using the whole compound library described above. During the whole procedure, the two binding sites were treated separately. For each site, compounds were docked using the funnel strategy described in Materials and Methods. After the first screening step, 1.01×10^6 compounds for the RNA cap site and 6.72×10^5 compounds for the SAM site were docked using the SP protocol; 1.09×10^5 compounds (RNA cap site) and 7.7×10^4 compounds (SAM site) were finally subjected to the XP procedure, and the 4000 top-ranking molecules from each binding site were selected for further refinement.

To prevent the lowest-energy pose finally retained from being influenced by artifacts of minimization on a 3D grid, each hit molecule was reconstructed and minimized under different conditions (solvent and force field). These additional input conformations were docked, and only the best-scored pose was retained for each compound (3392 compounds for the RNA cap site and 3365 compounds for the SAM site), discarding isomers along with suboptimal poses.

Candidate Selection and Inhibition Assays. Next, we calculated additional scores for the docked compounds. First, a correction term for ligand strain was applied to the GlideScore. Second, ligand binding energies were estimated using the MM-GBSA protocol in Prime version 1.6. The three scores were subsequently combined into a rank-based consensus score. Final short lists for each binding site were as follows: the 200 top compounds by consensus and the top 100 compounds by each of the individual scores. These overlapping criteria resulted in approximately 350 compounds per binding site, which were visually inspected and prioritized by a jury panel drawn from our institutions, involving four or five independent jurors. In this step,

different criteria were considered: diversity of chemical moieties covered by selection, credibility of pose based on experience in protein X-ray crystallography, similarity to known inhibitor poses, and presence of one or more key and additional intermolecular contacts with penalization of very close distance contacts. The consensus opinion of the jurors was used to produce a short list of 100 prioritized compounds for each binding site.

In total, 183 compounds could be obtained from vendors and were assayed for their ability to inhibit the transfer of a ³H-labeled methyl group from SAM to a short synthetic GTP-capped RNA oligonucleotide using a scintillation proximity assay.²³ Initial testing for MTase inhibition was conducted at a single concentration (25 or 100 μ M), and IC₅₀ concentrations were then determined for compounds showing substantial inhibition (>40%) in these experiments. Of the compounds tested, 23 were found to be inhibitors of DENV2 MTase with a spectrum of IC₅₀ values ranging from 2.62 to 37.46 μ M (Table 1, compounds 10–32).

Pharmacophore Screening. To retrieve further active compounds from the compound database, we next built a five-feature pharmacophore hypothesis from the predicted binding modes of compounds 10, 13, 17, 20, and 32 in the RNA cap site, reasoning that factors important for ligand binding may be inferred from the predicted binding poses of hit compounds. From a 3D superposition of these five compounds, a pharmacophore hypothesis was built using Phase version 2.5, consisting of pharmacophore features common to all compounds, resembling the protein–ligand interactions as predicted by docking. The hypothesis consists of five pharmacophoric features: two aromatic rings corresponding to (1) the diphenylamino ring stacking with Phe25, (2) the benzenesulfonate moiety, as well as hydrogen bond (3) donor and (4) acceptor features representing the diamino/amide moiety interacting with Ser150 and Ser214, and (5) a negatively charged group on the sulfonate moiety, interacting with Lys29 and Arg212 (Figure 3).

Searching through the compound database resulted in 4200 hits to the pharmacophore, which were subsequently docked into the RNA cap binding site and scored using Glide XP. To reduce the number of compounds but retain structural diversity, the 308 most diverse structures with the best GlideScore were selected on the basis of structural similarity. Resulting poses were visually evaluated by a jury as described above. Eighteen compounds were selected and subsequently tested in vitro as described above. Dose-dependent inhibition of compounds 33–35 was found to exhibit IC₅₀ values between 4.8 and 14.5 μ M (Table 1). Additional pharmacophore searches of the compound database were performed using six-feature models derived from either compound 10 only or the experimental structures of GMP and RTP in the binding pocket, retrieving 149 and 193 compounds, respectively. Of these, five and four compounds were selected for in vitro assays as described above, respectively, but no further inhibitors of MTase were found.

Testing Hit Compounds for Unspecific Inhibition. Of the compounds tested for inhibition, a large fraction exhibited a Hill coefficient substantially larger than unity. A large Hill coefficient signifies that a small increase in inhibitor concentration leads to an anomalously large change in inhibition, which can stem from positive ligand cooperativity, enzymes with equivalent binding sites, but also from non-ideal, nonspecific behavior that leads to abrupt enzyme inhibition above a critical concentration. One such mechanism

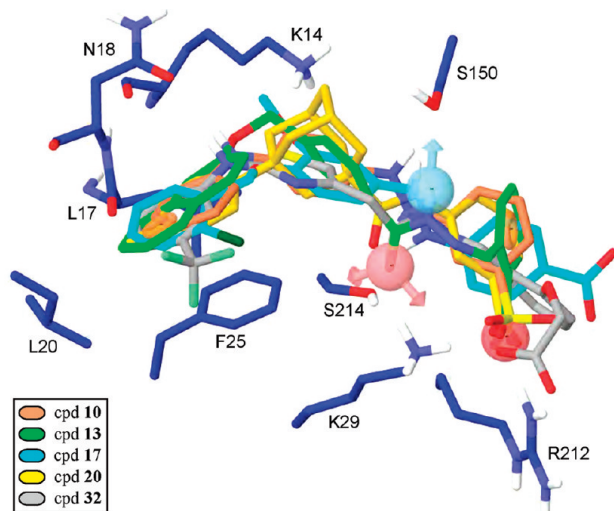


Figure 3. Pharmacophore query. Predicted binding modes of five compounds, obtained by docking to the RNA cap binding site and pharmacophore hypothesis created from the predicted binding modes. Ligands are shown as licorice sticks with colored carbons. Compound numbers in the legend refer to Table 1. Surrounding residues are labeled and shown with blue carbons. The pharmacophore hypothesis consists of five features: two aromatic rings (orange rings), one hydrogen bond acceptor and one donor (red and blue spheres with arrows, respectively), and one negatively charged group (red sphere).

is aggregation,⁴⁸ which has been recognized as a major cause of false positives in high-throughput screening.^{37,39,49,50} The addition of the nonionic surfactant Triton X-100 (0.1%) to the assay solution is often used to prevent the formation of compound aggregates without influencing enzymatic activity. While the addition of 0.1% Triton X-100 to our assay did not abolish inhibition by sinefungin, it had a marked effect on the majority of the previously chosen compounds.

Of the 35 compounds, 10 retained inhibitory activity ($IC_{50} < 100 \mu\text{M}$) in this assay (Table 1, compounds **2**, **8–10**, **14**, **25**, **28**, **29**, **31**, and **35**) (Figure 4), and **25** were rejected, exhibiting essentially flat dose–response curves in the presence of surfactant. The IC_{50} values of four of the actives changed ≤ 3 -fold compared to measurements without surfactant: $14.3 \mu\text{M}$ for **2**, $4.47 \mu\text{M}$ for **9**, $7.14 \mu\text{M}$ for **10**, and $4.91 \mu\text{M}$ for **35**. Compound **8** exhibited a 6.18-fold change with the addition of Triton X-100 but retained an IC_{50} value of $< 10 \mu\text{M}$. Compounds **14**, **25**, **28**, **29**, and **31** have IC_{50} values shifted more than 3-fold from the values obtained without Triton X-100 and do not saturate the MTase enzyme at the highest inhibitor concentrations (50 and $100 \mu\text{M}$) used in the assays (Figure 4). In Figure 5, the two-dimensional structures and predicted binding modes are shown for hits with IC_{50} values of $< 10 \mu\text{M}$. As these low-micromolar inhibitors are of particular interest, we performed two additional experiments to further rule out aggregation as the cause of inhibition.

First, compounds were assayed after centrifugation at 14000 rpm for 15 min (spin-down assay) to deplete the solution of any colloid particles. IC_{50} values obtained under these conditions are comparable to those obtained in previous assays with Triton X-100: $6.22 \mu\text{M}$ vs $9.46 \mu\text{M}$ (Cpd **8**), $10.52 \mu\text{M}$ vs $4.47 \mu\text{M}$ (Cpd **9**), $10.68 \mu\text{M}$ vs $7.14 \mu\text{M}$ (Cpd **10**), and $4.34 \mu\text{M}$ vs $4.91 \mu\text{M}$ (Cpd **35**). Second, IC_{50} values of the compounds were assayed in the presence of 8 or

80 nM MTase. Nonaggregating compounds should be insensitive to this shift in enzyme concentration. Compounds **8** and **10** exhibit relatively small changes in IC_{50} , which may be attributed to experimental variation: 2.3-fold for Cpd **8** [from $3.16 \mu\text{M}$ (8 nM) to $7.14 \mu\text{M}$ (80 nM)] and 1.3-fold for Cpd **10** [from $7.62 \mu\text{M}$ (8 nM) to $10.24 \mu\text{M}$ (80 nM)]. The two remaining compounds, **9** and **35**, exhibit larger shifts: 6.4-fold for Cpd **9** [from $3.27 \mu\text{M}$ (8 nM) to $20.82 \mu\text{M}$ (80 nM)] and 8.0-fold for Cpd **35** [from $1.32 \mu\text{M}$ (8 nM) to $10.6 \mu\text{M}$ (80 nM)].

All 10 active compounds were further assessed by the CF-I assay to examine their activities against dengue virus replication. The assay is based on quantitative immunodetection of dengue virus E protein production in target cells.⁴³ EC_{50} and CC_{50} values are reported in Table 1. Of the tested compounds, only compounds **14** ($EC_{50} = 12 \mu\text{M}$), **25** ($EC_{50} = 10.9 \mu\text{M}$), and **29** ($EC_{50} = 50 \mu\text{M}$) elicited a response in the cellular assays. Unfortunately, these compounds also exhibit cytotoxicity in the midmicromolar range (CC_{50} values of 22.7, 30.7, and $75.1 \mu\text{M}$, respectively).

Modeling of Aggregation Behavior. The unexpected elimination of many compounds that initially tested positive due to their action as aggregators prompted us to further investigate this nonspecific effect. Different computational approaches that are trained on experimental data to predict aggregation behavior of chemical compounds have been published.^{35,37} For our study, we were interested in how well similar predictors can be applied to different biological targets and assay conditions, as training and validation sets are usually measured under identical conditions. To assess the transferability of the proposed methods and evaluate them for our biological target, we have assembled three data sets of compounds with known aggregation properties in their respective assays: one based on the data from this work and two based on previously published large-scale aggregator detection assays.

We adapted the decision tree proposed by Seidler et al.³⁷ to molecular descriptors available to us and applied this predictor to the three test sets (Table S2 of the Supporting Information). Of the 182 compounds assayed from the large docking study, 23 initially appeared to be active in vitro. Of these, one compound was reliably not aggregating ($IC_{50} < 10 \mu\text{M}$ in the presence of detergent). In this data set, our decision tree classifier underestimated the overall aggregation tendency: 15 (65%) of these compounds were predicted to aggregate. Furthermore, compound **10**, the highest-affinity nonaggregator, was wrongly classified. Overall, the aggregation tendency was predicted correctly 61% of the time. Prior to in vitro testing of the 27 compounds selected from the pharmacophoric search, we ran the same aggregation predictor against this data set. Six (22%) of the compounds were predicted to aggregate. Despite the imperfections of the aggregation predictor, this value was sufficiently low, compared to the 65% prediction for the first-round actives, that we were confident that the second-round compounds would exhibit a significantly weaker aggregation tendency.

We also evaluated the use of a random forest classifier as described by Feng et al.³⁵ to predict aggregators in the three test sets, on the basis of calculated physicochemical properties. Validation within test sets yielded acceptable false positive (FP) and false negative (FN) rates (see the diagonal elements in Table S3 of the Supporting Information), comparable to results reported elsewhere.³⁵ When random forest

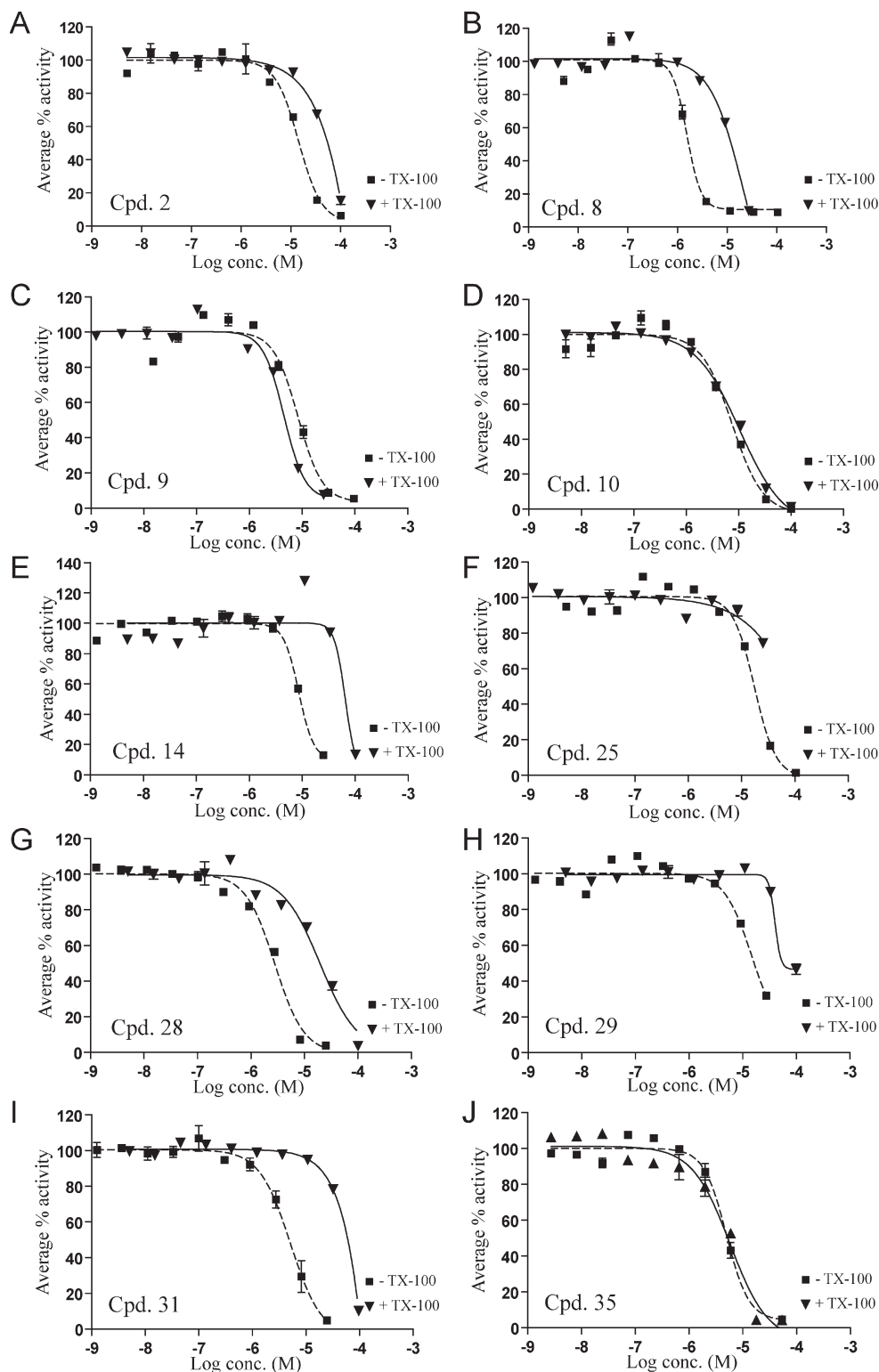


Figure 4. Dose-dependent inhibition of MTase activity. Recombinant NS5 MTase was preincubated with RNA cap analogue and varying concentrations of inhibitor. After addition of radiolabeled SAM, the transfer of the labeled methyl group to the RNA substrate was quantified using a scintillation proximity assay. Boxes and dashed lines show data for inhibition measured without addition of Triton X-100, and triangles and solid lines represent measurements in the presence of 0.1% Triton X-100. Measured counts per minute were normalized by dividing by the top curve value. (A–J) Inhibition of MTase by compounds **2**, **8–10**, **14**, **25**, **28**, **29**, **31**, and **35** (see Table 1).

models trained on the full data set for one assay condition were applied to data obtained with other biological targets or assay conditions (Table S3 of the Supporting Information), significantly larger false positive and false negative rates

resulted. In summary, we found both predictors to suffer from a weak ability to discriminate nonaggregating compounds (specificity), particularly when applied to conditions other than those on which they were trained.

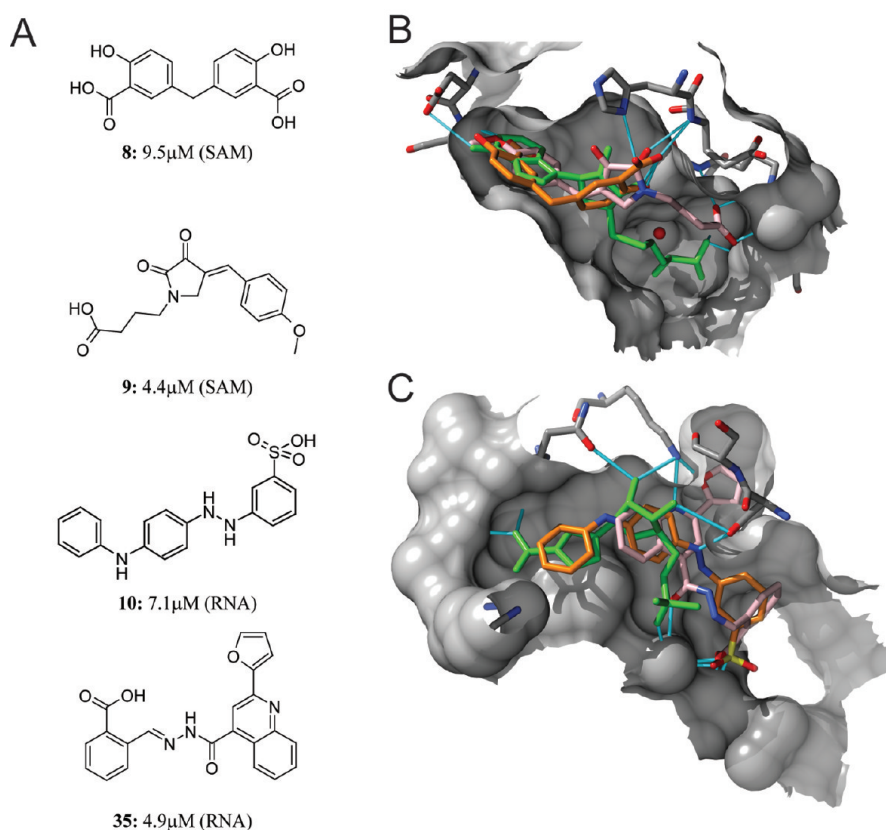


Figure 5. Structure and binding mode of DENV2 MTase inhibitors. (A) Structure of DENV2 MTase inhibitors with IC_{50} values of $< 10 \mu\text{M}$. Compound numbers refer to Table 1, followed by the IC_{50} value. The predicted target site is given in parentheses (SAM, SAM site; RNA, RNA cap site). (B) Crystallized pose of SAH shown as green licorice sticks. Compound **8** is shown with orange carbons, and compound **9** is shown with pink carbons. (C) Crystallized pose of RTP shown as green licorice sticks. Compounds **10** and **35** are shown with orange and pink carbons, respectively.

Of the 27 compounds selected for laboratory screening from our pharmacophoric searches, 23 came from models that included compound **10**, the best nonaggregating hit from the large-scale docking study. It is tempting to speculate that the reason this data set exhibited the lower fraction of aggregators among the apparent hits might be that the nonaggregator found in the earlier search was used to derive these models. It is not possible to draw this conclusion, given the limited number of compounds tested. However, the limited results we have obtained do suggest as a subject for future investigation that similarity to nonaggregating hits found in an early screen could be a useful criterion to incorporate into the development of a second-generation data set for a later screen.

Retrospective Analysis of Refinement and Rescoring Procedures. As the exact procedures for compound screening, refining, and rescoring evolved over the course of this work, we next performed a coherent retrospective analysis of the total set of 263 compounds tested for MTase inhibition in this work. In this analysis, the enrichment of compounds active in the inhibition assay (**2**, **8–10**, **14**, **28**, **29**, **31**, and **35**) was calculated at different steps of the procedure. All compounds were first passed through the final screening pipeline: (1) docking to the two binding sites using Glide XP, (2) refinement using resampled ligand conformations, and (3) generation of two additional scores (strain-corrected Glide XP scores and Prime MM-GBSA binding free energy estimates), as well as a consensus score. We next plotted enrichment of actives before and after the refinement

procedure (Figure 6A). Following the procedure applied in docking the large compound library, we next compared enrichment of actives achieved with each of the scores (Glide XP score of the refined pose, Glide XP score with internal strain correction, and Prime MM-GBSA binding free energy) and a consensus score, the rank-based average of individual scores (Figure 6B). As the number of actives in this data set is rather small, results from these analyses necessarily remain indicative, rather than conclusive. Nevertheless, our data suggest that the choice of rescoring scheme has an impact on the resulting hit list, whereas the impact of resampling ligand conformations is at least in our system less apparent.

Discussion

In this work, we describe the outcome of a combined computational and experimental study searching for novel inhibitors of dengue NS5 MTase among commercially available compounds. The viral methyltransferase possesses two binding sites that can be targeted in principle. However, the RNA cap site is rather shallow and solvent-exposed, so that molecules interacting firmly with this site are challenging to find. The second site binds the ubiquitous cofactor SAM, which invokes problems of specificity and off-target activity. For instance, the submicromolar inhibitor of dengue MTase sinefungin showed promise as an antibiotic, antiviral, and antiparasitic agent but was not further pursued because of its severe nephrotoxicity and lack of specificity.^{51–53}

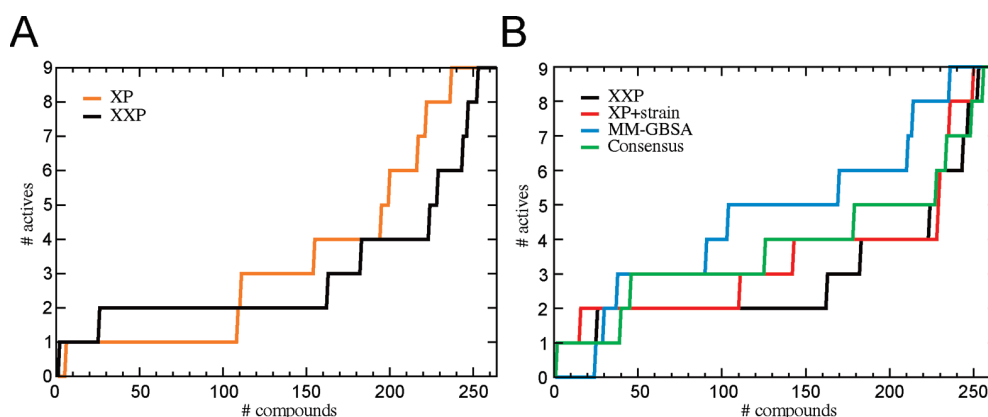


Figure 6. Enrichment of active vs inactive inhibitors. The number of actives recovered is plotted on the Y-axis against the size of the database screened on the X-axis. (A) Enrichment obtained with standard Glide XP docking (XP, orange) and with the best-scored Glide XP pose after increased sampling of starting conformations (XXP, black). (B) Enrichment obtained with individual scoring schemes and consensus score: Glide XP Score (XXP, black), Glide XP Score with internal strain correction (XP+strain, red), Prime MM-GBSA (MM-GBSA, blue), and rank-based consensus score (Consensus, green).

After an initial investigation of the target for suitability, we launched two virtual screening campaigns. Since little information was published about inhibitors of dengue MTase and their binding mode at the outset of the project, we pursued a broad-sweeping approach, using molecular docking to screen a library of several million compounds, rather than a small focused library. This approach was leveraged by the ready availability of computing resources through a computing grid that put more than 300 idle desktop PCs at our disposal.

The hits found must be considered starting points. The lack of an effect in cell-based assays observed with seven of the compounds may indicate cell permeation or stability problems, and the compounds (**14** and **29**) eliciting a response in these assays show an onset of cytotoxicity close to the half-maximal efficient dose. Nevertheless, these compounds inhibit the enzyme in *in vitro* assays, represent diverse chemotypes and predicted binding modes due to the study design, and therefore provide a number of inroads toward more focused approaches, which are currently under investigation.

Having at our disposal a set of 263 compounds characterized in inhibition assays, we retrospectively investigated the effect of decisive steps in our compound selection procedure on the resulting hit list. Notably, this set is a mixture of actives and very difficult decoys. All compounds initially docked well to one of the binding sites, were scored high, and were selected by a human panel as promising. Thus, absolute enrichment is not as interesting as are the differences between enrichments obtained by the various methods. In our system, increased sampling of starting conformations had no positive effect on the outcome, indicating that found poses and scores were not influenced by starting conformation biases. This is a positive finding. It indicates that the docking approach used here was not hampered by artifacts introduced by grid-based energy minimization. The comparison of different scoring procedures proves to be more interesting. In this case, where we try to discriminate between compounds scored closely together, Prime MM-GBSA binding free energies seem to provide the best enrichment of true actives for this target, closely followed by the consensus score obtained by the rank average of all three scoring methods used. Future virtual screening efforts against dengue or closely related MTases may benefit from this finding.

Why did our virtual screening result in the discovery of many aggregators? Since, in a virtual screen, there is no

opportunity for multiple copies of the ligand to interact with each other, there must be some other explanation for this observation. Most likely, the explanation lies in the fact that the binding sites studied here lend themselves to binding by long, flat molecules which, due to their shape and aromatic nature, have a strong tendency to aggregate. If this is correct, we would expect high-ranking virtual screening candidates obtained against binding sites more polar or more compact than those studied here to exhibit a much weaker tendency to aggregate. To put matters into perspective, the overall percentage of aggregators picked up in our study (25 of a total of 263 compounds tested, or 9.5%) is lower than that found in another study, where 19% of randomly chosen druglike compounds were acting as aggregators.³⁵ While docking does not select against aggregators, it does not select for them, either.

We tested two computational methods to predict compound aggregation to see if nonspecific binders can be identified before the *in vitro* stage. When a trained classifier is applied to a set of compounds assayed under different conditions, the level of misclassification is significantly increased in a manner independent of the training set used, making these predictions of limited practical use. As the tested models take compound properties but not the conditions of the assay into account, this is not surprising. Interestingly, the decision tree model proved to be more transferable than the random forest models.

As described earlier, we were confident, on the basis of the results of aggregation prediction, that the second-round compounds would exhibit a lower fraction of aggregates than the first-round actives. Laboratory results were in accord with this expectation. Of 27 compounds assayed, only two (7.4%) turned out to be aggregators. Of the three apparent hits, one (33%) turned out to be a nonaggregating hit, compared to one of 23 compounds (4.3%) from the large-scale docking screen. Apparent hits in the pharmacophoric screen were considerably more enriched in nonaggregators than the hits from the large-scale docking screen.

As the majority of compounds selected from the pharmacophoric searches came from models that included compound **10**, the best nonaggregating hit, it is tempting to speculate that similarity to nonaggregating hits found in an early screen could be a useful criterion to incorporate into the development

of a second-generation data set for a later screen, which would be a worthwhile topic for a future study.

In summary, our results, as well as the observations of others, do not suggest the broad use of computational methods in filtering out aggregators for two reasons. First, available methods lack accuracy and transferability between different assay conditions. Second, a number of active ligands and marketed drugs were found to be bona fide aggregators in inhibition assays³⁷ but pharmacologically active at concentrations lower than that at the onset of aggregation. This strongly suggests that the right way to control aggregation is by preventing it from interfering with the inhibition assay wherever possible. Aggregation prediction has a role, however, in the smart selection of limited follow-up compound sets from primary hits.

The cost of a screening campaign is a major concern, especially when neglected diseases are being targeted. We explored the extent to which in vitro experiments can be replaced by comparably inexpensive computational analysis. Roughly 5.5 million compounds were initially considered, and 263 molecules were assayed in the laboratory. Of these, 10 were initially characterized as inhibitors. When only compounds exhibiting IC₅₀ values of < 10 μM and a Hill coefficient of ≤ 2.5 were selected, four of the 10 compounds remain (**8–10** and **35**) as potentially interesting inhibitors. To further rule out aggregation as their mode of action, we applied two additional assays to these candidates. While the spin-down assay validated all four compounds, only two of the compounds (**8** and **10**) did not exhibit marked shifts in IC₅₀ when the enzyme concentration was changed 10-fold, strengthening the case for these two compounds in particular.

Compounds showing inhibition in the higher micromolar range (in particular, compounds **2** and **28**) are worth following up, as well, but may reveal themselves to be insufficiently specific in further tests. The future use of compounds **8–10** and **35** (Table 1 and Figure 6) should be considered in light of their properties. The object of this investigation was to find tool compounds and starting points for further optimization, using a set of commercially available compounds. While the compounds identified are not druglike in their properties and may not readily enter cells, they can serve as tools for cocrystallization or as starting points for scaffold hopping and bioisostere replacements. Interestingly, compound **8** exhibits a striking similarity to ATA, a low-micromolar inhibitor of DENV2 MTase found in a docking study published while this manuscript was being prepared.²⁴ Compound **9** is currently under further investigation in a program directed at identifying SAM competitive inhibitors.

In conclusion, the outcome of our study demonstrates that iterative combination of virtual screening and validation in the laboratory is a viable approach for the discovery of new hits against drug targets and can serve as a model for similar endeavors against other diseases. Computing power is becoming increasingly inexpensive or even free, as volunteers openly welcome requests for support of projects with a charitable aspect. Indeed, a number of grid-based drug discovery efforts have recently been launched.^{54–56} However, we strongly believe that the key to success is not access to virtually unlimited computing capacity, but rather establishing a tight interaction cycle between the computational and experimental parts of the project even before the first calculation takes place.

Acknowledgment. We acknowledge Jeremy R. Greenwood (Schrödinger LLC), Jürgen Kopp (formerly of University

of Basel), Eric Vangrevelinghe (Novartis Institutes for Biomedical Research), and Shahul Nilar (Novartis Institute of Tropical Diseases) for helpful discussions concerning methodology as well as valuable input in compound selection. We thank Hao Ying Xu and Boping Liu (Novartis Institute of Tropical Diseases) for their assistance in testing compounds in CF-I assays. Furthermore, we thank Alex Matter (formerly of Novartis Institute of Tropical Diseases), Manuel Peitsch (formerly of Novartis Institutes for Biomedical Research), and Jörg Weiser (Schrödinger LLC), who were instrumental in the origination of the collaboration.

Supporting Information Available: Implementation of decision tree and random forest aggregation predictors, and structures of tested compounds (Tables S1–S4). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Dengue and Dengue Hemorrhagic Fever: Information for Health Care Practitioners. <http://www.cdc.gov/NCIDOD/dvbid/dengue/dengue-hcp.htm> (accessed September 29, 2008).
- (2) *Dengue haemorrhagic fever: Diagnosis, treatment, prevention and control*, 2nd ed., World Health Organization: Geneva, 1997.
- (3) Morens, D. M.; Fauci, A. S. Dengue and hemorrhagic fever: A potential threat to public health in the United States. *JAMA, J. Am. Med. Assoc.* **2008**, *299* (2), 214–216.
- (4) Dengue and dengue haemorrhagic fever. <http://www.who.int/mediacentre/factsheets/fs117/en/> (accessed September 29, 2008).
- (5) Guzman, M. G.; Kouri, G. Dengue and dengue hemorrhagic fever in the Americas: Lessons and challenges. *J. Clin. Virol.* **2003**, *27* (1), 1–13.
- (6) Gould, E. A.; Solomon, T. Pathogenic flaviviruses. *Lancet* **2008**, *371* (9611), 500–509.
- (7) Whitehead, S. S.; Blaney, J. E.; Durbin, A. P.; Murphy, B. R. Prospects for a dengue virus vaccine. *Nat. Rev. Microbiol.* **2007**, *5* (7), 518–528.
- (8) Perera, R.; Kuhn, R. J. Structural proteomics of dengue virus. *Curr. Opin. Microbiol.* **2008**, *11* (4), 369–377.
- (9) Padmanabhan, R.; Mueller, N.; Reichert, E.; Yon, C.; Teramoto, T.; Kono, Y.; Takhampunya, R.; Ubol, S.; Pattabiraman, N.; Falgout, B.; Ganesh, V. K.; Murthy, K. Multiple enzyme activities of flavivirus proteins. *Novartis Found. Symp.* **2006**, *277*, 74–84, 251–253.
- (10) Mukhopadhyay, S.; Kuhn, R. J.; Rossmann, M. G. A structural perspective of the flavivirus life cycle. *Nat. Rev. Microbiol.* **2005**, *3* (1), 13–22.
- (11) Podvinec, M.; Schwede, T.; Peitsch, M. C. Docking for neglected diseases as community efforts. In *Computational Structural Biology: Methods and Applications*; Schwede, T., Peitsch, M. C., Eds.; World Scientific Publishing: Singapore, 2008; pp 683–704.
- (12) Dong, H.; Zhang, B.; Shi, P. Y. Flavivirus methyltransferase: A novel antiviral target. *Antiviral Res.* **2008**, *80* (1), 1–10.
- (13) Luzhkov, V. B.; Selisko, B.; Nordqvist, A.; Peyrane, F.; Decroly, E.; Alvarez, K.; Karlen, A.; Canard, B.; Qvist, J. Virtual screening and bioassay study of novel inhibitors for dengue virus mRNA cap (nucleoside-2′O)-methyltransferase. *Bioorg. Med. Chem.* **2007**, *15* (24), 7795–7802.
- (14) Cleaves, G. R.; Dubin, D. T. Methylation status of intracellular dengue type 2 40 S RNA. *Virology* **1979**, *96* (1), 159–165.
- (15) Furuichi, Y.; Shatkin, A. J. Viral and cellular mRNA capping: Past and prospects. *Adv. Virus Res.* **2000**, *55*, 135–184.
- (16) Ray, D.; Shah, A.; Tilgner, M.; Guo, Y.; Zhao, Y.; Dong, H.; Deas, T. S.; Zhou, Y.; Li, H.; Shi, P. Y. West Nile virus 5′-cap structure is formed by sequential guanine N-7 and ribose 2′-O methylations by nonstructural protein 5. *J. Virol.* **2006**, *80* (17), 8362–8370.
- (17) Eglhoff, M. P.; Decroly, E.; Malet, H.; Selisko, B.; Benarroch, D.; Ferron, F.; Canard, B. Structural and functional analysis of methylation and 5′-RNA sequence requirements of short capped RNAs by the methyltransferase domain of dengue virus NS5. *J. Mol. Biol.* **2007**, *372* (3), 723–736.
- (18) Eglhoff, M. P.; Benarroch, D.; Selisko, B.; Romette, J. L.; Canard, B. An RNA cap (nucleoside-2′-O)-methyltransferase in the flavivirus RNA polymerase NS5: Crystal structure and functional characterization. *EMBO J.* **2002**, *21* (11), 2757–2768.
- (19) Benarroch, D.; Eglhoff, M. P.; Mulard, L.; Guerreiro, C.; Romette, J. L.; Canard, B. A structural basis for the inhibition of the NS5

- dengue virus mRNA 2'-O-methyltransferase domain by ribavirin 5'-triphosphate. *J. Biol. Chem.* **2004**, *279* (34), 35638–35643.
- (20) PDB entries 1L9K, 2P1D, 1R6A, 2P3L, 2P3Q, 2P3O, 2P40, and 2P41.
- (21) Faumann, E.; Blumenthal, R.; Cheng, X. Structure and evolution of AdoMet-dependent methyltransferases. In *S-Adenosylmethionine-Dependent Methyltransferases: Structure and Functions*; Cheng, X., Blumenthal, R., Eds.; World Scientific Publishing: Singapore, 1999; pp 1–38.
- (22) Hodel, A. E.; Gershon, P. D.; Quijcho, F. A. Structural basis for sequence-nonspecific recognition of 5'-capped mRNA by a cap-modifying enzyme. *Mol. Cell* **1998**, *1* (3), 443–447.
- (23) Lim, S. P.; Wen, D.; Yap, T. L.; Yan, C. K.; Lescar, J.; Vasudevan, S. G. A scintillation proximity assay for dengue virus NS5 2'-O-methyltransferase: Kinetic and inhibition analyses. *Antiviral Res.* **2008**, *80*, 360–369.
- (24) Milani, M.; Mastrangelo, E.; Bollati, M.; Selisko, B.; Decroly, E.; Bouvet, M.; Canard, B.; Bolognesi, M. Flaviviral methyltransferase/RNA interaction: Structural basis for enzyme inhibition. *Antiviral Res.* **2009**, *83* (1), 28–34.
- (25) The universal protein resource (UniProt). *Nucleic Acids Res.* **2008**, *36* (Database issue), D190–D195.
- (26) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.
- (27) Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*; Wiley: New York, 2002; Chapter 2, Unit 2.3.
- (28) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera: A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.
- (29) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33.
- (30) Irwin, J. J.; Shoichet, B. K. ZINC: A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182.
- (31) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.
- (32) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5–6), 490–519.
- (33) PDB entries 1R6A and 2P1D.
- (34) Babaoglu, K.; Simeonov, A.; Irwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C. P.; Shoichet, B. K. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against β -lactamase. *J. Med. Chem.* **2008**, *51* (8), 2502–2511.
- (35) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **2005**, *1* (3), 146–148.
- (36) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **2007**, *50* (10), 2385–2390.
- (37) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* **2003**, *46* (21), 4477–4486.
- (38) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45* (1), 5–32.
- (39) Feng, B. Y.; Shoichet, B. K. A detergent-based assay for the detection of promiscuous inhibitors. *Nat. Protoc.* **2006**, *1* (2), 550–553.
- (40) Ryan, A. J.; Gray, N. M.; Lowe, P. N.; Chung, C. W. Effect of detergent on “promiscuous” inhibitors. *J. Med. Chem.* **2003**, *46* (16), 3448–3451.
- (41) Finney, D. J. Radioligand assay. *Biometrics* **1976**, *32* (4), 721–740.
- (42) Rodbard, D.; Hutt, D. M. *Statistical analysis of radioimmunoassays and immunoradiometric (labelled antibody) assays. A generalized weighted, iterative, least-squares method for logistic curve fitting*; 1974; pp 165–192.
- (43) Wang, Q. Y.; Patel, S. J.; Vangrevelinghe, E.; Xu, H. Y.; Rao, R.; Jaber, D.; Schul, W.; Gu, F.; Heudi, O.; Ma, N. L.; Poh, M. K.; Phong, W. Y.; Keller, T. H.; Jacoby, E.; Vasudevan, S. G. A small-molecule dengue virus entry inhibitor. *Antimicrob. Agents Chemother.* **2009**, *53* (5), 1823–1831.
- (44) Dunham, E. J.; Holmes, E. C. Inferring the timescale of dengue virus evolution under realistic models of DNA substitution. *J. Mol. Evol.* **2007**, *64* (6), 656–661.
- (45) Twiddy, S. S.; Holmes, E. C.; Rambaut, A. Inferring the rate and time-scale of dengue virus evolution. *Mol. Biol. Evol.* **2003**, *20* (1), 122–129.
- (46) Zhou, Y.; Ray, D.; Zhao, Y.; Dong, H.; Ren, S.; Li, Z.; Guo, Y.; Bernard, K. A.; Shi, P. Y.; Li, H. Structure and function of flavivirus NS5 methyltransferase. *J. Virol.* **2007**, *81* (8), 3891–3903.
- (47) DTP: 2D and 3D Structural Information. http://dtp.nci.nih.gov/docs/3d_database/Structural_information/structural_data.html (accessed October 5, 2008).
- (48) Copeland, R. A. *Evaluation of Enzyme Inhibitors in Drug Discovery: A Guide for Medicinal Chemists and Pharmacologists*; John Wiley & Sons: New York, 2005.
- (49) Shoichet, B. K. Screening in a spirit haunted world. *Drug Discovery Today* **2006**, *11* (13–14), 607–615.
- (50) Shoichet, B. K. Interpreting steep dose-response curves in early inhibitor discovery. *J. Med. Chem.* **2006**, *49* (25), 7274–7277.
- (51) Zweggarth, E.; Schillinger, D.; Kaufmann, W.; Rottcher, D. Evaluation of sinefungin for the treatment of *Trypanosoma (Nannomonas) congolense* infections in goats. *Trop. Med. Parasitol.* **1986**, *37* (3), 255–257.
- (52) Vedel, M.; Lawrence, F.; Robert-Gero, M.; Lederer, E. The antifungal antibiotic sinefungin as a very active inhibitor of methyltransferases and of the transformation of chick embryo fibroblasts by Rous sarcoma virus. *Biochem. Biophys. Res. Commun.* **1978**, *85* (1), 371–376.
- (53) Yebra, M. J.; Sanchez, J.; Martin, C. G.; Hardisson, C.; Barbes, C. The effect of sinefungin and synthetic analogues on RNA and DNA methyltransferases from *Streptomyces*. *J. Antibiot.* **1991**, *44* (10), 1141–1147.
- (54) Chang, M. W.; Lindstrom, W.; Olson, A. J.; Belew, R. K. Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries. *J. Chem. Inf. Model.* **2007**, *47* (3), 1258–1262.
- (55) Kasam, V.; Zimmermann, M.; Maass, A.; Schwichtenberg, H.; Wolf, A.; Jacq, N.; Breton, V.; Hofmann-Apitius, M. Design of new plasmepsin inhibitors: A virtual high throughput screening approach on the EGEE grid. *J. Chem. Inf. Model.* **2007**, *47* (3), 1818–1828.
- (56) Zhang, W.; Du, X.; Ma, F.; Zhang, J.; Shen, J. DGrid: Harness the Full Power of Supercomputing Systems. Fifth International Conference on Grid and Cooperative Computing Workshops (GCCW '06), **2006**.
- (57) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38* (3), 305–320.

2.2 Prediction of Non-Specific Inhibitors

2.2.1 Introduction

High-throughput (HTS) and virtual screening (VS) are commonly used techniques to discover novel hit compounds in the early phase of drug discovery. A major limitation of both methods is the fact that numerous false positive hits are obtained when compounds are tested experimentally. One contributing factor are nonspecific or 'promiscuous' inhibitors which affect the assay read-out but do not specifically inhibit the target protein.⁶⁸ One common mechanism for such promiscuous behaviour is the formation of colloidal aggregates of organic molecules which non-specifically inhibit enzymes.⁶⁹ Those compounds behave strangely with a weak structure-activity relationship, steep dose-response curves, poor selectivity and non-competitive inhibition.⁷⁰ To overcome this issue, numerous experimental techniques have been developed including high-throughput assays to detect detergent sensitivity or measure particle formation.^{71,72,73} In addition, different computational approaches that are trained on experimental data to predict aggregation behavior of chemical compounds have been published.^{71,69} Those methods are based on calculated physicochemical properties of the compounds and use supervised machine learning techniques to classify specific from non-specific inhibitors. They have been shown to predict aggregation with good accuracy within one dataset. It was unclear, however, how well such predictors can subsequently be transferred to different biological targets and assay conditions.

In our study to identify novel inhibitors for the dengue virus methyltransferase,⁷⁴ numerous compounds that initially tested positive were eliminated due to their action as aggregators as identified by an assay measuring detergent sensitivity. This prompted us to further investigate this nonspecific effect. The overall results of this study are summarized in Section 2.1 and detailed information about the employed method and the obtained results are given here.

For our study, we were interested in how well similar predictors can be applied to different biological targets and assay conditions, as training and validation sets published in the literature are usually measured under identical conditions.

2.2.2 Materials and Method

Test Sets

(A) *Dengue MTase test set (DenV)*: The 263 compounds tested during the screening for dengue MTase inhibitors described above were classified as 237 non-aggregators and 25 aggregators, based on detergent sensitivity in the inhibition assay described previously. Compounds losing their inhibitory activity in presence of 0.1% Triton-X 100 were classified as aggregators, whereas compounds either showing no inhibition or retaining their inhibitory activity in the presence of detergent were classified as non-aggregators.

(B) *Medium-size test set (Med)*: Data on the aggregation behavior of 1030 molecules has been published^{75,71,73} and the experimental results, based on dynamic light scattering and

a high-throughput detergent sensitive inhibition assay against AmpC β -lactamase, have been made available online (<http://shoichetlab.ucsf.edu>). From the 1030 molecules, all compounds showing ambiguous aggregation behavior were removed, leading to a set of 653 non-aggregators and 263 aggregators.

(C) *AmpC β -lactamase test set (AmpC)*: Recently, a set of 70563 molecules from the National Institutes of Health Chemical Genomics Center (NCGC) library was assayed in a high-throughput screen for detergent-dependent inhibition of AmpC β -lactamase.⁷³ Out of the 70563 molecules tested, 1204 were found to be unambiguously detergent-sensitive. From this dataset, obtained from <http://shoichetlab.ucsf.edu>, 402 non-aggregators and 82 aggregators were randomly picked as an additional test set.

Decision-tree based Aggregation Prediction

For a rapid attempt to predict aggregation behavior, we assembled a decision tree similar to that which Seidler, et al. derived using recursive partitioning.⁶⁹ Since we did not have access to machinery for generating the same set of descriptors, we employed descriptors that had similar meaning and used an iterative manual process to optimize the cut-off parameters for our own substitute descriptors to give the best agreement against the 111-compound training set supplied by Seidler et al. Table A.1 compares the descriptors and parameters we used with those of Seidler et al.

Random Forest based Aggregation Prediction

Compounds were predicted as aggregators or non-aggregators, based on calculated physicochemical descriptors, using a Random Forest (RF) model.⁷⁶ All compounds were prepared in their neutral form using the LigPrep protocol. For each compound, all 251 physicochemical descriptors available in MOE version 2007.09 (Chemical Computing Group, Montreal, Canada) were calculated. For each test set described above, 70% of the compounds were randomly selected to train a RF model, which was then tested on the remaining 30% of the data. To determine average false positive and negative rates, 100 iterations were performed per test set, each time producing 1000 unpruned trees from subsets of the 251 descriptors. To correct for the imbalanced dataset, the majority class was down-sampled during training of the RF model. For cross-validation between test sets, a RF model was trained on all compounds of a test set and then used to predict aggregators in the other two test sets. 100 iterations were carried out for each test set as described above. All calculations were performed using R version 2.5.16.⁷⁷

2.2.3 Results and Discussion

To assess the transferability of the proposed methods and evaluate them for our biological target, we have assembled three data sets of compounds with known aggregation properties in their respective assays. We evaluated both a decision tree based method adapted from Seidler

et al.⁶⁹ and a Random Forest based classifier as proposed by Feng et al.⁷¹ Both are described in Section 2.2.2.

Decision-tree based Aggregation Prediction

The decision tree proposed by Seidler et al.⁶⁹ was adapted to molecular descriptors available to us. Subsequently, this predictor was applied to the three test sets. Table 2.1 shows the results obtained on the training set of Seidler et al. using descriptors with the parameters shown in Section A.1 of the Appendix. With this decision tree, we achieved a prediction accuracy of 86.5%, which is not as good as the 93.4% accuracy reported by Seidler et al., but which was useful for our purposes. 43.2% of the compounds in the training set were aggregators; we predicted 42.3% aggregators, indicating that our false-positive and false-negative rates were approximately equal.

Table 2.1: Results of the decision-tree based prediction of aggregation behavior for three independent test sets.

	Training	Med	AmpC	DenV
FP	0.17	0.35	0.29	0.21
FN	0.11	0.03	0.11	0.06
Sensitivity: TP/(TP+FN)	0.88	0.95	0.87	0.93
Specificity: TN/(TN+FP)	0.84	0.74	0.75	0.82

Of the 182 compounds assayed from the large docking study, 23 initially appeared to be active in vitro. Of these, one compound was reliably not aggregating ($IC_{50} < 10 \mu M$ in the presence of detergent). In this data set, our decision tree classifier underestimated the overall aggregation tendency: 15 (65%) of these compounds were predicted to aggregate. Furthermore, compound 10, the highest affinity nonaggregator, was wrongly classified. Overall, the aggregation tendency was predicted correctly 61% of the time.

Random Forest based Aggregation Prediction

We also evaluated the use of a random forest classifier as described by Feng et al.⁷¹ to predict aggregators in the three test sets. Validation within test sets yielded acceptable false positive (FP) and false negative (FN) rates (see diagonal elements in Table 2.2), comparable to results reported elsewhere.⁷¹

However, when random forest models trained on the full data set for one assay condition were applied to data obtained with other biological targets or assay conditions (see off-diagonal elements in Table 2.2), significantly larger false positive and false negative rates resulted. In summary, we found both predictors to suffer from a weak ability to discriminate nonaggregating compounds (specificity), particularly when applied to conditions other than those on which they were trained.

Table 2.2: Results of Random Forest-based prediction of aggregation behavior for three test sets. Transferability of method is evaluated by performing all nine combinations of training and test set.

			test set					
			Med		AmpC		DenV	
			aver	stdev	aver	stdev	aver	stdev
training set	Med	FP	0.23	0.03	0.04	0.01	0.08	0.02
		FN	0.25	0.05	0.94	0.02	0.76	0.05
		Sensitivity: $TP/(TP + FN)$	0.76		0.50		0.55	
		Specificity: $TN/(TN + FP)$	0.77		0.56		0.76	
	AmpC	FP	0.88	0.01	0.42	0.05	0.95	0.01
		FN	0.03	0.01	0.39	0.11	0.06	0.04
		Sensitivity: $TP/(TP + FN)$	0.82		0.60		0.43	
		Specificity: $TN/(TN + FP)$	0.53		0.59		0.50	
	DenV	FP	0.59	0.03	0.39	0.02	0.38	0.06
		FN	0.09	0.01	0.62	0.04	0.37	0.22
		Sensitivity: $TP/(TP + FN)$	0.82		0.50		0.63	
		Specificity: $TN/(TN + FP)$	0.61		0.50		0.63	

2.2.4 Conclusion

To summarize, our results show that the accuracy of current computational approaches to predict compound aggregation is limited. In particular, both predictors suffer from a weak ability to discriminate nonaggregating compounds (specificity). In addition, the transferability of a predictor trained on one dataset to a different target set was investigated. The results suggest that prediction of aggregation behavior is not transferable between assay conditions or biological targets. Aggregation seems to be not only a property of the compound itself but rather depends on numerous other factors in the assay.

Thus, our results, as well as the observations of others, do not suggest the broad use of computational methods in filtering out aggregators but suggest to limit false-positive hits based on aggregation by directly preventing them from interfering with the inhibition assay. However, aggregation prediction might have a significant role in the clever selection of follow-up compounds from primary hits for which a particular classifier could be trained.

2.3 Experimental Characterization of Novel Inhibitors

2.3.1 Introduction

Based on a multi-stage virtual screening campaign for identifying inhibitors of the dengue methyltransferase, ten compounds were found to specifically inhibit dengue MTase with IC₅₀ values in the low μM range. In order to further characterize the identified inhibitors, a procedure for the production of purified recombinant dengue MTase was setup in-house, additional experimental assays were conducted and crystallization trials were performed on the five most promising compounds.

Compounds were followed-up using two inhibition assays to determine their inhibitory effect both on the 2'O and the N7 methylation reaction. In addition, an isothermal titration calorimetry based assay was developed in order to determine binding of the compounds to the MTase. Observing inhibition in two additional assays and determine compound binding could further help to validate their specific interaction with the dengue MTase.

2.3.2 Materials and Method

Protein Expression and Purification

Methyltransferase Plasmid The gateway vector pDEST14 which encode DENV2MTase Wild Type has been provided by Prof. Bruno Canard.¹

Transformation of *E. coli*

For plasmid isolation After mutagenesis PCR and DpnI digestion, 5 μl were used to transform 50 μl of competent *E. coli* TOP10 cells. Cells were thawed on ice, supplemented with plasmid DNA and incubated for 20 min on ice. Subsequently, cells were heat shocked for 90 sec at 42 °C in a thermomixer and incubated on ice for another 2 min. After addition of 250 μl of LB medium, the reaction mixture was incubated for 1 h at 37 °C with shaking and subsequently plated onto LB agar plates containing 34 $\mu\text{g/ml}$ chloramphenicol and 100 $\mu\text{g/ml}$ ampicillin for selection of positive transformants. Some transformed clones were chosen for plasmid isolation in order to sequence their insert and verify the nucleotide sequence.

LB Media: 10 g Bacto-Tryptone, 5 g Bacto-Yeast extract and 10 g NaCl in 900 ml H₂O. Adjust volume to 100 ml with H₂O and sterilize by autoclaving and store at RT.

For protein expression Prior to each protein expression experiments for preparative purposes, fresh transformants have been used to inoculate overnight cultures. Transformants have been generated as previously described (see for plasmid isolation part). Instead of *E. coli* TOP10 cells, Rosetta BL21(DE3) pLysS cells are used in this case.

¹Prof. Bruno Canard, AFMB, CNRS/University Aix-Marseille, 163 Avenue de Luminy, 13288 Marseille, France

Recombinant Protein Expression

Wild Type DENV2MTase Following transformation, a single colony was grown in 50 ml of LB medium at 37 °C overnight supplemented with the proper antibiotic. The culture was diluted 1/25 in the final volume of TB medium containing 34 $\mu\text{g/ml}$ chloramphenicol and 100 $\mu\text{g/ml}$ ampicillin and incubated with shaking at 37 °C until an O.D. of 0.5 was reached. After cooling down the cells to room temperature, expression was induced with 0.5 mM IPTG and incubated overnight at 25 °C. Cells were harvested by centrifugation at 5000g for 10 minutes at 4 °C. The bacterial pellet was resuspended in lysis buffer and stored at -20 °C.

TB Phosphate: Dissolve 2.31 g KH_2PO_4 (i.e. 0.17 M) and 12.54 g K_2HPO_4 (i.e. 0.72 M) in 90 ml H_2O . Adjust volume to 100 ml with H_2O and sterilize by autoclaving and store at RT.

TB Media: 12 g Bacto-Tryptone, 24 g Bacto-Yeast extract and 4 ml glycerol in 900 ml H_2O . Sterilize by autoclaving and cool to <60 °C. Add 100 ml of sterile 10 x TB phosphate and store at RT.

Lysis Buffer: 50 mM TrisHCl pH 8, 0.3 M NaCl, 5% Glycerol, 0.1% Triton X-100.

SDS-PAGE Electrophoresis To estimate protein solubility and expression level, SDS-PAGE electrophoresis have been performed. 12% acrylamide resolving gels were prepared and protein samples were mixed with 5 x SDS + 1 x DTT loading buffer and boiled at 95 °C for 5 min before loading. Electrophoresis run was carried out at room temperature applying 40 mA per gel. After electrophoresis, gels were stained with Coomassie solution and finally washed in Destaining solution. The same procedure was also used to evaluate the purity of protein fractions after purification.

Running Buffer: 25 mM TrisHCl pH 8.3, 0.2 M Glycine, 0.1% SDS

Coomassie solution: 100 ml acetic acid, 400 ml ethanol, 500 ml H_2O , 2.5 g Coomassie BB R250

Destaining solution: 50 ml acetic acid, 200 ml ethanol, 750 ml H_2O

Crude Extract Preparation

Cell disruption The thawed cell suspension was placed on ice, resuspended in Lysis Buffer with a homogenizer and complemented with a tablet of EDTA-free complete protease inhibitor cocktail (Roche) and 10 $\mu\text{g/ml}$ DNase prior to cell disruption. 2 rounds of microfluidizer at 12000 psi were performed to break the cells.

Ultracentrifugation Cells debris was removed by centrifugation at 35000 rpm for 30 min at 4 °C (Beckman, Ti70 rotor). The supernatant was sterile filtered (0.45 μ m, Millipore) before loading it on a Nickel-NTA affinity column.

Chromatographic Purification

The purification of DENV2MTase include two steps: Ni-NTA affinity and size exclusion chromatography. All chromatographic steps were performed at 4 °C using an ÄKTA purifier (GE Healthcare) and monitored at 280 nm.

Ni-NTA affinity chromatography The filtered crude extract was loaded on a HisTrap HP 5 ml column. The purification parameters are listed below:

Flow rate: 1 ml/min
Column Equilibration: 3 CV
Wash Out Unbound Samples: 3 CV

For WT MTase:

1st elution step: 6 CV of 15% Buffer B
2nd elution step: 15 CV of 45% Buffer B
Buffer A: 50 mM TrisHCl pH 8, 0.3 M NaCl, 20 mM imidazole
Buffer B: 50 mM TrisHCl pH 8, 0.3 M NaCl, 0.5 M imidazole

For Mutants:

1st elution step: 6 CV of 6% Buffer B supplemented with 10% glycerol
2nd gradient elution step: 21 CV of 6-60% of Buffer B supplemented with 10% glycerol
Buffer A: 50 mM TrisHCl pH 8, 0.5 M NaCl, 20 mM imidazole 10% glycerol
Buffer B: 50 mM TrisHCl pH 8, 0.5 M NaCl, 0.5 M imidazole 10% glycerol

15 ml fractions were collected and analyzed by SDS-PAGE. DENV2MTase containing fractions were pooled and concentrated using Vivaspin 20, MWCO 10k centrifugal concentrator (Sartorius) to a final volume of 15 ml for size exclusion chromatography.

Size exclusion chromatography Size exclusion chromatography was carried out with a preparative HiLoad 26/60 Superdex 75 prep grade column (GE Healthcare). 15 ml sample was injected to the column and eluted with 50 mM Bicine pH 7.5, 0.8 M NaCl, 10% glycerol, 1 mM DTT with a flow rate of 1 ml/min. 15 ml fractions were collected and analyzed by SDS-PAGE. Fractions containing pure DENV2MTase were pooled and concentrated using Vivaspin 20, MWCO 10k centrifugal concentrator (Sartorius) and depending on the following steps stored on ice, at 4 °C or at -20°C.

Isothermal Titration Calorimetry Experiments

Sample Preparation

Protein Buffer Exchange Isothermal titration calorimetry (ITC) is very sensitive and small buffer mismatches can generate large heats of dilution with each injection, which can mask heat changes from ligand binding. For this reason, protein buffer exchange has to be performed, the usual technique is dialysis. Unfortunately, this method could not be applied due to protein stability problems. Therefore, as an alternative, analytical SEC chromatography has been used with Superdex 75 10/300 GL column (GE Healthcare).

Due to solubility limitations of some ligands and stability problems of some mutants, different buffers have been tested for ITC experiments.

- 50 mM Bicine, 0.8 M NaCl, 10% Glycerol, 1 mM DTT, pH 7.5
- 20 mM Tris, 0.2 M NaCl, pH 7.5, 5 mM TECEP
- 20 mM Tris, 0.2 M NaCl, 10% Glycerol, pH 7.5, 5 mM TECEP
- 20 mM Tris, 0.2 M NaCl, 10% Glycerol, pH 7.5, 5 mM TECEP 10% DMSO

Ligand Preparation Selected inhibitors for ITC experiment are listed in Table 2.3. Ligands have been weighted and solubilized in the buffer used for protein size exclusion chromatography (see protein buffer exchange paragraph).

Table 2.3: Compounds selected for ITC, from the list of inhibitors identified by the screening approach described in Section 2.1.

^a IC₅₀ values are for the 2'O reaction.⁷⁴

Name	ID	MW (g/mol)	IC ₅₀ ^a (μ M)	predicted binding pocket	soluble
NSC14778	1	288	1.51	SAM	yes
ZINC02750651	8	566	2.81	RNA cap	no
ZINC02911543	7	377	7.56	RNA cap	yes
NSC140047	2	303	8.78	SAM	yes
ZINC03369470	9	365	4.80	RNA cap	no

To establish ITC experiment and to evaluate impact of mutations on the wild type DENV2MTase a number of ligands with known inhibitory effect have been used, as summarized in Table 2.4.

Table 2.4: Ligands used for establishing ITC assays and for evaluating effects on binding affinity of single point mutations.

Abbrev.	Name	MW	Spectroscopic Parameters	Binding Pocket
GTP γ S	Guanosine-5'-(γ -thio)-triphosphate, Lithium salt	539.24 g/mol	252 nm, 13700 cm ⁻¹ M ⁻¹	RNA cap
RTP	Ribavirin-5'-triphosphate, Sodium salt	484.14 g/mol	220 nm, 7700 cm ⁻¹ M ⁻¹	RNA cap
SAH	S-Adenosyl-L-homocysteine	384.40 g/mol	260 nm, 15400 cm ⁻¹ M ⁻¹	SAM

Protein and Ligand Quantification The concentration of proteins and binding ligands in solution was determined, according to the Lambert-Beer law, by measuring the absorption at the corresponding wavelength. For potential inhibitors, the molecular extinction coefficient is unknown. Therefore, the concentration was calculated from the weighted amount which have been weighted on a high-precision laboratory balance.

Cell Loading and Syringe Filling

The reference cell usually contains water or buffer and does not need to be changed after every experiment.

Prior to loading the sample cell and injection syringe, the protein and ligand solution are filtered (0.22 μm , Millipore), degassed to remove residual air bubbles and cooled down. The cell reactant, i.e. the protein solution in our case, is added to the sample cell of the calorimeter using a long needle glass syringe. 2 ml of protein solution is prepared to fill the cell which has a volume of 1.3-1.5 ml. Any excess solution remaining in the reservoir is removed.

The injection syringe is filled with approximately 300 μl of ligand solution and a purge/refill command is performed to push out potentially present air bubbles. 1 ml of solution is required to properly fill the injection syringe and to fully remove air bubbles.

ITC Measurement

All the measurements were performed at constant temperature of 22 °C. The first injection was set to a very small volume of 3 μL because heat disparities can appear due to a injection volume error arising from backlash in the drive screw mechanism of the syringe. To avoid this problem we also run a short down motion of the plunger to absorb the backlash from the up motion of the purge/refill command.

The chosen time interval between two consecutive injections was 200 sec in order to ensure that thermodynamic equilibrium was reached prior to the next injection.

Experimental parameters

Number of injections: 30

Run temperature: 22 °C

Reference power: 15 $\mu cal/s$

Initial delay: 200 s

Syringe solution: Ligand concentration depending on experiment

Cell solution: Protein concentration depending on experiment

Stirring speed: 300 rpm

Injection volume: 10 μL

Duration of injection: 20 s

Spacing: 200 s

Data analysis Data analysis was performed using an ITC analysis software based on ORIGIN 7.0. Blank experiments have been performed to validate the measured binding curves. The heat of dilution from the “end plateau” of the sigmoidal curve was used as reference data and the average of these values was subtracted from the entire sample data. After subtraction, thermodynamic parameters were determined with the “One Set of Sites” curve fitting model.

Inhibition Assays

2’O MTase Inhibition Assay Inhibition of 2’O MTase activity by inhibitor candidates was assayed in vitro based on an assay adopted from Luzhkov et al.⁵⁶

Briefly, 0.5 μM of purified recombinant dengue MTase in 50 mM Tris (pH 7.5) and 5 mM DTT was premixed with an inhibitor candidate at 20 or 100 μM final inhibitor concentration. The reaction was started by a premix of 5 μM [^3H]SAM (0.3 - 2 μCi) and 0.3 μM short capped RNA $^{7\text{Me}}\text{GpppAC}_5$ in the same buffer. The reaction was incubated at 30 °C for 3 h. To stop the reaction, 14 μl samples were spotted into 96 wellplates containing 100 μl of 20 μM ice cold SAH. Samples were transferred to glass-fiber filtermats and were washed twice with 0.01 M ammonium formate (pH 8.0) twice with water and once with ethanol. Scintillation fluid was added and RNA methylation was measured using a scintillation counter.

N7 MTase Inhibition Assay Inhibition of N7 MTase activity by inhibitor candidates was assayed in vitro based on an assay adopted from Milani et al.⁵⁷

0.5 μM of purified recombinant dengue MTase in 50 mM Tris (pH 7.0), 50 mM NaCl and 2 mM DTT was premixed with an inhibitor candidate at 16 or 80 μM final inhibitor concentration. The reaction was started by a premix of 80 μM [^3H]SAM (0.3 - 2 μCi) and 0.3 μM radiolabeled capped DENV₁₋₃₅₁ RNA in the same buffer. The reaction was incubated at 22 °C for 1 h. To stop the reaction, samples were heated to 70 °C for 5 min. Samples were transferred to glass-fiber filtermats and were washed twice with 0.01 M ammonium formate (pH 8.0) twice with water and once with ethanol. Scintillation fluid was added and RNA methylation was measured using a scintillation counter.

Protein Crystallization and Soaking

Crystals of the dengue MTase were grown by vapor diffusion at room temperature. 1 μl or 0.5 μl of protein solution at 18.5 mg/ml was mixed with 0.5 μl or reservoir solution and was allowed to equilibrate for two weeks in a hanging drop setup.

Different crystallization buffers were screened in 24 wellplates, consisting of ammonium sulfate (1.6 to 2.1 M or 1.9 to 2.4 M) and 0.15 M sodium citrate (pH 5.3 to 5.6). Crystals suitable for X-ray diffraction were obtained under the following conditions: 1.9 M ammonium sulfate, 0.15 M sodium citrate at pH 5.2 and 5.3 in a mixture of 1 μl protein and 0.5 μl crystallization buffer. Subsequently crystals were soaked for 1 h in 2.5 mM inhibitor solution containing 5% DMSO to enhance compound solubility.

All samples crystallized in P3₁21 symmetry, space group 152 with cell dimensions a=118, b=118, c=56 and angles=90°, 90°, 120°. Structures were solved by molecular replacement using the published structure 1R6A. Statistics are given in Table 2.5.

Table 2.5: Statistics of data collection of X-ray structures where the crystals were soaked with inhibitors.

Space group	P3121
Cell parameters (Å)	a = b = 117.739, c = 56.019
Resolution range (Å)	49.10 - 2.67
No. of total reflections	88020
No. of unique reflections	12761

2.3.3 Results and Discussion

Inhibition Assays

To further characterize the active compounds, to validate their specific interaction with the dengue MTase and to distinguish between inhibitory activity of the two catalyzed reactions, additional experimental assays were performed in a collaboration with Prof. Bruno Canard. To that end, inhibition assays for both the 2'O and the N7 reactions were further refined and selected inhibitor candidates were assayed. All results are given in Figure 2.1.

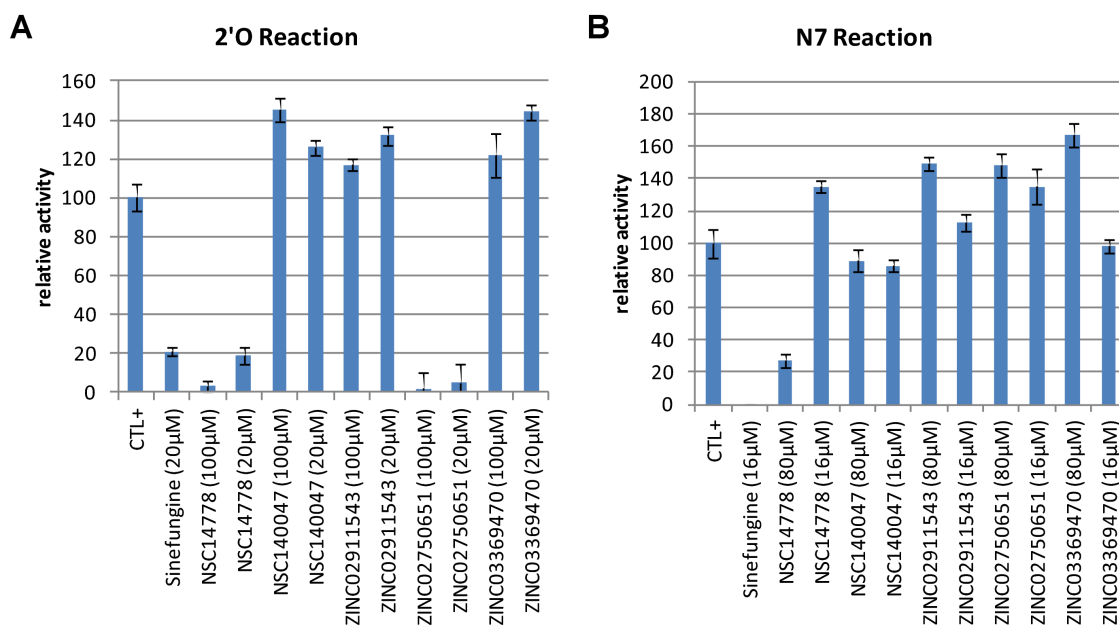


Figure 2.1: Inhibition of the dengue MTase 2'O reaction (A) and the N7 reaction (B) by selected compounds.

Compound NSC14778 inhibits the 2'O reaction at low concentrations (3% activity at 100 μM, 19% at 20 μM). In addition, at 80 μM in also inhibits the N7 reaction with a reduction of nativ methylation activity to 27%. At 16 μM concentration, no inhibition of the N7 reaction is observed. The fact that this compound inhibits both reaction is an indication that it binds

either to the SAM binding pocket or the RNA binding groove. This was correctly predicted by the original docking calculations which selected this compound based on docking to the SAM pocket.

Compound ZNC02750651 inhibits the 2'O reaction at low concentrations (2% activity at 100 μ M, 5% at 20 μ M). No inhibition of the N7 reaction is observed for this compound. This suggests that this compound binds to the RNA cap pocket which should influence only the 2'O reaction. Again, this was correctly predicted by the docking calculations which selected this compound based on docking to the RNA cap pocket.

Sinefungine, a known inhibitor of SAM dependent methyltransferases, was used as a positive control and shows strong inhibition at low concentrations with a reduction of native methylation activity to 20% and 0% for the 2'O and the N7 reaction.

All other compounds, do not show inhibition of either reaction in the assays present.

Crystallization

Obtaining X-ray crystal structures of inhibitors bound to the MTase would further our understanding of the important interactions governing protein-inhibitor binding and could validate predicted binding modes. This would be highly beneficial for future structure-based compound optimizations. Therefore, in a collaboration with Prof. Bruno Canard, we started efforts to obtain X-ray crystal structures of the inhibitors bound to the MTase.

Using the previously described purified protein of the dengue MTase domain, single crystals, suitable for X-ray diffraction were obtained using a hanging drop setup (Figure 2.2) and were subsequently used for soaking experiments with active compounds obtained by the previously described inhibitor screening efforts.

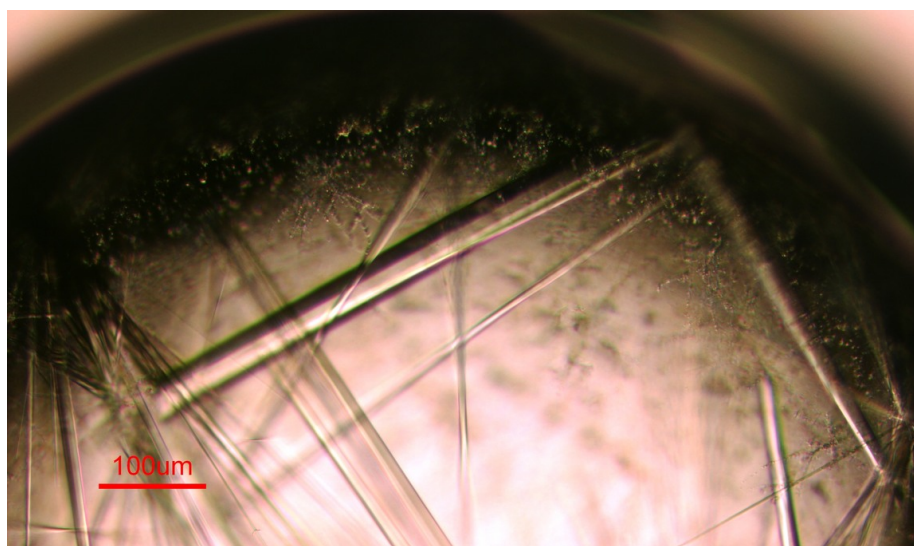


Figure 2.2: Crystals of the dengue MTase suitable for X-ray diffraction.

Structures were solved of crystals soaked with compounds NSC14778, NSC14047 and ZINC02911543 with resolutions from 2.5 to 3 Å. Samples containing the active compounds

aurintricarboxylic acid (ATA)⁵⁷ and sinefungine were of too low quality.

The structures of NSC14778 and NSC14047 show clearly absence of inhibitor molecules. In both structures, one molecule of SAH was observed in the SAM binding pocket. The structure of ZINC02911543 seems more promising with observed additional electron density. However, the actual data does not permit to say for sure that something else than SAH is bound to the MTase.

Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) based binding assays were performed for selected inhibitor candidates. For assay validation three compounds with known inhibitory effect were assayed, i.e. GTP γ S and ribavirin triphosphate, binding to the RNA cap pocket, as well as SAH, binding to the SAM pocket. For all compounds, a clear binding curve was observed.

Of the inhibitors, only three compounds (NSC14778, ZINC02911543, NSC140047) were soluble at the concentrations necessary for ITC measurements, whereas the two compounds ZINC02750651 and ZINC03369470 were insoluble. All ITC results are shown in Figure 2.3. Binding of the three assayed compounds could not be confirmed. A possible reason for this is discussed below.

Flaviviral MTases are often found to co-purify with one molecule of SAH or SAM bound to the protein. As neither of these co-factors are added during expression, purification or crystallization, SAH must originate from E.coli and co-purify with the methyltransferase.^{44,78} The fact that SAH stays bound to the protein during affinity and size exclusion chromatography purification steps indicate that SAH either binds very tightly to the protein or has a very slow off rate. In addition, Lim et al.⁵⁸ measured the content of bound SAH to the protein under assay conditions and found that 70% of MTase molecules were occupied by a SAH or SAM molecule.

Therefore, the ITC assay for compounds binding to the SAM pocket might be problematic since during a typical injection cycle (i.e. 3 min in our assay) no exchange between inhibitor and SAH is expected to occur. This is also indicated by the low stoichiometry (i.e. $N = 0.3$; expected $N = 1.0$) obtained for binding of SAH to wild type MTase.

Compound NSC14778 and NSC140047 are predicted to bind to the SAM pocket. In fact, for NSC14778 it was experimentally found that it inhibits both the N7 and the 2'O reaction, which indicates that it binds to the SAM pocket. Therefore, for those compounds, the ITC measurement might not be reliable.

Compound ZINC02750651 on the other hand, is the only inhibitor where inhibition assays indicated binding to the RNA cap pocket. This would be an excellent candidate for comparison, however, this compound is insoluble in the concentrations necessary for the ITC assay.

2.3.4 Conclusion

Of the ten active compounds identified through our virtual screening approach, five promising compounds were selected for further follow-up experimental assays. Of those, we were able to

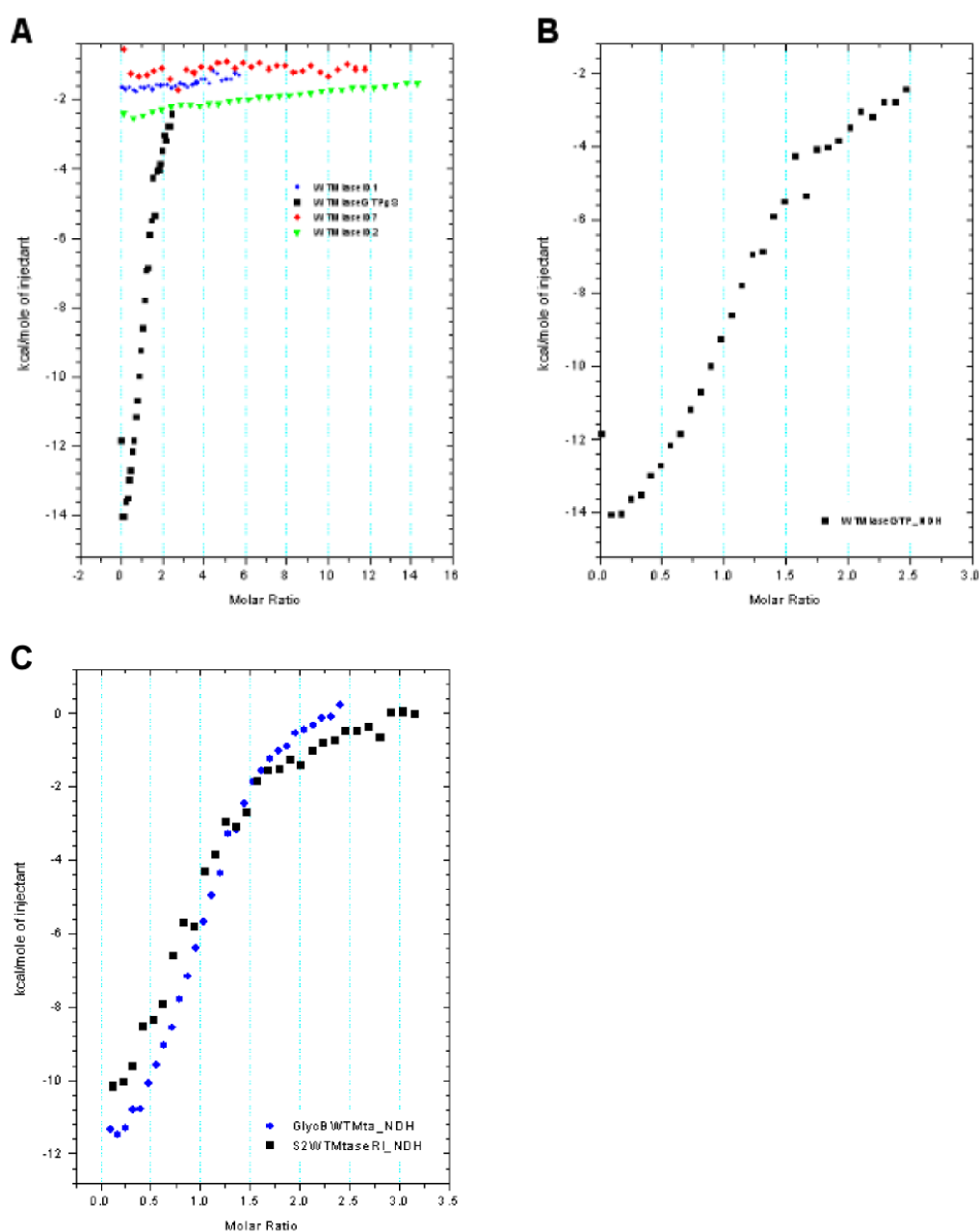


Figure 2.3: ITC measurements of inhibitors. (A) Three inhibitor candidates obtained from virtual screening efforts (red, blue, green) in comparison to GTP γ S (black). (B) GTP γ S individually to show the approximate sigmoidal curve shape. (C) Ribavarin triphosphate (blue) in comparison to GTP γ S (black) as positive controls.

confirm the inhibitory activity of the two most active compounds from our previous assays on the 2'O reaction. In addition, we were able to show that one of those compound inhibits also the N7 reaction.

The selected compounds were further used in crystal soaking experiments. For three of them, crystals suitable for X-ray diffraction were obtained. However, in two cases, the inhibitor candidate was not present in the structure whereas in the third case, additional electron density is observed, but the results do not permit to give a clear answer about the presence of an

inhibitor.

In addition, an assay based on isothermal titration calorimetry was developed to determine binding of a compound to the MTase. Of the selected compounds, three were assayed by ITC whereas the other two were insoluble under the necessary assay conditions. However, binding of the three assayed compounds could not be confirmed. One possible cause might be a pre-occupied SAH binding pocket which renders ITC based measurements unfeasible for compounds binding to the SAM pocket. To further quantify this effect, HPLC based determination of the SAH content and the quantification of the SAH exchange rate based on a pulse-chase experiment are necessary.

Chapter 3

Computational Analysis of the Methyltransferase Reaction

3.1 Introduction

Although flaviviral methyltransferases are attractive drug targets,^{61,79} little is known about the mechanism underlying their function. The enzyme is known to catalyze two distinct reactions on an RNA cap structure with a specific sequence,^{60,53,50} but so far the mechanism of the two methylation reactions at an atomistic level is not known. Thus, to obtain a better understanding of the molecular basis of this disease related enzymatic function, we aim at the elucidation of the underlying mechanisms using computational methods.

The enzyme is known to catalyze two distinct methylation reactions on an RNA cap structure with a specific sequence, but so far neither the structure nor the mechanism of the two methylation reactions are known at an atomistic level. Thus, we have modeled the protein in complex with the RNA for the 2'O reaction, based on available template structures and published mutagenesis data.

In order to characterize the underlying chemical reactions, high level ab initio electronic structure calculations were performed on model systems approximating the biological reactions. Reaction pathways and geometries were investigated and we found that the protein environment substantially lowers the reaction barrier, mediated by an active site lysine residue.

Based on the modeled complex structures, we applied a computational alanine scanning protocol employing molecular dynamics simulations and mixed QM/MM calculations in order to identify protein hot-spots and to further characterize the effect of mutations on different aspects of the system. We observed the influence of protein single point mutations on the geometric arrangement between methyl donor and acceptor, on the binding affinities of the SAM co-factor and on the reaction energetics.

Based on the results obtained by the computational alanine scanning procedure, previously uncharacterized protein residues were selected for further characterization using both computational and experimental methods.

Furthermore, to understand the RNA sequence specificity of the enzyme, we used the modeled structure in complex with different RNA cap structures of mutated sequences. Based

on molecular dynamics simulations, protein residues critical for RNA sequence specificity of the enzyme were identified and the possibility for forming an intramolecular hydrogen bond between distinct RNA elements was observed, whose absence might be detrimental for RNA sequence specificity.

3.2 Modeling of the Protein-RNA Complex

Since no structure of a flaviviral methyltransferase in complex with a short capped RNA is available, we have modeled the structure of the RNA bound DENV MTase, based on available template structures and published mutagenesis data, in order to investigate RNA binding, protein flexibility and enzymatic reaction mechanism.

3.2.1 Method

RNA Structure Modeling

For the 2'O-methylation, the VP39 MTase structure was used as a template. The structure of VP39 MTase (PDB code: 1av6) was superposed to the structure of DENV NS5MTase (PDB code: 2p41). The cap-guanosine and the first phosphate was taken from 2p41 and manually combined with the following two phosphates and the first three translated nucleotides obtained from 1av6. In an iterative manner, the obtained structure was manually adjusted and optimized in implicit solvent using the software MacroModel until a geometrically suitable structure was reached. This structure showed key protein-RNA interactions, similar to the interactions in 1av6, and no clashes between the protein and the RNA molecules were present. The nucleotides were replaced to yield the naturally occurring RNA sequence GpppAGU. The structure was solvated in a rectangular box of TIP3P waters. The structure was further optimized, first the solvent, then the whole system was heated to 300K and equilibrated for 300 ps before free molecular dynamics was performed. The structure stayed stable during 20 ns of MD simulation.

Ligand-Induced Structural Rearrangements

System Setup Structural rearrangements of the protein upon ligand binding was investigated using MD simulations. Two simulation systems were setup. First, the protein without ligands (*apo* simulation), second, the protein with SAM and GpppAGU RNA, modeled as described before (*holo* simulation).

The systems were solvated in a rectangular box of pre-equilibrated TIP3P waters and neutralized with chloride and potassium ions (0.15 M concentration). In the presence of the fixed solutes, the solvent was first minimized for 5000 steps of steepest descent (SD) minimization followed by an equilibration step of 300 ps at 298 K. Then the entire system was subjected to 5000 steps of SD minimization and equilibrated for 1 ns. Free molecular dynamics was then performed for 20 ns. All minimization and molecular dynamics runs were performed using NAMD (version 2.7b2)⁸⁰ with the CHARMM27 all-atom force field.^{81,82} MD simulations were performed at constant temperature (i.e. 298 K) and pressure (i.e. 1 bar) using particle-mesh Ewald (PME) electrostatics and a timestep of 1 fs.

For both systems, two individual free MD runs were carried out, using the same starting conformation, but different random seeds.

Analysis of Differences For each of the four MD runs, a frame was extracted every 100 ps, leading to 200 frames per trajectory.

When comparing two different runs, an all-against-all comparison was performed, where each frame extracted from the first trajectory was optimally superposed against all the frames from the second trajectory individually, leading to 40000 superpositions. For each superposition, a per-residue root mean square distance (RMSD) was calculated and the per-residue average over all superpositions was reported. All analysis was performed using the OpenStructure framework.⁸³

This comparison was performed within one set (e.g. both runs without ligands) and between the two sets (i.e. one run without ligands, one with ligand). To obtain ligand induced differences, the results obtained within the *apo* simulations was subtracted (or divided) from the results obtained between the *holo* and *apo* simulations.

3.2.2 Validation of the Structural Model

Since no structure of a flaviviral methyltransferase in complex with a short capped RNA is available, we have modeled the structure of the RNA bound DENV NS5MTase in order to investigate RNA binding, protein flexibility and enzymatic reaction mechanism, as described in section 3.2.1. The resulting conformation is shown in Figure 3.1 and structural stability is shown in Figure 3.2.

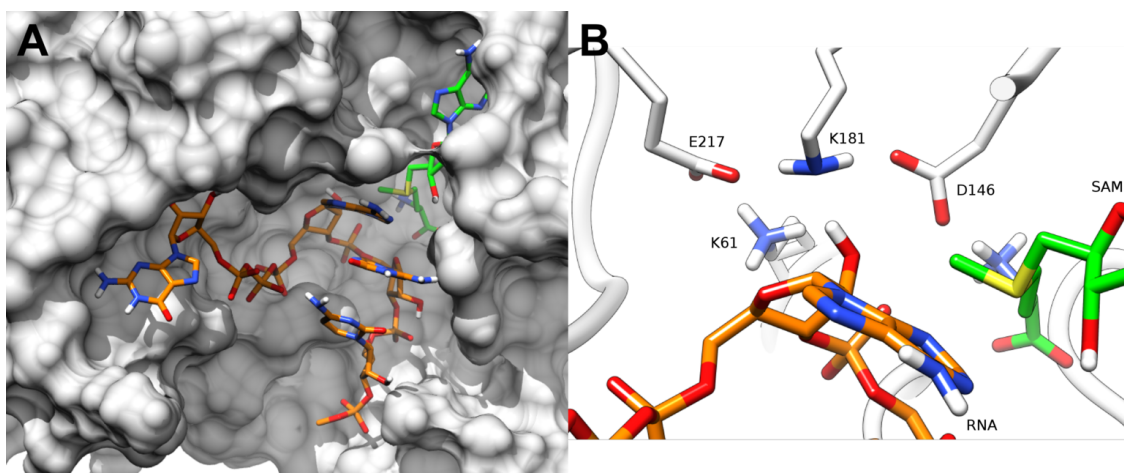


Figure 3.1: Modeled structure of RNA bound to dengue NS5MTase: (A) overall structure, (B) binding site. The RNA is shown in orange, SAM is shown in green.

For the N7 reaction, it has been shown that it can only take place on RNA templates comprising at least 74 nucleotides (nt) of the viral 5' UTR sequence.⁵⁰ However, the underlying reason is unclear. This RNA stretch is folded into a three-dimensional structure which might significantly influence the interaction with the MTase. A size comparison between the MTase and the first 74 nt of RNA is given in Figure 3.3. Therefore, it was not possible to obtain a structural model of a short RNA cap showing a conformation suitable for the N7 methyltransfer reaction.

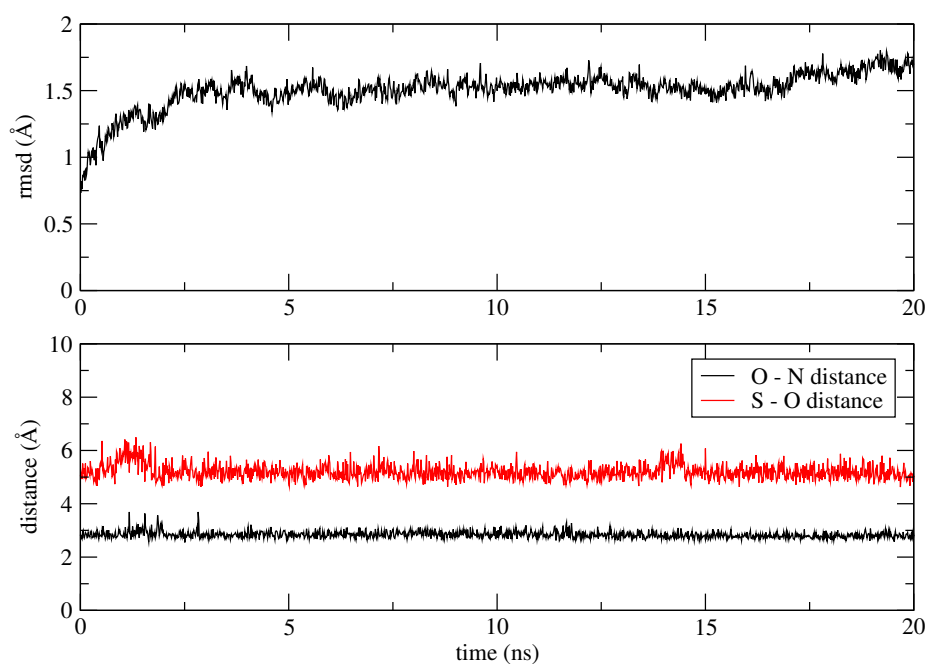


Figure 3.2: Stability of the modeled protein-RNA complex in MD simulations, measured by the rmsd of the complex (top panel) and the relevant distances (bottom panel) between methyl donor and acceptor (red) and between proton donor and acceptor (black).

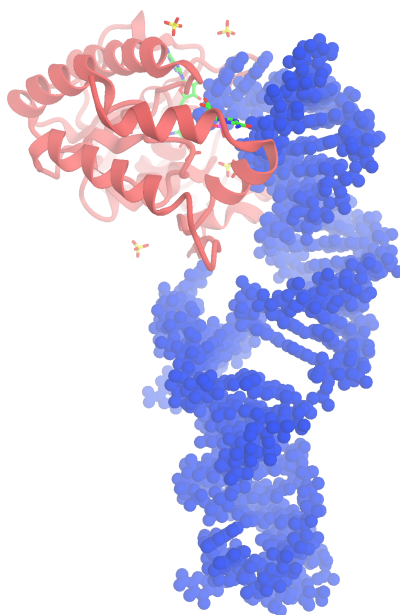


Figure 3.3: Schematic representation of a model of the complex between the DENV MTase (red) and the first 74 RNA nucleotides (blue) which are essential for the N7 reaction.

3.2.3 Ligand-Induced Structural Rearrangements

For analyzing structural rearrangements of the protein upon ligand binding, molecular dynamics simulations have been performed on two different systems: (A) the NS5MTase without any

ligands (*apo* system), (B) the NS5MTase in complex with SAM and RNA (*holo* system), modeled as described in Section 3.2.2.

Structural changes have been analyzed by comparing per-residue RMSDs between two MD simulations of the *apo* system with RMSDs between one MD simulation of the *holo* system and one of the *apo* system (see Section 3.2.1 for more detail). This comparison is shown in Figure 3.4 and the difference was mapped onto the protein structure as shown in Figure 3.5.

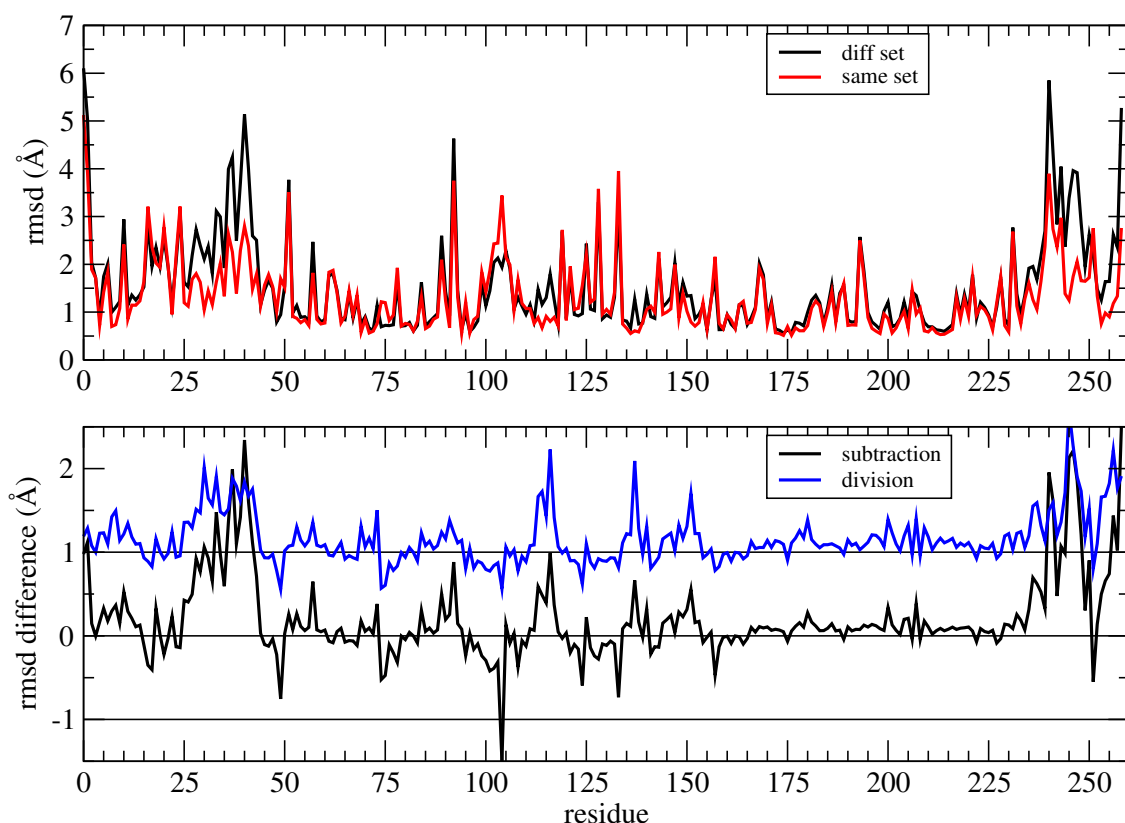


Figure 3.4: (A) Ligand induced structural changes, plotted as the per-residue RMSD difference between two sets (black: different sets (*apo/holo*), red: same sets (*apo/apo*)). (B) Subtraction (black) and division (blue) of the two curves in (A).

Overall, structural rearrangements are small, especially within the SAM and GTP binding pocket as well as in the active site. Most prominent, but still small structural changes are observed in helix A3 (residues 31 to 41) and the C-terminal loop structure (residues 239 to 249). Both substructures are interacting with each other, and the former is directly involved in binding of the 3' end of the RNA. Upon ligand binding, a slight movement away from the protein core is observed.

For the loop covering the SAM pocket (residues 99 to 106) it has been hypothesized that it could form a flap, which was present in an open and a closed form.⁶⁰ Although this loop shows higher flexibility, our calculations do not support this hypothesis, since no significant difference was observed between the *apo* and the *holo* simulations.

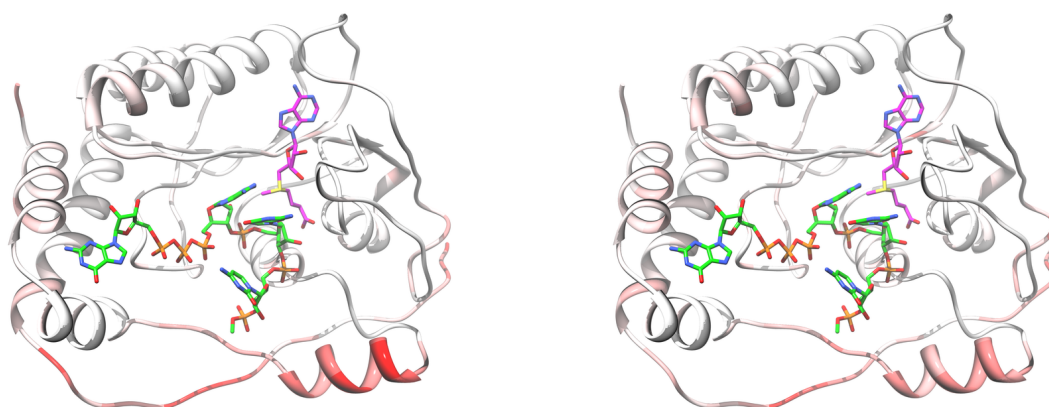


Figure 3.5: Ligand induced structural changes mapped onto the protein structure. (A) Rmsd difference (i.e. holo-apo) shown from white (difference: 0.0) to red (difference: 1.5). (B) Rmsd ratio (i.e. holo/apo) shown from white (ratio: 1.0) to red (ratio: 2.5).

3.3 Methylation of Guanosine N7 and Adenosine 2'O in Model Systems

In order to characterize the chemical reactions involved in the guanosine N7 and the adenosine 2'O methyl transfer reactions, ab initio electronic structure calculations were performed on model systems approximating the biological reactions.

Based on minimized structures of the reactants, the transition state and the products we have explored geometric arrangements, energetics of the reaction and transfer of charges both for the 2'O and the N7 reaction.

3.3.1 Method

Three different model systems were generated to study reaction energetics on simplified systems. One for the N7 reaction and two for the 2'O reaction where the proton acceptor was modeled either by a lysine like molecule or a water molecule.

Reaction energy profiles were calculated for the model system. All geometry optimizations and energy calculations were performed in Gaussian03⁸⁴ using the B3LYP⁸⁵ density functional theory method and the 6-311++G(d,p) basis set.⁸⁶ Solvation effects were included both during optimization and energy calculations based on the C-PCM implicit solvent model.⁸⁷ For all model systems, the reactant and the product complexes were fully optimized in implicit solvent first. Then, using the obtained structures, transition state optimization with QST3⁸⁸ was carried out. For all fully optimized structures, vibrational analyzes were performed on the same level of theory in order to confirm the nature of stationary points and to compute thermal corrections to the Gibbs free energy. From the obtained energies, a free energy profile for the reaction was created. In addition, point charges were computed based on natural bond orbital (NBO) analysis.⁸⁹

3.3.2 Geometry

For this study, model systems were used to study the methyl transfer reaction catalyzed by flaviviral methyltransferases using quantum chemical calculations. Two model systems were designed, one for the 2'O reaction and one for the N7 reaction.

The model for the 2'O reaction consists of the following three groups: (1) a model of the methyl donor SAM, which is truncated one carbon away from the reactive sulfur atom. (2) a model of the methyl acceptor ribose moiety, which is truncated two carbon atoms away from the reactive 2' hydroxy group. (3) A model of the side chain of Lys181, truncated one carbon atom away from the N ζ , which acts as a proton acceptor. The geometry optimized structures of this model are shown in the top panel of Figure 3.6.

The model of the N7 reaction consists of two groups: (1) the same model of the methyl donor SAM. (2) an N-Methylimidazole as a model of the methyl acceptor guanosine. For this reaction, no proton needs to be abstracted, thus, those two groups are sufficient to describe

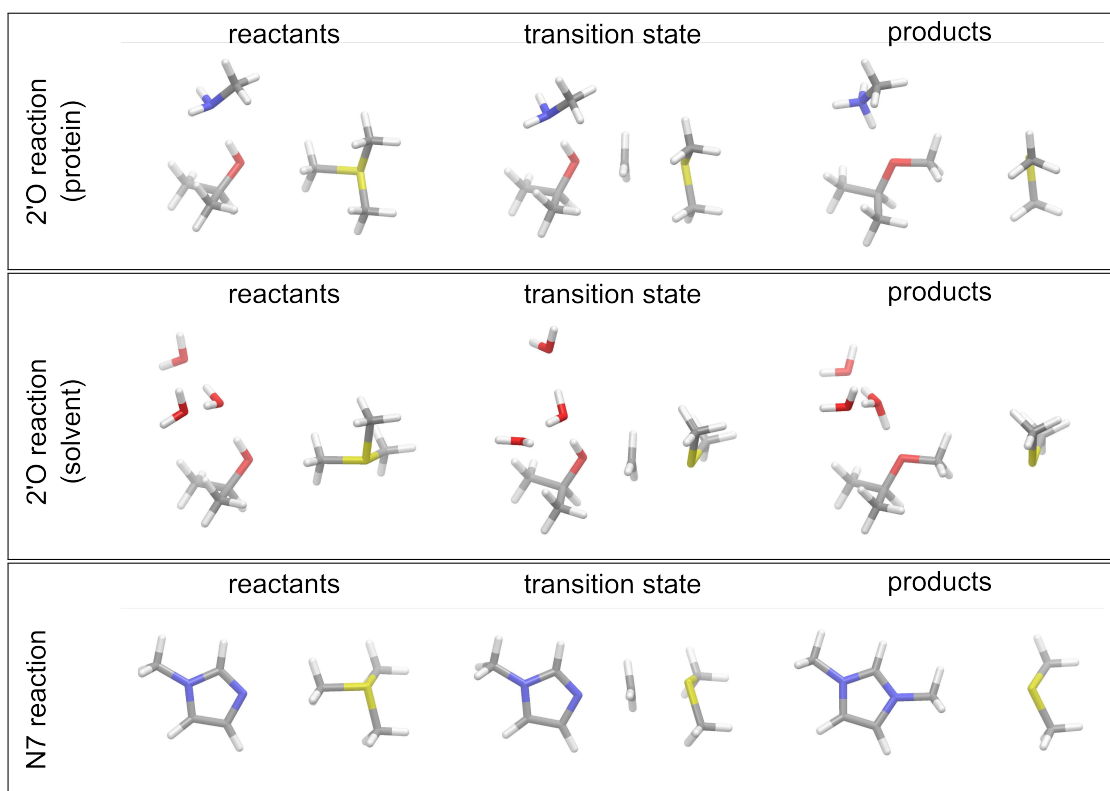


Figure 3.6: Optimized structures of reactants, products and transition states obtained by *ab initio* calculations for the three different model systems: (top) 2'O reaction with a NH_2CH_3 molecule to act as proton acceptor, (middle) 2'O reaction with a water molecule as proton acceptor, (bottom) N7 reaction

the reaction. The geometry optimized structures of this model are shown in the bottom panel of Figure 3.6.

In addition, a third model was designed to investigate the uncatalyzed 2'O reaction. This model consists of the same groups as the model for the catalyzed 2'O reaction, except that the proton acceptor moiety is replaced by a water molecule that accepts the transferred proton. To stabilize this water molecule, two additional waters were added, which are hydrogen bonded to the active water but do not directly participate in the reaction. The geometry optimized structures of this model are shown in the middle panel of Figure 3.6.

Figure 3.6 shows the optimized structures of all model systems for reactant state (left row), transition state (middle row) and product state (right row). In all reactions, the methyl groups are transferred in a $\text{S}_{\text{N}}2$ type nucleophilic substitution reaction, where the lone pair of the acceptor group attacks the positively charged methyl group. This reaction involves an inversion of the methyl group with a planar methyl conformation in the transition state. In all reactions, a nearly linear arrangement between the donor, the transferring methyl group and the acceptor is observed. Geometric parameters were extracted from the optimized structures as shown in Figure 3.7 with the values are given in Table 3.1 and Table 3.2 for the 2'O and the N7 model systems, respectively.

The distance between the methyl donor and acceptor atoms show similar behaviors for all reactions. Starting from the reactant complex, this distance is shortened in the transition state

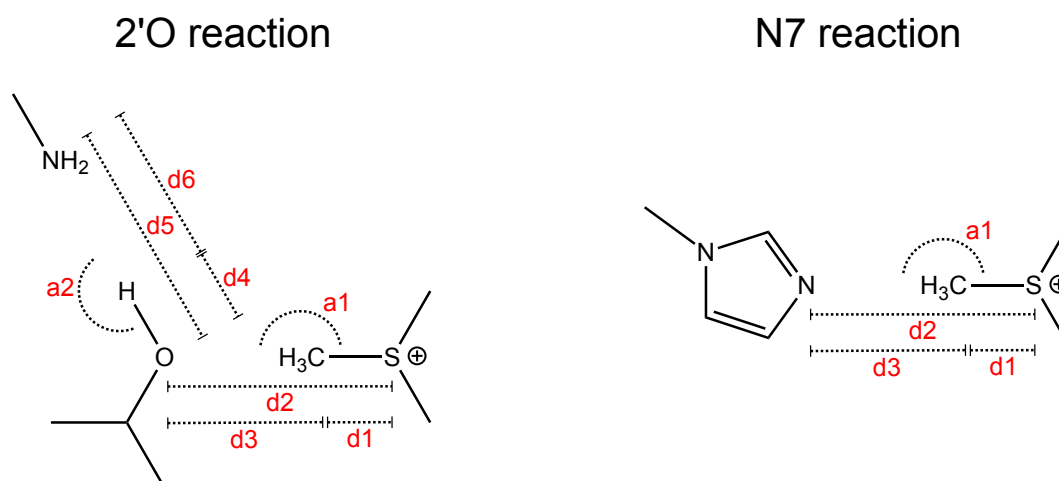


Figure 3.7: Schematic representation of the model systems for the 2'O and the N7 reaction. All reported distances and angles are indicated in read labels. For the 2'O reaction, only the system with a NH_2CH_3 molecule as proton acceptor is shown since all geometric parameters are directly transferable.

Table 3.1: Geometric parameters obtained from minimized structures for model systems of the 2'O reaction. Either NH_3CH_3 , mimicking Lys181 (left columns), or a water molecule (right columns) is present in the model system to act as a proton acceptor.

	NH_2CH_3 proton acceptor			H_2O proton acceptor		
	reactants complex	transition state	product complex	reactants complex	transition state	product complex
d1 (Å)	1.82	2.39	4.00	1.82	2.39	3.82
d2 (Å)	4.78	4.36	5.43	4.82	4.36	5.25
d3 (Å)	2.96	1.97	1.43	2.99	1.97	1.44
d4 (Å)	0.99	1.04	1.72	0.98	1.00	1.49
d5 (Å)	2.86	2.66	2.77	2.80	2.62	2.52
d6 (Å)	1.87	1.63	1.05	1.83	1.63	1.03
a1 (°)	176.9	177.0	176.2	177.7	177.1	175.7
a2 (°)	174.3	175.0	174.9	173.5	171.2	174.1

Table 3.2: Geometric parameters obtained from minimized structures for model systems of the N7 reaction.

	reactants complex	transition state	product complex
d1 (Å)	1.82	2.30	3.98
d2 (Å)	4.94	4.40	5.44
d3 (Å)	3.12	2.11	1.47
a1 (°)	177.4	180.0	176.6

by 0.54 Å and 0.42 Å for the N7 and the 2'O reaction, respectively. For the product state, a significant increase in the afore mentioned distance is observed compared to the reactant state. In the N7 reaction, it is increased by 0.5 Å and in the 2'O reaction by 0.65 Å. Thus, a significant compression of the structure in the transition state is observed which facilitates the reaction process by lowering the energy barrier (as discussed below).

For both models of the 2'O reactions, the methyl group is transferred onto a hydroxy moiety, yielding a methoxy group as the result of the reaction. Since a protonated methoxy moiety is unstable, the proton needs to be abstracted by a proton acceptor. In the protein environment, this is performed by the N ζ atom of Lys181 which was shown to be present in a deprotonated state.^{64,65} In the uncatalyzed reaction, however, this would be performed by a water molecule. The geometry of both model systems for the 2'O reaction are very similar for the reactant and the transition state. For the product state, however, significant differences can be observed. For the uncatalyzed reaction, the distance between the methyl donor and acceptor is decreased by 0.18 Å. In addition, the distance between the proton donor and acceptor is reduced by 0.25 Å which leads to a smaller distance between the proton donor and the transferred hydrogen atom, reduced by 0.23 Å. Thus, the transferred hydrogen atom is more strongly bound to the methoxy group which indicates that the water molecule is a worse proton acceptor compared to the model of the lysine. This effect is even stronger when the two additional water molecules are removed from the model system (i.e. leaving only the reactive water molecule). In that case, no proton transfer is observed yielding a protonated methoxy group in the product state (data not shown).

3.3.3 Energy Profiles

Figure 3.8 shows the Gibbs free energies calculated as described in Section 3.3.1 for all geometries of the optimized structures of all model systems for reactant (left), transition (middle) and product state (right). Table 3.3 summarizes all free energies.

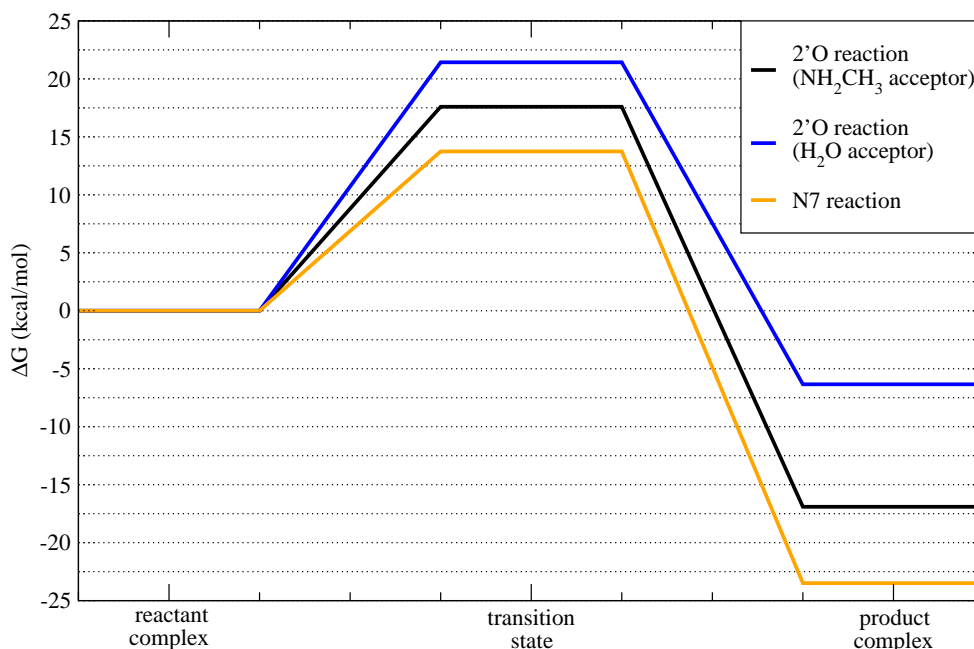


Figure 3.8: Gibbs free energy profile for reactants, product and transition state, obtained from *ab initio* model system calculations for the N7 (orange) and the 2'O methyltransfer reaction. For the 2'O reaction, either NH₃CH₃, mimicking Lys181 (black), or a water molecule (blue) is present in the model system to act as a proton acceptor.

Table 3.3: Summary of the Gibbs free energies for reactants, product and transition state, obtained from *ab initio* model system calculations for the N7 and the 2'O methyltransfer reaction. For the 2'O reaction, either NH_3CH_3 , mimicking Lys181, or a water molecule is present in the model system to act as a proton acceptor.

	2'O reaction (NH_2CH_3 proton acceptor)	2'O reaction (H_2O proton acceptor)	N7 reaction
reactant complex	0.00	0.00	0.00
transition state	17.59	21.43	13.74
product complex	-16.91	-6.34	-23.49

For both the 2'O as well as the N7 system, the calculations reveal that the reactions are exergonic processes where the product state is energetically significantly lowered compared to the reactant state with an energetically unfavorable transition state in between. The Gibbs free energy difference between product and reactant is -23.5 kcal/mol for the N7 reaction and -16.9 kcal/mol for the 2'O reaction. For the transition state, an energy barrier of 13.7 kcal/mol and 17.6 kcal/mol is observed for N7 and 2'O, respectively. From experimentally determined k_{cat} value,⁹⁰ the activation barrier of the 2'O reaction can be estimated using transition state theory. This yields an estimated activation barrier of 15.2 kcal/mol for the 2'O reaction. Thus, the calculated activation energy barriers are in a similar range indicating that the postulated $\text{S}_{\text{N}}2$ type methyl transfer reaction is energetically feasible.

Comparing the N7 to the catalyzed 2'O reaction, shows that the product state of the N7 reaction is lowered by 6.6 kcal/mol and the reaction barrier is reduced by 3.9 kcal/mol .

Comparing the catalyzed 2'O methyl transfer reaction to the uncatalyzed one, a significant reduction of the product state free energy can be observed when going from modeling the reaction in water to modeling the protein catalyzed reaction. The product state energy is lowered by 10.6 kcal/mol . In addition, the reaction energy barrier is lowered by 3.8 kcal/mol . This corresponds to a reaction rate enhancement of ~ 700 times.

In summary, for the 2'O reaction, the data suggests the importance of a lysine residue, representing Lys181 in the active site of the protein, which acts as a proton acceptor and significantly stabilizes the product state and reduces the activation barrier compared to the reaction in aqueous solution. The N7 methyl transfer reaction on the other hand, has a significantly lower energy barrier to overcome in aqueous solution and is thus more likely to occur without direct protein interactions. In conclusion, the data agrees well with the mechanistic hypothesis, where the 2'O reaction needs direct involvement of protein residues, whereas the N7 methylation does not need direct contact with the protein as long as the two reactants are in close proximity.⁶⁶

3.3.4 Energy Landscapes

Figure 3.9 shows the two dimensional potential energy landscapes computed at the B3LYP/6-311++G(d,p) level, for the model system of the N7 (left panel) and the 2'O reaction (right panel). Enthalpies of partially optimized systems are reported. They were computed on a grid

using two correlated reaction coordinates. First, the distance between the acceptor atom and the transferring methyl group (x-axis; distance $d3$ in Figure 3.7). Second, the distance between the acceptor and the donor atom (y-axis, distance $d2$ in Figure 3.7).

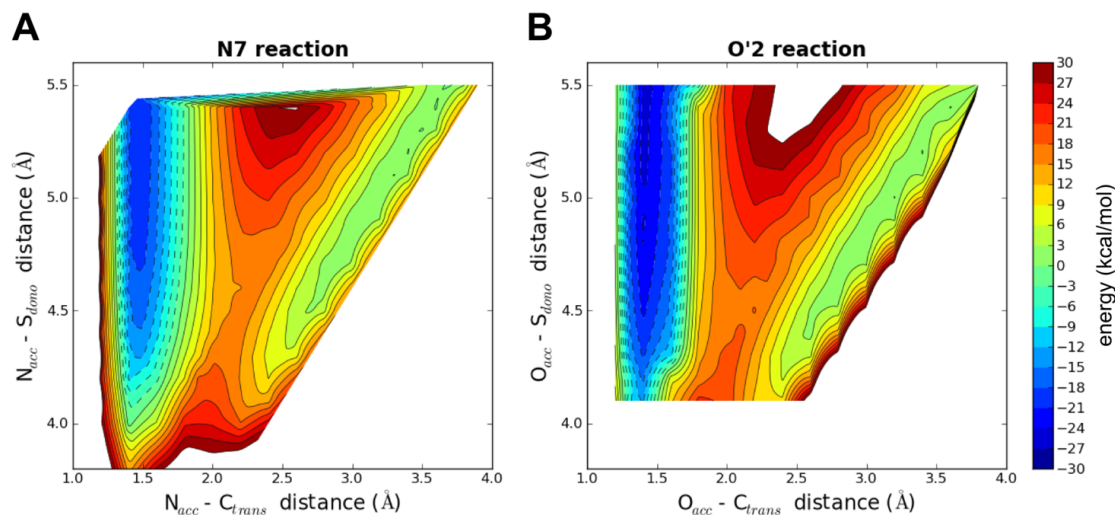


Figure 3.9: Two dimensional potential energy landscape for model systems of the N7 (left) and the 2'O reaction (right). Energies are shown for two reaction coordinates: (x-axis) distance between the acceptor atom and the transferring CH_3 group; (y-axis) distance between the acceptor and the donor atoms.

Although, the absolute values do not fully correspond to the afore mentioned Gibbs free energies obtained from fully optimized systems, they clearly indicate important features of the potential energy landscape. For both reactions, two distinct minima are shown. One corresponding to the reactant state with acceptor-methyl group distances above 2.5 \AA and an energetically more favored minimum corresponding to the product state with acceptor-methyl group distances around 1.5 \AA . The two minima are connected by a saddle point corresponding to the transition state. When the system is moved along its minimum energy pathway from the reactants through the transition state to the products, the system first compresses where the donor-acceptor distance (y-axis) reduces from $\sim 5 \text{ \AA}$ to $\sim 4.5 \text{ \AA}$ and then expands again to its product state with a donor-acceptor distance of $\sim 5.5 \text{ \AA}$. Thus, this compression significantly reduces the activation barrier compared to a direct movement without changing the donor-acceptor distance. This needs to be considered when evaluating effects on the reaction energy barrier caused by external perturbations, like single point mutations of the protein environment.

3.3.5 Point Charges

Partial charges obtained from natural population analysis⁸⁹ performed on the afore mentioned optimized model structures are shown in Figure 3.10. The charges are mapped onto the N7 and the 2'O systems using a color gradient from red ($-0.8e$) to green ($0.8e$). Partial charges for substructures of the model systems are summarized in Table 3.4.

The computed point charges clearly indicate that the methyl group is transferred as a cation with an overall change in the partial charges of the full SAM group of $0.98e$ for the N7 and

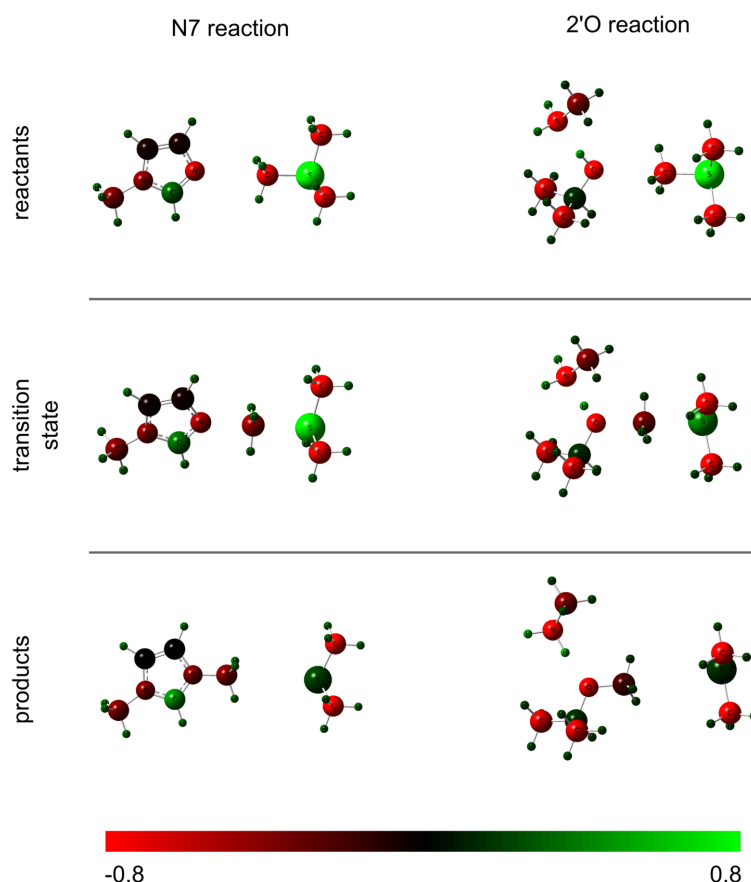


Figure 3.10: NBO point charges from *ab initio* calculations mapped onto the model systems for the N7 (left) and the 2'O reaction (right). The color gradient ranges from -0.8 (red) to $0.8e$ (green).

$0.99e$ for the 2'O reaction. The effect is more pronounced for the 2'O reaction, where the charge is localized more strongly on the transferred methyl group compared to the N7 reaction, yielding a higher charge on the methyl group both in the transition state and in the product. In comparison, for the N7 reaction, the charge is more distributed between donor, methyl group and acceptor, yielding a higher charge on SAM and the N-methylimidazol moiety in the transition state.

In the product state of the N7 reaction, the charge is located predominantly on the carbon atom at position 2 of the N-methylimidazol moiety with a point charge of $0.30e$, whereas for the 2'O reaction, the charge is located to a large extent on the proton which was transferred to the model lysine moiety, with a charge of $0.46e$.

3.3.6 Two Step Reaction

The catalytic lysine residue is envisioned to function as proton acceptor in two distinct ways: First, a two step mechanism involving the following steps: (1) deprotonation of the 2'-hydroxy group by the catalytic lysine residue prior to the methyl transfer reaction. This leads to the formation of a 2'-oxyanion. (2) transfer of the methyl group onto the 2'-oxyanion.

Table 3.4: Partial charges computed by natural bond orbital analysis for substructures of the model systems for the 2'O and the N7 reaction. The following substructures were used: (SAM) model of SAM without the reactive methyl group, thus consisting of CH_3SCH_3 , (CH_3) reactive methyl group, (ribose) model of the ribose moiety without the reactive methyl group, (lysine) model of the lysine side chain for the 2'O reaction, (guanine) model of the guanine acceptor for the N7 reaction.

	2'O reaction				N7 reaction		
	SAM	CH_3	ribose	lysine	SAM	CH_3	guanine
reactants	0.92	0.07	0.05	-0.05	0.90	0.08	0.01
TS	0.41	0.32	0.12	0.15	0.48	0.25	0.27
products	0.01	0.34	0.03	0.63	0.01	0.19	0.80

Second, a concerted mechanism where during the methyl transfer, the proton is transferred to the catalytic lysine. Thus, prior to the reaction, the proton acceptor does not deprotonate the 2'-hydroxy group but steers its orientation. NMR experiments indicate that the latter mechanism is used in the enzymatic reaction of vaccinia virus mRNA cap specific 2'O MTase VP39.⁶⁴

Geometry optimized model structures are in agreement with the latter mechanism, where the proton is bound to the ribose model compound both in the reactants and the transition state. The distance between the 2'-oxygen atom and the proton is 0.99 Å and 1.04 Å for reactants and transition state, respectively.

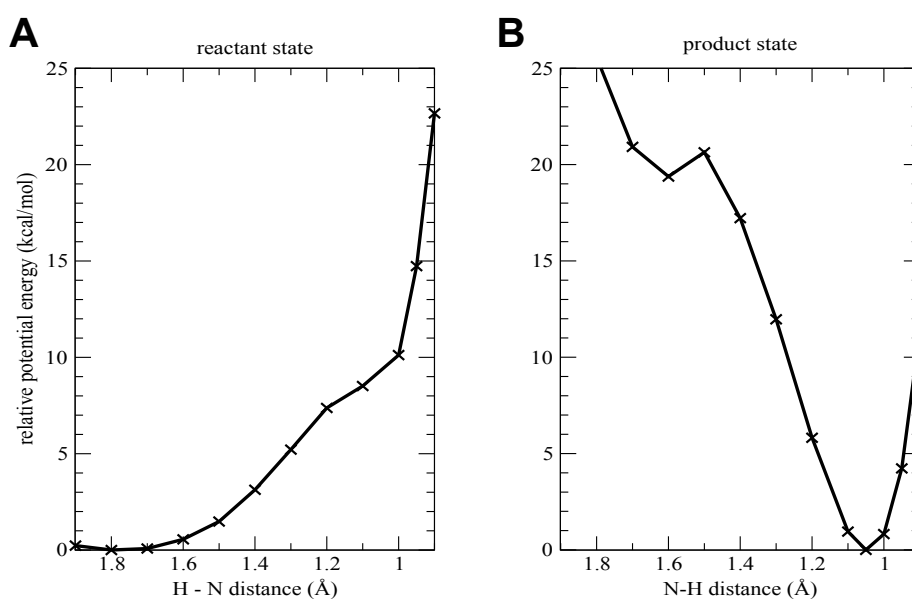


Figure 3.11: Potential energy curve for the proton transfer obtained by moving the proton from the ribose proton donor (H-N distance: 1.9 Å) to the lysine proton acceptor (H-N distance: 1.0 Å). (A) the system is in the reactant state (methyl group bound to the SAM donor). (B) the system is in the product state (methyl group is bound to ribose acceptor).

To confirm this, a linear scan was performed by moving the proton from the 2'O hydroxy group to the $\text{N}\zeta$ of the lysine model moiety. During the scan, the methyl group was kept in the reactant state and thus bound to the SAM model compound. Therefore, the endpoint of the scan yields a 2'-oxyanion. The energy profile is shown in Figure 3.11A. The calculations

estimate the costs of the proton transfer prior to the methyl transfer to ~ 10 kcal/mol. The profile increases monotonically and no significant minimum was observed which would stabilize the oxyanion conformation.

It should be noted that when the methyl group is moved to its product state, i.e. bound to the ribose moiety, minimization of the structure yields a spontaneous proton transfer with no observable barrier in between. This was confirmed by a linear scan of the proton position when the methyl group is in its product state (Figure 3.11B). This scan estimates the energy gain for the proton transfer in the product state to ~ 20 kcal/mol.

3.4 Impact of Single Point Mutations

Mutagenesis experiments can greatly help in understanding the mechanism of how an enzyme catalyzes a reaction by identifying the effect of single point mutations on different aspects of the system.

Although the effect on *methylation activity* of numerous single point mutations of DENV MTase were investigated experimentally, the underlying cause of a reduced methylation activity is still unknown since only relative *methylation activity* compared to wild type (WT) was measured. However, a reduced activity could have many different reasons, e.g. reduced substrate binding affinities, incorrect arrangement of the methyl-donor/acceptor pair, increased reaction energy barrier, misfolding of the protein.

Thus, we investigate the effect of mutations at an atomistic level using computer simulations. Therefore, numerous mutagenesis experiments were performed in-silico. To validate the simulations and to determine the contribution of each studied residue to the enzyme's function, different observables were computed from the simulations and compared to experimental measurements of methylation activity.

Based on a computational alanine scanning procedure, hot-spot residues were identified which significantly influence the reaction by modulating the geometric arrangement between methyl donor and acceptor, the methyl donor binding affinity or the reaction energy barrier. Selected hot-spot residues were further analyzed both computationally and experimentally in order to gain a better understanding of their role in the enzyme's function.

3.4.1 Materials and Methods

Plasmid and Primers

Mutants of the dengue MTase were produced based on the wild type plasmid described previously. Primers were designed according to the instructions from Stratagene manual and synthesized by Microsynth AG. The used primers are listed in Table 3.5.

Site Directed Mutagenesis

Mutants were prepared with Quick change II XL Site-Directed Mutagenesis (Stragagene), according to manufacturer instructions. 20 ng of plasmid DNA encoding Wild Type MTase were added to 50 μ l of a mixture containing 0.2 mM of dNTPs, 6% of DMSO, 1x PfuTurbo Buffer, 0.2 μ M of each primer and 2.5 U of PfuTurbo DNA polymerase. DNA was amplified with 18 cycles of PCR: denaturing at 95 °C for 50 sec, annealing at 55 °C for 50 sec, extension at 68 °C for 7 min. After mutagenesis PCR, 1 μ l of DpnI restriction enzyme was added to the amplification reaction and incubated for 1 h at 37 °C to digest non-mutated supercoiled dsDNA.

Table 3.5: Primers used for the production of DENV2MTase mutants.

Primer Name	T _m	5' - sequence - 3'
DENV2*Glu-79_F	74.7	GAA GGG AAA GTA GTG GAG CTC GGT TGC GGC
DENV2*Glu-79_R	82.4	GCC GCA ACC GAG CTC CAC TAC TTT CCC TTC
DENV2*Asn-79_F	73.8	GAA GGG AAA GTA GTG AAC CTC GGT TGC GGC
DENV2*Asn-79_R	80.9	GCC GCA ACC GAG GTT CAC TAC TTT CCC TTC
DENV2*Ser-79_F	74.7	GAA GGG AAA GTA GTG AGC CTC GGT TGC GGC
DENV2*Ser-79_R	82.4	GCC GCA ACC GAG GCT CAC TAC TTT CCC TTC
DENV2*Thr-88_F	71.2	GGC AGA GGA GGC TGG ACC TAC TAT TGT GGG
DENV2*Thr-88_R	75.6	CCC ACA ATA GTA GGT CCA GCC TCC TCT GCC
DENV2*Asn-88_F	69.6	GGC AGA GGA GGC TGG AAC TAC TAT TGT GGG
DENV2*Asn-88_R	74.0	CCC ACA ATA GTA GTT CCA GCC TCC TCT GCC
DENV2*Asp-217_F	68.0	CGA AAC TCC ACA CAT GAT ATG TAC TGG GTA TCC
DENV2*Asp-217_R	66.6	GGA TAC CCA GTA CAT ATC ATG TGT GGA GTT TCG
DENV2*Asn-217_F	72.2	CGA AAC TCC ACA CAT AAC ATG TAC TGG GTA TCC
DENV2*Asn-217_R	63.8	GGA TAC CCA GTA CAT GTT ATG TGT GGA GTT TCG
DENV2*Gln-217_F	73.4	CGA AAC TCC ACA CAT CAG ATG TAC TGG GTA TCC
DENV2*Gln-217_R	65.6	GGA TAC CCA GTA CAT CTG ATG TGT GGA GTT TCG
DENV2*Phe-219_F	67.8	CCA CAC ATG AGA TGT TCT GGG TAT CCA ATG CC
DENV2*Phe-219_R	67.8	GGC ATT GGA TAC CCA GAA CAT CTC ATG TGT GG
DENV2*His-219_F	71.4	CCA CAC ATG AGA TGC ACT GGG TAT CCA ATG CC
DENV2*His-219_R	65.8	GGC ATT GGA TAC CCA GTG CAT CTC ATG TGT GG

Protein Expression and Purification

Small scale expression and solubility screening of DENV2MTase Mutants 50 ml cultures of each mutant have been cultivated in the same conditions as the wild type (see above). 2 ml aliquots have been withdrawn before induction and after induction at following time points: 2 h, 4 h, 6 h and overnight. Extracted cell culture samples were centrifuged at 13000 rpm for 10 min and the supernatant was discarded. The remaining cell pellet was resuspended in 1 x Bugbuster + 200 $\mu\text{g/ml}$ lysosyme + 10 $\mu\text{g/ml}$ DNase solution, and incubated 20 min at room temperature with shaking. After centrifugation with 13000 rpm for 10 min, supernatants were collected and treated for SDS-PAGE analysis.

Protein Purification Mutated dengue MTase was expressed and purified as described in Section 2.3.2.

Isothermal Titration Calorimetry

Binding of S-adenosyl-L-methionine (SAH) and Guanosine-5'-(γ -thio)-triphosphate (GTP γ S) was determined using isothermal titration calorimetry as described in Section 2.3.2.

Computational Mutagenesis

Computational mutagenesis were performed using the following two distinct protocols.

Post-processing Protocol In the post-processing protocol, only simulations of the wild type MTase are performed. From a molecular dynamics trajectory of the unmutated WT protein-

ligand complex solvated in a cubic box of TIP3P water molecules, snapshots are extracted every 8 *ps* to obtain a representative ensemble of complex structures. For each snapshot the protein residue under consideration is mutated as follows. First, all side chain atoms which are not common in the WT and the mutated residue are removed. Second, all missing atoms of the mutated residue are added based on standard internal coordinates as implemented in CHARMM. Third, coordinates of the mutated residue are optimized in vacuum for 150 steps of steepest descent minimization. The structure obtained like this is used for further analysis.

Full MD Protocol In the full MD protocol, MD simulations both of the WT and the mutated protein-ligand complex are performed. Starting from the same initial structure as for WT MTase, all atoms not belonging to the protein, the RNA or the SAM molecule are removed. The protein residue under consideration is mutated using the same procedure as described in the post-processing protocol. The structure obtained after energy minimization, is subsequently used as the input structure for MD simulations using the same protocol as for WT MTase.

Molecular Dynamics Simulation

Molecular dynamics simulations are performed on the native complex as well as on single point mutants thereof, containing the dengue MTase, the capped RNA fragment of the form Gpp-pAGU and the methyl donor SAM.

For MD simulations, the initial starting conformation of the WT complex was obtained after MD based equilibration of the modeled complex structure as described in Section 3.2.1. For consistency, all atoms not belonging to the protein, the RNA or the SAM molecule are removed. Using the same structure, mutations were introduced as described before.

The systems were solvated in a rectangular box of pre-equilibrated TIP3P⁹¹ waters and neutralized with chloride and potassium ions (0.15 *M* concentration). In the presence of the fixed protein-ligand complex, the solvent was first minimized for 200 steps of steepest descent (SD) minimization. Second, the solute was minimized for 100 steps of SD minimization in the presence of fixed solvent in order to eliminate all unfavorable steric contacts possibly introduced by mutations. Subsequently, the solvent was further minimized for 1000 SD minimization steps followed by 2000 steps of full system minimization. Then, the solvent was equilibrated for 1 *ns* at 298 *K*. Subsequently, the entire system was equilibrated for 3 *ns*. Free molecular dynamics was then performed using two different approaches. First, a single long trajectory was computed for 20 *ns*. Second, ten individual short trajectories of 2 *ns* length were computed using the same input structure, but a different random number seed for the langevin thermostat. All minimization and molecular dynamics runs were performed using NAMD (version 2.7b2)⁸⁰ with the CHARMM27 all-atom force field.^{81,82} MD simulations were performed using periodic boundary conditions at constant temperature (i.e. 298 K) and pressure (i.e. 1 bar) using particle-mesh Ewald (PME) electrostatics and a timestep of 1 fs. Cutoffs for van der Waals (vdW) interactions were set to 12 Å using a switching scheme.

For the RNA and SAM ligands, published CHARMM force field parameters were used.^{92,93,94}

Trajectory Analysis

For the analysis of MD trajectories, snapshots were extracted from MD simulations. For WT and mutated MTase using the full MD mutation protocol, protein-ligand complex structures were extracted every 8 ps, yielding a total of 2500 snapshots for each system, obtained either from one long simulation or from ten individual simulations. Subsequently, all snapshots were analyzed as described below.

Binding Affinities Free energy of binding (ΔG_{bind}) between the SAM co-factor and the protein were computed using the MM-GBSA³² method.

In the MM-GBSA method, ΔG_{bind} is calculated as the ensemble average of the free energies of the complex system (G_{comp}) and the unbound protein (G_{prot}) and ligand (G_{lig}) as follows:

$$\langle \Delta G_{bind} \rangle = \langle G_{comp} \rangle - (\langle G_{prot} \rangle + \langle G_{lig} \rangle) \quad (3.1)$$

The configurational ensembles of the free protein and ligand were generated from a single simulation of the complex by extracting the unbound protein and ligand structures.

The binding free energy is computed as the sum of the molecular mechanics gas phase binding energy (ΔE_{MM}), the solvation free energy (ΔG_{solv}) and entropic contributions ($T\Delta S$).

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{solv} - T\Delta S \quad (3.2)$$

where ΔE_{MM} consists of electrostatic and van der Waals interactions:

$$\Delta E_{MM} = \Delta E_{ele} + \Delta E_{vdW} \quad (3.3)$$

The solvation free energy is separated into polar (ΔG_{polar}) and non-polar ($\Delta G_{non-polar}$) contributions, where former are approximated using the analytical generalized Born (GB) GB-MV2 model^{95,96} implemented in CHARMM and latter using the solvent accessible surface area (SASA):

$$\Delta G_{solv} = \Delta G_{polar} + \Delta G_{non-polar} \quad (3.4)$$

$$\Delta G_{solv} = \Delta G_{GB} + \Delta G_{SASA} \quad (3.5)$$

The entropic contributions ($T\Delta S$) consist of translational, rotational and vibrational contributions:

$$\Delta S = \Delta S_{trans} + \Delta S_{rot} + \Delta S_{vib} \quad (3.6)$$

ΔS_{trans} and ΔS_{rot} are functions of the mass and moments of inertia, whereas ΔS_{vib} can be calculated from a normal mode analysis. Since calculations of ΔS_{vib} are computationally very expensive, entropic contributions were calculated only every 25th snapshot, using normal

mode analysis as implemented in CHARMM on a fully minimized structure.

The relative free energy of binding between a mutant and the wild type complex ($\Delta\Delta G$) is computed as follows:

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{WT} \quad (3.7)$$

Structural Rearrangements Structural rearrangements of mutant MTases were quantified using the following three measures, based on trajectories obtained from the full MD mutagenesis protocol.

- relative donor-acceptor arrangement, defined as follows:

$$r = \|r_{CH_3-O}(WT) - r_{CH_3-O}(mutant)\| + \|r_{O-N}(WT) - r_{O-N}(mutant)\| \quad (3.8)$$

where r_{CH_3-O} is the distance between the transferring methyl group and the acceptor 2'-oxygen atom and r_{O-N} is the distance between the acceptor 2'-oxygen atom and the proton accepting N ζ atom of Lys181.

- relative RMSD of SAM compared to WT.
- relative RMSD of the RNA cap compared to WT. The RNA cap included the cap nucleotide and the following two translated nucleotides.

Intrinsic pKa of Active Site Lysine The intrinsic pKa value of the proton acceptor residue Lys181 was calculated for each mutant using the Poisson-Boltzmann electrostatic program APBS (version 1.3.0).⁹⁷ Every 10th snapshot extracted from multiple short simulations generated using the full MD mutagenesis protocol was subjected to the analysis.

Reaction Energy Barriers Reaction energy barriers were computed using molecular dynamics based umbrella sampling in a mixed quantum mechanics/molecular mechanics (QM/MM) system. The approximate DFT method, self-consistent charge density functional tightbinding (SCC-DFTB^{98,99}) was used for the QM level as implemented in CHARMM. The CHARMM27 all-atom force field was used for the MM level.

As starting structures, four snapshots were extracted from a standard MD simulation of the WT system. Mutations were introduced using the post-processing alanine scanning protocol described before. The system was set up in an analogous fashion as for MD simulations.

The QM system included all atoms of SAM, the side chain of Lys181 and the first translated RNA nucleotide. Link atoms between the QM and MM part were placed between C α and C β atoms of the Lys181 side chain and between C3' and O3' as well as C4' and C5' atoms of the RNA.

Reaction energy profiles were computed using umbrella sampling as implemented in CHARMM in order to sample the reaction coordinate. The following reaction coordinate was used:

$$rx = \frac{\text{dist}(A - CH_3)}{\text{dist}(A - D)} \quad (3.9)$$

where A is the RNA 2' oxygen methyl acceptor atom, D is the SAM sulfur donor atom and CH_3 is the transferred methyl group.

The reaction coordinate equals to 0.65 for the reactant state and to 0.30 for the product state. 33 equidistantly spaced umbrella potentials were used, applying a force constant k of $10000 \text{ kcal/mol/\AA}^2$. For each window, the system was minimized and equilibrated for 2 ps , followed by a 10 ps production run.

The final reaction energy profiles were obtained by combining the results from the individual simulations using the weighted histogram analysis method as implemented in the program WHAM (version 2.0.2).¹⁰⁰

For validation, two additional umbrella sampling runs were performed on the WT system. First, an umbrella sampling run with longer simulation times, i.e. 10 ps equilibration and 20 ps production run. Second, an umbrella sampling run with twice as many umbrella windows. For both runs, the observed reaction energy profile was the same within the error margin, determined as the standard deviation computed from individual simulations.

3.4.2 Results of Computational Alanine Scanning

To identify protein residues which significantly modulate the enzymatic reaction, and to further investigate the underlying cause for the observed effect, computational alanine scanning was performed. The idea behind computational alanine scanning¹⁰¹ originates from experimental mutagenesis studies,¹⁰² where the influence of a protein residue is approximated by mutating it to alanine. This relies on the assumption that alanine has a neutral influence due to its small and non-polar side chain while being chiral and structurally rigid compared to glycine.

In the following, results of the computational alanine scanning are reported, divided into their effects on the geometric arrangement between methyl donor and acceptor, the methyl donor binding affinity, the reaction energy barrier and the intrinsic pKa value of the active site proton acceptor group.

Geometric Arrangement

The geometric arrangement between the methyl donor (SAM), the methyl acceptor (RNA 2' hydroxy group) and the proton acceptor (protein Lys181) were investigated for wild type and single point mutations of dengue MTase using molecular dynamics simulations. The relative donor/acceptor arrangements of all alanine mutants compared to WT are plotted in Figure 3.12 against the experimental determined methylation activity obtained from the literature.^{103, 61, 104, 105, 45, 106, 58}

The figure shows the relative geometry obtained for each mutant on the x-axis and the experimental methylation activity on the y-axis. All mutants with no experimental results available are shown as red dots below the x-axis. The gray bar indicates uncertainties in

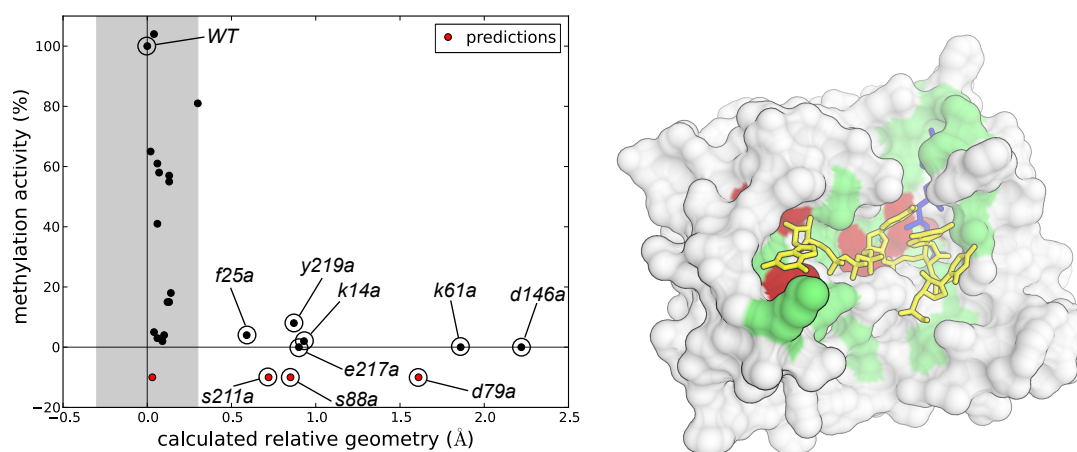


Figure 3.12: (A) Relative donor/acceptor arrangements of all alanine mutants plotted against the experimentally determined methylation activity. Mutants with no experimental results available are shown as red dots below the x-axis. The gray bar indicates uncertainties in computational results of WT. (B) Results mapped onto the surface of the MTase where all mutated residues are marked in green and residues yielding significant rearrangements are highlighted in red.

computational results of WT calculated as the standard deviation observed between 10 individual 2 *ns* simulations. Uncertainties observed in mutant simulations are in a comparable range as for WT but were omitted for clarity.

In addition, the results are mapped onto the surface of the MTase where all mutated residues are marked in green and residues yielding significant rearrangements are highlighted in red.

All results are in agreement with experimental values, since no false negatives are observed, i.e. mutants which are experimentally active but inactive in our simulations. On the other hand, we observe certain mutants which are inactive in experiment, but are predicted to be active from our simulations. This is expected, since the observable investigated here, i.e. orientation of donor/acceptor, is only one of many possible reasons to obtain a reduced methylation activity. Thus, for those mutants, we do not observe a difference which is significant. This indicates a different mechanism of inactivation for those mutants.

To further characterize the underlying cause of the structural rearrangement observed here, the relative RMSD of SAM and of the RNA cap was investigated as shown in Figures 3.13 and 3.14.

The results are again in agreement with experimental methylation activity. They indicate that the mutants S88A, D79A, D146A, K181A and E217A influence the arrangement of the SAM methyl donor and the mutants K14A, F25A, K61A, W87A, D146A and Y219A affect the conformation of the RNA cap structure.

Binding Affinities

Binding free energies of the methyl donor SAM to the MTase were investigated using the MM-GBSA approach. Two distinct computational mutagenesis protocols were used where MD simulations were performed either on the WT only (post-processing protocol) or on all mutants individually (full MD protocol) as described before in more detail. Similar methods have been

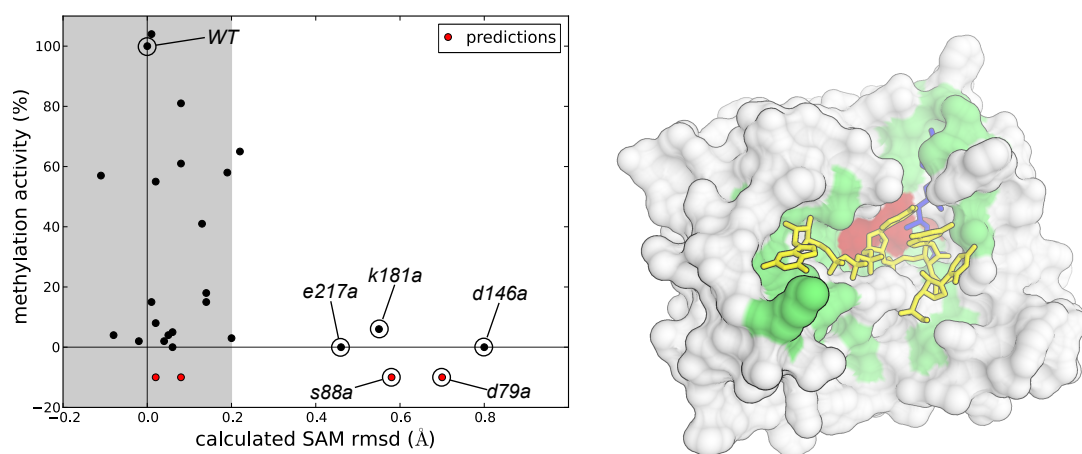


Figure 3.13: (A) Relative SAM RMSD of all alanine mutants plotted against the experimental determined methylation activity. Mutants with no experimental results available are shown as red dots below the x-axis. The gray bar indicates uncertainties in computational results of WT. (B) Results mapped onto the surface of the MTase where all mutated residues are marked in green and residues yielding significantly increased SAM RMSD are highlighted in red.

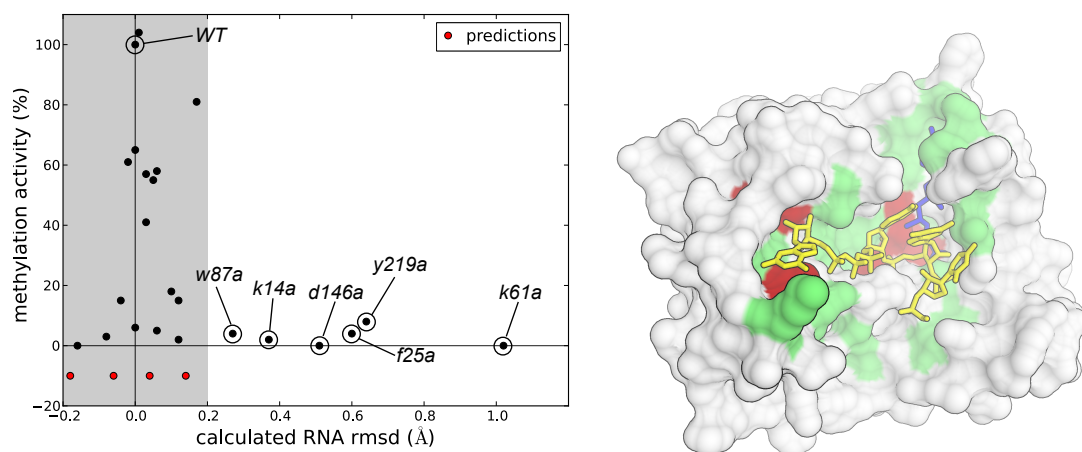


Figure 3.14: (A) Relative RNA cap RMSD of all alanine mutants plotted against the experimental determined methylation activity. Mutants with no experimental results available are shown as red dots below the x-axis. The gray bar indicates uncertainties in computational results of WT. (B) Results mapped onto the surface of the MTase where all mutated residues are marked in green and residues yielding significantly increased RNA cap RMSD are highlighted in red.

used for studying protein-ligand and protein-protein interactions.^{107,108,109} Results from the post-processing protocol are shown in Figure 3.15.

The MM-GBSA binding free energies based on the post-processing mutagenesis protocol are in agreement with experimental results as no false negatives were observed. In addition, for the K14A mutant, binding of SAM to the MTase was experimentally confirmed using the previously described ITC assay. Computationally, this was also observed in MM-GBSA binding free energies using the post-processing mutagenesis protocol, where $\Delta\Delta G = -0.3 \text{ kcal/mol}$ for the K14A mutant. In contrast, when using the full MD mutagenesis protocol and MM-GBSA energies were computed as the average of ten individual MD simulations, two false negatives were observed. For those cases, predicted binding affinities were significantly reduced for the

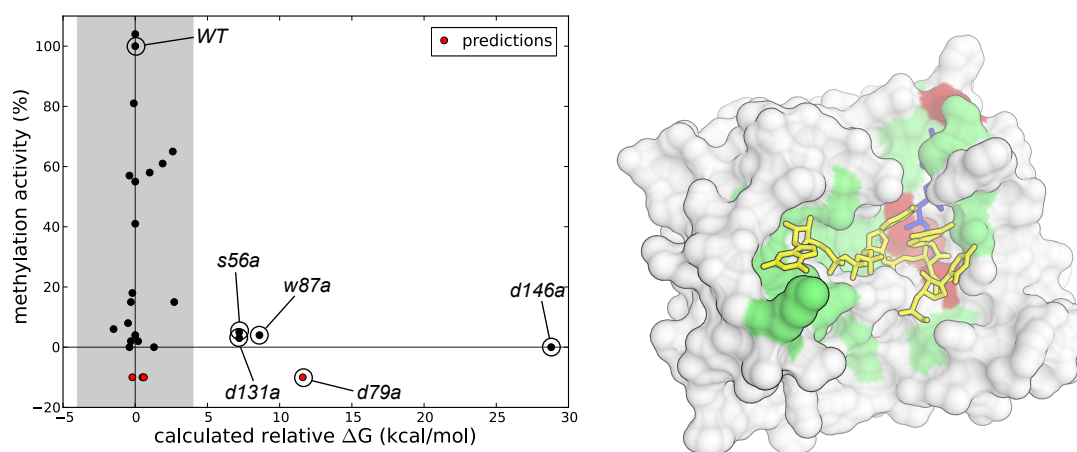


Figure 3.15: (A) Relative binding free energy of SAM to the MTase as computed by MM-GBSA. Results for all alanine mutants plotted are against the experimental determined methylation activity. Mutants with no experimental results available are shown as red dots below the x-axis. The gray bar indicates uncertainties in computational results of WT. (B) Results mapped onto the surface of the MTase where all mutated residues are marked in green and residues yielding significantly lowered binding affinities are highlighted in red.

following two mutants which retain their methylation activity in experiments: F133A (65% activity), E111A (57% activity). Similar findings, with more pronounced effects, were observed when using one long (20 ns) MD simulation instead of 10 individual short (2 ns) simulations. In addition, the experimentally examined mutant K14A, is predicted not to bind ($\Delta\Delta G = 12.2$ kcal/mol), which contradicts our experimental ITC binding results. Results on the geometric arrangement of the RNA cap and inspection of the MD trajectories suggest that the mutant K14A induces a structural rearrangement of the RNA cap and thus the protein's active site which in turn influences the calculated MM-GBSA binding free energies. This suggests that MM-GBSA binding free energies are less reliable if the protein or the ligand undergoes structural rearrangements.

To summarize, only the post-processing mutagenesis protocol yields binding affinity results in agreement with experimental results. Similar findings have been reported for computational alanine scanning when probing protein-protein interfaces.¹⁰⁷ Although this protocol is more accurate for our study, it does not incorporate structural rearrangements taking place upon mutation, which could lead to false positive predictions (i.e. mutants which are experimentally inactive but do not show reduced binding affinity in the simulations).

Overall, the results predict mutants S56A, D79A, W87A, D131A and D146A to significantly reduce the binding affinity of SAM to the MTase and thereby inhibiting the methyltransfer reaction.

Reaction Energetics

Reaction energy barriers were computed for a selected subset of mutations using molecular dynamics based umbrella sampling in a mixed quantum mechanics/molecular mechanics (QM/MM) system. The approximate DFT method self-consistent charge density functional

tightbinding (SCC-DFTB^{98,99}) was used for the QM level whereas the CHARMM27 all-atom force field was used for the MM level. SCC-DFTB is a computationally efficient method comparable to widely used semi-empirical methods such as AM1 and PM3 with reasonable accuracy. This makes it possible to perform extensive sampling of condensed phase systems. The SCC-DFTB method has been applied successfully to numerous biomolecular systems.^{110,111}

Since QM/MM calculations are computationally extremely costly, only WT and the following four protein residues were investigated as they are located closely to the reactive center: K61A, D146A, E217A, Y219A. All of these abolish methylation activity for the 2'O reaction. For the WT system, a energy barrier of 19.4 ± 2.0 kcal/mol was obtained. This is in a similar range as the barrier computed from experimentally observed k_{cat} using transition state theory ($\Delta G_{exp}^{TS} = 15.2$ kcal/mol) and agrees well with the results obtained from QM calculations on 2'O model systems ($\Delta G_{model}^{TS} = 17.6$ kcal/mol). Relative reaction energy barriers and fluctuations observed between four input structures are given in Figure 3.16.

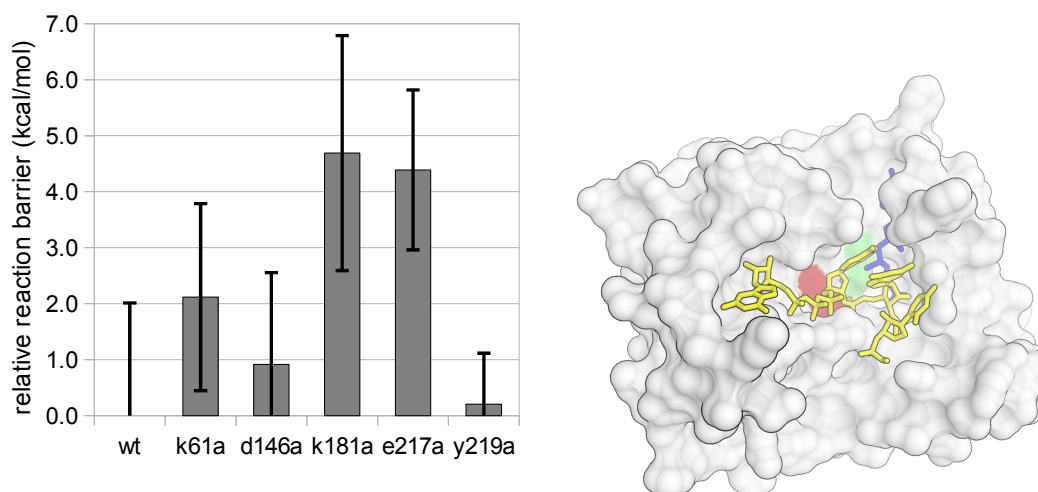


Figure 3.16: (A) Reaction energy barriers relative to WT, computed by umbrella sampling in a QM/MM system. Error bars were determined as the standard error obtained from four individual starting structures. (B) Results mapped onto the surface of the MTase where all mutated residues are marked in green and residues yielding significantly increased reaction energy barriers are highlighted in red.

The results show an increased reaction barrier for the mutants E217A and K61A. No significant increase is observed for mutant D146A and Y219A. In addition, for the mutant K181A, a significant increase in the reaction barrier by 4.7 kcal/mol is observed. In the native protein, Lys181 acts as the proton acceptor. Therefore, for the mutation K181A, a water molecule was modeled in the position of Lys181 N ζ and included into the QM region. This water molecule then acts as a proton acceptor. The increase in the reaction energy barrier for this mutant are comparable to the results obtained from QM calculations on 2'O model systems ($\Delta\Delta G_{model}^{TS} = 3.8$ kcal/mol).

Intrinsic pKa of Active Site Lysine

During the reaction, the protein residue Lys181 acts as a proton acceptor and thus, its protonation state is critical for the efficiency of the reaction, where a lowered pKa value improves the reaction. It has been shown in VP39 methyltransferase that this residue has a significantly lowered pKa value (pKa: 8.5) compared to an isolated lysine (pKa: 10.5).^{64,65}

Although VP39 MTase shows significant structural differences to DENV MTase, the active site is highly conserved both from a sequence as well as from a structural point of view. Therefore, intrinsic pKa values of Lys181 were computed for all complex structures of all studied single point mutants of DENV MTase based on the Poisson-Boltzmann continuum electrostatics methods. Calculations on WT MTase show a comparable reduction for DENV MTase with a predicted Lys181 pKa of 9.2.

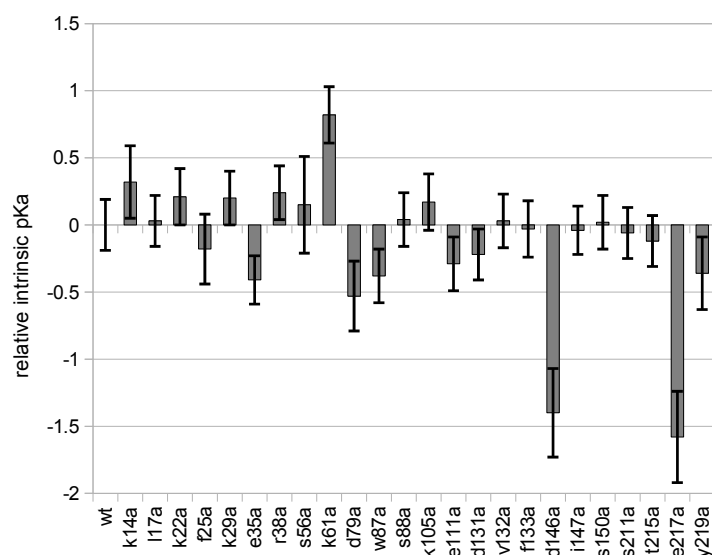


Figure 3.17: Calculated intrinsic relative pKa values of Lys181 for WT dengue MTase and all studied single point mutations.

Predicted pKa values for all mutants are given in Figure 3.17. No significant change of the pKa value was observed for most mutants. Only for two mutations a significant decrease over WT was observed: E217A (pKa: 7.6, WT pKa: 9.2), D146A (pKa: 7.8) whereas for one mutation a significant increase was observed: K61A (pKa: 10.0).

The enzyme's active site consists of the following four residues Lys181, Asp146, Lys61 and Glu217, where the two negatively charged residues are held together by the two lysines. For the reaction to occur, Lys181 must be deprotonated, and thus, the two negatively charged residues are only held together by one positive charge of Lys61. If this charge is removed (K61A), the protonated and therefore positively charged form of Lys181 is favored in order to stabilize the two negative charges. Thus, the pKa value of Lys181 is increased compared to WT. On the opposite, if one of the negative charges is removed (e.g. D146A, E217A), the deprotonated form of Lys181 is favored and thus, the pKa of Lys181 is reduced. However, as described previously, mutation of D146A, E217A or K61A leads to a structural rearrangement which inhibits the

reaction.

3.4.3 Summary of Computational Alanine Scanning

The results from all computational alanine scanning computations are summarized in Figure 3.18 and Table 3.6.

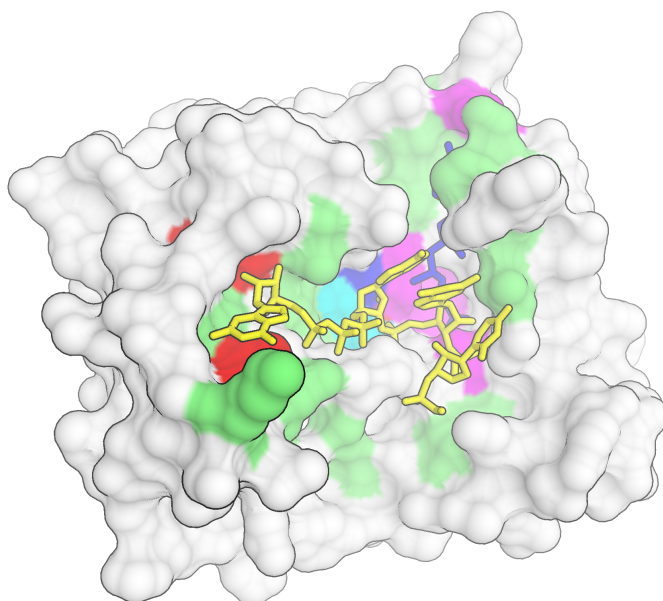


Figure 3.18: Summary of the effect of single point mutations on different aspects of the methyltransferase reaction. The following effects are mapped onto the protein surface: structural rearrangements of the RNA cap (red), structural rearrangements of the SAM co-factor (blue), reduced SAM binding affinity (magenta), increased reaction barrier (cyan). Residues where mutations to alanine did not have an effect are colored in green.

For most mutants with a significant effect on methylation activity, computations indicate an underlying cause, except for mutants K29A, E35A, R38A and I146A where all computed observables do not indicate any significant change. This discrepancy is caused by the fact that the former three residues are located at the far end of the RNA binding groove and are thus not appropriately reflected by the observables analyzed. The latter residue, on the other hand, is located in the SAM binding pocket but the chemical change of the isoleucine to alanine mutation is relatively small and thus, does yield a calculated binding free energy difference (2.7 *kcal/mol*) just below the significance threshold.

3.4.4 Experimental and Computational Analysis of Selected Mutants

Based on results from computational alanine scanning calculations, residues were selected for further computational and experimental mutagenesis experiments.

The following mutations were characterized experimentally using ITC based binding affinity measurements: D79E, D79N, S88T, E217D, E217Q, Y219F, Y219H. The obtained binding curves of SAH and GTP γ S are shown in Figures 3.19 and 3.20.

Table 3.6: Summary of the effect of single point mutations on different aspects of the methyltransferase reaction. ^a The “—” sign indicates that there is no data available in the literature.

mutant	experimental activity (%) ^a	SAM	SAM	RNA cap	protonation	reaction
	103, 61, 104, 105, 45, 106, 58	binding	orientation	orientation	state	barrier
K14A	2			x		
L17A	41					
K22A	81					
F25A	4			x		
K29A	15					
E35A	2					
R38A	18					
S56A	5	x				
K61A	0			x	x	x
D79A	—	x	x			
W87A	4	x		x		
S88A	—		x			
K105A	61					
H110A	58					
E111A	57					
D131A	3	x				
V132A	—					
F133A	65					
D146A	0	x	x	x		
I147A	15					
S150A	55					
K181A	6		x			x
S211A	—					
T215A	104					
E217A	0		x			x
Y219A	8			x		

As described previously, ITC based SAH binding measurements are difficult. Despite this fact, qualitative results can still be obtained when measuring SAH binding. The results show that SAH binds to the WT MTase (black dots) and all mutants of Tyr219 (i.e. Y219F, Y219H). However, all studied mutants of Asp79 (i.e. D79E, D79N), Ser88 (i.e. S88T) and Glu217 (i.e. E217D, E217Q) do not bind SAH.

Binding of the RNA was determined using GTP γ S as a short cap analog. For all observed mutants, the binding was retained. As a control experiment, binding of GTP γ S to the mutant K14A was measured. Lys14 is located in the RNA cap pocket and is known to abolish methylation activity. Our ITC experiments confirm that binding of GTP γ S is eliminated in this mutant.

Mutations of Asp79

Asp79 is located in the SAM binding pocket. The side chain of Asp79 interacts with the amino group of SAM through a water mediated interaction. Thus, mutation of Asp79 to alanine abolishes this interaction. So far, no published methylation activities are available for mutants

3.4. IMPACT OF SINGLE POINT MUTATIONS

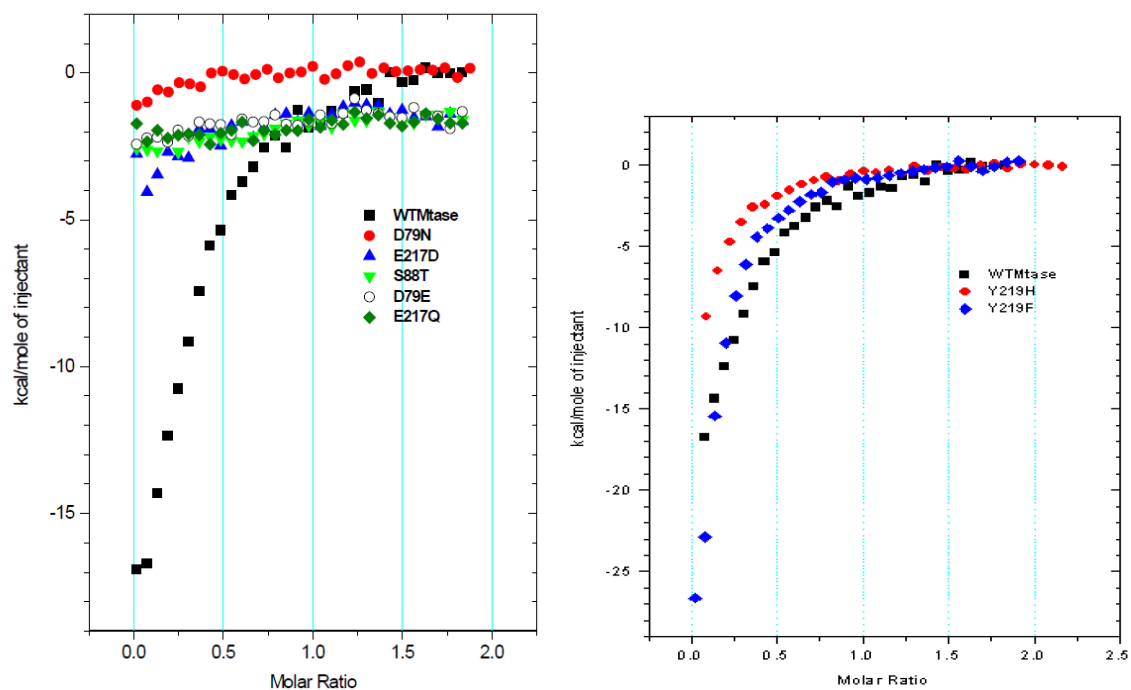


Figure 3.19: Experimentally determined ITC binding curves measured for SAH binding to WT DENV MTase and mutants thereof.

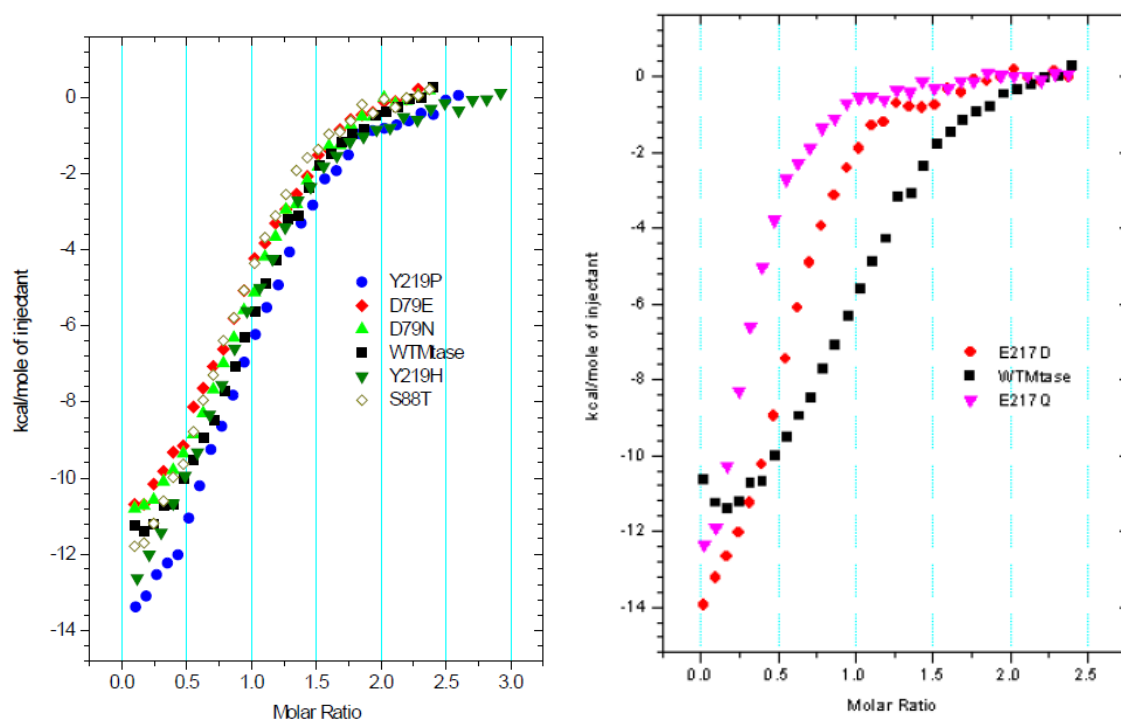


Figure 3.20: Experimentally determined ITC binding curves measured for $GTP\gamma S$ binding to WT DENV MTase and mutants thereof.

of Asp79. Simulations however, indicate a reduced SAM binding affinity and a conformational change of SAM, incompatible with the reaction.

Experimentally, ITC measurements of mutations to Glu and Asn were performed. The mutation to Glu elongates the side chain by one carbon atom but retains the charge. Thus, it probes the dependencies on size effects. On the other hand, the mutation to Asn changes the negatively charged native residue while retaining the same size. Both mutants were shown to abolish SAH binding in the ITC experiments.

Computational mutagenesis experiments are in good agreement with those findings. Significantly reduced SAM binding affinities were predicted with ΔG shifts of 14.4 *kcal/mol* and 18.7 *kcal/mol* for the mutants Asp79Glu and Asp79Asn, respectively. As opposed to Asp79Ala mutant, both mutants did not show significant structural rearrangements during MD simulations with relative SAM rmsds of -0.12 Å and 0.25 Å and relative RNA rmsds of -0.16 Å and -0.17 Å.

Mutations of Ser88

Ser88 is located in the second shell of the SAM binding pocket. It is hydrogen bonded to the previously described Asp79 and thereby it constrains the position of the latter which is involved in SAM binding. So far, no published methylation activities are available for mutants of Ser88.

Experimentally, mutation to Thr was performed, which retains the capability of forming hydrogen bonds while being sterically more demanding and conformationally less flexible. Due to the close proximity to Val100 and Val124, the hydroxy group of Thr needs to adopt a different orientation compared to the native Ser hydroxy group, where in this conformation, the hydrogen bond to Asp79 is not retained. In the ITC experiments, the Ser88Thr mutant loses the ability to bind SAH.

Computationally, no shift was observed in the SAM binding free energy compared to WT ($\Delta\Delta G = 0.5$ *kcal/mol*). Since the residue Ser88 does not directly interact with SAM it likely influences SAM binding affinity through structural rearrangements. This aspect cannot be covered by the employed computational method since the computational mutagenesis protocol does not adequately incorporate structural rearrangements.

Mutations of Glu217

Glu217 is located in the protein's active site and belongs to the conserved and essential residues of the catalytic tetrad consisting of Lys181, Asp146, Lys61 and Glu217. Glu217 interacts directly with the proton acceptor residue Lys181. The correct positioning of Lys181 is essential for an efficient reaction and Glu217 helps to constrain this position. Experimentally, mutant Glu217Ala was shown to abolish methylation activity.

ITC binding affinity measurements were performed on mutants Glu217Asp and Glu217Gln. Mutations to Asp retain the negative charge on the carboxy group but shorten the side chain by one carbon atom. Mutations to Gln retain the side chain length but render the residue neutral

while retaining the ability to form hydrogen bonds. Both mutants were found to eliminate SAH binding while binding to GTP γ S was retained.

Computationally, for both mutants, no significant shifts in SAM binding free energies were observed (Glu217Asp: $\Delta\Delta G = 0.0 \text{ kcal/mol}$, Glu217Gln: $\Delta\Delta G = 1.4 \text{ kcal/mol}$). Like for the mutants of Ser88, SAM binding is likely influenced through structural rearrangements induced by the mutation which are not adequately covered in the computational mutagenesis protocol.

By monitoring the geometry between methyl donor, acceptor and proton acceptor in ten individual MD trajectories for each mutant, significant conformational rearrangements were observed in the mutants. These can be attributed to conformational rearrangements of the proton acceptor Lys181 as indicated by a significant increased relative rmsd of the Lys181 residue compared to WT (Glu217Asp: $0.8 \pm 0.14 \text{ \AA}$, Glu217Gln: $0.7 \pm 0.09 \text{ \AA}$).

Mutations of Tyr219

Tyr219 is located in the protein's active site. It is hydrogen bonded to the carboxy group of Asp146 of the catalytic tetrad. No direct interactions with SAH or the RNA were observed. Published activity data of mutant Tyr219Ala shows no methylation activity.

Experimentally, SAH binding affinities were measured for the mutants Tyr219Phe and Tyr219His. In both cases, binding affinities are retained at WT levels. Calculated SAH binding free energies are in agreement, where no significant shifts in predicted SAM binding free energies were observed.

For the mutation Tyr219Ala, a change in the donor-acceptor geometry was observed. This is not the case for the Tyr219Phe mutation. However, for the latter mutation, a significant increase was observed in the relative reaction energy barrier by $2.8 \pm 1.33 \text{ kcal/mol}$. Therefore, computations suggest that the Tyr219Phe mutant inhibits the methylation reaction through modulating the reaction energy barrier.

3.4.5 Conclusion

Using complementary computational methods, we have build an in-silico approach to identify the effect of a single point mutation on different aspects governing the enzyme's catalytic activity.

Results of a computational alanine scanning procedure are qualitatively in good agreement with experimentally determined methylation activity. In addition, they indicate the role of each studied residue, thereby yielding information not easily accessible through experiments. In this way, protein residue patches were identified which modulate the geometric arrangement between methyl donor and acceptor, the methyl donor binding affinity or the reaction energy barrier.

Based on a computational alanine scanning procedure, previously uncharacterized hot-spot residues were identified which were predicted to significantly influence the catalyzed reaction. Selected hot-spot residues were further analyzed both computationally and experimentally in order to gain a better understanding of their role in the enzyme's function. In addition, the

results agree well with experimental binding free energies in cases where mutations do not induce significant structural rearrangements.

3.5 RNA Sequence Specificity

Although flaviviral methyltransferases are attractive drug targets,^{61,79} little is known about the atomistic details of the mechanism underlying their function. The enzyme is known to catalyze two independent methyltransfer reactions onto two distinct positions of the RNA 5'-cap structure: the guanine N7 position on the cap nucleotide and the ribose 2'O hydroxy group on the first transcribed nucleotide.^{60,53,50,106} Dong et al.¹¹² have shown that distinct RNA elements with a specific sequence are required for the methylation of the two RNA positions, indicating that specific RNA binding conformations and methyltransfer mechanisms are required. So far, the underlying reason for this sequence specificity, however, has not been elucidated in detail and the mechanism of the two methylation reactions at an atomistic level is not known. Therefore, we are investigating effects on the methyltransfer reaction introduced by mutated RNA sequences at an atomistic level using molecular dynamics simulations.

Since structural details about the guanine N7 reaction are largely unknown, we focus on the ribose 2'O methyltransfer reaction. For the 2'O reaction, it has been experimentally determined that the identity of the cap nucleotide as well as the following two transcribed nucleotides, i.e. GpppAG-RNA, are essential for the reaction.¹¹² Those nucleotides are conserved within all flaviviruses. A chemical drawing of the cap structure and its interactions with the protein are shown in Figure 3.21.

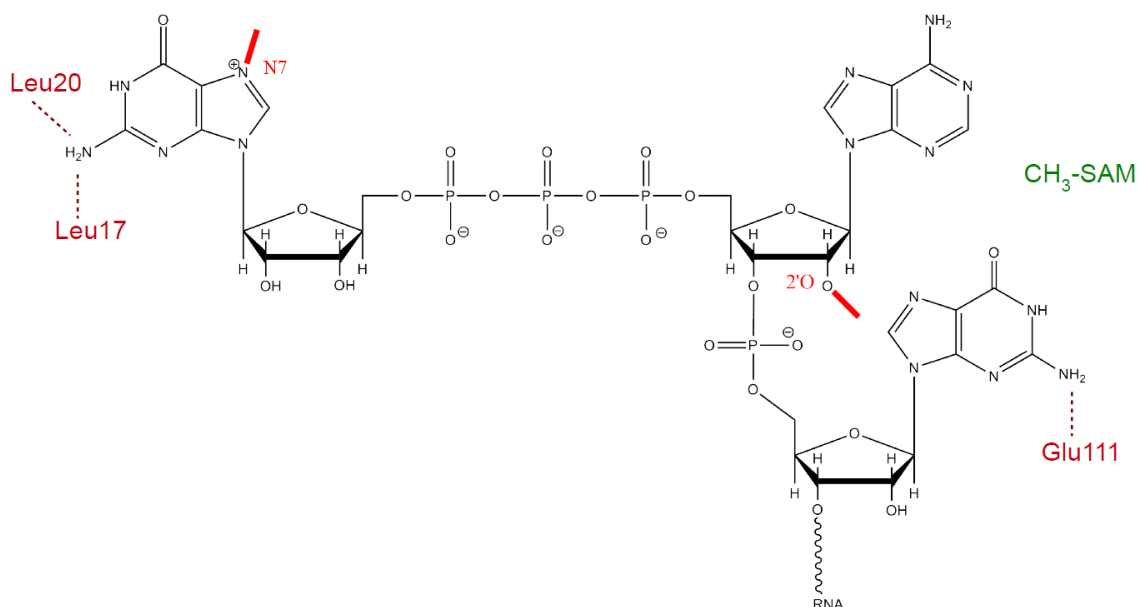


Figure 3.21: Schematic representation of wild type dengue RNA cap-1 structure located at the 5' termini. Specific interactions with the protein observed during molecular dynamics simulations are indicated in red. The location of the methyl donor SAM is indicated in green letters.

3.5.1 Method

System Setup

Wild type (WT) and nine mutated RNA sequences were modeled in the binding site of the dengue NS5 methyltransferase. Starting from the native sequence, the structures of the nine mutated sequences were generated by mutating the appropriate nucleotide to all other nucleotides, using Maestro (Schrodinger, LLC). Each base was subjected to a full conformational search using MacroModel (Schrodinger, LLC) with default parameters, while keeping the protein and the rest of the RNA fixed. The minimum energy structure was used as a starting point for molecular dynamics simulations. System setup and simulation procedure was the same as described in Section 3.2.1.

Protein-RNA Contact Distances

An RNA-protein contact map was computed as follows. First, the set of atoms in contact with the protein was defined as all atom of protein residues where at least one atom of the residue was closer than 8 Å from the RNA in the initial WT structure. Second, for each of the atoms in this set, the closest distance to any atom of the RNA was computed. This was repeated for every 10th frame of the MD trajectory. Mean and standard deviation were computed over the whole trajectory. Subsequently, for each mutant those values were plotted against the same distance in WT. All distances deviating by more than 30% were labeled by their residue number. In addition, for a global measure of the agreement between WT and mutant, an rmsd between the corresponding distances was computed. All analysis was done in OpenStructure (version 1.2.1).⁸³

RNA Per-Residue RMSD

For each nucleotide, the RMSD between every 10th frame of the MD trajectory and the initial structure was computed. Mean and standard deviation were computed over the whole trajectory.

Hydrogen Bonding Distance

The hydrogen bonding distance between the 2nd nucleotide and the γ -phosphate group of the triphosphate linker was computed for every 10th frame of the MD trajectory and mean and standard deviation over the whole trajectory.

Reaction Energy Barriers

Reaction energy barriers were estimated using QM/MM based umbrella sampling simulations with the SCC-DFTB method as described in Section 3.4.1.

3.5.2 Results and Discussion

Overall Structural Rearrangements

Since a correct alignment of the methyl donor and acceptor is essential for the reaction to occur efficiently, structural effects of mutated RNA sequences were investigated first. Therefore, molecular dynamics simulations were performed for MTase complexed to mutated RNA cap structures. For each position in the RNA sequence that is essential for the 2'O reaction, all possible RNA mutants at that position were modeled. Subsequently, those simulations were compared to simulations performed for MTase complexed to wild type RNA cap structure.

To quantify structural rearrangements of mutant RNA sequences in comparison to WT RNA sequence, a contact map between the RNA and the surrounding protein atoms was computed as described in Section 3.5.1. Plots for the three RNA mutants most closely related to the WT RNA sequence (i.e. ApppAGU, GpppGGU, GpppAAU) are shown in Figure 3.22, all plots are given in Appendix A.1. Points on the diagonal indicate no difference between WT and the mutant, whereas points above the diagonal indicate an elongated distance in the mutant compared to WT. In addition, a global measure of the agreement between WT and mutant was computed as the rmsd between the corresponding distances in mutant and WT distance maps. The value is shown in the plot.

As shown in Figure 3.22 the most significant structural rearrangements are observed for all mutants of the cap-nucleotide, indicated by significant distance changes and the highest overall distance rmsd of 0.92 Å for the RNA sequence ApppAGU. All mutants of the 1st nucleotide show the smallest structural rearrangements with an distance rmsd of 0.34 Å for the RNA sequence GpppGGU. Mutants of the 2nd nucleotide show some structural rearrangements with a distance rmsd of 0.51 Å for the RNA sequence GpppAAU.

Cap Nucleotide Mutations

In the native structure, the cap nucleotide is strongly interacting with the protein by forming direct hydrogen bonds from the guanosine amine group to the backbone carbonyl groups of residues Leu20 and Leu17. When mutating this position to any other nucleotide, the RNA is no longer capable of forming this direct interaction with the protein. Thus, a less stable binding of this residue to the GTP pocket is observed, leading to a significant structural rearrangement of the RNA during the 20 ns of MD simulation as observed by the significantly elongated distances between RNA and protein residues 20, 21, 22 and 24 shown in Figure 3.22. All those residues are located in the GTP pocket, indicating the largest structural rearrangements to occur in that subpocket.

To quantify structural rearrangements of the cap-nucleotide, Table 3.7 shows the rmsd values of the cap nucleotide obtained during MD simulations, both for WT and all RNA mutants. It is clear that all mutants of the cap nucleotide show a significantly increased cap nucleotide rmsd compared to WT or all mutants of all other nucleotides.

Figure 3.23 shows a superposed snapshot obtained from the MD simulations of WT (blue)

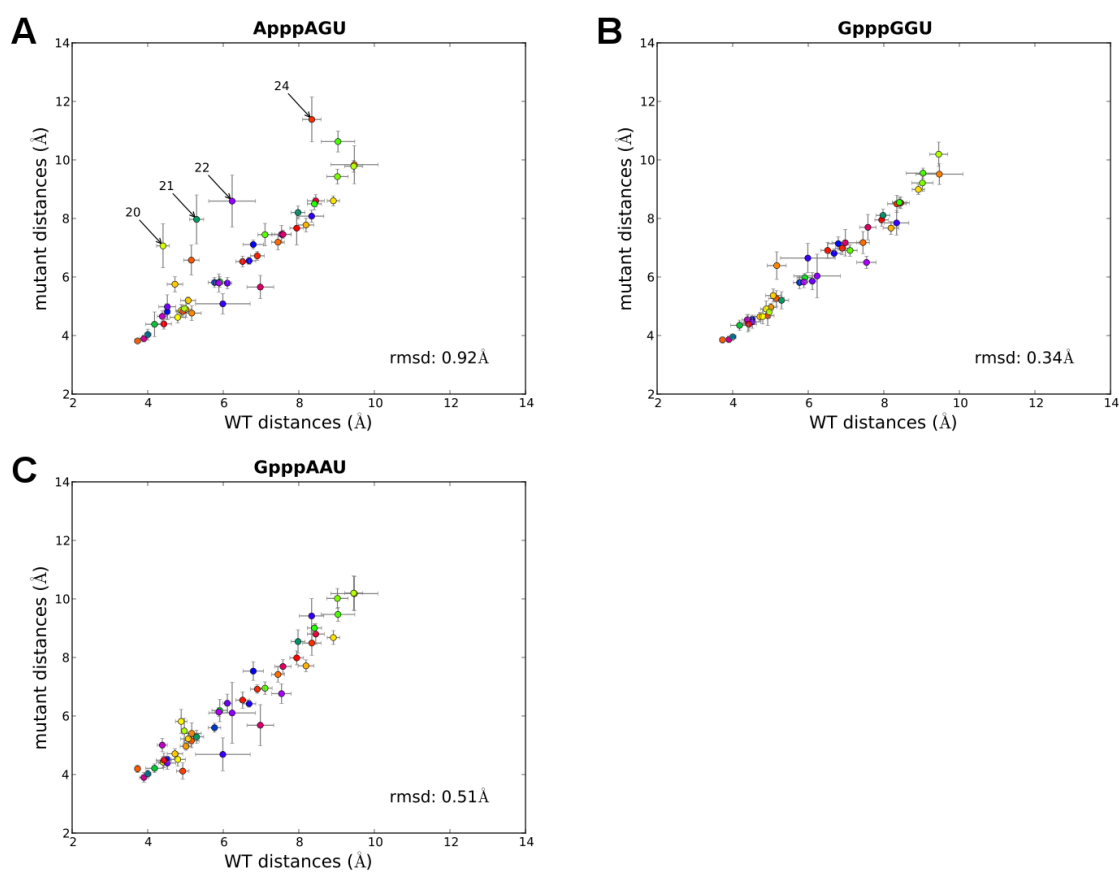


Figure 3.22: Differences in RNA-MTase contacts between WT and mutated RNA. Contacts are measured by their distances observed in MT simulations of the MTase bound to SAM and WT or mutated RNA. The three RNA mutants most closely related to the WT RNA sequence are shown: (A) cap-nucleotide (ApppAGU), (B) 1st nucleotide (GpppGGU), (C) 2nd nucleotide (GpppAAU). Residues with significantly modified distances are labeled by their number.

and ApppAGU mutant (green) which clearly visualizes the weaker interactions in the mutant leading to a partial unbinding of the cap nucleotide in the GTP pocket and therefore to a rearrangement of the overall RNA.

1st Nucleotide Mutations

Although the first transcribed nucleotide is located closely to the active side where the reaction takes place, no significant structural rearrangement was observed when mutating this residue. However, when investigating the interactions with the protein, it was found that the amine group of the guanine in the mutated RNA is pointing directly towards the reactive sulfur of SAM. Thus, the environment of the reaction is modified considerably which might have a significant effect on the energetics of the catalyzed reaction. However, this hypothesis could not be confirmed by reaction energy barrier calculations where no significant change was observed ($\Delta\Delta G^{TS} = 0.5 \pm 1.6 \text{ kcal/mol}$).

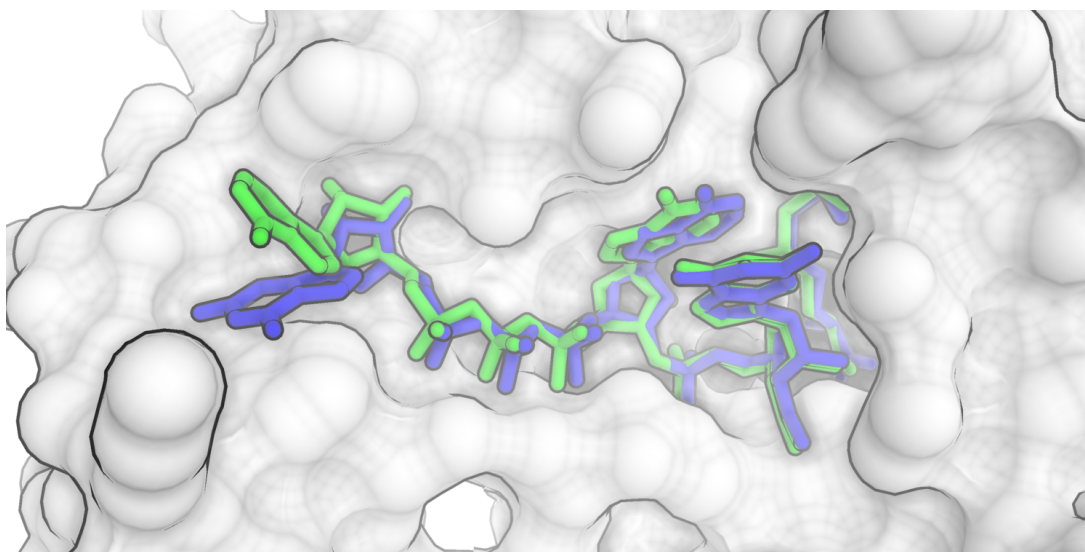


Figure 3.23: Snapshots extracted from MD simulations of MTase in complex with SAM and WT RNA GpppAGU (blue) or cap mutant RNA ApppAGU (green). The MTase structure is represented by its molecular surface whereas the RNA is shown in sticks. For clarity only the cap and the first two RNA nucleotides are shown.

2nd Nucleotide Mutations

For mutants of the second transcribed nucleotide, some structural rearrangements are observed when visual inspecting the MD trajectories. This rearrangement leads to a non-optimal geometry between the methyl-donor and acceptor.

During the whole simulation of the native structure, the guanosine at the second transcribed position is hydrogen bonded through the amine group at the 2 position to the side chain carboxyl group of Glu111. This interaction might participate in constraining the position of the bound RNA to a conformation suitable for the reaction. When this nucleotide is mutated to an adenosine, the hydrogen bond can no longer be formed. Recently, however, data for the WNV single point mutation of Glu111 to alanine was published which shows a reduction of the methylation efficiency to 57% of the WT efficiency.⁴⁵ This indicates, that the loss of the aforementioned interaction is insufficient to fully describe the almost complete loss of activity for the RNA mutants at this position.

Inspection of the MD trajectories showed that the mutated RNA takes up a more compact conformation compared to WT RNA. This seems to be induced by an intramolecular hydrogen bond between the second nucleotide and the γ -phosphate group of the triphosphate linker, as indicated in Figure 3.24 This hydrogen bond is possible in all mutants of the second nucleotide, however not in the WT RNA. To quantify this finding, the above mentioned distance is measured over the whole trajectory and the mean values are given in Table 3.7. The observed distance is significantly shorter in all mutants of the second nucleotide compared to wild type RNA or all mutants of the other RNA positions.

Although the data does not allow to draw the following conclusion significantly, due to the large fluctuations observed in the simulations, it is interesting to note that within the mutants

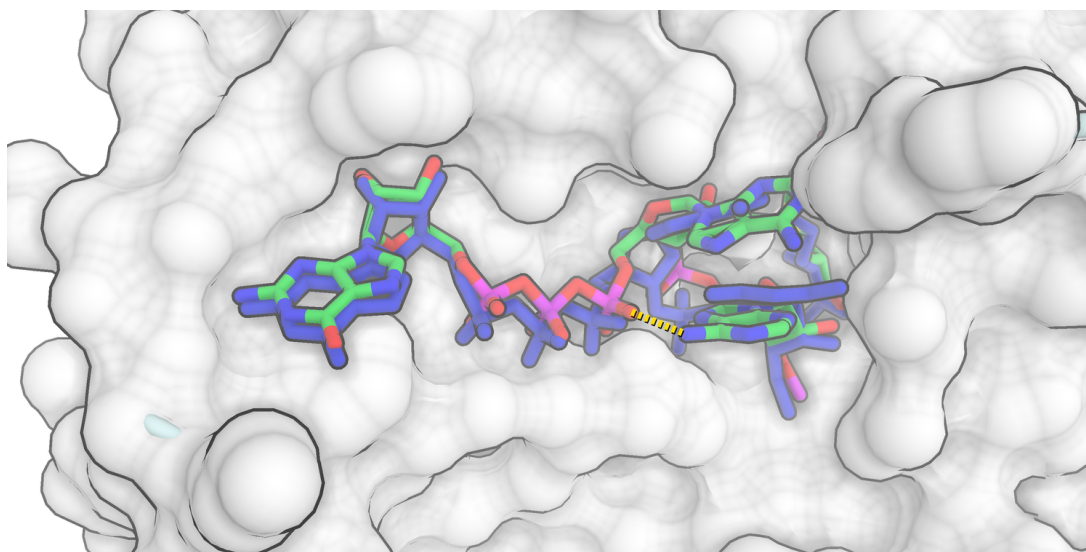


Figure 3.24: Snapshots extracted from MD simulations of MTase in complex with SAM and WT RNA GpppAGU (blue) or 2nd-nucleotide mutant RNA GpppAAU (green/element color). The MTase structure is represented by its molecular surface whereas the RNA is shown in sticks. The hydrogen bond between the 2nd-nucleotide and the γ -phosphate of the triphosphate linker is indicated with a dashed yellow line. For clarity only the cap and the first two RNA nucleotides are shown.

Table 3.7: RNA sequence specificity of dengue methyltransferase. Enzymatic activity is obtained from the literature. Two observables obtained from molecular dynamics simulations are shown: (A) RMSD of the cap-nucleotide from the starting geometry, (B) hydrogen bonding distance between the 2nd-nucleotide and the γ -phosphate of the triphosphate linker.

RNA sequence	experimental activity ^{44, 112}	cap nucleotide RMSD (Å)	h-bonding distance (Å) $\gamma\text{PO}_4 - 2^{\text{nd}}\text{-nucleotide}$
GpppAGU (WT)	100%	0.56	4.17
ApppAGU	0%	1.87	4.20
CpppAGU		1.51	4.02
UpppAGU		1.45	4.11
GpppGGU	0%	0.63	3.54
GpppCGU		0.63	4.28
GpppUGU		0.68	4.45
GpppAAU	33%	0.61	3.21
GpppACU	0%	0.53	2.40
GpppAUU	19%	0.65	2.89

of the second nucleotide, there is a good agreement between the observed hydrogen bonding distance and the experimentally observed methylation activity, i.e. the distance is shorter the lower the activity is.

3.5.3 Conclusion

Based on molecular dynamics simulations, we have investigated RNA sequence specificity and the effects of mutated RNA cap sequences on the methyltransfer reaction.

We have found that the guanine cap nucleotide interacts strongly with the protein's backbone in the RNA cap pocket. Any mutation of this nucleotide leads to an RNA structure which is unable to form these interactions and therefore, starts to dissociate from the protein structure within 20 ns of molecular dynamics simulations.

In addition, simulations highlight that the identity of the 2nd translated RNA nucleotide must be a guanine since any mutations thereof results in RNA conformations where the methyl acceptor group is positioned in locations unsuitable for the 2'O methylation reaction. Molecular dynamics simulations suggest that in all mutants of this nucleotide an intramolecular hydrogen bond is formed between the 2nd nucleotide and the γ -phosphate group of the triphosphate linker. The only nucleotide where this is not possible is guanine where no hydrogen bond donor is located in the appropriate region. To validate this hypothesis, experimental characterization of mutated RNA structures with non-natural nucleotides should be performed where formation of an the afore mentioned intramolecular hydrogen bond is inhibited.

3.6 Conclusion

Using a diverse set of computational methods, we were able to address a number of open questions concerning the mechanism of the catalyzed methyltransfer reactions.

First, a structure of the enzyme bound to the RNA and the SAM co-factor has been modeled, based on available template structures and published mutagenesis data. The model was validated and structural rearrangements induced by binding of SAM or the RNA were assessed by molecular dynamics simulations. The results suggest that overall, structural rearrangements upon ligand binding are small, especially within the SAM and the RNA binding pocket as well as in the active site. Therefore, those sites can be considered structurally stable targets for structure-based drug discovery efforts.

Second, in order to characterize the underlying chemical reactions, *ab initio* electronic structure calculations were performed on model systems mimicking the biological reactions. Calculations on such model systems reveal that both the 2'O and the N7 reaction are energetically favored processes, where the N7 reaction produces a significantly more stable product and shows a lower activation barrier than the 2'O reaction. Comparison between the catalyzed and uncatalyzed 2'O reaction, revealed the importance of a lysine residue which acts as a proton acceptor and significantly stabilizes the product state and reduces the activation barrier.

Furthermore, an *in-silico* approach was developed to identify the effects of single point mutations on different aspects of the catalyzed reaction. Using this approach in a computational alanine scanning procedure helped to identify protein residue patches which modulate the geometric arrangement between methyl donor and acceptor, the methyl donor binding affinity or the reaction energy barrier. In addition, previously uncharacterized hot-spot residues were identified and analyzed further using computational and experimental methods in order to gain a better understanding of their role in the enzyme's function.

In addition, we have investigated RNA sequence specificity of the enzyme and effects of mutated RNA cap sequences on the methyltransfer reaction. Based on molecular dynamics simulations, protein residues critical for RNA sequence specificity of the enzyme were identified and the possibility for forming an intramolecular hydrogen bond between distinct RNA elements was observed, whose absence might be detrimental for RNA sequence specificity.

Chapter 4

Ligand Binding Site Prediction

The reliable prediction of ligand binding sites is crucial for characterizing proteins with unknown function. Therefore, the use of computational predictions of protein function and ligand binding sites for proteins without experimental structures are assessed in a blind and objective way. Limitations in the current prediction methods are analyzed and suggestions for a more reliable evaluation are given. Following those suggestions, an extended and fully automated assessment is implemented in the Continuous Automated Model EvaluatiOn framework.

4.1 Introduction

In the post-genomic era, the number of protein sequences with unknown structure is constantly growing at an exponential rate. Similarly, the number of protein structures with unknown biological function is steadily increasing. To bridge this rapidly growing gap between known sequences and unknown function, numerous computational and experimental techniques have been developed to help predicting the biological function.^{5,1,6}

Among these methods, computational approaches for identifying the precise location of ligand binding sites and protein residues involved in ligand interaction are of high relevance for life science research, with applications in functional characterization of novel proteins, drug design and enzyme engineering. Various approaches for the prediction of ligand-binding sites have been proposed, based on sequence conservation, geometric criteria of the protein surface, or homology transfer from known structures.⁷

Relevant biological questions, however, can only be addressed if predictions are specific and accurate. Therefore, evaluating the performance of prediction methods in a blind and objective way is crucial. To achieve this goal, ligand binding site prediction methods have been assessed since the 6th edition of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment.^{113,114}

4.1.1 Critical Assessment of Protein Structure Prediction

CASP is a community-wide experiment, with the goal to evaluate and advance the methods for protein structure prediction. CASP is a blind experiment where the predictors receive a set

of protein sequences for which the structure is unknown. When the experimentally determined structures are released, predictions are evaluated by an independent assessor to identify the current state of the art in the protein structure prediction field.

It should be noted that CASP is a double blind experiment where all predictors are given the same question at the same time, ensuring that the same data is available to all predictors. This is important to guarantee a fully independent, objective and comparable evaluation of current methods. The experiment is held biannually.

The main focus of CASP lies in the evaluation of template based modeling techniques (TBM) but it includes a number of additional categories for the evaluation of template free modeling, model refinement, model quality prediction, disorder prediction and function prediction. In the latter category, the prediction of ligand binding sites are assessed. Here, the predictors are given a set of protein sequences with unknown structures and are asked to identify the residues involved in ligand binding.

4.2 Assessment of Ligand Binding Site Prediction in CASP9

In the following, a published manuscript is included:

“Assessment of ligand-binding residue predictions in CASP9”

Assessment of Other Categories of Predictions

Assessment of ligand-binding residue predictions in CASP9

Tobias Schmidt,^{1,2} Jürgen Haas,^{1,2} Tiziano Gallo Cassarino,^{1,2} and Torsten Schwede^{1,2*}

¹ Biozentrum, University of Basel, Switzerland

² SIB Swiss Institute of Bioinformatics, Basel, Switzerland

ABSTRACT

Interactions between proteins and their ligands play central roles in many physiological processes. The structural details for most of these interactions, however, have not yet been characterized experientially. Therefore, various computational tools have been developed to predict the location of binding sites and the amino acid residues interacting with ligands. In this manuscript, we assess the performance of 33 methods participating in the ligand-binding site prediction category in CASP9. The overall accuracy of ligand-binding site predictions in CASP9 appears rather high (average Matthews correlation coefficient of 0.62 for the 10 top performing groups) and compared to previous experiments more groups performed equally well. However, this should be seen in context of a strong bias in the test data toward easy template-based models. Overall, the top performing methods have converged to a similar approach using ligand-binding site inference from related homologous structures, which limits their applicability for difficult *de novo* prediction targets. Here, we present the results of the CASP9 assessment of the ligand-binding site category, discuss examples for successful and challenging prediction targets in CASP9, and finally suggest changes in the format of the experiment to overcome the current limitations of the assessment.

Proteins 2011; 79(Suppl 10):126–136.
© 2011 Wiley-Liss, Inc.

Key words: protein function; protein structure; evaluation; assessment; binding site; active site; co-factor; ligand; CASP.

INTRODUCTION

To perform their functions, proteins interact with a plethora of small molecules within the cell. Most of these interactions are unspecific and transient in nature (e.g., interactions with water and ions), some are persistent and may play a structural or functional role (e.g., certain metal ions), and others might be transient but nevertheless highly specific, often resulting in essential changes of the protein or the ligand (e.g., enzyme-substrate complexes or receptor-ligand complexes). Hence, the identification of a protein's functionally important residues, such as ligand-binding sites or catalytic active residues, is a crucial step toward the goal of understanding the protein's molecular function and its biological role in the cell. Although protein ligand interactions are crucial for the function of a protein, in many cases they are unknown. Although the kind of ligands interacting with a protein is often known from biochemical analyses, elucidating the structural details of these interactions requires elaborate and time-consuming studies by X-ray crystallography or NMR. Therefore, computational tools have been developed aiming at predicting the precise location of binding sites, and specifically which amino acid residues in a protein are directly interacting with ligands. Various approaches for the prediction of ligand-binding sites have been proposed,¹ both from structure and

The authors state no conflict of interest.

Additional Supporting Information may be found in the online version of this article.

Abbreviations: FM, free modeling; MCC, Matthews' correlation coefficient; TBM, template-based modeling.

Grant sponsor: SIB Swiss Institute of Bioinformatics; Grant sponsor: National Institute of General Medical Sciences; Grant number: U01 GM093324-01.

*Correspondence to: Torsten Schwede, Biozentrum University of Basel, SIB Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland. E-mail: torsten.schwede@unibas.ch

Received 23 May 2011; Revised 29 July 2011; Accepted 4 August 2011

Published online 12 September 2011 in Wiley Online Library (wileyonlinelibrary.com).

DOI: 10.1002/prot.23174

from sequence, based on sequence conservation,^{2–7} geometric criteria of the protein surface,^{8–12} or homology transfer from known structures.^{13–17}

The function prediction category (FN) was introduced in the 6th Critical Assessment of Protein Structure Prediction (CASP), where predictions for Gene Ontology molecular function terms, Enzyme Commission numbers, and ligand-binding site residues were evaluated.^{18,19} Because very little new functional information becomes available during and after the experiment, the first two categories were difficult to assess. Therefore, since CASP8, the prediction task has been to identify functionally important residues such as ligand-binding residues or catalytic residues.²⁰ Here, we present the assessment of 33 groups participating in the recent CASP9 experiment. In the ligand-binding site prediction category (FN), the sequence of a protein with unknown structure was provided to predictors. The task was to predict the residues directly involved in ligand binding in the experimental control structure. This approach differs significantly from typical ligand-binding studies (like docking or virtual screening), where the chemical identity of the ligand is given, and the correct geometric orientation of the molecule in the receptor protein is to be determined.^{21–25} In CASP, however, the chemical identity of the ligand is unknown at the time of prediction, and only the interacting residues are predicted.

In summary, all top performing groups have applied a similar approach, using ligand information derived from homologous structures in the PDB.²⁶ In comparison with CASP8,²⁰ we could not observe a significant progress by the top groups, but rather a larger number of groups performing at the same level. We believe that this observation is caused on one side by the bias in the data set to “easy” template-based predictions with only a very small number of difficult de novo targets in recent rounds of CASP. This gives strong advantage to methods using PDB information directly, but discourages the development of methods addressing the more challenging de novo cases. Another limiting factor is the binary format of the prediction task, which does not allow specifying probabilities for specific residues or differentiating between types of ligands.

MATERIALS AND METHODS

Prediction targets

All CASP9 target structures were analyzed for nonsolvent, nonpeptidic ligand groups in the deposited protein structures. Based on literature information, UniProt²⁷ annotations, structures of closely related homologues (Table SI, Supporting Information), and conservation of functionally important residues, we aimed at identifying ligands with biological/functional relevance for the specific protein. All targets, including those containing ligands classified as “nonbiologically relevant,” were further analyzed to identify cases where a ligand clearly

mimicked the interactions of known biologically relevant ligands for this target.

Binding site definition

For each prediction target, binding site residues were defined as those residues in direct contact with the ligand in the target structure, that is, all protein residues with at least one heavy atom within a certain distance from any heavy atom of the ligand. The distance cutoff was defined by the CASP organizers as the sum of the van der Waals radii of the involved atoms plus a tolerance of 0.5 Å. In addition, different tolerance values ranging from 0 to 2.0 Å were evaluated.

In cases where multiple chains with bound ligands were present in the target structure (e.g., homo-oligomeric assemblies), the definition of the binding site residues for individual chains were combined into a single binding site definition. For targets where ligands were observed to bind in the interface between multiple chains, the oligomeric structure as defined by the authors and PISA²⁸ (five cases) or only PISA (1 case) was used for the binding site definition. Analysis of structures and ligand-binding sites was performed using OpenStructure (version 1.1).²⁹

For targets in which only part of the relevant ligand was present, the binding site definition was extended to include the entire biologically relevant ligand. In these cases, two separate evaluations of the prediction performance were conducted. First, denoted as “extended binding site,” all atoms of the partial and the extended ligand were used to define the binding site in the same way as described earlier. Second, denoted as “partial binding site,” only atoms of the partial ligand were used to define the binding site, whereas all residues exclusively in contact with the extended part of the ligand were treated as neutral and excluded from the evaluation.

Binding site prediction evaluation

As in the previous assessment,²⁰ binding site prediction performance was measured using the Matthews correlation coefficient³⁰ (MCC), which accounts both for over and under predictions. For each target, residue predictions were classified as true positives (TP: correctly predicted binding site residues), true negatives (TN: correctly predicted nonbinding site residues), false negatives (FN: incorrectly under predicted binding site residues), and false positives (FP: incorrectly over predicted nonbinding site residues) based on the binding site definition described before. The MCC was computed using Eq. (1):

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}}$$

MCC ranges from +1 (perfect prediction), over 0 (random prediction) to –1 (inverse prediction). Empty submissions that did not include any binding site predictions and missing predictions were assigned a MCC score of zero.

To reduce the effects of target difficulty on the ranking, MCC scores were standardized by computing Z scores among all predictions P for a given target T using Eq. (2):

$$z_{P,T} = \frac{MCC_{P,T} - \overline{MCC}_T}{\sigma_T}$$

In this equation, $MCC_{P,T}$ is the raw MCC score for target T given by predictor P , \overline{MCC}_T is the mean MCC score for target T , and σ_T is the standard deviation of MCC scores for target T . The overall performance for each predictor was computed as the mean of Z scores over all targets, which was subsequently used for obtaining a final ranking. In addition to the MCC score, we computed the recently published binding site distance test (BDT).³¹ BDT takes the actual three-dimensional locations of the predicted residues into account and scores residues differently, according to the distance between the predicted and the observed binding site. Predictions close to the binding site score higher than more distant predictions. The BDT score ranges from 0, for a random prediction to 1, for a perfect prediction.

Robustness and significance

Statistical significance of the ranking and robustness with regard to composition of the target data set was assessed using two different methods. First, two-tailed Student's paired t -tests as well as Wilcoxon signed rank tests³² between all predictor groups were performed based on MCC scores for each target. Both t -tests and Wilcoxon signed rank tests were performed using R (version 2.11.1).³³ Second, bootstrapping was performed, where scores were computed on a randomly selected subset of three-fourth of all targets (i.e., 23 of 30 targets). Seventy-five rounds of bootstrapping were executed for different target subsets, and for each bootstrapping experiment, mean, minimum, and maximum Z scores per group were calculated as previously described. Additionally, the rank for each prediction group was calculated, and mean, minimum, and maximum ranks over all bootstrapping experiments were computed.

To assess the performance of groups on different types of ligands, we have analyzed the prediction performance separately on targets including only metal ions (10 targets) and on targets including only nonmetal ligands (17 targets). Mixed targets including both metal and nonmetal ligands (three targets) were not considered in this subanalysis.

RESULTS AND DISCUSSION

Overall performance

In the CASP9 protein-binding sites prediction category (FN), the predictors were given a protein sequence with unknown structure and asked to identify the residues involved in ligand binding. According to the CASP format,

the predictions were binary, and, thus, classified each residue as either binding-site or nonbinding-site residue. As defined by the organizers, only protein-small molecule interactions were considered in this category. The assessment of this category consisted of the following three steps: (1) identification of biologically relevant ligands in the target structures, (2) definition of binding site residues, and (3) assessment of the prediction performance.

One dominant factor in assessing the correctness of ligand-binding site prediction is the availability of experimental data and the evaluation of the biological relevance of the specific ligand binding. Whether a certain ligand is observed in an experimental structure is first and foremost determined by the specific purification procedure, by the experimentalist's choice of using this ligand for a co-crystallization experiment, and the specific experimental conditions (ligand concentration, pH and buffer conditions, ionic strength, precipitant, etc.). If a ligand is not observed in a specific experimental structure, it could still bind under different conditions, that is, it cannot be considered as a "true negative" data point for the assessment. On the other hand, if a certain ligand is observed in a target structure, we can classify the residues within this structure into "binding" and "nonbinding" with regard to this specific ligand. Note that a target protein might be able to bind different ligands under different experimental conditions, and only a subset of them might be present in the target structure at hand. For example, the structure of an enzyme might be crystallized in complex with the cofactor, but without substrate or product molecules.

Although the identification of ligands in CASP9 was based only on experimentally observed ligands, it was still not straightforward to categorize their biological relevance. Although in 73% of the target structures in CASP9, various ligands were present, most of them were not considered biologically relevant but rather as originating, for example, from solvent, crystallization precipitant, or buffers. For the assessment, however, we included only ligands that we considered to be biologically relevant. The decision on biological relevance was done by manual curation, primarily based on the type and location of the ligand, literature information, and UniProt²⁷ annotations. In addition, information from structurally closely related homologues and conservation of functionally important residues was used to guide the selection process. Using this approach, 16 target structures with biologically relevant ligands were selected of the 109 targets available in CASP9 for the assessment.

In addition, we have analyzed all remaining heteroatomic groups, if they occupied binding sites that mimicked the interactions of a known biologically relevant ligand for this protein. In these cases, we defined an "extended binding site" consisting of all residues in contact with the known biologically relevant ligand. We were careful to include only targets where the assignment was unambiguous in order to avoid the inclusion of false bind-

Table I

Summary of CASP9 Targets with Bound Ligands

Target	PDB	Partial ligand	Extended ligand	Chemical class	Interface	CASP category
T0515	3MT1	S04	PLP, LYS	Nonmetal	A–B	TBM
T0516	3N06	IMD	PF1	Nonmetal		TBM
T0518	3NMB	NA		Metal		TBM
T0521	3MSE	CA, CA		Metal		TBM
T0524	3MWX	GOL	GAL	Nonmetal		TBM
T0526	3NRE	PEG	GLA	Nonmetal		TBM
T0529	3MWT	MN		Metal		TBM
T0539	2L0B	ZN, ZN		Metal		TBM
T0547	3NZP	PLP	PLP, LYS	Nonmetal	A–B	TBM
T0548	3NNQ	ZN		Metal		TBM
T0565	3NPF	CSA	DGL, ALA	Nonmetal		TBM
T0570	3N03	Mg, GOL		Metal, nonmetal		TBM
T0582	3O14	ZN		Metal		TBM
T0584	3NF2	S04	DST, IPR	Nonmetal		TBM
T0585	3NE8	ZN		Metal		TBM
T0591	3NRA	LLP		Nonmetal	A–B	TBM
T0597	3NIE	ANP		Nonmetal		TBM
T0599	3OS6	S04	ISC	Nonmetal		TBM
T0604	3NLC	FAD		Nonmetal		TBM / FM
T0607	3PFE	ZN	ZN, BES	Metal, nonmetal		TBM
T0609	3OS7	TLA	GAL	Nonmetal		TBM
T0613	3OBI	EDO	GAR, NHS	Nonmetal		TBM
T0615	3NQW	MN, S04	MN, GPX	Metal, nonmetal		TBM
T0622	3NKL	S04	NAD	Nonmetal		TBM
T0625	3ORU	ZN		Metal		TBM
T0629	2XGF	FE, FE, FE, FE, FE, FE, FE		Metal	A–B–C	FM
T0632	3NWZ	COA		Nonmetal	A–B–C	TBM
T0635	3N1U	CA		Metal		TBM
T0636	3P1T	TLA	HSA, PLP	Nonmetal	A–B	TBM
T0641	3NYI	STE		Nonmetal		TBM

ing site definitions. Using this approach, the number of target structures in the FN category was extended by 14, yielding a total of 30 targets in this category (Table I).

Within the selected targets, 10 were found in complex with metal ions (Ca, Fe, Mg, Mn, Na, and Zn) and further 17 targets in complex with nonmetal ligands (Table I). The latter included amino acids and derivatives, nucleotides, sugars, fatty acids, and others. Additionally, in three cases, nonmetal ligands were coordinated to metal ions (Mg, Mn, and Zn). In most of the targets, the ligand-binding site was located within a monomer, while, for six targets, the ligand was bound in the interface between multiple chains: T0515, T0547, T0591, T0636 (dimeric structures), T0629 (trimeric structure), and T0632 (tetrameric structure). The ligands were bound between all chains of the oligomeric structure, except for T0632 where the ligand is bound to only three of the four chains.

Following the identification of biologically relevant ligands, the binding site residues for those targets were defined as those residues directly in contact with the ligand. Atoms were considered to be in contact if they were within a distance of the sum of their van der Waals radii plus a tolerance distance. The list of binding site residues used in the assessment for each target is provided in Table SI (Supporting Information). The tolerance distance was defined as 0.5 Å by the CASP organizers. We tested the influence of different values for the tol-

erance distance of the binding site definition and their influence on the assessment of prediction performance. No significant differences in the overall prediction performances were observed for different tolerance distances (Fig. S1, Supporting Information).

The majority of FN targets in CASP9 were classified as template-based modeling targets (TBM), and only two targets were free modeling (FM) targets: (1) target T0629, where the ligand binding domain had no template structure Figure 8(C), (2) target T0604, where the ligand was bound between two domains where one was a template-based modeling (constituting 90% of the binding site residues) and one a free modeling domain (constituting 10% of the binding site residues). This strong bias in the data set has direct consequences for the assessment, as it is to be expected that template-based prediction methods will perform much better than de novo methods in this context.

In total, 33 groups submitted predictions in the CASP9 FN category. A summary of the predictions is given in Figure 1. Among the participating groups, 18 were registered as “human predictors” and 15 as “servers” (Table II). Most groups predicted at least 25 of the assessed 30 targets, that is, 12 groups (6 humans and 6 servers) predicted between 25 and 29 of the assessed targets and 15 groups (6 humans and 9 servers) predicted all 30 targets; six human groups returned predictions for only six or less targets. Binding site prediction perform-

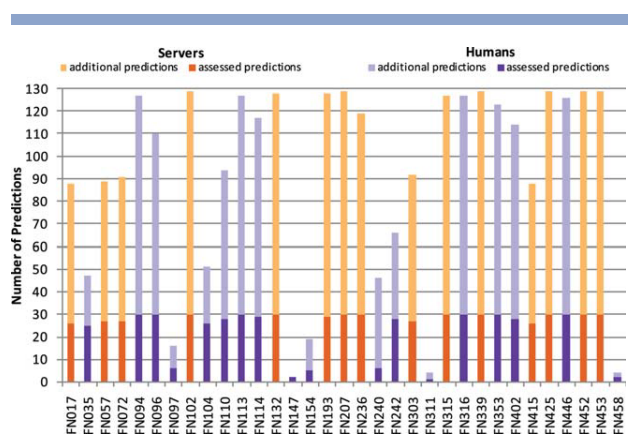


Figure 1

Overview of predictions per group. Predictions for targets which were assessed in the FN category (i.e., targets with a relevant binding site) are displayed in dark colors, additional predictions, which were not assessed (i.e., targets without an experimentally confirmed binding site) are displayed in light colors. Human groups are shown in purple, servers in orange.

ance was measured using Z scores of Matthews correlation coefficients (see Methods section).^{**} The comparison between all groups is shown in Figure 2, where the error bars indicate minimum and maximum Z scores obtained by bootstrapping on a randomly selected subset of three-fourth of the targets. The error bars indicate a fluctuation in the average Z score for each group. However, in case of a correlated movement in the score, this would not influence the groups ranking. Therefore, the rank for each prediction group was computed in each bootstrapping experiment and the average, minimum, and maximum rank over all bootstrapping experiments is shown in Figure 3.

The top 12 predictors clearly distinguished themselves from the following 21 groups and show a significantly better performance. Two predictors from the Zhang group (FN096, Zhang and FN339, I-TASSER_FUNCTION) show a better performance in terms of MCC compared to the following 10 participants, whereas the performance among the latter group is comparable. Because many predictors seemed to perform similarly, statistical tests were used to assess the significance of the differences between these groups. Paired t -tests on all targets between all pairs of predictors were performed. The results are shown in Table III, with cells shaded according to computed P values. According to the t -test, the differences between the top

^{**}As described in Materials and Methods section, the authors decided that assigning a MCC score of zero to empty submissions, which did not include any binding site predictions and to missing predictions would most appropriately reflect a “real life” prediction situation in the assessment. Please note that this policy has consequences for the final ranking as it penalizes methods, which are not able to make predictions for some targets, and encourages the more risky development of novel methods as there is no implicit penalty for making predictions for challenging targets.

Table II
Groups Participating in the FN Category in CASP9

ID	Rank	Name	Type	Group
FN017	22	3DLIGANDSITE1	S	Michael Sternberg
FN035	5	CNIO-FIRESTAR	H	Gonzalo Lopez
FN057	21	3DLIGANDSITE3	S	Michael Sternberg
FN072	23	3DLIGANDSITE4	S	Michael Sternberg
FN094	8	MCGUFFIN	H	Liam McGuffin
FN096	1	ZHANG	H	Yang Zhang
FN097	30	KOCHANCZYK	H	Marek Kochanczyk
FN102	15	BILAB-ENABLE	S	Shugo Nakamura
FN104	7	JONES-UCL	H	David Jones
FN110	6	STERNBERG	H	Michael Sternberg
FN113	9	FAMSSEC	H	Katsuichiro Komatsu
FN114	10	LEE	H	Jooyoung Lee
FN132	27	MN-FOLD	S	Chris Kauffman
FN147	28	GENESILICO	H	Janusz Bujnicki
FN154	33	JAMMING	H	Gabriel del Rio
FN193	24	MASON	S	Huzefa Rangwala
FN207	26	ATOME2_CBS	S	Jean-Luc Pons
FN236	12	GWS	S	Jooyoung Lee
FN240	32	TMD3D	H	Hiroshi Tanaka
FN242	4	SEOK	H	Chaok Seok
FN303	20	FINDSITE-DBDT	S	Jeffrey Skolnick
FN311	31	ALADEGAP	H	Kei Yura
FN315	3	FIRESTAR	S	Gonzalo Lopez
FN316	18	LOVELL_GROUP	H	Simon Lovell
FN339	2	I-TASSER_FUNCTION	S	Yang Zhang
FN353	17	SAMUDRALA	H	Ram Samudrala
FN402	13	TASSER	H	Jeffrey Skolnick
FN415	25	3DLIGANDSITE2	S	Michael Sternberg
FN425	19	INTFOLD-FN	S	Liam McGuffin
FN446	16	KIHARALAB	H	Daisuke Kihara
FN452	11	SEOK-SERVER	S	Chaok Seok
FN453	14	HHPREDA	S	Johannes Soeding
FN458	29	BILAB-SOLO	H	Mizuki Morita

ranked group (FN096, Zhang) and groups FN339 (I-TASSER_FUNCTION), FN242 (Seok), and FN035 (CNIO-Firestar) are not statically significant, while the differences between FN096 and the remaining predictors are significant. In addition, the nonparametric Wilcoxon signed rank test was performed, which yielded comparable results to the t -tests (Table SII, Supporting Information).

Recently, McGuffin and coworkers published an alternative binding site distance test (BDT).³¹ Opposed to MCC, BDT takes the actual three-dimensional positions of the predicted residues into account and scores residues differently, according to the distance between the predicted and the observed binding site. Hence, BDT limits the boundary effects originating from ambiguous definition of binding sites. When applying the BDT score on the predictions (Fig. S2, Supporting Information), for the top ranked groups, no significant deviations to the MCC-based prediction assessment were observed.[†]

As described earlier, for 14 targets, the partial binding sites were individually extended around the observed ligand to reflect a binding site accommodating the most

[†]The largest change in ranking by three positions would be for group FN110.

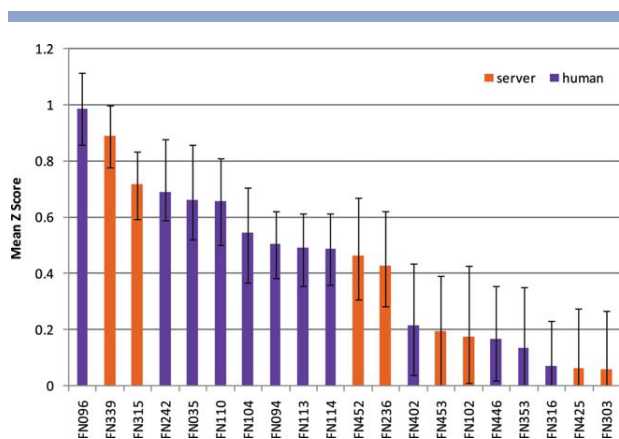


Figure 2

Mean Z scores over all targets for the top 20 predictor groups. Error bars show minimum and maximum average Z scores obtained from bootstrapping experiment. Human predictor groups are shown in purple, servers in orange.

probable biologically relevant ligand. To investigate the influence of this extension, the assessment was performed both on all residues of the extended binding site and separately on all the residues of the partial binding site while treating the residues exclusively in the extended binding site as “neutral” for the analysis. For the top-ranked groups, no significant differences in the overall prediction performances were observed between partial and extended binding site definitions (Fig. S3, Supporting Information).^{††}

Assessment by type of binding sites

In addition to the overall performance, subsets of the targets were evaluated individually, according to the ligand’s chemotype. The distinct chemical properties of metal ions and organic ligands give rise to diverse binding sites. Thus, it could be expected that various prediction methods perform differently. To address this question, we have analyzed the prediction performance separately on all targets including only metal ligands (10 targets) and on targets including only nonmetal ligands (17 targets). The mean Z score per group separated into metal and nonmetal targets are shown in Figure 4. Within the top 10 groups, most of them show a better performance for nonmetal targets, with the exception of FN242 (Seok) and FN114 (Lee). Especially group FN114 shows a better performance on metal ligands, compared to an average performance on the full set of targets.

Among the CASP9 FN targets, in six cases, the ligand binds in the interface between multiple chains of an oligomeric protein complex. Although, the number of inter-

face targets is very limited, we were interested in the question if the prediction of ligand-binding sites of interface targets is more difficult than noninterface targets. We compared the average prediction performance, both according to mean MCC values, as well as the number of very good predictions ($MCC > 0.85$), for interface versus noninterface targets. No significant difference was observed; thus, on average, in those target categories, it seems equally difficult to predict the binding site residues. However, it should be considered that four of the six targets are “trivial” oligomers, where a simple blast query returns a homologues template-ligand complex with the correct oligomeric state.

Human versus server predictions

Looking at the top 10 groups, 8 of them were registered as “humans,” and only 2 as “servers.” Overall, there is a striking difference between the average performance of human groups and server groups with a mean Z score of 0.47 and 0.15, respectively. Although predictor groups registered as “human” performed considerably better than “servers,” the role of human beings in the prediction process was difficult to evaluate. Several aspects seemed to contribute to this observation: human predictors had access to multiple servers for structure modeling and various server binding site predictions, while server predictors have to rely on their own method only. While human predictors can make use of additional annotation from biological knowledge bases and scientific literature, servers have to rely on structured machine-readable information. A major bottleneck in this context seems the lack of consistent annotation of ligands found in PDB entries with respect to their biological relevance. It

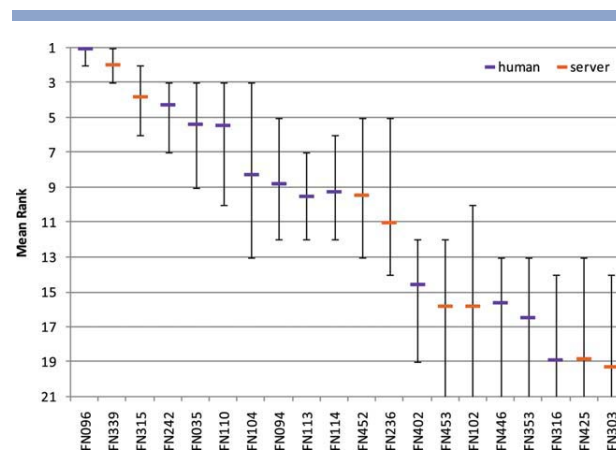


Figure 3

Mean rank based on bootstrapping experiment for the top 20 predictor groups. Error bars show minimum and maximum rank obtained from bootstrapping experiment. Human predictors are shown in purple, servers in orange.

^{††}The largest difference was observed for group FN113, which would change rank by three positions.

Table III
P Values Computed by Paired *t*-Test of All Against All Predictors

	FN096	FN339	FN315	FN242	FN035	FN110	FN104	FN094	FN113	FN114	FN452	FN236
FN096	—	0.24	0.01	0.08	0.06	0.01	0.01	0.00	0.00	0.00	0.00	0.00
FN339	0.24	—	0.27	0.20	0.28	0.20	0.05	0.04	0.02	0.05	0.02	0.02
FN315	0.01	0.27	—	0.81	0.56	0.63	0.17	0.20	0.03	0.14	0.12	0.07
FN242	0.08	0.20	0.81	—	0.85	0.90	0.31	0.28	0.27	0.19	0.10	0.09
FN035	0.06	0.28	0.56	0.85	—	0.88	0.44	0.52	0.38	0.45	0.45	0.31
FN110	0.01	0.20	0.63	0.90	0.88	—	0.33	0.28	0.27	0.30	0.33	0.18
FN104	0.01	0.05	0.17	0.31	0.44	0.33	—	0.88	0.88	0.89	0.93	0.93
FN094	0.00	0.04	0.20	0.28	0.52	0.28	0.88	—	0.99	0.98	0.94	0.79
FN113	0.00	0.02	0.03	0.27	0.38	0.27	0.88	0.99	—	0.99	0.95	0.76
FN114	0.00	0.05	0.14	0.19	0.45	0.30	0.89	0.98	0.99	—	0.96	0.56
FN452	0.00	0.02	0.12	0.10	0.45	0.33	0.93	0.94	0.95	0.96	—	0.83
FN236	0.00	0.02	0.07	0.09	0.31	0.18	0.93	0.79	0.76	0.56	0.83	—

Significant differences between two groups are indicated by cells with white background. For clarity, only the 12 top performing predictors are shown, sorted by their overall performance.

appears that human predictors benefit from the longer prediction time mainly by their ability to distinguish relevant from irrelevant ligand predictions.

Prediction methods have converged to a similar approach

When comparing the methods of the top performing groups, it seems that they have converged to similar approaches, which are based on homology transfer from related structures in the PDB. By identifying homologous protein structures with bound ligands, putative binding site residues in the target model are classified by spatial proximity after alignment or superposition. The methods differ in their specific implementations with regards to the underlying structure databases (PDB vs. curated binding site libraries), target representation (alignment to structure vs. full atomic models), superposition to related structures to identify putative binding sites, and the use of residue conservation information in the prediction process. The major drawback of these homology-based inference methods is that they rely on the availability of related protein structures with bound ligands and are thus unable to make predictions for novel proteins without prior ligand information.

Although many groups have used similar approaches to make their predictions, we observed a surprising heterogeneity of performance within targets. As shown in Figure 5 (and Figure S4), the 12 top performing groups show overall a similar spectrum of results, with a few nearly perfectly predicted targets and some poorly predicted targets. Interestingly, when analyzing the results for individual targets, at least one good prediction was achieved across all groups (MCC value of at least 0.56; on average 0.84; see Fig. 6), and even predictors with a poor overall performance, can yield the best individual prediction for certain targets, as shown in Figure 7. Thus, either the performance of the different methods is highly target specific, or there is a considerable random

component in the prediction process in combination with a strong influence by the small and biased target data set.

Prediction examples

Obviously, target T0604 was the most difficult target in the FN category in CASP9, with a maximum MCC score of 0.56 for the best prediction, and an average score of 0.29. The protein is a putative FAD-dependent oxidoreductase with a bound FAD molecule (PDB : 3nlc). The protein is monomeric and forms a large binding pocket for the ligand. The structure is shown in Figure 8(A) together with the binding site predictions of group FN035 (CNIO-FIRESTAR) as one of the best predictions for this target. The top performing methods were able to accurately predict the lower part of the binding site around the adenine moiety, whereas all of them failed for the upper part of the binding site around the flavin moiety.

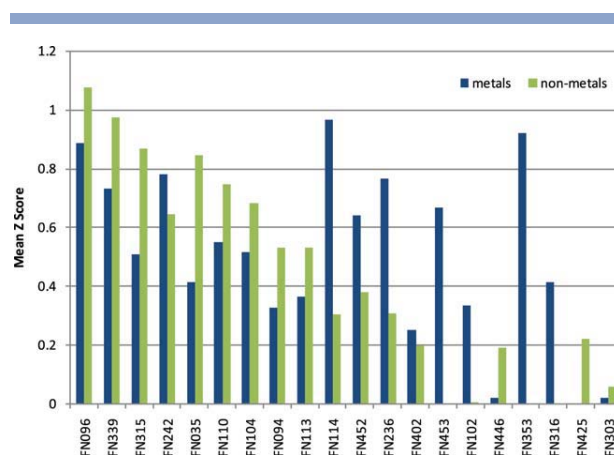
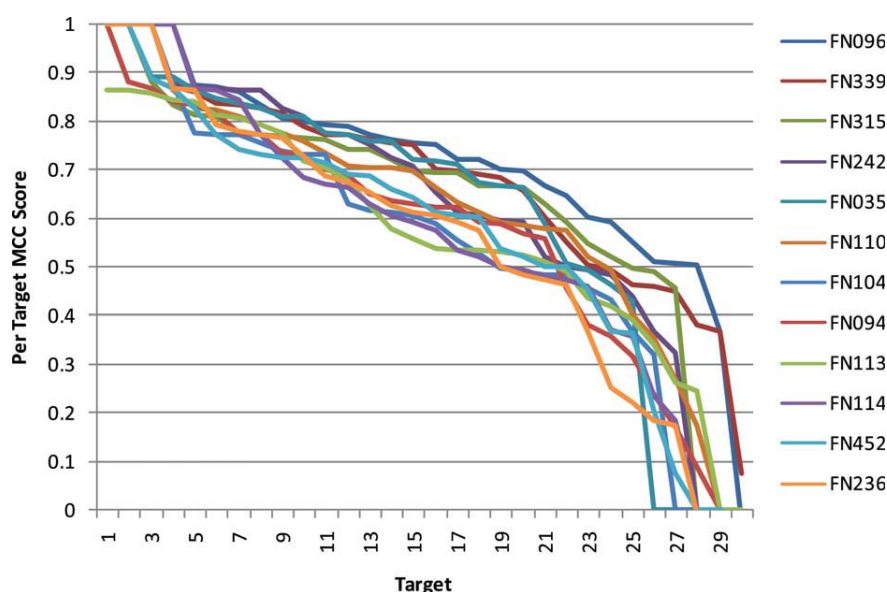


Figure 4

Mean Z scores of the top 20 groups, separated by the ligand's chemotype. Metals are shown in blue, nonmetals are shown in green.

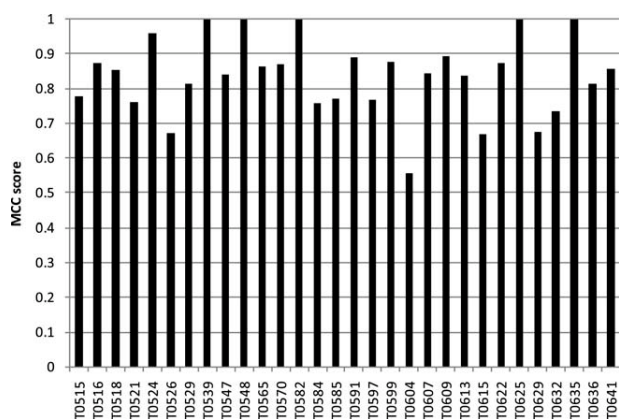

Figure 5

MCC scores for the 12 top performing groups for all targets. Targets were sorted by their respective MCC score, individually for each group.

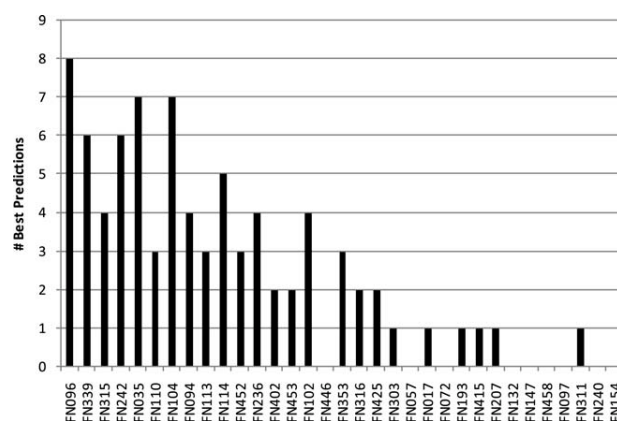
This stems from the fact that this target structure has only remote homologues, which differ significantly in the flavin binding site region. This example clearly demonstrates the limitations of prediction methods that are based on homology transfer.

Target T0629 is the only target in the current ligand-binding target set, which was classified as free modeling target and thus has no template structure. The protein (PDB: 2xgf) is the bacteriophage T4 long tail fiber receptor-binding tip. It contains a long fiber like structure, which is formed by three chains and binds seven iron

atoms. Each iron atom is complexed by six histidine residues. Each protein chain contributes two histidines to each binding site, where the two histidines are in a His-X-His motive, with X being either Ser, Thr, or Gly. The target structure is shown in Figure 8(C) together with the binding site predictions of group FN114 (Lee), the best predictor for this target among the top 10. Common to all predictions for this target is that they correctly predicted a subset of the seven binding sites—most likely due to local similarity to another metal binding protein


Figure 6

Overall target difficulty. MCC value of the best overall prediction for each target.


Figure 7

Number of targets where a particular group returned the best prediction. Groups are sorted by their overall performance. For one target, multiple groups can perform equally.

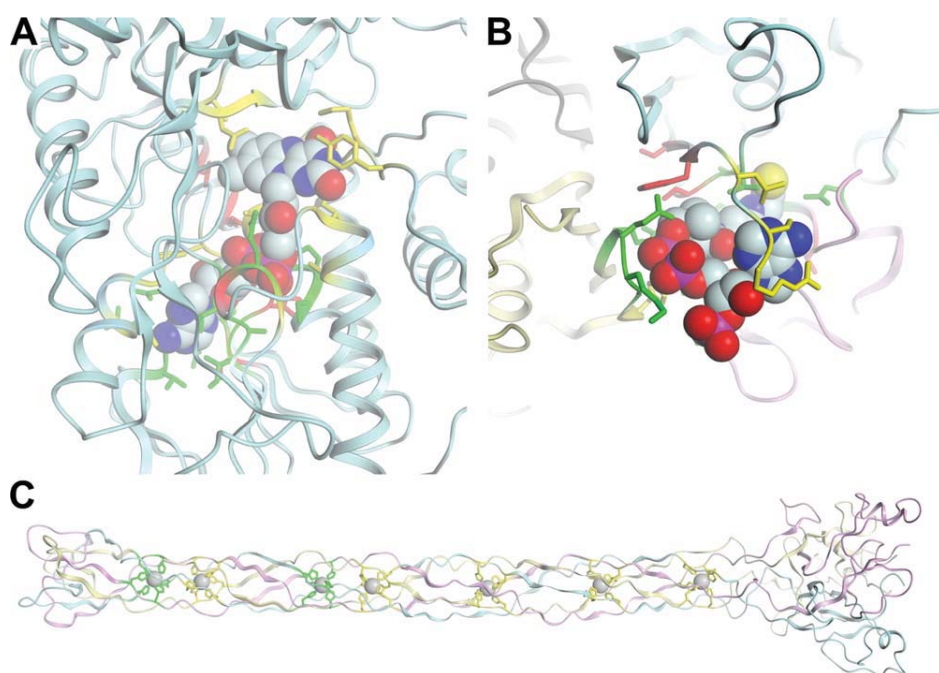


Figure 8

Examples of binding site predictions. All ligands are shown in spheres render mode. The protein backbone is shown in cartoon mode with each chain colored separately. All side chains of observed and predicted binding site residues are shown in licorice sticks. Correctly predicted residues (true positives) are colored in green, incorrectly under predicted binding site residues (false negatives) in yellow and incorrectly over predicted nonbinding site residues (false positives) in red. A: Target T0604 with predictions of group FN035. B: Predictions of group FN096 for target T0632. C: Group FN114's predictions for target T0629.

with a His-X-His motif - but no predictor identified all sites correctly.

The structure of target T0632 (PDB : 3nwz) is a homo-tetramer, which binds coenzyme-A. This ligand is interacting with three of the four chains of the protein, which seems to present a challenge for binding site residue prediction observed by a low average MCC of 0.22. An excellent prediction was obtained by group FN096 (Zhang) with an MCC of 0.72, which is depicted in Figure 8(B) along with the target structure. Many residues were well predicted despite originating from different chains. In this prediction, the largest errors originate from missing some binding site residues due to an elongated terminus compared to structurally closely related templates.

CONCLUSION

The task of predicting binding sites from a protein's sequence is of high relevance for life science research, ranging from functional characterization of novel proteins to applications in drug design, and consequently the ligand-binding site prediction category in CASP has received increasing attention over the past years. In CASP9, it attracted a total of 33 predictors—10 more

groups than in CASP8. In contrast to the previous CASPs, where only three predictors yielded reliable predictions,²⁰ in this assessment, nearly half of the prediction groups yielded reliable predictions for the majority of targets. Two groups (FN096, Zhang; FN339, I-TASSER_FUNCTION) performed better than the rest (when accounting for missing target predictions in the assessment), while the following ten prediction groups performed comparably well. This is not very surprising with respect to the observation that in this round all top performing groups based their methods on approaches, which are similar to the best performing strategy in previous CASP experiments (i.e., Sternberg³⁴ and Lee¹⁵).

Limitations of the current format and recommendations for future experiments

The very low number of target structures with relevant ligands is a major limitation to the assessment as it does not allow to draw significant conclusions on the specific strengths and weakness of different prediction methods, for example, with regard to target difficulty or type of the ligands. Only 30 of the total 109 CASP9 targets (28%) were considered to have a biologically relevant ligand bound in the target structures and were thus assessed in the FN category. It is likely that some of the

remaining target proteins would bind interesting ligands under different experimental conditions, but such conclusions can not be made with the available data. In the previous CASP8 experiment, the total number of targets in this category was 27, illustrating that this is a recurring problem—and not specific to this round of CASP. Another rather drastic limitation of the FN category is the binary prediction format that classifies residues as either ligand binding/nonbinding based on a hard distance cutoff. Consequently, all ligands are currently treated uniformly, independent of their chemical type, and all potential binding sites are treated uniformly, independent of their affinity (or binding probability) for different ligands. Moreover, most targets in the FN category were straightforward TBM targets with numerous, closely related template structures, and only one of the 30 targets was categorized as free modeling (FM). However, exactly this class of target structures is of highest interest for computational ligand-binding site prediction, where no obvious information about the location of their binding sites is available. We suggest the following modifications to the assessment of ligand-binding site predictions to enable the community to benefit even further from future rounds of this experiment:

- In order to accumulate a sufficiently large number of prediction targets, the assessment of this category should be done continuously based on a weekly PDB prerelease. This would allow assessing the performance in different ranges of target difficulty, similar to other CASP categories and facilitate analyzing the strengths and weakness of different approaches. During the CASP meeting in Asilomar, we have suggested that the CAMEO project (Continuous Automated Model EvaluatiOn) of the Protein Model Portal³⁵ could contribute to this effort.
- Binding sites differ chemically and structurally from each other, for example, a metal ion binding site has different characteristics compared to, for example, a sugar-binding site. We therefore suggest that the assessment of binding site residue predictions should be made according to chemotype categories of the ligand expected to be bound. We propose the following categories: “metal ions” (e.g., Na, Ca, Zn, Fe, Mn, and Mg), “inorganic anions” (e.g., SO₄ and PO₄), “DNA/RNA” for poly-ribonucleic acid binding sites, and “organic ligands” for cofactors, substrates, and receptor agonists/antagonists (e.g., NAD, FAD, ATP, SAM, CoA, and PLP). More fine-grained assessment categories might be necessary if more specific prediction methods emerge in the future.
- The binary prediction of binding site residues should be replaced by a continuous probability measure, thus reflecting the likelihood for a residue to be involved in binding a ligand of a certain type. For example, a certain residue might be predicted as

having a high probability to bind a metal ion, but a low probability to bind an organic ligand. The assessment of continuous prediction variable (e.g., using ROC type analysis) would better reflect the spectrum of “high affinity” and “low affinity” sites of different types.

- The experimentalist solving a protein structure typically will have more insights and experimental evidence for the biological role and relevance of ligands observed in a protein structure than the information, which is publicly available to assessors during the CASP experiment. It would therefore be beneficial to capture the information about the biological role of “HETATM” records during PDB deposition.

Predicting binding sites from a protein’s sequence has the potential for yielding high impact on life science research—if the predictions are specific and accurate enough to help addressing relevant biological questions. We hope that with the suggested modifications, the assessment of ligand-binding site predictions will be more suited to evaluate the current state of the art of prediction methods, identify possible bottlenecks, and further stimulate the development of new methods.

ACKNOWLEDGMENTS

The authors thank the experimental groups for providing the target structures for the CASP9 experiments and all predictors for their participation. We are especially grateful to Mike Sternberg and Johannes Söding for fruitful discussions on ligand-binding site prediction and assessment.

REFERENCES

1. Gherardini PF, Helmer-Citterich M. Structure-based function prediction: approaches and applications. *Brief Funct Genomic Proteomic* 2008;7:291–302.
2. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 2004;20:1322–1324.
3. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. *Nat Struct Biol* 1995;2:171–178.
4. del Sol A, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. *J Mol Biol* 2003;326:1289–1302.
5. Fischer JD, Mayer CE, Soding J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* 2008;24:613–620.
6. Innis CA. siteFiNDER!3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res* 2007;35(Web Server issue):W489–W494.
7. Pazos F, Rausell A, Valencia A. Phylogeny-independent detection of functional residues. *Bioinformatics* 2006;22:1440–1448.
8. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM. A method for localizing ligand binding pockets in protein structures. *Proteins* 2006;62:479–488.
9. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 1997;15:359–363,389.

10. Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res* 2009;37(Web Server issue):W413–W416.
11. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* 2006;6:19.
12. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 1995;13:323–330,307–328.
13. Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci USA* 2008;105:129–134.
14. Lopez G, Valencia A, Tress ML. Firestar—prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res* 2007;35:W573–W577.
15. Oh M, Joo K, Lee J. Protein-binding site prediction based on three-dimensional protein modeling. *Prot-Struct Funct Bioinform* 2009;77:152–156.
16. Pandit SB, Brylinski M, Zhou H, Gao M, Arakaki AK, Skolnick J. PSiFR: an integrated resource for prediction of protein structure and function. *Bioinformatics* 2010;26:687–688.
17. Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res* 2010;38:W469–W473.
18. Soro S, Tramontano A. The prediction of protein function at CASP6. *Proteins* 2005;61(Suppl 7):201–213.
19. Lopez G, Rojas A, Tress M, Valencia A. Assessment of predictions submitted for the CASP7 function prediction category. *Prot-Struct Funct Bioinform* 2007;69:165–174.
20. Lopez G, Ezkurdia I, Tress ML. Assessment of ligand binding residue predictions in CASP8. *Proteins* 2009;77(Suppl 9):138–146.
21. Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. *J Med Chem* 2006;49:6789–6801.
22. Huang SY, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys Chem Chem Phys* 2010;12:12899–12908.
23. Leach AR, Shoichet BK, Peishoff CE. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J Med Chem* 2006;49:5851–5855.
24. Shoichet BK. Virtual screening of chemical libraries. *Nature* 2004;432:862–865.
25. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. *J Med Chem* 2006;49:5912–5931.
26. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 2007;35(Database issue):D301–D303.
27. The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research* 2011;39(Suppl 1):D214–D219.
28. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 2007;372:774–797.
29. Biasini M, Mariani V, Haas J, Scheuber S, Schenk AD, Schwede T, Philippsen A. OpenStructure: a flexible software framework for computational structural biology. *Bioinformatics* 2010;26:2626–2628.
30. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975;405:442–451.
31. Roche DB, Tetchner SJ, McGuffin LJ. The binding site distance test score: a robust method for the assessment of predicted protein binding sites. *Bioinformatics* 2010;26:2920–2921.
32. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull* 1945;1:80–83.
33. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
34. Wass MN, Sternberg MJ. Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins* 2009;77(Suppl 9):147–151.
35. Arnold K, Kiefer F, Kopp J, Battey JN, Podvenc M, Westbrook JD, Berman HM, Bordoli L, Schwede T. The Protein Model Portal. *J Struct Funct Genomics* 2009;10:1–8.

4.3 CAMEO Ligand Binding

4.3.1 Introduction

In the post-genomic era, the number of protein structures with unknown biological function is steadily growing. To bridge this rapidly growing gap between known sequences and unknown function, numerous computational and experimental techniques have been developed.^{5,1,6} Among these methods, structural comparison of the three dimensional shape to homologous proteins with known function can often help to assign the function of unknown proteins.¹⁹ However, the accuracy of such an approach is often insufficient since proteins with the same global fold can have different biological functions. In addition, for many applications knowing the global function of a protein is insufficient and it is critical to elucidate the structural details of how the protein interacts with its ligands.¹¹⁵ Hence, the identification of a protein's functionally important residues, such as ligand-binding sites or catalytic active residues, is a crucial step toward the goal of understanding the protein's molecular function and its biological role in the cell. Although those interactions are crucial for the function of a protein, they are often unknown and require elaborate and time-consuming studies by X-ray crystallography or NMR. To facilitate this process, numerous computational methods have been developed with the goal of identifying the precise location of ligand binding sites and the protein residues directly involved in interacting with the ligand.⁷

Predicting binding sites from a protein's sequence is of high relevance for life science research, ranging from functional characterization of novel proteins to applications in drug design and enzyme engineering. Consequently, the development of automated methods for predicting ligand-binding sites has received increasing attention over the past years.

However, relevant biological questions can only be addressed if predictions are specific and accurate. Therefore, evaluating the performance of prediction methods in a blind and objective way is crucial. To achieve this goal, ligand binding site prediction methods have been assessed since the 6th edition of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment.^{113,114} This process has been extended and fully automated in the Continuous Automated Model EvaluatiOn (CAMEO) framework as a new category.

In CAMEO Ligand Binding we continuously evaluate the accuracy and reliability of ligand binding site prediction services in a blind and fully automated manner to assess the current state of the art of prediction methods, identify possible bottlenecks, and further stimulate the development of new methods.

As mentioned before, ligand binding site (LB) predictions have been assessed in recent CASP experiments.^{25,24,23,113} However, the setup in CASP has shown some major limitations which prevented to draw significant conclusions:

- A very low number of challenging target structures with relevant ligands.
- A limited prediction format which treats all ligands uniformly, independent of their chemical type and treats all potential binding sites uniformly, independent of their affinity for different ligands.

Hence, in CAMEO a number of changes were made to alleviate the shortcomings of previous CASP LB experiments:

- First, in CAMEO participating servers are evaluated continuously, every week, based on all newly released PDB structures in order to accumulate a sufficiently large number of prediction targets.
- Second, in CAMEO we have modified the ligand binding site prediction format to allow a more fine-grained prediction and a more detailed assessment.
- Third, binding sites differ chemically and structurally from each other e.g. a metal ion binding site has different characteristics compared to a sugar binding site. We therefore assess ligand binding site predictions according to chemotype categories of the ligand expected to be bound.
- Fourth, the prediction of binding site residues employs continuous probability measures as opposed to the binary prediction format used in CASP, thus reflecting the likelihood for a residue to be involved in binding a ligand of a certain type.

4.3.2 CAMEO Workflow

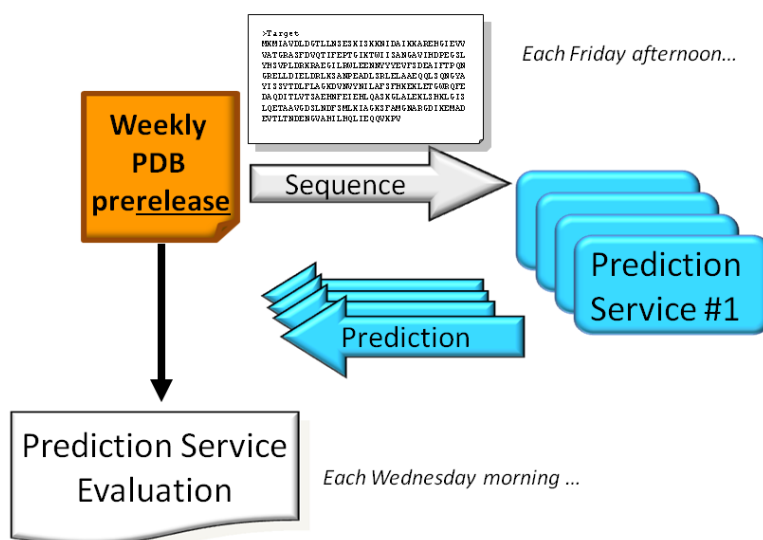


Figure 4.1: Schematic overview of the general CAMEO workflow. Every Friday afternoon, CAMEO sends the sequences of the weekly PDB prerelease to all registered servers. Each Wednesday, upon release of the structures, CAMEO Ligand Binding evaluates all predictions received prior to that date.

The workflow used in CAMEO is schematically depicted in Figure 4.1. It consists of three independent steps.

Collecting all PDB prerelease sequences (i.e. sequences ready to be released within the next release cycle), filtering them to remove duplicates and subsequently submitting them to all participating servers.

Collecting the responses from the participating servers and checking for format errors.

Evaluating automatically all predictions received prior to the new PDB release and presenting the results on the CAMEO web pages.

It should be noted that CAMEO functions in a blind and fully automated fashion. This is important to guarantee a fully independent, objective and comparable evaluation of current methods.

For CAMEO Ligand Binding, the biological role of ligands can be annotated by the users through a web-based, consensus “wisdom of the crowd” approach available as the CAMEO Structure Annotation System.

4.3.3 Prediction Targets

All protein sequences from the weekly PDB prerelease with a minimum length of 30 amino acids are submitted to the participating ligand binding site prediction servers five days before the structures are published. Predictions are only accepted if deposited before the new PDB release. Upon release of the target structures, all protein structures containing biologically relevant ligands are used for further evaluation of the accuracy of the prediction servers.

4.3.4 Ligand Annotation

All ligands present in the newly released PDB structures are classified based on their biological relevancy (i.e. 'biological relevant' or 'biological irrelevant'). Classification is based on manual inspection of all newly released protein structures. The user is assisted by an automatic ligand classification algorithm which suggests the biological relevancy for each ligand in a protein structure, based on the following criteria:

- distance between protein and ligand atoms must be within a certain cutoff distance

- commonly observed irrelevant ligands (e.g. common buffers, crystallization molecules) are excluded

- covalently bound post-translational modifications are removed

All automatic assignments can be overwritten by the user through a publicly available, web-based structure annotation platform, which we have developed for straightforward inspection of protein-ligand complex structures.¹¹⁶ This allows the user to easily navigate through all ligands of a structure using 3D visualization. Additional structural information is displayed and a direct access to the corresponding publication is given in order to facilitate the user's decisions. All ligands can be annotated according to the ligand classification scheme (see Section 4.3.5) and all annotations are stored in a database for further access. In addition, literature references can be added and discussions among users is possible within the annotation system.

Assignment of the biological role of a ligand in a structure has been one of the major limiting factors in the field of binding site and function prediction but is not limited to this

field. Thus, building a publicly available resource of manually curated ligand annotations has a great potential for improving current methods for protein structure and function prediction. Annotating ligands in CAMEO is a community effort. Thus, all people in the research community are invited to annotate their favorite structures. By using such a “wisdom of the crowd” approach a consensus answer can be obtained directly and controversial cases can be identified.

4.3.5 Ligand Classification Scheme

For the functional annotation of ligands present in CAMEO target structures, a new ligand classification scheme has been developed as shown in Figure 4.2. This classification scheme is based on a ligand-centric functional annotation approach as opposed to a binding-site-centric approach. For each ligand class a corresponding ligand classification number (LC number) was assigned consisting of a number for each depth level of the tree, separated by dots in order to easily determine the super class for a given LC number and to facilitate possible extensions of the current ontology.

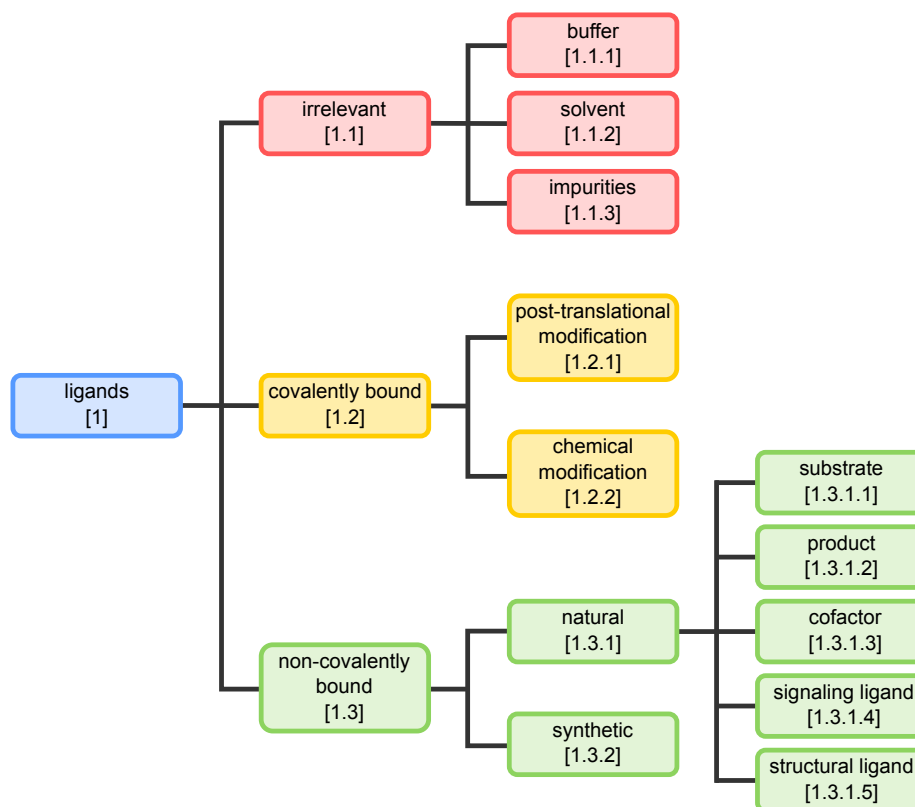


Figure 4.2: Current ligand classification ontology used in the CAMEO Structure Annotation System. For each ligand class its classification number is shown in square brackets.

Currently, the ontology consists of three super classes: irrelevant ligands [1.1], covalently bound ligands [1.2] and non-covalently bound ligands [1.3]. The first super class includes buffers [1.1.1], solvents [1.1.2] and other impurities [1.1.3]. The second super class includes natural post-translational modifications [1.2.1] like glycosylations or phosphorylation and chemical modifications [1.2.2] like addition of MTSL spinlabels. The third super class includes natural [1.3.1]

and synthetic [1.3.2] non-covalently bound ligands where the former is further divided into substrates [1.3.1.1] and products [1.3.1.2] of enzymatic reactions, functional cofactors [1.3.1.3] like ATP or SAM, signaling ligands [1.3.1.4] like steroid hormones and structural ligands [1.3.1.5] like structurally important metals ions. For CAMEO ligand binding, the evaluation is currently limited to the last super class.

4.3.6 Ligand Categorization

All ligands are categorized into four predefined categories and predictions are assessed individually for each category where at least one ligand was observed in the target structure. Ligand categorization is based on the preliminary ligand classification (chemtype) of the PDB, as defined in the chemical component dictionary¹¹⁷ under the item '*_chem_comp.pdbx_type*'.

Table 4.1: Definition of ligand categories in CAMEO LB and rules that are applied to categorized ligands according to the ligand classification (chemtype) of the PDB.

^aAll chemtypes of the organic category are included if covalently linked to ligands in the respective category.

Category	Description	Enclosed PDB Chemtype	Examples
I	ions	HETAI, HETIC (non polymeric)	ZN, SO4, ACT, NH4, IOD
O	organics	HETAIN, ATOMS, ATOMN, ATOMP, HETAC, HETAD (non-polymeric: <i>resnum</i> ≤ 2)	ATP, FAD, SAM, GLA, F3S
N	poly-nucleotides	ATOMN + organics ^a (polymeric: <i>resnum</i> > 2)	A, DA, G, U, T
P	poly-peptides	ATOMP + organics ^a (polymeric: 2 < <i>resnum</i> ≤ 10)	ALA, KCX, LLP, PTR

The following four CAMEO ligand categories are currently evaluated: ions (*I*), organics (*O*), poly-nucleotides (*N*) and peptides (*P*). Each category is defined as a subset of the available PDB chemtypes, as summarized in Table 4.1. All ions must be monomeric and thus, consist of a single residue. Organics can be monomeric or dimeric. Peptides and poly-nucleotides must consist of at least three residues and at maximum of 10 and 200 residues, respectively. In order to include capped peptides and poly-nucleotides, residues of the organic type are allowed in the latter two categories given that they are covalently linked to the peptide or nucleotide.

4.3.7 Assessment

Preparation of Target Structures

Upon release, all target structures are obtained from the PDB¹¹⁸ in the mmCIF file format, are filtered and prepared for further use through the following steps.

All structures which fail to align to the sequence which was sent to the predictors (i.e. SEQRES sequence) are discarded.

All non-standard amino acids are converted to their respective standard amino acid to match the exact sequence that was sent to the predictors.

All biological assemblies defined by the PDB are considered.

Covalently linked residues are grouped together into ligand entries.

Each ligand entry is uniquely categorized into one ligand class.

Calculation of the Reference

Based on the prepared target structure, a reference binding site prediction is calculated for each biological assembly of each protein-ligand complex. The reference contains a binding site probability value for each heavy atom of the protein which is observed in the released structure. This value is computed based on the protein-ligand distance (r) using a sigmoidal curve of the following form:

$$p_{ref} = \frac{1}{1 + \exp(1.5 * r - 7.5)}$$

Thus, a distance of 5 Å yields a binding site probability value (p_{ref}) of 0.5 whereas a distance of 3 Å and 7 Å leads to a probability close to 1.0 or 0.0, respectively.

p_{ref} values are computed for each ligand category individually by considering only the subset of ligands corresponding to that category. In addition, for ligands belonging to the categories I and O, p_{ref} values are computed individually for each specific ligand found in the structure.

In case multiple protein chains are present in the biological unit, the reference is computed for each protein chain individually. In this step, all ligands of a particular category are included, not just the ones contained in the current chain. This allows to compute probabilities for ligands bound in the interface between multiple chains.

In case multiple biological assemblies are defined, an individual reference is generated for each biounit as previously described.

Compare Prediction with Reference

Each prediction is evaluated by comparing it to the reference as described in Section 4.3.8. A number of considerations apply for numerous special cases.

Multiple chains If the *prediction* contains multiple chains with matching sequences, each chain is scored against the reference and the mean score of all chains is reported.

If the *reference* contains multiple chains with the same SEQRES sequence, the prediction is evaluated against each chain of the reference and the highest score is reported.

In case that both the *reference* and the *prediction* contain multiple matching chains, each prediction chain is evaluated against each reference chain. To obtain the overall score, the best score for each chain in the prediction is averaged over all available prediction chains.

Multiple biological assemblies In case where multiple biological assemblies are defined in the mmCIF file, a reference is generated for each biounit. Since none of these assemblies is a priori more correct than the others, the prediction is assessed against each of those references and the value for the biounit with the best score is reported.

Level of Prediction When comparing a prediction against a reference, the reference always contains p-values for each atom present in the structure. The prediction however, can contain p-values for all atoms or for all residues or mixed for some atoms and some residues. Therefore, for the comparison, prediction values at the atom level are used if available. Otherwise, for each atom present in the reference, the p-values of the corresponding residue are used. If no p-values are given both for the atom as well as the corresponding residue, all p-values are set to 0.0.

Specific Compound Predictions When predictions are given for specific compounds in addition to the ligand categories, the reported score is exclusively computed based on the compound predictions, if the following condition applies: For each ligand category individually, all compounds observed in the target structure must be present in the prediction. Predictions for additional compounds which are not present in the structure are neglected. If the condition is not fulfilled, the reported value is based solely on the category prediction values.

4.3.8 Scoring

Predictions are currently scored based on four different methods. Receiver operating characteristics (ROC) area under the curve,¹¹⁹ Pearson's correlation coefficient, Spearman's rank correlation coefficient¹²⁰ and Matthew's correlations coefficient.¹²¹

ROC area under curve is a measure for the ability of a classifier to produce relative scores, i.e. predicting higher values for residues in the binding site compared to those not in the binding site. For ROC curves, a cutoff p-value must be defined on the reference. For CAMEO this cut-off is set to a p-value of 0.5 corresponding to a distance of 5 Å.

Pearson's correlation coefficient is a measure of the linear dependence between the prediction and the reference and does not need any cutoff. However, it has two drawbacks: (1) it is highly dependent on the mathematical function used to generate the reference, (2) it is heavily influenced by p-values for residues that are far away from the binding site, where the correct ordering, from a biological point of view, is of limited importance.

Spearman's rank correlation coefficient quantifies the non-parametric statistical dependency between the prediction and the reference. Opposed to Pearson's correlation coefficient, it is independent of the mathematical function used to generate the reference, however, it is still heavily influenced by p-values for residues that are far away from the binding site.

Matthew's correlation coefficient is implemented for comparison with the results of previous CASP experiments^{24,25} where it was used as a measure to evaluate the ligand binding site prediction category. It is a measure that accounts both for over and under-prediction. However, it needs both a cutoff value for the reference as well as one for the prediction to classify all p-values into positives and negatives. Thus, the advantage gained by introducing continuous probability prediction values is lost. As with ROC, the cutoff p-value for the prediction and the reference was chosen as 0.5 corresponding to a distance of 5 Å in the target structure.

4.3.9 Baseline Servers

For comparison of server performances, three naïve servers were added to establish a baseline for what a straightforward method would predict. Each of these servers focuses on a specific approach: (1) sequence conservation, (2) homology transfer and (3) geometric binding pocket identification.

Naïve Conservation

A server that computes the conservation of each residue in a query sequence using sequence conservation information only. This server focuses on the most difficult case where no homologue structure of the query protein is detectable.

Given a query sequence, the server performs the following three operations: (1) sequence search in a protein database: a blast¹²² search is performed on the NR-90 database. All HSPs with an E-value of less than 0.1 are accepted. (2) multiple sequence alignment: using the list of HSPs obtained from blast, a multiple sequence alignment is built using ClustalW.¹²³ (3) computation of conservation values: conservation scores for each amino acid in the query sequence are computed from the multiple sequence alignment according to the ConSurf method.¹²⁴ The raw conservation scores are used to assign ligand binding site probabilities with equal values for all four ligand categories.

Naïve Homology

A server that computes ligand binding sites by homology transfer based on ligands found in homologous structures in the PDB. This server focuses on the most straightforward case where there is at least one homologue structure which contains biological relevant ligands.

The ligand binding predictions are created from models generated by SWISS-MODEL.^{125,126} Ligands present in the template structure are inserted into the protein model if certain criteria are met:

If a compound is included in a list of selected biologically important compounds¹, its binding site, defined as all residues within 3 Å of the ligand, is superposed onto the model and all ligand

¹currently the following small molecules are taken into account in the SWISS-MODEL pipeline:
cations: CA, CO, CU, CU2, FE, FE2, MG, MN, MO, NA, NI, ZN
cofactors: ADP, AMP, ATP, BTN, COA, BGC, GLC, GDP, GMP, GTP, GSH, FAD, FMN, HEM, HEA, HEB, NAD, NAP, NDP, NAI, PLP, SAM, THG, TPP, UDP, CDP, SF4, FES

heavy atoms are copied into the model. If the ligand clashes with the model, i.e. any protein-ligand distance is less than 1.5 Å, or the RMSD of the binding site residues is larger than 2 Å, the ligand is rejected.

For each atom in the protein the distance between its position and the position of all the ligand atoms is computed and used to calculate the prediction score. If multiple ligands of the same category are present, the maximum of the computed score for this category or that specific compound is reported.

The score is calculated with a capped linear function bounded between zero and one using the following formula:

$$\text{score}(d) = \begin{cases} 1 & \text{if } d < 3 \\ 2 - (\frac{1}{3} * d) & \text{if } 3 \leq d \leq 6 \\ 0 & \text{if } d > 6 \end{cases}$$

where d corresponds to the distance between the protein and ligand atom.

Naïve Pocket

This baseline server predicts ligand binding sites according to an analysis of the shape of the protein structure as computed based on a homology model of the input sequence. This server focuses on an intermediate case where there is at least one homologue structure which however, does not contain any biological relevant ligands.

First, a comparative model is built using SWISS-MODEL.^{125,126} Second, all ligands in the modeled structure are discarded and the molecular surface of the protein is computed using MSMS.¹²⁷ A three dimensional grid spanning the whole protein is generated. On every point of that grid, the number of surface points within a distance of 10 Å is counted and stored as a value on that grid point. The computed grid is then smoothed using a Gaussian filter with a smoothing radius of 4 Å. Subsequently, for each atom of the protein structure, the probability of being a ligand binding site is computed from a trilinear interpolation of the values of the 8 closest grid points. The values are then scaled to match the range of 0.0 to 1.0. The method is described in detail in Section 4.4.

4.3.10 Prediction Format

Although, predictions in the format used by the predictioncenter¹²⁸ during previous CASP experiments^{25,24,23} are accepted, a new format which follows the suggestions from the last assessment for the ligand binding category during CASP9²⁵ is implemented, allowing much more detailed predictions.

This new format consists of three sections separated by the "|" symbol:

The first section is a unique identifier for a residue or atom. It has two mandatory fields, the residue name ("r") and the residue number ("n"). In addition, two optional fields can be specified, the chain name ("c") and/or the atom name ("a").

The second section contains predicted p-values for four ligand categories: ions ("I"), organics ("O"), poly-nucleotides ("N") and peptides ("P"). Predictions for all four categories are mandatory. The values are probabilities resembling the likelihood of binding a ligand belonging to a specific category.

The last section is optional and allows the assignment of p-values for specific ligands denoted by their three letter PDB code.

The following additional details about the new prediction format should be noted:

- All values must be specified as a key-value pair, where the value is separated from the key by a "=" sign. All key-value pairs must be terminated by a ";" sign.
- The order of the key-value pairs within one of the three sections is irrelevant, whereas the order of the three sections is fixed.
- All predicted values must be in a range from 0.0 to 1.0.
- To simplify the prediction format, lines can be omitted if they contain only zero values for the predictions, both in the categories and the compounds section.
- Predictions can be made at the residue and/or atom level. In the case of the latter, the atom name must be specified in the unique identifier section (key: "a").
- Predictions are mandatory for all four ligand categories, whereas for specific compounds they are optional.

A general format description is given here:

```
r=<resname>; n=<resnum>; [c=<chainname>;] [a=<atomname>;] | ↔
I=<ion prob>; O=<org prob>; N=<nucl prob>; P=<pep prob>; | ↔
[<compound ID1>=<compound prob>;] [<compound ID2>=<compound prob>;] ...
```

where text in between "< >" brackets are placeholders for actual values and all key-value pairs in between "[]" brackets are optional. The ↔ sign indicates that this represents one line in a file.

Two examples of the new prediction format are given in Appendix A.2.1.

Conversion from CASP format Predictions using the binary CASP format are accepted and internally converted to the new prediction format. Since the latter is much more detailed, numerous limitations apply when converting.

Predictions are converted into p-values of 0.0 (non-binding site) and 1.0 (binding site), with no values in between.

Predictions for all four categories are set to the same value.

Predictions for specific compounds are not assigned.

Predictions are assigned on the residue level only and no specific predictions at the atom level are done.

Predictions containing residue numbers outside of the sequence range are discarded, except for residue number 0, indicating that no binding site was predicted.

4.3.11 Results and Discussion

Number of Relevant Targets

A major limitation of previous CASP LB experiments was the limited number of target structures that could be evaluated for the ligand binding site category. Although, the number of target structures in recent CASPs was around 120, only about one fourth of those (i.e. around 30 structures) had biologically relevant ligands bound, and only very few of them were hard cases where either no structure or no ligand information was present in the PDB. To alleviate this shortage of ligand binding sites, CAMEO evaluates all servers weekly based on all newly released PDB structures to obtain a large number of target structures.

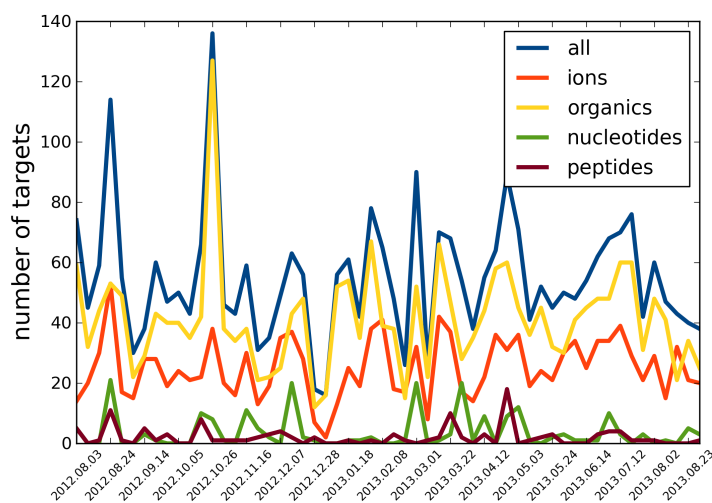


Figure 4.3: Number of targets per week evaluated by CAMEO Ligand Binding.

We have analyzed the number of CAMEO LB targets with biologically relevant ligands over a period of 35 weeks. The results are summarized in Figure 4.3 where the number of targets for each category is shown over time. Overall, 1647 structures were evaluated. On average 47 structures were assessed every week, with a minimum of 1 and a maximum of 114. Of those, 48.1% contained ions, 72.8% organic ligands, 6.8% poly-nucleotides and 5.1% poly-peptides.

It should be noted that using CAMEO LB, the number of targets available in a single week is larger than the number of targets present in a full round of CASP.

Category Predictions

Figure 4.4 shows the CAMEO LB target 4EQP_1 (CAMEO date: 2012.04.27). The structure is a Staphylococcal nuclease which is bound to a calcium ion and the organic inhibitor thymidine-3',5'-diphosphate (THP). The target contains two binding sites. The first contains the calcium ion which is bound through the side chains of Asp21, Asp40 and Glu43 and the backbone carbonyl of Thr41. The second binds the organic ligand THP which is in direct contact with Arg35, Lys84, Tyr85, Arg87, Leu89, Tyr113, Tyr115. The prediction from the homology based baseline server is mapped onto the structure. Binding site residues are shown in sticks representation and are colored according to the predicted p-values using a gradient from red (non-binding site) to green (binding site).

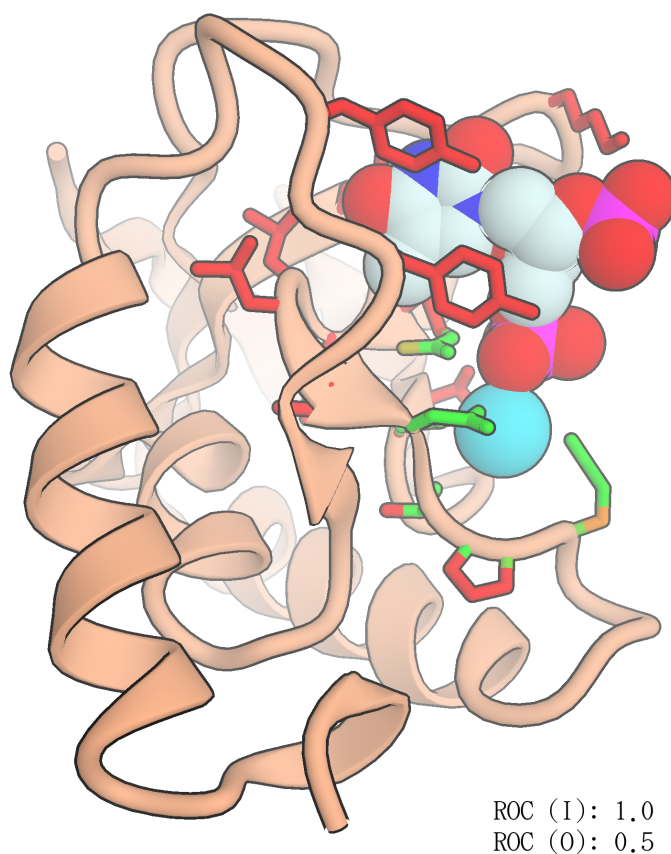


Figure 4.4: Structure of a Staphylococcal nuclease (4EQP) bound to a calcium ion and the organic inhibitor thymidine-3',5'-diphosphate (THP). The ligands are shown in sphere representation, with the calcium ion colored in cyan, and THP in element coloring. Predictions from homology based baseline server (naïve homology) are mapped onto the structure using a gradient from red (non-binding site) to green (binding site). The ROC score of both ligand categories is reported at the lower right hand side.

This target is an excellent example of the strength of using the new prediction format allowing to give different predictions for different ligand categories. The naïve homology server predicted only the binding site of the calcium ion but not of the organic ligand THP, since the latter is not in the list of commonly observed, relevant organic ligands used by this server for

prediction. From Figure 4.4 it is obvious that the binding site of the ion is perfectly predicted with an ROC auc of 1.0. However, since no prediction for the organic category was given, the performance is assigned a score of a random prediction (ROC auc of 0.5). If this target was predicted and assessed as a whole, as was done in recent rounds of CASP, a ROC auc score of 0.69 would be obtained. This example clearly shows the strength of predicting and assessing multiple ligand categories individually. The overall score would indicate a rather weak prediction performance (ROC auc of 0.69) which clearly does not appropriately reflect the real prediction performance, which in this example is excellent for the ion binding site but poor for the organic ligand binding site.

Specific Compound Predictions

If specific compound predictions are given for all ligands present in the target structure, the prediction performance will be assessed on these specific predictions instead of the corresponding category. This has the advantage in the case where, for example a protein has two binding sites for two distinct metal ions but in the target structure only one of the two is occupied.

To demonstrate the effect, we have generated an example based on the CAMEO ligand binding target 3RVH_1 (CAMEO date: 2012.05.04). The structure is a lysine-specific histone demethylase JMJD2A. The structure contains two metal ions: a structurally important zinc ion and a nickel ion substituting for the catalytically relevant iron ion. In addition, the protein is bound to an organic small-molecule inhibitor (HQ2) which is in direct contact with the catalytic metal ion. The two metal ions bind to two distinct binding sites located 15 Å apart. The naïve homology prediction server gives predictions at the specific compound level and predicts both metal ion binding sites correctly. To investigate the difference in the prediction performance between specific compound predictions and category predictions, the structural zinc ion in the target structure was removed prior to the assessment, yielding a target structure where only one metal ion binding site was present. Subsequently, the prediction was evaluated using both the category or the compound predictions exclusively.

Table 4.2: Comparison between the evaluated prediction performance when assessed either based on categories or based on specific compounds exclusively. All scores are computed for the ion category.

	assessed on	
	categories	compounds
ROC auc	0.994	0.997
Spearman's correlation coefficient	0.395	0.538
Pearson's correlation coefficient	0.474	0.651

From Table 4.2 a significant increase can be observed for Spearman's and Pearson's correlation coefficient when specific compounds are predicted. A much smaller increase is observed for ROC auc, since ROC only accounts for relative scores and in both cases, all high probability binding site residues are well scored.

Oligomeric State

Figure 4.5 shows the CAMEO LB target 4G8S_1 (CAMEO date: 2012.08.03). The structure is a nitroreductase from *Geobacter sulfurreducens* PCA which forms a homodimer. It binds the organic cofactor riboflavin-5'-phosphate (FMN) which is located in the interface between the two protein chains and is in contact with 13 residues of chain A and 9 of chain B. The target is a relatively easy case, where the close homologue 4DN2 (sequence identity: 76%, superposition rmsd: 1.0 Å), also a homodimer, binds the same ligand in the same location. The prediction from the INTFOLD-FN server is mapped onto the structure. Binding site residues are shown in sticks representation colored according to the predicted p-values using a gradient from red (non-binding site) to green (binding site).

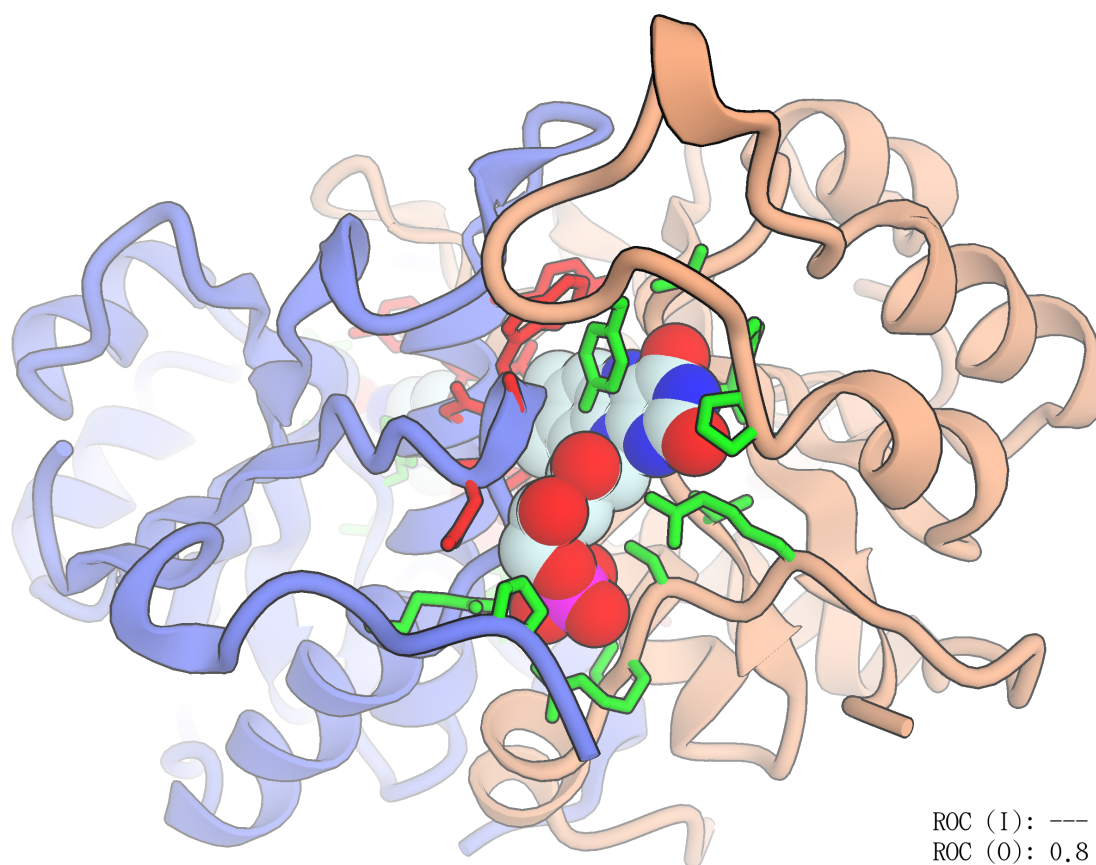


Figure 4.5: Homodimeric structure of a nitroreductase from *Geobacter sulfurreducens* PCA (4G8S) bound to the organic cofactor riboflavin-5'-phosphate (FMN). Predictions from the server INTFOLD-FN are mapped onto the structure using a gradient from red (non-binding site) to green (binding site). The ROC score of both ligand categories is reported at the lower right hand side.

From Figure 4.5 it is obvious that INTFOLD-FN only predicts binding site residues from one single chain and thus does not consider the correct oligomeric state. This is also true for all other registered non-naïve servers. The performances are generally good as shown in Table 4.3. It is obvious that the performance could be further improved by considering the oligomeric state of the target structure for the prediction as clearly demonstrated by the naïve homology server,

Table 4.3: Comparison of the prediction performance for CAMEO ligand binding target 4G8S_1.

	INTFOLD-FN	HHfunc	server5	naïve homology
ROC auc	0.83	0.71	0.83	0.98
Spearman's correlation coefficient	0.55	0.42	0.55	0.77
Pearson's correlation coefficient	0.70	0.54	0.70	0.81
Matthew's correlation coefficient	0.78	0.59	0.77	1.00

which is the only server that currently takes the correct oligomeric state into account.

Web-based Representation of the Results

All servers are evaluated based on a number of different scores and all raw numbers are reported on the CAMEO Ligand Binding web site. To facilitate the interpretation of these numbers, CAMEO Ligand Binding presents them in more user friendly graphical representation as the averaged performance (according to ROC auc) of each server against the number of targets that were predicted by the server. Thus, a server on the upper left hand corner performs very well on a small number of selected targets but does not give predictions for all the other targets. On the other hand, a server on the right hand side, might on average perform less good than the previously described server, but gives predictions for nearly all targets. Depending on the users application, both approaches might be valid and thus the user can select what is needed or combine predictions from multiple servers in order to obtain the best suited predictions. In addition, to observe the development of a server, plots of its performance over time are given on the CAMEO Ligand Binding web site.

4.3.12 Conclusion

CAMEO Ligand Binding has been publicly released recently.¹²⁹ Despite the short period since the announcement, it is already used by the ligand binding site prediction community. In this short time, three external servers, coming from the McGuffin and the Söding groups, have joined.

CAMEO Ligand Binding has already shown significant improvements over recent CASP ligand binding site prediction experiments by evaluating a significantly increased number of targets and by allowing to predict binding sites in a more fine grained manner through categorization and continuous prediction values.

CAMEO Ligand Binding has already proven to be useful by identifying possible bottlenecks in current prediction methods like for example the incorrect use of the oligomeric state of target and template structures. CAMEO LB also helps the server administrators by testing the server on a weekly basis, by notifying them in case where no predictions are received and by reporting server response times to obtain a measure for the technical performance.

CAMEO Ligand Binding helps the users of ligand binding site prediction methods by identifying which methods are best suited for a particular use-case. This is achieved by assessing the

prediction performance on individual ligand categories and by identifying strength and weaknesses of each server, like the difference between highly accurate servers which predict only few targets and less accurate servers which predict nearly all targets. The information obtained through CAMEO could further help the user to combine multiple server predictions in order to obtain the best predictions for a particular purpose.

CAMEO Ligand Binding is continuously extended and improved in order to challenge all participating servers and to provide the most useful information to the community. Ultimately, CAMEO should be able to determine which methods should be chosen in order to yield the best predictions for a particular target sequence.

4.4 Geometry Based Ligand Binding Site Prediction

4.4.1 Introduction

A quantitative comparison of the performance of ligand binding site prediction methods is not always straightforward. In particular, different methods perform differently based on the data available for a particular prediction case. For example, a method that depends on homology transfer from related structures might not perform well in the case where either no homologue structure is available or where no ligands are bound in the homologue structures. Whereas, in such a case, a method depending only on sequence conservation might yield a good performance. However, comparing methods with each other that use completely different underlying methodologies might be problematic.

To alleviate this problem, methods might be compared to a naïve baseline method which uses the same methodology in a conservative manner, without incorporating the latest developments in the field. Thus, for CAMEO Ligand Binding, we have developed three baseline servers, each focusing on a individual approach as described in Section 4.3.9: (1) sequence conservation, (2) homology transfer and (3) geometric binding pocket identification. Here, we describe the latter, naïve geometric binding pocket identification server, named naïve pocket. This server focuses on an intermediate case where homologue structures are present, but they do not contain any ligand information.

4.4.2 Method

The geometric prediction server is based on BScore (see 5). Briefly, BScore uses the molecular surface to define the protein shape and analyzes the distribution of surface vertices to identify binding pockets. The following steps are performed:

The solvent excluded surface of a protein structure is computed by MSMS (version 2.6.1)¹²⁷ using a probe radius of 1.4 Å and a sampling density of 6.

A 3-dimensional orthogonal grid is produced with a grid spacing of 1 Å and the dimension of the protein lengths plus a margin of 3 Å on each side. For each point on this grid, the number of surface points that are within a certain cutoff distance r_{max} is computed. In addition, surface points can be excluded based on the direction of the surface normal \vec{n}_j and a cutoff angle γ_{max} .

The computed grid is smoothed using a Gaussian filter with a smoothing radius σ .

The score $b(\vec{r}_i)$ on grid point i at position \vec{r}_i is computed as

$$b(\vec{r}_i) = \sum_j \delta_j(\vec{P}_j, \vec{r}_i) \quad (4.1)$$

where

$$\delta_j(\vec{P}_j, \vec{r}_i) = \begin{cases} 1 & \text{if } \|\vec{P}_j, \vec{r}_i\| \leq r_{max} \text{ and } \gamma \leq \gamma_{max} \\ 0 & \text{if } \|\vec{P}_j, \vec{r}_i\| > r_{max} \text{ or } \gamma > \gamma_{max} \end{cases} \quad (4.2)$$

where \vec{P}_j is the position of surface vertex j and γ is the angle between the vector from the grid point \vec{r}_i to the surface vertex \vec{P}_j and the surface normal \vec{n}_j computed as

$$\gamma = \arccos \left(\frac{(\vec{r}_i - \vec{P}_j) \cdot \vec{n}_j}{\|\vec{r}_i - \vec{P}_j\| \cdot \|\vec{n}_j\|} \right) \quad (4.3)$$

4.4.3 Results and Discussion

Set of Protein-Ligand Complexes

All crystal structures used in CAMEO Ligand Binding from December 2011 to January 2012 containing ligands in the organic category were used as a training set. The crystal structures were obtained from the PDB,^{130,118} all ligands were removed and predictions were performed on these structures. Reference prediction entities were generated for all biologically relevant ligands (according to the CAMEO LB classification) and the predictions were assessed using the CAMEO LB pipeline.

Parameter Optimization

As shown in Section 4.4.2 the method depends on the parameters r_{max} , γ_{max} and σ). To optimize the method to obtain the best performance, those parameters were systematically varied as follows:

r_{max} : 2 to 10 in steps of 2

γ_{max} : 20,45,90,135,180°

σ : 0 to 20 in steps of 2

The performance of the method was evaluated for each parameter combination by computing the area under the curve (AUC) of the receiver operating characteristics (ROC), Spearman's rank correlation coefficient (rank) and Pearson's correlation coefficient (correl) for each structure of the set of protein-ligand complexes described previously. For each of the three scores, the overall performance was computed by averaging the score over all structures in the test set.

Prediction Performance on Crystal Structures

The performance of the naïve pocket server, as evaluated on the previously described set of crystal structures, is shown in Figure 4.6. Independent of the metric used for evaluating the performance (i.e. ROC auc, rank, correl), similar behavior is obtained. Overall, the performance depends highly on all three parameters as follows:

Cutoff Distance (r_{max}) Generally, good performance is obtained for a broad range of cutoff distances from 5 to 10 Å. Independent of the other parameters, an increase of the cutoff distance

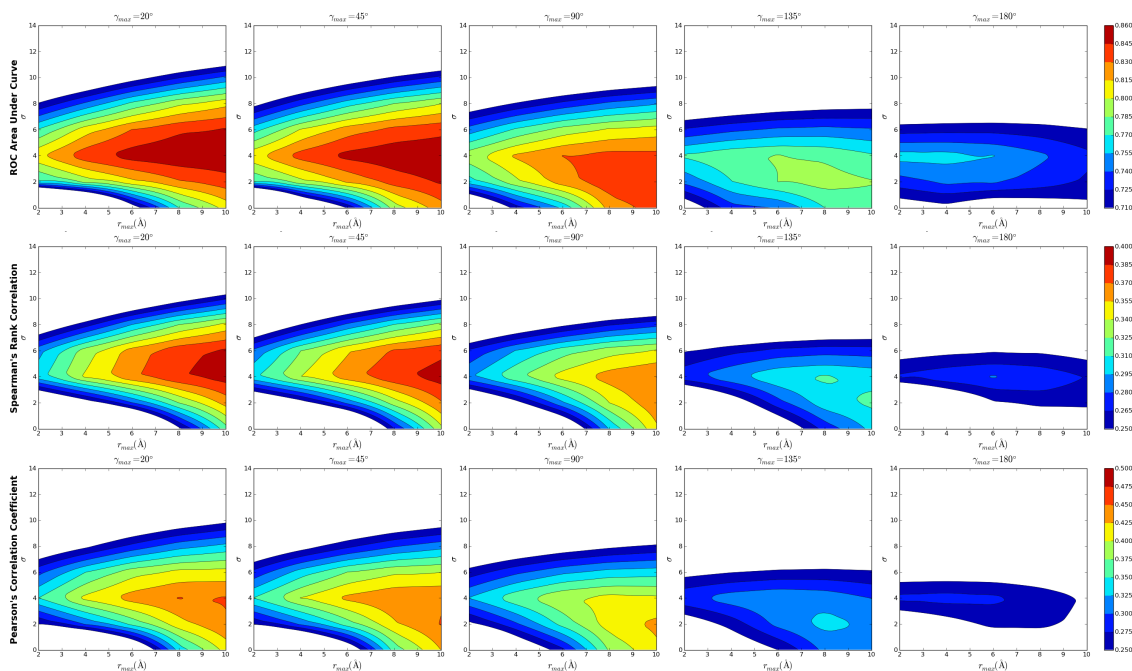


Figure 4.6: Performance of the naive pocket server evaluated on X-Ray structures using three different scores: ROC area under curve (top panel), Spearman's rank correlation coefficient (middle panel), Pearson's correlation coefficient (bottom panel). The scores are shown using a color gradient from blue (lower performance) to red (better performance) and are plotted against the three parameters of the scoring function: cutoff angle γ_{max} (individual plots from left to right), cutoff distance r_{max} (x-axis within one plot), smoothing factor σ (y-axis within one plot).

yields an increase in performance, with a maximum at 10 Å.

Cutoff Angle (γ_{max}) When discarding surface points where their surface normal vector points away from the grid point, i.e. decreasing the cutoff angle, the performance of the method is improved. The optimal performance is obtained using a cutoff angle between 20 and 45°.

Smoothing Factor (σ) Using a Gaussian function to smooth the scores of the grid has a significant impact on the binding site prediction performance, where both no smoothing (i.e. $\sigma = 0$) or extensive smoothing (e.g. $\sigma > 10$) yields poor performance but intermediate smoothing shows a significant improvement with a maximum performance at 4.0.

Overall, prediction performances are encouraging where optimal parameters yield a good ROC auc of 0.855. For the naive pocket server included in CAMEO Ligand Binding, the following parameters were chosen which give rise to the optimal performance without severely limiting the number of included surface vertex points: $\gamma = 45^\circ$, $r_{max} = 10 \text{ \AA}$, $\sigma = 4.0$.

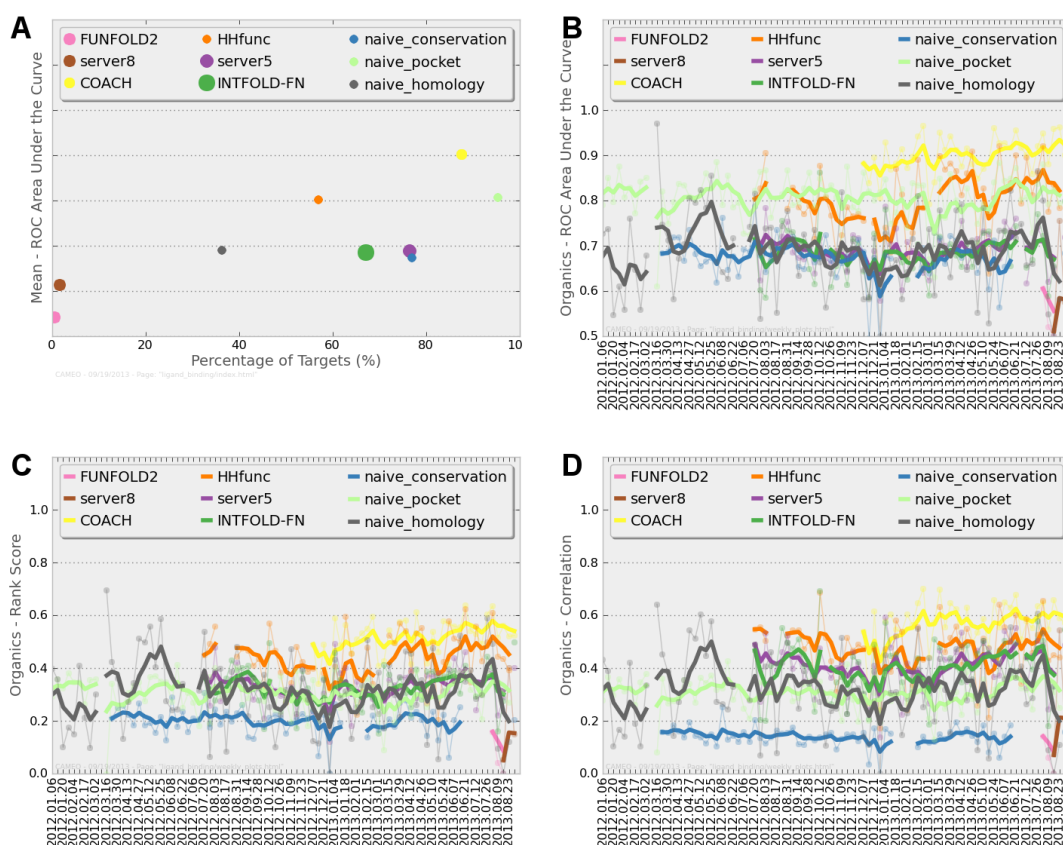


Figure 4.7: Performance of the naïve pocket server in the organic ligand binding site category of CAMEO LB. (A) The ROC AUC averaged over all CAMEO LB targets of the last seven month is plotted against the number of modeled targets for all participating servers. (B-D) Comparison of weekly scores between all participating servers: (B) ROC AUC, (C) Spearman's rank correlation coefficient, (D) Pearson's correlation coefficient.

Prediction Performance on Models

In addition to crystal structures, the naïve pocket server is applicable to homology models and was evaluated on comparative protein structure models produced by SWISS-MODEL.^{125,126} The server was added to CAMEO Ligand Binding in January 2012, and its performance over the first seven month is shown in Figure 4.7 (server name: naïve_pocket). Since the server only predicts binding sites for organic ligands, only this category is evaluated.

The accuracy according to ROC auc is comparable to non-naïve servers. However, naïve_pocket is one of the few servers which predicts nearly all targets (97%), significantly more than all non-naïve servers currently included in CAMEO Ligand Binding (Figure 4.7 A,B). The performance according to Spearman's rank correlation and Pearson's correlation coefficient is relatively low with a correlation coefficient of 0.32 and 0.31, respectively (Figure 4.7 C,D). The high ROC score but the low correlation scores highlight, that naïve_pocket is able to correctly classify residues into binding site and non-binding site residues, but has limited ability to correctly rank the residues within the two classes.

4.4.4 Conclusion

The naïve geometric prediction server depends on a reliable protein structure to determine binding pockets and is both applicable to experimentally determined protein structures as well as protein structure models originating from comparative modeling.

Although, the methodology used by the naïve pocket predictor for identifying ligand binding sites is straightforward, its capability to predict binding sites for organic ligands is promisingly high and outperforms most other servers. This clearly illustrates that current binding site prediction methods have to advance further in order to predict binding sites accurately and reliably.

In addition, by combining such a straightforward prediction server with other methodologies like homology transfer, further improvements of the prediction performance could be easily achieved which might lead to one of the top performing methods in the field.

Chapter 5

BEscore: a Novel Method for Rapid Scoring of Protein-Ligand Complexes

5.1 Introduction

Computational methods have a significant impact on the drug discovery process for example by greatly accelerating the identification of early hit compounds.^{28, 131} In cases where 3-dimensional structural information is present, or where accurate models can be built, virtual screening using molecular docking is often the method of choice, due to its good compromise between accuracy and computational efficiency.

In spite of major progress in recent years docking is often still hampered by the generation of a large number of false positive hits. Amongst other shortcomings, one underlying reason for false positive predictions is the generation of many incorrect, but highly scored, ligand conformations. This problem becomes particularly important when applying post-filtering methods such as protein-ligand interaction footprint filtering¹³² or re-scoring based on free energy methods^{133, 134} which rely on finding the best poses, i.e. those which are close to the experimentally observed placement of the ligand. Despite those limitations of scoring functions, an expert's eye, however, can often distinguish the best poses from decoys by visual inspection.¹³⁵ For example, a set of docking results usually offers a number of poses with different yet plausible hydrogen bonding patterns. With visual inspection most of the unreliable poses can be filtered out. In our experience these are often the less buried ones,⁷⁴ and in most cases an expert would prefer placements with an intimate binding of the ligand in sub-pockets, even at the expense of fewer polar interactions. This is in agreement with the notion that, generally, steric complementarity plays a larger role in molecular recognition than polar and electrostatic interactions.¹³⁶ Such a visual inspection approach has been commonly applied in successful docking studies^{137, 138, 139} where manual inspection of selected ligand poses was performed as a manual post-filtering step. Although it is a promising approach, it clearly cannot be performed in a high-throughput docking experiment due to its high time demand and its subjectivity.

The inability of scoring functions to always detect the best poses is just one part of the scoring function problem. This problem has been extensively discussed and reviewed.^{136, 140, 141, 142, 143, 144} A number of validation studies on scoring functions and docking have also been published,^{136, 144, 145, 146, 147, 148, 149,}

complemented by articles on the intrinsic difficulties and flaws associated with such studies.^{152,153}

As a consequence we have developed a scoring function that aims to identify the best poses by quantifying the degree of burial and the electrostatic interactions of the ligand poses in a binding site. Here, we present the initial results of these investigations.

5.2 Method

5.2.1 Shape Term (Degree of Burial)

In the approach pursued here the ligand binding cavity is converted into a representation that reflects just the shape of the protein surface. The method relies only on a 3-dimensional protein structure, obtained from experiment or through comparative modeling. It does not require affinity data, atom typing or elaborate deduction of potentials. The shape term in the new scoring function aims to assess docking poses based on their degree of burial in a ligand binding pocket. Such measures have so far been used for detecting cavities on a protein surface,^{15,154} which implies that the degree of burial can be useful for identifying protein surface depressions and quantitatively describing their geometric characteristics. Here, the degree of burial is used for re-scoring the poses of docking calculations.

The starting point for deriving the local degree of burial $b(\vec{r}_i)$ for a given protein at a given position \vec{r}_i is the shape of its ligand binding site, which is represented by the points of the solvent excluded surface calculated by MSMS (version 2.6.1).¹²⁷ A schematic representation is shown in Figure 5.1. From the surface points the $b(\vec{r}_i)$ values are calculated over the \vec{r}_i coordinates of an orthogonal three-dimensional grid. For each position \vec{r}_i on this grid, the local degree of burial $b(\vec{r}_i)$ in the binding site is estimated by counting the surface points \vec{P}_j , within a given distance r_{max} from the orthogonal grid point \vec{r}_i :

$$b(\vec{r}_i) = \sum_j \delta_j(\vec{P}_j, \vec{r}_i) \quad (5.1)$$

where

$$\delta_j(\vec{P}_j, \vec{r}_i) = \begin{cases} 1 & \text{if } \|\vec{P}_j, \vec{r}_i\| \leq r_{max} \\ 0 & \text{if } \|\vec{P}_j, \vec{r}_i\| > r_{max} \end{cases} \quad (5.2)$$

Gaussian Weighting

As an alternative to applying a simple distance cutoff, Gaussian functions can also be used for calculating the local degree of burial. The Gaussian weighted version of this score is designated $b_g(\vec{r}_i)$:

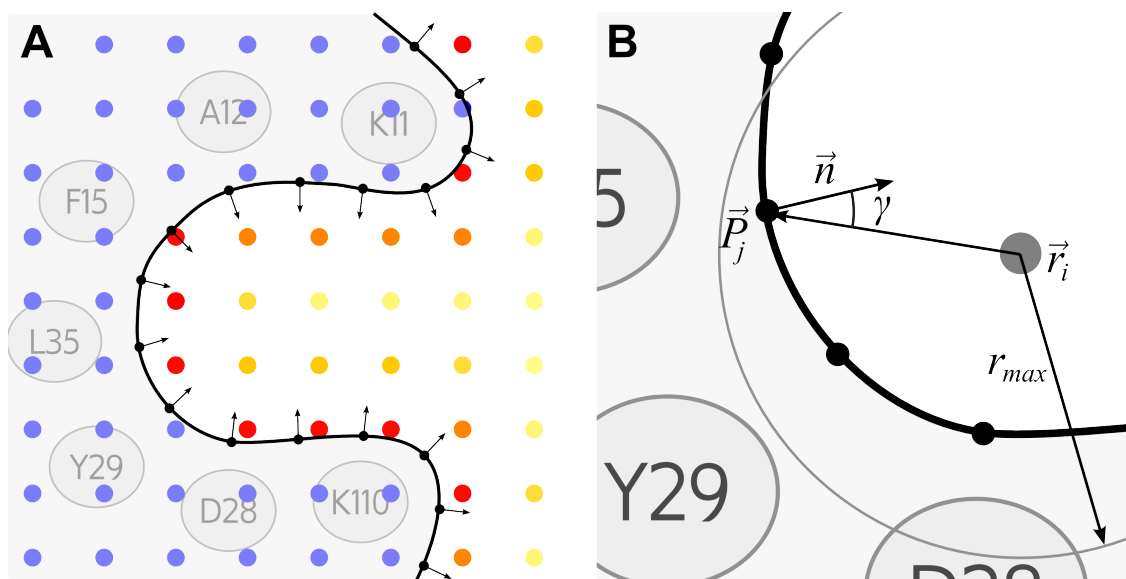


Figure 5.1: Schematic representation of the $BScore$ and $BScore_g$ method. (A) The protein is represented in light gray with its surface (black line) as defined by surface vertices (black dots) and their normal vectors (black arrows). The score is computed on each grid point (colored dots) where the value is shown in a color gradient from yellow (low score) to red (high score). (B) Cut-out of plot (A) representing the different terms used in Equation 5.1- 5.10

$$b_g(\vec{r}_i) = \sum_j C \cdot e^{-\alpha(\|\vec{P}_j - \vec{r}_i\| - r_{max})^2} \quad (5.3)$$

where C is a normalization factor to make $b_g(\vec{r}_i)$ independent of α

$$C = \frac{1}{\sqrt{2\pi} \sqrt{\frac{1}{2\alpha}}} \quad (5.4)$$

Gaussians allow the surface to be scanned within a given distance range. In this way a distance-weighted measure for the protein-ligand contact area is calculated, which is more related to the vdW interaction energy than the simple distance cutoff functional form.

Surface Directionality

To introduce surface directionality, surface points are excluded based on the direction of the surface normal \vec{n}_j relative to the current grid point \vec{r}_i . This applies to both versions of the score:

$$b_{dir}(\vec{r}_i) = \delta_{dir}(\gamma) \cdot b(\vec{r}_i) \quad (5.5)$$

$$b_{g,dir}(\vec{r}_i) = \delta_{dir}(\gamma) \cdot b_g(\vec{r}_i) \quad (5.6)$$

where

$$\delta_{dir}(\gamma) = \begin{cases} 1 & \text{if } \gamma \leq \gamma_{max} \\ 0 & \text{if } \gamma > \gamma_{max} \end{cases} \quad (5.7)$$

where γ_{max} is an adjustable cutoff angle and γ is the angle between $\vec{r}_i - \vec{P}_j$ and the surface normal \vec{n}_j computed as

$$\gamma = \arccos \left(\frac{(\vec{r}_i - \vec{P}_j) \cdot \vec{n}_j}{\|\vec{r}_i - \vec{P}_j\| \|\vec{n}_j\|} \right) \quad (5.8)$$

Atomic and Total Score

The atomic score $b(\vec{x}_k)$ for an atom k at position \vec{x}_k in a ligand pose is calculated as its local degree of burial, as given in the scoring grid. $b_{dir}(\vec{x}_k)$ is calculated using a trilinear interpolation of the eight orthogonal grid points closest to the atom. The total shape score $Bscore$ for a particular ligand pose is then the sum of the atomic contributions $b(\vec{x}_k)$ at each atom position \vec{x}_k (see equation 5.9). The same applies to $b_{g,dir}(\vec{x}_k)$ leading to the total shape score $Bscore_g$ of equation 5.10.

$$Bscore = \sum_k^{atoms} b_{dir}(\vec{x}_k) \quad (5.9)$$

$$Bscore_g = \sum_k^{atoms} b_{g,dir}(\vec{x}_k) \quad (5.10)$$

Direct Scoring

The orthogonal scoring grid is used in order to reduce computational costs when scoring many ligand poses. The scores b_{dir} or $b_{g,dir}$ need to be evaluated only on each grid point and the atomic scores can be computed extremely fast by trilinear interpolation. However, interpolation introduces an error which is dependent on the grid spacing and, in case of b_g , on the width of the Gaussian weighting function (α). Therefore, the scores can also be computed directly on the atom positions which avoids this error. However, in this case, the scores must be evaluated for each atom of each ligand pose separately which can lead to a significant increase in computational costs.

5.2.2 Electrostatic Term

Since $Bscore$ and $Bscore_g$ can be seen as approximations of the intermolecular van der Waals (vdW) energy between the ligand and the receptor, the introduction of an additional electrostatic correction term seems to be reasonable. For this purpose, the molecular electrostatic potential $\varphi(\vec{r}_i)$ generated by the protein is calculated and stored over the \vec{r}_i coordinates of an orthogonal three-dimensional grid using the Poisson-Boltzmann methods implemented in APBS

(version 1.3).⁹⁷ The MMFF94 force field,^{155,156,157,158,159} as implemented in OpenBabel (version 2.3.1),¹⁶⁰ is used for calculating partial charges q_k for all ligand and protein atoms. The electrostatic interaction energy between the receptor and a ligand atom k at position \vec{x}_k with a partial charge q_k is approximated as the product of the partial charge q_k and the value of φ at position \vec{x}_k . In this case the same interpolation procedure is used as for the estimation of $b(\vec{x}_k)$ and $b_g(\vec{x}_k)$. This leads to the following form for E_{score} , the electrostatic part of our scoring function:

$$E_{score} = \sum_k^{atoms} q_k \cdot \varphi(\vec{x}_k) \quad (5.11)$$

5.2.3 BEscore

By linearly combining the shape and the electrostatic terms described above, we introduce the new scoring functions BE_{score} and BE_{score}_g :

$$BE_{score} = B_{score} - \omega \cdot E_{score} \quad (5.12)$$

$$BE_{score}_g = B_{score}_g - \omega_g \cdot E_{score} \quad (5.13)$$

where ω and ω_g are linear weighting factors balancing the relative contributions of the two terms.

The new scoring functions are implemented using OpenStructure (version 1.2.2).⁸³

5.3 Sets of Receptor-Ligand Complexes

We applied our new scoring function to three different sets of receptor-ligand complexes. One set consists of the same receptor bound to different ligands (i.e. cross-docking), while the other two sets includes different receptors, each binding a different ligand (i.e. re-docking).

5.3.1 Thrombin Set

Thrombin inhibitor complex structures with a resolution of less than 2.0 Å were selected from the PDB.^{130,118,161} Structures with covalently bound ligands and peptides were ignored. All structures were aligned by superimposing the protein structures on the PDB structure 1etr¹⁶² using the binding site superposition algorithm implemented in Maestro (version 9.3, Schrodinger, LLC, New York, NY, 2012). The target structure 1etr was prepared by the Protein Preparation Wizard in Maestro (Suite 2011, Schrodinger, LLC). Hydrogen atoms were added to all ligands and the most stable protonation states and tautomers at pH 7.0 was generated using Epik (version 2.2, Schrdinger, LLC).¹⁶³ Standard precision (SP) Glide (version 5.7, Schrdinger, LLC)⁶⁷ was employed as the docking tool using default parameters. As in previous work of Graves et

al¹⁶⁴ the scoring function is supposed to discriminate ligand poses which are far away from the X-ray ligand structure (decoys) from the *best* poses.

Therefore, we generated a large set of differently placed protein-ligand complex conformations. For this purpose the experimentally observed ligand structure was used as starting geometry in the subsequent docking calculations. All poses were generated in various docking runs with 1etr as receptor. All ligands were excluded for which no acceptable conformation was found by the Glide docking run (i.e. symmetry corrected rmsd $\leq 2.0\text{\AA}$). This yielded a set of 66 thrombin inhibitor structures. On average 164 complex structures were generated with a minimum of 56 and a maximum of 276. Of those, on average, 17 are considered correct with an *rmsd* $\leq 2.0\text{\AA}$, with a minimum of 2 and a maximum of 90. Thus, this dataset gave rise to a retrieval rate of 10.2% when selecting ligand poses randomly. For re-scoring purposes, the functions $b(\vec{r}_i)$, $b_g(\vec{r}_i)$ and $\varphi(\vec{r}_i)$ (see equations 5.1, 5.3 and 5.11) were evaluated on the points of a three-dimensional orthogonal grid for the PDB structure 1etr, and were then applied to all 66 thrombin structures.

5.3.2 Astex Diverse Set

This set is a high-quality test set designed for the validation of protein-ligand docking performance.¹⁶⁵ It contains the structures of 85 diverse protein-ligand complexes retrieved from the PDB. These complexes share the following main characteristics: the ligands are drug-like; no particular target is represented more than once; the proteins are all drug discovery or agrochemical targets; only high quality structures are included. Major protein families are represented in the set, with 11 kinases, 9 nuclear receptors, 5 serine proteases and 3 members of the phosphodiesterase family.

Although the Astex diverse set is a high quality test set, a number of significant problems were observed, mainly due to wrongly parametrized co-factors observed in the structure. Thus, all complexes were visually inspected and manually cleaned prior to the docking. Subsequently, for each protein-ligand complex we applied the same preparation, docking and re-scoring procedure as for the thrombin test set using Glide SP. On average 141 poses were generated with a minimum of 22 and a maximum of 447 of which, on average, 36 were correctly placed, with a minimum of 3 and a maximum of 152. This yields a random retrieval rate of 28.8%. No alignment had to be performed and the functions $b(\vec{r}_i)$, $b_g(\vec{r}_i)$ and $\varphi(\vec{r}_i)$ (see equations 5.1, 5.3 and 5.11) were individually calculated on the points of three-dimensional orthogonal grids for each receptor.

5.3.3 S3DB

S3DB¹⁶⁶ is a database of manually curated protein-ligand complex structures, based on the ligand-protein database¹⁶⁷ which contains a diverse set of protein families and ligand chemotypes.

Similar to the Astex diverse set, all complex structures in S3DB were prepared for docking by the Protein Preparation Wizard in Maestro (Suite 2011, Schrodinger, LLC). Structures

generating errors in Maestro were excluded from further analysis. For each complex structure the corresponding ligand was docked using standard precision Glide. Structures where no correct pose (i.e. $rmsd \leq 2.0\text{\AA}$) was generated were discarded. This led to a set of 145 protein-ligand complex structures. On average, 127 poses were generated (minimum: 20, maximum: 311) of which, on average, 29 are correctly placed (minimum: 1, maximum: 150). This yields a random retrieval rate of 24.7%. No alignment had to be performed and the functions $b(\vec{r}_i)$, $b_g(\vec{r}_i)$ and $\varphi(\vec{r}_i)$ (see equations 5.1, 5.3 and 5.11) were individually calculated on the points of three-dimensional orthogonal grids for each receptor.

5.4 Validation

In order to assess the prediction performance we computed four commonly used measures: retrieval rate,¹⁶⁴ enrichment area under curve (AUC), Δ RMSD and Pearson's correlation coefficient.

Retrieval Rate

Given a set of receptor-ligand complexes such as the thrombin or the Astex diverse sets described in the previous section, the retrieval rate is the percentage of complexes where the ligand pose ranked best by the scoring function is within 2.0 Å symmetry corrected root mean square deviation (rmsd) of the X-ray structure.

Enrichment AUC

Enrichment area under the curve is a measure for the ability of a scoring function to rank correct poses higher than incorrect ones, where correct poses are defined as conformations with an symmetry corrected rmsd $\leq 2.0\text{\AA}$.

Δ RMSD

For a set of docking poses for a particular ligand, Δ RMSD is the difference in symmetry corrected rmsd between the best scored pose and the crystal structure.

In the following paragraphs all four scores resulting from $Bscore$, $Bscore_g$, $Escore$, $BEscore$ and $BEscore_g$ have been calculated for the thrombin, the S3DB and the Astex diverse set.

5.4.1 Shape Term

According to our model the shape term of the scoring function can have two functional forms: $Bscore$ (equation 5.9) and $Bscore_g$ (equation 5.10) both of which depend on different parameters (equations 5.2 and 5.3).

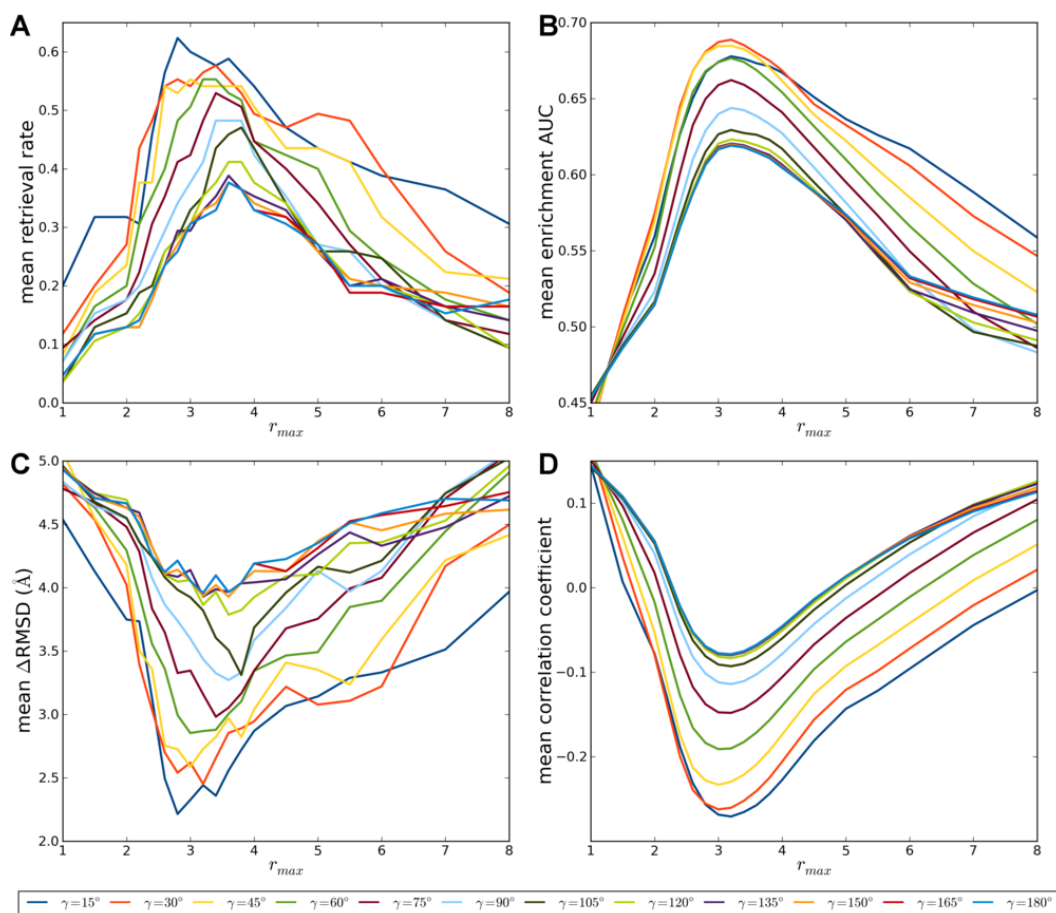


Figure 5.2: Evaluation of the performance of *Bscore* based on the Astex diverse set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

Bscore

Equation 5.2 depends only on the parameter r_{max} , whereas when including surface directionality (Equation 5.5), the parameter γ_{max} is added. Figure 5.2 and 5.3 show the performance for the Astex diverse set and the thrombin set (see Appendix A.3 Figure A.3 for the S3DB test set) for different γ_{max} values. For the Astex diverse set and the S3DB set the highest performance values were obtained for an r_{max} value between 3.2 and 3.4 Å and γ_{max} between 15 and 30°, with retrieval rates of 56.5% and 52.4% and enrichments of 68.9% and 65.3%, respectively. For the thrombin set, a broader maximum is observed with the same optimal γ_{max} value (30°) but slightly larger optimal r_{max} values (in the range of 4.0 to 4.5 Å) were observed, with retrieval rates of 52.5% and enrichments of 86.4%.

Since the Astex diverse set and the S3DB set represent a broader spectrum of binding site properties and ligand chemotypes and the optimal values obtained for those two sets are in good agreement, the optimal values for *Bscore* parameters were chosen as follows:

$$r_{max} = 3.4 \text{ \AA} \text{ and } \gamma_{max} = 30^\circ$$

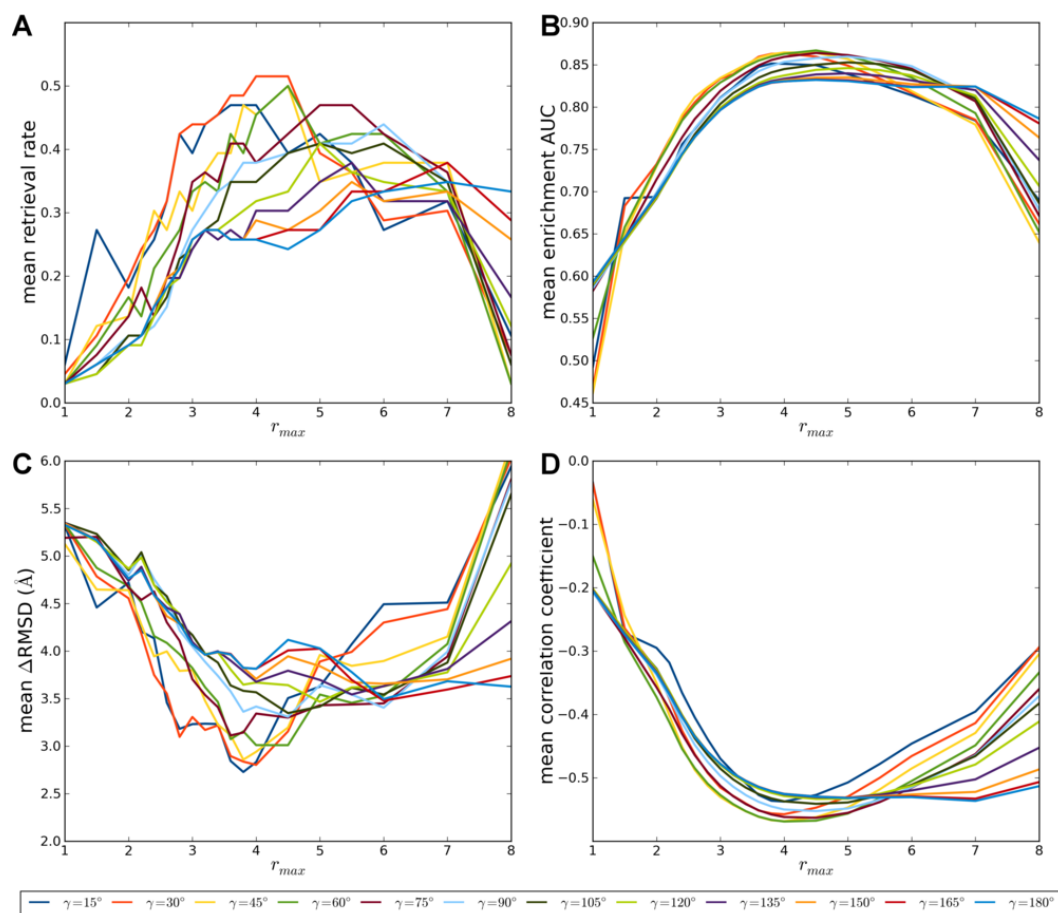


Figure 5.3: Evaluation of the performance of $Bscore$ based on the thrombin set using (A) retrieval rate, (B) enrichment AUC, (C) $\Delta RMSD$ and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

$Bscore_g$

Equation 5.3 depends on two parameters, r_{max} and α , and when including surface directionality (Equation 5.6), the parameter γ_{max} is added. Figures 5.4 and 5.5 show the retrieval rates as a function of r_{max} and α for different γ_{max} values for the Astex diverse set and the thrombin set, respectively (see Appendix A.3 Figure A.4 for the S3DB test set). For the thrombin set the highest retrieval rates (56.1%) and enrichments (75.2%) are achieved for r_{max} values around 3.0 Å, $\ln(\alpha)$ values between 0 and 3 and γ_{max} value 45°. In the case of the Astex diverse set the highest retrieval rates (up to 65.9%) and enrichments (69.6%) are observed for r_{max} values between 2.3 and 2.4 Å, $\ln(\alpha)$ values between 2 and 4 and a γ_{max} value 45°. For the S3DB set similar optimal parameters were obtained as for the Astex diverse set (r_{max} between 2.3 and 2.4 Å, $\ln(\alpha)$ between 0 and 4, γ_{max} 45°) with highest retrieval rates (60.0%) and enrichments (65.9%).

The distribution of the retrieval rates for the sets depicted in Figures 5.4, 5.5 and Figures A.4 and the fact that the Astex diverse set and the S3DB set represent a broader spectrum of binding site properties and ligand chemotypes, led us to choose optimal values for $Bscore_g$ parameters

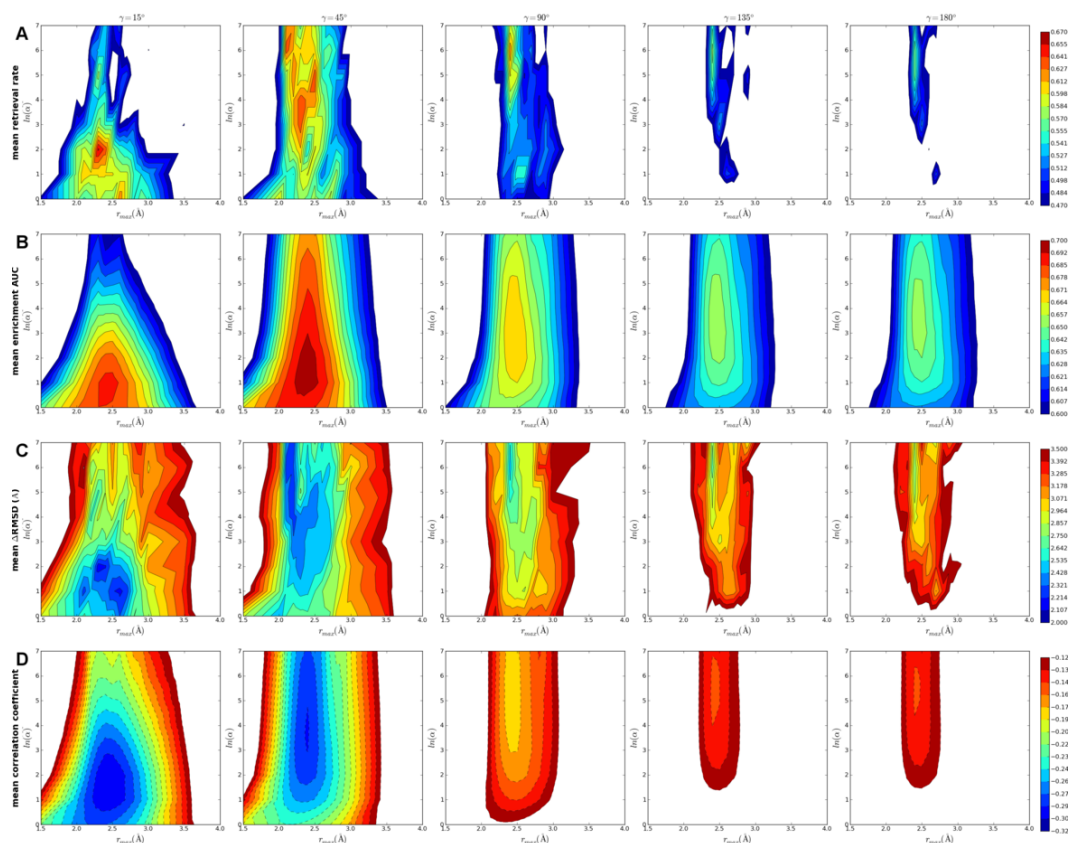


Figure 5.4: Evaluation of the performance of $Bscore_g$ based on the Astex diverse set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

with the following values:

$$r_{max} = 2.4 \text{ \AA}, \ln(\alpha) = 4.0 \text{ and } \gamma_{max} = 45^\circ$$

The optimal r_{max} values for $Bscore$ and for $Bscore_g$ ($r_{max} = 3.4 \text{ \AA}$ and $r_{max} = 2.4 \text{ \AA}$, respectively) are not far from typical vdW contact distances. This suggests that the $Bscore$ and $Bscore_g$ values approximate the vdW energy, which correlates with the size of the contact surface area.

Directionality

When adding the directionality term γ_{max} to $Bscore$ or $Bscore_g$, a significant improvement in the performance scores can be observed compared to when using no directionality restrictions (i.e. $\gamma_{max} = 180^\circ$).

For $Bscore$ the retrieval rate increases from 38% to 56.5% (Astex diverse set, Figure 5.2), from 34% to 52.4% (S3DB set, Figure A.3) and from 28% to 52.5% (thrombin set, Figure 5.3). The enrichments increase from 62% to 68.9% (Astex diverse set), from 62% to 65.3% (S3DB set) and from 84% to 86.4% (thrombin set).

For $Bscore_g$ the retrieval rate increases from 55% to 65.9% (Astex diverse set, Figure 5.4),

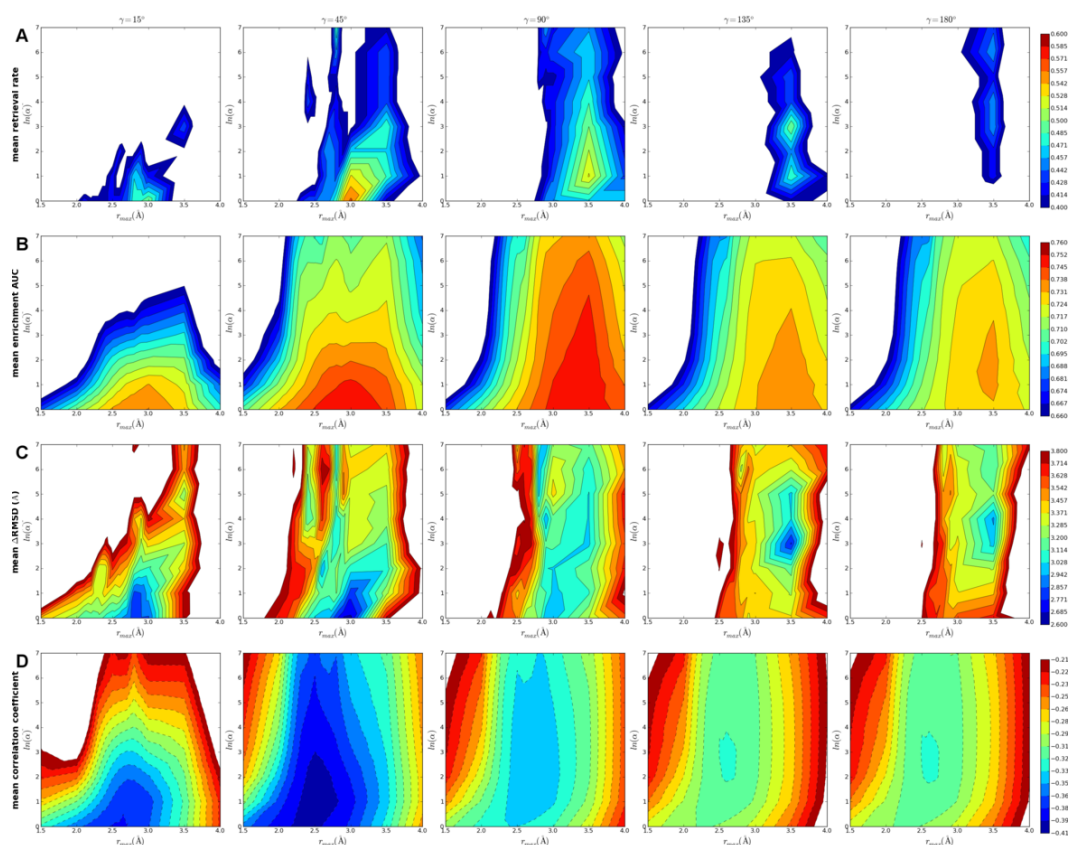


Figure 5.5: Evaluation of the performance of $Bscore_g$ based on the thrombin set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

from 46% to 60.0% (S3DB set, Figure A.4) and from 44% to 56.1% (thrombin set, Figure 5.5). The enrichments increase from 65% to 69.6% (Astex diverse set), from 63% to 65.9% (S3DB set) and from 73% to 75.2% (thrombin set).

5.4.2 Electrostatic Term

The electrostatic potential $\varphi(\vec{r}_i)$ in equation 5.11 (and consequently $Escore$) depends, among other parameters, on the value of the dielectric constant ϵ which is assigned to the interior of the protein, and on the type and concentration of mobile ions. We tested the effect of these two parameters on the scoring performance of the electrostatic term alone. For ϵ we tried different values ranging from 1 to 80. For the mobile ion charge density we tested the following three configurations: the absence of mobile ions (in this case the Poisson-Boltzmann equation becomes a Poisson equation) and the presence of 0.15 mol/L or 0.30 mol/L of sodium chloride (NaCl).

The performance of $Escore$ is shown in Figure 5.6 for the Astex diverse set and Figure 5.7 for the thrombin set (see Appendix A.3 Figure A.5 for S3DB). The highest retrieval rates and enrichments for the Astex diverse set (60.0% and 63.8%, respectively) and the S3DB set (49.1% and 66.9%, respectively) are obtained for a dielectric constant between ϵ between 20

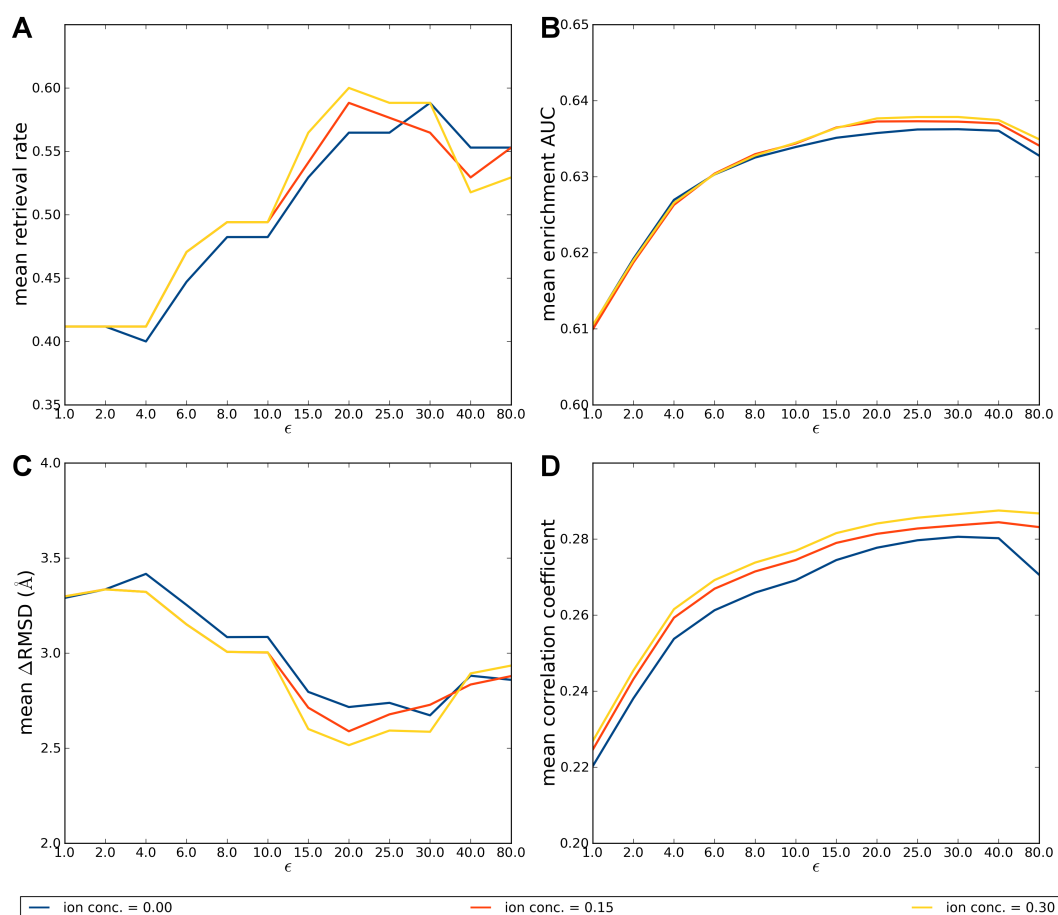


Figure 5.6: Evaluation of the performance of *E_{score}* based on the Astex diverse set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

and 30 and in the presence of 0.15 to 0.30 mol/L NaCl. For the thrombin set, *E_{score}* based retrieval rates are low (21.2%) whereas enrichments are in a similar range as for the other sets. Optimal retrieval rates are obtained for ϵ values around 6.0 and in the presence of 0.15 to 0.30 mol/L NaCl. Those parameters yield also optimal enrichments (63.5%) when not considering enrichments at a mobile ion concentration of 0.00 mol/L, in which case the average is dominated by a small number of outliers.

Since the Astex diverse set and the S3DB set represent a broader spectrum of binding site properties and ligand chemotypes, the optimal values for *E_{score}* parameters were chosen as follows:

$$\epsilon = 25, \text{ mobile ion concentration} = 0.15 \text{ mol/L}$$

5.4.3 Summary of Individual Terms

The optimized retrieval rates and enrichment values are summarized in Table 5.1. We emphasise the fact that the retrieval rates obtained using the local degree of burial alone (65.9% for the

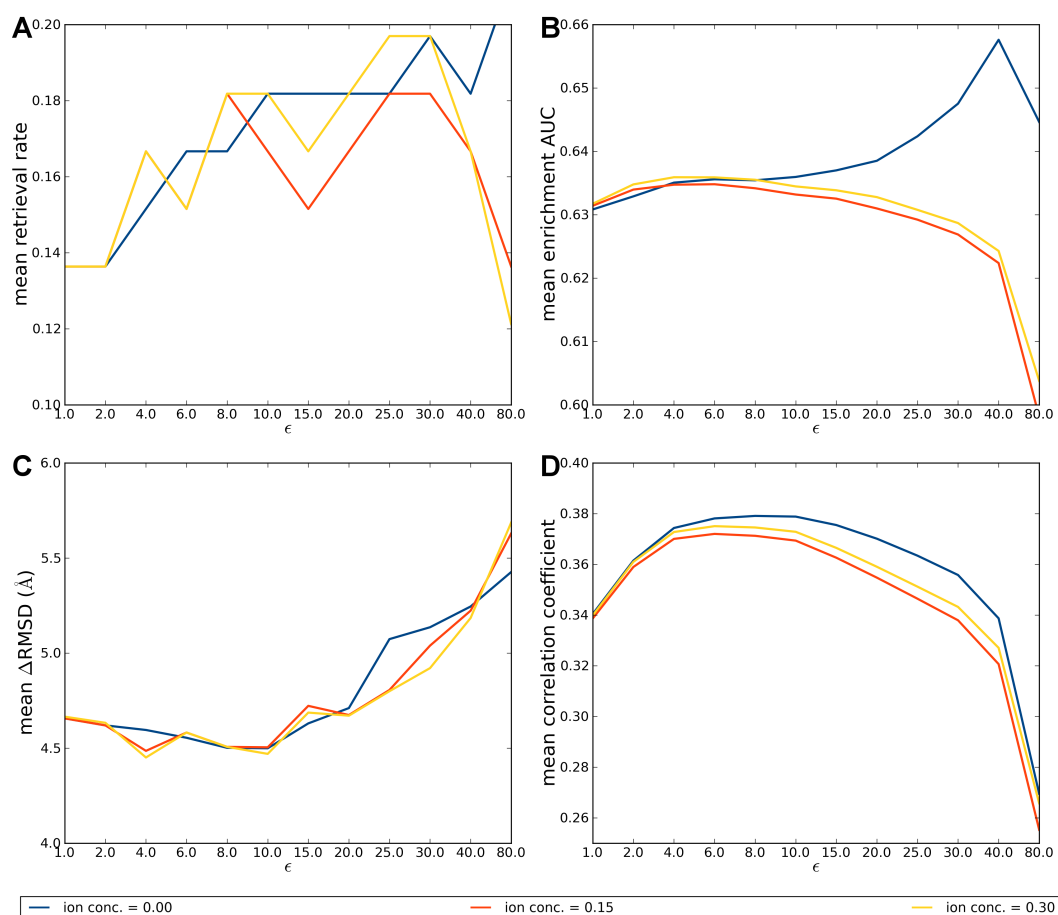


Figure 5.7: Evaluation of the performance of *E_{score}* based on the thrombin set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

Astex diverse set and 56.1% for the thrombin set, using the Gaussian functional form) are higher than the retrieval rates obtained using the electrostatic term alone (60.0% for the Astex diverse set and 21.2% for the thrombin set). The same holds true for enrichment values. This confirms the notion that, generally, steric complementarity plays a larger role in molecular recognition than polar and electrostatic interactions.¹³⁶ It should be mentioned that reasonable polar interactions were already generated by the docking program prior to these re-scoring calculations.

5.4.4 Comparison to Van der Waals Interaction Energies

Since the shape terms are related to standard van der Waals (vdW) interactions and might simply approximate the latter, we compare the results obtained with *B_{score}* or *B_{score_g}* for the Astex diverse set with vdW interaction energies as computed by Glide SP. Correlations are computed for all poses of each target of the Astex diverse set and a histogram of the correlation coefficients is shown in Figure 5.8. On average, a correlation coefficient of -0.51 and -0.48 is obtained for *B_{score}* and *B_{score_g}*, respectively. This intermediate average correlation

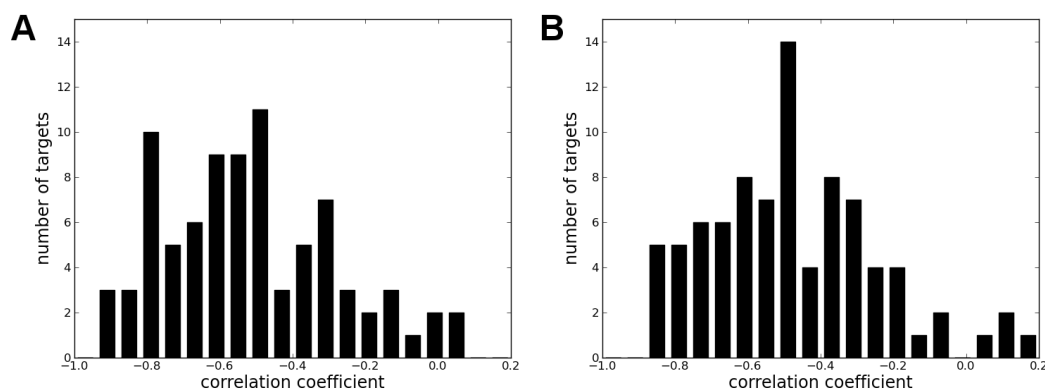


Figure 5.8: Histogram of the correlation between the shape terms (A: $Bscore$, B: $Bscore_g$) and standard van der Waals interactions as computed by Glide SP. Correlations are computed for all poses of each target of the Astex diverse set.

coefficient indicates a certain degree of similarity between the two methods but clearly shows that there are significant differences between the two.

5.4.5 BEScore

As defined in equations 5.12 and 5.13, $BEScore$ or $BEScore_g$ linearly combines $Bscore$ or $Bscore_g$ with $Escore$ via the ω or ω_g parameter. In order to determine the optimal value of ω and ω_g for both equations, we used the optimised parameters for $Bscore$, $Bscore_g$ and $Escore$ as described in the previous paragraphs:

$$Bscore: r_{max} = 3.4 \text{ \AA}, \gamma_{max} = 30^\circ$$

$$Bscore_g: r_{max} = 2.4 \text{ \AA}, \ln(\alpha) = 4.0, \gamma_{max} = 45^\circ$$

$$Escore: \varepsilon = 25, \text{ mobile ion concentration} = 0.15\text{mol/L}$$

Figure 5.9 shows the retrieval rate of $BEScore$ as a function of ω for the Astex diverse sets, the thrombin set and the S3DB set. The highest retrieval rates and enrichments for the Astex diverse set (65.9% and 73.3%, respectively) are obtained for $\omega = 360$, for the thrombin set (53.0% and 86.3%, respectively) for $\omega = 180$, while for the S3DB set (67.2% and 72.1%, respectively) for $\omega = 160$. Figure 5.10 shows the retrieval rate of $BEScore_g$ as a function of ω_g for both sets. For the Astex diverse set the highest retrieval rates and enrichments (71.8% and 73.5%, respectively) are observed at $\omega_g = 330$, for the thrombin set (45.5% and 84.3%, respectively) for $\omega_g = 230$ and for the S3DB set (69.3% and 72.7%, respectively) for $\omega_g = 230$. The optimized retrieval rates and enrichment values are summarized in Table 5.1.

On average $BEScore_g$ exhibits higher retrieval rates and enrichments compared to $BEScore$. From the distributions in Figures 5.9 and 5.10 we chose as optimal value for $\omega = 180$ and $\omega_g = 240$.

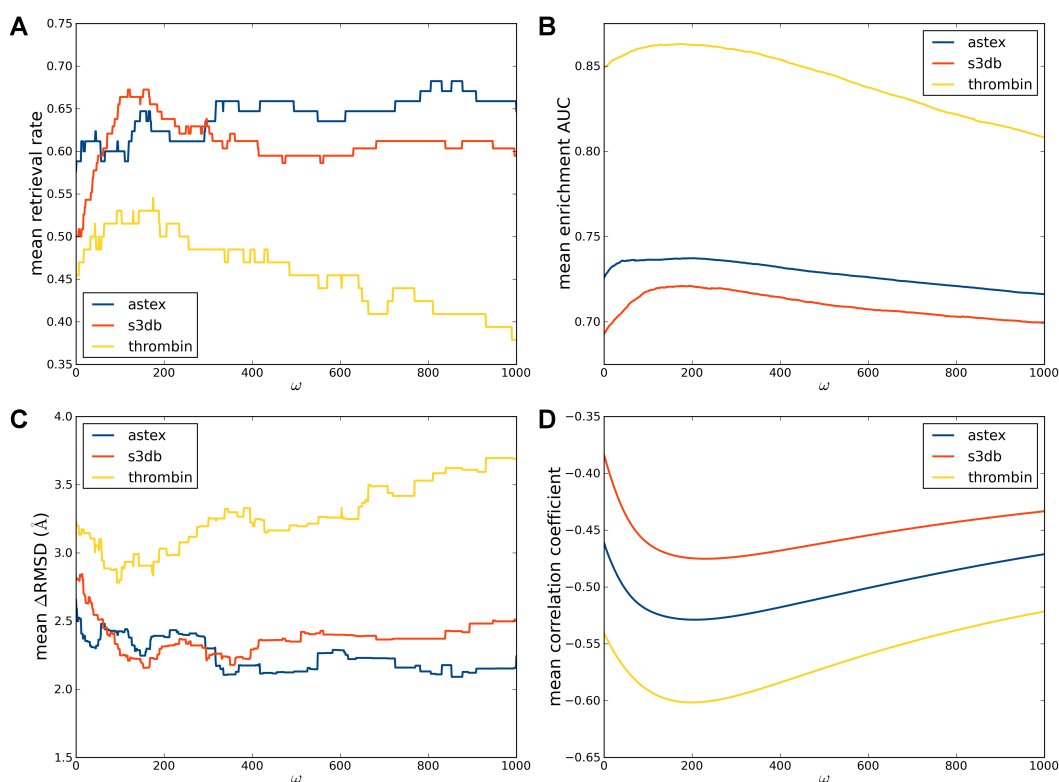


Figure 5.9: Evaluation of the performance of *B*Score based on the Astex diverse set, the S3DB set and the thrombin set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

Table 5.1: Summary of the results obtained from optimizing the parameters of *B*score, *B*score_g, *E*score, *B*EScore and *B*EScore_g. The retrieval rates and enrichments obtained with the optimal parameters are shown for the three test sets. ^a Best parameters were optimized for each set individually. ^b Selected parameters were selected to perform well on all sets.

		best parameters ^a			selected parameters ^b		
		Astex	S3DB	Thrombin	Astex	S3DB	Thrombin
<i>B</i> score	retrieval	56.5%	52.4%	52.5%	57.6%	52.4%	45.5%
	enrichment	68.9%	65.3%	86.4%	68.5%	65.3%	84.9%
<i>B</i> score _g	retrieval	65.9%	60.0%	56.1%	62.4%	57.2%	43.9%
	enrichment	69.6%	65.9%	75.2%	68.9%	64.9%	71.9%
<i>E</i> score	retrieval	60.0%	49.1%	21.2%	57.6%	49.1%	18.2%
	enrichment	63.8%	66.9%	63.5%	63.7%	66.8%	62.9%
<i>B</i> EScore	retrieval	65.9%	67.2%	53.0%	62.4%	65.5%	53.0%
	enrichment	73.3%	72.1%	86.3%	73.7%	72.1%	86.3%
<i>B</i> EScore _g	retrieval	71.8%	69.3%	45.5%	69.4%	69.3%	45.5%
	enrichment	73.5%	72.7%	84.3%	73.7%	72.7%	84.2%

5.5 Analysis of Surface Point Distribution

The distribution of surface points around ligand atoms was analyzed for the 85 protein-ligand complexes of the Astex diverse set.

For each complex, all ligand poses were divided into correct and incorrect poses according to

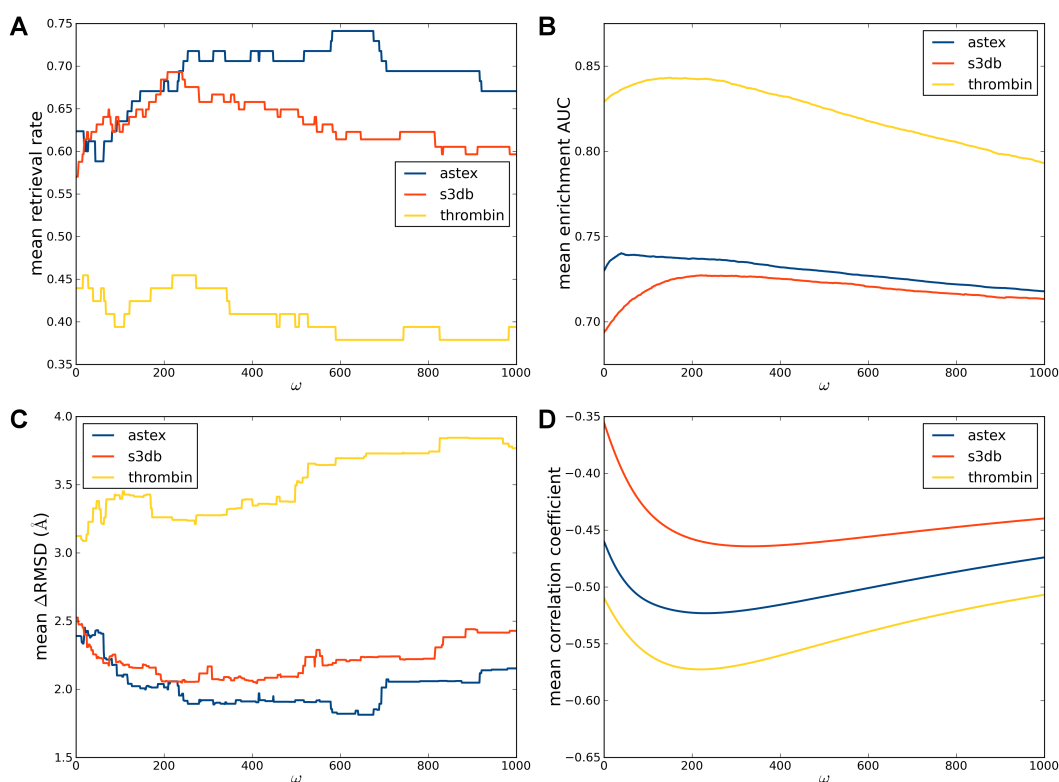


Figure 5.10: Evaluation of the performance of $BScore_g$ based on the Astex diverse set, the S3DB set and the thrombin set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

their all-atom RMSD values of $\leq 2.0 \text{ \AA}$ (correct) and $> 4.0 \text{ \AA}$ (incorrect). For each atom of each ligand pose, a histogram of the distribution of surface points versus the distance from the atom is computed. For each ligand pose the per-atom distributions are summed and a normalized histogram is computed. For both, the correct and the incorrect poses, the distributions from all contributing ligand poses are averaged and scaled to the range from 0.0 to 1.0. The difference between the distribution of correct and incorrect poses is plotted in Figure 5.11. The actual distributions are shown in Figure A.2 in Appendix A.3. The procedure was repeated with different cutoff values for angle between the vector from the atom to the surface vertex and the normal vector of the surface vertex (γ_{max}).

For all different γ_{max} values, a maximum in the distribution difference is observed at $r = 2.3 \text{ \AA}$. This corresponding to a distance range where the correct poses show an increased surface point density compared to the incorrect poses. Thus, exclusively selecting distances in a range around 2.3 \AA should yield the highest separation between correct and incorrect poses and thus the best scoring performance.

Figure 5.12 shows an enlarged part of the surface point distribution difference for $\gamma_{max} = 180^\circ$ as a representative example. The maximum of the difference between the distributions (black line) can be fitted well by a Gaussian curve (red line) with a peak (r_{max}) at 2.3 \AA and a width ($\ln(\alpha)$) of 2.1. This indicates that selecting a Gaussian shaped function to weight the individual contribution of surface points to the per-atom $BScore_g$ seems reasonable.

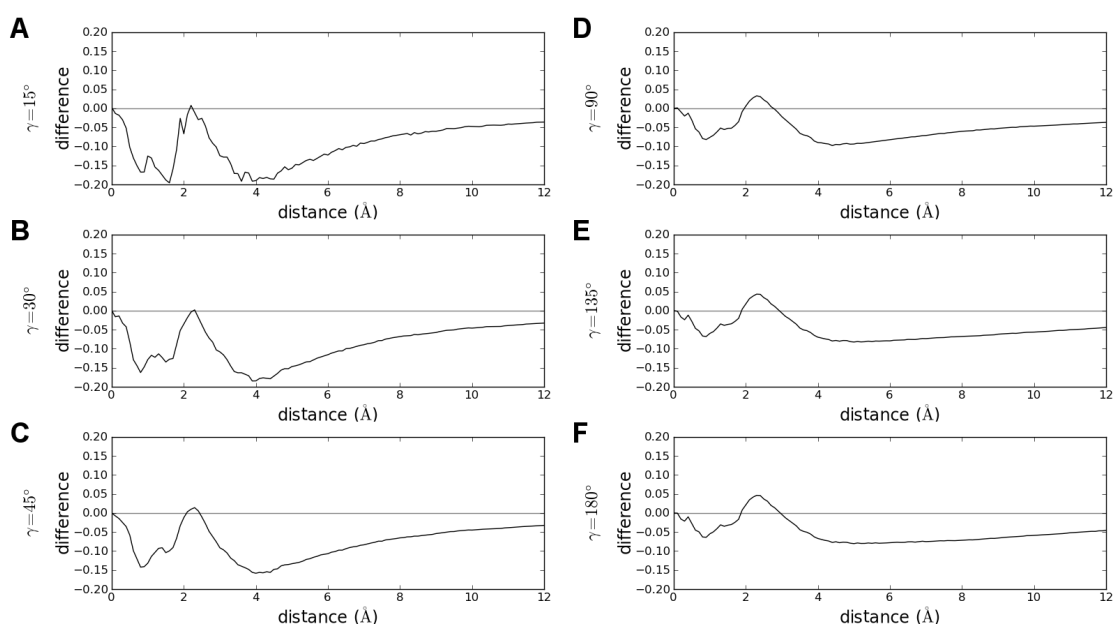


Figure 5.11: Difference between the distributions of surface vertex points around the ligands of the Astex diverse set. Ligands are categorized into correct ($rmsd \leq 2.0 \text{ \AA}$) and incorrect poses ($rmsd > 4.0 \text{ \AA}$) based on their rmsd to the crystal structure. The difference between the scaled distributions of correct and incorrect poses is plotted for different γ_{max} angles: (A) 15° , (B) 30° , (C) 45° , (D) 90° , (E) 135° , (F) 180° .

When comparing the distributions for different γ_{max} values, a more pronounced difference is observed for small γ_{max} values up to $\gamma_{max} = 45^\circ$. This indicates that small γ_{max} values should better discriminate correct from incorrect ligand poses. Since the number of included surface points decreases with smaller γ_{max} , the noise in the distributions increases. Thus, an optimal γ_{max} is in the range of $30 - 45^\circ$.

It is worth noting that the values for r_{max} and α obtained through the surface point distribution are in excellent agreement with the values obtained from parameter optimization of $Bscore_g$ ($r_{max} = 2.4 \text{ \AA}$, $\ln(\alpha) = 4.0$).

5.6 Results

5.6.1 Astex Diverse Set

Figure A.6 (Appendix A.3) shows $Bscore_g$ as a function of the rmsd for each docked pose from the relative X-ray ligand position for the 85 complexes in the Astex diverse set. The following combination of parameters was used: $r_{max} = 2.4 \text{ \AA}$, $\ln(\alpha) = 4.0$, $\gamma_{max} = 45^\circ$ and $\omega_g = 240$. For those parameters the retrieval rate is 69.4%, the enrichment is 73.7%, $\Delta RMSD$ is 1.95 \AA and the correlation coefficient is -0.52.

Figure 5.13 shows an example for the target structure 2bm2 of the Astex diverse set with atomic $Bscore_g$ values mapped onto the ligand structures using a gradient from blue (low score) to red (high score). The ligand pose (A) resembles closely the X-ray ligand conformation ($rmsd = 0.2 \text{ \AA}$) and has a high score ($Bscore_g = 2160$) whereas pose (B) is incorrectly placed with an rmsd of 3.0 \AA and has a lower score ($Bscore_g = 1499$). The coloring represents well,

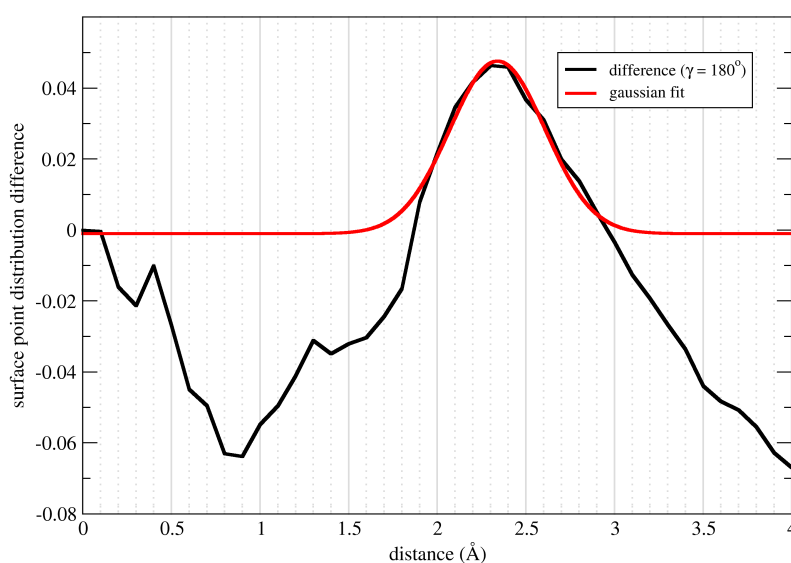


Figure 5.12: Fit of a Gaussian curve (red line) to the difference between the distributions of surface vertex points around the ligands of the Astex diverse set (black line) for $\gamma_{max} = 180^\circ$.

which parts of the ligand are incorrectly placed and not in intimate contact with the protein surface.

In 59 cases the top ranked structure was within 2.0 Å rmsd. In three additional cases the best ranked structure was between 2.0 and 2.5 Å rmsd and in three more cases between 2.5 and 3.0 Å rmsd, while in the remaining 20 cases the top ranked structure deviated more than 3.0 Å rmsd from the X-ray structure.

If not only the best ranked pose, but the two, five or ten best ranked poses are considered, the retrieval rate increases significantly from 69.4% to 76.5%, 87.1% and 94.1%, respectively. This illustrates that in most cases accurate poses are found within the first few top ranked conformations. In fact, only five complex structures did not yield a correctly placed ligand pose within the ten best ranked structures.

5.6.2 Thrombin Set

Figure A.8 (Appendix A.3) shows BE_{score_g} as a function of the rmsd for each docked pose from the relative X-ray ligand position for the 66 thrombin inhibitor complex structures. The same combination of parameters as in the paragraph above was used. The retrieval rate is 45.5%, the enrichment is 84.2%, $\Delta RMSD$ is 3.24Å and the correlation coefficient is -0.57.

In 30 cases the top ranked structure was within 2.0 Å rmsd of the X-ray ligand crystal structure. For additional six cases the deviation was found to be between 2.0 and 2.5 Å rmsd and in four more cases between 2.5 and 3.0 Å. The remaining 26 cases the top ranked structure deviated more than 3.0 Å rmsd from the X-ray structure.

The thrombin set shows a relatively low retrieval rate but a high enrichment, which suggests that when including not only the best ranked pose in the calculation, the retrieval rates should increase significantly. Retrieval rates increase from 45.5% to 50.0%, 74.2% and 87.9% for

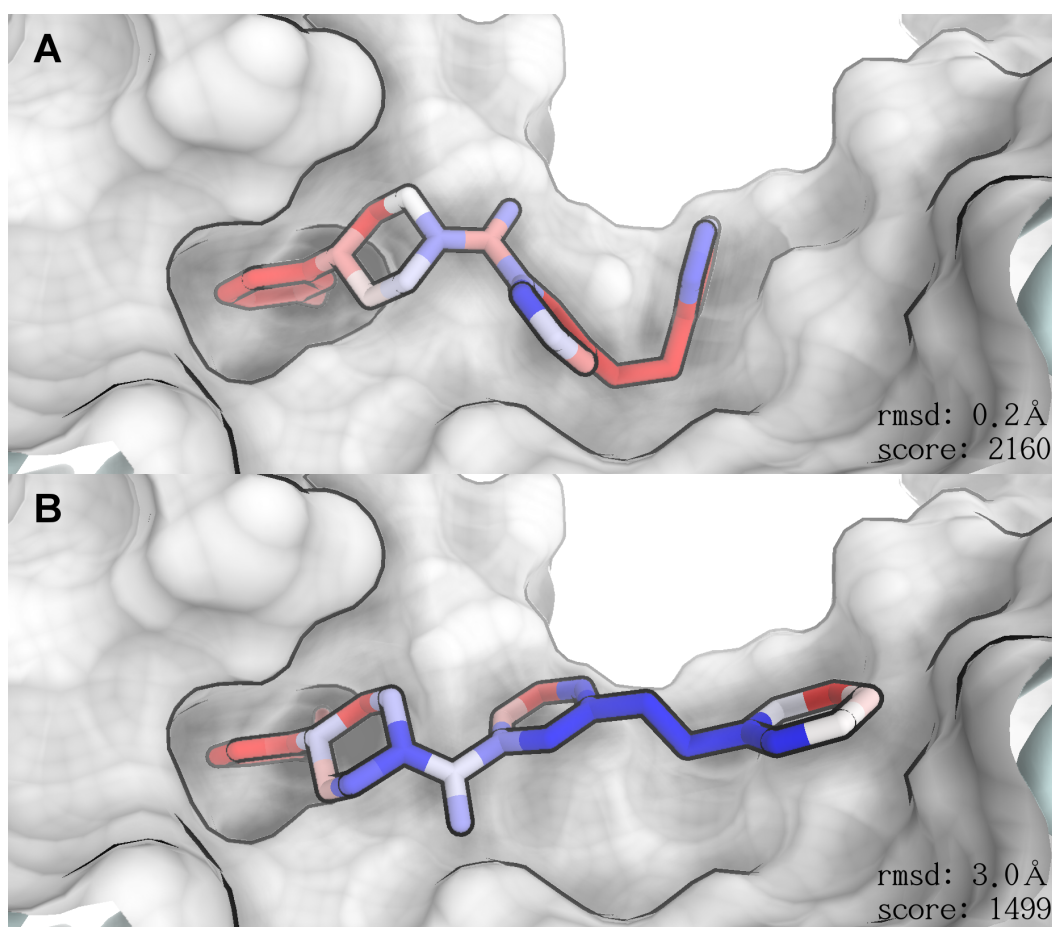


Figure 5.13: Two arbitrarily chosen poses for target 2bm2 of the Astex diverse set with atomic $Bscore_g$ values mapped onto the ligand structures using a gradient from blue (low score) to red (high score). The protein is represented as a molecular surface, whereas the ligand is shown in stick representation. The ligand pose (A) resembles closely the X-ray ligand conformation ($rmsd = 0.2 \text{ \AA}$) whereas pose (B) is incorrectly placed with an $rmsd$ of 3.0 \AA . The overall score is computed as the sum of atomic $Bscore_g$ contributions.

including the two, five and ten best ranked poses, respectively. Thus, when including the ten best ranked poses, only eight out of the 66 thrombin structures did not yield a correctly placed ligand pose.

5.6.3 S3DB

Figure A.7 (Appendix A.3) shows $BEscore_g$ as a function of the rmsd for each docked pose from the relative X-ray ligand position for the 66 thrombin inhibitor complex structures. The same combination of parameters as in the paragraph above was used. The retrieval rate is 69.3%, the enrichment is 72.7%, $\Delta RMSD$ is 2.05 \AA and a correlation coefficient of -0.46 was found.

In 100 cases the top ranked structure was within 2.0 \AA rmsd of the X-ray ligand structure. In ten additional cases the rmsd of the top ranked pose is between 2.0 and 2.5 \AA and in twelve cases between 2.5 and 3.0 \AA . The remaining 23 structures exhibit an rmsd greater than 3.0 \AA .

As with both the Astex diverse set as well as the thrombin set, the retrieval rates increase

significantly when not only the top ranked pose is considered. Retrieval rates increase from 69.3% to 77.2%, 90.4% and 95.6% when including the two, five and ten best scored poses, respectively.

5.7 Comparison to X-Score and Glide SP

For a direct comparison, we compared our results to the Glide SP⁶⁷ results and additionally applied the X-Score scoring function¹⁶⁸ to the Astex diverse sets.

For X-Score, retrieval rates of 68.2% and enrichments of 73.5% were obtained. Those values are very similar to what we obtained with $BEscore_g$ (retrieval rate: 69.4%, enrichment: 73.7%). Thus, $BEscore_g$ seems to be able to compete well with X-Score and thus other comparable scoring functions.

For Glide standard precision, retrieval rates of 83.5% and enrichments of 77.2% were obtained. Although the performance values obtained using $BEscore_g$ (retrieval rate: 69.4%, enrichment: 73.7%) are slightly lower than what was obtained using Glide SP, the performance of $BEscore_g$ is still very promising, especially when considering its simplicity compared to a full blown scoring function as Glide SP.

5.8 Combining with X-Score and Glide SP

Our new scoring function $BEscore_g$ is intended for re-scoring of docking poses and thus as a post docking filtering step. As such, ligand poses generated by a docking/scoring method would then be filtered based on $BEscore_g$ with the aim of reducing false positive placements.

To validate such a strategy, $BEscore_g$ has been combined with the two scoring functions X-Score and Glide SP on the Astex diverse set. All ligand poses, generated as described above, were first scored using X-Score or Glide SP score and were subsequently re-scored using $BEscore_g$. Only the ten poses ranked best according to $BEscore_g$ were retained and retrieval rates according to X-Score or Glide SP score were computed. Using this strategy, retrieval rates could be significantly improved for X-Score from 68.2% to 76.5% and slightly for Glide SP from 83.5% to 84.7%.

5.9 Discussion

The retrieval rates obtained for the Astex diverse set (69.4%) and for the S3DB set (69.3%) using $BEscore_g$ are very promising and they outperform a number of commonly used scoring functions. Retrieval rates between 70 and 80% for identifying the best docking poses from a number of poses can be reached with a number of scoring functions, including CHARMM,¹⁶⁹ DOCK,¹⁷⁰ DOCK6,¹⁷¹ ChemScore,¹⁷² DrugScore^{PDB},¹⁷³ AutoDock,¹⁷⁴ EADock¹⁷⁵ and Lead Finder¹⁷⁶ while the PMF function¹⁷⁷ achieved about 60%. A study by Velec et al. presented DrugScore^{CSD}, a knowledge-based scoring function with a retrieval rate of 87% for 100 protein-

ligand complexes,¹⁷⁸ and even higher retrieval rates of 91% have been claimed for ICM.^{146,179} These are, to the best of our knowledge, the highest figures reported to date. The study by Velec also reported retrieval rates for a number of scoring functions, including, for example, Cerius2/PLP (76%), DrugScore^{PDB} (72%), AutoDock (62%), Cerius2/PMF (42%), and SYBYL/Chemscore (35%). These few examples highlight the difficulties of comparing pose retrieval rates, which can differ significantly depending on the targets, the number of decoys, and the way in which the pool of decoy poses are generated. In our study, we used up to 447 poses generated by Glide SP, in contrast to up to 100 decoys used in Velec's work, and an unknown number in the ICM study. This underlines the influence of the pose/decoy generation method, and the need for a publicly available and commonly agreed standardised test set that comprises a well defined set of decoys for different protein-ligand complex structures which adequately cover the space of potential poses, and allow for more objective benchmarking procedures.

Since comparing different methods on different sets of protein-ligand complex structures or different numbers of correct and decoy poses is difficult,¹⁵² we compare our results directly with other commonly used scoring functions. We compared our results to the Glide SP⁶⁷ results and additionally applied the X-Score scoring function¹⁶⁸ to the Astex diverse sets. X-Score gave comparable results whereas Glide SP slightly outperformed *BEscore_g*. The performance of *BEscore_g* is thus very promising, especially when considering its simplicity compared to a full blown scoring function as Glide SP. In addition, two points should be considered. First, the Astex diverse set is a very commonly used test set to evaluate and train scoring functions. Since this set was released previous to the current version of Glide SP, it is likely that the Glide SP scoring function has been trained in order to obtain good results on this test set. Second, Glide SP might have an advantage since it was used in order to generate the ligand poses and thus, Glide SP scores were optimized in contrast to *BEscore_g* values which were only used for re-scoring.

Since *BEscore_g* is intended for re-scoring of docking poses and thus as a post docking filtering step, we evaluated the performance of *BEscore_g* when used to filter docking poses scored with the scoring functions X-Score and Glide SP. Using this strategy, retrieval rates were significantly improved for X-Score and slightly for Glide SP. This first proof of concept shows that such a strategy has high potential and that *BEscore_g* could significantly improve the current methodologies by reducing the problem of a large number of false placements.

Another important aspect is the general applicability of the scoring function parameters which could in principle be adjusted for a particular protein family or for a particular class of ligands. An ideal scoring function should of course be invariant of target, cavity and ligand characteristics. To better reach this goal we have chosen to work with the three different sets of protein-ligand complexes presented in section 5.3. In section 5.4 we have pursued an extensive analysis and validation of all the scoring function parameters: r_{max} , α , γ_{max} , ω , ω_g , ε and the ion salts concentration. This has shown that in the case of *BEscore_g* the optimal values for α , γ_{max} , the ion salts concentration and ω are very similar for the three sets of protein ligand complexes. The optimal values of r_{max} , α and ε are nearly identical for the Astex diverse

set and the S3DB set but show some small differences for the thrombin set. However, using an intermediate value for all parameters the retrieval rate for both sets drops only about 2.5% whereas the enrichments stays constant. This proves that our scoring function is robust towards its parameters.

The method opens up a number of new options for structure-based drug design. For example, re-scoring could be done on selected portions of the cavity, in order to find fragments that can bind to particular sub-pockets of a binding site. A ligand efficiency measure^{180,181} for the degree of burial for ligands could be introduced. Re-scoring grids derived from a series of superimposed structures could be easily derived. These would, for example, allow representation of the highest possible degree of burial in a binding site exhibiting flexibility, and could help to account for small induced fit effects. In addition, multiple scoring grids could be easily derived for protein structures including different solvent molecules. Thus, the effect of displaceable water molecules could be accounted for. Obviously, another option is to complement the array of existing scoring functions for use with consensus scoring approaches.^{182,183,184,185}

These prospects, the simplicity of the method and the promising results proves that our scoring function is a potentially powerful tool for structure-based drug design that could complement the array of computational methods, and provide an alternative way of looking at protein-ligand shape complementarity in virtual screening.

Chapter 6

Design and Evaluation of a Novel, Intuitive Human-Computer Interface Device

6.1 Introduction

Structural biology is a branch of biology interested in understanding the three dimensional structure of biological macromolecules. Characterizing macromolecular structures is one of the key approaches for obtaining a detailed understanding of their biological function, how they are regulated, how they interact with other molecular systems and how modifications of their structure affect their function. Since biological macromolecules, like proteins and nucleic acids, are detrimental to most processes in living cells, a detailed understanding of their structure and function leads to a broad range of applications in life-science research, ranging from functional characterization of novel proteins to applications in drug design and protein engineering.

Since structural biology provides vast amount of data, it is often challenging to access and interpret this data without being overwhelmed.¹⁸⁶ Most often, structural information is accessed through means of molecular visualization, which is used not only by structural experts but has become widely accepted in the biological research community.¹⁸⁷ In order to efficiently access structural information, sophisticated visualization techniques based on computer graphics and intuitive means to interact with the molecular representation are required. Although, visualization tools have significantly improved over the last decades, both in their rendering quality as well as their user-friendliness,¹⁸⁸ however, systems to interact with those tools in an intuitive and natural way are largely missing.

Picking up small objects, moving them around and reorienting them is clearly a straightforward task in the real world that can be done within seconds. However, performing similar movements in a virtual world is often much more challenging, since current human-computer interface devices do not reflect those natural movements.

Traditionally, interaction with a virtual three-dimensional environment is performed using standard computer peripherals like keyboard and mouse. Those input devices are inexpensive and commonly available, but have been designed to perform two-dimensional movements. The virtual scene, however, requires movements along six degrees of freedom. Therefore, using standard equipment to interact with a three-dimensional scene, is often unintuitive and imposes

a substantial hurdle to non-expert users. Over the years, numerous dedicated input devices, designed for movements in a three-dimensional environments, have been developed,^{189, 190, 191, 192} of which currently, to our knowledge, only the SpaceMouse™ is commercially available.¹⁹³

In a recent attempt to introduce fundamental principles of structural biology to a non-expert audience, we have experimented with both a user interface based on a SpaceMouse™ as well as a standard computer mouse/keyboard combination. Using a stereographic display, a simplified three-dimensional representation of a protein structure, i.e. its molecular surface, was shown to the user. In addition, a known small molecular inhibitor of that protein was displayed. The user could manipulate the position and orientation of the inhibitor by applying translation and rotation movements. A snapshot of the system is shown in Figure 6.1

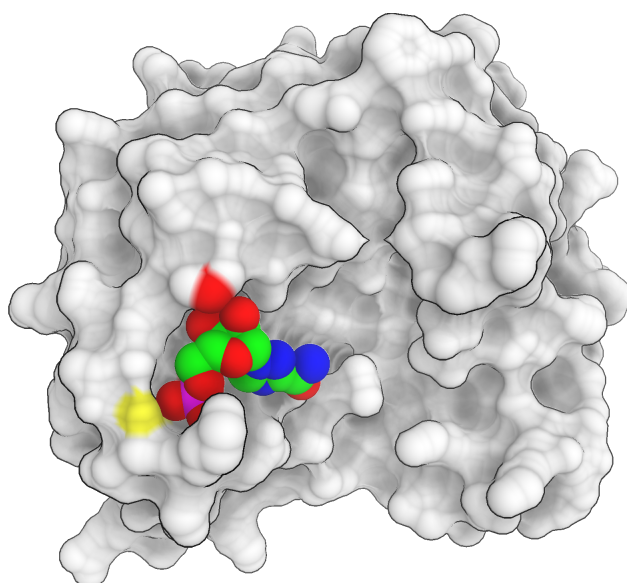


Figure 6.1: Visualization of the protein ligand system used during the demonstration of the basics of structure based drug discovery. The molecular surface of the protein is shown in gray, the small molecular inhibitor is shown in a sphere representation, where each sphere indicates an atom and the colors indicate different elements.

For most users, independent of their age and computer experience, a substantial hurdle was introduced by either of the afore mentioned user interface devices. In fact, this barrier was so high that most users could only focus on how to interact with the virtual environment while being completely distracted from the scientific content that was displayed.

This experiment clearly indicates that there is a need for a natural and intuitive user interface device designed specifically for tasks common to the field of structural biology and related areas. A transparent human-computer interface could substantially facilitate the interaction and improve the user experience. This would attract a broader community of life scientists and therefore could lead to a better use of the available data produced by structural biology studies.

Therefore, we have designed and evaluated a new user interface device which allows to use natural movements to interact intuitively with the afore mentioned virtual biological system.

6.2 Interface Device Design

The new user interface device uses a combination of an inertial measurement system (IMU) to follow the rotation of the user's hand and an optical motion sensing system, i.e. a Microsoft Kinect™ sensor to track the hand's position. The device represents an isotonic sensor, measuring its travel and applies direct position control with a linear relationship between the sensor and the virtual object movement. It is designed with large screen environments in mind and is based on inexpensive consumer grade equipment.

It has been shown that a number of design decisions are crucial for designing an intuitive user interface device:

Movement Separation Positioning of an object in a three-dimensional environment requires at least six degrees of freedom (DOF), three for object translation and three for rotation. A strict separation between the three translational and the three rotational DOFs was shown to be an integral part of a user-friendly device.¹⁹⁴ This point has only been poorly incorporated in the SpaceMouse™, where translation often triggers a slight rotation and vice versa.

Movements of real objects are clearly separated into a rotational and a translational component performed by individual body parts. Translations, are mainly performed by movements of the elbow and shoulder. Rotations however, are often executed by the finger tips and the wrist – for example when rolling an object between the fingers. The latter movements have been shown to yield precise rotational input.¹⁹⁵ Therefore, the new device was designed to mimic the same movements as used for real object manipulation.

Shape Mismatch Generally, the device held in the users hand is nothing like the object manipulated in the virtual environment. Although numerous attempts were taken to match the shape of the input device to the virtual object,^{196,197} we decided to use a spherical device for three reasons: First, a spherical input device behaves isotropically and thus can easily be rotated and does not interfere with the user. Second, an input device in the shape of the virtual object requires auto-fabrication of a new device for each represented system.^{198,199} Although, nowadays this is feasible, it significantly limits the interchangeability of the device. Third, such a device would need accurate alignment of the absolute orientation between the real and the virtual object in order to produce a realistic impression on the user.

Device Location It has been shown that the location of the input device can have an effect on its usability.^{200,201} Usually, the input device is held to the side of the computer and therefore movements do not correspond directly to the natural movements involved for manipulating real objects. However, having an input device where the movements correspond to natural movements requires lifting of the device off the ground. Although this might be fatiguing, in our experience, it significantly improves the design.

6.2.1 Hardware Architecture

The hardware used in our new input device is separated into two devices. Translation detection is based on an off-the-shelf Microsoft Kinect™ sensor. Rotation detection however, is based on a custom made hardware as described in more detail below.

Rotation Input Device

The core of the system is based on an 8-bit microcontroller (Atmel ATmega324-PA) which is interfaced to different hardware modules. The block diagram of the hardware design is shown in Figure 6.2 and the board layout in Figure 6.3. The direction of the communication process is indicated by arrows between the components labeled with the respective interface bus. The core microcontroller has three main tasks: communication with the inertial measurement unit (IMU), communication with the host computer and processing/filtering of the IMU data stream.

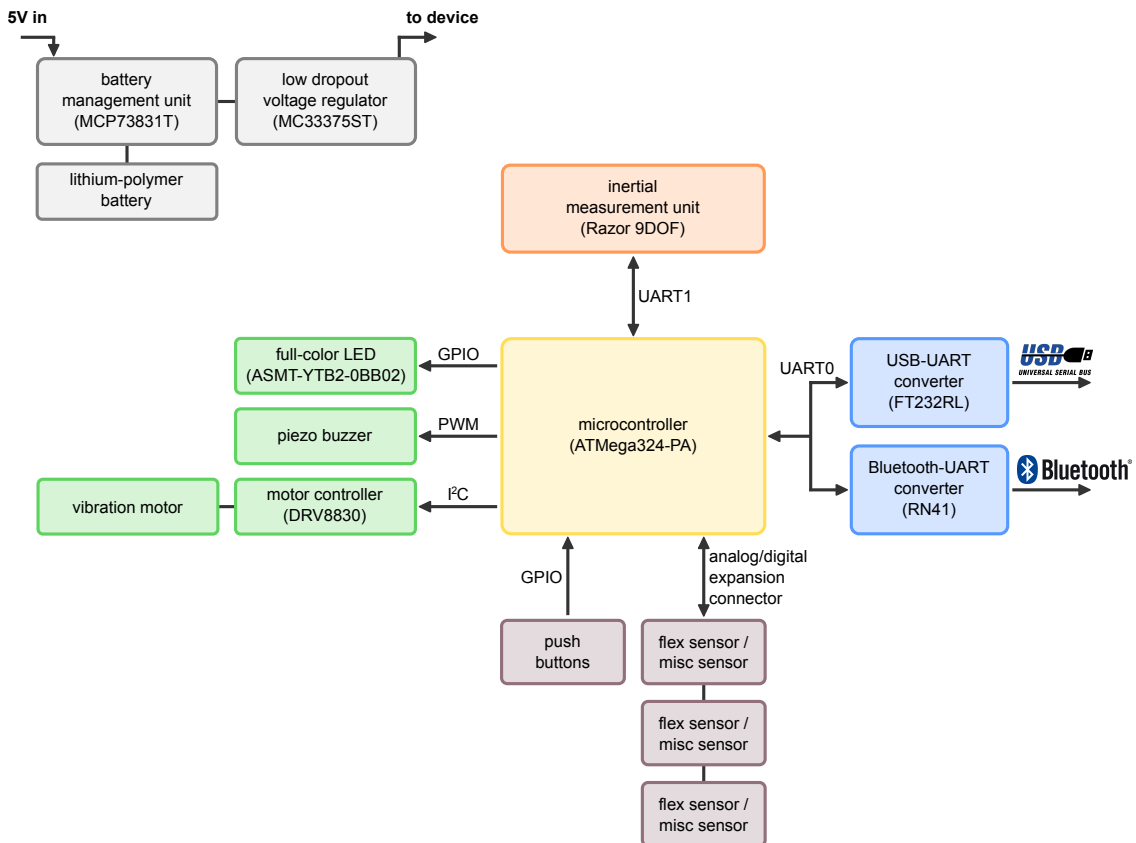


Figure 6.2: Schematic representation of the hardware architecture of the rotation input device. The core of the system is based on a microcontroller which interfaces different hardware modules. The direction of the communication process is indicated by arrows between the components labeled with the respective interface bus.

The inertial measurement unit (Sparkfun Razor 9DOF SEN-10736) senses both linear and angular acceleration as well as the absolute magnetic field in three axis each. It uses low-cost consumer-grade MEMS based sensor components, i.e. an Analog Devices ADXL345 triple-axis accelerometer, an InvenSense ITG-3200 triple-axis gyroscope and a Honeywell HMC5883L triple-

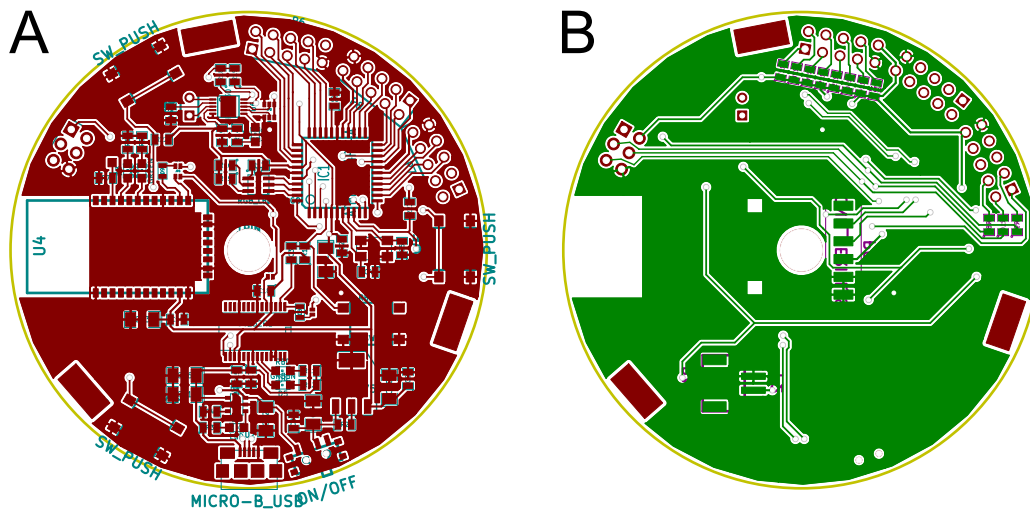


Figure 6.3: Front (A) and back (B) copper layer and silkscreen showing the layout of the custom made printed circuit board of the rotation input device.

axis magnetometer. The sensor values are processed by an on-board ATmega328 microcontroller which uses the firmware by Bartz et al.²⁰² (version 1.4.0) to compute heading, attitude and yaw information from the raw sensor values. The results are transmitted to the core microcontroller's UART1 port using the RS-232 protocol.

The core microcontroller communicates with the host computer through two distinct interfaces. First, a cable-based USB interface and second, a wireless class 1 Bluetooth interface with a range of up to 100m. Both interfaces are bidirectional. The incoming signals from the PC are translated to the RS-232 protocol using either an FT232RL USB-UART converter chip or an RN41 Bluetooth-UART converter module, for USB and Bluetooth communication protocols, respectively. Since the ATmega324-PA has two separate UART ports, but one is occupied by the IMU, the communication with both converter chips share the UART0 port. Since RS-232 is a point-to-point interface, sharing of a port between multiple devices is not natively supported and would even lead to the destruction of the interface chips if both devices communicate at the same time. Thus, we implemented a simple solution, using two Schottky diodes and a $100k\Omega$ pull-up resistor to emulate open collector output ports on the two interface chips. Schematics of the electronic circuit are shown in Appendix A.4. This solution allows the core microcontroller to send data through both interfaces in parallel and to receive data from both interfaces in a sequential manner.

The microcontroller interfaces with an array of push buttons to sense a compression of the input device casing. This allows for an easy, omnidirectional input signal which can be operated independent of the orientation of the device. The output signal caused by a compression event is arbitrarily configurable and can be used for example to activate/deactivate the input device. In addition, the system has an expansion connector which allows to connect up to eight analog or digital sensors, like flex or force sensors or additional push buttons. Additionally, both the I²C as well as the SPI bus are connected to an output header in order to add any input or output modules based on those protocols.

The device is able to produce optic, acoustic and haptic feedback. To accomplish this, three output modules are interfaced to the microcontroller. First, a full-color light emitting diode (ASMT-YTB2-0BB02) connected through a parallel interface. This allows to produce any color combination to communicate status information and for general visual feedback. Second, a piezoelectric buzzer, interfaced through a general purpose input-output (GPIO) pin where the frequency is controlled by a hardware pulse-width modulated (PWM) signal to emit a variable pitch acoustic signal. Third, a vibration motor connected to an appropriate motor controller (DRV8830) which is interfaced to the microcontroller through the I²C bus. This allows to produce haptic feedback through vibration of the input device. All output modules can be steered by signals from the input modules, events on the IMU data stream or commands from the host computer which are all interpreted by the core microcontroller.

The system also incorporates a 1000 mAh lithium polymer battery and the corresponding battery management unit (MCP73831T). Thus, charging of the battery is simply performed by connecting the device through the USB interface. The battery voltage of 3.7 V is then regulated to the appropriate 3.3 V used by all components by a low dropout linear voltage regulator (MC33375ST). Thus, the device incorporates all modules for power management and can be charged without removing the battery. The lifetime of the battery is in the range of days.

6.2.2 Software Architecture

The software used in our new device interfaces with the two hardware devices. A schematic overview of the software architecture is given in Figure 6.4.

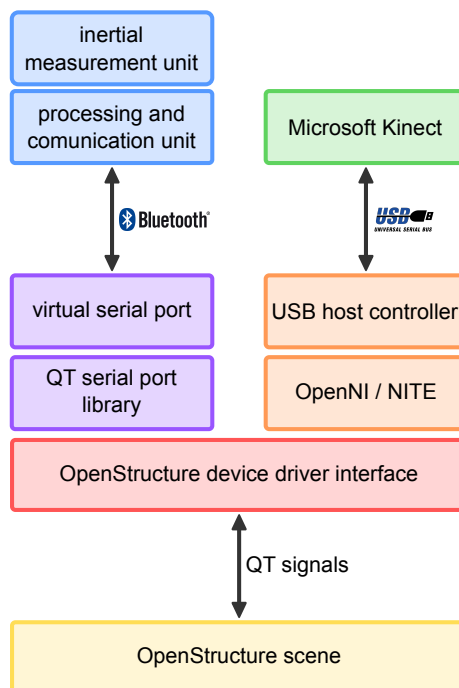


Figure 6.4: Schematic representation of the software architecture.

As described above, the inertial measurement unit, used for rotation input, is connected to a microcontroller platform responsible for signal processing and communication with the host computer through a wireless Bluetooth communication link. On the host computer, the Bluetooth kernel module emulates a serial port. Based on the QT serial port library `QExtSerialPort`²⁰³ a general device driver interface was implemented in `OpenStructure`,⁸³ allowing to connect to any device using the RS-232 serial port communication protocol.

The Microsoft Kinect™ sensor, used for translation input, is connected to the host computer through the USB interface. Based on the `OpenNI` framework²⁰⁴ and the `NITE` middleware²⁰⁵ skeleton tracking is performed. In `OpenStructure` a driver was implemented that interfaces with the `NITE` skeleton tracking and extracts the three-dimensional coordinates of the user's joints.

The driver interface now allows to attach either any graphical object or the camera itself to any hardware controller. This in turn allows the object's position and orientation to be controlled by the input devices. The flexibility of this interface has the advantage, that multiple input devices can be combined or different graphical objects, e.g. protein and ligand, can be controlled by different input devices.

6.3 User Experience

In order to illustrate computational approaches in structural biology and their application in modern life science to the general public, we have developed "Drug The Bug" – a 3D projection display installation with intuitive, interactive structure manipulation capabilities based on the newly developed user interface. This system demonstrates the basic ideas of structure based drug design approaches.

The software is written in Python and is based on `OpenStructure`,⁸³ an open-source, modular, flexible, molecular modeling and visualization environment. Within the application, different molecular systems can be chosen, each consisting of a known protein-ligand complex structure obtained from the PDB.¹¹⁸ The user can interactively manipulate the position and orientation of the ligand with the aim of finding the orientation which fits best into the binding pocket of the corresponding target protein. For simplicity, the position of the protein atoms as well as the internal degrees of freedom of the ligand are fixed. This is thus, a simplified version of what molecular docking programs like `Autodock`,¹⁷⁴ `Glide`⁶⁷ or `Dock`^{170,171} do, which are commonly used in structure based drug discovery.

The biological system is visualized as two individual objects. First, the protein structure is represented by its molecular surface computed by `MSMS`.¹²⁷ This is a reduced representation of the actual structure which allows for easier identification of surface exposed binding pockets. Second, the ligand is represented in sphere mode where each atom is drawn as a sphere with its radius set to the standard van der Waals radius of the corresponding element.

Molecular recognition between a protein and its ligand is manifold. To estimate binding affinities computationally, current methods employ numerous approximations.²⁹ Binding affinities are often estimated based on a combination of scores accounting for geometric fit,

electrostatic complementarity or van der Waals interaction energy. To illustrate these concepts, “Drug The Bug” displays different scores to the user. First, steric clashes between the ligand and the protein are displayed by coloring the protein surface. A color gradient is used to indicate the severeness, where white corresponds to no clash and red to a severe clash. Second, geometric complementarity and electrostatic interactions between the ligand and the protein are computed using BEScore (see Section 5) and the ligand atoms are colored accordingly using a gradient from blue (no interaction) to green (good interaction). All scores are continuously updated to give an immediate feedback to the user. To allow this rapid score update, BEScore is precomputed on all points of a three dimensional grid encompassing the whole protein. The score for each atom of a given ligand orientation can then be computed extremely fast through trilinear interpolation of the values at the eight closest grid points.

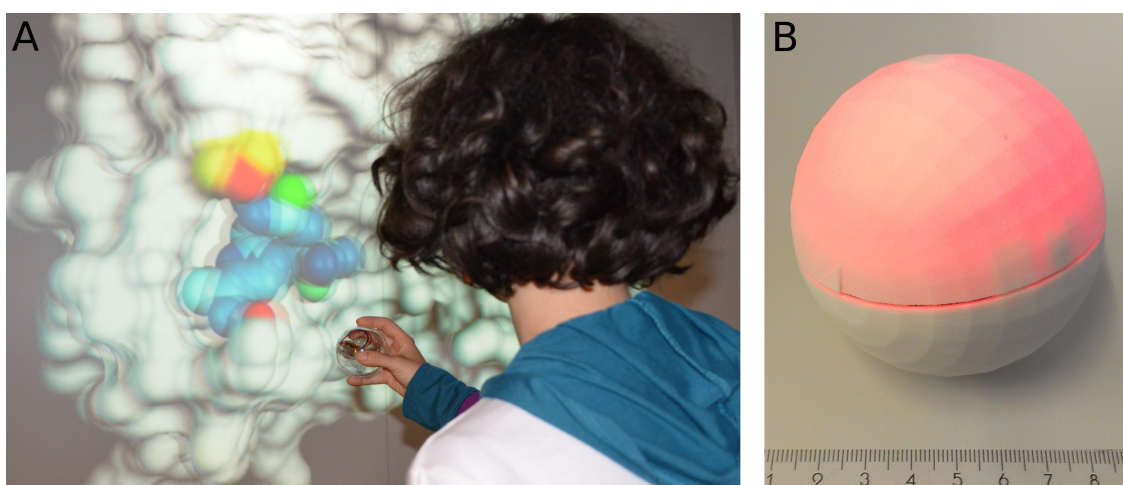


Figure 6.5: (A) A participant of the recent demonstration holding a prototype of the new input device in her right hand. (B) Photo of the current prototype.

The installation has been displayed on several occasions to the general public and was used for teaching of undergraduate students. A snapshot of a participant using a prototype of the input device is shown in Figure 6.5. The project was very well received by a highly diverse range of people with different age, computer experience and scientific or technical education. The barrier that was observed in previous experiments, where commercially available input devices like a standard computer mouse or the dedicated SpaceMouse™ was used, were significantly reduced. This was clearly indicated by the fact that most users did neither need a thorough explanation of how the interface worked nor did they need any training. It was commonly observed that the user forgot about the device and was fully immersed into the virtual world.

6.4 Conclusion

We have described here a new user interface device for three-dimensional virtual environments which fully relies on natural movements that the user knows from everyday motions. The new user interface device uses a combination of an inertial measurement system and an optical

motion sensing system. The system is an isotonic sensor which directly translates real space movements into the same movements in the virtual world. This renders it highly intuitive and is a step forward to a fully transparent user interface which does not distract the attention of the user from what is on the screen but allows him to fully dive into the virtual world.

The device has been intensively used in our newly developed interactive 3D projection display installation – “Drug The Bug” – which illustrates computational approaches in structural biology and their applications in modern life science to the general public. Due to the use of this device, the barrier between the user and the virtual environment was substantially reduced compared to earlier events where other input methods were used. This allowed the users to focus entirely on the scientific content of the demonstration. “Drug The Bug” is an excellent basis for further developments of interactive structure visualization and manipulation techniques including approaches like continuous energy minimization²⁰⁶ or interactive molecular dynamics simulations.²⁰⁷

The device opens up a number of new options for the intuitive interaction with three-dimensional biological structures. It is able to produce optic, acoustic and haptic feedback through the incorporation of a full-color light-emitting diode, a piezoelectric speaker and a vibration motor. Thus, a bidirectional communication between the user and the software application is possible where different feedback channels can be activated depending on the event. For example, when the device is used for docking of a ligand, unfavorable steric clashes between the ligand and the protein could be reported to the user by vibrating the input device, with the strength of the vibration depending on the magnitude of the clash. Future developments could incorporate additional communication channels, like olfactory or temperature output, to exploit the full spectrum of human sensory channels.

Chapter 7

Summary and Outlook

In this study we have successfully employed a combination of structure-based virtual screening methods and enzymatic inhibition assays in order to discover inhibitors of the dengue virus methyltransferase. Ten hit compounds were initially identified and the inhibitory activity of the two most active compounds was confirmed in additional inhibition assays. Due to solubility issues only a subset of three compounds was assayed in a subsequent ITC experiment, however, their binding to the dengue MTase could not be confirmed. One underlying reason for this might be a pre-occupied SAM binding pocket which renders ITC based measurements unfeasible for compounds binding to the SAM pocket. This effect should to be further quantified, for example by determination of the co-purified SAH content and by quantification of the SAH exchange rate.

The investigation of the catalytic mechanism of the dengue MTase addressed a number of open questions concerning the mechanism of the methyltransfer reactions.

By characterizing the underlying chemical reactions based on ab initio electronic structure calculations applied to model systems approximating the biological reactions, it was found that the 2'O and the N7 reactions are energetically favored processes, where the N7 reaction produces a significantly more stable product and shows a lower activation barrier than the 2'O reaction. Comparison between the catalyzed and uncatalyzed 2'O reaction, revealed the importance of a lysine residue which acts as a proton acceptor and significantly stabilizes the product state and reduces the activation barrier.

Furthermore, an in-silico approach was developed to identify the effects of single point mutations on different aspects of the catalyzed 2'O reaction. In a computational alanine scanning procedure protein residue patches were identified which modulate either the geometric arrangement between methyl donor and acceptor, the methyl donor binding affinity or the reaction energy barrier. In addition, previously uncharacterized hot-spot residues were identified and analyzed further using computational and experimental mutagenesis in order to gain a better understanding of their role in the enzyme's function. The analysis indicate that those mutations either modulate SAM binding or increase the reaction energy barrier. With the knowledge obtained in this study, we hope to facilitate the rational development of inhibitors against dengue fever and related diseases caused by flavivirus.

The evaluation of methods for predicting ligand binding sites for proteins with unknown structure during the CASP9 experiment highlighted the state of the art of current prediction methods. The results demonstrate that all top performing methods are based on homology transfer from known structures. However, such methods are limited to cases where a closely related protein structure with bound ligand is available which is not commonly the case. Thus, there is a clear need for the development of new methods allowing de-novo predictions of ligand binding sites. The setup of the ligand binding site prediction category in CASP has shown some major limitations primarily caused by a very low number of challenging target structures with relevant ligands and a restricted prediction format which treats all ligands uniformly, independent of their chemical type. To overcome those limitations, an extended assessment of ligand binding site predictions was implemented into the CAMEO framework introducing a format more suited for high accuracy predictions. This allows now to continuously evaluate the accuracy and reliability of ligand binding site prediction services in a blind and fully automated manner to assess the current state of the art of prediction methods, identify possible bottlenecks, and further stimulate the development of new methods.

We have successfully developed a rapid scoring function to identify the best ligand poses out of an ensemble of pre-docked poses. By quantifying the degree of burial and the electrostatic interactions of the ligand in a binding site promisingly high retrieval rates were achieved for selecting the best poses from a pool of decoy poses. Inspection of the scoring functions results indicates some limitations of the current method and suggests possible improvements to be addressed by further development. The main issue being an appropriate inclusion of water mediated interactions to further improve the scoring function's performance on protein-ligand complexes where molecular recognition is governed by water mediated interactions.

Acknowledgments

I would like to thank my two supervisors, Prof. Torsten Schwede and Prof. Markus Meuwly for giving me the opportunity to perform my PhD studies in their groups and for their continuous support and fruitful discussions.

I would also like to say thanks to all my colleagues from both groups for the pleasant and fruitful working environment. In particular, I thank Dr. Juergen Haas for numerous stimulating discussions, guidance and encouragement during all phases of this work and for critically reading my thesis and providing constructive feedback. Special thanks goes to Mohamed-Ali Mahi for his competent work in the wet-lab, conducting all experimental measurements. Furthermore, I would like to thank Marco Biasini for many (scientific) discussions and for his continuous efforts in developing OpenStructure, without this much of the work would have been impossible.

Additionally, I am grateful to all collaboration partners from Schrodinger, the NITD and the group of Prof. Bruno Canard.

Last but foremost, I would like to express my thanks to my wife Nicole, my daughter Anina and my family for their endless support and love without whom this work would not have been possible.

Bibliography

- [1] Jones, S. and Thornton, J. M. Searching for functional sites in protein structures. *Current Opinion in Chemical Biology* **8**(1), 3–7 (2004).
- [2] Kalyaanamoorthy, S. and Chen, Y. P. P. Structure-based drug design to augment hit discovery. *Drug Discovery Today* **16**(17-18), 831–839 (2011).
- [3] Mulholland, A. J. Modelling enzyme reaction mechanisms, specificity and catalysis. *Drug Discovery Today* **10**(20), 1393–1402 (2005).
- [4] Chothia, C. and Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *Embo Journal* **5**(4), 823–826 (1986).
- [5] Gabaldon, T. and Huynen, M. A. Prediction of protein function and pathways in the genome era. *Cellular and Molecular Life Sciences* **61**(7-8), 930–944 (2004).
- [6] Wolfson, H. J., Shatsky, M., Schneidman-Duhovny, D., et al. From structure to function: Methods and applications. *Current Protein and Peptide Science* **6**(2), 171–183 (2005).
- [7] Gherardini, P. F. and Helmer-Citterich, M. Structure-based function prediction: approaches and applications. *Brief Funct Genomic Proteomic* **7**(4), 291–302 (2008).
- [8] Berezin, C., Glaser, F., Rosenberg, J., et al. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* **20**(8), 1322–4 (2004).
- [9] Casari, G., Sander, C., and Valencia, A. A method to predict functional residues in proteins. *Nat Struct Biol* **2**(2), 171–8 (1995).
- [10] del Sol, A., Pazos, F., and Valencia, A. Automatic methods for predicting functionally important residues. *J Mol Biol* **326**(4), 1289–302 (2003).
- [11] Fischer, J. D., Mayer, C. E., and Soding, J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* **24**(5), 613–20 (2008).
- [12] Innis, C. A. siteFiNDER—3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res* **35**(Web Server issue), W489–94 (2007).
- [13] Pazos, F., Rausell, A., and Valencia, A. Phylogeny-independent detection of functional residues. *Bioinformatics* **22**(12), 1440–8 (2006).

- [14] Glaser, F., Morris, R. J., Najmanovich, R. J., et al. A method for localizing ligand binding pockets in protein structures. *Proteins* **62**(2), 479–88 (2006).
- [15] Hendlich, M., Rippmann, F., and Barnickel, G. Ligsite: Automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling* **15**(6) (1997).
- [16] Hernandez, M., Ghersi, D., and Sanchez, R. SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res* **37**(Web Server issue), W413–6 (2009).
- [17] Huang, B. and Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* **6**, 19 (2006).
- [18] Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **13**(5), 323–30, 307–8 (1995).
- [19] Brylinski, M. and Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* **105**(1), 129–34 (2008).
- [20] Oh, M., Joo, K., and Lee, J. Protein-binding site prediction based on three-dimensional protein modeling. *Proteins-Structure Function and Bioinformatics* **77**, 152–156 (2009).
- [21] Pandit, S. B., Brylinski, M., Zhou, H., et al. PSiFR: an integrated resource for prediction of protein structure and function. *Bioinformatics* **26**(5), 687–8 (2010).
- [22] Wass, M. N., Kelley, L. A., and Sternberg, M. J. E. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Research* **38**, W469–W473 (2010).
- [23] Lopez, G., Rojas, A., Tress, M., et al. Assessment of predictions submitted for the CASP7 function prediction category. *Proteins-Structure Function and Bioinformatics* **69**, 165–174 (2007).
- [24] Lopez, G., Ezkurdia, I., and Tress, M. L. Assessment of ligand binding residue predictions in CASP8. *Proteins-Structure Function and Bioinformatics* **77**, 138–146 (2009).
- [25] Schmidt, T., Haas, J., Cassarino, T. G., et al. Assessment of ligand-binding residue predictions in CASP9. *Proteins-Structure Function and Bioinformatics* **79** (2011).
- [26] Ko, J., Park, H., Heo, L., et al. Galaxyweb server for protein structure prediction and refinement. *Nucleic Acids Research* **40**(W1), W294–W297 (2012).
- [27] Roy, A., Yang, J. Y., and Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research* **40**(W1), W471–W477 (2012).
- [28] Kitchen, D. B., Decornez, H., Furr, J. R., et al. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery* **3**(11) (2004).

- [29] Gilson, M. K. and Zhou, H. X. Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure* **36**, 21–42 (2007).
- [30] Leach, A. R., Shoichet, B. K., and Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J Med Chem* **49**(20), 5851–5 (2006).
- [31] Brooijmans, N. and Kuntz, I. D. Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure* **32**, 335–373 (2003).
- [32] Kollman, P. A., Massova, I., Reyes, C., et al. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research* **33**(12), 889–897 (2000).
- [33] Aqvist, J., Medina, C., and Samuelsson, J. E. New method for predicting binding-affinity in computer-aided drug design. *Protein Engineering* **7**(3), 385–391 (1994).
- [34] Senn, H. M. and Thiel, W. Qm/mm methods for biomolecular systems. *Angewandte Chemie-International Edition* **48**(7), 1198–1229 (2009).
- [35] Kastner, J. Umbrella sampling. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **1**(6), 932–942 (2011).
- [36] *The Togaviruses: Biology, Structure, Replication*, chapter Togavirus morphology and morphogenesis, 241–316. Schlesinger R.W. (1980).
- [37] Grubler, D. Dengue/dengue haemorrhagic fever: history and current status. In *New Treatment Strategies For Dengue And Other Flaviviral Diseases*, volume 277 of *Novartis Foundation Symposium*, 3–16. John Wiley and Sons, Ltd, (2006).
- [38] Mackenzie, J. S., Gubler, D. J., and Petersen, L. R. Emerging flaviviruses: the spread and resurgence of japanese encephalitis, west nile and dengue viruses. *Nature Medicine* **10**(12), S98–S109 (2004).
- [39] Halstead, S. B. Dengue. *Lancet* **370**(9599), 1644–1652 (2007).
- [40] Whitehead, S. S., Blaney, J. E., Durbin, A. P., et al. Prospects for a dengue virus vaccine. *Nature Reviews Microbiology* **5**(7), 518–528 (2007).
- [41] Mukhopadhyay, S., Kuhn, R. J., and Rossmann, M. G. A structural perspective of the flavivirus life cycle. *Nature Reviews Microbiology* **3**(1), 13–22 (2005).
- [42] Harris, E., Holden, K., Edgil, D., et al. Molecular biology of flaviviruses. In *New Treatment Strategies For Dengue And Other Flaviviral Diseases*, volume 277 of *Novartis Foundation Symposium*, 3–16. John Wiley and Sons, Ltd, (2006).
- [43] Perera, R. and Kuhn, R. J. Structural proteomics of dengue virus. *Current Opinion in Microbiology* **11**(4), 369–377 (2008).

- [44] Egloff, M. P., Benarroch, D., Selisko, B., et al. An rna cap (nucleoside-2'-o)-methyltransferase in the flavivirus rna polymerase ns5: crystal structure and functional characterization. *Embo Journal* **21**(11), 2757–2768 (2002).
- [45] Dong, H. P., Ren, S. P., Zhang, B., et al. West nile virus methyltransferase catalyzes two methylations of the viral rna cap through a substrate-repositioning mechanism. *Journal of Virology* **82**(9), 4295–4307 (2008). Dong, Hongping Ren, Suping Zhang, Bo Zhou, Yangsheng Puig-Basagoiti, Francesc Li, Hongmin Shi, Pei-Yong.
- [46] Issur, M., Geiss, B. J., Bougie, I., et al. The flavivirus ns5 protein is a true rna guanylyltransferase that catalyzes a two-step reaction to form the rna cap structure. *Rna-a Publication of the Rna Society* **15**(12), 2340–2350 (2009).
- [47] Sampath, A. and Padmanabhan, R. Molecular targets for flavivirus drug discovery. *Antiviral Research* **81**(1), 6–15 (2009).
- [48] Cleaves, G. and Dubin, D. Methylation stratus of intracellular dengue type 2 40 S RNA. *Virology* **96**, 159–165 (1979).
- [49] Furuichi, Y. and Shatkin, A. J. Viral and cellular mrna capping: Past and prospects. *Advances in Virus Research, Vol 55* **55**, 135–184 (2000).
- [50] Ray, D., Shah, A., Tilgner, M., et al. West nile virus 5'-cap structure is formed by sequential guanine n-7 and ribose 2'-o methylations by nonstructural protein 5. *Journal of Virology* **80**(17), 8362–8370 (2006).
- [51] Martin, J. L. and McMillan, F. M. Sam (dependent) i am: the s-adenosylmethionine-dependent methyltransferase fold. *Current Opinion in Structural Biology* **12**(6), 783–793 (2002).
- [52] Benarroch, D., Egloff, M. P., Mulard, L., et al. A structural basis for the inhibition of the ns5 dengue virus mrna 2'-o-methyltransferase domain by ribavirin 5'-triphosphate. *Journal of Biological Chemistry* **279**(34), 35638–35643 (2004).
- [53] Egloff, M. P., Decroly, E., Malet, H., et al. Structural and functional analysis of methylation and 5'-RNA sequence requirements of short capped RNAs by the methyltransferase domain of dengue virus NS5. *Journal of Molecular Biology* **372**(3), 723–736 (2007).
- [54] Stevens, A. J., Gahan, M. E., Mahalingam, S., et al. The medicinal chemistry of dengue fever. *Journal of Medicinal Chemistry* **52**(24), 7911–7926 (2009).
- [55] Lim, S. P., Wen, D. Y., Yap, T. L., et al. A scintillation proximity assay for dengue virus ns5 2'-o-methyltransferase-kinetic and inhibition analyses. *Antiviral Research* **80**(3), 360–369 (2008).

- [56] Luzhkov, V. B., Selisko, B., Nordqvist, A., et al. Virtual screening and bioassay study of novel inhibitors for dengue virus mrna cap (nucleoside-2'o)-methyltransferase. *Bioorganic and Medicinal Chemistry* **15**(24), 7795–7802 (2007).
- [57] Milani, M., Mastrangelo, E., Bollati, M., et al. Flaviviral methyltransferase/rna interaction: Structural basis for enzyme inhibition. *Antiviral Research* **83**(1), 28–34 (2009). Milani, Mario Mastrangelo, Eloise Bollati, Michela Selisko, Barbara Decroly, Etienne Bouvet, Mickael Canard, Bruno Bolognesi, Martino EU IP Project Vizier [Cr 2004-511960]; Italian Ministry for University and Scientific Research FIRB Project 30 ELSEVIER SCIENCE BV 464BP.
- [58] Lim, S. P., Sonntag, L. S., Noble, C., et al. Small molecule inhibitors that selectively block dengue virus methyltransferase. *Journal of Biological Chemistry* **286**(8), 6233–6240 (2011).
- [59] Stahla-Beek, H. J., April, D. G., Saeedi, B. J., et al. Identification of a novel antiviral inhibitor of the flavivirus guanylyltransferase enzyme. *Journal of Virology* **86**(16), 8730–8739 (2012).
- [60] Zhou, Y. S., Ray, D., Zhao, Y. W., et al. Structure and function of flavivirus ns5 methyltransferase. *Journal of Virology* **81**(8), 3891–3903 (2007).
- [61] Dong, H. P., Zhang, B., and Shi, P. Y. Flavivirus methyltransferase: A novel antiviral target. *Antiviral Research* **80**(1), 1–10 (2008). Times Cited: 11.
- [62] Hodel, A. E., Gershon, P. D., and Quioco, F. A. Structural basis for sequence-nonspecific recognition of 5'-capped mrna by a cap-modifying enzyme. *Molecular Cell* **1**(3), 443–447 (1998).
- [63] Hager, J., Staker, B. L., Bugl, H., et al. Active site in rrmj, a heat shock-induced methyltransferase. *Journal of Biological Chemistry* **277**(44), 41978–41986 (2002).
- [64] Li, C., Xia, Y., Gao, X., et al. Mechanism of rna 2'-o-methylation: Evidence that the catalytic lysine acts to steer rather than deprotonate the target nucleophile. *Biochemistry* **43**(19), 5680–5687 (2004).
- [65] Li, C. and Gershon, P. D. pk(a) of the mrna cap-specific 2'-o-methyltransferase catalytic lysine by hsqc nmr detection of a two-carbon probe. *Biochemistry* **45**(3), 907–917 (2006).
- [66] Fabrega, C., Hausmann, S., Shen, V., et al. Structure and mechanism of mrna cap (guanine-n7) methyltransferase. *Molecular Cell* **13**(1), 77–89 (2004).
- [67] Friesner, R. A., Banks, J. L., Murphy, R. B., et al. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry* **47**(7) (2004).

- [68] McGovern, S. L., Caselli, E., Grigorieff, N., et al. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *Journal of Medicinal Chemistry* **45**(8), 1712–1722 (2002).
- [69] Seidler, J., McGovern, S. L., Doman, T. N., et al. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *Journal of Medicinal Chemistry* **46**(21), 4477–4486 (2003).
- [70] Shoichet, B. K. Screening in a spirit haunted world. *Drug Discovery Today* **11**(13-14), 607–615 (2006).
- [71] Feng, B. Y., Shelat, A., Doman, T. N., et al. High-throughput assays for promiscuous inhibitors. *Nature Chemical Biology* **1**(3), 146–148 (2005).
- [72] Feng, B. Y. and Shoichet, B. K. A detergent-based assay for the detection of promiscuous inhibitors. *Nature Protocols* **1**(2), 550–553 (2006).
- [73] Feng, B. Y., Simeonov, A., Jadhav, A., et al. A high-throughput screen for aggregation-based inhibition in a large compound library. *Journal of Medicinal Chemistry* **50**(10), 2385–2390 (2007).
- [74] Podvynec, M., Lim, S. P., Schmidt, T., et al. Novel inhibitors of dengue virus methyltransferase: Discovery by in vitro-driven virtual screening on a desktop computer grid. *Journal of Medicinal Chemistry* **53**(4) (2010).
- [75] Babaoglu, K., Simeonov, A., Lrwin, J. J., et al. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase. *Journal of Medicinal Chemistry* **51**(8), 2502–2511 (2008).
- [76] Breiman, L. Random forests. *Machine Learning* **45**(1), 5–32 (2001).
- [77] Ihaka, R. and Gentleman, R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**(3), 299–314 (1996).
- [78] Yap, L. J., Luo, D. H., Chung, K. Y., et al. Crystal structure of the dengue virus methyltransferase bound to a 5'-capped octameric rna. *Plos One* **5**(9), 9 (2010).
- [79] Bollati, M., Alvarez, K., Assenberg, R., et al. Structure and functionality in flavivirus ns-proteins: Perspectives for drug design. *Antiviral Research* **87**(2), 125–148 (2010).
- [80] Phillips, J., Braun, R., Wang, W., et al. Scalable molecular dynamics with namd. *J. Comput. Chem.* **26**, 1781–1802 (2005).
- [81] MacKerell, A. D., Bashford, D., Bellott, M., et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* **102**(18), 3586–3616 (1998).

- [82] MacKerell, A. D., Feig, M., and Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry* **25**(11), 1400–1415 (2004).
- [83] Biasini, M., Mariani, V., Haas, J., et al. Openstructure: a flexible software framework for computational structural biology. *Bioinformatics* **26**(20) (2010).
- [84] Frisch, M. J., Trucks, G. W., Schlegel, H. B., et al. Gaussian 03, revision d.02. *Gaussian, Inc., Wallingford CT* (2004).
- [85] Becke, A. Density-functional thermochemistry. III. the role of exact exchange. *J. Chem. Phys.* **7**, 5648–5652 (1993).
- [86] Ochterski, J., Petersson, G., and Montgomery, J. A complete basis set model chemistry. V. extensions to six or more heavy atoms. *J. Chem. Phys.* **104**, 2598–2619 (1996).
- [87] Cossi, M., Rega, N., Scalmani, G., et al. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem* **24**, 669–681 (2003).
- [88] Peng, C. Y. and Schlegel, H. B. Combining synchronous transit and quasi-newton methods to find transition-states. *Israel Journal of Chemistry* **33**(4), 449–454 (1993).
- [89] Reed, A. E., Weinstock, R. B., and Weinhold, F. Natural-population analysis. *Journal of Chemical Physics* **83**(2), 735–746 (1985).
- [90] Selisko, B., Peyrane, F. F., Canard, B., et al. Biochemical characterization of the (nucleoside-2' o)-methyltransferase activity of dengue virus protein ns5 using purified capped rna oligonucleotides (7me)gpppac(n) and gpppac(n). *Journal of General Virology* **91**, 112–121 (2010).
- [91] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., et al. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **79**(2), 926–935 (1983).
- [92] Foloppe, N. and MacKerell, A. D. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry* **21**(2), 86–104 (2000).
- [93] MacKerell, A. D. and Banavali, N. K. All-atom empirical force field for nucleic acids: II. application to molecular dynamics simulations of DNA and RNA in solution. *Journal of Computational Chemistry* **21**(2), 105–120 (2000).
- [94] Feng, M. H., Philippopoulos, M., MacKerell, A. D., et al. Structural characterization of the phosphotyrosine binding region of a high-affinity sh2 domain-phosphopeptide complex by molecular dynamics simulation and chemical shift calculations. *Journal of the American Chemical Society* **118**(45), 11265–11277 (1996).

- [95] Lee, M. S., Salsbury, F. R., and Brooks, C. L. Novel generalized born methods. *Journal of Chemical Physics* **116**(24), 10606–10614 (2002).
- [96] Lee, M. S., Feig, M., Salsbury, F. R., et al. New analytic approximation to the standard molecular volume definition and its application to generalized born calculations. *Journal of Computational Chemistry* **24**(11), 1348–1356 (2003).
- [97] Baker, N. A., Sept, D., Joseph, S., et al. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **98**(18) (2001).
- [98] Elstner, M., Porezag, D., Jungnickel, G., et al. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Physical Review B* **58**(11), 7260–7268 (1998).
- [99] Cui, Q., Elstner, M., Kaxiras, E., et al. A qm/mm implementation of the self-consistent charge density functional tight binding (scc-dftb) method. *Journal of Physical Chemistry B* **105**(2), 569–585 (2001).
- [100] Kumar, S., Bouzida, D., Swendsen, R. H., et al. The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method. *Journal of Computational Chemistry* **13**(8), 1011–1021 (1992).
- [101] Massova, I. and Kollman, P. A. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *Journal of the American Chemical Society* **121**(36), 8133–8143 (1999).
- [102] Cunningham, B. C. and Wells, J. A. High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science* **244**(4908), 1081–1085 (1989).
- [103] Geiss, B. J., Thompson, A. A., Andrews, A. J., et al. Analysis of flavivirus ns5 methyltransferase cap binding. *Journal of Molecular Biology* **385**(5), 1643–1654 (2009).
- [104] Dong, H., Chang, D. C., Xie, X., et al. Biochemical and genetic characterization of dengue virus methyltransferase. *Virology* **405**(2), 568–578 (2010).
- [105] Dong, H. P., Liu, L. H., Zou, G., et al. Structural and functional analyses of a conserved hydrophobic pocket of flavivirus methyltransferase. *Journal of Biological Chemistry* **285**(42), 32586–32595 (2010).
- [106] Kroschewski, H., Lim, S. P., Butcher, R. E., et al. Mutagenesis of the dengue virus type 2 ns5 methyltransferase domain. *Journal of Biological Chemistry* **283**(28) (2008).
Kroschewski, Helga Lim, Siew Pheng Butcher, Rebecca E. Yap, Thai Leong Lescar, Julien Wright, Peter J. Vasudevan, Subhash G. Davidson, Andrew D.

- [107] Bradshaw, R. T., Patel, B. H., Tate, E. W., et al. Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Protein Engineering Design and Selection* **24**(1-2), 197–207 (2011).
- [108] Hao, G. F., Yang, G. F., and Zhan, C. G. Computational mutation scanning and drug resistance mechanisms of hiv-1 protease inhibitors. *Journal of Physical Chemistry B* **114**(29), 9663–9676 (2010).
- [109] Huo, S., Massova, I., and Kollman, P. A. Computational alanine scanning of the 1 : 1 human growth hormone-receptor complex. *Journal of Computational Chemistry* **23**(1), 15–27 (2002).
- [110] Riccardi, D., Schaefer, P., Yang, Y., et al. Development of effective quantum mechanical/molecular mechanical (qm/mm) methods for complex biological processes. *Journal of Physical Chemistry B* **110**(13), 6458–6469 (2006).
- [111] Elstner, M., Frauenheim, T., and Suhai, S. An approximate dft method for qm/mm simulations of biological structures and processes. *Theochem* **632**, 29–4141 (2003).
- [112] Dong, H. P., Ray, D., Ren, S. P., et al. Distinct rna elements confer specificity to flavivirus rna cap methylation events. *Journal of Virology* **81**(9), 4412–4421 (2007).
- [113] Soro, S. and Tramontano, A. The prediction of protein function at CASP6. *Proteins-Structure Function and Bioinformatics* **61**, 201–213 (2005).
- [114] Pellegrini-Calace, M., Soro, S., and Tramontano, A. Revisiting the prediction of protein function at CASP6. *Febs Journal* **273**(13), 2977–2983 (2006).
- [115] Schwede, T., Sali, A., Honig, B., et al. Outcome of a workshop on applications of protein models in biomedical research. *Structure* **17**(2), 151–159 (2009).
- [116] Waterhouse, A., Schmidt, T., Biasini, M., et al. CAMEO Structure Annotation System. <http://www.cameo3d.org/annotation>, (2012).
- [117] wwPDB. Chemical Component Dictionary. <http://www.wwpdb.org/ccd.html>, (2012).
- [118] Berman, H., Henrick, K., and Nakamura, H. Announcing the worldwide protein data bank. *Nature Structural Biology* **10**(12) (2003).
- [119] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**(8), 861–874 (2006).
- [120] Spearman, C. The proof and measurement of association between two things. *American Journal of Psychology* **15**, 72–101 (1904).
- [121] Matthews, B. W. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Et Biophysica Acta* **405**(2), 442–451 (1975).

- [122] Altschul, S. F., Gish, W., Miller, W., et al. Basic local alignment search tool. *Journal of Molecular Biology* **215**(3), 403–410 (1990).
- [123] Larkin, M. A., Blackshields, G., Brown, N. P., et al. Clustal W and clustal X version 2.0. *Bioinformatics* **23**(21), 2947–2948 (2007).
- [124] Armon, A., Graur, D., and Ben-Tal, N. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology* **307**(1), 447–463 (2001).
- [125] Arnold, K., Bordoli, L., Kopp, J., et al. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**(2), 195–201 (2006).
- [126] Peitsch, M. C. Protein modeling by e-mail. *Bio-Technology* **13**(7), 658–660 (1995).
- [127] Sanner, M. F., Olson, A. J., and Spehner, J. C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **38**(3) (1996).
- [128] Prediction Center. Prediction Center. <http://www.predictioncenter.org>, (2012).
- [129] Haas, J. CAMEO. <http://www.cameo3d.org>, (2012).
- [130] Berman, H. M., Westbrook, J., Feng, Z., et al. The protein data bank. *Nucleic Acids Research* **28**(1) (2000).
- [131] Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **7**(20) (2002).
- [132] Deng, Z., Chuaqui, C., and Singh, J. Structural interaction fingerprint (sift): A novel method for analyzing three-dimensional protein-ligand binding interactions. *Journal of Medicinal Chemistry* **47**(2) (2004).
- [133] Graves, A. P., Shivakumar, D. M., Boyce, S. E., et al. Rescoring docking hit lists for model cavity sites: Predictions and experimental testing. *Journal of Molecular Biology* **377**(3), 914–934 (2008).
- [134] Thompson, D. C., Humblet, C., and Joseph-McCarthy, D. Investigation of mm-pbsa rescoring of docking poses. *Journal of Chemical Information and Modeling* **48**(5) (2008).
- [135] Kroemer, R. T. Structure-based drug design: Docking and scoring. *Current Protein and Peptide Science* **8**(4) (2007).
- [136] Ferrara, P., Gohlke, H., Price, D. J., et al. Assessing scoring functions for protein-ligand interactions. *Journal of Medicinal Chemistry* **47**(12) (2004).
- [137] Friedman, R. and Caflich, A. Discovery of plasmepsin inhibitors by fragment-based docking and consensus scoring. *Chemmedchem* **4**(8) (2009).

- [138] Doman, T. N., McGovern, S. L., Witherbee, B. J., et al. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1b. *Journal of Medicinal Chemistry* **45**(11) (2002).
- [139] Lyne, P. D., Kenny, P. W., Cosgrove, D. A., et al. Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. *Journal of Medicinal Chemistry* **47**(8) (2004).
- [140] Cross, J. B., Thompson, D. C., Rai, B. K., et al. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *Journal of Chemical Information and Modeling* **49**(6) (2009).
- [141] Gohlke, H. and Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie (International ed. in English)* **41**(15) (2002).
- [142] Krovat, E. M., Steindl, T., and Langer, T. Recent advances in docking and scoring. *Current Computer-Aided Drug Design* **1**(1) (2005).
- [143] Schulz-Gasch, T. and Stahl, M. Scoring functions for proteinligand interactions: a critical perspective. *Drug Discovery Today: Technologies* **1**(3), 231–239 (2004).
- [144] Warren, G. L., Andrews, C. W., Capelli, A.-M., et al. A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry* **49**(20) (2006).
- [145] Bissantz, C., Folkers, G., and Rognan, D. Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **43**(25) (2000).
- [146] Chen, H. M., Lyne, P. D., Giordanetto, F., et al. Evaluating molecular-docking methods for pose prediction and enrichment factors. *Journal of Chemical Information and Modeling* **46**(1) (2006). Times Cited: 109 4th Indo-US Workshop on Mathematical Chemistry JAN 08-12, 2005 Pune, INDIA.
- [147] Kellenberger, E., Rodrigo, J., Muller, P., et al. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins-Structure Function and Bioinformatics* **57**(2) (2004).
- [148] Kontoyianni, M., McClellan, L. M., and Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *Journal of Medicinal Chemistry* **47**(3) (2004).
- [149] Perola, E., Walters, W. P., and Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins-Structure Function and Bioinformatics* **56**(2) (2004).
- [150] Stahl, M. and Rarey, M. Detailed analysis of scoring functions for virtual screening. *Journal of Medicinal Chemistry* **44**(7) (2001).

- [151] Wang, R. X., Lu, Y. P., and Wang, S. M. Comparative evaluation of 11 scoring functions for molecular docking. *Journal of Medicinal Chemistry* **46**(12) (2003).
- [152] Jain, A. N. and Nicholls, A. Recommendations for evaluation of computational methods. *Journal of Computer-Aided Molecular Design* **22**(3-4) (2008).
- [153] Cole, J. C., Murray, C. W., Nissink, J. W. M., et al. Comparing protein-ligand docking programs is difficult. *Proteins-Structure Function and Bioinformatics* **60**(3) (2005).
- [154] Schmitt, S., Kuhn, D., and Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *Journal of Molecular Biology* **323**(2) (2002).
- [155] Halgren, T. A. Merck molecular force field .1. basis, form, scope, parameterization, and performance of mmff94. *Journal of Computational Chemistry* **17**(5-6) (1996).
- [156] Halgren, T. A. Merck molecular force field .2. mmff94 van der waals and electrostatic parameters for intermolecular interactions. *Journal of Computational Chemistry* **17**(5-6) (1996).
- [157] Halgren, T. A. Merck molecular force field .3. molecular geometries and vibrational frequencies for mmff94. *Journal of Computational Chemistry* **17**(5-6) (1996).
- [158] Halgren, T. A. and Nachbar, R. B. Merck molecular force field .4. conformational energies and geometries for mmff94. *Journal of Computational Chemistry* **17**(5-6) (1996).
- [159] Halgren, T. A. Merck molecular force field .5. extension of mmff94 using experimental data, additional computational data, and empirical rules. *Journal of Computational Chemistry* **17**(5-6) (1996).
- [160] O'Boyle, N. M., Banck, M., James, C. A., et al. Open babel: An open chemical toolbox. *Journal of Cheminformatics* **3** (2011).
- [161] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., et al. Protein data bank - computer-based archival file for macromolecular structures. *Journal of Molecular Biology* **112**(3) (1977).
- [162] Brandstetter, H., Turk, D., Hoeffken, H. W., et al. Refined 2.3-angstrom x-ray crystal-structure of bovine thrombin complexes formed with the benzamidine and arginine-based thrombin inhibitors napap, 4-tapap and mqpa - a starting point for improving antithrombotics. *Journal of Molecular Biology* **226**(4) (1992).
- [163] Shelley, J. C., Cholleti, A., Frye, L. L., et al. Epik: a software program for pk (a) prediction and protonation state generation for drug-like molecules. *Journal of Computer-Aided Molecular Design* **21**, 681–691 (2007).

- [164] Graves, A. P., Brenk, R., and Shoichet, B. K. Decoys for docking. *Journal of Medicinal Chemistry* **48**(11) (2005).
- [165] Hartshorn, M. J., Verdonk, M. L., Chessari, G., et al. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry* **50**(4) (2007).
- [166] Grosdidier, A., Zoete, V., and Michielin, O. Blind docking of 260 protein-ligand complexes with eadock 2.0. *Journal of Computational Chemistry* **30**(13) (2009).
- [167] Roche, O., Kiyama, R., and Brooks, C. L. Ligand-protein database: Linking protein-ligand complex structures to binding data. *Journal of Medicinal Chemistry* **44**(22), 3592–3598 (2001).
- [168] Wang, R. X., Lai, L. H., and Wang, S. M. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design* **16**(1) (2002).
- [169] Momany, F. A. and Rone, R. Validation of the general-purpose quanta(r)3.2/charmm(r) force-field. *Journal of Computational Chemistry* **13**(7) (1992).
- [170] Ewing, T. J. A., Makino, S., Skillman, A. G., et al. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design* **15**(5) (2001).
- [171] Brozell, S. R., Mukherjee, S., Balius, T. E., et al. Evaluation of dock 6 as a pose generation and database enrichment tool. *Journal of Computer-Aided Molecular Design* **26**(6), 749–773 (2012).
- [172] Eldridge, M. D., Murray, C. W., Auton, T. R., et al. Empirical scoring functions .1. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design* **11**(5) (1997).
- [173] Gohlke, H., Hendlich, M., and Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology* **295**(2) (2000).
- [174] Morris, G. M., Goodsell, D. S., Huey, R., et al. Distributed automated docking of flexible ligands to proteins: Parallel applications of autodock 2.4. *Journal of Computer-Aided Molecular Design* **10**(4) (1996).
- [175] Zoete, V., Grosdidier, A., Cuendet, M., et al. Use of the facts solvation model for protein-ligand docking calculations. application to eadock. *Journal of Molecular Recognition* **23**(5), 457–461 (2010).
- [176] Novikov, F. N., Stroylov, V. S., Zeifman, A. A., et al. Lead finder docking and virtual screening evaluation with astex and dud test sets. *Journal of Computer-Aided Molecular Design* **26**(6), 725–735 (2012).

- [177] Muegge, I. and Martin, Y. C. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *Journal of Medicinal Chemistry* **42**(5) (1999).
- [178] Velec, H. F. G., Gohlke, H., and Klebe, G. Drugscore(csd)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *Journal of Medicinal Chemistry* **48**(20) (2005).
- [179] Neves, M. A. C., Totrov, M., and Abagyan, R. Docking and scoring with icm: the benchmarking results and strategies for improvement. *Journal of Computer-Aided Molecular Design* **26**(6), 675–686 (2012).
- [180] Abad-Zapatero, C. and Metz, J. T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discovery Today* **10**(7) (2005).
- [181] Hopkins, A. L., Groom, C. R., and Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **9**(10) (2004).
- [182] Baber, J. C., William, A. S., Gao, Y. H., et al. The use of consensus scoring in ligand-based virtual screening. *Journal of Chemical Information and Modeling* **46**(1) (2006).
- [183] Charifson, P. S., Corkery, J. J., Murcko, M. A., et al. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry* **42**(25) (1999).
- [184] Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discovery Today* **11**(9-10) (2006).
- [185] Oda, A., Tsuchida, K., Takakura, T., et al. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *Journal of Chemical Information and Modeling* **46**(1) (2006).
- [186] O'Donoghue, S. I., Gavin, A. C., Gehlenborg, N., et al. Visualizing biological data-now and in the future. *Nature Methods* **7**(3), S2–S4 (2010).
- [187] O'Donoghue, S. I., Goodsell, D. S., Frangakis, A. S., et al. Visualization of macromolecular structures. *Nature Methods* **7**(3), S42–S55 (2010).
- [188] Goodsell, D. S. Visual methods from atoms to cells. *Structure* **13**(3), 347–354 (2005).
- [189] Ware, C. Using hand position for virtual object placement. *Visual Computer* **6**(5) (1990).
- [190] Venolia, D. Facile 3D direct manipulation, (1993).
- [191] Balakrishnan, R., Baudel, T., Kurtenbach, G., et al. The Rockin'Mouse: integral 3D manipulation on a plane, (1997).

- [192] Froehlich, B., Hochstrate, J., Skuk, V., et al. The GlobeFish and the GlobeMouse: two new six degree of freedom input devices for graphics applications, (2006).
- [193] 3dconnexion. 3dconnexion. <http://www.3dconnexion.com>.
- [194] Jacob, R. J. K., Sibert, L. E., McFarlane, D. C., et al. Integrality and separability of input devices. *ACM Trans. Comput.-Hum. Interact.* **1**(1), 3–26 (1994).
- [195] Zhai, S., Milgram, P., and Buxton, W. The influence of muscle groups on performance of multiple degree-of-freedom input, (1996).
- [196] Hinckley, K., Pausch, R., Goble, J. C., et al. Passive real-world interface props for neurosurgical visualization, (1994).
- [197] Gillet, A., Sanner, M., Stoffler, D., et al. Tangible interfaces for structural molecular biology. *Structure* **13**(3), 483–491 (2005).
- [198] Gillet, A., Sanner, M., Stoffler, D., et al. Tangible augmented structural molecular interfaces for biology. *Ieee Computer Graphics and Applications* **25**(2), 13–17 (2005).
- [199] Herman, T., Morris, J., Colton, S., et al. Tactile teaching - Exploring protein structure/function using physical models. *Biochemistry and Molecular Biology Education* **34**(4), 247–254 (2006).
- [200] Graham, E. D. and MacKenzie, C. L. *Physical versus virtual pointing*, 292–299. *Acm* (1996).
- [201] Ware, C. and Rose, J. Rotating virtual objects with real handles. *ACM Transactions on Computer-Human Interaction* **6**(2) (1999).
- [202] Bartz, P. and Spors, S. Razor AHRS firmware. <http://dev.qu.tu-berlin.de/projects/sf-razor-9dof-ahrs>, (2012).
- [203] QExtSerialPort. QExtSerialPort. <http://code.google.com/p/qextserialport>, (2011).
- [204] OpenNI. OpenNI Framework. <http://openni.org>, (2011).
- [205] PrimeSense. NITE middleware. <http://www.primesense.com/nite>, (2011).
- [206] Surles, M. C., Richardson, J. S., Richardson, D. C., et al. Sculpting proteins interactively - Continual energy minimization embedded in a graphical modeling system. *Protein Science* **3**(2), 198–210 (1994).
- [207] Delalande, O., Ferey, N., Grasseau, G., et al. Complex molecular assemblies at hand via interactive simulations. *Journal of Computational Chemistry* **30**(15), 2375–2387 (2009).

Appendix A

Appendix

A.1 Dengue

A.1.1 Identification and Validation of Novel Dengue Methyltransferase Inhibitors

Table A.1: Comparison of the decision-tree criteria used by the original work of Seidler et al and by our work.

Seidler et al. Criterion	Criterion, our work
Daylight clogp \leq 3.633	QikProp clogPow $<$ 3.1
Electrotopological S _{sssN} \leq 2.287	Max Epik pKa for tertiary N $<$ 7
Max_conj_path \leq 18.5	Largest contiguous set of sp ² atoms $<$ 19.5
Contains COOH	Contains COOH
Daylight clogP \geq 5.389	QikProp clogPow \geq 4.7

A.1.2 Computational Analysis of the Methyltransferase Reaction

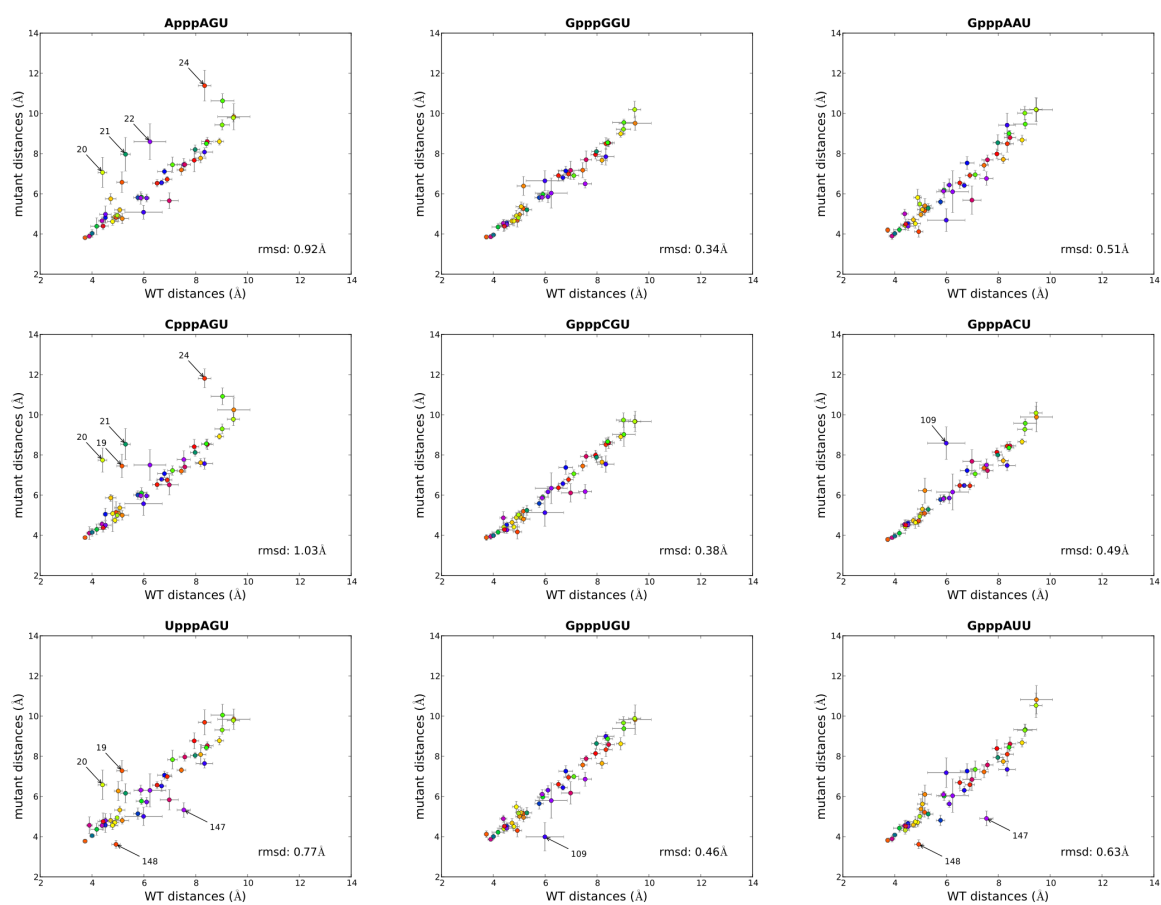


Figure A.1: Differences in RNA-MTase contacts between WT and mutated RNA. Contacts are measured by their distances observed in MT simulations of the MTase bound to SAM and WT or mutated RNA. All RNA mutants are shown: (left column) mutants of cap-nucleotide, (middle column) mutants of 1st nucleotide, (right column) mutants of 2nd nucleotide. Residues with significantly modified distances are labeled by their number.

A.2 CAMEO Ligand Binding

A.2.1 CAMEO Ligand Binding Format Examples

Minimum example of ion binding site

Listing A.1 shows an example of the new CAMEO ligand binding site prediction format using the bare minimum required for a valid prediction that format.

Listing A.1: Example format of a ion (zinc) binding site in structure 3ZTT.

```

1 r=SER; n=198; | I=0.000; O=0.000; N=0.000; P=0.000; |
2 r=GLU; n=199; | I=1.000; O=0.000; N=0.000; P=0.000; |
3 r=GLY; n=200; | I=0.000; O=0.000; N=0.000; P=0.000; |
4 r=ALA; n=201; | I=0.513; O=0.000; N=0.000; P=0.000; |

```

Extended example of ATP and Mn binding site

Listing A.2 shows an extended example of the new CAMEO ligand bindings site prediction format. In addition to the minimum fields required, predictions are given for each atom individually (key: a=) and predictions for specific compounds are given (keys: ANP= and MN=)

Listing A.2: Example format of an ion (manganese) and an organic ligand (ATP) binding site in structure 3QAM.

```

1 r=GLU; n=170; a=N; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
2 r=GLU; n=170; a=CA; | I=0.047; O=0.412; N=0.000; P=0.000; | ANP=0.412; MN=0.047;
3 r=GLU; n=170; a=C; | I=0.337; O=0.668; N=0.000; P=0.000; | ANP=0.668; MN=0.337;
4 r=GLU; n=170; a=O; | I=0.372; O=1.000; N=0.000; P=0.000; | ANP=1.000; MN=0.372;
5 r=GLU; n=170; a=CB; | I=0.249; O=0.424; N=0.000; P=0.000; | ANP=0.424; MN=0.249;
6 r=GLU; n=170; a=CG; | I=0.000; O=0.077; N=0.000; P=0.000; | ANP=0.077; MN=0.000;
7 r=GLU; n=170; a=CD; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
8 r=GLU; n=170; a=OE1; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
9 r=GLU; n=170; a=OE2; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
10 r=ASN; n=171; a=N; | I=0.331; O=0.353; N=0.000; P=0.000; | ANP=0.353; MN=0.331;
11 r=ASN; n=171; a=CA; | I=0.401; O=0.307; N=0.000; P=0.000; | ANP=0.307; MN=0.401;
12 r=ASN; n=171; a=C; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
13 r=ASN; n=171; a=O; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
14 r=ASN; n=171; a=CB; | I=0.528; O=0.251; N=0.000; P=0.000; | ANP=0.251; MN=0.528;
15 r=ASN; n=171; a=CG; | I=0.987; O=0.584; N=0.000; P=0.000; | ANP=0.584; MN=0.987;
16 r=ASN; n=171; a=OD1; | I=1.000; O=0.939; N=0.000; P=0.000; | ANP=0.939; MN=1.000;
17 r=ASN; n=171; a=ND2; | I=0.859; O=0.637; N=0.000; P=0.000; | ANP=0.637; MN=0.859;
18 r=LEU; n=173; a=N; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
19 r=LEU; n=173; a=CA; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
20 r=LEU; n=173; a=C; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
21 r=LEU; n=173; a=O; | I=0.000; O=0.000; N=0.000; P=0.000; | ANP=0.000; MN=0.000;
22 r=LEU; n=173; a=CB; | I=0.000; O=0.175; N=0.000; P=0.000; | ANP=0.175; MN=0.000;
23 r=LEU; n=173; a=CG; | I=0.000; O=0.509; N=0.000; P=0.000; | ANP=0.509; MN=0.000;
24 r=LEU; n=173; a=CD1; | I=0.000; O=0.898; N=0.000; P=0.000; | ANP=0.898; MN=0.000;
25 r=LEU; n=173; a=CD2; | I=0.000; O=0.696; N=0.000; P=0.000; | ANP=0.696; MN=0.000;

```

A.3 BEscore

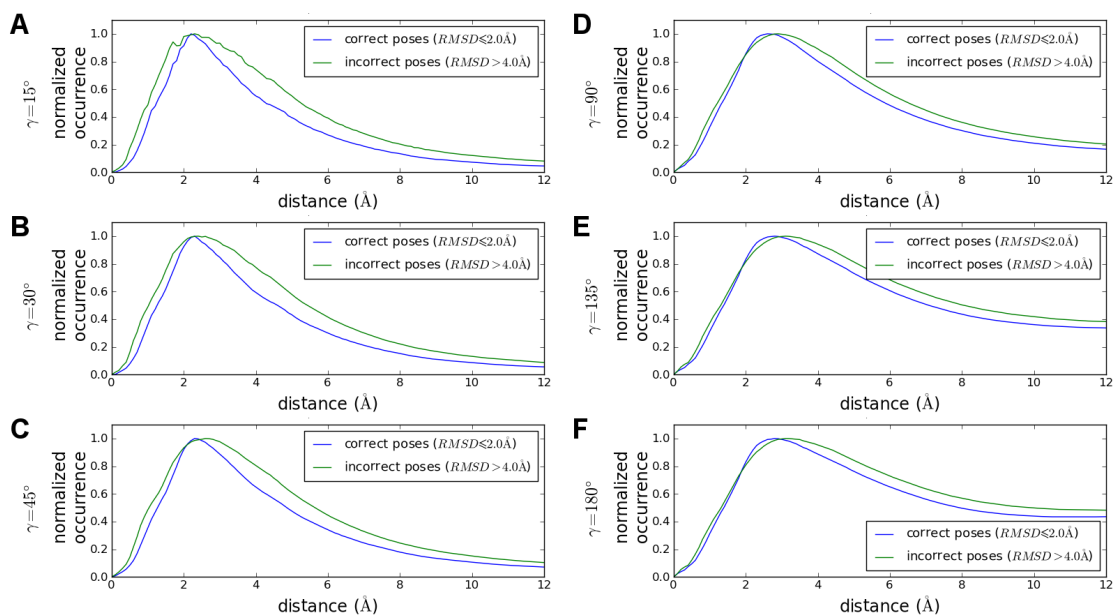


Figure A.2: Distributions of surface vertex points around the ligands of the Astex diverse set. Ligands are categorized into correct and incorrect poses based on their rmsd to the crystal structure. The scaled distributions are plotted for both categories for different γ_{max} angles: (A) 15° , (B) 30° , (C) 45° , (D) 90° , (E) 135° , (F) 180° .

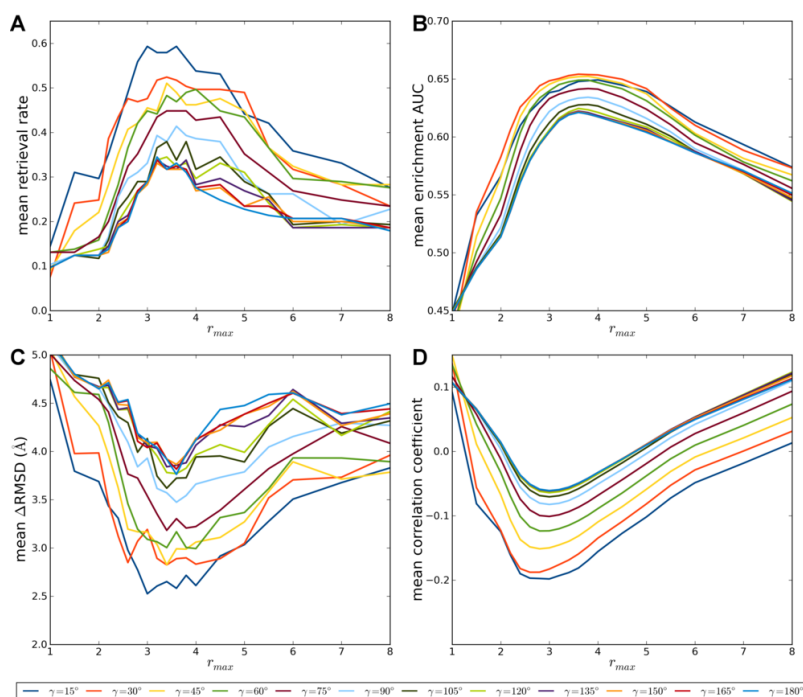


Figure A.3: Evaluation of the performance of *Bscore* based on the S3DB test set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

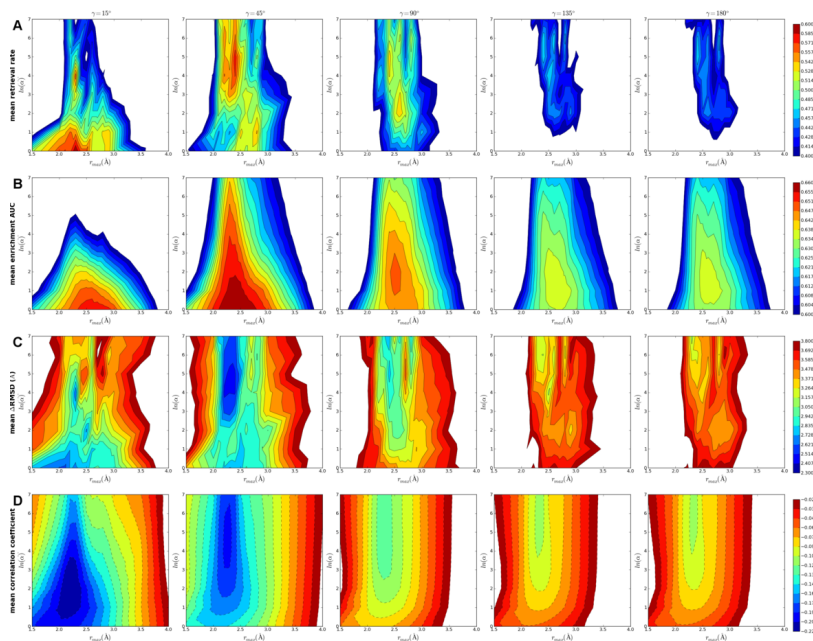


Figure A.4: Evaluation of the performance of *Bscore_g* based on the S3DB test set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

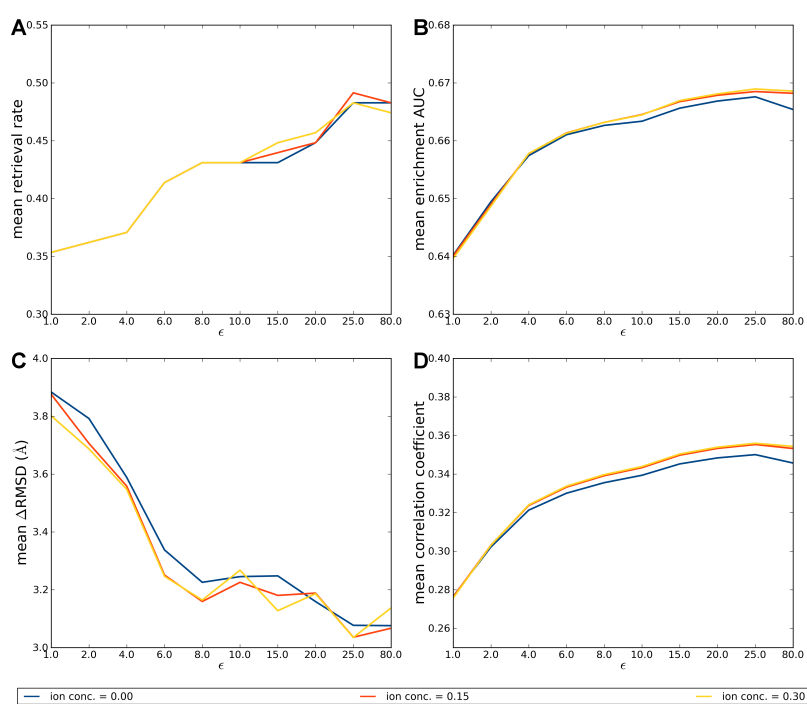


Figure A.5: Evaluation of the performance of *Escore* based on the S3DB set using (A) retrieval rate, (B) enrichment AUC, (C) Δ RMSD and (D) Pearson's correlation coefficient averaged over all complexes in the test set.

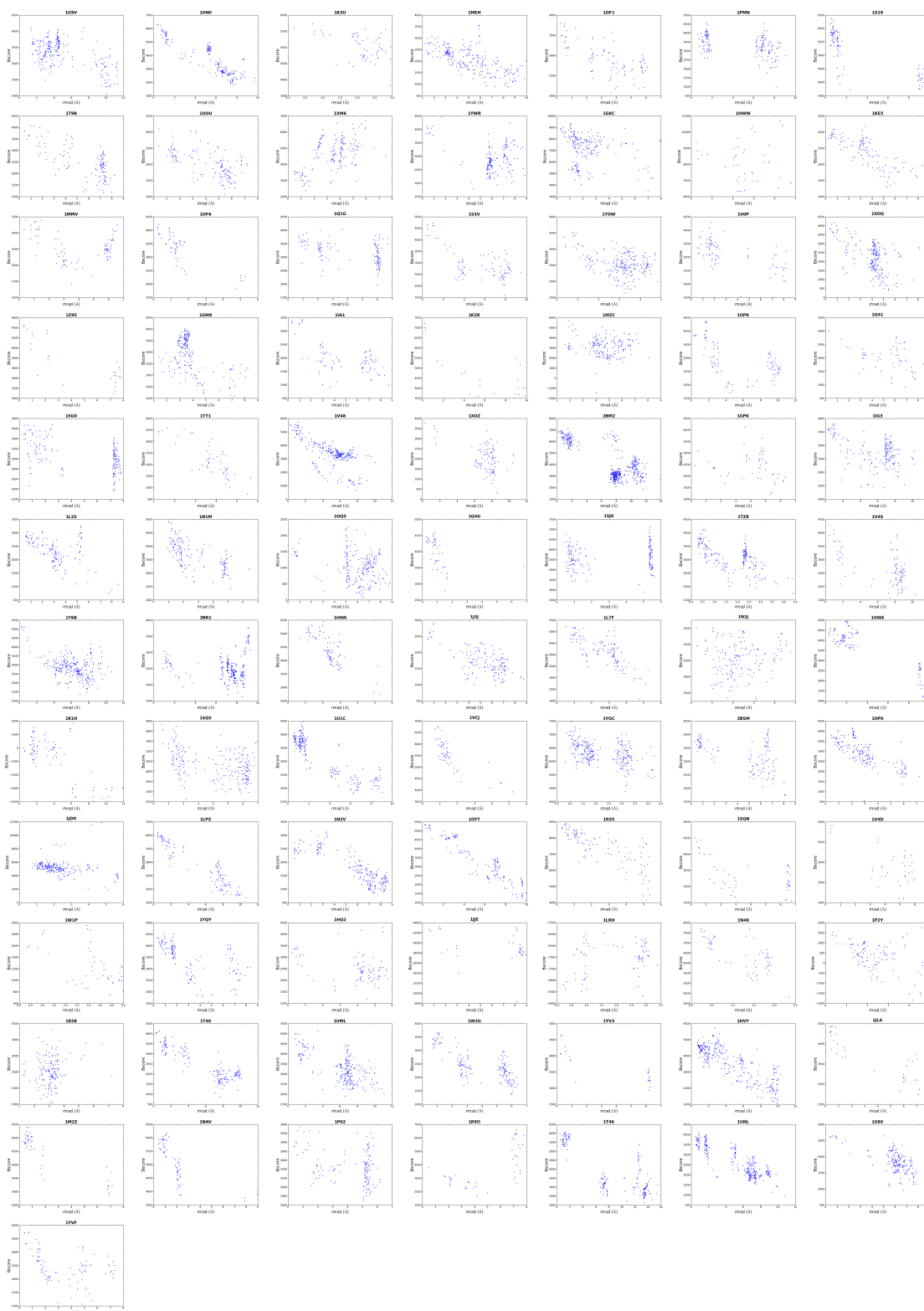


Figure A.6: $BEscore_g$ values (y-axis) plotted against symmetry corrected rmsd for each docked pose against the respective X-ray ligand conformation (x-axis) for the 85 complex structures in the Astex diverse set.

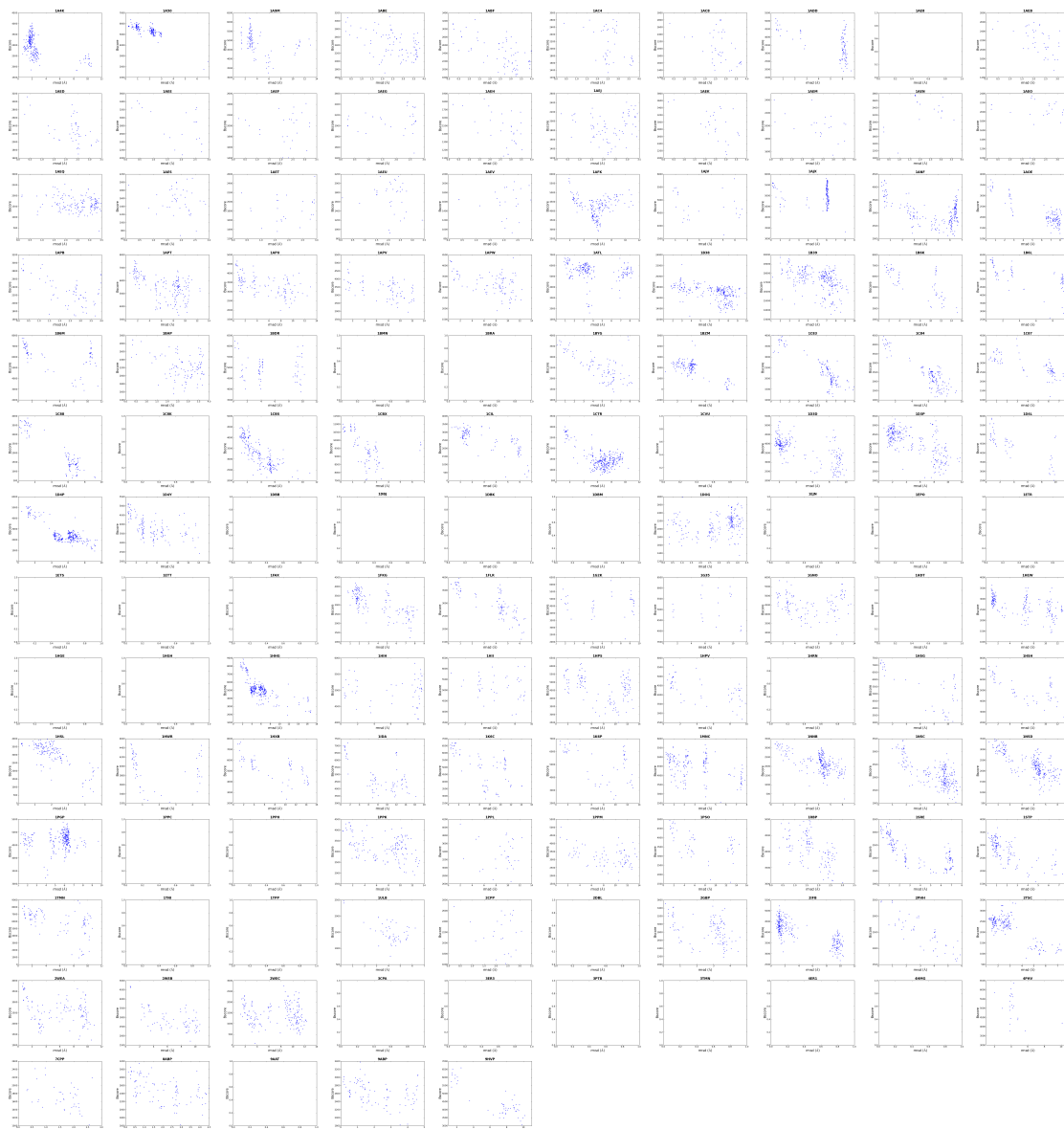


Figure A.7: $BEscore_g$ values (y-axis) plotted against symmetry corrected rmsd for each docked pose against the respective X-ray ligand conformation (x-axis) for the 145 complex structures in the S3DB set.

A.4 Human-Computer Interface Schematics

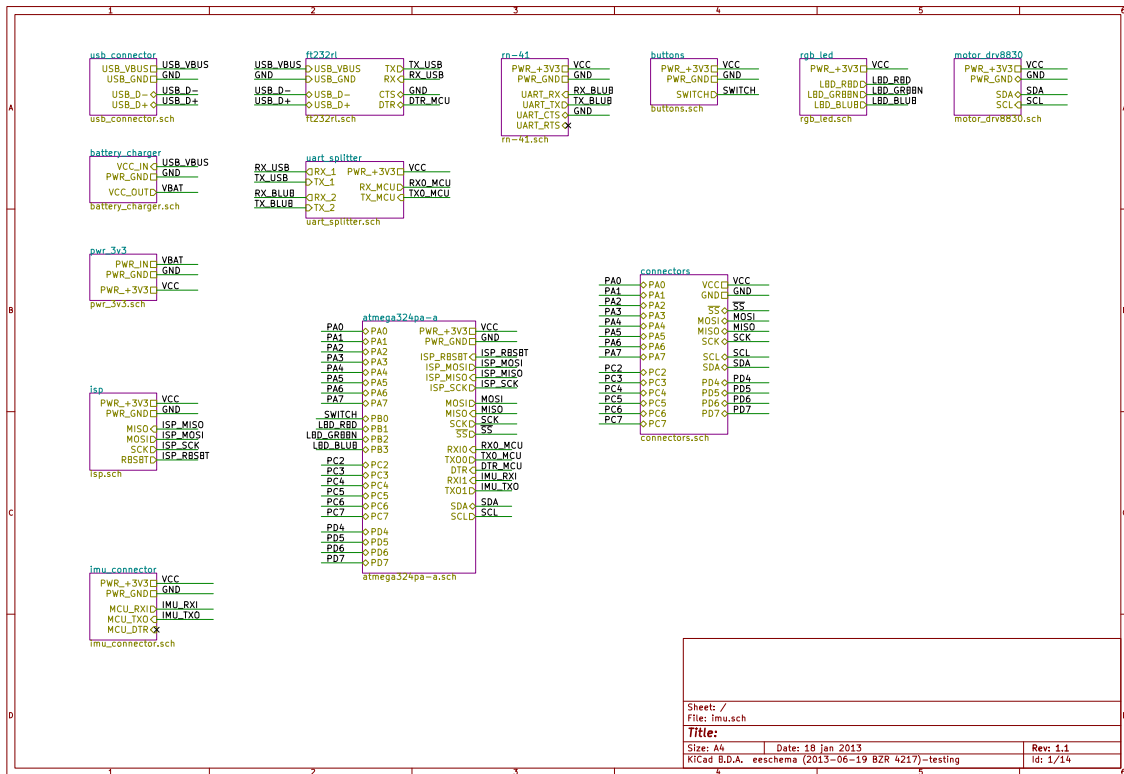


Figure A.9: Overview schematic drawing of the rotation input device.

