2013

Understanding Molecular Interactions: Application of HINT-based Tools in the Structural Modeling of Novel Anticancer and Antiviral Targets, and in Protein-Protein Docking

Hardik Parikh
Virginia Commonwealth University

UNDERSTANDING MOLECULAR INTERACTIONS: APPLICATIONS OF HINT-BASED TOOLS IN THE STRUCTURAL MODELING OF NOVEL ANTICANCER AND ANTIVIRAL TARGETS, AND IN PROTEIN-PROTEIN DOCKING

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commomwealth University

by

Hardik I. Parikh
Bachelor of Pharmacy, University of Pune, India, 2007

Director: Dr. Glen E. Kellogg, Ph. D.
Associate Professor,
Department of Medicinal Chemistry

Virginia Commomwealth University
Richmond, Virgina
May 2013

।। જય શ્રી કૃષ્ણ ।।

ACKNOWLEDGMENT

This dissertation would not have been possible without the guidance, help and support of several individuals. It is my pleasure to convey my gratitude to them all.

I would like to express my deepest appreciation, first and foremost, to my advisor, Dr. Glen E. Kellogg, for his absolute support, supervision and innovative ideas during the course of my graduate studies. The past five years have been a great learning experience for me, and I shall carry the lessons with me throughout my life. I am extremely fortunate to have him as my advisor. I am thankful to him for patiently correcting my writing, and also for financially supporting my research work. Without his guidance, patience and persistent help, this dissertation would have remained a dream.

My parents, Mr. Ishwar S. Parikh and Mrs. Amita I. Parikh, have always been there to support me in times of need. Thank you both for your unconditional love and encouragement at every stage of my life. I dedicate this work to them. My sister Mrs. Harshal A. Patwa and my brother-in-law Mr. Aanak G. Patwa deserve my wholehearted thanks as well. I would also like to thank members of my extended family for their best wishes.

I gratefully acknowledge Dr. Shijun Zhang, Dr. Michael A. McVoy, Dr. Umesh R. Desai and Dr. H. Tonie Wright for serving as members of my graduate student committee. I thank them for taking time out of their busy schedules for reviewing and evaluating my research. I am grateful for their helpful insights and invaluable suggestions throughout.

I would also like to thank current and former members of *Kellogg's Molecular Modeling and Drug Design Group* for sharing their research skills and experiences with me, in particular Dr. Aurijit Sarkar, Dr. Ashutosh Tripathi, Dr. Vishal Koparde, Dr. Philip Mosier, Dr. Alexander Bayden, Chenxiao "Max" Da, Saheem Zaidi, Mostafa Ahmed, and Jeremy Chojnacki.

My friends have helped me stay sane through these difficult years. Their love, care and support have helped me stay focused on my graduate studies. I would like to thank Swati, Kushal, Shreyas, Aagam, Dhawal, Mayank, Manali and Sayali, who have always been there for me. Words are not enough to acknowledge their influence in my life.

It gives me immense pleasure to thank all my friends in Richmond, especially Atul, Rio, Akul, Sweety, Shrenik, Pratik, Soumya, Khushboo, Shilpa, Priyanka, Suditi, Farhana, Tanvi, Divya, Batul, Shrinal, Arvind, Harshad, Shankar, Dipen, Jayul, Soundarya, Vidya, Della, Rajkumar, Jigar, Ronak, Rakesh, Dharik, Jugal and Vivek for being a family away from home. It is because of them that my graduate experience has been one that I will cherish forever. Thank you all!

Finally, I would like to acknowledge the School of Pharmacy, VCU for giving me the opportunity to undertake my graduate studies. I would also like to thank the VCU Graduate School for the 2012-2013 Thesis/Dissertation Assistantship for providing the financial support for the final year of my graduate studies.

Thank you, Lord, for always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

UNDERSTANDING MOLECULAR INTERACTIONS: APPLICATIONS OF HINT-BASED TOOLS IN THE STRUCTURAL MODELING OF NOVEL ANTICANCER AND ANTIVIRAL TARGETS, AND IN PROTEIN-PROTEIN DOCKING

By Hardik I. Parikh, Ph. D.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commomwealth University

Virginia Commomwealth University, 2013

Major Director: Dr. Glen E. Kellogg, Ph. D.
Associate Professor, Department of Medicinal Chemistry

Computationally driven drug design/discovery efforts generally rely on accurate assessment of the forces that guide the molecular recognition process. HINT (Hydropathic INTeraction) is a natural force field, derived from experimentally determined partition coefficients that quantifies all non-bonded interactions in the biological environment, including hydrogen bonding, electrostatic and hydrophobic interactions, and the energy of desolvation. The overall goal of this work is to apply the HINT-based atomic level description of molecular systems to biologically important

proteins, to better understand their biochemistry – a key step in exploiting them for therapeutic purposes.

This dissertation discusses the results of three diverse projects: i) structural modeling of human sphingosine kinase 2 (SphK2, a novel anticancer target) and binding mode determination of an isoform selective thiazolidine-2,4-dione (TZD) analog; ii) structural modeling of human cytomegalorvirus (HCMV) alkaline nuclease (AN) UL98 (a novel antiviral target) and subsequent virtual screening of its active site; and iii) explicit treatment of interfacial waters during protein-protein docking process using HINT-based computational tools.

SphK2 is a key regulator of the sphingosine-rheostat, and its upregulation /overexpression has been associated with cancer development. We report structural modeling studies of a novel TZD-analog that selectively inhibits SphK2, in a HINT analysis that identifies the key structural features of ligand and protein binding site responsible for isoform selectivity.

The second aim was to build a three-dimensional structure of a novel HCMV target – AN UL98, to identify its catalytically important residues. HINT analysis of the interaction of 5’ DNA end at its active site is reported. A parallel aim to perform *in silico* screening with a site-based pharmacophore model, identified several novel hits with potentially desirable chemical features for interaction with UL98 AN.

The majority of current protein-protein docking algorithms fail to account for water molecules involved in bridging interactions between partners, mediating and stabilizing

their association. HINT is capable of reproducing the physical and chemical properties of such waters, while accounting for their energetic stabilizing contributions. We have designed a solvated protein-protein docking protocol that explicitly models the Relevant bridging waters, and demonstrate that more accurate results are obtained when water is ***not*** ignored.

CHAPTER 1

INTRODUCTION

1.1 The Hydrophobic Effect

In chemistry, hydrophobicity (coming from Greek words **hydro**, meaning *water*, and **phobos**, meaning *fear*) is the physical property/tendency of a non-polar molecule to form aggregates in order to reduce their surface area exposed to a surrounding polar environment. This hydrophobic effect affects a number of diverse systems – from something as simple as immiscibility of oil in water to more complex phenomenon at the molecular level like protein folding and ligand binding.[1] Decades of research into understanding the biomolecular environment has established the fact that four major types of non-covalent interactions – hydrogen bonds, ionic interactions, Van der Waals interactions and hydrophobic interactions, govern nearly all processes at the molecular level, which is at the core of biological action.

While a "bond" is definitely not created, Kauzmann coined the term – *hydrophobic bond,* to describe the adhesion tendency of non-polar molecules within an aqueous solution.[2] Through the mid-twentieth century, the concept of a bond between two non-polar molecules was widely accepted, since this attractive force was unusually

strong. However, with a better understanding of the physical properties of dilute solutions of hydrophobic molecules in water, it was recognized that this attraction is a more complex phenomenon involving the configurational rearrangement of polar water molecules as the two hydrophobic species come together.[3] The association/aggregation of non-polar molecules due to the hydrophobic effect is energetically favorable due to the increase in entropy associated with the release/scattering of ordered water molecules surrounding them. The manifestations of the hydrophobic effect have been well reviewed in the literature.[1,4,5]

A large number of biological processes like protein folding; absorption, distribution, metabolism, and excretion of biological molecules; molecular recognition; protein-ligand interactions; protein-protein interactions; and more are governed by hydrophobicity. The hydrophobic effect plays a central role in guiding protein structure. Water-soluble proteins organize themselves such that amino acids with hydrophobic side-chains are retained within the core and buried from water, while amino acids with charged and polar side-chains are located on the solvent exposed surface, where they are capable of interacting with the surrounding water molecules. The hydrophobic effect makes a significantly large contribution towards the stability of globular proteins, and together with hydrogen bonding interactions within the core, drives the protein folding process.[6] The basic physical principles of molecular recognition are governed by thermodynamics, especially the Gibbs free energy (ΔG), which is given by –

$$\boldsymbol{\Delta G = \Delta H - T\Delta S} \quad (1)$$

described by the sum of changes in enthalpy (ΔH) and entropy ($-T\Delta S$) of the system.

The hydrophobic effect largely contributes towards the entropic term. Protein-ligand or protein-protein binding events involve bringing two molecules together, mainly via formation of electrostatic interactions like hydrogen bonds or hydrogen-bonding networks formed through water molecules. Those water molecules that are unable to locate within the binding interface are displaced into the bulk solvent, increasing entropy of the system.

For medicinal chemists engaged in drug discovery, the atomistic level understanding of how biomolecules associate (both intra- or intermolecular) and the precise chemical and physical features responsible for mediating them is of utmost importance. Structure-based drug design efforts rely on the knowledge of 3D structure of therapeutically relevant biomacromolecules and complexes. On the other hand, ligand-based drug design efforts rely only on the knowledge of physicochemical properties of small-molecules/ligands that bind to these biological targets of interest. These efforts aim to identify and optimize the biomolecular interactions, in search of novel, more potent drugs. The following section will review the different types of molecular interactions present in a single protein-ligand complex, from the vantage point of a medicinal chemist.

1.2 Molecular Interactions

A comprehensive review on the specific types of intermolecular interactions between ligands and their host molecules has been described elsewhere,[7] and should

be referred to for intricate details. Here, we will focus on some of the more frequently observed interactions observed in protein-ligand crystal structures.

One of the most important specific molecular interactions in ligand binding, biomacromolecular associations and related phenomenon is *hydrogen bonding*. Its structural and functional role in processes like protein secondary structure stabilization, protein folding, molecular recognition, enzymatic reactions, has been well documented.[8] It is a well-accepted fact that hydrogen bonds are primarily electrostatic and highly directional in nature.[9,10] The prevalence of different types of hydrogen bonding interactions within protein structures and in their complexes with ligands, like interactions between NH and carbonyl groups; between OH and carbonyl, ether and ester groups; and those with aromatic heterocycles; has been studied from crystal structure databases.[7,11-13] The traditional distance preference, the 'van der Waals distance cutoff', for identifying a hydrogen bond is too limiting and X–H•••A interactions with median distances between the proton and acceptor atom up to 3.0 Å have been observed.[14] Also, there are pronounced angular preferences for hydrogen bonds – with linear interactions (angles > 150°) preferred, although the location of electron density, molecular dipole and other neighboring intermolecular forces may influence the geometry to deviate.[7] While extremely important in conveying specificity to a recognition process, the contribution of hydrogen bonding towards net binding free energy gain is minimal in most cases as desolvation of the donor and acceptor atoms must occur for the interaction to form, and as a result, the effects of hydration and hydrogen bond formation counterbalance each other.[15] This also holds true for salt bridges, where the hydrogen bond distances are comparatively shorter and the interaction stronger.[16] Since

hydrogen bonds differ significantly in their intrinsic strengths, *Laurence, C. et al.* have recently introduced the pK_{BHX} scale, the hydrogen-bond basicity scale, to determine the relative strength of a hydrogen-bond acceptor.[17] This scale is created by measuring the equilibrium constant of the reaction of formation of hydrogen bond for a series of bases under the same conditions. It provides medicinal chemists a tool to probe the strength of a hydrogen bonding interaction systematically in an attempt to design more potent and selective analogues.[18,19] Other aspects of hydrogen bonds often observed in intra- and intermolecular interactions like cooperative hydrogen bonding, referring to additional hydrogen bonds in the vicinity that mutually strengthen each other;[20] multicentered/branched hydrogen bonds, referring to the stabilization of a hydrogen bond by additional partners to satisfy the hydrogen bonding potential;[21] and neighboring acceptor and donor groups that might weaken a hydrogen bond;[22] also require attention of the designer of drugs.

Numerous instances of *weaker hydrogen bonds*, involving non-classical donors and acceptors, have been observed in proteins.[21] The *π*-electron cloud of an aromatic ring can act as hydrogen bond acceptor to classical amide NH and hydroxyl OH donors, as well as hydrogens of aromatic XCH units polarized by neighboring heteroatoms (X = O, N).[23] Other frequently observed weak hydrogen bonds in crystal structures are CF•••HN and CF•••HO interactions, the Cα-H•••O=C interactions, the Cα-H•••F interactions, a SH•••*π-system* interaction, etc.[24-27] A different kind of hydrogen bond-like interaction is the *cation-π interaction*. Cations, from small ions like Li^+ to complex groups like guanidinuim and ammonium, are strongly attracted to *π* electrons of aromatic side-chains of Phe, Tyr, and Trp. Stacking interactions of the guanidino-group

of arginine with the aromatic rings of nucleic acid are conserved in some protein-nucleic acid complexes. Not just cations, but methyl groups bound to an electronegative atom (like alkylammonium group) are also capable of interacting with the π face of aromatic rings.[28,29]

Another type of non-covalent interaction observed between protein and ligand is the *halogen bond*. Halogen atoms of small-molecule ligands, when bound to aryl or electron withdrawing alkyl groups, show attraction towards carbonyl groups and other classical hydrogen-bond acceptors found in proteins, resulting in a C–X•••B type of interaction (X – halogen atom; B – electronegative acceptor atom).[30] The strength of a halogen bond is highly dependent on a number of things – the size of the halogen atom (the larger the halogen atom, the stronger the interaction), the electronegativity of the carbon substituent in the C–X partner (the higher the electronegativity, the stronger the interaction), and the electron density of the binding partner.[31,32] Similar to the *cation–π* interactions, a weak favorable interaction has also been observed between aromatic rings and halogen substituents.[33] Various groups have performed detailed analyses of different interactions involving halogen atoms, with results emphasizing the fact that halogens should not be viewed as only lipophilic groups, but can be utilized to form electrostatic interactions within the protein binding site under appropriate conditions.

One of the most important non-covalent interactions prevalent in almost all protein-ligand/protein-protein complexes is the *hydrophobic interaction*. The formation of strong interactions between non-polar ligands and a lipophilic protein pocket, formed by side chains of non-polar amino acids like leucine and phenylalanine, can be attributed to the hydrophobic effect. Similar to aggregation/micelle formation, a non-polar ligand

prefers to bind to a hydrophobic protein pocket resulting in decreased surface area exposed to waters, for both the ligand and the protein, and a subsequent gain in entropy, thus making the entire process energetically favorable. Hydrophobic interactions involving aryl-aryl/aryl-alkyl groups in host and guest molecules are facilitated by the electronic properties of the interacting aromatic rings. In case of aryl-aryl interactions, the aromatic side chains of protein residues interact with aryl rings of the ligand in a parallel-displaced stacking arrangement, maximizing overlap of the π-systems. The stacking arrangement between electron-rich hosts (due to electron-donating substituents) and complementary electron-deficient guests (due to electron-withdrawing substituents) affords charge transfer, thereby strengthening the interaction. In case of heterocyclic aromatic rings, the orientation of the interaction is controlled by the complementary alignment of partial charges on atoms and molecular dipoles.[34] In contrast to aryl-aryl interactions, the interaction between an alkyl group and an aromatic ring is more biased towards an edge-to-face geometry. The interaction energy of this type of interaction can be increased with increasing the acidity of the interacting CH unit of the alkyl partner.[35] Significant binding energy gains can be achieved by promoting intermolecular hydrophobic interactions, as evident from various studies showing correlation between binding affinity and the amount of hydrophobic surface buried upon ligand binding.[36]

Knowledge of the various interactions discussed above leads to better understanding of protein-ligand complexes, and draws the attention of a medicinal chemist / drug designer to the fact that there are multiple interactions involved in the binding process and any particular interaction must not be overemphasized.

1.3 Interfacial Structural Waters

Interfacial waters are not just “spectators” to ligand-binding or protein-protein association process; every such event involves displacement of water molecules from the binding site. The majority of biomolecular interactions occur in aqueous medium – each interacting partner is surrounded by water molecules, and the changes in the water structure upon their association contributes significantly towards the entropic component of the Gibbs free energy. Also, due to its high polarity, the presence of water molecules significantly changes the electrostatic interactions at the binding interface, and thereby contributes towards the enthalpic component too. Although having just three atoms, a single water molecule can engage in four hydrogen bonds (two as donor and two as acceptor). This enables water to mediate binding between protein and its ligand (both small-molecules and other proteins) via a hydrogen-bonding network.[37] In fact, analysis of thousands of protein-ligand crystal structures has revealed the presence of at least one water-mediated bridging interaction in each binding site.[38] Detailed analysis of the interface of protein complexes has shown that 40.1% of interfacial residues interact through water.[39] These studies highlight the importance of including the contribution of water molecules in rational ligand design.

Several studies have shown the utilization of structural waters at binding sites in ligand optimization efforts – either promoting water mediated interaction or designing an analogue that displaces it. One such example that illustrates the importance of water molecules at active site and its exploitation in inhibitor design is that of HIV-1 protease. The unliganded crystal structure of HIV-1 protease[40] shows the presence of catalytically important water Wat300, coordinated to an Asp residue in its active site. This water

molecule is consistently displaced when a ligand binds, disrupting the enzyme's catalytic activity. Another water molecule in its active site, Wat301, that forms hydrogen bonds with backbone amides of the two symmetry related Ile residues, is detected in crystal structures of the protease in its free form and in complexes with different ligands.[41] The conservation of this water molecule initially led to efforts directed at designing ligands that provide hydrogen bond acceptors to the protons of Wat301, resulting in potent inhibitors. However, ligands designed to displace this water proved to be even more tight binders, owing to the gain in entropy achieved by releasing that water into bulk solvent.[42,43] In another study, a nitrile substitution in the quinalozine and benztriazine inhibitors of scytalone dehydrogenase with an aim to displace water molecule from the active site, again resulted in more potent inhibitors.[44] Similarly, introduction of a nitrile group into quinazoline-based inhibitors of EGFR kinase led to more potent inhibitors, with the cyano-group interacting with the active site Thr in a similar manner to a water molecule.[45] Such studies and many more, have indicated that water molecules can be viewed as an extension of or addition to protein structural features, and they should be treated explicitly to assist rational ligand design and also guide modeling techniques like ligand docking / protein-protein docking.

1.4 Free Energy and Binding Affinity

The previous sections shed light on the different molecular interactions prevalent in protein-ligand complexes – direct molecular interactions, as well as the influence of water molecules. Drug discovery efforts largely rely on the accurate assessment of all

different effects that might influence the binding of small-molecule ligands with the target protein. The quantitative knowledge of the forces that guide the binding process thus becomes very important. The binding affinity of two molecules to form a complex must be rationalized and understood in terms of Gibb's Free Energy (ΔG).

A non-covalent, reversible association of protein (P) and ligand (L) to form a protein-ligand complex ($P'L'$) usually occurs in an aqueous solution, and can be represented by the following equation –

$$\boldsymbol{P_{aq.} + L_{aq.} \rightleftarrows P'L'_{aq.}} \tag{2}$$

Under thermodynamic equilibrium conditions, the association constant K_a (or the dissociation constant K_d) is given by the following equation –

$$\boldsymbol{K_a = K_d^{-1} = \frac{[P'L']}{[P][L]}} \tag{3}$$

The experimentally determined dissociation constant K_d (in case of enzyme inhibition, the inhibition constant K_i) can be related to the standard Gibb's free energy change of the dissociation ($\Delta_d G^\circ$) of $P'L'$ as –

$$\boldsymbol{\Delta_d G^\circ = RTlnK_d} \tag{4}$$

where R is the gas constant and T is the absolute temperature (T = 298 K).

The more negative the value of ΔG°, the smaller the dissociation constant K_d, and the stronger the binding. This relationship shows that the affinity of a molecule towards a target protein can be determined by calculating the associated changes in the thermodynamic parameters of the system – the changes in standard enthalpy and standard entropy upon complex formation.

It is generally accepted that protein-ligand binding events are determined by not only electrostatic interactions like hydrogen bonds, salt bridges, dipole-dipole interactions, interactions with metal ions, but also contributions from solvation/desolvation processes and spatial complementarity in van der Waals interactions.[46] The binding process is a complex phenomenon, and the enthalpy and entropy changes associated with direct interaction between protein and ligands are often not sufficient to describe the free energy change of the entire system. In the biological environment, both interacting partners are solvated before binding. The first event in the binding process is the desolvation of the ligand molecule and the protein-binding site, which contributes to free energy changes. Next is the conformational change in the protein side chains and ligand molecule, which also results in a change of entropy. This is followed by the energy gain attained from the molecular interactions forming between the interacting partners. Finally, if solvent accessible, the protein-ligand binding site is resolvated; with a favorable free energy if water molecules are set around polar/ionic groups and an unfavorable free energy if they are largely around hydrophobic groups. The total free energy change of the system (ΔG_{bind}) arises from contributions from each step of the binding process, and should be taken into account in calculations. The master equation can be written as –

$$\Delta \boldsymbol{G_{bind}} = \Delta \boldsymbol{G_{solvent}} + \Delta \boldsymbol{G_{conf}} + \Delta \boldsymbol{G_{int}} + \Delta \boldsymbol{G_{motion}} \quad (5)$$

where, $\Delta G_{solvent}$ is the hydration free energy, ΔG_{conf} free energy contribution due to conformational changes in the protein and ligand, ΔG_{int} is the free energy change due

to specific interaction between protein and ligand, and ΔG_{motion} is the free energy change due to "motion" in protein and ligand once they are proximal.[47]

1.4.1 *Experimental measurement of binding free energy*

Experimental determination of binding affinities can be achieved by indirect methods like binding assays. For enzyme reactions, the influence of ligand binding on enzyme kinetics results in a change in some physical property (like absorption, fluorescence, fluorescence polarization), which is subsequently measured. For receptor binding studies, a suitably labeled ligand is used. In both cases, the measured property is used to indirectly determine the binding constant and the standard free energy change (ΔG°). Recently, physicochemical techniques like Surface Plasmon Resonance spectroscopy,[48] NMR spectroscopy,[49] Mass spectroscopy,[50] and also Atomic-Force microscopy[51] have been used for indirect measurements of binding constants. A more direct measurement of binding affinities can be accomplished through microcalorimetric measurements, such as using Isothermal Titration Calorimetry (ITC). ITC is a highly versatile technique that allows determination of the complete thermodynamic profile (binding constant *K*, stoichiometry *n*, enthalpy change ΔH, entropy change ΔS) of a protein-ligand interaction from a single label-free experiment.[52]

For medicinal chemists, the affinity of small-molecule ligand for a macromolecular protein is of utmost importance as it serves as a benchmark criterion to define its biological activity. Lead discovery and optimization process is aimed at attaining better binding affinities by improving the intermolecular interactions, thereby leading to more potent drug candidates. However, synthesizing every potential analogue of a prototype candidate molecule and its subsequent experimental binding

affinity determinations are impractical, as they demand substantial resources. A more rapid, cost-efficient, approach would be to computationally predict the binding affinities of these analogues, which in turn will aid in selection of lead compounds or drug candidates with more favorable chemical characteristics.

1.4.2 *Theoretical calculation of binding free energy*

Various theoretical approaches for prediction of binding affinities have been well documented in literature.[15,46,53] Classical molecular mechanics forcefield-based scoring functions quantify protein-ligand interaction by focusing mostly on the steric and electrostatic forces involved. The solvation/desolvation effects and entropic contributions to binding events are often poorly described, and in some cases ignored.[54] Empirical scoring functions, on the other hand, are designed to approximate the binding affinities based on the individual interactions within a protein-ligand complex. The individual interaction terms accounting for favorable enthalpic contributions arising from electrostatic and hydrophobic contacts, as well as unfavorable entropic contributions arising from immobilization of rotatable bonds upon complex formation, may be implemented in the scoring function. The dependence of these methods on the quality of experimental data sets used to perform regression analysis and fitting, is a major reason for lack of accuracy in some approaches.[54] A third category of scoring functions – knowledge-based scoring functions – are based on simple atomic interaction-pair potentials derived from the observed frequencies of atom-pairing within crystal structures of known protein-ligand complexes.[55] Similar to empirical methods, a knowledge-based scoring function attempts to implicitly account for enthalpic and entropic contributions to binding. Despite the increasing number of scoring functions

being developed for accurate predictions of binding affinities, no general-purpose function is available. Most make various assumptions and simplifications for faster calculations, and do not completely describe every physical phenomenon involved in molecular recognition.[54] Indeed, the binding phenomenon itself is complex, with many moving parts, and only partially understood.

1.5 The HINT (Hydropathic INTeraction) Model[56-58]

Kellogg and Abraham designed a novel empirical force field named HINT (Hydropathic INTeraction) for calculating intermolecular interactions and free energies, based on experimentally determined partition coefficients $Log\ P_{o/w}$.

The octanol/water partition coefficient ($Log\ P_{o/w}$) of a molecule A is the ratio of its equilibrium concentration in a mixture of two immiscible solvents –

$$\boldsymbol{Log\ P_{o/w} = Log\ \frac{[A]_{octanol}}{[A]_{water}}} \tag{6}$$

1-octanol is a hydrophobic solvent that serves as a model environment to represent the biological phospholipid membrane. Thus, the distribution of a compound between water and 1-octanol provides accurate approximation of its partitioning between the cytosol and lipid membranes of living systems. The solvent partitioning phenomenon, much like the molecular recognition process, is governed by the same set of forces – the polar and electrostatic interactions guide the polar part of a molecule towards hydrophilic solvent (water) and the hydrophobic interactions guide the hydrophobic part of the molecule towards hydrophobic solvent (1-octanol). As a result, $Log\ P_{o/w}$ values implicitly

include the effects of entropy and solvation, along with other non-covalent interactions like hydrogen bonding, Coulombic, acid-base, hydrophobic interactions, etc. The HINT model uses the $Log\ P_{o/w}$ values for classification and quantitative scoring of molecular interactions, thereby incorporating both polar and hydrophobic complementarity, collectively referred to as hydropathy, between biomolecules. The quantitative solvent partitioning measurement (using experiments like the shake-flask method) can be viewed as free energy experiments and encode thermodynamic information, such that the standard free energy change of the solute transfer process ($\Delta G°$) can be related to its equilibrium constant ($Log\ P_{o/w}$) using the following equation –

$$\boldsymbol{Log\ P_{o/w} = -\Delta G°/2.303\ RT} \quad (6)$$

where R is the gas constant and T is the absolute temperature. Thus:

$$\boldsymbol{Log\ P_{o/w} = k\ \Delta G°} \quad (7)$$

where k = –0.733 kcal/mol at 298K.

This relationship shows that HINT force field, which is based on $Log\ P_{o/w}$, can be used to estimate the free energy of a binding process, and thereby predict the binding affinities of ligands.

The HINT molecular interaction model calculates free energy scores for the hydropathic interactions within/between biomolecules by quantifying each atom-atom pair interaction using the following equation –

$$\boldsymbol{b_{ij} = a_i\ S_i\ a_j\ S_j\ T_{ij}\ R_{ij} + r_{ij}} \quad (8)$$

where b_{ij} is the hydropathic interaction score between atoms i and j. a is the hydrophobic atom constant, S is the solvent accessible surface area (calculated using a H_2O probe), T_{ij} is a logic function assuming +1 or –1 value depending on the character of the interacting polar atoms, and the distance dependent functions R_{ij} and r_{ij} are simple exponential function e^{-r} and an implementation of the Lennard-Jones potential function, respectively. The total score of the system is then calculated by taking a double sum over every atom-atom pair –

$$\boldsymbol{HINT_{total} = \Sigma\Sigma\, b_{ij}} \tag{9}$$

The HINT force field scores favorable interactions with $b_{ij} > 0$, and unfavorable interactions with $b_{ij} < 0$. The value of logic function T_{ij} depends on the type of interacting atoms. A favorable hydrophobic-hydrophobic interaction gets a $T_{ij} = +1$ value; an unfavorable hydrophobic-polar interaction gets a $T_{ij} = -1$ value; for a polar-polar interaction, $T_{ij} = +1$ if the interacting atoms are an acid and a base, whereas $T_{ij} = -1$ for an unfavorable acid-acid/base-base type interaction. Table 1.1 shows the T_{ij} interaction matrix.

The hydrophobic atom constant, a, is calculated by an adaptation of a partitioning algorithm CLOP of Hansch and Leo,[59] which calculates the total solvation partition constant for a molecule by summation of fragment constants into a single value. HINT, on the other hand, uses a slightly different approach – it distributes and assigns

Table 1.1 – T_{ij} interaction matrix[57]

Atom Type [atom constant]	***H** (apolar)* [a > 0]	***H*** (polar) [a > 0]	***C*** (apolar) [a > 0]	***Polar*** (N, O, etc) [a < 0]
***H** (apolar)* [a > 0]	+ 1	- 1	+ 1	- 1
H (polar) [a > 0]	- 1	- 1	- 1	+ 1
C (apolar) [a > 0]	+ 1	- 1	+ 1	- 1
Polar (N, O, etc) [a < 0]	- 1	+ 1	- 1	- 1

Colors code: Green – favorable hydrophobic-hydrophobic; Red – unfavorable hydrophobic-polar; Blue – favorable acid-base or hydrogen bond; Yellow – unfavorable acid-acid; Orange – generally unfavorable base-base, but may depend on charge.

hydrophobic atom constants and factors to each atom in the molecule. HINT uses values from a functional group primitive dataset of small molecules and bio-macromolecules, with re-parameterized force field atom types; modified factors for bond, branching, ring and chain factors; and polar proximity factors. The modifications represent real biophysical phenomena related to the molecular structure and properties. The hydrophobic atom constant for every atom of the molecule is calculated by modifying the factors based on the atom's structural connectivity and proximity to other atoms. The $Log\ P$ of the molecule can be considered as the sum of individual hydrophobic atom constants (a_i) –

$$\boldsymbol{Log\ P = \Sigma\, a_i} \tag{10}$$

Most importantly, the hydrophobic atom constant is a thermodynamic parameter whose sign and magnitude reveals the potential type and strength of the interaction that the atom may engage in.

One of the most significant aspects of the HINT model is that it is empirical in nature and approximates all non-covalent interactions in the biological environment, including hydrogen bonding, electrostatic and hydrophobic interactions. Moreover, entropy and solvation/desolvation effects are also implicitly encoded in the $Log\ P$ data. The hydrophobic atom constants (a_i) are parameters directly derived from the free energy of atom transfer between two solvents – which means that these solvents serve as model environments for hydrophobic and polar regions of biomolecules and the free energy of atom transfer between hydrophobic and polar regions of biomolecules is the same as that between 1-octanol and water. That is, the a_i values of each atom indicate

how it will interact with other atoms in a biological micro-environment, much as how it interacts with solvent molecules/atoms. For example, for atoms i and j, positive values of a_i and a_j imply that they are hydrophobic and the HINT algorithm would score the i/j interaction favorably ($b_{ij} > 0$). If a_i and a_j are both negative, and one is a Lewis acid while the other is a Lewis base, the HINT algorithm would score the i/j interaction favorably also ($b_{ij} > 0$). However, if a_i is positive and a_j is negative, the i/j hydrophobic-polar type interaction would be scored unfavorably ($b_{ij} < 0$). Table 1.2 shows the matrix of atom-atom interaction types characterized and scored by the HINT algorithm.

The HINT force field and its free energy scoring form the basis for quantitative assessment of molecular interactions, which as we have discussed before, has a direct consequence in drug design. Within a homogeneous biological set (*i.e.*, within families of different ligands binding to the same protein site), HINT scores can be easily correlated to the binding free energy associated with protein-ligand complex formation. *Kellogg et al.* have shown that total HINT interaction scores correlate with the $\Delta G_{binding}$ (Figure 1.1) for a diverse set of 76 protein-ligand complexes at resolution better than 3.2 Å, with a standard error of ± 2.33 kcal mol^{-1} using the equation –

$$\Delta \boldsymbol{G_{binding}} = -\mathbf{0.0019}\, \boldsymbol{H_{total}} - \mathbf{3.927} \qquad (11)$$

A better correlation was achieved within a subset of 56 complexes structurally determined at a resolution better than 2.5 Å (r = 0.85, SE = ± 1.8 kcal mol^{-1}).[60]

Table 1.2 – Types of atom-atom interactions characterized and scored by HINT force fields[57]

	Hydrophobic	**Polar** – Lewis Acid (H-bond donor)	**Polar** – Lewis Base (H-bond acceptor)
Hydrophobic	Hydrophobic interaction	Hydrophobic – Polar (desolvation energy)	Hydrophobic – Polar (desolvation energy)
Polar – Lewis Acid (H-bond donor)	Hydrophobic – Polar (desolvation energy)	Coulombic repulsion	Acid – Base (Hydrogen bond)
Polar – Lewis Base (H-bond acceptor)	Hydrophobic – Polar (desolvation energy)	Acid – Base (Hydrogen bond)	Coulombic repulsion

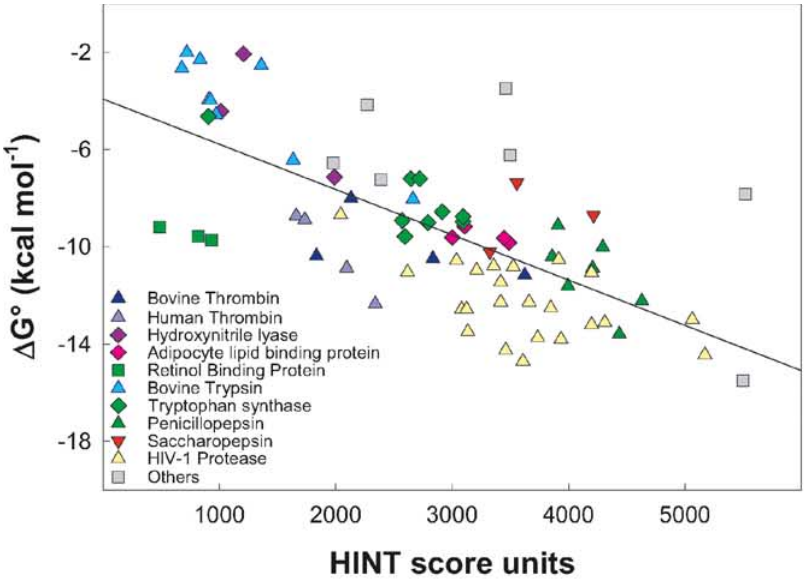


Figure 1.1 Correlation between calculated HINT scores and measured free energy of binding for 76 diverse protein-ligand complexes (crystallographic resolution better than 3.2 Å). Reprint from ref 60.

The capability of predicting the free energy of binding, especially since it has terms for hydrophobic interactions, in addition to the electrostatic, van der Waals and desolvation, makes HINT a suitable tool for scoring predicted binding modes of ligands in a docking experiment, in order to identify the best conformation. Several studies have successfully used HINT to distinguish active molecules from inactive ones.[61-64]

One of the earliest applications of HINT was incorporation of the HINT hydropathic fields into 3D-QSAR program CoMFA,[65] which originally had just two fields, steric and electrostatic. The introduction of hydropathic field enables full description of binding events and aids in designing new molecules based on such QSAR studies, especially in cases where ligands and/or active sites are predominantly non-polar in nature. Several studies based on the HINT–CoMFA are reported in the literature.[66-69]

An important consideration for reliable modeling results is the exact designation of protonation state of ionizable groups of both protein side chains and ligands, which can have significant influence on binding affinities. The hydropathic analysis can be performed using the HINT force field to assign the location of hydrogens on functional groups within the binding pocket, using an extension of the HINT model known as Computational Titration.[70] Multiple potential ionization states for protein and ligand are enumerated and analyzed by the HINT model, with the best scoring complex representing the optimum state of binding and its corresponding protonation ensemble.

As discussed before, interfacial water molecules mediate and stabilize the molecular recognition phenomenon – either directly via hydrogen bonding network or indirectly via solvation/desolvation processes. The contribution of bulk solvent towards the hydrophobic effect (the desolvation energy, entropic in nature) is implicitly encoded

within the $Log\ P_{o/w}$ data. However, the solvent molecules forming bridging interactions between ligand and protein must be explicitly accounted for. The HINT force field can calculate the global interaction score for a water-mediated protein-ligand interaction (HS_{total}) by incorporating the contribution made by water –

$$\boldsymbol{HS_{total} = HS_{prot-lig} + HS_{lig-water} + \left[HS_{prot-water}\right]} \quad (12)$$

where, $HS_{prot-lig}$ is the interaction score between protein and ligand; $HS_{lig-water}$ is the interaction score between ligand atoms and water molecules; and $HS_{prot-water}$ is the interaction score between protein atoms and water molecules (at the binding site), which can be ignored if those waters are preexisting, *i.e.*, part of the protein. This global interaction score was shown to correlate with experimentally determined binding constants better than the scores when waters were ignored (SE = ± 0.98 kcal mol^{-1}).[71]

The capabilities of HINT force field, and tools based on it, make it a highly versatile tool with numerous applications in drug designing. The HINT toolkit, a set of linkable subroutines that access HINT energy scoring functions and HINT 3D grid map objects, is made available to the scientific community and can be incorporated into programs for computer aided drug discovery.[72]

1.6 Research Plan

In addition to quantitatively scoring molecular interactions, the HINT force field can – (i) create hydropathic fields for ligands within a protein environment, (ii) rationally evaluate the correct ionization states of functional groups within a protein binding site,

(iii) incorporate energetic contribution of interfacial waters, and (iv) enhance crystallographic data by optimizing the protein residue interaction environment. The overall aim of this research work was to utilize the HINT force field and HINT-based computational tools in various aspects of molecular modeling. Our overarching goal is to apply this atomistic-level simulation technology to important biological proteins in order to gain structural insights into their mechanism of action that can be exploited for designing compounds intended to inhibit them.

Sphingosine Kinase (SphK) is a key regulator of the sphingosine rheostat, which maintains optimum levels of a lipid metabolite sphingosine-1-phosphate (an anti-apoptotic agent).[73] Overexpression and/or upregulation of SphK have been associated with various aspects of cancer development.[74-76] Biological characterization of a thiazolidine-2,4-dione (TZD) analogue identified it as an isoform-selective SphK2 inhibitor.[77] We asked ourselves if the key structural features of the ligand and the protein binding site, which makes it isoform selective, could be identified using the HINT force field. This is important to optimize the lead compound for future drug development. In Chapter 2, we will discuss the protein structure building process for the two human isoforms of SphK (SphK1 and SphK2). This was followed by molecular docking of the ligand to its putative binding site on the kinases. Using molecular docking and HINT free energy scoring to identify the probable native-like conformation of ligand within the binding pocket, we propose a binding mode for the TZD-analogue showing a preference for SphK2.

Human cytomegalovirus (HCMV) is a human pathogen responsible for diseases in immune-compromised and HIV patients, and severe birth defects when acquired

during pregnancy.[78] The HCMV Alkaline Nuclease (AN) UL98, vital for viral replication, represents a novel target for development of antivirals.[79] In the absence of a crystallographic structure, we wanted to see if molecular modeling techniques could be used to build a structural model to identify UL98's catalytically important residues. In Chapter 3, we will discuss the homology-based structural modeling of UL98 AN. The computational model has been experimentally validated, and subsequently used to perform a structure-based virtual screening with an aim of identifying novel agents capable of inhibiting UL98.

Finally, in Chapter 4 we will discuss the utility of HINT based tools in designing a solvated protein-protein docking protocol. With increasing interest in targeting protein-protein associations to interrupt biochemical pathways, and relatively few crystal structures of protein-protein complexes available, it becomes very important to have computational tools for accurate prediction of these biomacromolecular associations. Various HINT-based studies have shown the importance of interfacial waters in mediating and stabilizing protein-protein complexes.[71,80-82] The majority of current docking algorithms include the effects of solvent by introducing desolvation energy terms in their scoring functions; however, they fail to account for the water molecules involved in bridging interactions. With HINT-based tools that can explicitly account for interfacial waters at our disposal, we wanted to check the influence of accounting for their energetic contributions on the outcome of docking predictions. Using these tools, we have designed a solvated protein-protein docking protocol that explicitly models the Relevant bridging interfacial waters, and demonstrate that more accurate results are obtained.

References

1. Meyer, E. E.; Rosenberg, K. J.; Israelachvili, J. Recent progress in understanding hydrophobic interactions. *Proc. Natl. Acad. Sci. U.S.A.* **2006,** *103,* 15739-15746.
2. Kauzmann, W. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **1959,** *14,* 1-63.
3. Israelachvili, J.; Pashley, R. The hydrophobic interaction is long range, decaying exponentially with distance. *Nature* **1982,** *300,* 341-342.
4. Blokzijl, W.; Engberts, J. B. Hydrophobic effects. Opinions and facts. *Angew. Chem. Int. Ed.* **1993,** *32,* 1545-1579.
5. Sarkar, A.; Kellogg, G. E. Hydrophobicity - Shake flasks, protein folding and drug discovery. *Curr. Top. Med. Chem.* **2010,** *10,* 67-83.
6. Pace, C. N.; Shirley, B. A.; McNutt, M.; Gajiwala, K. Forces contributing to the conformational stability of proteins. *FASEB J.* **1996,** *10,* 75-83.
7. Bissantz, C.; Kuhn, B.; Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **2010,** *53,* 5061-5084.
8. Derewenda, Z. S.; Lee, L.; Derewenda, U. The occurrence of C–H•••O hydrogen bonds in proteins. *J. Mol. Biol.* **1995,** *252,* 248-262.
9. Legon, A.; Millen, D. Directional character, strength, and nature of the hydrogen bond in gas-phase dimers. *Acc. Chem. Res.* **1987,** *20,* 39-45.
10. Mitchell, J.; Price, S. On the electrostatic directionality of N–H•••O=C hydrogen bonding. *Chem. Phys. Lett.* **1989,** *154,* 267-272.
11. Taylor, R.; Kennard, O.; Versichel, W. Geometry of the imino-carbonyl (NH•••O:C) hydrogen bond. 1. Lone-pair directionality. *J. Am. Chem. Soc.* **1983,** *105,* 5761-5766.
12. Lommerse, J. P.; Price, S. L.; Taylor, R. Hydrogen bonding of carbonyl, ether, and ester oxygen atoms with alkanol hydroxyl groups. *J. Comput. Chem.* **1997,** *18,* 757-774.
13. Nobeli, I.; Price, S.; Lommerse, J.; Taylor, R. Hydrogen bonding properties of oxygen and nitrogen acceptors in aromatic heterocycles. *J. Comput. Chem.* **1997,** *18,* 2060-2074.
14. Steiner, T. The hydrogen bond in the solid state. *Angew. Chem. Int. Ed.* **2002,** *41,* 48-76.
15. Hunter, C. A. Quantifying intermolecular interactions: Guidelines for the molecular recognition toolbox. *Angew. Chem. Int. Ed.* **2004,** *43,* 5310-5324.
16. Luo, R.; David, L.; Hung, H.; Devaney, J.; Gilson, M. K. Strength of solvent-exposed salt-bridges. *J. Phys. Chem. B* **1999,** *103,* 727-736.
17. Laurence, C.; Brameld, K. A.; Graton, J.; Le Questel, J.-Y.; Renault, E. The pK(BHX) database: Toward a better understanding of hydrogen-bond basicity for medicinal chemists. *J. Med. Chem.* **2009,** *52,* 4073-4086.
18. Morris, J. J.; Hughes, L. R.; Glen, A. T.; Taylor, P. J. Non-steroidal antiandrogens. Design of novel compounds based on an infrared study of the

dominant conformation and hydrogen-bonding properties of a series of anilide antiandrogens. *J. Med. Chem.* **1991,** *34,* 447-455.
19. Bingham, A. H.; Davenport, R. J.; Gowers, L.; Knight, R. L.; Lowe, C.; Owen, D. A.; Parry, D. M.; Pitt, W. R. A novel series of potent and selective IKK2 inhibitors. *Bioorg. Med. Chem. Lett.* **2004,** *14,* 409-412.
20. DeChancie, J.; Houk, K. N. The origins of femtomolar protein-ligand binding: Hydrogen-bond cooperativity and desolvation energetics in the biotin-(strept)avidin binding site. *J. Am. Chem. Soc.* **2007,** *129,* 5419-5429.
21. Panigrahi, S. K.; Desiraju, G. R. Strong and weak hydrogen bonds in the protein-ligand interface. *Proteins* **2007,** *67,* 128-141.
22. Jorgensen, W. L.; Pranata, J. Importance of secondary interactions in triply hydrogen bonded complexes: Guanine-cytosine vs uracil-2,6-diaminopyridine. *J. Am. Chem. Soc.* **1990,** *112,* 2008-2010.
23. Steiner, T.; Koellner, G. Hydrogen bonds with pi-acceptors in proteins: Frequencies and role in stabilizing local 3D structures. *J. Mol. Biol.* **2001,** *305,* 535-557.
24. Dunitz, J. D.; Taylor, R. Organic fluorine hardly ever accepts hydrogen bonds. *Chem.-Eur. J.* **1997,** *3,* 89-98.
25. Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. How Strong Is the Calpha-H•••O=C Hydrogen Bond? *J. Am. Chem. Soc.* **2000,** *122,* 4750-4755.
26. Bohm, H.-J.; Banner, D.; Bendels, S.; Kansy, M.; Kuhn, B.; Muller, K.; Obst-Sander, U.; Stahl, M. Fluorine in medicinal chemistry. *ChemBioChem* **2004,** *5,* 637-643.
27. Ringer, A. L.; Senenko, A.; Sherrill, C. D. Models of S/pi interactions in protein structures: Comparison of the H2S-benzene complex with PDB data. *Protein Sci.* **2007,** *16,* 2216-2223.
28. Mecozzi, S.; West Jr, A. P.; Dougherty, D. A. Cation-pi interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. U.S.A.* **1996,** *93,* 10566-10571.
29. Biot, C.; Buisine, E.; Kwasigroch, J.-M.; Wintjens, R.; Rooman, M. Probing the energetic and structural role of amino acid/nucleobase cation-pi interactions in protein-ligand complexes. *J. Biol. Chem.* **2002,** *277,* 40816-40822.
30. Ramasubbu, N.; Parthasarathy, R.; Murray-Rust, P. Angular preferences of intermolecular forces around halogen centers: Preferred directions of approach of electrophiles and nucleophiles around carbon-halogen bond. *J. Am. Chem. Soc.* **1986,** *108,* 4308-4314.
31. Glaser, R.; Chen, N.; Wu, H.; Knotts, N.; Kaupp, M. 13C NMR study of halogen bonding of haloarenes: Measurements of solvent effects and theoretical analysis. *J. Am. Chem. Soc.* **2004,** *126,* 4412-4419.
32. Sarwar, M. G.; Dragisic, B.; Salsberg, L. J.; Gouliaras, C.; Taylor, M. S. Thermodynamics of halogen bonding in solution: Substituent, structural, and solvent effects. *J. Am. Chem. Soc.* **2010,** *132,* 1646-1653.
33. Prasanna, M.; Guru Row, T. C-halogen•••pi interactions and their influence on molecular conformation and crystal packing: A database study. *Cryst. Eng.* **2000,** *3,* 135-154.

34. Meyer, E. A.; Castellano, R. K.; Diederich, F. o. Interactions with aromatic rings in chemical and biological recognition. *Angew. Chem. Int. Ed.* **2003,** *42,* 1210-1250.
35. Sinnokrot, M. O.; Sherrill, C. D. Substituent effects in pi-pi interactions: Sandwich and T-shaped configurations. *J. Am. Chem. Soc.* **2004,** *126,* 7690-7697.
36. Boehm, H.-J.; Klebe, G. What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs? *Angew. Chem. Int. Ed.* **1996,** *35,* 2588-2614.
37. Li, Z.; Lazaridis, T. Water at biomolecular binding interfaces. *Phy. Chem. Chem. Phy.* **2007,** *9,* 573-581.
38. Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003,** *326,* 607-620.
39. Teyra, J.; Pisabarro, M. T. Characterization of interfacial solvent in protein complexes and contribution of wet spots to the interface description. *Proteins* **2007,** *67,* 1087-1095.
40. Pillai, B.; Kannan, K.; Hosur, M. 1.9 A x-ray study shows closed flap conformation in crystals of tethered HIV-1 PR. *Proteins* **2001,** *43,* 57-64.
41. Priestle, J. P.; Fassler, A.; Rosel, J.; Tintelnot-Blomley, M.; Strop, P.; Grutter, M. G. Comparative analysis of the X-ray structures of HIV-1 and HIV-2 proteases in complex with CGP 53820, a novel pseudosymmetric inhibitor. *Structure* **1995,** *3,* 381-389.
42. Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; et al. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* **1994,** *263,* 380-384.
43. Schaal, W.; Karlsson, A.; Ahlsen, G.; Lindberg, J.; Andersson, H. O.; Danielson, U. H.; Classon, B.; Unge, T.; Samuelsson, B.; Hulten, J.; Hallberg, A.; Karlen, A. Synthesis and comparative molecular field analysis (CoMFA) of symmetric and nonsymmetric cyclic sulfamide HIV-1 protease inhibitors. *J. Med. Chem.* **2001,** *44,* 155-169.
44. Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. Structure-based design of potent inhibitors of scytalone dehydratase: Displacement of a water molecule from the active site. *Biochemistry* **1998,** *37,* 17735-17744.
45. Wissner, A.; Berger, D. M.; Boschelli, D. H.; Floyd, M. B., Jr.; Greenberger, L. M.; Gruber, B. C.; Johnson, B. D.; Mamuya, N.; Nilakantan, R.; Reich, M. F.; Shen, R.; Tsou, H. R.; Upeslacis, E.; Wang, Y. F.; Wu, B.; Ye, F.; Zhang, N. 4-Anilino-6,7-dialkoxyquinoline-3-carbonitrile inhibitors of epidermal growth factor receptor kinase and their bioisosteric relationship to the 4-anilino-6,7-dialkoxyquinazoline inhibitors. *J. Med. Chem.* **2000,** *43,* 3244-3256.
46. Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.* **2002,** *41,* 2644-2676.
47. Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* **1995,** *38,* 4953-4967.

48. Rich, R. L.; Myszka, D. G. Advances in surface plasmon resonance biosensor analysis. *Curr. Opin. Biotechnol.* **2000,** *11,* 54-61.
49. Hicks, R. P. Recent advances in NMR: Expanding its role in rational drug design. *Curr. Med. Chem.* **2001,** *8,* 627-650.
50. Veenstra, T. D. Electrospray ionization mass spectrometry in the study of biomolecular non-covalent interactions. *Biophys. Chem.* **1999,** *79,* 63-79.
51. Janshoff, A.; Neitzert, M.; Oberdorfer, Y.; Fuchs, H. Force spectroscopy of molecular systems-single molecule spectroscopy of polymers and biomolecules. *Angew. Chem. Int. Ed.* **2000,** *39,* 3212-3237.
52. Olsson, T. S.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. The thermodynamics of protein-ligand interaction and solvation: Insights for ligand design. *J. Mol. Biol.* **2008,** *384,* 1002-1017.
53. Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Molecular mechanics methods for predicting protein-ligand binding. *Phys. Chem. Chem. Phys.* **2006,** *8,* 5166-5177.
54. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **2004,** *3,* 935-949.
55. Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006,** *49,* 5895-5902.
56. Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput. Aided Mol. Des.* **1991,** *5,* 545-552.
57. Kellogg, G. E.; Abraham, D. J. Hydrophobicity: is LogP(o/w) more than the sum of its parts? *Eur. J. Med. Chem.* **2000,** *35,* 651-661.
58. Kellogg, G. E.; Burnett, J. C.; Abraham, D. J. Very empirical treatment of solvation and entropy: A force field derived from Log Po/w. *J. Comput. Aided Mol. Des.* **2001,** *15,* 381-393.
59. Hansch, C.; Leo, A. Substituent constants for correlation analysis in chemistry and biology. J. Wiley and Sons Inc.: NY, 1979.
60. Kellogg, G. E.; Fornabaio, M.; Spyrakis, F.; Lodola, A.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J. Getting it right: Modeling of pH, solvent and "nearly" everything else in virtual screening of biological targets. *J. Mol. Graph. Model.* **2004**, *22*, 479-486.
61. Wang, S.; Liu, M.; Lewin, N. E.; Lorenzo, P. S.; Bhattacharrya, D.; Qiao, L.; Kozikowski, A. P.; Blumberg, P. M. Probing the binding of indolactam-V to protein kinase C through site-directed mutagenesis and computational docking simulations. *J. Med. Chem.* **1999,** *42,* 3436-3446.
62. Spyrakis, F.; Amadasi, A.; Fornabaio, M.; Abraham, D. J.; Mozzarelli, A.; Kellogg, G. E.; Cozzini, P. The consequences of scoring docked ligand conformations using free energy correlations. *Eur. J. Med. Chem.* **2007,** *42,* 921-933.
63. Tripathi, A.; Fornabaio, M.; Kellogg, G. E.; Gupton, J. T.; Gewirtz, D. A.; Yeudall, W. A.; Vega, N. E.; Mooberry, S. L. Docking and hydropathic scoring of polysubstituted pyrrole compounds with antitubulin activity. *Bioorg. Med. Chem. Lett.* **2008,** *16,* 2235-2242.

64. Da, C.; Telang, N.; Barelli, P.; Jia, X.; Gupton, J. T.; Mooberry, S. L.; Kellogg, G. E. Pyrrole-based antitubulin agents: Two distinct binding modalities are predicted for C-2 analogs in the colchicine site. *ACS Med. Chem. Lett.* **2012,** *3,* 53-57.
65. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988,** *110,* 5959-5967.
66. Opera, T. I.; Waller, C. L.; Marshall, G. R. 3D-QSAR of human immunodeficiency virus (I) protease inhibitors. III. Interpretation of CoMFA results. *Drug Des. Discov.* **1994,** *12,* 29-51.
67. Pajeva, I.; Wiese, M. Molecular modeling of phenothiazines and related drugs as multidrug resistance modifiers: A comparative molecular field analysis study. *J. Med. Chem.* **1998,** *41,* 1815-1826.
68. Bursi, R.; Grootenhuis, P. D. Comparative molecular field analysis and energy interaction studies of thrombin-inhibitor complexes. *J. Comput. Aided Mol. Des.* **1999,** *13,* 221-232.
69. Debnath, A. K. Application of 3D-QSAR techniques in anti-HIV-1 drug design - An overview. *Curr. Pharm. Des.* **2005,** *11,* 3091-3110.
70. Fornabaio, M.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 2. Computational titration and pH effects in molecular models of neuraminidase-inhibitor complexes. *J. Med. Chem.* **2003,** *46,* 4487-4500.
71. Fornabaio, M.; Spyrakis, F.; Mozzarelli, A.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. *J. Med. Chem.* **2004,** *47,* 4507-4516.
72. eduSoft LC developers' toolkits. http://www.edusoft-lc.com/toolkits/ (accessed 03/29/2013).
73. Spiegel, S.; Milstien, S. Sphingosine-1-phosphate: An enigmatic signalling lipid. *Nat. Rev. Mol. Cell Biol.* **2003,** *4,* 397-407.
74. Hait, N. C.; Sarkar, S.; Le Stunff, H.; Mikami, A.; Maceyka, M.; Milstien, S.; Spiegel, S. Role of sphingosine kinase 2 in cell migration toward epidermal growth factor. *J. Biol. Chem.* **2005,** *280,* 29462-29469.
75. Johnson, K. R.; Johnson, K. Y.; Crellin, H. G.; Ogretmen, B.; Boylan, A. M.; Harley, R. A.; Obeid, L. M. Immunohistochemical distribution of sphingosine kinase 1 in normal and tumor lung tissue. *J. Histochem. Cytochem.* **2005,** *53,* 1159-1166.
76. Pyne, N. J.; Pyne, S. Sphingosine-1-phosphate and cancer. *Nat. Rev. Cancer* **2010,** *10,* 489-503.
77. Liu, K.; Guo, T. L.; Hait, N. C.; Allegood, J.; Parikh, H. I.; Xu, W.; Kellogg, G. E.; Grant, S.; Spiegel, S.; Zhang, S. Biological characterization of 3-(2-amino-ethyl)-5-[3-(4-butoxyl-phenyl)-propylidene]-thiazolidine-2,4-dione (K145) as a selective Sphingosine Kinase-2 inhibitor and anticancer agent. *PLOS ONE* **2013,** *8,* e56471.
78. Britt, W., Virus entry into host, establishment of infection, spread in host, mechanisms of tissue damage. In *Human herpesviruses: Biology, Therapy, and*

Immunoprophylaxis, 2011/02/25 ed.; Arvin, A.; Campadelli-Fiume, G.; Mocarski, E.; Moore, P. S.; Rioizman, B.; Whitley, R.; Yamanishi, K., Eds. Cambridge University Press: Cambridge, 2007.

79. Kuchta, A. L.; Parikh, H. I.; Zhu, Y.; Kellogg, G. E.; Parris, D. S.; McVoy, M. A. Structural modelling and mutagenesis of human cytomegalovirus alkaline nuclease UL98. *J. Gen. Virol.* **2012,** *93,* 130-138.
80. Amadasi, A.; Spyrakis, F.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Mapping the energetics of water-protein and water-ligand interactions with the "natural" HINT forcefield: Predictive tools for characterizing the roles of water in biomolecules. *J. Mol. Biol.* **2006,** *358,* 289-309.
81. Spyrakis, F.; Cozzini, P.; Bertoli, C.; Marabotti, A.; Kellogg, G. E.; Mozzarelli, A. Energetics of the protein-DNA-water interaction. *BMC Struct. Biol.* **2007,** *7,* 4.
82. Ahmed, M. H.; Spyrakis, F.; Cozzini, P.; Tripathi, P. K.; Mozzarelli, A.; Scarsdale, J. N.; Safo, M. A.; Kellogg, G. E. Bound water at protein-protein interfaces: Partners, roles and hydrophobic bubbles as a conserved motif. *PLOS ONE* **2011,** *6,* e24712.

CHAPTER 2

THIAZOLIDINE-2,4-DIONE (TZD) ANALOGUE K145 – SELECTIVE SPHK2 INHIBITOR

2.1 Introduction

2.1.1 *Sphingolipids – Structure and Function*

Sphingolipids and glycosphingolipids are complex lipids containing the sphingoid backbone, *i.e.*, a long-chain aliphatic C_{18} – C_{20} backbone linked to a fatty acid via its acyl group, and attached to a charged head group such as ethanolamine, serine or choline through an O-linkage. Sphingosine and dihydrosphingosine, which are just long-chain sphingoid bases are the simplest possible functional sphingolipids. Ceramide, a complex sphingolipid, has a fatty acid linked to the base by an amide bond. More complex sphingolipids are formed by addition of head groups to ceramide, such as sphingomyelins and cerebrosides that contain phosphocholine/phosphoethanolamine and glucose/galactose attached to the 1-hydroxy group of ceramide by an ester linkage and *β*-glycosidic linkage, respectively. Figure 2.1 shows the general structures of some sphingolipids.

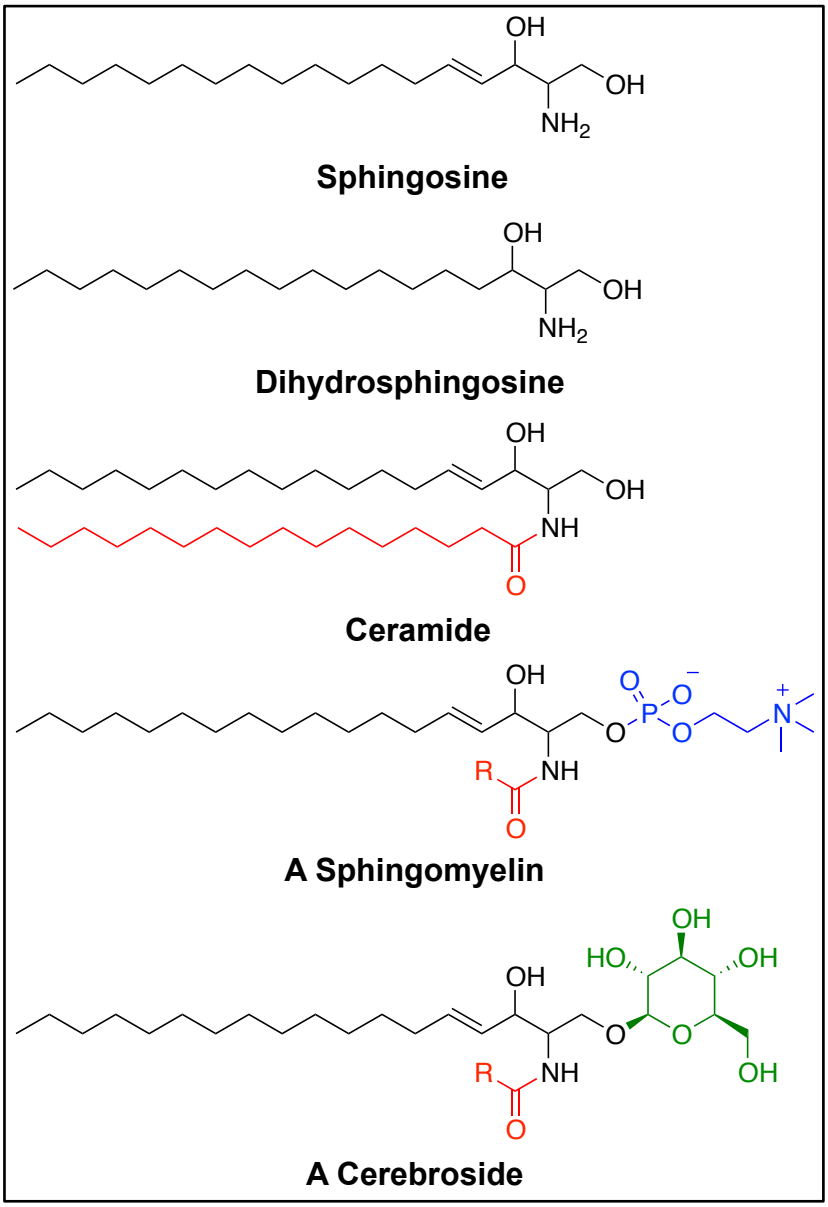


Figure 2.1 – General structures of common sphingolipids

Sphingolipids play very important structural and functional roles in the plasma membranes of eukaryotic cells.[1] Over the last decade, significant studies have also shown their importance in non-structural roles. Sphingolipids have been demonstrated to be an important signaling mediator for vital cellular and physiological processes such as cell motility, invasion, proliferation, angiogenesis and apoptosis.[2-6] Sphingolipids and their metabolites contribute to various cellular signaling pathways either by directly interacting with GPCRs or by acting as intracellular second messengers capable of interacting with a plethora of targets.[7] Plasma membranes of many cell types contain lipid rafts, which are involved in various cellular processes like signal transduction, membrane trafficking, cytoskeletal organization, and, inside the nervous system, implicated in neuronal adhesion, axon guidance and synaptic transmission.[8,9] These lipid rafts are enriched in sphingomyelin, ceramide and glycosphingolipids. Sphingolipids function as secondary messengers by interacting with a number of proteins and are capable of modifying the activity of various receptors, enzymes and ion channels, as well as mobilizing intracellular calcium.[7] Figure 2.2 shows an overview of the roles of sphingolipids in cell biology.

The bioactive sphingolipids – sphingosine, sphingosine-1-phosphate and ceramide – are the central players of sphingolipid-mediated biology. Ceramide and sphingosine have been associated with growth arrest and apoptosis induced by tumor-necrosis factor (TNF) α and Fas ligand.[10,11] In contrast, sphingosine-1-phosphate (S1P) has been demonstrated to play pro-survival roles. S1P induces mitogenesis and acts as a secondary messenger in cellular proliferation induced by platelet-derived growth factor

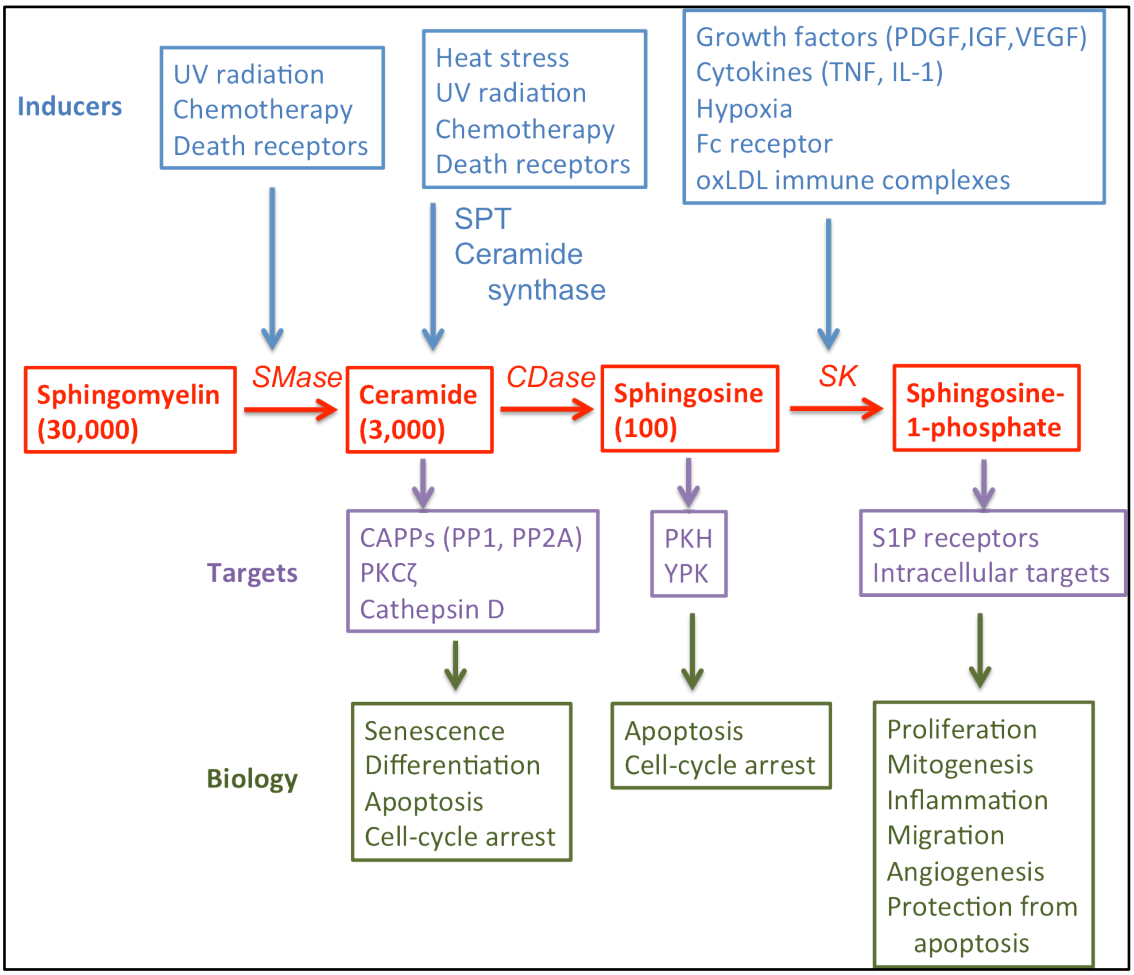


Figure 2.2 – Scheme showing the participation of bioactive sphingolipids – ceramide, sphingosine and sphingosine-1-phosphate (S1P) in cell biology.
Breakdown of various sphingomyelins by sphingomyelinases (SMases) generates ceramide. Ceramide can also be synthesized *de novo* by serine palmitoyl transferase (SPT) and ceramide synthase. Sphingosine and S1P are generated by ceramidases (CDases) and sphingosine kinases (SKs). These sphingosine metabolites interact with specific targets like phosphatases, kinases and GPCRs, which in turn mediate the effects of these lipids. Adapted from ref 11.

and serum.[12-14] The intracellular levels of these metabolites and their regulatory effects on the members of MAPKs, rather than their absolute amounts, determine cell fate.[10] Regulation of the levels of these sphingolipids, a so-called *"sphingolipid rheostat"*, is complex and a number of enzymes have been demonstrated to be important (Figure 2.3).[6,11]

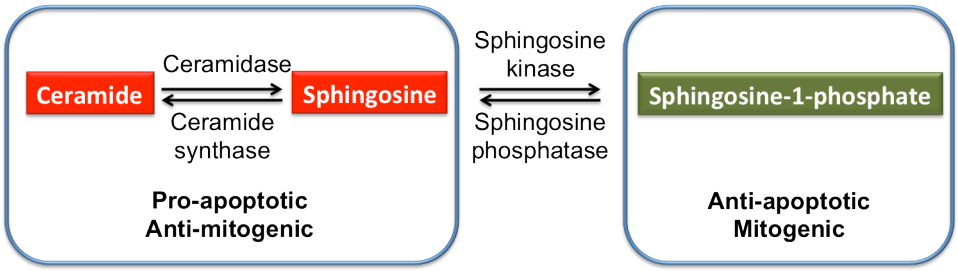


Figure 2.3 – The *Sphingosine Rheostat*: Cell fate determinant.

Studies have suggested the possible role of S1P signaling and the *Sphingosine rheostat* in carcinogenesis, with sphingolipid metabolism often disregulated.[15] Exploiting the opposing effects of these interconvertible metabolites on cell proliferation for therapeutic benefit has emerged as an exciting strategy against cancer.

2.1.2 *Sphingosine Kinases and Cancer*

A key regulator of the sphingosine rheostat is **Sphingosine Kinase (SphK)**, the enzyme that phosphorylates sphingosine to sphingosine-1-phosphate (S1P), thereby producing the pro-growth, anti-apoptotic messenger. On the other hand, SphK decreases the levels of pro-apoptotic sphingosine, and in turn, ceramide. Various studies have shown that a cell is protected against ceramide-induced apoptosis with

increased levels of S1P, whereas depleted levels of S1P enhance ceramide-induced apoptosis.[10,16-18] Human sphingosine kinase exists as two isoforms – Sphingosine kinase 1 (SphK1) and Sphingosine Kinase 2 (SphK2). Although SphK1 and SphK2 share a high degree of homology, they differ significantly in size, tissue distribution and subcellular localization.[19] SphK1 and SphK2 have five conserved domains (C1 – C5) sharing approximately 50% identity, with SphK2 having about 200 more amino acids than SphK1. SphK1 mainly resides in the cytosol while SphK2 is present in different intracellular compartments, including the nucleus, endoplasmic reticulum and mitochondria.[20,21]

Over the past few years, evidence has accumulated that suggest associations of sphingosine kinases with various aspects of cancer development and progression, such as proliferation, migration, invasion and angiogenesis.[22] The very first observation that proposed the possibility of SphK1 as an oncogene was the transformation of SphK1-transfected NIH3T3 fibroblasts to form fibrosarcoma cells, accompanied by increased S1P formation.[23] SphK1 expression has been reported to be upregulated in many different solid tumor types including breast, lung, kidney, stomach, ovary, uterine and colon.[24-27] *Spiegel et al.* showed that enforced expression of SphK1 increased S1P levels and blocked breast adenocarcinoma MCF7 cell death induced by anti-cancer drugs, sphingosine and TNF-α.[28] In another study investigating the significance of SphK1 in gastric cancer progression, it was observed that SphK1 protein levels were upregulated in gastric cancer lesions compared to that in adjacent noncancerous tissues, and patients with higher SphK1 expression have shorter overall survival times.[29] Much less is known about SphK2. Recently, however, it has been shown that

downregulation of SphK2 inhibits the proliferation and migration of tumor cells, such as glioblastoma and breast cancer cells.[30,31] A recent RNA interference study showed that tumor cell proliferation and migration/invasion were suppressed more by SphK2-selective ablation compared to SphK1 ablation.[32]

There is substantial evidence that S1P, an anti-apoptotic, mitogenic sphingolipid metabolite generated by the sphingosine kinases, is involved in cancer. A number of studies have implicated the signaling pathways of S1P in cancer. It regulates processes such as inflammation, neovascularization, cell growth and survival; all of which are important for tumor growth/proliferation and motility. For a detailed review, please refer to *Pyne et al.*[33] Figure 2.4 summarizes the role of S1P in cancer.

As previously mentioned, cell fate is regulated by the sphingosine rheostat and sphingosine kinases play a major role in maintaining the balance in levels of involved sphingolipid metabolites. Strategies that shift the ceramide-sphingosine-S1P rheostat towards the pro-apoptotic/anti-mitogenic ceramide and that inhibit the activity of SphKs below the optimum level for cancer-cell survival are potential avenues for combating cancer. The presented evidence of involvement of SphKs in cancer makes them an ideal target for modulating the sphingolipid-mediated signaling for therapeutic effects. Although a number of potent and selective SphK1 inhibitors have been developed and reported,[6,34-36] only a few SphK2 inhibitors with moderate potency (Figure 2.5), such as ABC294640,[37] SG-12,[38] R-FTY720-OMe[39] and trans-12,[40] have been reported.

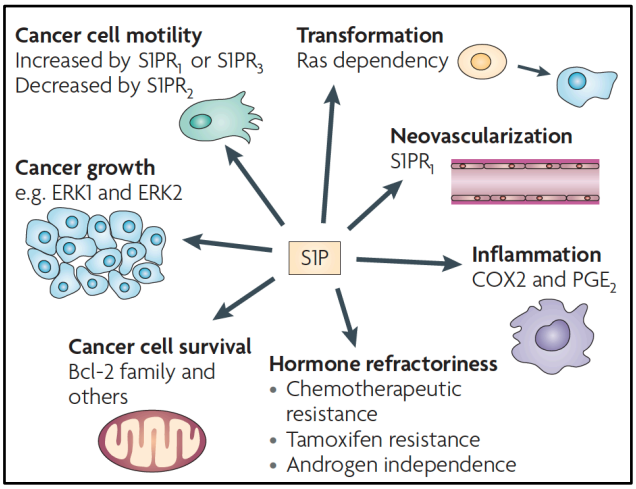


Figure 2.4 – Summary of role of S1P in cancer.
S1P interacts with a family of GPCRs and regulates processes involved in cancer cell motility and proliferation. Reprint from ref. 33

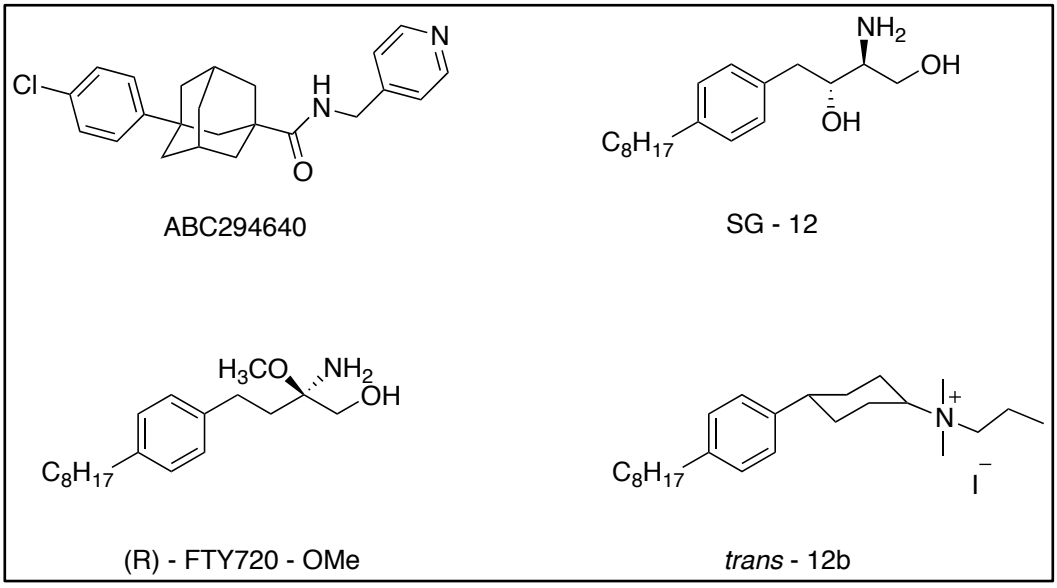
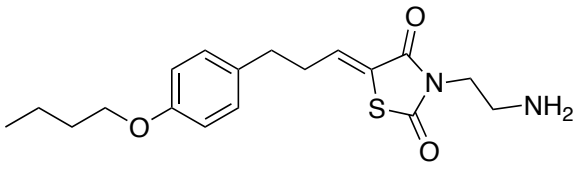


Figure 2.5 - Structures of known SphK2 selective inhibitors.

2.1.3 *Thiazolidine-2,4-dione Analog K145 – Selective SphK2 Inhibitor*[41]



(*Z*)-3-(2-aminoethyl)-5-(3-(4-butoxyphenyl)propylidene)thiazolidine-2,4-dione

K145

Recently, *Dr. Zhang* and his research group initiated the development of thiazolidine-2,4-dione (TZD) analogs as dual-pathway inhibitors of the ERK and Akt signaling pathways.[42,43] The TZD scaffold has emerged as a privileged template in drug discovery and design because of its frequent appearance in hits in various potential anticancer agents.[44,45] The 3-(2-aminoethyl)-TZD moiety of these inhibitors may be able to mimic the amino-hydroxyl sphingoid base suggesting the possibility of them being sphingosine kinase inhibitors (Figure 2.6). Also, it has been shown that an aromatic ring with an alkyl chain is an important structural feature of SphK inhbitors.[35] All these observations led to the hypothesis that K145, a TZD analog, could be a SphK inhibitor.[41]

Following the synthesis of K145, biochemical assays as well as *in vitro* and *in vivo* studies were performed in the labs of *Dr. Zhang*, *Dr. Spiegel* and *Dr. Grant* (Virginia Commomwealth University) to determine its inhibitory activity towards SphKs and its nature of inhibition, and to examine its apoptotic effects on human leukaemia U937 cells and demonstrate its in vivo efficacy as a potential lead anticancer agent.

Notably, K145 inhibited the activity of SphK2 in a dose-dependent manner with an IC_{50} of 4.40 ± 0.05 μM, while no inhibition of SphK1 at concentrations up to 10 μM was observed (Figure 2.7A). In contrast, DMS (10 μM) a non-selective SphK inhibitor,

showed inhibition of both SphK1 and SphK2. This indicated that K145 is a selective SphK2 inhibitor. Lineweaver-Burk analysis revealed a K_i of 6.4 ± 0.7 μM for SphK2 and indicated that K145 is a substrate (sphingosine) competitive inhibitor (Figure 2.7B). Further screening against ceramide kinase, Akt kinase, ERK1/2, PI3K, PKA and other kinases also demonstrated the relatively high selectivity for SphK2.[41]

Biological characterization using human leukemia U937 cells demonstrated that K145 accumulated in U937 cells and inhibited the phosphorylation of FTY720, and also inhibited the growth of U937 cells, mainly through apoptotic effects. Furthermore, K145 was shown to significantly suppress the growth of U937 tumors in nude mice and inhibited growth of JC tumor cells in BALB/c mice without apparent toxicity.[41]

These results strongly indicate that K145 is a selective SphK2 inhibitor and encourage further optimization of K145 as a novel lead compound to develop more potent and selective SphK2 inhibitors. This would be of great value as a pharmacological tool to complement the ongoing molecular and genetic studies, and help unravel the roles of SphK2 in different pathological and physiological conditions. It would be useful to have new and adaptable chemical scaffolds available as selective SphK2 inhibitors that can provide valuable information regarding structural requirements for designing new SphK2 inhibitors.

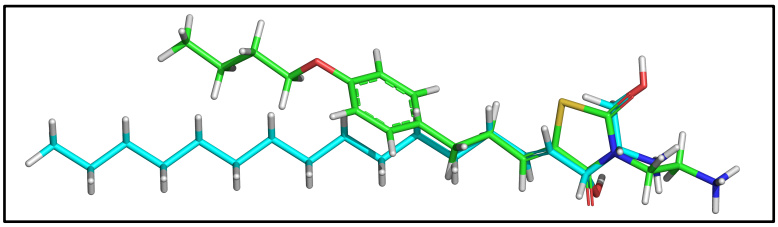


Figure 2.6 – Overlay of K145 (carbon: green) with Sphingosine (carbon: cyan) showing chemical similarity between the heterocycle of K145 and the amino-hydroxyl sphingoid head of sphingosine.

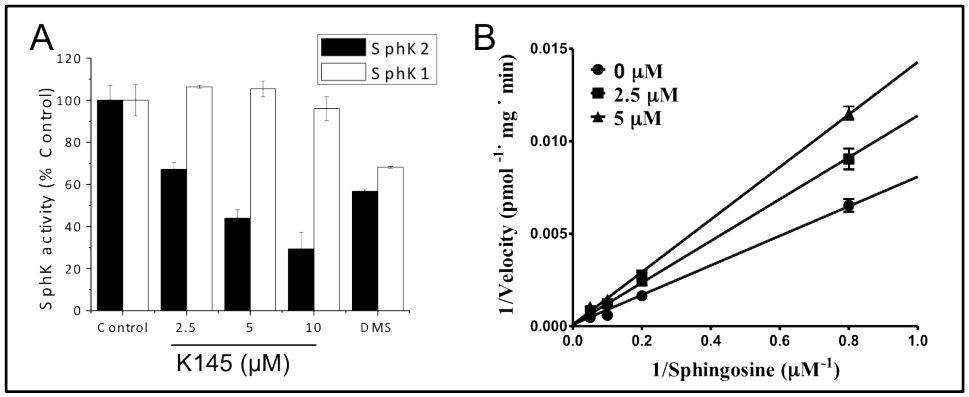


Figure 2.7 – Biochemical assays showing K145 as a selective, substrate-competitive SphK2 inhibitor.
(A) SphK1 and SphK2 activities were measured with 10 μM sphingosine in the absence or presence of the indicated concentrations of K145 or 10 μM DMS. Data expressed as percentage SphK activity in the absence of inhibitor. (B) SphK2 activity was measured with increasing concentrations of sphingosine and indicated concentrations of K145. Lineweaver-Burk analysis revealed a Vmax of 10820 ± 210 pmol/min per mg of protein, and a K_i of 6.4 ± 0.7 μM for SphK2.

2.1.4 *Sphingosine Kinase C4 domain – A putative Sphingosine-binding domain*

As mentioned before, both sphingosine kinase isoforms have five conserved domains C1 – C5, sharing about 50% identity (Figure 2.8). The domains, C1 – C3 and C5, share homology with other kinases like diacylglycerol kinase (DGK) and ceramide kinase (CerK), with the C2 domain containing the ATP-binding consensus sequence $SGDGX_{17\text{-}21}K$.[46,47]

The C4 domain is highly conserved only in SphKs making it the only domain that might be specific for sphingosine binding. *Yokota et al.* constructed various mutants of mouse SphK1, within the C4 domain, in order to identify the residues important for sphingosine recognition. The negatively charged Asp175, Asp177, Glu179 and Glu181 residues, in the C4 domain of mSphK1, were mutated to Asn and Gln. Also, double mutants mSphK1a$^{D175N/D177N}$ and mSphK1a$^{D177N/E179Q}$ were prepared. Each mutant was analyzed for SphK activity using D-erythro-sphingosine and ATP as substrates (Figure 2.9). The results demonstrated that the highly conserved Asp177 is involved in sphingosine recognition.[48]

In order to identify the putative Asp residue involved in sphingosine-recognition in human SphKs, we performed a multiple sequence alignment between mouse SphK1a and human SphK1 and SphK2. Asp178 in hSphK1 and Asp344 in hSphK2 were identified as the residues corresponding to the Asp177 of mouse SphK1 in the C4 domain (Figure 2.10).

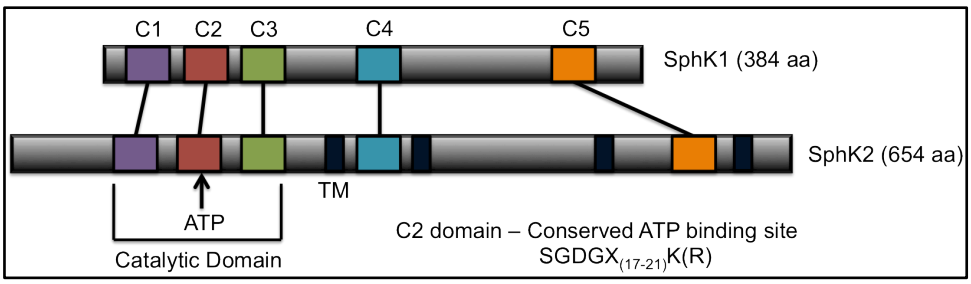


Figure 2.8 – Schematic representation of human sphingosine kinase isoforms

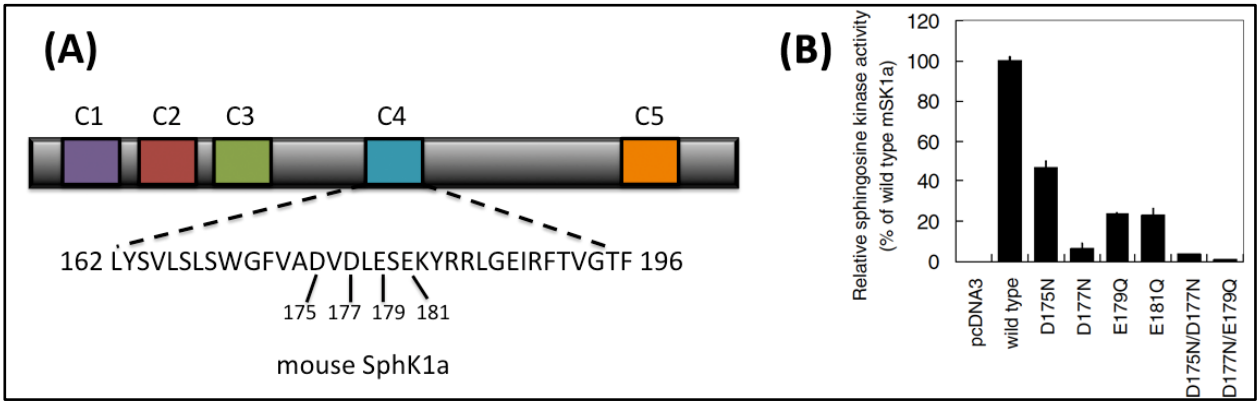


Figure 2.9 – Relative sphingosine kinase activities of mouse SphK1a mutants.
(A) Schematic representation of mSphK1 showing the C4 domain and the residues that were mutated. (B) Relative SphK activities of mSphK1a mutants. The mutants were analyzed for SphK activity using D-erythro-sphingosine and ATP as substrates.[48]

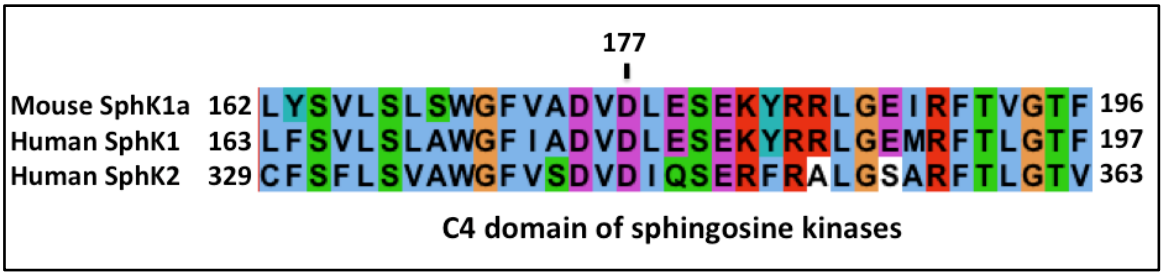


Figure 2.10 – Sequence alignment of the C4 domain of sphingosine kinases. Asp178 (hSphK1) and Asp344 (hSphK2) were identified as the corresponding residues important for sphingosine-recognition. Sequence alignment was performed using Clustal X v2.0,[49] image obtained using Jalview,[50,51] sequences are colored using the default Clustal X color scheme (Appendix A.1).

2.1.5 *Specific Aim*

To further understand the SphK2 selectivity of K145, we conducted molecular modeling studies to identify the structural features of K145 that interact with the key residues of SphK2. Since the biochemical assays showed that K145 is a substrate (Sphingosine) competitive inhibitor of SphK2 (Section 2.1.3), we performed our studies on the C4 domain of both isoforms. In this chapter, we report the results of protein structure building followed by the docking studies of various reported Sphingosine kinase inhibitors. Finally, a binding mode for K145 within the putative sphingosine-binding domain of SphK1 and SphK2 has been proposed.

2.2 Methods

2.2.1 *Structural Modeling of SphK1 and SphK2*

2.2.1.1 Template Identification and Alignment

Human SphK1 (Accession: Q9NYA1) and SphK2 (Accession: NP_001191088) sequences were obtained from the NCBI database (www.ncbi.nih.gov/protein/). A Position Specific Iterated BLAST[52,53] search against the database of Protein Data Bank proteins was performed to identify a template structure. Diacylgycerol kinase from *Bacillus anthracis* str. Sterne (PDB ID: 3T5P) was identified as the closest match to both isoforms of SphK. Sequence alignments of each SphK1 and SphK2 with 3T5P were performed using Clustal X v2.0.[49] Unaligned regions in both the proteins were deleted.

2.2.1.2 Homology Modeling and Refinement

A total of 100 homology models for each isoform were generated based on these alignments, using the *automodel* class of MODELLER 9v10.[54] A DOPE (Discrete Optimized Protein Energy) Score[55] and a GA341 Score[56] was calculated for each model using MODELLER. The top 5 models for each kinase with lowest DOPE scores and molpdf scores (a MODELLER object function score) and with GA341 scores closest to 1 were chosen for further refinement. The side chains for each model were optimized using SCWRL[57] (dunbrack.fcc.edu/scwrl4/). Hydrogens were added to these top models using SYBYL v8.1 (TRIPOS Inc.) and subsequently subjected to Powell minimization for 10000 iterations in Tripos force field with a 0.005 kcal/mol-Å termination gradient. The quality of minimized models was evaluated using MolProbity,[58] which performs an all-atom contact analysis to give a ‘clashscore’ that is indicative of the number of serious

clashes (>0.4 Å) per 1000 atoms. Poor side-chain rotamers and unreasonable bond lengths and angles were checked. Ramachandran plots were also generated using MolProbity to check the backbone-geometry of the models. Atom clashes and bad bond lengths and angles were optimized with further minimization. Sphingosine, the natural substrate for both kinases, was docked into the C4 domain (putative Sph binding domain, L163 – F197 for SphK1 and C329 – V363 for SphK2) of each model, using GOLD v5.1.[59] The docked poses were scored using HINT.[60] The "best" model of both SphK1 and SphK2 was then chosen based on its overall stability and its ability to accommodate Sphingosine in its C4 domain.

2.2.2 *Inhibitor Docking*

The optimized models of both SphK1 and SphK2 were used for the docking studies. The structures of inhibitors were sketched using SYBYL v8.1, and subjected to minimization to get a low energy structure. The docking simulations were performed using GOLD v5.1. The binding site was defined to encompass all atoms within 20 Å of CA of Asp178 of SphK1 (Asp344 of SphK2). Fifty solutions for each inhibitor molecule were generated with a protein hydrogen-bond constraint that the carboxylate of Asp178 of SphK1 (Asp344 of SphK2) forms a hydrogen bond with ligand, since the Asp is important for recognition of sphingosine (Section 2.1.4). The docked poses were scored using HINT. The poses with the best HINT scores were complexed with the protein and the protein-ligand complex was subjected to minimization (2500 iterations, termination gradient of 0.005 kcal/mol-Å), to remove steric clashes and get an induced-fit model. The binding modes of the ligands after minimization were re-scored using HINT.

2.3 Results and Discussion

2.3.1 *Structural Modeling of SphK1 and SphK2*

Since no crystal structure is currently available for either SphK1 or SphK2, we generated their structural models using MODELLER, a comparative protein structure-modeling program. First, a PSI-BLAST search was performed against the database of PDB proteins that identified a structure of Diacylglycerol kinase from *Bascillus anthracis* str. Sterne (PDB ID: 3t5p) as the template (Figure 2.11). The template structure shares ~25% sequence identity and ~46% homology in the aligned regions to both SphK1 and SphK2. The primary amino acid sequences were aligned using Clustal X v2.0, the results of which are shown in Figure 2.12 and Figure 2.13.

Based on the sequence alignment, structural models of both isoforms of SphKs were generated using MODELLER. The program sets up spatial restraints on Cα-Cα distances, main chain N-O distances and main chain and side chain dihedral angles obtained from the template crystal structure,[54] which is followed by modeling all non-hydrogen atoms by violating these restraints as little as possible. The natural substrate for SphKs, sphingosine, was docked into the C4 domain of top 5 models of each kinase (based on lowest energy scores, refer Section 2.2.1.2) using GOLD (Genetic Optimization for Ligand Docking). GOLD performs automated ligand docking, using genetic algorithm to explore full ligand flexibility within the neighborhood of protein binding site.[59] The generated binding poses are ranked by a simple scoring function that comprises of a hydrogen bonding term, a pairwise dispersion potential and a term for

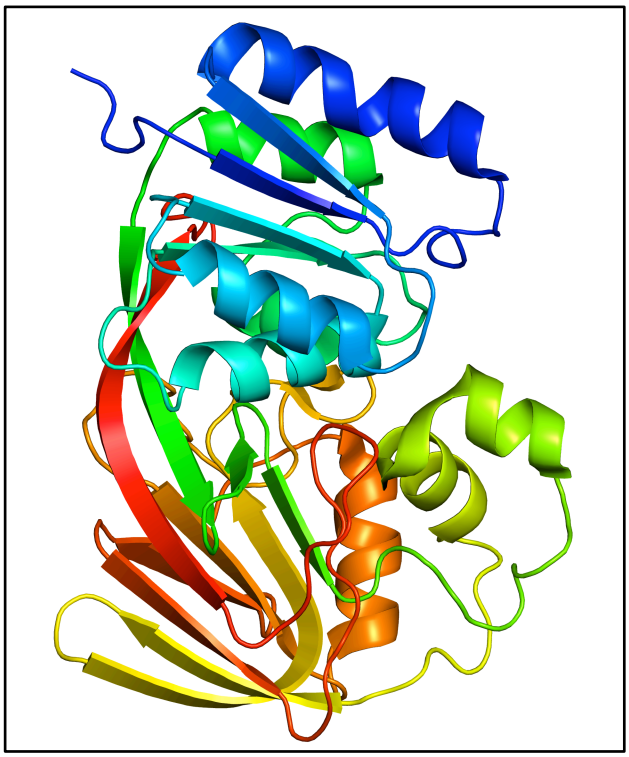


Figure 2.11 – Overall fold of the template structure – Diacylglycerol kinase from *Bascillus anthracis* str. Sterne (PDB ID: 3t5p).
Image prepared using PyMOL.[61]

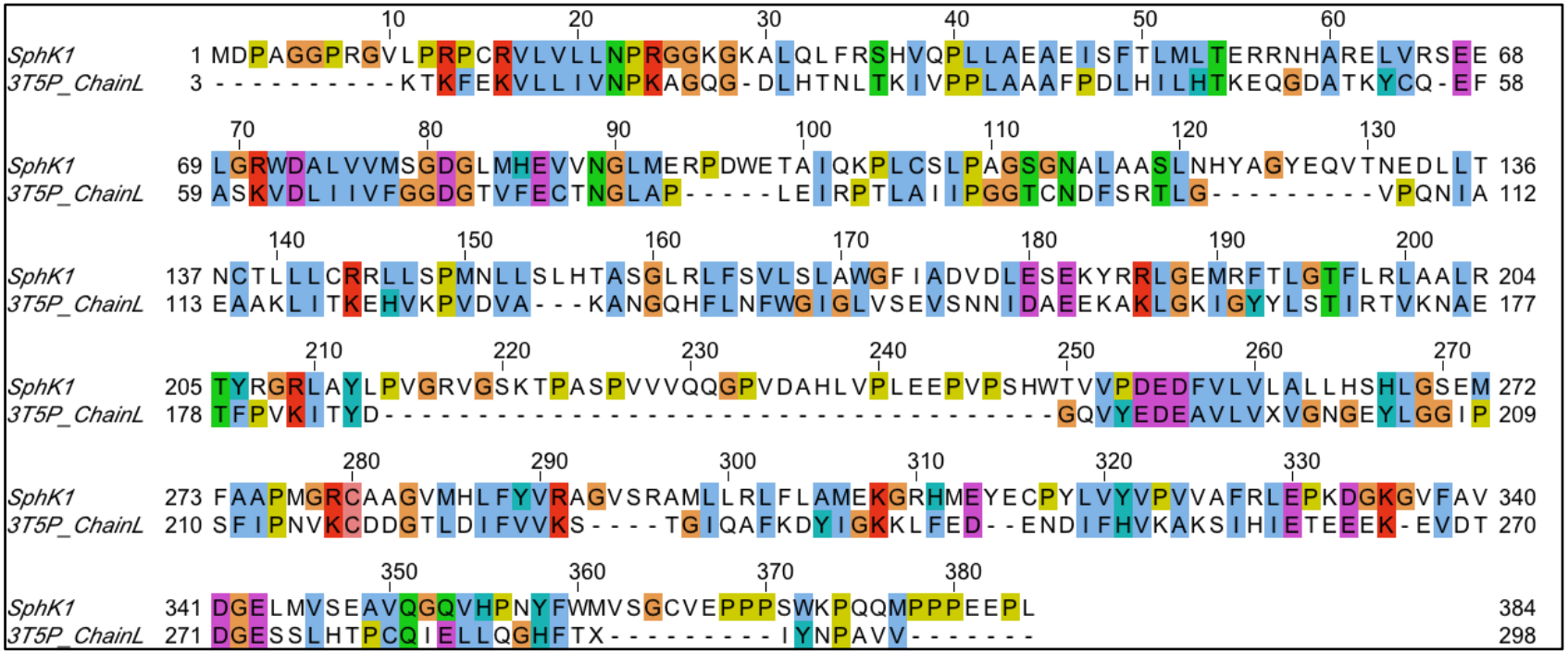


Figure 2.12 – Sequence alignment of SphK1 with the template 3T5P (Diacylglycerol Kinase from *Bacillus anthracis* str. Sterne).
Alignment performed using Clustal X v2.0. The alignment is colored based on the Clustal X color scheme (Appendix A.1). Image prepared using Jalview.

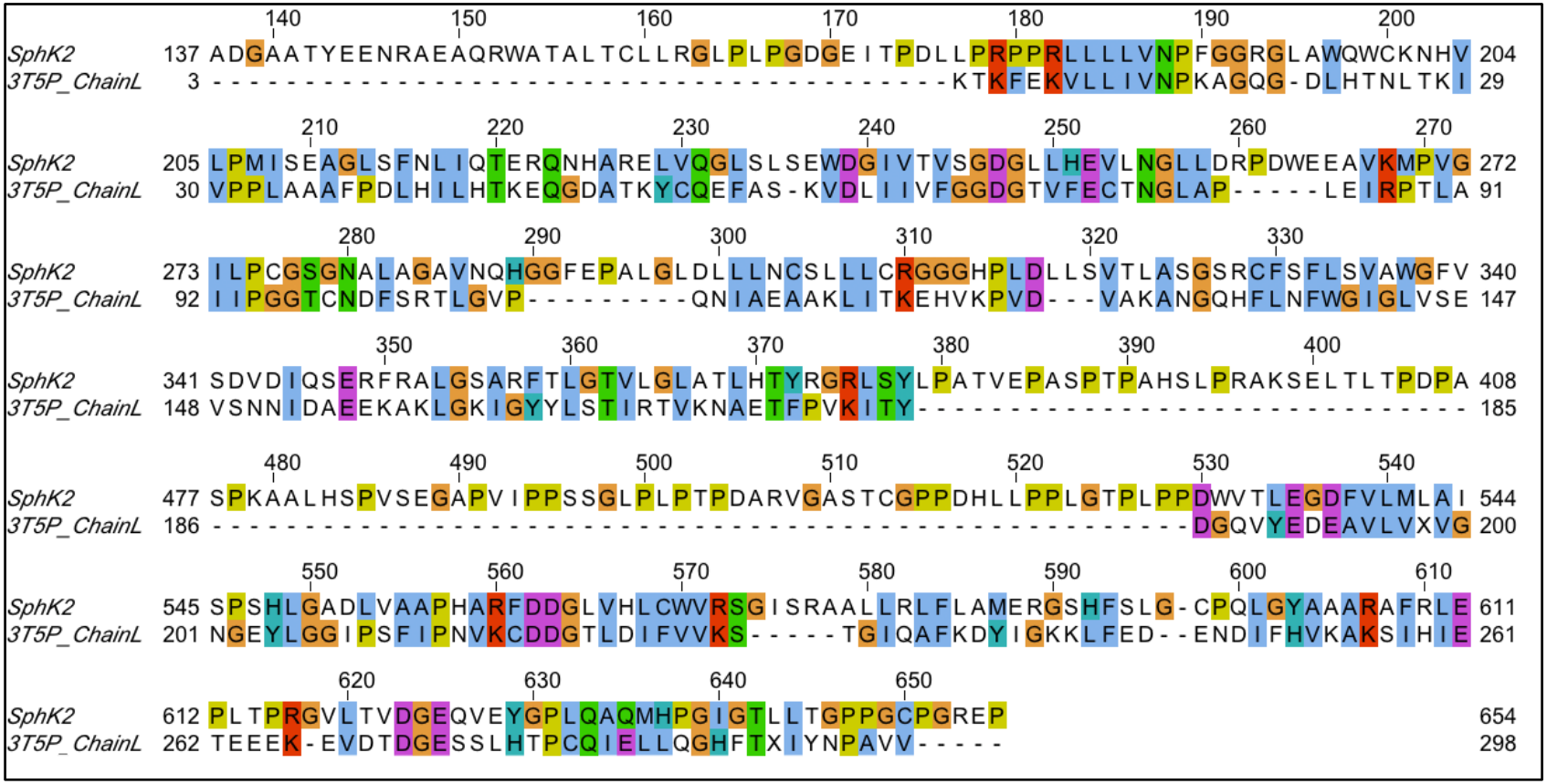


Figure 2.13 – Sequence alignment of SphK2 with the template 3T5P (Diacylglycerol Kinase from *Bacillus anthracis* str. Sterne).
Alignment performed using Clustal X v2.0. The alignment is colored based on the Clustal X color scheme (Appendix A.1). Image prepared using Jalview.

internal energy of the ligand conformation.[59] For our study, all the docked poses were scored using the HINT force field, as it accounts for all the hydropathic forces involved in the protein–ligand binding event. The "best" model of both SphK1 and SphK2 was then chosen based on its overall stability and its ability to accommodate Sphingosine in its C4 domain. Model044 was the best model for SphK1, with a clashscore of 3.2 (97th percentile) and with 96.8% residues in allowed regions on Ramachandran plots. Model055 was the best model for SphK2, with a clashscore of 2.03 (99th percentile) and with 98.5% residues in allowed regions on Ramachandran plots. Figure 2.14 shows the sphingosine-binding site of both SphK1 and SphK2.

2.3.2 *Model Validation by Inhibitor Docking*

The final optimized models for both proteins were further validated by docking a panel of inhibitors including the reported SphK2 selective compounds shown in Figure 2.5, a SphK1 selective inhibitor, SK1-I,[35] and FTY720, a compound known to bind to both SphKs.[62,63] Docking and scoring studies were performed as described in Section 2.2.2. As shown in Table 2.1, the HINT score results indicate that both FTY720 and its o-methoxy derivative (R)-FTY720-OMe bind to SphK2 preferably, over SphK1. SK1-I binds more favorably to SphK1, consistent with the reported biological results.[35] We also docked SG-12, ABC294640 and trans-12b (Figure 2.5); the relative ordering of HINT scores (H_{TOTAL}) are more or less in concordance with the reported binding/inhibitory observations (Table 2.1).[37-40] While specific HINT score values generally should be calibrated for the specific biomacromolecular-ligand system,

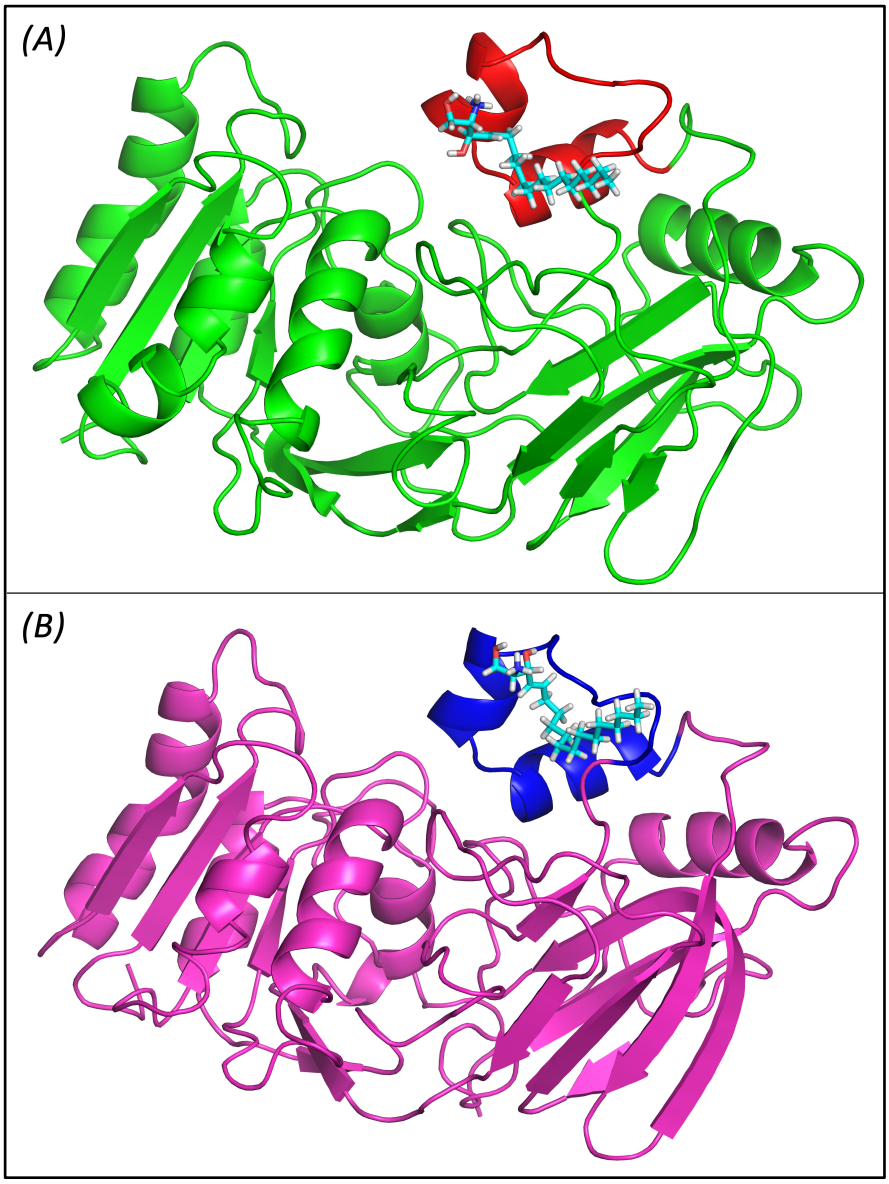


Figure 2.14 – Structural models of SphK1 (A) and SphK2 (B).
Both kinases are shown in cartoon representation, with their C4 domains in different color. The docked sphingosine molecule is shown as sticks. Image prepared using PYMOL.

previous studies consistently indicate that Δ(HINT score)/ $\Delta\Delta G_{binding}$ ~ –515, *i.e.* HINT score differences of around 515 correspond to 1 kcal/mol differences in binding energies.[64,65]

Despite low overall homology to the template, there is considerable sequence and structural similarity at the sphingosine-binding domain (C4 domain), and we believe that these models will provide valuable structural information.

Table 2.1 – HINT Scores for docked molecules (previously reported inhibitors) into the C4 domain of SphK1 and SphK2

Ligand	HINT Scores[a]	
	SphK1	SphK2
FTY720	2347	2751
(R)-FTY720-OMe	138	1878
SG-12	1626	1876
ABC294640	-73	153
Trans-12b	-935	218
SK1-I	2080	679

[a]Previous studies have shown that ~515 score units correspond to ΔΔG = -1.0 kcal/mol. In the absence of a reference point from a calibration for this specific biomolecular system, the HINT score *difference* between ligands and/or between SphK1 and SphK2 are more meaningful than their specific values.

2.3.3 *Proposed Binding Mode for K145*

We then docked K145 to the two kinase models. The docking results revealed that K145 binds preferentially to SphK2 (Table 2.2), as it shows more favorable interactions in the sphingosine-binding pocket of SphK2 than that of SphK1. Specifically, as shown in Figure 2.15, our model indicates that the terminal $–NH_2$ of K145 forms strong salt-bridge interactions with the carboxylate group of Asp344 (the putative sphingosine recognizing residue). Other favorable hydrogen bonding interactions are also formed between the guanidino-group of Arg351 and Gln346 with the carbonyl oxygens of the TZD heterocycle. The TZD ring of K145 shows favorable *π*-stacking interactions with Phe350 and the 4-butoxy-phenyl ring of K145 fits into a hydrophobic pocket formed by the sidechains of Ala336, Val340, Val343, Arg617 and Val619. K145 shows a very similar binding mode within the C4 domian of SphK1. The terminal $–NH_2$ of K145 forms salt-bridge interaction with Asp178 and the carbonyl oxygen at 2-position of TZD rings forms hydrogen-bonding interaction with Arg185. The TZD ring *π*-stacks with the aromatic ring of Tyr184 and the hydrophobic interactions of the tail with the surrounding hydrophobic residues are more or less conserved. In contrast to its binding mode in SphK2, the carbonyl oxygen at 4-position of the TZD ring showed an unfavorable base/base interaction with the carboxylate group of Glu180 (that corresponds to Gln346 of SphK2). Judging from the sequence similarity in the sphingosine-binding domains of both isoforms, the Gln→Glu change in SphK1 is the only significant difference in this region and might be the reason for K145 showing selectivity towards SphK2.

Table 2.2 – HINT Scores for K145 docked into the C4 domain of SphK1 and SphK2

Ligand	HINT Score	
	SphK1	SphK2
K145	1506	3011

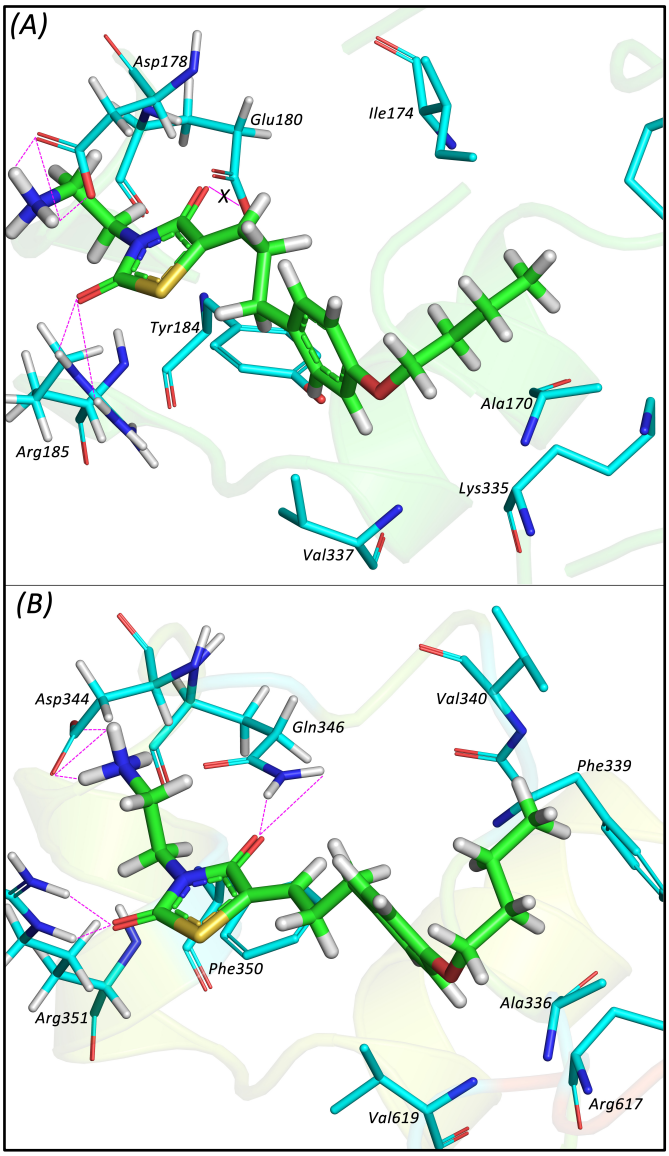


Figure 2.15 – Binding mode of K145 in SphK1 (A) and SphK2 (B).
K145 is shown as sticks with carbon in green, while the interacting residues of both kinases are shown as sticks with carbon in cyan. For simplicity, hydrogens are only shown on residues forming hydrogen-bonding interactions with K145. Image prepared using PyMOL.

2.4 Conclusion

Dr. Zhang and his group identified a thiazolidine-2,4-dione analog, K145, as a selective SphK2 inhibitor. Biochemical assays using recombinant SphK1 and SphK2 established that K145 selectively inhibited SphK2 and not SphK1.[41] In absence of any crystallographic data available for either isoform, we performed *in-silico* studies to gain structural insights into protein-ligand interaction. We successfully generated structural models of both sphingosine kinase isoforms, SphK1 and SphK2, using a known structure of diacylglycerol kinase from *Bacillus anthracis* str. Sterne as the template structure. K145 was then docked into the C4 domain of each model, and the protein-ligand intermolecular interactions were elucidated using HINT scoring.

Although these are tentative models built from a kinase of bacterial origin, their sphingosine-binding domains are more conserved than other regions, and the models are more than adequate as hypothesis generators for compound design. Nonetheless, our docking results do support the experimental assertion that K145 is a selective SphK2 inhibitor.

References

1. Guschina, I. A.; Harwood, J. L., *Lipids: Chemical Diversity*. J. Wiley and Sons.: NY, 2008.
2. Spiegel, S.; Cuvillier, O.; Edsall, L. C.; Kohama, T.; Menzeleev, R.; Olah, Z.; Olivera, A.; Pirianov, G.; Thomas, D. M.; Tu, Z.; Van Brocklyn, J. R.; Wang, F. Sphingosine-1-phosphate in cell growth and cell death. *Ann. N. Y. Acad. Sci.* **1998,** *845,* 11-18.
3. Cuvillier, O. Sphingosine in apoptosis signaling. *Biochim. Biophys. Acta.* **2002,** *1585,* 153-162.
4. Hannun, Y. A.; Obeid, L. M. The Ceramide-centric universe of lipid-mediated cell regulation: Stress encounters of the lipid kind. *J. Biol. Chem.* **2002,** *277,* 25847-25850.
5. Spiegel, S.; Milstien, S. Sphingosine-1-phosphate: An enigmatic signalling lipid. *Nat. Rev. Mol. Cell. Biol.* **2003,** *4,* 397-407.
6. Pitson, S. M. Regulation of sphingosine kinase and sphingolipid signaling. *Trends Biochem. Sci.* **2011,** *36,* 97-107.
7. Colombaioni, L.; Garcia-Gil, M. Sphingolipid metabolites in neural signalling and function. *Brain Res. Rev.* **2004,** *46,* 328-355.
8. Brown, D. A.; London, E. Functions of lipid rafts in biological membranes. *Annu. Rev. Cell Dev. Biol.* **1998,** *14,* 111-136.
9. Paratcha, G.; Ibanez, C. F. Lipid rafts and the control of neurotrophic factor signaling in the nervous system: Variations on a theme. *Curr. Opin. Neurobiol.* **2002,** *12,* 542-549.
10. Cuvillier, O.; Pirianov, G.; Kleuser, B.; Vanek, P. G.; Coso, O. A.; Gutkind, S.; Spiegel, S. Suppression of ceramide-mediated programmed cell death by sphingosine-1-phosphate. *Nature* **1996,** *381,* 800-803.
11. Hannun, Y. A.; Obeid, L. M. Principles of bioactive lipid signalling: Lessons from sphingolipids. *Nat. Rev. Mol. Cell Biol.* **2008,** *9,* 139-150.
12. Olivera, A.; Spiegel, S. Sphingosine-1-phosphate as second messenger in cell proliferation induced by PDGF and FCS mitogens. *Nature* **1993,** *365,* 557-560.
13. Coroneos, E.; Martinez, M.; McKenna, S.; Kester, M. Differential regulation of sphingomyelinase and ceramidase activities by growth factors and cytokines. Implications for cellular proliferation and differentiation. *J. Biol. Chem.* **1995,** *270,* 23305-23309.
14. Spiegel, S.; Milstien, S. Sphingolipid metabolites: Members of a new class of lipid second messengers. *J. Membr. Biol.* **1995,** *146,* 225-237.
15. Loh, K. C.; Baldwin, D.; Saba, J. D. Sphingolipid signaling and hematopoietic malignancies: to the rheostat and beyond. *Anticancer Agents Med. Chem.* **2011,** *11,* 782-793.
16. Edsall, L. C.; Pirianov, G. G.; Spiegel, S. Involvement of sphingosine 1-phosphate in nerve growth factor-mediated neuronal survival and differentiation. *J. Neurosci.* **1997,** *17,* 6952-6960.
17. Cuvillier, O.; Rosenthal, D. S.; Smulson, M. E.; Spiegel, S. Sphingosine 1-phosphate inhibits activation of caspases that cleave poly(ADP-ribose)

polymerase and lamins during Fas- and ceramide-mediated apoptosis in Jurkat T lymphocytes. *J. Biol. Chem.* **1998,** *273,* 2910-2916.
18. Xia, P.; Wang, L.; Gamble, J. R.; Vadas, M. A. Activation of sphingosine kinase by tumor necrosis factor-alpha inhibits apoptosis in human endothelial cells. *J. Biol. Chem.* **1999,** *274,* 34499-34505.
19. Siow, D.; Wattenberg, B. The compartmentalization and translocation of the sphingosine kinases: mechanisms and functions in cell signaling and sphingolipid metabolism. *Crit. Rev. Biochem. Mol. Biol.* **2011,** *46,* 365-375.
20. Johnson, K. R.; Becker, K. P.; Facchinetti, M. M.; Hannun, Y. A.; Obeid, L. M. PKC-dependent activation of sphingosine kinase 1 and translocation to the plasma membrane. Extracellular release of sphingosine-1-phosphate induced by phorbol 12-myristate 13-acetate (PMA). *J. Biol. Chem.* **2002,** *277,* 35257-35262.
21. Igarashi, N.; Okada, T.; Hayashi, S.; Fujita, T.; Jahangeer, S.; Nakamura, S. Sphingosine kinase 2 is a nuclear protein and inhibits DNA synthesis. *J. Biol. Chem.* **2003,** *278,* 46832-46839.
22. Gault, C. R.; Obeid, L. M. Still benched on its way to the bedside: sphingosine kinase 1 as an emerging target in cancer chemotherapy. *Crit. Rev. Biochem. Mol. Biol.* **2011,** *46,* 342-351.
23. Xia, P.; Gamble, J. R.; Wang, L.; Pitson, S. M.; Moretti, P. A.; Wattenberg, B. W.; D'Andrea, R. J.; Vadas, M. A. An oncogenic role of sphingosine kinase. *Curr. Biol.* **2000,** *10,* 1527-1530.
24. French, K. J.; Schrecengost, R. S.; Lee, B. D.; Zhuang, Y.; Smith, S. N.; Eberly, J. L.; Yun, J. K.; Smith, C. D. Discovery and evaluation of inhibitors of human sphingosine kinase. *Cancer Res.* **2003,** *63,* 5962-5969.
25. Johnson, K. R.; Johnson, K. Y.; Crellin, H. G.; Ogretmen, B.; Boylan, A. M.; Harley, R. A.; Obeid, L. M. Immunohistochemical distribution of sphingosine kinase 1 in normal and tumor lung tissue. *J. Histochem. Cytochem.* **2005,** *53,* 1159-1166.
26. Kawamori, T.; Osta, W.; Johnson, K. R.; Pettus, B. J.; Bielawski, J.; Tanaka, T.; Wargovich, M. J.; Reddy, B. S.; Hannun, Y. A.; Obeid, L. M.; Zhou, D. Sphingosine kinase 1 is up-regulated in colon carcinogenesis. *FASEB J.* **2006,** *20,* 386-388.
27. Antoon, J. W.; Beckman, B. S. Sphingosine kinase: A promising cancer therapeutic target. *Cancer Biol. Ther.* **2011,** *11,* 647-650.
28. Nava, V. E.; Hobson, J. P.; Murthy, S.; Milstien, S.; Spiegel, S. Sphingosine kinase type 1 promotes estrogen-dependent tumorigenesis of breast cancer MCF-7 cells. *Exp. Cell Res.* **2002,** *281,* 115-127.
29. Li, W.; Yu, C. P.; Xia, J. T.; Zhang, L.; Weng, G. X.; Zheng, H. Q.; Kong, Q. L.; Hu, L. J.; Zeng, M. S.; Zeng, Y. X.; Li, M.; Li, J.; Song, L. B. Sphingosine kinase 1 is associated with gastric cancer progression and poor survival of patients. *Clin. Cancer Res.* **2009,** *15,* 1393-1399.
30. Hait, N. C.; Sarkar, S.; Le Stunff, H.; Mikami, A.; Maceyka, M.; Milstien, S.; Spiegel, S. Role of sphingosine kinase 2 in cell migration toward epidermal growth factor. *J. Biol. Chem.* **2005,** *280,* 29462-29469.

31. Van Brocklyn, J. R.; Jackson, C. A.; Pearl, D. K.; Kotur, M. S.; Snyder, P. J.; Prior, T. W. Sphingosine kinase-1 expression correlates with poor survival of patients with glioblastoma multiforme: Roles of sphingosine kinase isoforms in growth of glioblastoma cell lines. *J. Neuropathol. Exp. Neurol.* **2005,** *64,* 695-705.
32. Gao, P.; Smith, C. D. Ablation of sphingosine kinase-2 inhibits tumor cell proliferation and migration. *Mol. Cancer Res.* **2011,** *9,* 1509-1519.
33. Pyne, N. J.; Pyne, S. Sphingosine 1-phosphate and cancer. *Nat. Rev. Cancer* **2010,** *10,* 489-503.
34. French, K. J.; Upson, J. J.; Keller, S. N.; Zhuang, Y.; Yun, J. K.; Smith, C. D. Antitumor activity of sphingosine kinase inhibitors. *J. Pharmacol. Exp. Ther.* **2006,** *318,* 596-603.
35. Paugh, S. W.; Paugh, B. S.; Rahmani, M.; Kapitonov, D.; Almenara, J. A.; Kordula, T.; Milstien, S.; Adams, J. K.; Zipkin, R. E.; Grant, S.; Spiegel, S. A selective sphingosine kinase 1 inhibitor integrates multiple molecular therapeutic targets in human leukemia. *Blood* **2008,** *112,* 1382-1391.
36. Kennedy, A. J.; Mathews, T. P.; Kharel, Y.; Field, S. D.; Moyer, M. L.; East, J. E.; Houck, J. D.; Lynch, K. R.; Macdonald, T. L. Development of amidine-based sphingosine kinase 1 nanomolar inhibitors and reduction of sphingosine 1-phosphate in human leukemia cells. *J. Med. Chem.* **2011,** *54,* 3524-3548.
37. French, K. J.; Zhuang, Y.; Maines, L. W.; Gao, P.; Wang, W.; Beljanski, V.; Upson, J. J.; Green, C. L.; Keller, S. N.; Smith, C. D. Pharmacology and antitumor activity of ABC294640, a selective inhibitor of sphingosine kinase-2. *J. Pharmacol. Exp. Ther.* **2010,** *333,* 129-139.
38. Kim, J. W.; Kim, Y. W.; Inagaki, Y.; Hwang, Y. A.; Mitsutake, S.; Ryu, Y. W.; Lee, W. K.; Ha, H. J.; Park, C. S.; Igarashi, Y. Synthesis and evaluation of sphingoid analogs as inhibitors of sphingosine kinases. *Bioorg. Med. Chem.* **2005,** *13,* 3475-3485.
39. Lim, K. G.; Sun, C.; Bittman, R.; Pyne, N. J.; Pyne, S. (R)-FTY720 methyl ether is a specific sphingosine kinase 2 inhibitor: Effect on sphingosine kinase 2 expression in HEK 293 cells and actin rearrangement and survival of MCF-7 breast cancer cells. *Cell. Signal.* **2011,** *23,* 1590-1595.
40. Raje, M. R.; Knott, K.; Kharel, Y.; Bissel, P.; Lynch, K. R.; Santos, W. L. Design, synthesis and biological activity of sphingosine kinase 2 selective inhibitors. *Bioorg. Med. Chem.* **2012,** *20,* 183-194.
41. Liu, K.; Guo, T. L.; Hait, N. C.; Allegood, J.; Parikh, H. I.; Xu, W.; Kellogg, G. E.; Grant, S.; Spiegel, S.; Zhang, S. Biological characterization of 3-(2-amino-ethyl)-5-[3-(4-butoxyl-phenyl)-propylidene]-thiazolidine-2,4-dione (K145) as a selective sphingosine kinase-2 inhibitor and anticancer agent. *PLoS One* **2013,** *8,* e56471.
42. Li, Q.; Al-Ayoubi, A.; Guo, T.; Zheng, H.; Sarkar, A.; Nguyen, T.; Eblen, S. T.; Grant, S.; Kellogg, G. E.; Zhang, S. Structure-activity relationship (SAR) studies of 3-(2-amino-ethyl)-5-(4-ethoxy-benzylidene)-thiazolidine-2,4-dione: development of potential substrate-specific ERK1/2 inhibitors. *Bioorg. Med. Chem. Lett.* **2009,** *19,* 6042-6046.

43. Liu, K.; Rao, W.; Parikh, H.; Li, Q.; Guo, T. L.; Grant, S.; Kellogg, G. E.; Zhang, S. 3,5-Disubstituted-thiazolidine-2,4-dione analogs as anticancer agents: design, synthesis and biological characterization. *Eur. J. Med. Chem.* **2012,** *47,* 125-137.
44. Tomasic, T.; Masic, L. P. Rhodanine as a privileged scaffold in drug discovery. *Curr. Med. Chem.* **2009,** *16,* 1596-1629.
45. Knight, S. D.; Adams, N. D.; Burgess, J. L.; Chaudhari, A. M.; Darcy, M. G.; Donatelli, C. A.; Luengo, J. I.; Newlander, K. A.; Parrish, C. A.; Ridgers, L. H. Discovery of GSK2126458, a highly potent inhibitor of PI3K and the mammalian target of rapamycin. *ACS Med. Chem. Lett.* **2010,** *1,* 39-43.
46. Pitson, S. M.; Moretti, P. A.; Zebol, J. R.; Zareie, R.; Derian, C. K.; Darrow, A. L.; Qi, J.; D'Andrea, R. J.; Bagley, C. J.; Vadas, M. A.; Wattenberg, B. W. The nucleotide-binding site of human sphingosine kinase 1. *J. Biol. Chem.* **2002,** *277,* 49545-49553.
47. Sugiura, M.; Kono, K.; Liu, H.; Shimizugawa, T.; Minekura, H.; Spiegel, S.; Kohama, T. Ceramide kinase, a novel lipid kinase. Molecular cloning and functional characterization. *J. Biol. Chem.* **2002,** *277,* 23294-23300.
48. Yokota, S.; Taniguchi, Y.; Kihara, A.; Mitsutake, S.; Igarashi, Y. Asp177 in C4 domain of mouse sphingosine kinase 1a is important for the sphingosine recognition. *FEBS Lett.* **2004,** *578,* 106-110.
49. Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007,** *23,* 2947-2948.
50. Clamp, M.; Cuff, J.; Searle, S. M.; Barton, G. J. The Jalview java alignment editor. *Bioinformatics* **2004,** *20,* 426-427.
51. Waterhouse, A. M.; Procter, J. B.; Martin, D. M.; Clamp, M.; Barton, G. J. Jalview Version 2 - A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009,** *25,* 1189-1191.
52. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997,** *25,* 3389-3402.
53. Altschul, S. F.; Wootton, J. C.; Gertz, E. M.; Agarwala, R.; Morgulis, A.; Schaffer, A. A.; Yu, Y. K. Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* **2005,** *272,* 5101-5109.
54. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M. Y.; Pieper, U.; Sali, A., Comparative protein structure modeling using MODELLER. In *Current Protocols in Protein Science*, 2008/04/23 ed.; J. Wiley and Sons, Inc.: 2007; Vol. 50, pp 2.9.1-2.9.31.
55. Shen, M. Y.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006,** *15,* 2507-2524.
56. Melo, F.; Sanchez, R.; Sali, A. Statistical potentials for fold assessment. *Protein Sci.* **2002,** *11,* 430-448.
57. Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **2009,** *77,* 778-795.

58. Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B., 3rd; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007,** *35,* W375-383.
59. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007,** *50,* 726-741.
60. Kellogg, G. E.; Abraham, D. J. Hydrophobicity: is LogP(o/w) more than the sum of its parts? *Eur. J. Med. Chem.* **2000,** *35,* 651-661.
61. Schrodinger, LLC *The PyMOL Molecular Graphics System, Version 1.5.0.4*, 2010.
62. Zemann, B.; Kinzel, B.; Muller, M.; Reuschel, R.; Mechtcheriakova, D.; Urtz, N.; Bornancin, F.; Baumruker, T.; Billich, A. Sphingosine kinase type 2 is essential for lymphopenia induced by the immunomodulatory drug FTY720. *Blood* **2006,** *107,* 1454-1458.
63. Tonelli, F.; Lim, K. G.; Loveridge, C.; Long, J.; Pitson, S. M.; Tigyi, G.; Bittman, R.; Pyne, S.; Pyne, N. J. FTY720 and (S)-FTY720 vinylphosphonate inhibit sphingosine kinase 1 and promote its proteasomal degradation in human pulmonary artery smooth muscle, breast cancer and androgen-independent prostate cancer cells. *Cell. Signal.* **2010,** *22,* 1536-1542.
64. Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. *J. Med. Chem.* **2002,** *45,* 2469-2483.
65. Marabotti, A.; Spyrakis, F.; Facchiano, A.; Cozzini, P.; Alberti, S.; Kellogg, G. E.; Mozzarelli, A. Energy-based prediction of amino acid-nucleotide base recognition. *J. Comput. Chem.* **2008,** *29,* 1955-1969.

CHAPTER 3

HOMOLOGY MODELING OF HUMAN CYTOMEGALOVIRUS ALKALINE NUCLEASE UL98 AND IDENTIFICATION OF POTENTIAL LEADS BY VIRTUAL SCREENING

3.1 Introduction

3.1.1 *Human Cytomegalovirus (HCMV) – A human pathogen*

Cytomegaloviruses, one of the major lineages in the *Herpesviridae* family, are present in a wide range of mammalian species. Since they have a higher tendency to infect the salivary gland, they are also termed as “salivary gland viruses”. Out of the eight human pathogens belonging to *Herpesviridae* family, human cytomegalovirus (CMV), also known as human herpes virus 5 (HHV-5), is the most extensively characterized member. It is a prototypic member of this subfamily with a large double-stranded DNA genome of about 235 kbp in size, that consists of unique long and short segments, each surrounded by inverted repeats.[1]

HCMV infection is existent in most populations around the world, with overall HCMV seroprevalence ranging from 20%–100% in different countries. Higher seroprevalence, nearly up to 100%, has been observed in individuals from resource-constrained countries; in contrast, adults from well-developed countries from Northern Europe and North America are associated with lower rates of

CMV infection.[2] The overall age-adjusted CMV seroprevalence in the United States is ~50%.[3]

Infectious cytomegalovirus can be found in various body fluids like tears, saliva, blood, urine and semen. CMV is spread easily with extended and repeated exposure to virus in infants/young children attending day-care facilities, hospitalized patients and hospital staff.[4] Prolonged virus shedding from infected individuals seems to be a common source of CMV acquisition in the community. CMV transmission easily occurs by person-to-person contact or contact with contaminated surfaces. Infants, toddlers and young children are an important reservoir of the virus. The virus is efficiently transmitted among children attending the same child-care centers by constant physical contact, hand-to-mouth contact or by limited personal hygiene practices. A susceptible seronegative mother or a pregnant woman visiting the group care facilities is at a very high risk of infection, which in turn leads to congenital CMV infection of the fetus[4]. Lactating mothers act as a source of CMV to newborn babies due to ingestion of infected breast milk.[5] Infected children readily transmit the virus to adults. The shedding of CMV virus via the genitourinary tract leads to its transmission among adults during sexual activity.[6] Sources of CMV in hospitalized patients include blood product transfusions and transplanted organs from seropositive sources. Health-care workers are also at a risk of acquiring infection since they are in such environments for prolonged periods of time.

CMV infections acquired in healthy people are usually mild and asymptomatic, and typically go unnoticed. On the other hand, CMV infections can have severe clinical and pathological manifestations leading to life threatening conditions in

immunocompromised patients, especially transplant recipients and HIV-positive patients.[7] CMV pneumonitis is a major cause of morbidity and mortality in severely immunocompromised patients following solid organ transplantation or hematopoietic stem cell transplantation, even with treatment.[8] In HIV-infected patients, a CMV infection causes serious ocular complications like retinitis, gastric complications leading to hemorrhage/perforations, and in some cases neurological damage causing paralysis or fatal encephalitis.[9,10] Congenital and neonatal CMV infections are more serious and in many cases are life-threatening. In newborns congenital CMV infections are the leading cause of sensorineural hearing loss (SNHL), which may be present at birth or develop later in childhood.[11,12] CMV infections acquired during pregnancy have been reported to cause learning disabilities and mental retardation.[13] More recently, reports of CMV infections in immunocompetent patients with prolonged and relapsing illnesses involving fever, sweats, and in some cases abnormal liver functions, have been well documented.[14,15]

3.1.2 *Antiviral Therapy for CMV Infection*

Over the past few years, major advances have been made in the treatment and prevention of CMV infections by the development of antiviral agents. Currently available licensed antivirals for CMV (Figure 3.1) include ganciclovir (Cytovene®), its valine ester prodrug valganciclovir (Valcyte®), foscarnet (Foscavir®), cidofovir (Vistide®), and Cytomegalovirus Immune Globulin (CytoGam®) (summarized in Table 3.1).

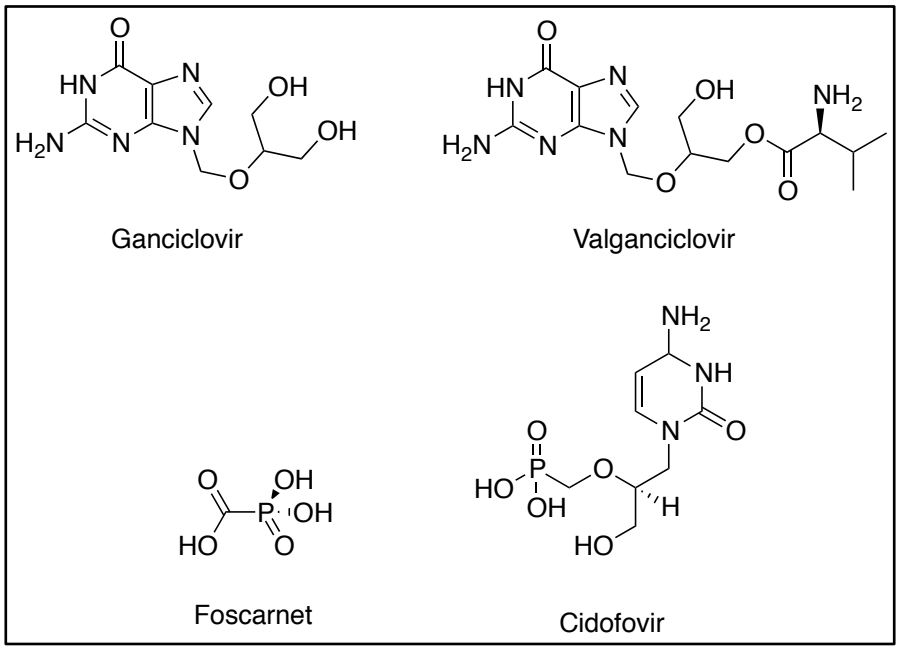


Figure 3.1 – Antivirals for cytomegalovirus infections

Table 3.1 – Summary of currently licensed antivirals for CMV infections (adapted from ref 1,11).

Available Drug	Dose	Major Indications	Associated Toxicities
Ganciclovir Cytovene®	10–12 mg/kg/day IV	Prophylaxis; Congenital and perinatal CMV infection; CMV retinitis; End-organ disease in immunosuppressed patients	Hematotoxicity (mainly neutropenia, thrombocytopenia); Carcinogenic; Teratogenic
Valganciclovir Valcyte®	–	Prophylaxis; CMV retinitis	Identical to its active metabolite – ganciclovir
Foscarnet Foscavir®	90 mg/kg q 12h IV	Prophylaxis; CMV retinitis; End-organ disease in immunosuppressed patients with ganciclovir-resistant infections	Nephrotoxicity; Electrolyte imbalances; Teratogenic
Cidofovir Vistide®	5 mg/kg q weekly	End-organ disease in immunosuppressed patients with ganciclovir-resistant infections	Nephrotoxicity; Ocular toxicity; Carcinogenic; Teratogenic
Cytomegalovirus Immune Globulin Cytogam®	100–150 mg/kg, IV post-transplant	Prophylaxis; Adjunctive treatment for CMV pneumonitis in immunocompromised patients	Minimal toxicity

Ganciclovir (and its prodrug – vanciclovir) is a guanosine analog, which upon phosphorylation by the CMV UL97 kinase, acts as a chain terminator during viral DNA replication. Ganciclovir therapy is the treatment of choice for CMV related retinitis and pneumonitis in transplant patients.[16-18] Foscarnet is a structural analog of pyrophosphate and inhibits the CMV DNA polymerase by binding at its pyrophosphate-binding site and halts DNA chain elongation. DNA modifying enzymes in CMV produces pyrophosphate as one of the products, thus making Foscarnet a product inhibitor; and so unlike ganciclovir, it does not compete with natural nucleotides. It does not require activation by the phosphorylative enzymes of either the host or the virus. Foscarnet is an effective second-line treatment for CMV infections in HIV-patients and in cases which develop ganciclovir resistance.[19] Cidofovir, like foscarnet, is a second-line therapy antiviral. It is an acyclic deoxycytidine monophosphate analog, which upon conversion to diphosphoryl metabolite by cellular enzymes inhibits the CMV DNA polymerase.[20] Cytomegalovirus Immune Globulin is an intravenous immunoglobulin preparation that is indicated for prophylaxis against CMV infections in transplant recipients. It is prescribed either alone, or in combination with any of the above-mentioned antiviral agents.[11]

All of the above drugs share a similar mechanism of action, i.e., inhibition of viral DNA polymerase and interfering with its DNA synthesis process. There are also significant toxicities associated with each of these drugs. In immunocompromised patients, ganciclovir is associated with bone marrow suppression, mainly granulocytopenia and thrombocytopenia. Ganciclovir, when used for a prolonged period

of time, is considered a potential human carcinogen, teratogen, and mutagen.[21] The most important clinical problem that may emerge during foscarnet/cidofovir therapy is nephrotoxicity.[22,23] Although suited for prophylaxis and treatment of CMV related infections in adults, evidence of teratogenicity in lab animals makes these drugs not ideal for treatment in pregnant women.[24] Also, there is limited evidence for using these antivirals to treat congenital and perinatal HCMV infections.[25] Patients receiving prolonged therapy are at the risk of developing resistance to these antivirals.[26] A ganciclovir resistance study, on a group of AIDS patients suffering from CMV retinitis by *Jabs D. A., et al.,* showed that about 11% of the patients developed resistant-strain virus within 6 months of treatment, and about 28% by 9 months.[27] Being structurally so similar, it would be expected that a virus resistant to one compound would exhibit cross-resistance to others.

Although these antiviral drugs have been useful in treating a range of CMV-related infections, the toxicities and chance of developing resistance associated with them emphasize the need for developing less toxic novel antivirals, especially those that target alternate processes essential for viral survival. Safer, nonteratogenic antivirals could be used during pregnancy to treat fetal infections or in neonates to prevent CMV related mental retardation and SNHL.

3.1.3 *HCMV Alkaline Nuclease UL98 – A novel target*

The general idea behind developing antivirals against CMV is to identify and disable viral proteins that are important for its development. Novel compounds targeting different processes are desirable because of their potential to be used in combination therapy and to avoid resistant-strain viral development. Alkaline nucleases, encoded in all herpes viruses, represent one such target.

The herpesvirus alkaline nucleases are DNA-modifying enzymes that possess both 5' – 3' exonucleolytic activity and a slightly moderate endonucleolytic activity.[28] They have a high *in vitro* pH optimum and hence are termed "alkaline". These enzymes can process both single-stranded and double-stranded DNA substrates, with a notable preference for supercoiled substrates. Different members of the *Herpesviridae* family – viz. Herpes Simplex Virus (HSV), Cytomegalovirus (CMV), Kaposi's Sarcoma associated herpesvirus (KSHV) and Epstein-Barr Virus (EBV), encode the alkaline exonuclease. The majority of data on the biological role of these enzymes at different stages of viral development comes from studies on Herpes Simplex Virus type I (HSV-1) Alkaline Nuclease – UL12, which is the first described, expressed and mapped herpesvirus alkaline nuclease.[29] Although the exact function of alkaline nucleases in viral survival is unknown, studies have shown that HSV-1 AN is required for efficient processing of viral DNA replication intermediates, suggesting its role in maturation and packaging of viral DNA into capsids.[30] Additional reports have indicated the importance of HSV-1 AN in the efficient egress of capsids from the nucleus.[31]

Different genes involved in viral DNA replication, cleavage and packaging, are highly conserved among HSV-1 and CMV[32]; even though the HSV-1 genome is significantly smaller in size (~150 kb compared to ~235 kb of CMV). Predictions based on gene arrangement and amino acid sequence homology have indicated the HCMV *UL98* gene to be the counterpart of HSV-1 *UL12* gene. *Trans*-complementation experiments have demonstrated functional conservation of proteins encoded by the homologous genes of HSV1 and CMV.[33] The UL98 protein encoded by the CMV *UL98* gene is indeed the conserved enzyme homolog of α- and γ- herpesvirus alkaline nucleases.[34]

In vitro, HCMV UL98 AN shows both 5' – 3' exonuclease (exo) and endonuclease (endo) activities that are optimal at alkaline pH, similar to other ANs. It also can hydrolyze dsDNA and ssDNA substrates and requires a divalent cation for activity, with preference for magnesium ion.[34] The role of UL98 alkaline kinase in CMV infection is extremely important, suggested by the fact that its synthesis starts at early stages of infection, with significant increase in its levels after the onset of viral DNA replication.[35] In more recent studies, the global functional analysis of HCMV genome has identified the *UL98* gene, which encodes the UL98 AN, as *essential* for viral growth.[36,37] The fact that HCMV UL98 gene complemented an HSV-1 UL12 deletion mutant functionally can be used to assume a similar role of UL98 alkaline nuclease in CMV DNA modification, capsid stability and egress.[33] Although the alkaline nuclease is

not required for viral DNA synthesis,[38] there is enough evidence to emphasize the importance of UL98 AN in the different stages of CMV viral development.

The plausible impact of UL98 AN on CMV viral growth, and the presence of an alkaline nuclease homologue in every herpesvirus, makes it a unique antiviral target for CMV infections. Unlike traditional approaches that target viral DNA synthesis by inhibiting DNA polymerase, this approach targets the late events in viral replication like viral DNA packing, capsid stability or egress. This would be beneficial to treat resistant infections and also prevent their occurrence with combination therapy. Also, identifying novel inhibitors of UL98 AN may help to overcome the modest antiviral activity of currently approved drugs and their dose-related toxicities, and result in a highly efficacious antiviral therapy, intended to be used in patients of all ages and conditions (infants, adults, pregnant women, transplant recipients, AIDS patients).

3.1.4 *Alkaline Nucleases – Structural insights*

A phylogenetic analysis, by *Rychlewski, L., et. al.*, has classified the herpesvirus alkaline nucleases within the PD-(D/E)XK superfamily of DNases[39]. This family includes structurally well-characterized, functionally diverse members like restriction endonucleases[40] (EcoRI and EcoRV), DNA-nicking enzymes (Vsr[41] and MutH[42]), and bacteriophage λ exonuclease[43] (λ-exo). Although these enzymes share little overall sequence similarity, they have a common core fold and a conserved, well-defined PDX_{10-30}(D/E)XK motif that is involved in metal-binding and catalysis. The central core

region is formed of four-stranded, mixed β-sheets flanked by two α-helices on either side, with αββββαβ topology (Figure 3.2). The catalytic sites of these enzymes contain a conserved aspartic acid (D), an aspartic or glutamic acid (D/E) and a lysine (K).[44] To date, crystal structures of two herpesvirus alkaline nucleases have been solved – the Kaposi's sarcoma-associated herpesvirus Shut-off and Exonuclease[45] (KSHV-SOX) and its Epstein-Barr virus (EBV) homolog BGLF5.[46] They confirm the structural similarity of herpesvirus ANs to the PD-(D/E)XK superfamily, sharing a comparable central core fold and active-site formation.

Very little is known about the structure of CMV UL98 alkaline nuclease, apart from the fact that it too is a herpesvirus alkaline nuclease, and therefore should be structurally similar to proteins of PD-(D/E)XK superfamily. The knowledge of the three-dimensional structure of UL98 AN is required for better understanding of its functional mechanism and to elucidate its exact role in virus growth and development. The recent computational advances in homology protein structure modeling have enabled reliable prediction of unknown protein structure based on a homologous protein of known structure. The model can then be potentially used for structure-based design of a novel class of antiviral drugs.

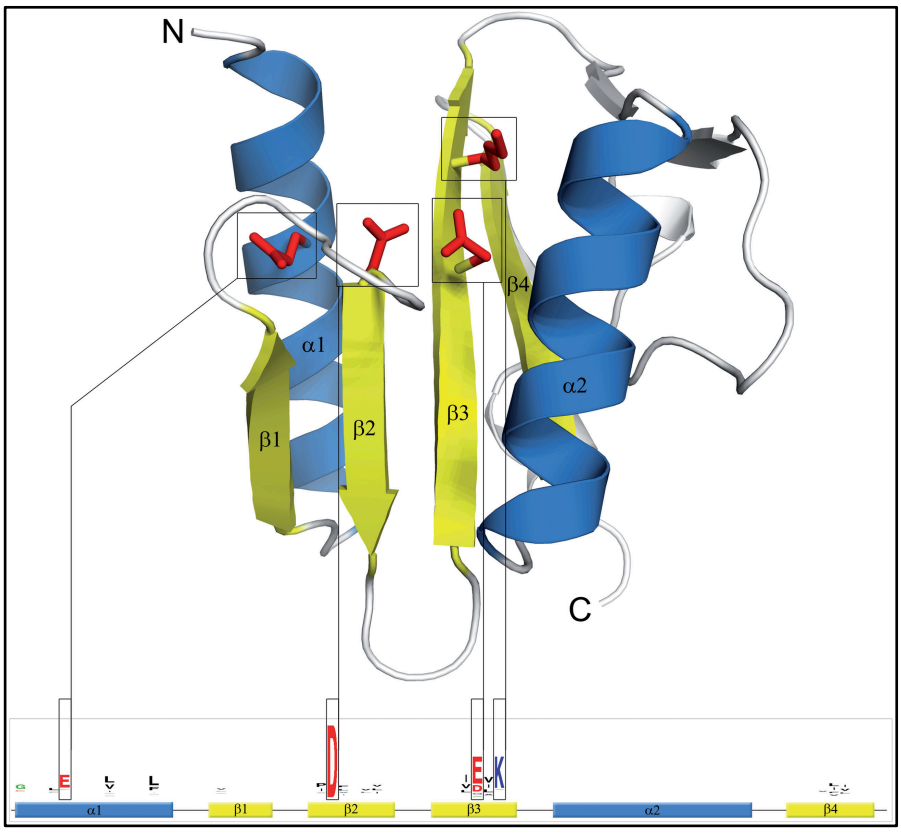


Figure 3.2 – The conserved PD-(D/E)XK core fold with active site formation (Reprint from ref 44).

3.1.5 *Specific Aims*

In the current work, we aim at identifying novel antiviral agents that target CMV UL98 AN using traditional computational approaches like homology modeling and 3D virtual screening on the protein model. The working hypothesis that underlies this project is that the nuclease activity of UL98 is critical for HCMV viral development and targeting it may lead to a novel, efficacious antiviral therapy for CMV infections, with reduced toxicities. The work has been divided into two specific aims:

1) Structural modeling of CMV UL98 AN, and
2) Virtual screening on the active-site of UL98 AN model

The first part of this work involved homology-based structural modeling of UL98 AN in order to identify the active-site residues. The active-site residues were validated using mutagenesis, in the lab of *Dr. Michael McVoy*, Professor of Pediatrics, Molecular Biology and Genetics, School of Medicine, VCU. Mutant viral construction was also performed to show the importance of UL98 in CMV replication. The computational work has been described in detail in the following sections, and the findings have been recently published (Kuchta *et al*).[47]

The second part of the work involved structure-based drug discovery on the UL98 AN model with an aim to identify novel small-molecules that might inhibit its nuclease activity. We have performed 3D virtual screening on the active site, the results of which will be discussed in detail in the following sections.

3.2 Methods

3.2.1 *Structural Modeling of HCMV UL98 AN*

3.2.1.1 Template Identification and Alignment

PSI-BLAST (Position-Specific Iterated BLAST)[48,49] was performed on the amino acid sequence of HCMV UL98 AN (strain AD169) using the PDB Protein Database to identify a template structure based on which its structure can be built. The first Expect value (E-value) was set to 1 for the initial BLAST search and the second E-value, which is the threshold value for inclusion in the position specific matrix (BLOSUM62) used for PSI-BLAST iterations, was set to 0. The Shutoff and Exonuclease Protein from KSHV (KSHV-SOX) (PDB ID: 3FHD)[45] was chosen as the *"template"* structure for model building, based on its max score of 104 and E-value of 1e-22. Clustal X v2.0[50] was used to align the amino acid sequence of UL98 AN (query sequence) with the sequence of KSHV-SOX (template). The active site and the 5' phosphate binding residues in UL98 AN were identified based on their alignment with corresponding residues of KSHV-SOX.

3.2.1.2 Homology Modeling of UL98 AN

The homology model of UL98 AN was constructed based on the available crystal structure of KSHV-SOX (PDB ID: 3FHD) and the sequence alignment obtained from Clustal X v2.0 using Comparative Modeling Program MODELLER 9v7.[51-54] MODELLER sets up spatial restraints on Cα-Cα distances, main chain N-O distances and main chain and side chain dihedral angles obtained from the template crystal structure.[51] A total of

100 models were generated using the *automodel* class of MODELLER, which builds models by violating these restraints as little as possible. The final model was selected based on its ability to accommodate a DNA fragment in its active site crevice.

3.2.1.3 Docking of dsDNA into the active site of UL98 AN

Protein structure preparation. From all 100 generated models, the unaligned regions that were built by MODELLER without a template were deleted using the general molecular modeling program SYBYL v8.1 (TRIPOS Inc.); i.e., residues Met1-Glu25, Arg83-Ile90, Thr344-Leu355, Lys381-Ser419, Asp431-Val446, and Ser569-Pro584. The region starting at Asp189 and ending at Phe211, which is near the active site, aligned with 'the bridge' region of the template (Pro164-Phe179), which could not be crystallized owing to disorder,[45] was deleted from the alignment. All the other regions that were deleted were distant from the active site crevice of UL98 AN and should not affect later modeling in this region. Hydrogens were added to all the protein structures, charges were calculated by the Gasteiger-Hückel method, and all models subjected to Powell minimization for 1000 steps, keeping the coordinates of all non-hydrogen atoms fixed.

Substrate preparation. A B-DNA dodecamer (PDB ID: 3BNA) was chosen as the ligand to be docked into the active site crevice of UL98 AN homology models using GOLD v4.1. Since this ligand was too big, a 4bp long dsDNA fragment (A-C-G-T) was built in SYBYL v8.1 (TRIPOS Inc.), and used as a reference ligand for docking. The positions of the 5'-phosphate group and the phosphodiester bond joining the final two

bases as predicted by GOLD, were then used to overlay the corresponding groups of 3BNA in order to get the final protein-ligand complex.

Docking. The ACGT reference ligand was docked into the active site crevice of all MODELLER generated UL98 AN homology models using GOLD v4.1. The binding site was defined to encompass all atoms within a 15 Å radius of Oδ1 of Asp254, which is located at the center of the active site. All rotatable bonds, except terminal, in the ACGT reference ligand were fixed to make sure that GOLD would not disrupt the double stranded helix structure by rotating the other bonds in the ligand structure to find a better binding mode. Default Genetic Algorithm parameters were used. A total of 30 solutions were generated for each UL98 AN homology model. The solutions for each model were scored and ranked using the GOLD Fitness Scoring Function. Only 8 out of all the homology models gave positive GOLD Fitness Scores, suggesting that the other models were unable to accommodate the ACGT ligand. The binding modes of the ACGT ligand in the remaining 8 models were visually analyzed to see which UL98 AN model was best able to accommodate the reference ligand and explain its exonuclease activity. HCMV UL98 AN homology model065 was chosen, as it was best able to accommodate the 5’ phosphate group among all the models. This final model was subjected to further refinement. The 5’ phosphate group and P2 group of 3BNA were overlaid onto the positions of the corresponding groups in the ACGT reference ligand, in its best conformation for UL98 AN model065 as predicted by GOLD, in order to obtain the best possible conformation for 3BNA in the active site of UL98 AN.

3.2.1.4 Final UL98 AN Model – Refinement and Validation

UL98 AN model065 was evaluated using the DOPE (Discrete Optimized Protein Energy) scoring function[55] and GA341 assessment score,[56] incorporated within MODELLER 9v7. A DOPE per-residue score was calculated for this model065 and compared to that of the KSHV-SOX crystal structure.

KSHV-SOX is a PD-(D/E)XK nuclease that uses a magnesium ion and a water molecule for activity.[45] The relative coordinates of a putative active site magnesium ion and water molecule for UL98 AN model were obtained from the crystal structure of KSHV-SOX, by aligning their structures based on homology within SYBYL v8.1. Model065, now containing active site magnesium ion and water molecule, was complexed with 3BNA (in a position analogous to the best conformation of the ACGT reference ligand). The side-chain of Ser252 in the active site of UL98 AN was manually rotated towards the 5' phosphate group to show better interaction, mimicking the position of the corresponding active site Ser219 of KSHV-SOX. This protein-ligand complex was then subjected to Powell Minimization to a gradient of 0.005 kcal mol^{-1} $Å^{-1}$ to account for any steric clashes and to obtain an induced fit model of UL98 AN with its substrate. Following minimization, all the ligands were deleted from the protein and Ramachandran plots were calculated using MolProbity.[57]

3.2.1.5 Multiple Sequence Alignment of Herpesvirus ANs

In order to check that the active site residues identified for HCMV UL98 AN are conserved within other Herpesviridae ANs, the primary amino acid sequences for Herpesvirus ANs, including Human Cytomegalovirus (HCMV) Alkaline Nuclease UL98, ORF37 of Kaposi's Sarcoma Associated Herpesvirus (KSHV) (ABD28888), BGLF5 of Epstein-Barr Virus (EBV) (ABB89261), U70 of Human Herpesvirus 6 (HHV6) (NP-050249), U70 of Human Herpesvirus 7 (HHV7) (AAC40784) and UL12 of Human Herpesvirus 1 (HHV1) (BAA84005) were obtained from the NCBI (http://www.ncbi.nlm.nih.gov/) and aligned using Clustal X v2.0.[50]

3.2.1.6 Estimation of Changes in Free Energies of Association of UL98 AN with dsDNA upon Mutation

Intermolecular interactions between the UL98 AN and dsDNA (3BNA), in their minimized complex form, were calculated using HINT (Hydropathic INTeraction). The HINT score (H_{total}) was calculated and converted to free energy using the equation: $\Delta G_{binding} = -0.00195 H_{total} - 5.543$.[58] The structures of mutants of UL98 AN (R164A, S252A, D254A, E278A and K280A) were generated within the structure of UL98 – dsDNA complex, using SYBYL 8.1 and H_{total} scores were calculated for each mutant-ligand pair, and subsequently converted to $\Delta G_{binding}$ free energy values. On the basis of the predicted $\Delta G_{binding}$ for each UL98 mutant, $\Delta\Delta G_{binding}$ values of differences in free energy between wild type and each UL98 mutant were calculated.

3.2.2 *Validation of UL98 AN Model using Mutagenesis*

All the experimental work was performed in *Dr. Michael McVoy's* lab (Department of Pediatrics, Virginia Commonwealth University, VA) and *Dr. Deborah S Parris's* lab (Department of Molecular Virology, Immunology and Medical Genetics, The Ohio State University, OH). For detailed experimental procedures, please refer to our recent publication – Kuchta, A., *et al.*[47] Briefly, recombinant UL98 and proteins containing single alanine substitutions (R164A, S252A, D254A, E278A or K280A) were expressed with amino-terminal hexahistidine (His6) tags. The proteins were expressed and partially purified by immobilized metal affinity chromatography (IMAC). The exo activities of the IMAC-purified proteins were quantified by release of acid-soluble radioactivity from ^{14}C labelled DNA. Endo activity was determined qualitatively by incubating IMAC-purified proteins with closed-circular pUC19 plasmid DNA and evaluating its conversion to open-circular and linear forms by agarose gel electrophoresis. To quantify endo activity independent of exo activity, a fluorescence-based assay was developed in which fluorescence increases when endonucleolytic cleavage of an ssDNA substrate releases a 3' quencher from a 5' fluorophore.

3.2.3 *3D Virtual Screening on the Active-Site of UL98 AN Model*

The model of UL98 AN – dsDNA complex was used to identify the important molecular interactions involved in the association. These were used as the basis to

generate a pharmacophore model describing features desirable in ligands that might bind at the active site and inhibit its activity.

3.2.3.1 Design of Query

The *UNITY* module within SYBYLx1.2 (http://www.tripos.com/) was used to create a 3D query that can be used to scan databases of diverse chemical compounds. A "*negative center*", sphere of 1 Å radius, with the center at a position analogous to the 5' phosphate atom, was defined. A "*donor atom*" feature was defined at a position near the catalytic metal, sphere of 0.5 Å radius, such that any hydrogen-bond donor group in the ligand would engage the carboxylate groups of D254 and E278 residues and disrupt the enzyme's catalytic activity. Another feature defined in the catalytic region was an "*acceptor atom*", sphere of 0.5 Å radius, complementary to K280. A central "*aromatic hydrophobic*" core region, sphere of 1.5 Å radius, was defined in the vicinity of the deoxyribose ring of the first nucleotide. Next, a *distance constraint*, with a tolerance of ±0.5Å, was set up between the negative center and the donor atom. Finally, *receptor-site constraints* were defined around heavy atoms of the protein surrounding the cavity.

The generated 3D query was used to screen a library of over 250,000 compounds belonging to the National Cancer Institute Open Database (http://cactus.nci.nih.gov/ncidb2.1/). A "flexible 3D search" that uses a torsional minimizer was performed. This technique identifies molecules that might fit the defined query. Also, all the compounds are screened for their drug-like properties based on Lipinski's rule of five.[59]

The entire data mining process reduced the candidates to a more manageable number, all of which were subjected to docking into the active site of UL98AN.

3.2.3.2 Docking of Hits

The hits obtained at the end of first stage of screening were docked into the active site of UL98 AN using GOLD v5.1. This approach leads to identification of hits having a better fit in the active site, along with all the structural features necessary for binding. Before this step, all the hits were visualized to check for incorrect atom-types, bond-types, unsatisfied valencies; all such were corrected if necessary, and saved in a ready-to-dock, 3D format. The binding site was defined to encompass all atoms within 15 Å radius of Oδ1 of Asp254, which is located at the center of the active site. Default genetic algorithm parameters were used. A total of 100 solutions per ligand were generated (no early termination, no constraints) in order to obtain multiple poses within the binding site.

3.2.3.3 Scoring of Docked Poses

To identify the most likely binding mode of a ligand within UL98 AN active site, out of all possible solutions generated from docking, each protein-ligand complex was scored using the HINT forcefield. Unlike the default scoring function of the docking software, HINT scores are known to correlate with the binding free energy.[58,60,61] The binding modes of the top scored solution for each ligand was visually analyzed. Hits that showed significant interaction with the important residues of the active site (R164, S252, D254, E278, K280) were retained.

3.3 Results and Discussion

3.3.1 *The Template – Kaposi's Sarcoma Associated Herpesvirus Shut-off and Exonuclease (KSHV-SOX)*

Herpesvirus ANs are classified within the λ exo family of DNases, which is within the PD-(D/E)XK superfamily of DNA-modifying enzymes.[43] The catalytic sites of these enzymes contain a conserved aspartic acid (D), an aspartic or glutamic acid (D/E) and a lysine (K). To identify a potentially related sequence of known structure that can be used as a template to build the structure for UL98 AN, PSI-BLAST was performed on the primary amino acid sequence of UL98 using the PDB Protein Database. The Shutoff and Exonuclease Protein from Kaposi's Sarcoma Associated Herpesvirus (KSHV-SOX) (PDB ID: 3FHD), which belongs to the same PD-(D/E)XK superfamily of proteins, was the closest structural match identified, and shares 26% sequence identity and 40% homology with it. The catalytic residues of KSHV-SOX are D221 and E244, which coordinate Mg^{2+}, and K246, which stabilizes the leaving group.[45] An adjacent 5'-phosphate-binding pocket formed by R139, S146 and S219 was suggested by the presence of a sulphate ion in the crystal structure.[45] Figure 3.3 shows the overall fold of KSHV-SOX and its active site residues.

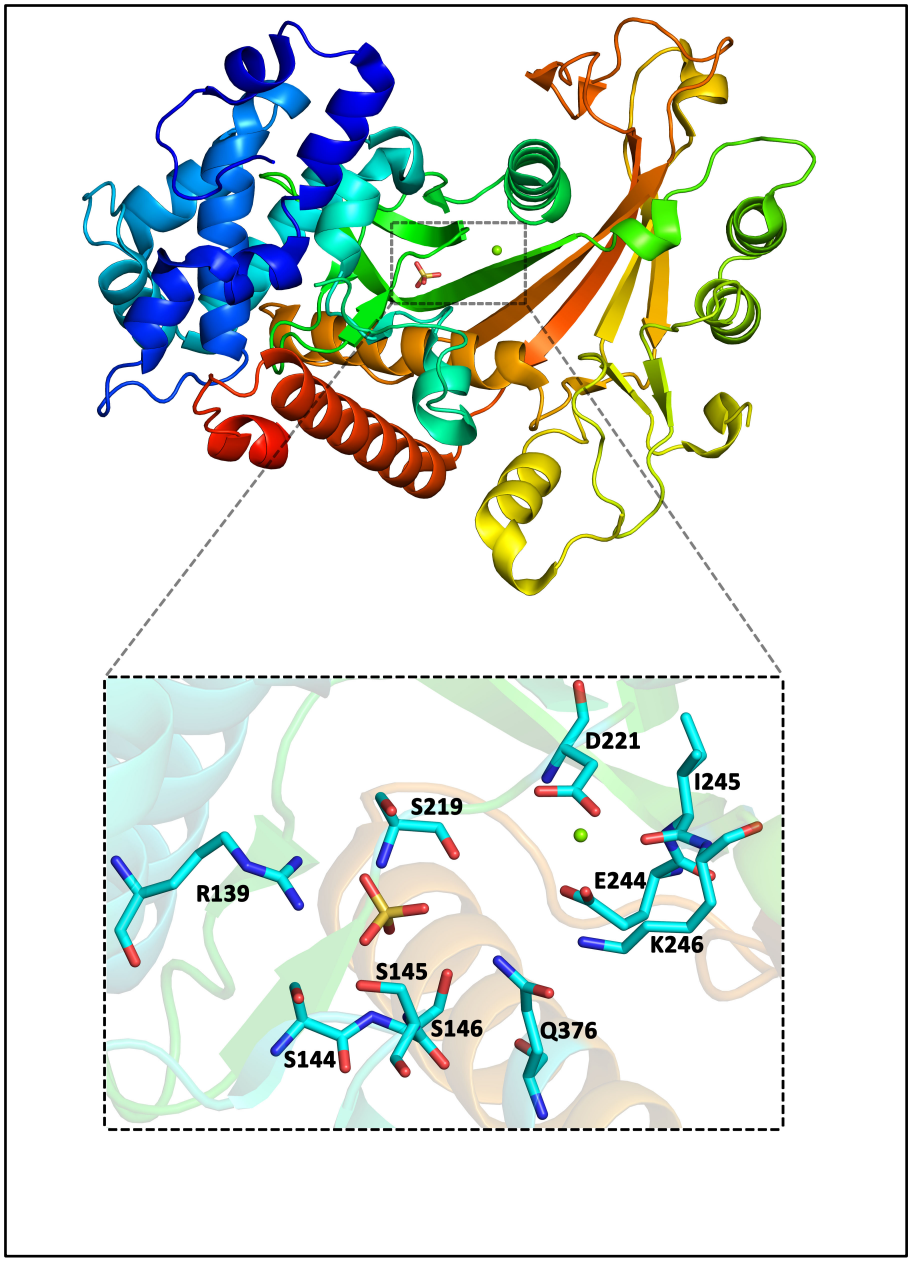


Figure 3.3 – Template structure KSHV-SOX (PDB ID: 3fhd).
The catalytically important Mg^{2+} ion (green sphere) and sulfate ion (sticks) at a position analogous to 5' phosphate binding site. The inset shows the active-site residues (shown as sticks) that are involved in coordination of sulfate ion and Mg^{2+} ion. Image prepared using PyMOL.[62]

3.3.2 *HCMV UL98 AN Model*

Pairwise sequence alignment between the UL98 AN primary amino acid sequence and that of KSHV-SOX was performed using Clustal X v2.0, which helped to identify the regions of similarity (Figure 3.4). Amino acids D254, E278 and K280 of UL98 AN are the corresponding catalytic residues, and R164, T171 and S252 form the putative 5'-phosphate-binding pocket.

A structural model of UL98 was constructed based on the crystal structure of KSHV-SOX and *in silico* docking was used to determine the optimal binding of a dsDNA dodecamer in the predicted UL98 binding site crevice.

A total of 100 models of UL98 AN were generated using the comparative homology modeling program MODELLER 9v7, based on the sequence alignment with KSHV-SOX. The docking was initially performed using a 4 bp long nucleotide (ACGT) as the reference ligand. The most reasonable accommodation of the ligand in the active site was shown by model065. The model shows that 5' phosphate end is held in place deep into the active site crevice through strong hydrogen bonding interactions with side chains of Arg164 and Ser252 and weak interactions with the backbone of Ala170 and Thr171. Also, the phosphodiester bond, where nucleophilic attack takes place for hydrolysis, is situated very close to Asp254, Glu278 and Lys280, which is the metal coordination site.

Model065 was evaluated using the Discrete Optimized Protein Energy (DOPE) scoring function and GA341 assessment score within MODELLER 9v7. DOPE per-

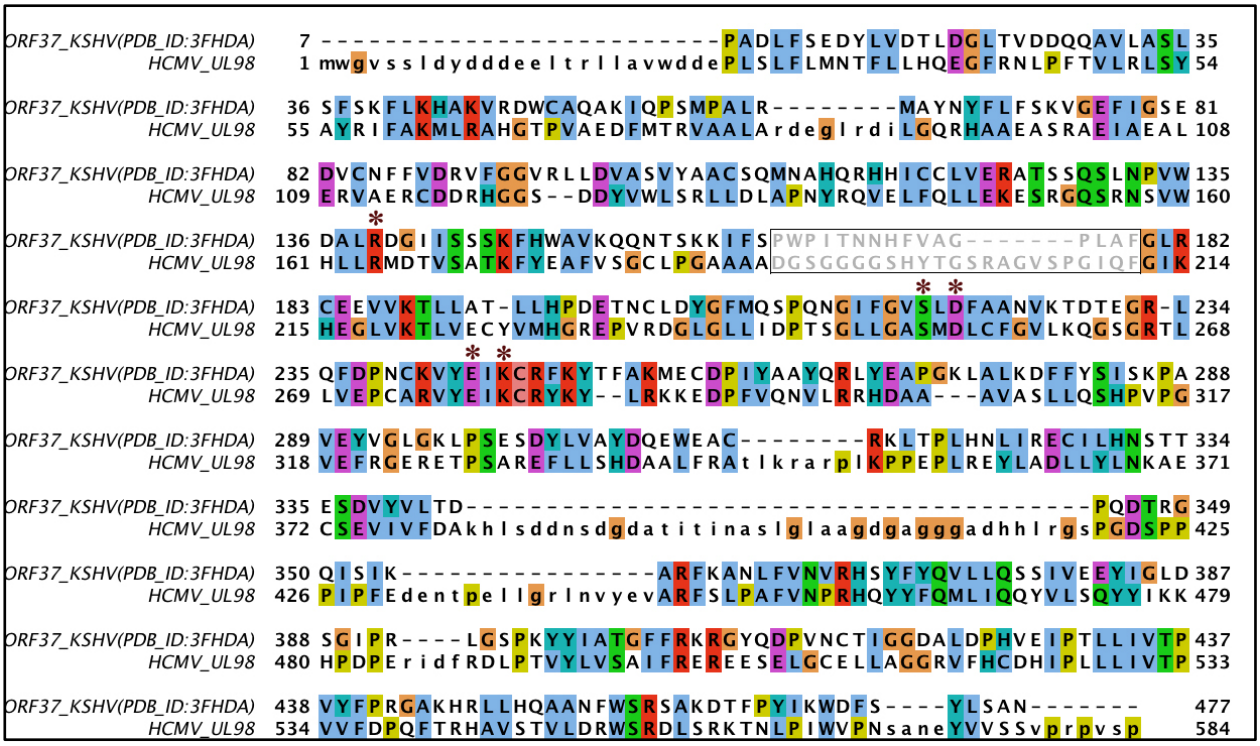


Figure 3.4 – UL98 AN amino acid sequence in pairwise alignment with KSHV-SOX, using Clustal X v2.0.
Asterisks indicate putative active site residues and the default Clustal X v2.0 color scheme is applied. Apparent insertions in UL98 AN relative to KSHV-SOX (lower case) and residues that align with a region in KSHV-SOX that did not crystallize (grey box) were deleted from the UL98 model. Figure prepared using Jalview,[63,64] sequences colored using the default Clustal X color scheme (Appendix A.1).

residue scores calculated for both model065 and 3FHD indicated that model065 depicts UL98 AN in a native-like form, as the two DOPE profiles were similar in the active site region. The GA341 method, which uses percentage sequence identity with the template as a parameter to assess the quality of the predicted model and ranges from 0.0 (worst-model) to 1.0 (native-like), yielded a score for model065 of 0.9999.

The coordinates for a putative Mg^{2+} ion and active site water molecule for model065 were obtained by aligning the template structure to the model. The protein-ligand complex was then subjected to an all-atom minimization to a gradient of 0.005 kcal mol^{-1} $Å^{-1}$ to account for steric clashes and to obtain an induced-fit model of UL98 with its substrate. All the ligands were then deleted from the model and Ramachandran plots were calculated using MolProbity. 83.1% of the residues were in the favored region and 96.8% of residues in the allowed region (Figure 3.5). Out of 438 residues, there were 14 outliers; each was visually inspected and all were found to reside in regions remote from the active site region.

The final UL98 AN structural model was generated following refinement (discussed above) on model065. Figure 3.6 shows the overall structural fold of the protein, comprising an amino-terminal domain consisting of ten α-helices and a carboxy-terminal domain formed by five-stranded β-sheets flanked by five α-helices. The putative Mg^{2+} ion and the active-site water molecule are also shown, along with the active-site residues.

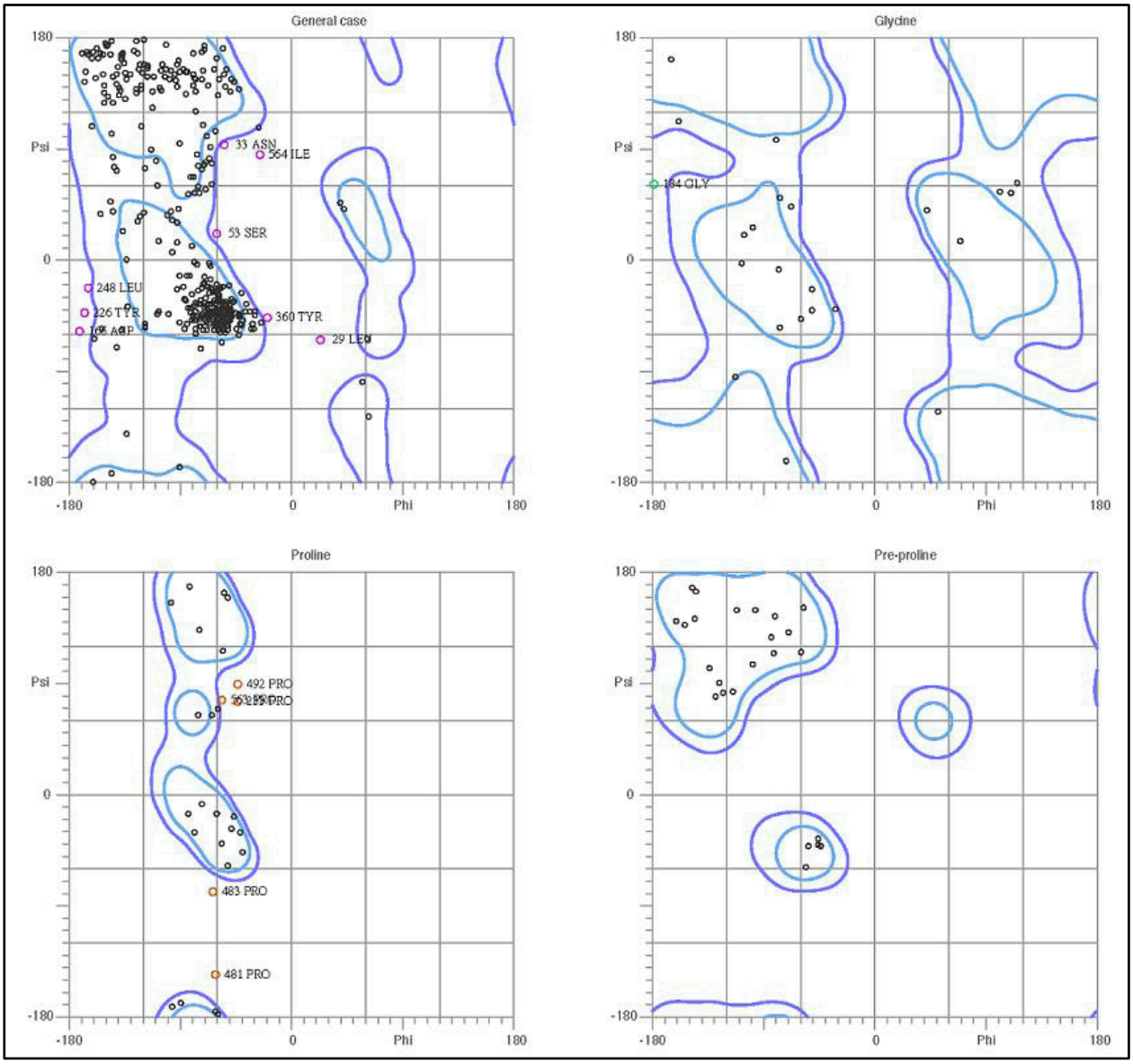


Figure 3.5 – Ramachandran plots for UL98 AN homology model as determined by MolProbity. Outliers are shown in colors.

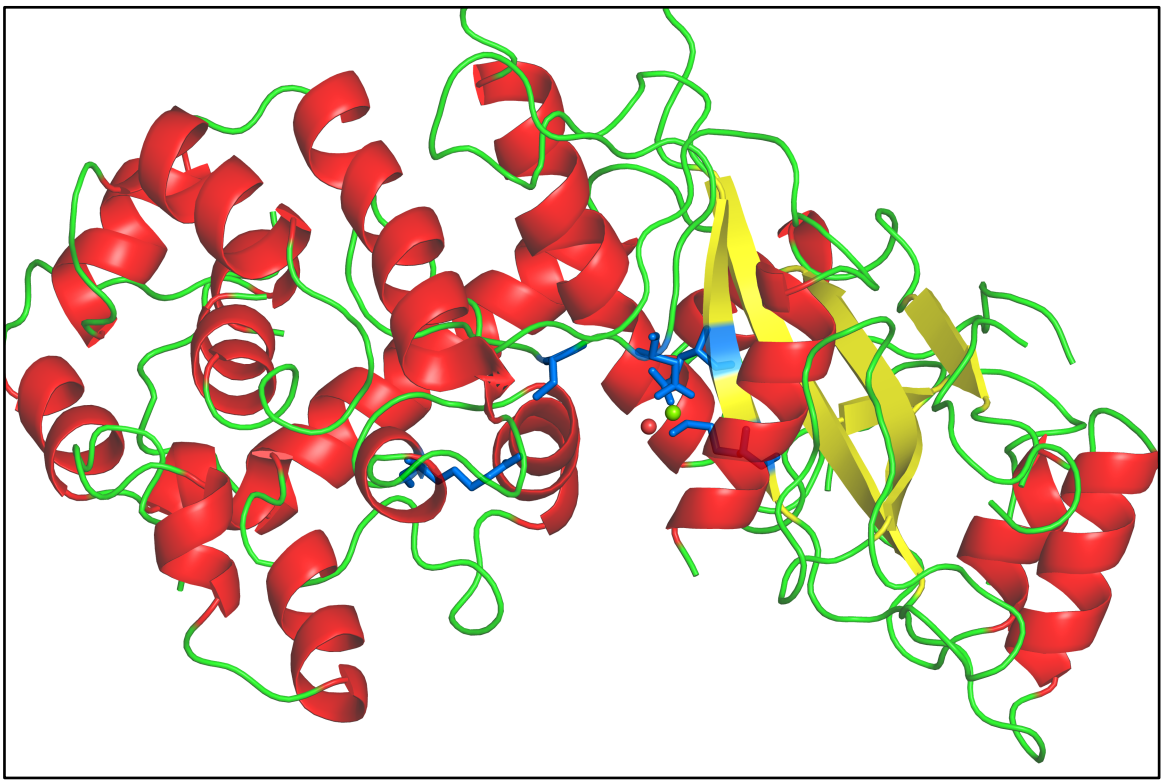


Figure 3.6 – UL98 AN Model.
Active site residues are shown in blue sticks. The catalytically important Mg^{2+} ion and water molecule are shown as spheres. Image prepared using PyMOL.

Guided by the positions of the 5' phosphate group and the P2 group, the structure of crystallized dsDNA (PDB ID: 3BNA) was overlaid on ACGT reference ligand, in its best-docked position within the active site crevice of HCMV-UL98. The predicted position of DNA shows the 5' phosphate held deep into the active-site crevice through hydrogen bonding with R164 and S252. The scissile phosphodiester bond is appropriately positioned proximal to the metal coordination region formed by D254, E278 and K280 (Figure 3.7).

Important active site residues should be strongly conserved within the Herpesviridae ANs. Previous alignments have identified seven amino acid motifs that are conserved among ANs from the α-herpesvirus subfamily and extend to ANs from the β and γ-herpesvirus subfamilies.[65] To determine if the five residues predicted by structural modeling are conserved, amino acid sequences for ANs representing all three subfamilies were aligned using Clustal X v2.0 (Figure 3.8). The putative active site was found to span motifs II and III. Residues Arg164, Ser252, Asp254, Glu278, and Lys280 were fully conserved.

Active-site models for DNA interactions of wild-type and mutant UL98 proteins were generated from the UL98 model. Fig. 3.9 shows predicted active-site structures for wild-type UL98 and each of the five mutants. Loss of hydrogen-bonding interactions resulting from substitutions R164A or S252A may result in imprecise alignment of the DNA in the active site, resulting in improper positioning of the scissile phosphodiester bond relative to the catalytic K280 residue (Figure 3.9(B) and 3.9(C)). This was further

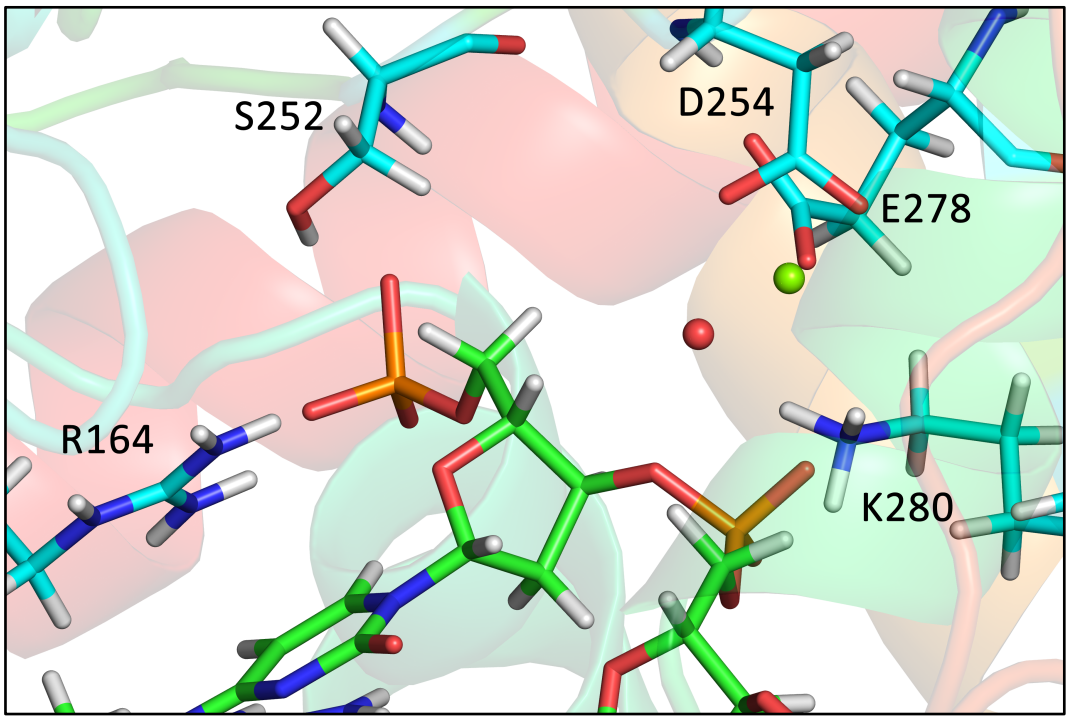


Figure 3.7 – UL98 AN active site model (residues – carbons in cyan; dsDNA – carbons in green).
Mg^{2+} (green sphere) is coordinated by the carboxylate groups of D254 and E278 and holds the water molecule (red sphere) at the active site. K280 is proximal to the phosphodiester group and the 5' phosphate shows strong hydrogen-bond interactions with side chains of R164 and S252. Image prepared using PyMOL.

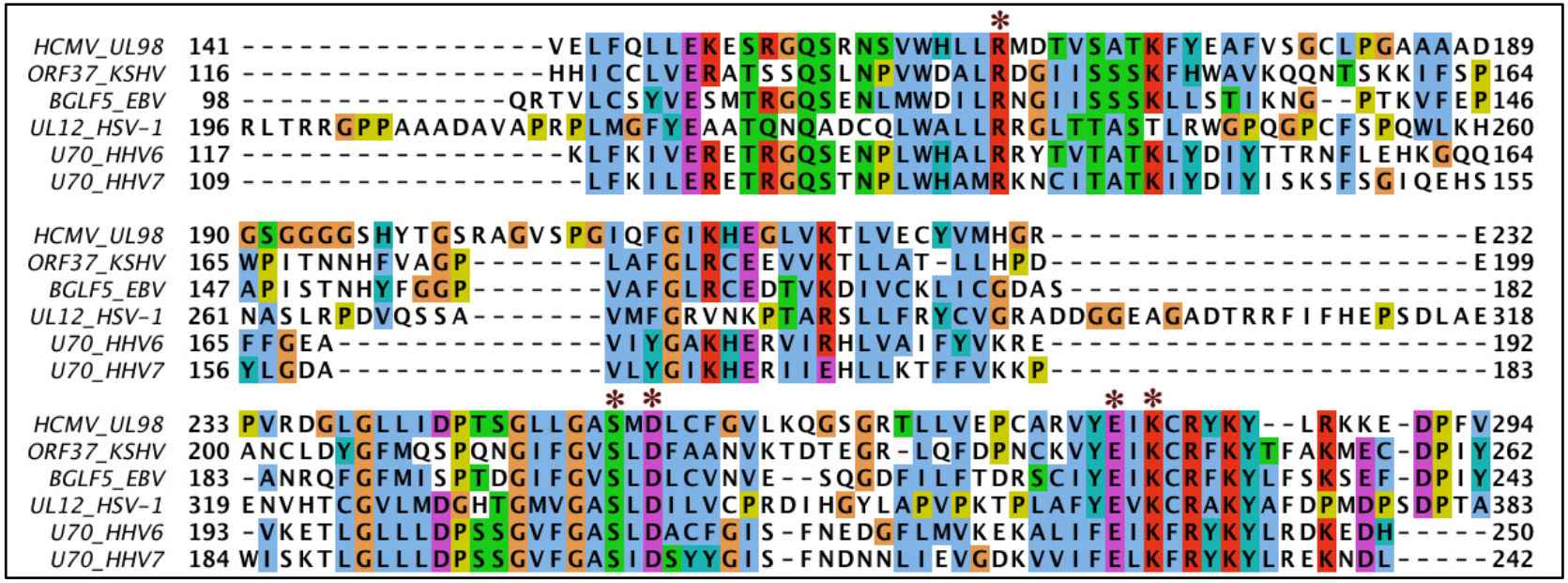


Figure 3.8 – Multiple sequence alignment between Herpesvirus alkaline nucleases.
Asterisks indicate active-site residues that are fully conserved among the entire family. Figure prepared using Jalview, sequence colored using the default Clustal X color scheme (Appendix A.1).

supported by significant changes in the predicted DNA-binding energies for R164A, S252A and K280 mutants (Table 3.2). Similarly, the loss of charged carboxylate residues in D254A or E278A mutants would result in failure to coordinate the magnesium ion and water molecule (Figure 3.9(D) and 3.9(E)). Changes in predicted DNA-binding energies for mutants D254A and E278A are minimal (Table 3.2), consistent with these residues interacting with Mg^{2+} rather than by directly interacting with the DNA. The K280A mutation would be unable to carry out the hydrolysis of the phosphodiester bond and stabilize the leaving group (Fig. 3.9(F)).

Table 3.2 - H_{TOTAL} scores and corresponding $\Delta\Delta G_{binding}$ energies calculated for wild-type and UL98 AN mutants

Protein	**H_{TOTAL}**	**$\Delta\Delta G_{binding}$ (kcal mol^{-1})**
WT	5401	0.00
R164A	3190	4.31
S252A	4159	2.42
D254A	5267	0.26
E278A	5303	0.19
K280A	3204	4.28

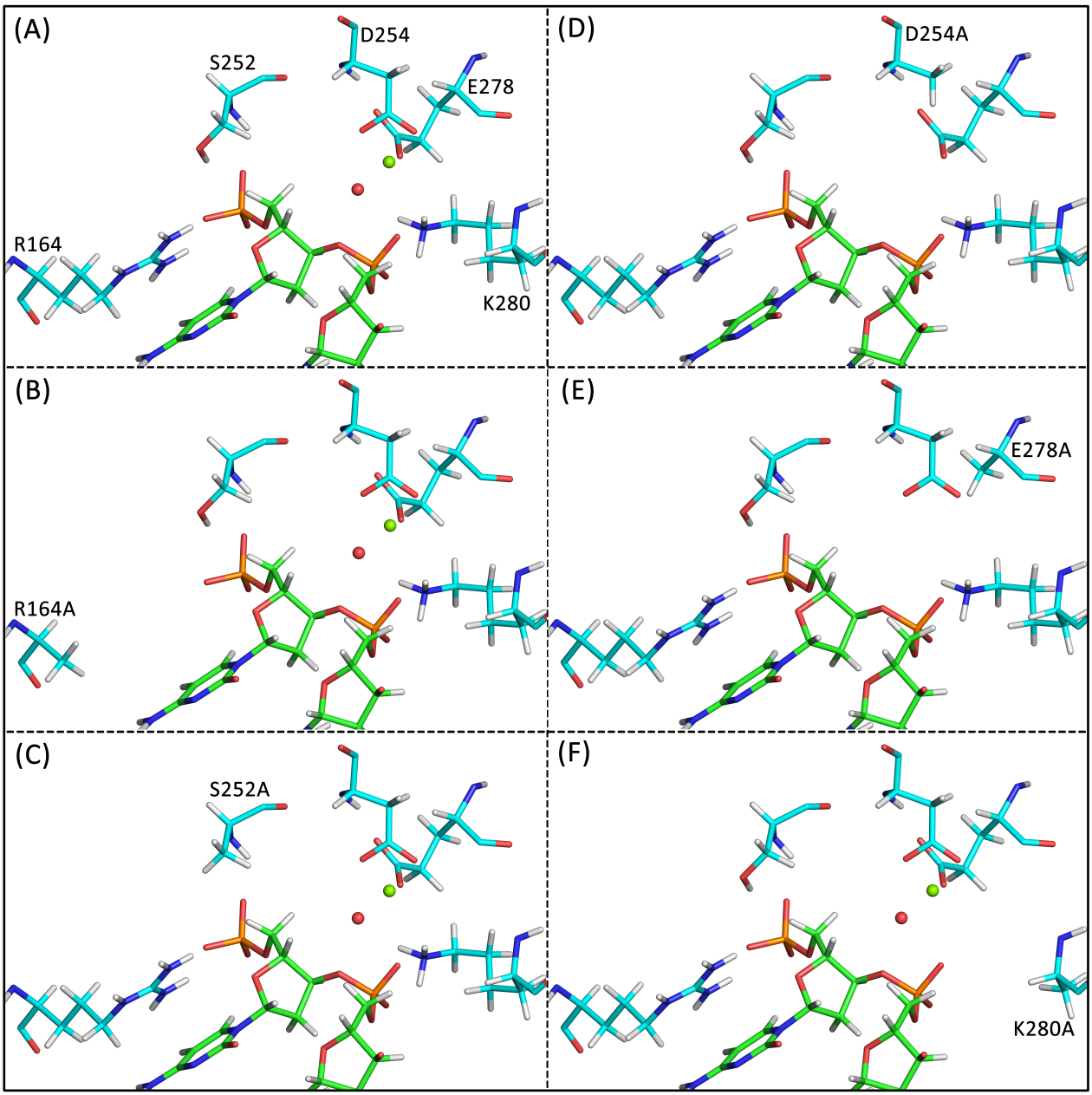


Figure 3.9 – Active site models for DNA interactions of wild-type and mutant UL98. Active site residues are shown as sticks with carbon atoms in cyan, the DNA is shown as sticks with carbons in green. Image prepared using PyMOL.

The region starting at Asp189 and ending at Phe211 in UL98 AN, the bridge region that connects the N-terminal domain to the C-terminal domain, was not modeled due to its absence in the template owing to disorder. There have been suggestions that the bridge architecture plays an important role in exonucleolytic cleavage by inducing strand separation.[46,66] The bridge region is also proposed to promote endonucleolytic processing by developing local regions of ssDNA within the substrate, which are the likely targets for endo cleavage. However, the high degree of disorder in the structure of the bridge region in both KSHV-SOX and the KSHV-SOX–DNA complex structures made it difficult to model the corresponding region in UL98. Thus, our modeling focused on the active-site region and cannot be extrapolated to the role of 'the bridge' in the functions of UL98.

3.3.3 *Validation of UL98 AN Model using Mutagenesis*

The experimental data are generally consistent with the modeling predictions. For details, please refer to our recent publication – Kuchta, A., *et al.*[47]

All of the single-alanine mutant proteins failed to digest substantial amounts of substrate DNA in a one hour period. Exo activity of the R164A mutant was 10.6% of wild-type and statistically higher than those of the other mutants and GUS (negative control); whereas activities for D254A, E278A, K280A and S252A were <5% that of wild-type and not statistically different from each other or from GUS (Figure 3.10).

For the D, E or K ➔ A mutants, endo activity was greatly reduced, but low level activity was still apparent with the supercoiled dsDNA substrate. Although D254A and E278A preparations were able to convert some of the supercoiled DNA into open-circular and full-length linear DNA, the majority of the supercoiled DNA substrate remained intact after the 12 hour incubation. In contrast, the K280A and S252A mutant proteins nicked essentially all of the supercoiled DNA, although large amounts of nicked open-circular and full length linear forms remained undigested. Following 12 hour incubation, the wild-type and R164A proteins were not only capable of nicking the supercoiled substrate, they had degraded all of the DNA substrate, presumably to products too small to visualize on the gel.

The quantitative fluorescence-based assay showed no significant difference between the endo activities of R164A and wild-type UL98 and both wild-type and R164A endo activities were significantly different from those exhibited by GUS and the other mutant proteins. The activities of S252A, D254A or E278A were not different from each other or from GUS (Figure 3.11).

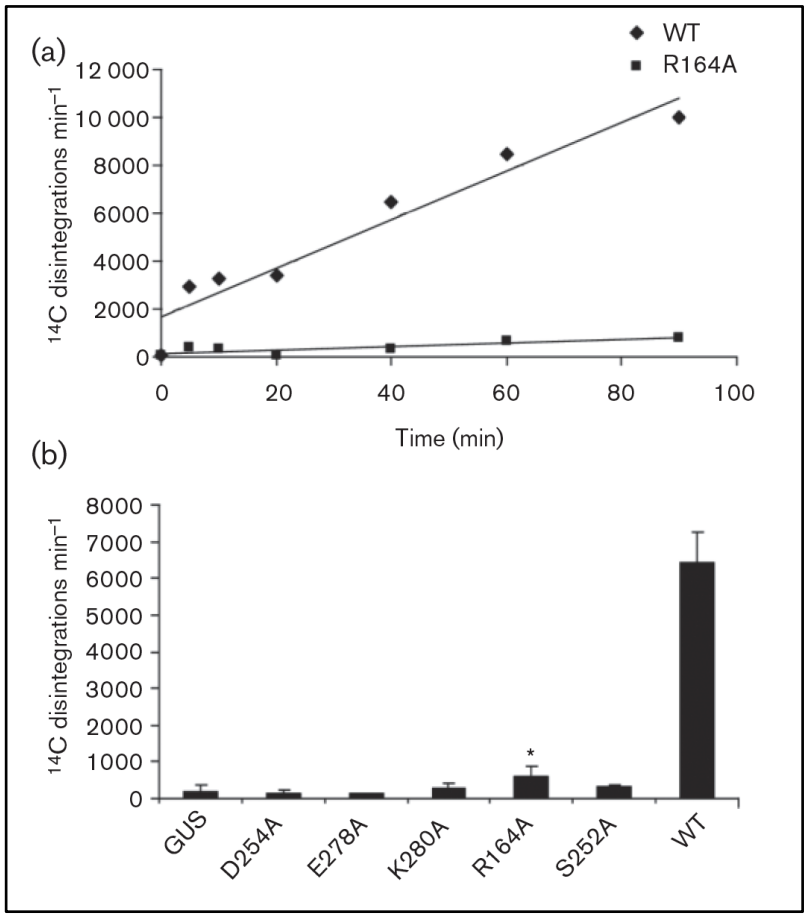


Figure 3.10 – UL98 AN mutants – Exo activity.
(a) Exo activities were determined as acid-soluble radioactivity released during incubation of 14C-labelled DNA with 2.5 µg of IMAC-purified UL98 WT or UL98 R164A. (b) Exo activities of mutants were measured after one hour incubation. Data are means of disintegrations per minute obtained from three experiments. Error bars represent means ±1 SD.
* R164A differs from GUS and other mutants; unpaired *t*-test, P ≤ 0.05
Reprint from ref 47.

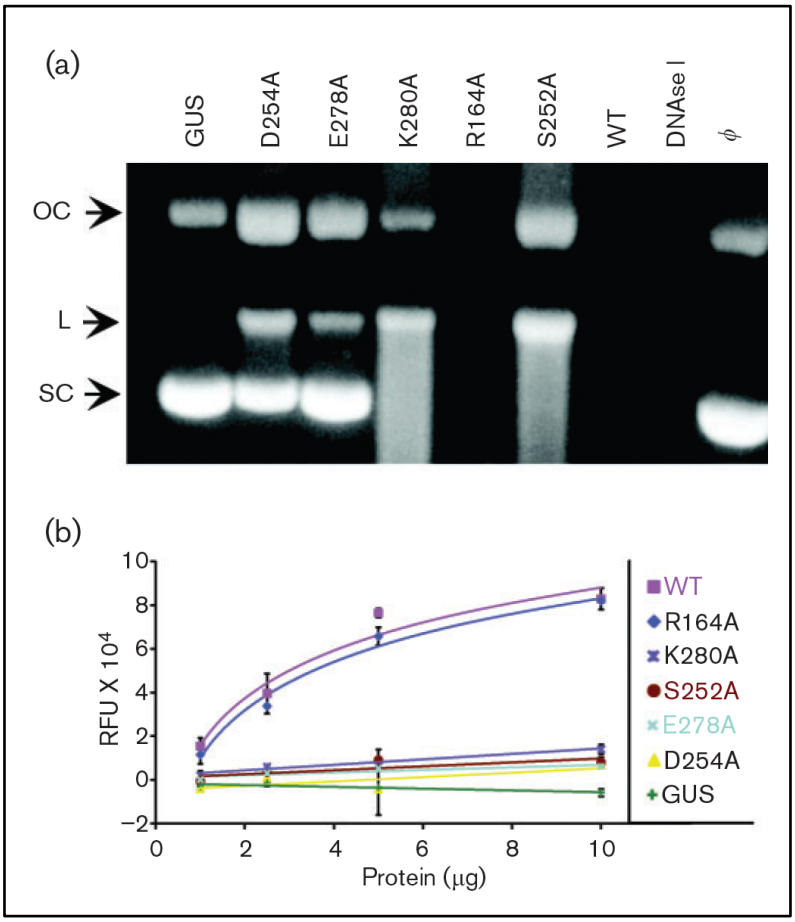


Figure 3.11 – UL98 AN mutants – Endo activity.
(a) Supercoiled plasmid DNA (250 ng) was incubated for 12 hours with buffer only (*ϕ*), 5 μg of IMAC-purified protein of 1 U of DNase I. Products were analyzed by agarose gel electrophoresis. Arrows indicate the positions of supercoiled (SC), open circular (OC) and linear (L) DNA. (b) A 35 nt synthetic ssDNA substrate with a fluorescent emitter at the 5' end and a quencher at 3' end was incubated for 14 hours with increasing amounts of each protein; fluorescence was measured as relative fluorescence units (RFU).
Reprint from ref 47.

The estimation of changes in free energies of association between UL98 AN and ds DNA upon mutation using HINT (refer section 3.3.1.7) predicts that D, E or K ➔ A substitution mutants should lack both exo and endo activity and that R or S ➔ A mutants may have impaired exo activity. As no 5' phosphate is present on the substrate in endonucleolytic cleavage, endo activity might be unaffected in R or S ➔ A mutants. Consistent with this, we observed experimentally that the D, E or K ➔ A mutants lacked measureable exo activity. Endo activity was greatly reduced, but low-level activity was still apparent with the supercoiled dsDNA substrate. Of the two UL98 mutations targeting putative phosphate-binding residues, R164A behaved as predicted. Exo activity was reduced by 90%, but there was no impact on endo activities detected with supercoiled dsDNA or end blocked ssDNA substrates. In contrast, the S252A mutation eliminated exo activity, and reduced both ssDNA and supercoiled dsDNA endo activities to levels similar to those of the K280A mutant. We believe that the proximity of the S ➔ A mutation, one residue from the catalytic aspartic acid, may have perturbed the local geometry of the active site enough to impair catalytic function, and hence to have impacted not only exo but also endo activity.

3.3.4 *3D Virtual Screening Hits*

We have performed a virtual screening study on the active site of UL98 AN for possible identification of its novel inhibitors. The model of UL98 AN complexed to dsDNA, validated by mutagenesis, identified important residues in the enzymes' active-site. Based on the interactions of the 5' phosphate group and the scissile phosphodiester group in the active site of UL98 AN, it was clear that negatively charged groups were desirable in the ligand. The model also helped predict other electronic and structural features that would appear to be necessary for a molecule to bind at the active site.

A pharmacophore model describing molecular features that are necessary for recognition by the enzyme was generated. One of the important features in the active site of the enzyme is the anionic site formed complementary to the side chains of R164 and S252, which accommodates the 5' end phosphate group of DNA. This site was defined as a "negative center" in the query. The second feature defined was a "donor atom" feature in order to look for a hydrogen-bond donor group in the ligand, capable of interacting with the carboxylate groups of D254 and E278. An "acceptor atom" was defined in the vicinity of K280, with an aim at engaging the amino group that is important for stabilizing the leaving group of substrate. The central hydrophobic aromatic core was defined to impart rigidity to the molecule. A larger region was defined so that compounds with bicyclic and tricyclic aromatic scaffolds can also be identified as hits. Also, an aromatic scaffold affords synthetic ease during compound modifications at

the stage of lead optimization. A distance constraint was set up between the two features for better defining the 3D shape of the cavity and to eliminate many undesirable compounds. Receptor site constraints were set up around the atoms surrounding the cavity, in order to avoid steric clashes. The compounds VDW volume must not intrude into these exclusion spheres to be considered a hit. The final three-dimensional query is shown in Figure 3.12.

The compounds in NCI Open Database were screened against the pharmacophore model using the “flex search” algorithm within UNITY. This algorithm generates all possible conformations of a given candidate structure, performs minimization in torsion angle space, and attempts to determine if it can reasonably flex into a conformation that matches the query.[67] The compounds were also checked against the pre-defined set of rules, Lipinski’s Rule of Five (that gives a molecule favorable permeation and absorption characteristics), *viz.* the molecule should have less than 5 hydrogen-bond donors and 10 hydrogen-bond acceptors, its molecular weight should be less than 500 D and its ClogP less than 5.

The process of virtual screening narrowed down the number of compounds from ~250,000 to a more manageable number of 72. All these ‘hits’ were docked into the active site of UL98 AN, and the docked poses were rescored using HINT. The best binding mode of each compound, based on HINT scores, was analyzed visually to identify a reasonable binding in the active site such that the catalytically important

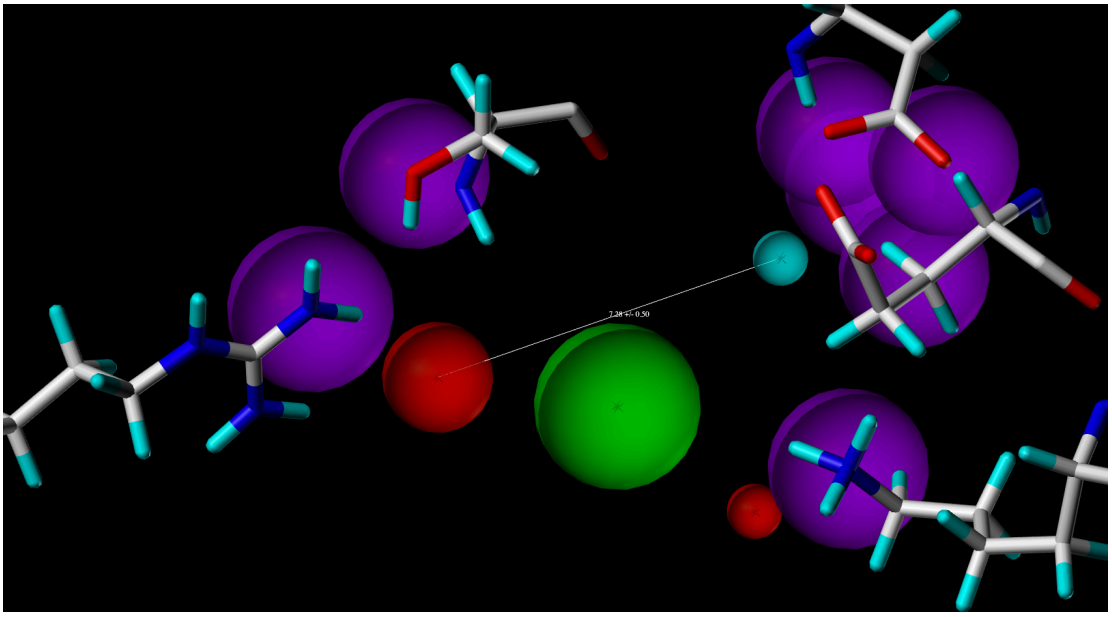


Figure 3.12 – Pharmacophore Model.
Negative-center (radius 1 Å) showed as red sphere, donor-atom (radius 0.5 Å) showed as cyan sphere, acceptor-atom (radius 0.5 Å) showed as red sphere, aromatic hydrophobic center (radius 1.5 Å) showed as green sphere. Distance constraint (±0.5 Å) set up between negative-center and acceptor-atom. Purple spheres are the receptor-site constraints.

residues would be engaged. Table 3.3 / Figure 3.13 shows the list of top 15 chemically diverse compounds identified as “hits”, along with their NSC numbers and HINT scores.

Figure 3.14 shows compound NSC 120634, which is the second best scored compound, within the active site of UL98 AN. Even though NSC 238165 had a higher score, NSC 120634 showed perfect fit into the pharmacophoric model, satisfying all the features defined to be desirable in a candidate structure. Also, its binding mode predicts interactions with all the catalytically important residues, D254, E278, K280, as well as those involved in 5’-phosphate-binding, R164 and S252.

All compounds listed in Table 3.3 were obtained from the NCI and are currently being evaluated for their exonuclease and endonuclease inhibitory activity. Suitable candidates shall also be evaluated for their antiviral activity in cell culture and for toxicity. Some of these compounds have HINT scores significantly higher than that of dsDNA in the active site (refer Table 3.2), and hence might be expected to be stronger binders of UL98 AN.

Table 3.3 – Top 15 hits from virtual screening (ranked based on HINT scores)

Rank	NSC #	HINT Score
1	NSC 238165	9665
2	NSC 120634	8171
3	NSC 175852	6639
4	NSC 342023	6259
5	NSC 60279	5975
6	NSC 37413	5283
7	NSC 37053	5210
8	NSC 163	4670
9	NSC 226640	3943
10	NSC 132073	3582
11	NSC 163091	3510
12	NSC 44630	2900
13	NSC 129478	2880
14	NSC 329204	2747
15	NSC 41439	2656

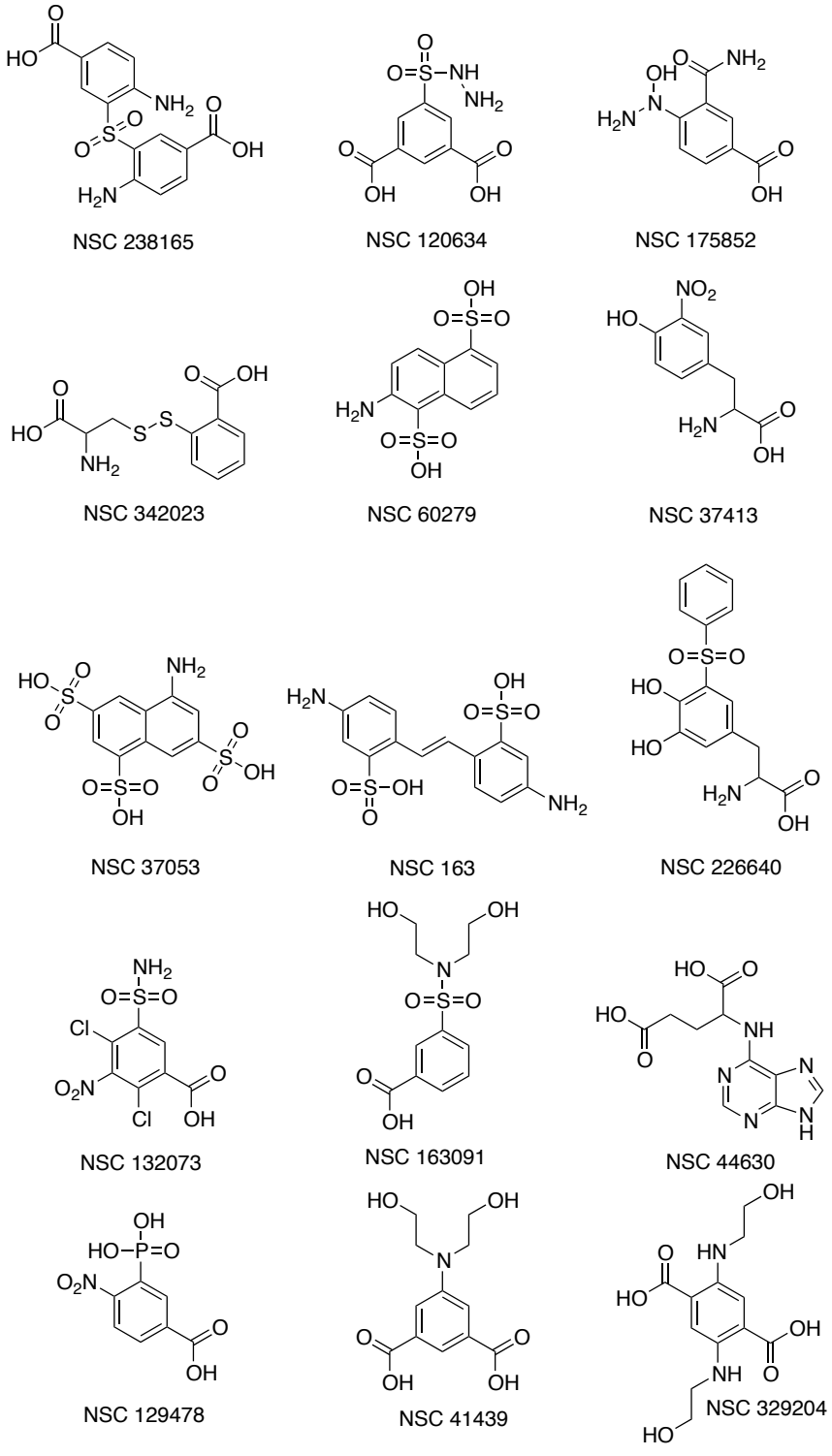
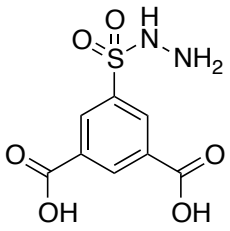


Figure 3.13 – Structures of virtual screening hits (top 15)



NSC 120634

5-(hydrazinylsulfonyl)isophthalic acid

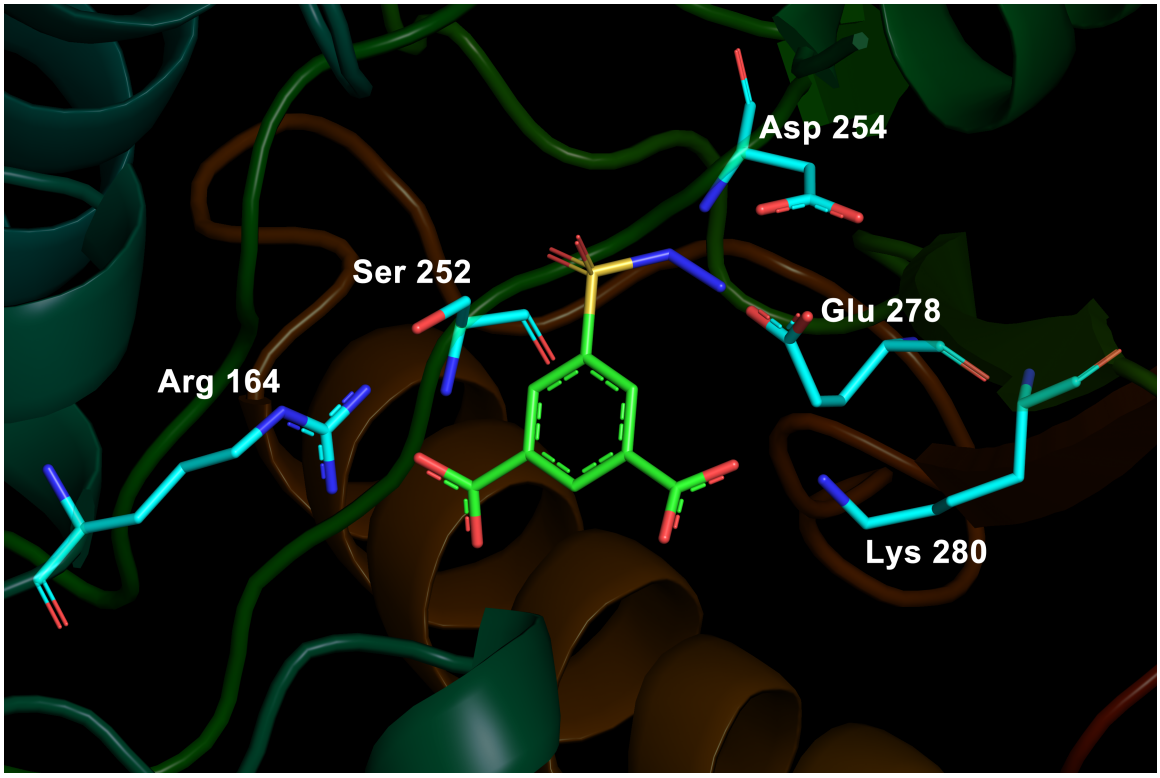


Figure 3.14 – Binding mode of *hit NSC120634*.
The carboxylates of isophthalic aromatic nucleus shows strong salt-bridge interactions with guanidine- and amino- groups of side chains of R164 and K280 respectively, occupying positions analogous to the phosphate groups of DNA substrate. The hydrazine moiety shows strong interactions with the carboxylate groups of D254 and E278.

3.4 Conclusion

In the absence of a crystallographic structure for UL98 AN, *in silico* modeling was used to successfully build a structural model of the protein that predicted the residues important for its nucleolytic functions. The predicted active-site residues – R164, S252, D254, E278 and K280, were validated by mutagenesis, where alanine scanning showed abolition of activity. Identification of residues involved in 5'-phosphate binding that affect the exo, but not the endo activity, may provide a useful tool for exploring the biological roles of each activity in herpesvirus replication.

The structural model provided a basis for performing rational drug discovery to identify novel antiviral agents for CMV infections. A virtual screening for active site inhibitors of UL98 AN was conducted on a pharmacophore, built based on our model, followed by molecular docking and HINT scoring. We have identified a number of novel scaffolds that have shown promise in our computational studies. These candidates are currently being evaluated experimentally.

References

1. Boeckh, M.; Geballe, A. P. Cytomegalovirus: pathogen, paradigm, and puzzle. *J. Clin. Invest.* **2011,** *121*, 1673-1680.
2. Alford, C. A.; Stagno, S.; Pass, R. E.; Huang, E. S. Epidemiology of cytomegalovirus. In *The Human Herpesviruses: An Interdisciplinary Perspective*, Nahmais, A.; Dowdle, W.; Schinazi, R., Eds. Elsevier: New York, 1981; pp 159-171.
3. Bate, S. L.; Dollard, S. C.; Cannon, M. J. Cytomegalovirus seroprevalence in the United States: The National Health and Nutrition Examination Surveys, 1988-2004. *Clin. Infect. Dis.* **2010,** *50*, 1439-1447.
4. Britt, W. Virus entry into host, establishment of infection, spread in host, mechanisms of tissue damage. In *Human Herpesviruses: Biology, Therapy and Immunoprophylaxis*, Arvin, A.; Whitley, R., Eds. Cambridge University Press: New York, 2007; pp 737-764.
5. Vochem, M.; Hamprecht, K.; Jahn, G.; Speer, C. P. Transmission of cytomegalovirus to preterm infants through breast milk. *Pediatr. Infect. Dis.* **1998,** *17*, 53-58.
6. Coonrod, D.; Collier, A. C.; Ashley, R.; DeRouen, T.; Corey, L. Association between cytomegalovirus seroconversion and upper genital tract infection among women attending a sexually transmitted disease clinic: a prospective study. *J. Infect. Dis.* **1998,** *177*, 1188-1193.
7. Ray, C. G.; Ryan, K. J. Herpesvirus. In *Sherris Medical Mocrobiology*, 5th ed.; Ray, C. G.; Ryan, K. J., Eds. McGraw-Hill: New York, 2010.
8. Boeckh, M.; Nichols, W. G.; Papanicolaou, G.; Rubin, R.; Wingard, J. R.; Zaia, J. Cytomegalovirus in hematopoietic stem cell transplant recipients: current status, known challenges, and future strategies. *Biol. Blood Marrow Transplant.* **2003,** *9*, 543-558.
9. Farman, J.; Lerner, M. E.; Ng, C.; Balthazar, E.; Megibow, A.; Herlinger, H.; Grimes, M. Cytomegalovirus gastritis: protean radiologic features. *Gastrointest. Radiol.* **1992,** *17*, 202-206.
10. Jabs, D. A.; Natta, M. L. V.; Holbrook, J. T.; Kempen, J. H.; Meinert, C. L.; Davis, M. D. Longitudinal Study of the Ocular complications of AIDS: 1. Ocular Diagnoses at Enrollment. *Opthalmol.* **2007,** *114*, 780-786.
11. Schleiss, M. R. Antiviral therapy of congenital cytomegalovirus infection. *Semin. Pediatr. Infect. Dis.* **2004,** *16*, 50-59.
12. Flower, K. B.; Boppana, S. B. Congenital cytomegalovirus (CMV) infection and hearing deficit. *J. Clin. Virol.* **2006,** *35*, 226-231.
13. Vancikova, Z.; Dvorak, P. Cytomegalovirus infection in immunocompetent and immunocompromised individuals - a review. *Curr. Drug. Targets Immune. Endocr. Metabol. Disord.* **2001,** *1*, 179-187.

14. Wreghitt, T. G.; Teare, E. L.; Sule, O.; Devi, R.; Rice, P. Cytomegalovirus infection in immunocompetent patients. *Clin. Infect. Dis.* **2003,** *37*, 1603-1606.
15. Rafailidis, P. I.; Mourtzoukou, E. G.; Varbobitis, I. C.; Falagas, M. E. Severe cytomegalovirus infection in apparently immunocompetent patients: a systematic review. *J. Virol.* **2008,** *5*, 1-7.
16. Boeckh, M.; Gooley, T. A.; Myerson, D.; Cunningham, T.; Schoch, G.; Bowden, R. A. Cytomegalovirus pp65 antigenemia-guided early treatment with ganciclovir versus ganciclovir at engraftment after allogeneic marrow transplantation: a randomized double-blind study. *Blood* **1996,** *88*, 4063-4071.
17. Paya, C.; Humar, A.; Dominquez, E.; Washburn, K.; Blumberg, E.; Alexander, B.; Freeman, R.; Heaton, N.; Pescovitz, M. D. Efficacy and safety of valganciclovir vs. oral ganciclovir for preventation of cytomegalovirus disease in solid oragn transplant recipients. *Am. J. Transplant.* **2004,** *4*, 611-620.
18. Martin, D. F.; Sierra-Madero, J.; Walmsley, S.; Wolitz, R. A.; Macey, K.; Georgiou, P.; Robinson, C. A.; Stempien, M. J. A controlled trial of valganciclovir as induction therapy for cytomegalovirus retinitis. *N. Engl. J. Med.* **2002,** *346*, 1119-1126.
19. Palestine, A. G.; Polis, M. A.; De-Smet, M. D.; Baird, B. F.; Falloon, J.; Kovacs, J. A.; Davey, R. R.; Zurlo, J. J.; Zunich, K. M.; Davis, M.; Hubbard, L.; Brothers, R.; Ferris, F. L.; Chew, E.; Davis, J. L.; Rubin, B. I.; Mellow, S. D.; Metcalf, J. A.; Manischewitz, J.; Minor, J. R.; Nussenblatt, R. B.; Masur, H.; Lane, H. C. A randomized, controlled trial of foscarnet in the treatment of cytomegalovirs retinitis in patients with AIDS. *Ann. Intern. Med.* **1991,** *115*, 665-673.
20. Prichard, M. N.; Kern, E. R. The search for new therapies for human cytomegalovirus infections. *Virus Res.* **2011,** *157*, 212-221.
21. Hochster, H.; Dieterich, D.; Bozzette, S.; Reichman, R. C.; Connor, J. D.; Liebes, L.; Snoke, R. L.; Spector, S. A.; Valentine, F.; Pettinelli, C.; Richman, D. D. Toxicity of combined ganciclovir and zidovudine for cytomegalovirus disease associated with AIDS: An AIDS clinical trials group study. *Ann. Intern. Med.* **1990,** *113*, 111-117.
22. Jacobson, M. A. Review of the toxicities of foscarnet. *J. Acquir. Immune Defic. Syndr.* **1992,** *5*, S11-17.
23. Ho, E. S.; Lin, D. C.; Mendel, D. B.; Cihlar, T. Cytotoxicity of anitiviral nucleotides adefovir and cidofovir is induced by the expression of human renal organic anion transporter 1. *J. Am. Soc. Nephrol.* **2000,** *11*, 383-393.
24. Klug, S.; Lewandowski, C.; Merker, H. J.; Stahlmann, R.; Wildi, L.; Neubert, D. In vitro and in vivo studies on the prenatal toxicity of five virustatic nucleoside analogues in comparison to aciclovir. *Arch. Toxicol.* **1991,** *65*, 283-291.
25. Sharland, M.; Luck, S.; Griffiths, P.; Cotton, M. Antiviral therapy of CMV disease in children. *Adv. Exp. Med. Biol.* **2011,** *697*, 243-260.
26. Hakki, M.; Chou, S. The biology of cytomegalovirus drug resistance. *Curr. Opin. Infect. Dis.* **2011,** *24*, 605-611.

27. Jabs, D. A.; Enger, C.; Dunn, J. P.; Forman, M. Cytomegalovirus retinitis and viral resistance: ganciclovir resistance. CMV retinitis and viral resistance study group. *J. Infect. Dis.* **1998,** *177*, 770-773.
28. Hoffmann, P. J.; Cheng, Y.-C. DNase induced after infection of KB cells by Herpes Simplex Virus type 1 or type 2. Characterization of an associated endonuclease activity. *J. Virol.* **1979,** *32*, 449-457.
29. Morrison, J. M.; Keir, H. M. A new DNA-exonuclease in cells infected with herpes virus: Partial purification and properties of the enzyme. *J. Gen. Virol.* **1968,** *3*, 337-347.
30. Martinez, R.; Sarisky, R. T.; Weber, R. C.; Weller, S. K. Herpes simplex virus type 1 alkaline nuclease is required for efficient processing of viral DNA replication intermediates. *J. Virol.* **1996,** *70*, 2075-2085.
31. Shao, L.; Rapp, L. M.; Weller, S. K. Herpes simplex virus 1 alkaline nuclease is required for efficient egress of capsids from the nucleus. *Virology* **1993,** *196*, 146-162.
32. Riozman, B. Genome variation and evolution among herpes viruses. *Ann. NY. Acad. Sci.* **1980,** *354*, 472-483.
33. Gao, M.; Robertson, B. J.; McCann, P. J.; O'Boyle, D. R.; Weller, S. K.; Newcomb, W. W.; Brown, J. C.; Weinheimer, S. P. Functional Conservations of the alkaline nuclease of herpes simplex type 1 and human cytomegalovirus. *Virology* **1998,** *249*, 460-470.
34. Sheaffer, A. K.; Weinheimer, S. P.; Tenney, D. J. The human cytomegalovirus UL98 gene encodes the conserved herpesvirus alkaline nuclease. *J. Gen. Virol.* **1997,** *78*, 2953-2961.
35. Adam, B.; Jervey, T. Y.; Kohler, C. P.; Wright, G. L.; Nelson, J. A.; Stenberg, R. M. The human cytomegalovirus UL98 gene transcription unit overlaps with the pp28 true late gene (UL99) and encodes a 58-kilodalton early protein. *J. Virol.* **1995,** *69*, 5304-5310.
36. Dunn, W.; Chou, C.; Li, H.; Hai, R.; Patterson, D.; Stolc, V.; Zhu, H.; Liu, F. Functional profiling of a human cytomegalovirus genome. *Proct. Natl. Acad. Sci. U S A* **2003,** *100*, 14223-14228.
37. Yu, D.; Silva, N. C.; Shenk, T. Functional map of human cytomegalovirus AD169 defined by global mutational analysis. *Proct. Natl. Acad. Sci. U S A* **2003,** *100*, 12396-12401.
38. Weller, S. K.; Seghatoleslami, M. R.; Shao, L.; Rowse, D.; Carmichael, E. P. The herpes simplex virus type 1 alkaline nuclease is not essential for viral DNA synthesis: ioslation and characterization of a lacZ insertion mutant. *J. Gen. Virol.* **1990,** *71*, 2941-2952.
39. Bujnicki, J. M.; Rychlewski, L. The herpesvirus alkaline exonuclease belongs to the restriction endonuclease PD-(D/E)XK superfamily: Insights from molecular modeling and phylogenetic analysis. *Virus Genes* **2000,** *22*, 219-230.

40. Kovall, R. A.; Matthews, B. W. Type II restriction endonucleases: structural, functional and evolutionary relationships. *Curr. Opin. Chem. Biol.* **1999,** *3*, 578-583.
41. Tsutakawa, S. E.; Muto, T.; Kawate, T.; Jingami, H.; Kunishima, N.; Ariyoshi, M.; Kohda, D.; Nakagawa, M.; Morikawa, K. Crystallographic and functional studies of very short patch repair endonuclease. *Mol. Cell* **1999,** *3*, 621-628.
42. Ban, C.; Yang, W. Structural basis for MutH activation in *E.coli* mismatch repair and relationaship of MutH to restriction endonucleases. *EMBO J.* **1998,** *17*, 1526-1534.
43. Kovall, R. A.; Matthews, B. W. Structural, functional, and evolutionary relationships between lambda-exonuclease and the type II restriction endonucleases. *Proct. Natl. Acad. Sci. U S A* **1998,** *95*, 7893-7897.
44. Steczkiewicz, K.; Muszewska, A.; Knizewski, L.; Rychlewski, L.; Ginalski, K. Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Res.* **2012,** *40*, 7016-7045.
45. Dahlroth, S.-L.; Gurmu, D.; Haas, J.; Erlandsen, H.; Nordlund, P. Crystal structure of the shutoff and exonuclease protein from the oncogenic Kaposi's sarcoma-associated herpesvirus. *FEBS J.* **2009,** *276*, 6636-6645.
46. Buisson, M.; Geoui, T.; Flot, D.; Tarbouriech, N.; Ressing, M. E.; Wiertz, E. J.; Burmeister, W. P. A bridge crosses the active-stie canyon of the Epstein-Barr virus nuclease with DNase and RNase activities. *J. Mol. Biol.* **2009,** *391*, 717-728.
47. Kuchta, A. L.; Parikh, H. I.; Zhu, Y.; Kellogg, G. E.; Parris, D. S.; McVoy, M. A. Structural modeling and mutagenesis of human cytomegalovirus alkaline nuclease UL98. *J. Gen. Virol.* **2012,** *93*, 130-138.
48. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997,** *25*, 3389-3402.
49. Altschul, S. F.; Wooron, J. C.; Gertz, M.; Agarwal, R.; Morgulis, A.; Schaffer, A. A.; Yu, Y.-K. Protein database seraches using compostitionally adjusted substitution matrices. *FEBS J.* **2005,** *272*, 5101-5109.
50. Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007,** *23*, 2947-2948.
51. Sali, A.; Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993,** *234*, 779-815.
52. Fiser, A.; Do, R. K.; Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000,** *9*, 1753-1773.
53. Marti-Renom, M. A.; Stuart, A.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000,** *29*, 291-325.

54. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M.-y.; Pieper, U.; Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **2007,** *50*, 1-31.
55. Shen, M. Y.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006,** *15*, 2507-2524.
56. Melo, F.; Sanchez, R.; Sali, A. Statistical potentials for fold assessment. *Protein Sci.* **2002,** *11*, 430-448.
57. Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B.; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007,** *35*, W375-W383.
58. Fornabaio, M.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 2. Computational tritration and pH effects in molecular models of neuraminidase-inhibitor complexes. *J. Med. Chem.* **2003,** *46,* 4487-4500.
59. Lipinski, C. A. Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discov. Today Technol.* **2004**, *1*, 337-341.
60. Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. *J. Med. Chem.* **2002,** *45*, 2469-2483.
61. Fornabaio, M.; Spyrakis, F.; Mozzarelli, A.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free enrgy of binding for protein-ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. *J. Med. Chem.* **2004,** *47*, 4507-4516.
62. Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.5.0.4.
63. Clamp, M.; Cuff, J.; Searle, S. M.; Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* **2004,** *20*, 426-427.
64. Waterhouse, A. M.; Procter, J. B.; Martin, D. M.; Clamp, M.; Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009,** *25*, 1189-1191.
65. Liu, M. T.; Hu, H. P.; Hsu, T. Y.; Chen, J. Y. Site-directed mutagenesis in a conserved motif of Epstein-Barr virus DNase that is homologous to the catalytic centre of type II restriction endonucleases. *J. Gen. Virol.* **2003,** *84*, 677-686.
66. Bagneris, C.; Briggs, L. C.; Savva, R.; Ebrahimi, B.; Barrett, T. E. Crystal structure of a KSHV-SOX-DNA complex: Insights into the molecular mechanisms underlying DNase activity and host shutoff. *Nucleic Acids Res.* **2011,** *39*, 5744-5756.
67. Hurst, T. Flexible 3D searching: The directed tweak technique. *J. Chem. Inf. Comput. Sci.* **1994,** *34*, 190-196.

CHAPTER 4

INCLUSION OF "RELEVANT" INTERFACIAL WATERS IMPROVE PROTEIN-PROTEIN DOCKING PREDICTIONS

4.1 Introduction

4.1.1 *Protein-Protein Interactions: Need for Computational Prediction Tools*

Protein-protein interactions play a fundamental role in most biological events and many pathological processes. Virtually every molecular process in a cell is carried out via interactions between two macromolecules; for example, DNA synthesis, gene expression, post-translational modifications, transport, signal transduction, etc. Various genetic, biochemical, or bioinformatics studies have identified tens of thousands of proteins interacting with each other forming millions of putative complexes. A detailed atomic understanding of the nature of these often-transient interactions is a key step to exploiting/inhibiting these biomolecular associations as potential new routes to disease therapeutics.

A large number of datasets exist that contain experimentally verified protein-protein interactions, like HPRD (Human Protein Reference Database),[1] BIND (Biomolecular Interaction Network Database),[2] MINT (Molecular Interactions Database),[3] etc. These databases contain a total of about 26,500 binary protein-protein interactions,[4] and provide a wealth of information pertaining to the human proteome. These data are related to thousands of protein-protein interactions, post-translational modifications, enzyme/substrate relationships, disease associations and more.

In contrast, the RCSB Protein Data Bank (http://www.rcsb.org/pdb/) only contains a few hundred protein-protein complex structures. One of the reasons for this lack of structural information is that experimental structural determinations using techniques like X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Electron Microscopy (EM) are very demanding.[5]

During the past couple of decades, there has been a rapid emergence of novel computational algorithms to predict, model and understand these interactions – also known as protein-protein docking methods. These predictive methods have been very essential for progress, since the experimental structural determination techniques, although powerful, definitely have low throughput. Accurate and reliable computational predictions are very useful in inferring how two proteins bind, give valuable functional information about the interacting proteins and also help guide new genetic and biochemical experiments. A number of docking algorithms and different scoring functions have been developed in recent times, and this endeavor has gained wide popularity as seen by the CAPRI (Critical Assessment of PRedicted Interactions), a communitywide experiment.[6] Wodak and Janin were the first to develop a predictive

algorithm, in the late 1970s that generated possible orientations of one protein relative to another.[7]

4.1.2 *Protein-Protein Docking: The Process*

As the field developed, protein-protein docking algorithms have become more sophisticated partly due to the rapid progress in computer hardware, but also due to our improved understanding of structure and interactions. The docking protocols have evolved from simple *rigid-body* docking (where both interacting partners are considered as rigid solid bodies), to *soft-body* docking (which incorporates side-chain and backbone flexibility in either one or both molecules), to inclusion of short MD simulations (to obtain an induced-fit model), to inclusion of explicit solvent models within their docking protocols; all of which are probably directly attributed to the availability of increased computational power.

The process of docking macromolecules is multi-step, involving accurate representation of the system, conformational space search, protein flexibility upon association and ranking of potential solutions.[8]

In most of the current docking approaches, the description of the protein surface is the atomic representation of its solvent-exposed residues, using mathematical models such as geometric shape descriptors like the Connolly surface.[9] The geometric descriptors that accurately represent the maxima (holes) and the minima (knobs) of the shape function[10,11] are usually combined with other computed properties designed to have physicochemical meaning like the affinity grids that are calculated based on the force field potentials for Van der Waals and electrostatic interactions.[12] Precise

representation of the system is followed by a search of the available conformational space. The majority of programs employ an *"ab-inito"* approach: one of the docking partners (usually the larger one) is fixed in space and the second protein is rotated and translated around the fixed one. Different search methods have been used: matching surface complementarity at the protein-protein interface;[13] combining geometric complementarity with amino acid pairwise affinities at the interface;[14] Fourier correlation techniques such as the FFT algorithm, first described by *Katchalski-Katzir et al.*,[15] which evaluates the interface contacts between binding partners while penalizing protrusions into the protein core. Other algorithms like geometric hashing,[10,16-18] Genetic algorithms,[19] Brownian dynamics simulations,[20] and simulations combined with energy minimization[21] have also been applied to the docking problem.

The induced fit model suggested by Koshland shows that conformational changes could occur upon binding.[22] A variety of studies followed that supported this theory of flexibility. It is impractical to treat molecular flexibility in an explicit way in the case of protein-protein docking, due to the large number of atoms and degrees of conformational freedom involved. However, a number of approaches have tried to address this shortcoming. A recent review by *Andrusier, et al.* describes how protein flexibility is treated during different stages of the docking process.[23] Some docking protocols adopt a two-stage approach that combines rigid-body search with molecular dynamics to account for backbone and/or side-chain flexibility. Side-chain optimization has been shown to discriminate near-native conformations from false positives.[24] Another algorithm, SOFTSPOTS identifies interfacial residues most likely to undergo conformational change at an interface, and generates the corresponding rotamers,

before the docking calculations.[25] Incorporating full backbone flexibility is highly challenging, because too large a conformational change in a backbone may lead to deformation of global structure. Various docking methods handle backbone flexibility differently – some methods utilize "soft docking" protocols that initially allow steric clashes, followed by a refinement step; some perform ensemble docking, using different conformations generated *a priori*; some methods deal with hinge-bending motions, while other methods perform a wide conformational space search to identify energetically the most favorable one.[23]

Comprehensive search algorithms generate a huge number of potential solutions, from which the one corresponding to the lowest free energy of binding must be found. An ideal scoring function should be able to distinguish between native-like predictions from false-positives. During the complex prediction step, the degree of shape complementarity of the interacting protein surfaces is used as an initial filter to eliminate incorrect predictions; but that alone is not sufficient to take into account the complete energetics of protein-protein associations. In most of the algorithms developed so far the initial filtering is then followed by ranking the predicted solutions using scoring functions that take into account geometric complementarity, electrostatic interactions, hydrogen bonding and/or desolvation energy.[26-28] Most scoring functions are designed to predict the free energy of binding $\Delta G_{binding}$, which is not a trivial task, since the individual components within them are imperfectly able to completely characterize the biomacromolecular association process.[12] Despite efforts to identify the correct binding modes using free energy as a reliable guide, scoring still remains a major challenge in the docking process.

4.1.3 *Current Protein-Protein Docking Algorithms*

With the emergence of novel docking approaches to predict biomacromolecular associations, a community-wide experiment to evaluate their capabilities was designed – called CAPRI (Critical Assessment of PRedicted Interactions).[6] CAPRI is a data-driven blind experiment wherein participating groups submit predictions of a target complex using their docking algorithms. The experiments would start when an unpublished X-ray crystal structure or NMR structure of target protein-protein or protein-DNA complex is made available by experimentalists to the CAPRI management. The atomic coordinates of the interacting partners are provided to participating structural biologists, who within 4-6 weeks, submit a set of 10 models that are compared to the experimental structure. The quality of the submitted models is evaluated based on the standard CAPRI criteria (that will be discussed in detail in Section 4.2.4) and classified as models of "high", "medium" and "acceptable" accuracy.[29]

The most recently concluded 4th CAPRI Evaluation Meeting held at Mare Nostrum, Barcelona (2009) evaluated rounds 13-19 that took place over a period of 2007-2009 and comprised a total of 14 targets. 76 participating groups submitted a total of 4420 docking predictions.[30] Rounds 1-2 of CAPRI with a total of 7 targets were evaluated in 2002, rounds 3-5 with a total of 10 targets were evaluated in 2005, and rounds 6-12 with a total of 9 targets were evaluated in 2007.[6,29,31] Each evaluation round has been well reported in special issues of the journal *Proteins: Structure, Function and Bioinformatics (Proteins 2003:52; Proteins 2005:60 and Proteins 2007:69)*.

The comparison of various docking programs based on their relative performance in the CAPRI experiments over the last decade is very difficult since the

algorithms differ in the methods used to perform the conformational search and score the predictions, and may perform better for different types of targets. A recent review on protein-protein docking by *Moreira, I. S. et al.*[12] analyzed the performance of the participants by giving quantitative measures to the quality of submitted models for each target.

Globally, ICM-DISCO,[32] ZDOCK,[28] HADDOCK[33,34] and RosettaDock[27] have been the best predictors over the past decade. The same review article also compares the popularity of these programs based on the citations per year, which suggests that the most popular ones are HADDOCK,[33,34] RosettaDock,[27] ClusPro[35] followed by PatchDock[36] and ZDOCK.[28] Table 4.1 summarizes the software characteristics, along with some advantages and disadvantages, of the top 5 programs (based on their performance in the CAPRI experiments and software popularity).

4.1.4 *Solvated Docking*

Water is a vital component in all living organisms and plays a crucial role in all biological processes. Particularly for proteins, the dynamics of water-protein interactions govern various molecular phenomena – like protein folding and molecular recognition,[37] as well as maintenance of structural integrity.[38] A water molecule can act both as a donor and as an acceptor of hydrogens, capable of forming four directional hydrogen bonds. This allows for easy and rapid reorientation and reconfiguration into different three-dimensional structures. Due to this unique property of water, the strongly bound or

Table 4.1 – Protein-Protein Docking Softwares: Characteristics, advantages and disadvantages (adapted from ref 12)

Software	Conformational Search Algorithm	Filtering at search stage	Flexibility	
			Search stage	Refinement stage
ICM - DISCO (Docking and Interface Side-Chain Optimization)	Rigid Body Docking; pseudo-Brownian Monte Carlo minimization	Specific filtering criterion on a case-by-case basis	–	Fully-flexible interface ligand side-chains
ZDOCK	Rigid body search using Fast Fourier Transform (FFT) algorithm	Biological data driven-docking. Allows definition of interfacial and blocking residues	–	Optimizes full atoms internal energy and vdW
HADDOCK (High Ambiguity Driven protein-protein DOCKing)	Three stage process - (i) randomization of orientations and rigid-body energy minimization (EM), (ii) semi-rigid simulated annealing in torsion angle space (TAD-SA), (iii) refinement in Cartesian space with explicit solvent	Data driven docking - Use of chemical shift perturbation data and NOEs/RDCs data from NMR experiments	Interfacial side-chain flexibility	Side-chain and backbone flexibility in the simulated annealing and minimization stages
RosettaDock	Rigid-body Monte Carlo search followed by minimizations	–	Side chain minimizations	–
ClusPro	Rigid body search using Fast Fourier Transform (FFT) algorithm	Filtering using empirical free energy functions	–	–

Table 4.1 – Protein-Protein Docking Softwares: Characteristics, advantages and disadvantages (cont.)

Software	Scoring		Advantages
	Search stage	Refinement stage	
ICM - DISCO (Docking and Interface Side-Chain Optimization)	Truncated vdW potential, Electrostatic potential corrected for solvation, hydrogen-bonding potential, hydrophobicity potential	Truncated vdW potential, Electrostatic potential corrected for solvation, hydrogen-bonding potential, hydrophobicity potential, rotamer probability	Global procedure, fully-automated, handles induced changes of interface side-chains
ZDOCK	Pairwise shape complementarity (PSC)	PSC, desolvation and electrostatics	Performed effectively for antibody-antigen test cases in CAPRI experiments
HADDOCK (High Ambiguity Driven protein-protein DOCKing)	Clustering, based on intermolecular energies	Average interaction energies (Sum of electrostatic potential, vdW potential, ambiguous interaction restraints AIR derived from experimental information available) and Average buried surface area	Side chain and backbone flexibility, use of experimental data restraints narrows the space search to relevant regions
RosettaDock	Residue-residue interaction potential	van der Waal's potential, desolvation potential, hydrogen-bond potential, electrostatics	Protocol mimics the physical process of docking, with refinement stage optimizing the interfacial side-chain packing
ClusPro	Shape complementarity, desolvation and electrostatics	Shape complementarity, desolvation and electrostatics	Fully-automated program, performs docking, filtering and scoring rapidly

Table 4.1 – Protein-Protein Docking Softwares: Characteristics, advantages and disadvantages (cont.)

Software	Disadvantages	Web address	References
ICM - DISCO (Docking and Interface Side-Chain Optimization)	Less successful for cases with significant backbone rearrangements	http://www.molsoft.com/docking.html	32
ZDOCK	Ineffective for cases with large conformational change	http://zlab.bu.edu/zlab/index.shtml	28
HADDOCK (High Ambiguity Driven protein-protein DOCKing)	(i) Ineffective in cases without additional experimental data, (ii) highly dependent on accuracy of biological information available *a priori*	http://www.nmr.chem.uu.nl/haddock/	33, 34
RosettaDock	Less successful for cases with significant backbone rearrangements	http://graylab.jhu.edu/docking/rosetta/	27
ClusPro	Cannot introduce additional information to drive correct docking	http://cluspro.bu.edu/home.php	35

"conserved" waters (those that are consistently observed in several crystallographic structures and not easily displaced by ligands) are capable of modifying the protein surface properties like its shape and charge.

Even though *Bogan and Thorn (1998)* proposed an O-ring hypothesis for interfaces claiming that occlusion of solvent by hot spot residues is found to be a necessary condition for energetically favorable interactions,[39] the abundant presence of water at protein-protein and protein-DNA interfaces underscores the vital role played by water in the polar interactions that stabilize the complexes.[40] *Jannin's* closer examination of structural data on protein-protein and protein-DNA recognition sites revealed that the associated interfaces contain at least as many water-mediated interactions as direct hydrogen bonds or salt bridges.[40]

Some research groups have made efforts to incorporate water-molecules in protein-ligand docking protocols, both explicitly and implicitly. The very first hurdle in this effort is the determination of bound water molecules at the ligand binding site. Several methods available for identifying/predicting protein-ligand interfacial waters have shown promise. GRID[41] performs well in predicting ligand-binding site water molecules, by calculating interaction energy using Lennard-Jones potential, electrostatic and hydrogen-bond terms. AQUARIUS,[42] a knowledge-based approach, identifies water sites in proteins from the experimentally generated electron density maps. CS-Map predicts the most favorable binding position of water molecules on protein surface based on an interaction potential that accounts for van der Waals, electrostatic and solvation contributions.[43] The Fold-X force field allows the prediction of positions of bound water molecules that show interaction with two or more polar atoms of proteins.[44]

A more recently available tool WaterMap,[45,46] predicts active-site bound water molecules by solvating the site and calculating its thermodynamic properties. A few protein-ligand docking programs like Flex-X,[47] Autodock,[48] GOLD,[49] and GLIDE[50,51] have shown significant improvements in docking performances by developing algorithms to include contributions from interfacial waters.[52]

All these approaches have shown a lot of promise in protein-ligand systems; however, water has been *neglected* in almost all protein-protein docking algorithms. Most of the development in solvated protein-protein docking has been focused on implicit treatment of solvent molecules, as it reduces the computational cost associated with explicit treatment. *Chen et al.*[53] have reviewed the progress from in-vacuo to in-solution docking, using implicit solvent-based methods. Although promising, a more detailed understanding of a protein-protein complex interface can perhaps be achieved from explicit treatment of water molecules. HADDOCK is one of the few docking programs designed to account for explicitly added water molecules in the docking process. Its solvated docking protocol starts with hydrating individual protein molecules, followed by rigid-body docking process resulting in a water layer in between the two proteins. All non-interfacial water molecules are then removed and a fraction of resulting interfacial waters is subsequently removed in a biased Monte Carlo procedure based on water-mediated contact probabilities. This methodology resulted in noticeable improvements both in quality and scoring than unsolvated docking, for most of the 10 studied cases that included examples of both wet and dry interfaces.[54]

4.1.5 *Explicit Hydropathic Approach*

Our lab, led by *G. E. Kellogg*, along with our collaborators have been interested in understanding the energetic contribution of water molecules in various biological environments. The empirically derived HINT forcefield (described in detail in Section 1.5), which models both hydrophobic and polar non-covalent interactions between two molecules, forms the basis for our analyses.

In a study on five *β*37 mutant hemoglobin crystal structures, *Burnett et al.* showed the contribution of crystallographically important water molecules in the dimer-tetramer assembly.[55] *Fornabaio et al.* developed an approach based on HINT energy function to map the energetics of water-protein and water-ligand interactions at protein-ligand interfaces. They analyzed the protein-ligand interactions in the active site of 23 HIV-1 protease-ligand complexes and showed significant improvement in correlation between HINT scores and experimentally determined binding constants when appropriate bridging water molecules are taken into account.[56]

In another study by *Amadasi et al.*, protein-water and water-ligand interactions in the binding site of sets of uncomplexed and ligand-complexed proteins were evaluated using the HINT forcefield.[57] Also, each water molecule was scored using the Rank algorithm,[58] which assigns a higher rank to a water molecule that is capable of a maximum of four hydrogen-bonds (≤2 donors and ≤2 acceptors). The HINT free energy scoring model and the Rank algorithm were combined to develop a statistically validated Water Relevance Metric[59] that classifies water molecules in protein binding sites that are generally conserved (between unliganded and ligand-bound states) as “Relevant” waters. These high Relevance waters are not likely to be displaced by the

ligand and should be explicitly considered when building geometrically and functionally correct models of the binding site. This also has implications in structure-based drug design, as it helps to identify key polar interactions within a protein's binding site that can be utilized to design more potent ligands with polar functional groups capable of mimicking water's hydrogen bonds.

Spyrakis et al. calculated HINT interaction scores for 39 crystallographic protein-DNA complexes, taking into account the contributions from interfacial waters that act as linkers between amino acid side-chains and nucleotide bases. The study quantified the key energetic role of bridging waters in protein-DNA associations.[60] Recently, *Ahmed et al.* performed a comprehensive study on the role of bound water at protein-protein interfaces.[61] Analysis of a total of 4741 water molecules at the interface of 179 heterodimeric protein-protein complex crystal structures revealed that 21% of the bound water is involved in bridging interactions with both proteins. Their analysis also showed that the total energetic contribution of bridging water ranges up to -11.35 kcal mol^{-1} per protein pair. Another, more subtle, role that these bridging waters serve at the interface is act as nano-scale pH buffers owing to their ability to easily swap between acting as hydrogen-bond donors and acceptors and thus maintain the integrity of the interface. This comprehensive study emphasized the importance of characterizing the behavior of biological waters as their presence at the interface may influence the assembly of biomacromolecular complexes, and begins to establish a basis for including the effects of individual waters in macromolecular docking algorithms.

4.1.6 *Specific Aims*

Protein-protein interactions involved in various biological pathways can be exploited as novel targets for rational drug discovery. A detailed atomic level understanding of interactions at protein-protein interfaces is crucial to identify hot-spot residues guiding their recognition. In the absence of experimentally determined 3D structure of a protein complex, a docking algorithm aims at predicting it starting from atomic resolution structures of the individual components. Although the challenges remain significant, different tools for protein-protein docking have been reasonably successful at modeling biomolecular associations, as seen from the recent CAPRI evaluation.[30] Overall, 67% of the participating groups produced acceptable models for at least one target. No evident correlation was seen between the ranks of models and their accuracy,[62] underscoring again the weakness of current scoring function methodology.

Several studies in our lab, including Chapter 2 and Chapter 3 of this work, have successfully shown the utility of the HINT paradigm in calculations of free-energy of binding for protein-ligand complexes[56,63,64] as well as protein-DNA complexes.[65] The Computational Titration (CT) algorithm, based on HINT force field, is capable of exploring the protonation states of protein active site residues and ligand functional groups.[66,67] The Water Relevance metric[59] has been shown to accurately (92% in cases with ≤2.0 Å resolution) predict the conserved water molecules at protein-ligand interfaces based on only the unliganded protein structure. It is our current long-range goal to incorporate HINT force field along with these tools into a protein-protein docking algorithm. While the process of docking is principally a two-stage search and score problem, the HINT-based tools can initially be applied easily in the latter stage of

complex refinement and scoring. We believe that a protein-protein docking algorithm accounting for not only the physical effects of biomolecular association like shape complementarity and residue flexibility, but also the associated chemical effects like hydropathic complementarity, correct residue ionization states and explicit inclusion of interfacial water molecules will yield more realistic models, but these effects must be incorporated during the search stage and will be reported in other reports from our laboratory. The current work addresses the issue of explicit solvent accounting during a protein-protein docking process.

In this study, we investigate the effect of bridging waters on docking performance. ZDOCK, a rigid-body protein-protein docking program, was used for the purpose of this study, as it has consistently been a top performer at the CAPRI experiments.[12] The main aim of this project was to check the influence of explicit water accounting on the accuracy of predictions, and not necessarily to improve upon the current ZDOCK algorithm. Interfacial waters relevant to both interacting partners were identified using the HINT Relevance metric. We forced ZDOCK to include these waters as atoms in one of the two interacting proteins, and show that more accurate results are obtained when water is ***not*** ignored.

4.2 Methods

4.2.1 *Data Set*

A non-redundant benchmark for protein-protein docking algorithms, which contains test cases for which the 3D structures of the complex and both unbound components are available, was developed by *Weng's* group at University of Massachusetts.[68] To obtain a set of well-structured interfacial waters, the data set for this study was limited to those cases for which the bound complex resolution is ≤ 2.0 Å.

The coordinates for all the complexes in our data set were obtained from the Protein Data Bank (http://www.pdb.org/). First, all ligands and cofactors were deleted from each complex structure using SYBYL v8.1 (TRIPOS, Inc.). For cases with "multi"-mer assembly, only one chain of each component forming the complex was retained. Hydrogen atoms were added and minimized under the Tripos force field (1000 iterations, 0.01 kcal mol^{-1} $Å^{-1}$ gradient, Gasteiger-Hückel charges), keeping the coordinates of all heavy-atoms fixed. Interfacial waters, those that are within 4 Å from atoms on both interacting proteins, were retained with each protein-protein complex. For each test case, the larger protein of the two was defined as the "receptor" protein, which would be kept fixed during the docking process, and the smaller one as the "ligand" protein.

4.2.2 *Determination of "Bridging" Interfacial Waters* –

Intermolecular interaction score was calculated between each receptor-ligand pair using HINT scoring function.[69] In principle, the HINT model scores each atom-atom interaction (b_{ij}) between atoms i and j using –

$$\boldsymbol{H_{TOTAL} = \Sigma_i \Sigma_j b_{ij} = \Sigma_i \Sigma_j \left(a_i S_i a_j S_j T_{ij} R_{ij} + r_{ij} \right)} \quad \textbf{(1)}$$

where, *a* is the hydrophobic atom constant, *S* is the solvent accessible surface area, T_{ij} is a logic function assuming +1 or -1 value depending on the character of the interacting polar atoms, and the distance dependent functions R_{ij} and r_{ij} are simple exponential function e^{-r} and an implementation of the Lennard-Jones potential function,[70,71] respectively. A direct HINT interaction score was calculated for every protein-protein complex, without accounting for the contributions made by interfacial waters. The HINT parameters and controls used were similar to those in the previous studies[57,60,65] – the protein molecules were partitioned using the *dictionary* method, with *essential hydrogen* treatment (where polar hydrogens are treated explicitly and non-polar ones are treated implicitly), and a 30 $Å^2$ correction used for calculations of the *S*-values for backbone amide nitrogens.

Next, the crystallographic orientation of every water molecule at the interface was optimized using an algorithm that performs an exhaustive search of its orientation space to assign H-atom positions.[72] This algorithm, developed around the HINT empirical force field, treats every water molecule as a 'small-ligand' and the surrounding atoms (within 6-8 Å) from both proteins as its 'binding-site'. HINT scores are calculated between the water molecule and its surroundings, allowing rotation of H-atoms around the three axes and limited translation of O-atom centroid with an aim at maximizing the interaction score.

The Rank algorithm[72] was applied to the optimized water molecules. Ideally, a water molecule is capable of forming a maximum of four hydrogen-bonds (≤2 donors and ≤2 acceptors) with its surrounding atoms. Rank represents the weighed number of

potential hydrogen bonds that each optimized water molecule forms, and is calculated as shown

$$\boldsymbol{Rank = \Sigma_n\{(2.80\ \text{Å}/r_n) + [\Sigma_m\ cos(\theta_{Td} - \theta_{nm})]/6\}} \quad (2)$$

where, r_n is the distance between the water O-atom and target heavy atom atom n (n = 1 to number of valid targets), θ_{Td} is the ideal tetrahedral angle (109.5°) and θ_{nm} is the angle between targets n and m (n = m to number of valid targets). The Rank algorithm yields values ranging from 0 for waters that do not form any hydrogen bonds with non-water molecules to about 6 for waters forming four high quality hydrogen bonds with excellent bond length and bond angle geometry.

In order to classify the water molecules, its HINT score and Rank were combined to give a Relevance value.[59] The Relevance of a water molecule is calculated using the weighed probability equation –

$$\boldsymbol{P_A = \frac{P_R(|W_R| + 1)^2 + P_H(|W_H| + 1)^2}{(|W_R| + 1)^2 + (|W_H| + 1)^2}} \quad (3)$$

where, P_A is the overall probability or the "Relevance" value for a water molecule, P_R and P_H are the percent probabilities for water conservation based on Rank and HINT score, and W_R and W_H are the weights for these probabilities, respectively. A water molecule at a protein interface with $P_A \geq 50\%$ is considered "conserved" water, meaning it would be present in ligand-bound complex. This water Relevance metric, although trained on protein-ligand complexes, was extended to protein-protein complexes in order to identify the waters contributing towards bridging interaction, as we showed earlier.[61]

The Relevance value for interfacial waters in all test cases was calculated using the above described model. We propose that an interfacial water molecule that is involved in bridging interactions should be Relevant to both proteins. As previously used,[61] our criteria for considering a water molecule to be "truly bridging" was a Relevance score of ≥ 0.25 with respect to both proteins (thus giving it a total value of ≥ 0.5). With this definition, the HINT scores, Rank and Relevance scores for interfacial waters in every test case was calculated.

4.2.3 *Solvated Docking using ZDOCK* –

For the present study, ZDOCK v3.0.2 (that incorporates a 3D convolution library to improve its efficiency) was obtained from http://zdock.umassmed.edu/software/. A total of 100 solutions were generated for each receptor-ligand pair in the data set. Since bound-bound docking was performed, a seed integer was specified for randomization of the starting coordinates for ligand structures. Also, rotational sampling was set to dense, which means the rotational search was performed in 6° steps. The receptor protein coordinates were fixed, preventing its rotation or switching with ligand during execution.

Using these parameters, two different docking protocols were evaluated on every case in the data-set of 15 protein-protein complexes:

1. Unsolvated Docking: Standard rigid-body docking, absence of interfacial water molecules.
2. Solvated Docking: Rigid-body docking, explicit inclusion of "bridging" water molecules identified using the HINT-based water Relevance metric.

For solvated docking, the interfacial bridging water molecules identified were added to the receptor file, and considered as a part of the protein. Since the ZDOCK program is not parameterized to include explicit waters in its algorithm, the ACE type, atom radius and atom charge for each water molecule was manually updated in the input file.

The ZDOCK output file provides information regarding the rotation and translation for the ligand with respect to its initial positioning. The protein-protein complex file was generated for each prediction; hydrogens were added and subjected to minimization under the Tripos force field (1000 iterations, 0.01 kcal mol^{-1} $Å^{-1}$ gradient, Gasteiger-Hückel charges).

HINT interaction scores for unsolvated docking were calculated between the receptor protein and ligand protein for each prediction. In the case of solvated docking, the water molecules at the interface were first optimized using the water-optimization algorithm, as described earlier, followed by the HINT interaction score calculation that accounts for contributions from interfacial waters. For the purpose of comparison, the HINT scores were normalized with respect to the top HINT score for each case. The HINT scores for each prediction were represented as a fraction of top HINT score for each individual test case. The predictions were then ranked based on their scaled HINT scores.

4.2.4 *The Assessment Protocol* –

A standard CAPRI assessment criteria was used to evaluate the predictions against the target crystallographic structures.[29] A number of different characteristics of

predicted complexes were evaluated – not just the root mean square deviations (interfacial and ligand), but also the identification of correct residue-residue contact pairs, which is extremely important for inhibitor design. Figure 4.1 and 4.2 shows the schematic illustration of the three different parameters used to assess the quality of the predictions.

Residue-Residue Contact Pairs (*fnat*) – Residues on either protein at the interface were considered to be in contact if any of their atoms were within 5 Å of each other. The total number of residue-residue contact pairs was calculated for the target structure (crystallographic complex structure) and for each prediction using SYBYL v8.1. '*fnat*', defined as the fraction of number of native/correct contacts identified in the predicted complex structure with respect to the target structure, was then computed.

Ligand Root Mean Square Deviation (*l-RMSD*) and Interface Root Mean Square Deviation (*i-RMSD*) – Two more parameters were calculated to evaluate the 3D fit between the predicted complexes and target structures. The global geometric fit was calculated by computing the *l-RMSD*, which is defined as the RMSD of the ligand backbone atoms in the predicted complexes versus the target structure, after superimposing the receptor protein. The fit within the interfacial region was quantified by calculating the *i-RMSD*, defined as the RMSD of the backbone atoms of all interfacial residues of predicted complexes versus target structure. For this calculation, interfacial residues were defined as those within 10 Å of the partner molecule. The interfacial residues were identified using SYBYL v8.1 and the l-RMSD and i-RMSD calculations were performed using the McLachlan algorithm[73] as implemented in the program ProFit (Martin, A. C. R. and Porter, C. T., http://www.bioinf.org.uk/software/profit/).

The predictions were classified into four categories – incorrect models, acceptable models (*), medium-accuracy models (**) and high-quality models (***) based on the acceptance criteria shown in Table 4.3.

Average Hit Count – A prediction with medium-accuracy or better (** or ***) was considered a "hit". An average hit count was calculated for the top *N* predictions for both unsolvated and solvated docking protocols.

Weighted Score – For the purpose of giving quantitative measures to the success of the two docking protocols, a weighted-scored was calculated for each test case by giving a value of 0, 1, 2, or 3 to the incorrect, acceptable, medium and high accuracy predictions within top *N* models ranked based on HINT scores, respectively.

Statistical Analysis – All the statistical analysis were performed at the level of significance $\alpha = 0.05$ using the software JMP v10.[74]

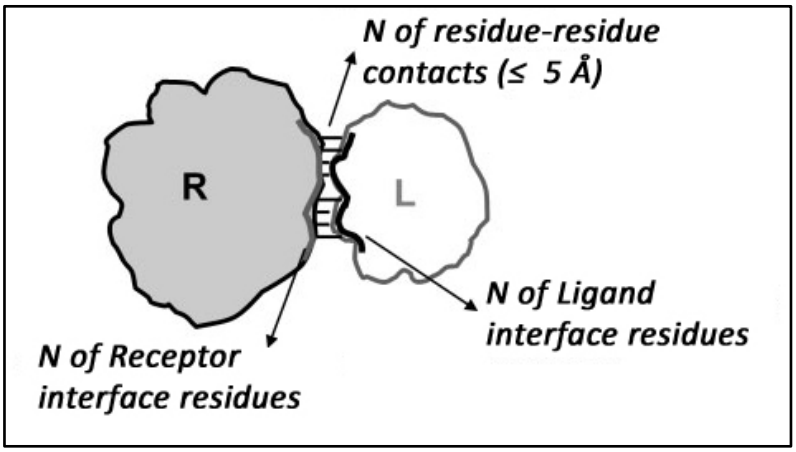


Figure 4.1 – Schematic illustration of residue-residue contact pairs. A receptor residue is considered to be in contact with a ligand residue if any of its atoms were within 5 Å of each other. *fnat* is the fraction of native/correct residue-residue contact pairs identified in a prediction.

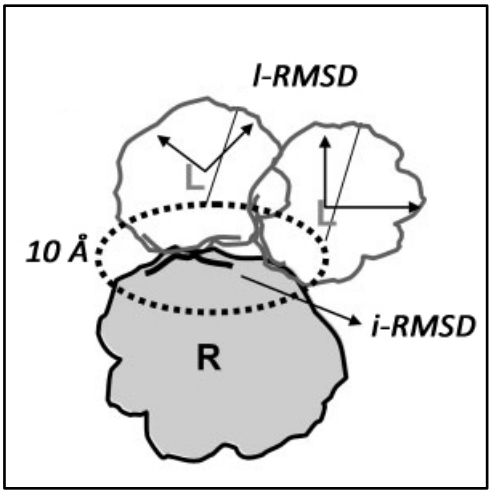


Figure 4.2 – Schematic illustration of ligand and interface RMSDs. *l-RMSD* is the RMSD of ligand backbone atoms between prediction and target structure, after superimposing receptor atoms. *i-RMSD* is the RMSD of interfacial residues, those within 10 Å of partner molecule.

Table 4.2 – Predicted model quality classification criteria[29]

Model Quality	**Criteria**
Incorrect	*fnat* < 0.1 OR (*l-RMSD* > 10.0 Å AND *i-RMSD* > 4.0 Å)
Acceptable (*)	*(fnat* ≥ 0.1 AND *fnat* < 0.3) AND (*l-RMSD* ≤ 10.0 Å OR *i-RMSD* ≤ 4.0 Å)
	OR
	fnat ≥ 0.3 AND *l-RMSD* > 5.0 Å AND *i-RMSD* > 2.0 Å
Medium (**)	*(fnat* ≥ 0.3 AND *fnat* < 0.5) AND (*l-RMSD* ≤ 5.0 Å OR *i-RMSD* ≤ 2.0 Å)
	OR
	fnat ≥ 0.5 AND *l-RMSD* > 1.0 Å AND *i-RMSD* > 1.0 Å
High (***)	*fnat* ≥ 0.5 AND (*l-RMSD* ≤ 1.0 Å OR *i-RMSD* ≤ 1.0 Å)

4.3 Results

4.3.1 *Data Set*

The protein-protein docking benchmark, developed by *Weng's* group,[68] comprises a total of 176 cases that are classified into three classes based on the extent of conformational change at the interface upon complex formation – rigid body cases (123), medium difficulty cases (29) and difficult cases (24). The high-resolution (< 2.0 Å) subset of 42 complexes chosen from the data set contains cases from all the three classes defined by *Weng* and also represents a good sampling of the protein interface sizes,[75] with change in accessible surface areas (ΔASA) on complex formation ranging from 808 to 3347 $Å^2$.[68] A complete hydropathic analysis of the protein-protein interface for each test case was performed using HINT. Only those water molecules that were relevant to both proteins, the so-called "Relevance 2" waters or "Bridging" waters, were retained with their protein-protein complexes, while those that were relevant to just one protein or neither were ignored for the present study. 12 out of 42 cases did not show the presence of any bridging waters and were removed from the data set. HINT interaction scores were now calculated, using the same parameters as described before, now taking into account the contribution of bridging waters. Table 4.2 lists the set of randomly selected 15 protein-protein complexes used for this study, along with their PDB ID, crystallographic resolution, chain IDs of receptor protein and ligand protein, the total numbers of interfacial waters and relevance-2/bridging waters, and their HINT interaction scores calculated with and without accounting for bridging waters.

Table 4.3 – Solvated protein-protein docking data set

#	PDB ID	Resolution	Chain ID Rec/Lig	Water		HINT Score	
				Interfacial	Bridging HOH	Without HOH	Accounting for Bridging HOH
1	1avx	1.90	A/B	8	2	2122.30	2364.61
2	1clv	2.00	A/I	30	8	-167.35	2319.10
3	1dqj	2.00	AB/C	17	7	1786.58	2863.55
4	1fle	1.90	E/I	12	1	1063.10	1035.44
5	1iqd	2.00	AB/C	25	6	2011.56	2589.28
6	1jiw	1.74	P/I	29	3	-2855.97	-1540.40
7	1jps	1.85	HL/T	8	3	1172.06	1458.27
8	1klu	1.93	AB/D	8	3	1249.55	1585.62
9	1pxv	1.80	A/C	24	5	450.66	2509.35
10	1r0r	1.10	E/I	27	6	-1361.36	-496.70
11	1r8s	1.46	E/A	24	5	118.12	1974.93
12	1wej	1.80	HL/F	10	6	1847.78	2735.49
13	1zhh	1.94	A/B	32	10	-411.06	2223.41
14	2hqs	1.50	A/H	46	7	-420.71	1523.90
15	2sic	1.80	E/I	17	1	-1.08	159.11

4.3.2 *ZDOCK – A rigid body docking program*

ZDOCK uses the Fast Fourier Transform (FFT) algorithm to find the 3D structure of a protein complex, starting from structures of individual components, by optimizing three parameters – shape complementarity, electrostatics and desolvation free energy.[76] Each individual protein file is first parsed through the *mark_sur* program that calculates the amount of accessible surface area (ASA) of each atom using a water probe and marks the atom type based on its atomic contact energy (ACE).[77] This is followed by a search in the 3D translational space using a FFT approach with the ligand protein, the smaller of the two, rotated in either 15° or 6° steps resulting in a total of 3,600 or 54,000 angles, respectively. The top scoring translation is retained for each angle. Each receptor-ligand complex is scored using physical and biochemical properties: i) pair-wise shape complementarity (PSC) that is composed of a favorable term coming from number of atoms pairs between receptor and ligand protein within a cutoff distance and a penalty term for number of overlapping grid points; ii) an electrostatic energy term that correlates the electric potential generated by receptor with the charges of ligand; and iii) desolvation free energy term calculated based on atomic contact energies.[76]

4.3.3 *HINT scores predict correct geometry*

As previously mentioned, HINT scores have been successfully correlated to the free energy of interaction in case of protein-ligand systems. The use of HINT scoring function in distinguishing active molecules from inactive ones is well documented. The first key question that must be answered before building protein-protein docking

algorithms based on HINT scoring is whether that function accurately predicts geometry. To test this, we performed a rigid-body docking on the data set of 15 high-resolution protein-protein complexes (crystallographic resolution ≤ 2.0 Å). The coordinates of the individual partners were obtained from the complex structure, and 100 predictions for each test case were generated using ZDOCK (the unsolvated protocol). Intermolecular interaction score for each prediction was calculated using the HINT, followed by ranking them based on their scaled HINT scores. The accuracy of each prediction was evaluated using the standard CAPRI criteria by calculating three parameters – the *fnat* value, the *l-RMSD and i-RMSD*. Predictions were classified as incorrect, acceptable-accuracy, medium-accuracy and high-accuracy models according to the cut-offs described in Table 4.3.

A prediction with high *fnat* value (close to 1) indicates correct identification of the interface. Figure 4.3 shows a plot of *fnat* values vs scaled HINT score for all predictions (n = 1500). A significant positive linear correlation was observed between the scaled HINT scores and *fnat* values ($r = 0.307$, $p < 0.0001$). That is, predictions with high scaled HINT scores have *fnat* values close to 1 and those with lower scaled-HINT scores have *fnat* values close to 0. We also checked the *fnat* values of top 10 and lowest 10 HINT-ranked predictions for each test case. A total of n = 92 out of 150 (61%) of the top 10 ranked predictions have a *fnat* value of ≥ 0.3 (one of the criteria for medium accuracy, or better); with 10 out of 15 (67%) top ranked predictions having a value of > 0.5 (high accuracy criteria) (Figure 4.4a). Also, a total of n = 112 out of 150 (75%) of the lowest 10 ranked predictions have a *fnat* value of < 0.3; with 12 out of the 15 (80%) lowest ranked predictions having a fnat value of < 0.1 (incorrect prediction)

(Figure 4.4b). In addition, out of all the predictions with *fnat* value of 0, i.e. for all predictions where not a single native residue-residue contact was identified, a total of n = 622 out of 691 predictions (90%) have a scaled HINT score of < 0.5. In general, high HINT-ranked predictions have *fnat* values close to 1 and vice versa – indicating the utility of HINT scores as an effective filter for pose selection.

4.3.4 *Unsolvated Docking vs Solvated Docking* –

Despite the progress made in the development of protein-protein docking programs, most predictions still lack considerable accuracy due to the complexity of the problem. When compared to docking small molecules into a pocket, there are many more degrees of freedom involved in bringing two proteins together. One of the most critical factors that influence the assembly of proteins – *water*, is almost always ignored. We believe that a protein-protein docking approach that explicitly accounts for interfacial waters, correct ionization states of residue side-chains and extensive flexibility during both the search and score stages will generate more accurate models.

To test the first of these, Relevant interfacial waters, i.e., “bridging” waters, were identified for the 15 protein-protein complexes in the data-set using the HINT Relevance Metric, and a rigid-body solvated docking was performed by forcing ZDOCK to include “bridging” waters as atoms in the receptor protein. 100 solutions were generated for every test case. The interfacial waters in every prediction were optimized for their correct orientation using the HINT-based water-optimization algorithm. HINT interaction scores were then calculated taking into account the energetic contribution of these

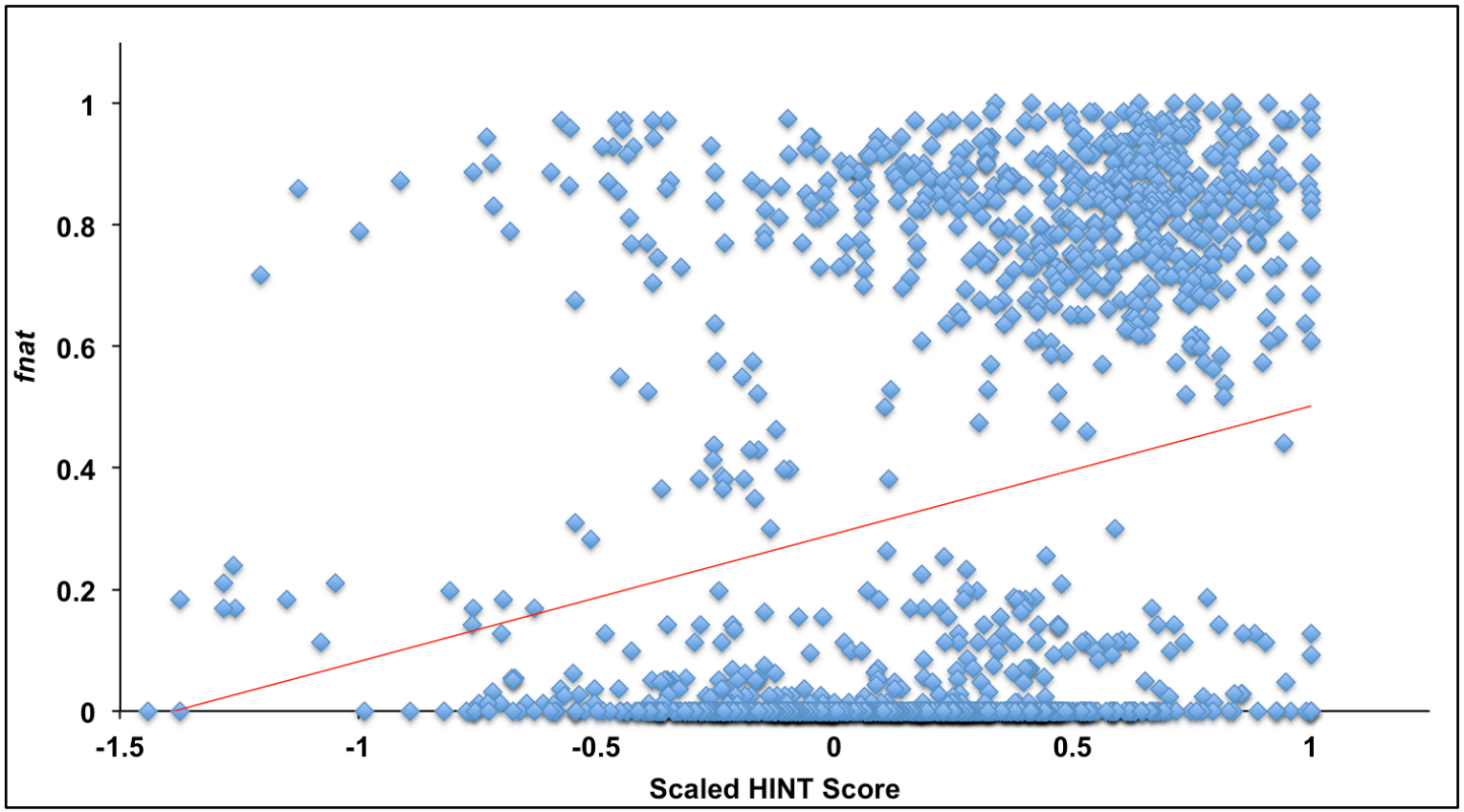


Figure 4.3 – Scatterplot showing positive linear correlation between scaled HINT Scores and *fnat* values of all predictions for each test case (n = 1500) ($r = 0.307$, $p < 0.0001$).
Predictions with high HINT scores have *fnat* values close to 1, and vice versa – indicating the ability of HINT to identify correct poses. For clarity purposes, points with scaled HINT scores below -1.5 are not shown.

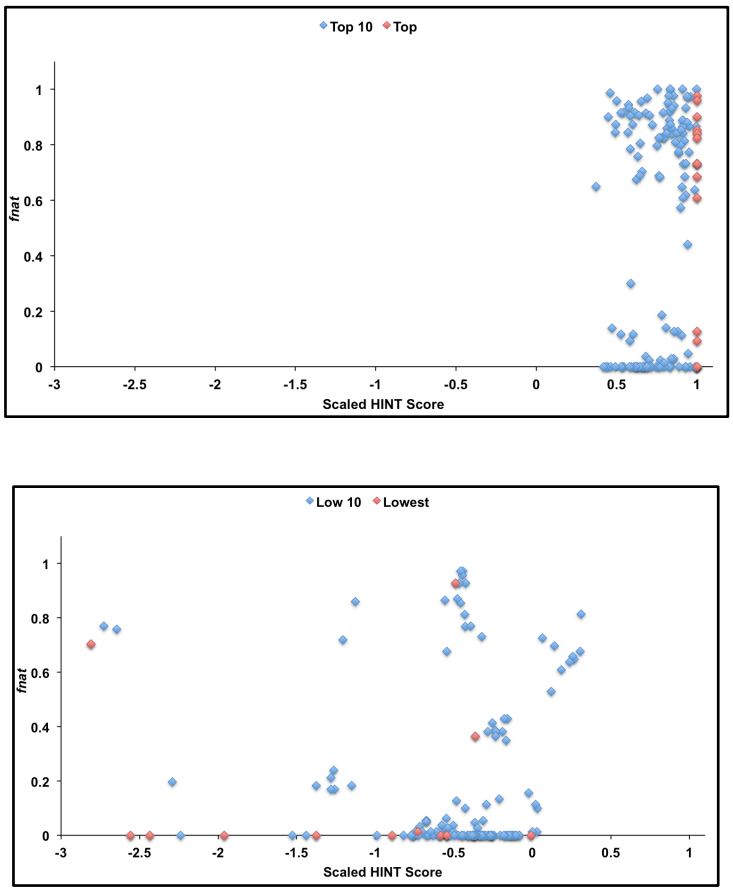


Figure 4.4 – Scatterplot of *fnat* vs scaled HINT scores for top 10 and lowest 10 predictions.

(A) Scatterplot of *fnat* values vs scaled HINT scores for top 10 predictions of every test case (n = 150). More than 60% of all top 10 predictions (92 out of 150) have *fnat* values corresponding to medium accuracy or better models. The points shown in red are the top predictions; 10 out of 15 have *fnat* values of > 0.5 – corresponding to high accuracy prediction.
(B) Scatterplot of *fnat* values vs Scaled HINT scores for the lowest 10 predictions of every test case (n = 150). 75% of all lowest 10 predictions (112 out of 150) have *fnat* values corresponding to acceptable or incorrect models. The points shown in red are the lowest ranked predictions; 12 out of 15 have *fnat* values < 0.1 – corresponding to incorrect prediction.

waters, and the predictions were ranked based on their scaled HINT scores. The accuracy of the predictions was evaluated according to the standard CAPRI criteria (Table 4.3), by computing three parameters – the *fnat* value, *l-RMSD* and *i-RMSD*; and classified as incorrect, acceptable-accuracy, medium-accuracy and high-accuracy models.

Two measures are calculated to compare the overall performance of unsolvated and solvated docking over the entire data-set – *average hit-count* and *weighted-score*. A prediction of medium accuracy or better was considered a *hit* and an *average hit-count* was calculated for the top *N* predictions. Also, every prediction was given a score of 0, 1, 2 or 3 for incorrect, acceptable, medium or high accuracy, respectively. A total *weighted-score* was calculated for the top *N* predictions of each test case.

Table 4.4 shows the *hit-count* for the top *N* predictions in case of unsolvated and solvated docking. As it can be seen, solvated docking performs better overall than unsolvated docking in terms of average hit-count. Especially in the top 10 predictions, which are of more importance compared to lower ranked predictions, there was significant improvement in the *hit-count* for each test case (paired t-test, $p < 0.05$) (Figure 4.5), with an overall improvement of 24.72 % (7.40 for solvated docking, compared to 5.93 for unsolvated docking) in the *average hit-count*. On comparing the *average hit-count* for the top, top 10, top 25, top 50 and top 100 predictions, a significant improvement (paired t-test, $p < 0.05$) was observed in the number of hits generated for solvated docking (Table 4.5, Figure 4.6).

Table 4.6 shows the *weighted-score* for the top *N* predictions in the cases of unsolvated and solvated docking. As can be seen, solvated docking performs better

overall than unsolvated docking in terms of the quality of predictions. There was a statistically significant improvement in the *weighted-score* for the top 10 predictions of each test case (paired t-test, $p < 0.05$) (Figure 4.7), with an overall improvement of 22.94 % in the average *weighted-score* (17.87 for solvated docking, compared to 14.53 for unsolvated docking). Similar to the *average hit-count*, the average *weighed-score* for top, top 10, top 25, top 50 and top 100 predictions showed statistically significant improvement (paired t-test, $p < 0.05$) for solvated docking (Table 4.7, Figure 4.8). When the number of high-accuracy predictions was compared (Table 4.8), similar improvements were observed. Most notable was the improvement seen in the success rate for the top ranked prediction: while the top prediction was a high-accuracy model for just n = 2 out of 15 (13%) test cases for unsolvated docking, the top prediction was a high-accuracy model for n = 6 out of 15 (40%) test cases for solvated docking. For the top ranked model, the quality of the prediction improved from incorrect/medium accuracy to high-accuracy for n = 5 out of total 15 (33%) test cases.

Table 4.4 – *Hit-count* for top *N* predictions for different docking protocols

PDB ID	Unsolvated Docking					Solvated Docking				
	Top 1	**Top 10**	**Top 25**	**Top 50**	**Top 100**	**Top 1**	**Top 10**	**Top 25**	**Top 50**	**Top 100**
1avx	1	8	20	36	54	0	9	22	41	63
1clv	1	10	25	50	95	1	10	25	50	99
1dqj	0	1	10	16	16	0	8	19	31	32
1fle	1	5	5	5	5	1	5	8	9	9
1iqd	1	10	25	50	76	1	10	25	49	85
1jiw	0	4	15	25	32	1	7	17	27	33
1jps	0	0	1	1	1	0	2	2	2	2
1klu	1	2	2	2	2	1	4	5	5	5
1pxv	1	10	25	49	59	1	10	24	43	62
1r0r	0	0	0	0	13	0	0	0	0	25
1r8s	1	10	24	48	96	1	10	25	50	93
1wej	0	4	4	4	4	1	10	11	11	11
1zhh	1	6	8	9	9	1	7	16	20	20
2hqs	1	9	16	19	19	1	9	16	18	18
2sic	1	10	24	47	74	1	10	22	46	72

Table 4.5 – Comparison of *average hit-count* for different docking protocols

	Docking Protocol	n (test-cases)	Top N Predictions				
			1	**10**	**25**	**50**	**100**
Total *hit-count*	Unsolvated	15	10	89	204	361	555
	Solvated	15	11	111	237	402	629
Difference			1	22	33	41	74
Average *hit-count*	Unsolvated		0.67	5.93	13.60	24.07	37.00
	Solvated		0.73	7.40	15.80	26.80	41.93
Difference			0.07	1.47	2.20	2.73	4.93

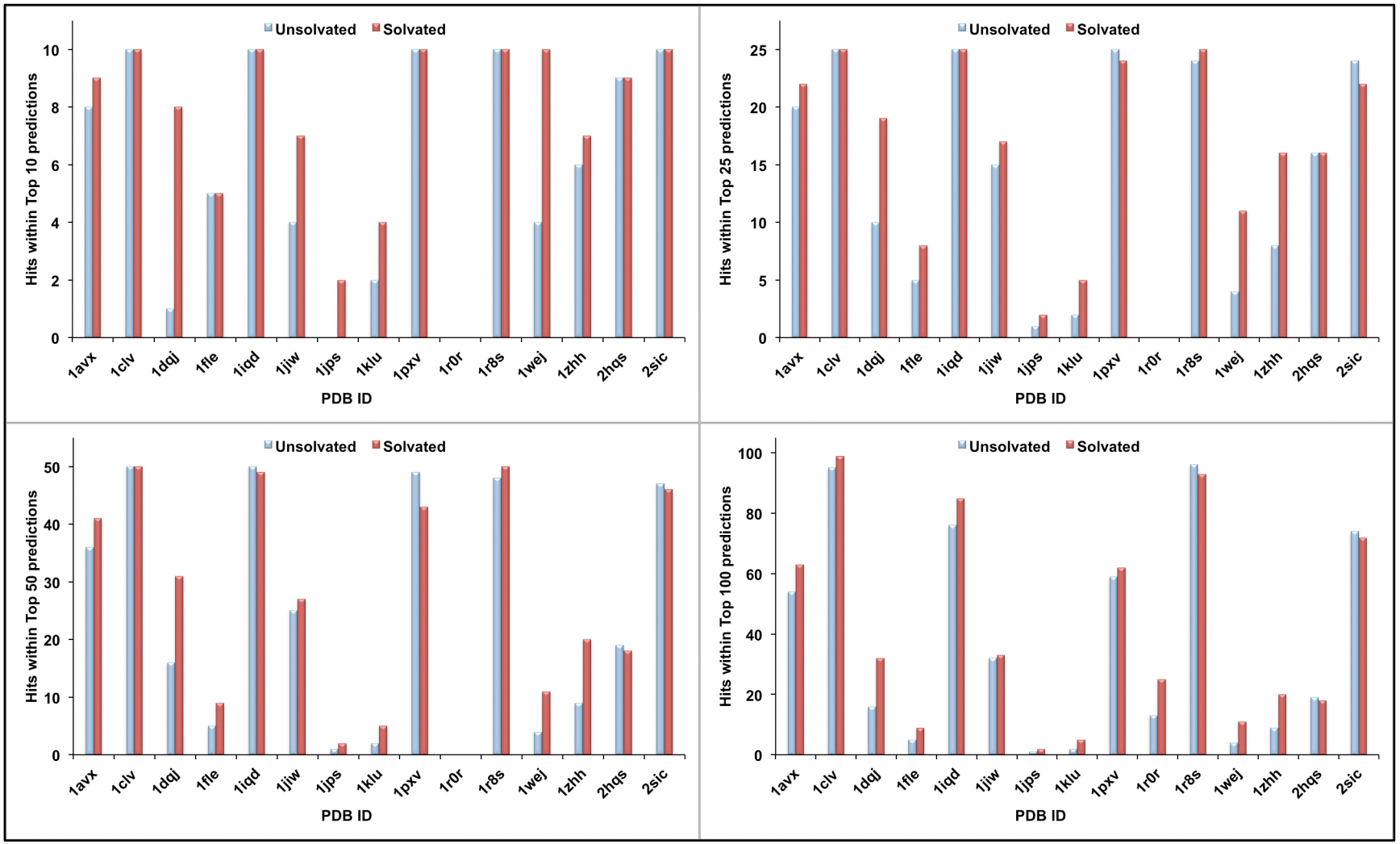


Figure 4.5 a – Number of *hits* in the top *N* predictions in unsolvated and solvated docking

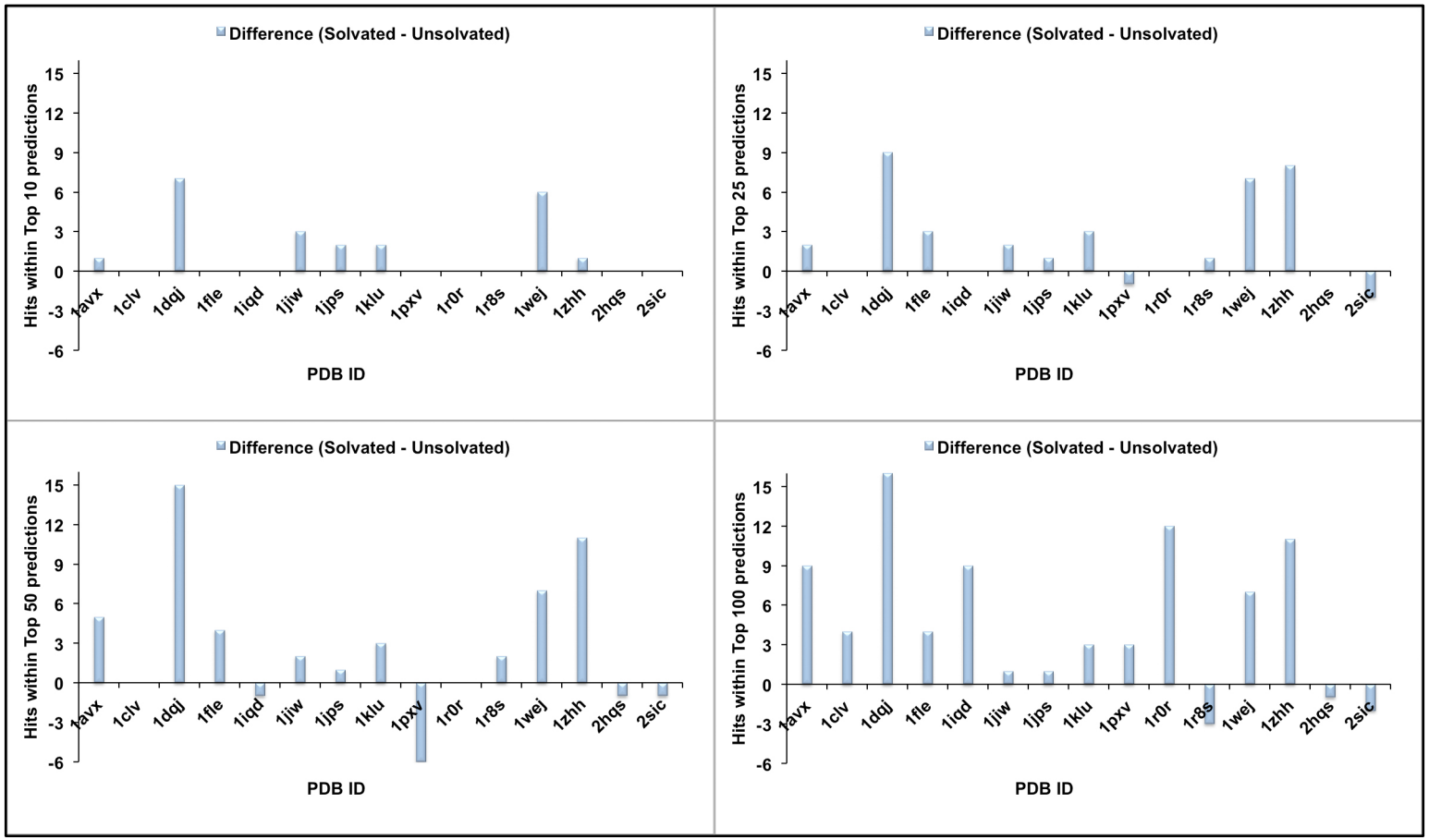


Figure 4.5 b – Difference in number of *hits* in the top *N* predictions in unsolvated and solvated docking for each test case. ($\Delta = N_{solvated} - N_{unsolvated}$).

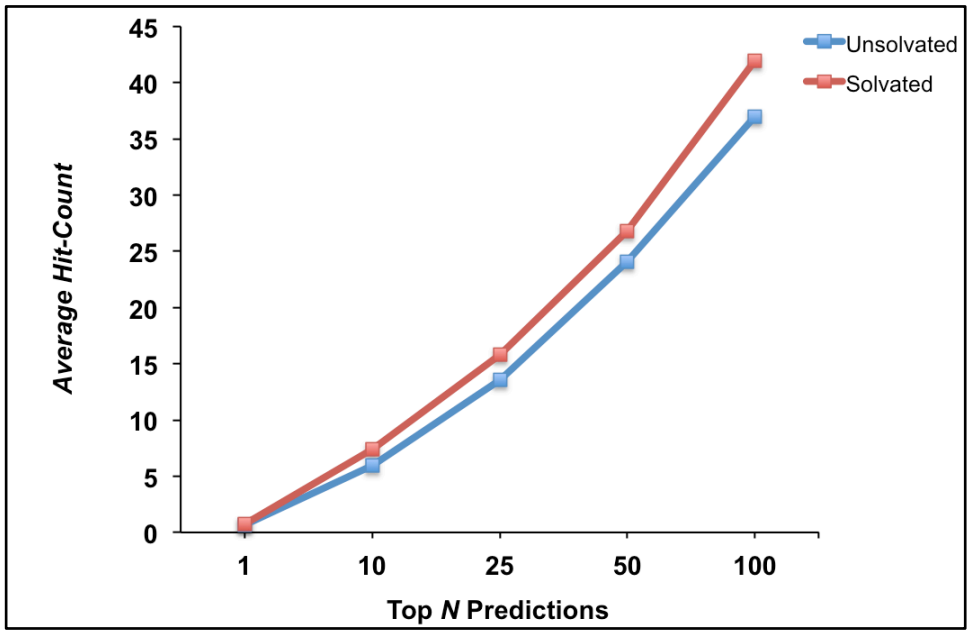


Figure 4.6 – Comparison of average hit-count in top N predictions for unsolvated and solvated docking.

Table 4.6 – *Weighted-score* for top *N* predictions for different docking protocols

PDB ID	Unsolvated Docking					Solvated Docking				
	Top 1	Top 10	Top 25	Top 50	Top 100	Top 1	Top 10	Top 25	Top 50	Top 100
1avx	2	19	48	93	143	1	19	52	97	155
1clv	3	30	75	149	284	3	30	75	150	296
1dqj	0	2	24	36	39	0	18	43	69	72
1fle	2	14	14	14	15	3	15	24	27	28
1iqd	2	23	59	112	170	2	21	56	109	188
1jiw	0	10	37	63	79	2	17	41	68	80
1jps	0	0	3	3	3	0	4	4	4	4
1klu	3	6	6	6	6	3	12	16	16	16
1pxv	2	21	51	101	122	2	21	49	91	130
1r0r	0	0	0	0	37	0	0	0	0	71
1r8s	2	20	49	97	196	2	20	51	102	189
1wej	0	10	13	14	14	3	26	32	35	36
1zhh	2	14	18	20	20	2	14	33	41	41
2hqs	2	23	40	49	49	3	23	40	45	45
2sic	2	26	60	112	190	3	28	59	113	189

Table 4.7 – Comparison of *average weighted-score* for different docking protocols

	Docking Protocol	n (test-cases)	Top *N* Predictions				
			1	**10**	**25**	**50**	**100**
Total *weighted-score*	Unsolvated	15	22	218	497	869	1367
	Solvated	15	29	268	575	967	1540
Difference			7	50	78	98	173
Average *weighted-score*	Unsolvated		1.47	14.53	33.13	57.93	91.13
	Solvated		1.93	17.87	38.33	64.47	102.67
Difference			0.47	3.33	5.20	6.53	11.53

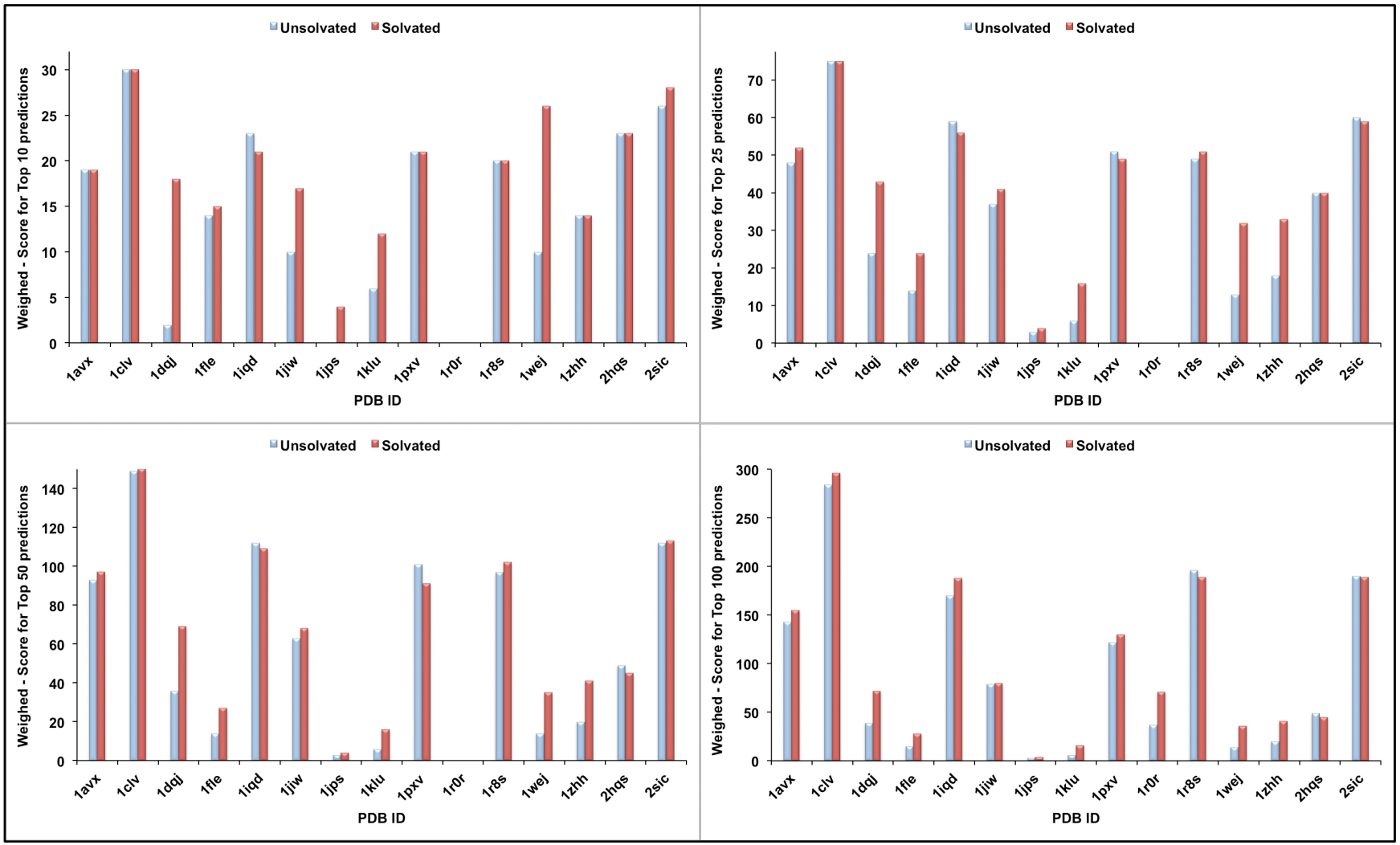


Figure 4.7 a – Weighted – Score for the top *N* predictions in unsolvated and solvated docking

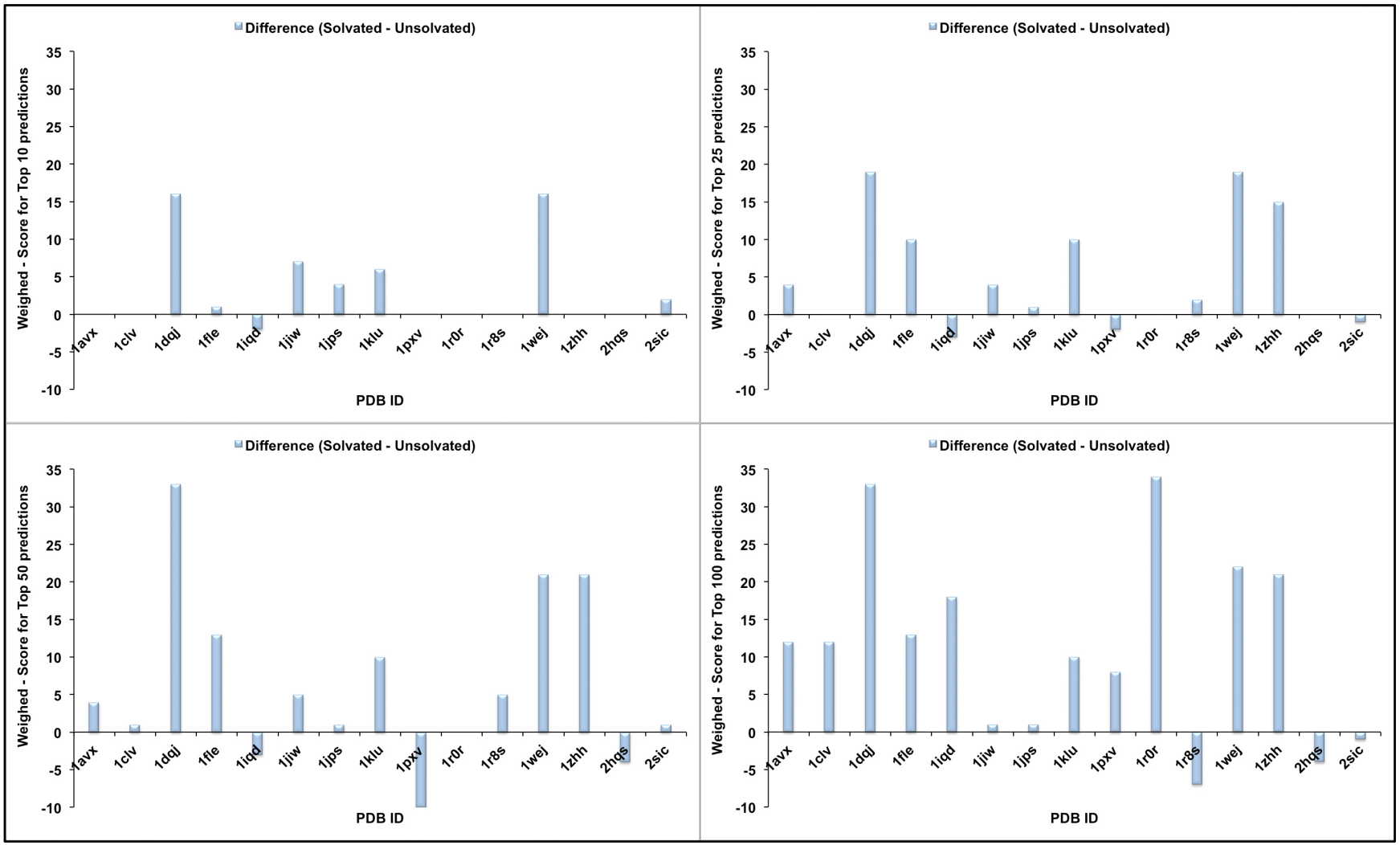


Figure 4.7 b – Difference in weighted-score for top *N* predictions in unsolvated and solvated docking for each test case. ($\Delta = S_{solvated} - S_{unsolvated}$).

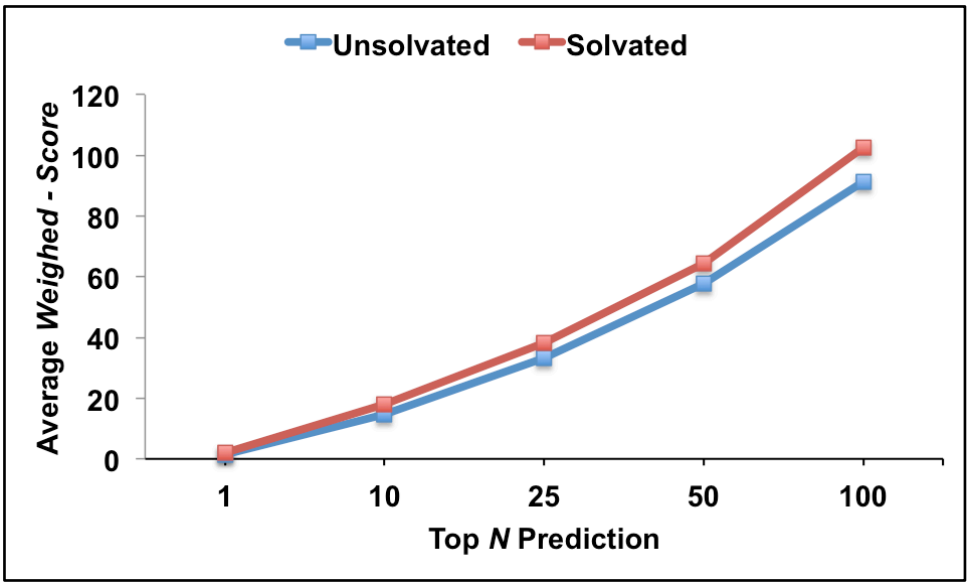


Figure 4.8 – Comparison of average weighted – score for top N predictions for unsolvated and solvated docking.

Table 4.8 – Count of high-accuracy (***) models in top *N* predictions

PDB ID	Unsolvated Docking					Solvated Docking				
	Top 1	Top 10	Top 25	Top 50	Top 100	Top 1	Top 10	Top 25	Top 50	Top 100
1avx	0	1	3	7	8	0	0	5	6	7
1clv	1	10	25	49	94	1	10	25	50	98
1dqj	0	0	4	4	4	0	2	5	7	7
1fle	0	4	4	4	4	1	5	8	9	9
1iqd	0	3	9	12	16	0	1	6	10	14
1jiw	0	0	4	9	9	0	2	5	11	11
1jps	0	0	1	1	1	0	0	0	0	0
1klu	1	2	2	2	2	1	4	5	5	5
1pxv	0	1	1	3	3	0	1	1	5	6
1r0r	0	0	0	0	11	0	0	0	0	21
1r8s	0	0	0	0	3	0	0	1	2	2
1wej	0	2	2	2	2	1	6	6	6	6
1zhh	0	2	2	2	2	0	0	1	1	1
2hqs	0	5	8	11	11	1	5	8	9	9
2sic	0	6	12	18	19	1	8	15	20	21
Total	2	36	77	124	189	6	44	91	141	217
Average	0.13	2.40	5.13	8.27	12.60	0.40	2.93	6.07	9.40	14.47

The comparison between the predictions obtained from both docking protocols show that there is a significant improvement in the total number of *hits* and the *quality* of predictions for solvated docking. This trend indicates that more accurate and reliable results are obtained when bridging interfacial waters are ***not*** ignored.

4.4 Discussion

Detailed analysis of protein-protein complexes has revealed that water molecules form hydrogen bonds with interfacial side-chains, mediating and stabilizing the biomolecular association. Most water molecules are not randomly trapped in the protein–protein interface, but are part of the recognition code facilitating interactions that are less favorable in its absence.[78] However, most current docking programs only take into account the underlying physics of protein-protein interactions, ignoring the role of water molecules. The conformational search step of the docking process is generally performed in vacuum, not accounting for the presence of water molecules. Some docking algorithms incorporate a desolvation term in their scoring functions, implicitly accounting for water. This improves the ranking of docked predictions and subsequent identification of correct configuration.[79] However, implicit treatment of waters introduces approximations and the description of energetics is coarser than explicit models.

In the current study, we demonstrated that water can be explicitly introduced into protein-protein docking protocols. Using HINT-based tools to identify Relevant bridging water molecules and incorporating them into the docking protocol improves the quality of predictions. Figure 4.9 shows the plot of *i-RMSD* vs scaled HINT scores for all predictions obtained by our solvated docking protocol, grouped based on quality. A total of n = 111 out of 150 (74%) top 10 predictions were of medium accuracy or better (Table 4.5); showing improvements in not only the number of hits, but also in scoring with high/medium accuracy models ranking much better than incorrect ones.

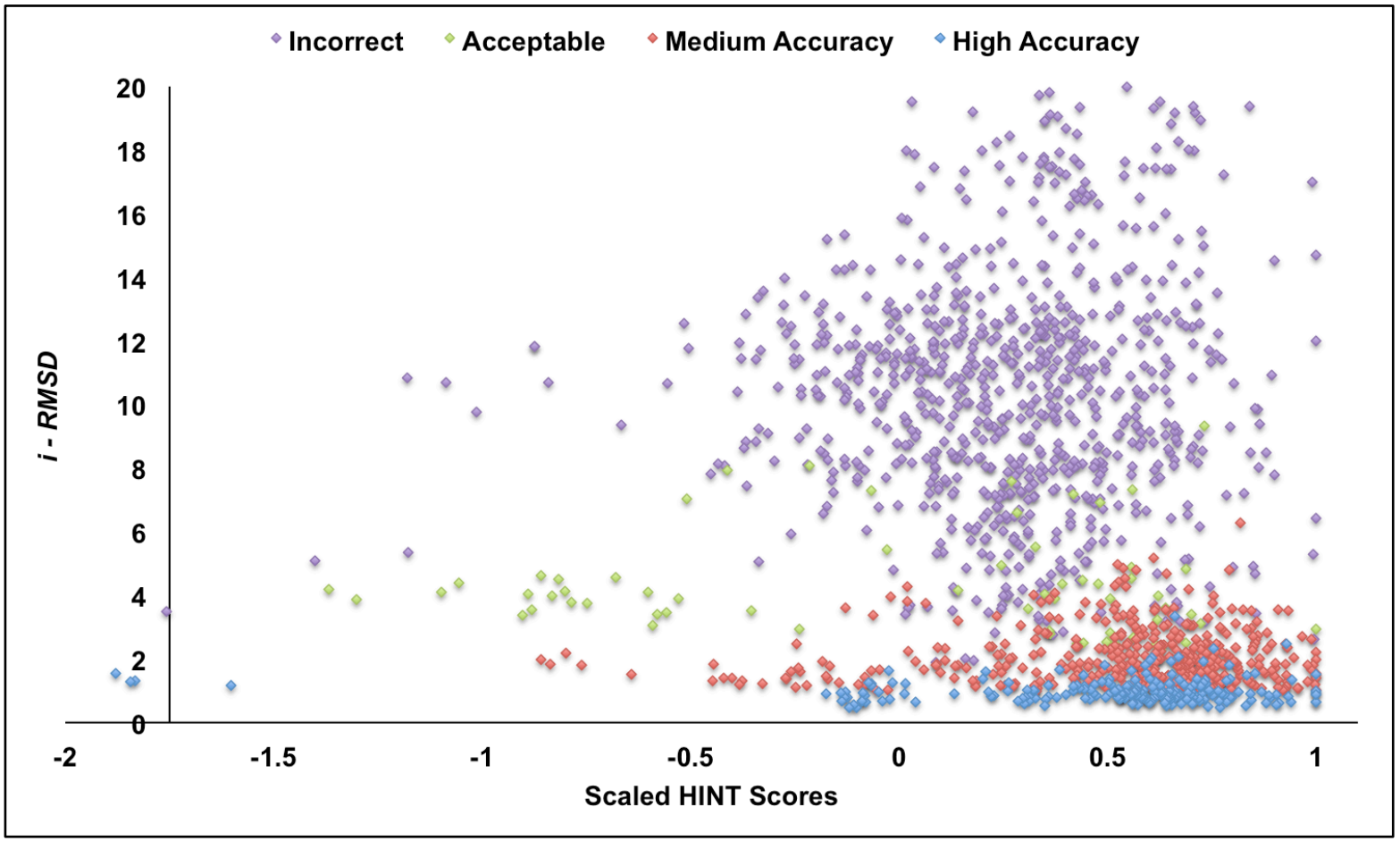


Figure 4.9 – Scatterplot of *i-RMSD* vs scaled HINT Scores for all predictions obtained from solvated docking, grouped based on their quality.
On average, predictions of high/medium accuracy rank much better than the incorrect ones.

We can now illustrate how docking results improve when a solvated docking approach is applied, using as an example the anti-lysozyme antibody HyHEL-63 complexed with hen egg white lysozyme HEL (test case # 3, PDB ID: 1dqj). The crystal structure of the complex has been determined at 2.0 Å resolution, with the presence of 17 interfacial waters (within 4.0 Å of both molecules).[80] The orientation of each water molecule was optimized using the HINT-based optimization algorithm, as described before. The relevance of these waters was determined using HINT Relevance metric, which identified 7 waters to be Relevant to both proteins forming bridging interactions with interfacial residues (Table 4.9, Figure 4.10). The individual proteins were separated from the complex structure and subsequently docked using both the unsolvated and solvated docking protocols; the 7 bridging waters were considered as a part of receptor protein for the latter. As described before, the predictions were ranked based on their HINT interaction scores and also evaluated for accuracy based on standard CAPRI criteria.

Table 4.9 – HINT Water Relevance Report for interfacial waters* in Anti-Lyzozyme Antibody HyHEL-63 – Lyzozyme HEL complex crystal structure (test case #3, PDB ID – 1dqj).

Water #	Monomer Name	Target One - Anti-Lyzozyme Antibody HyHEL-63 Chain A/B			Target Two -Lyzozyme HEL Chain C		
		Rank	Score	Relevance	Rank	Score	Relevance
1	**HOH130**	**4.19**	**68.60**	**0.65**	**1.51**	**72.20**	**0.37**
2	HOH131	1.26	122.40	0.39	2.34	-314.10	-0.47
3	**HOH133**	**2.74**	**124.60**	**0.56**	**2.78**	**-36.50**	**0.39**
4	HOH134	1.03	-5.70	0.25	2.58	-218.80	-0.24
5	HOH138	1.42	-58.90	0.21	3.59	-195.60	-0.19
6	HOH140	0.00	-64.00	-0.04	2.21	196.70	0.57
7	**HOH141**	**2.50**	**48.70**	**0.45**	**1.32**	**16.80**	**0.30**
8	**HOH143**	**1.44**	**117.60**	**0.41**	**3.92**	**48.20**	**0.62**
9	HOH146	0.90	-42.40	0.20	1.20	272.00	0.44
10	HOH152	1.03	-43.40	0.22	1.14	121.50	0.36
11	**HOH182**	**2.40**	**123.30**	**0.52**	**1.32**	**29.50**	**0.31**
12	**HOH222**	**3.00**	**270.70**	**0.71**	**2.85**	**113.00**	**0.55**
13	HOH243	1.02	-78.60	0.19	1.08	116.60	0.35
14	HOH263	1.13	60.30	0.32	0.00	-153.80	-0.07
15	HOH327	0.96	7.00	0.24	0.93	42.50	0.27
16	HOH335	1.03	166.90	0.34	0.00	-79.10	-0.04
17	**HOH388**	**1.11**	**43.70**	**0.30**	**0.92**	**36.50**	**0.26**

*Relevant/bridging waters (having relevance ≥ 0.25 for both proteins) are shown in bold.

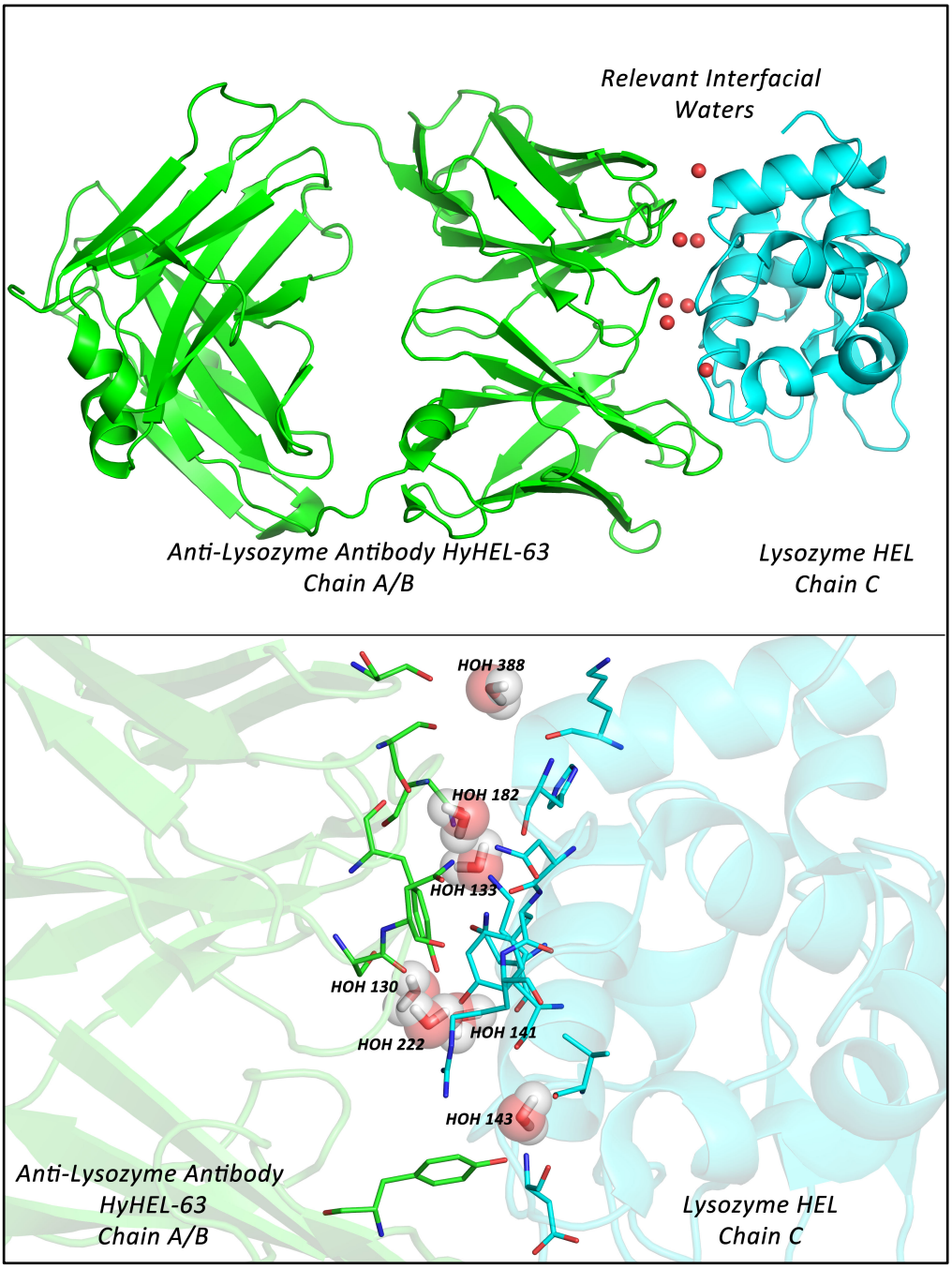


Figure 4.10 – (Top) Crystal Structure of Anti-Lysozyme Antibody HyHEL-63 (green) – Lysozyme HEL (cyan) complex (PDB ID: 1dqj). The image shows the presence of Relevant interfacial waters (red spheres). (Bottom) Detailed view of the interface showing bridging interactions of Relevant waters with residues on both proteins. Image prepared using PyMOL.[81]

In the case of unsolvated docking, a total of n = 16 out of 100 (proportion = 0.16) predictions were of medium accuracy or better (Table 4.4), with just one prediction of medium accuracy (rank = 10) within the top 10. Clustering the top 10 predictions showed three different poses for the ligand proteins – Cluster 1 consisting of n = 5 predictions (rank – 1, 2, 3, 4, 8) was an incorrect pose; Cluster 2 consisting of n = 4 predictions (rank – 5, 6, 7, 9) was also an incorrect pose; Cluster 3 consisting of n = 1 prediction (rank – 10) was a medium accuracy pose (Table 4.10). Solvated docking, on the other hand, resulted in n = 32 (proportion = 0.32) predictions of medium accuracy or better (Table 4.4). More notable was the improvement in the accuracy of the top 10 predictions. Clustering showed two different poses – Cluster 1 consisting of n = 2 predictions (rank – 1, 3) was an incorrect pose; while Cluster 2 consisting of the remaining n = 8 predictions was the native-like pose, with 6 predictions of medium accuracy and 2 predictions (rank – 7, 8) of high-accuracy (Table 4.11)

Table 4.10 – Unsolvated docking results for HyHEL-63 – HEL complex

Prediction	HINT Score	Scaled HINT Score	Rank	*i-RMSD*	*l-RMSD*	*fnat*	Cluster/ Pose	Model Quality	Hit
model 1	4805.926	1.000	1	8.549	11.004	0.127	1	-	n
model 2	4355.065	0.906	2	9.864	12.264	0.113	1	-	n
model 3	4236.699	0.882	3	9.806	12.167	0.127	1	-	n
model 4	4124.475	0.858	4	8.835	11.374	0.127	1	-	n
model 5	4115.901	0.856	5	7.954	12.733	0.028	2	-	n
model 6	4105.278	0.854	6	8.256	13.208	0.028	2	-	n
model 7	4036.962	0.840	7	8.388	13.442	0.028	2	-	n
model 8	3880.502	0.807	8	9.573	12.326	0.141	1	-	n
model 9	3835.744	0.798	9	8.235	13.061	0.014	2	-	n
model 10	3796.472	0.790	10	1.716	2.321	0.915	3	**	y

Top 10 predictions, ranked based on their HINT scores. *i-RMSD, l-RMSD* and *fnat* values calculated against the complex crystal structure. Three clusters/poses for the ligand protein (within 2.0 Å of each other) were seen within the top 10 models, as indicated by the number. Asterisks in the model quality column correspond to CAPRI quality criteria – high accuracy (***), medium accuracy (**), acceptable (*) and incorrect (-). *Hit* – a prediction with medium accuracy or better.

Table 4.11 – Solvated docking results for HyHEL-63 – HEL complex

Prediction	HINT Score	Scaled HINT Score	Rank	*i-RMSD*	*l-RMSD*	*fnat*	Cluster/ Pose	Model Quality	Hit
model 1	6098.512	1.000	1	6.436	7.036	0	1	-	n
model 2	5327.556	0.874	2	1.617	2.223	0.944	2	**	y
model 3	5045.993	0.827	3	7.194	7.474	0	1	-	n
model 4	4982.928	0.817	4	1.497	1.208	0.986	2	**	y
model 5	4855.246	0.796	5	1.079	1.305	0.958	2	**	y
model 6	4852.371	0.796	6	1.049	2.931	0.761	2	**	y
model 7	4774.186	0.783	7	0.948	1.717	0.901	2	***	y
model 8	4715.043	0.773	8	0.899	0.640	0.958	2	***	y
model 9	4638.276	0.761	9	2.480	2.463	0.873	2	**	y
model 10	4634.395	0.760	10	1.414	3.055	0.775	2	**	y

Top 10 predictions, ranked based on their HINT scores. *i-RMSD, l-RMSD* and *fnat* values calculated against the complex crystal structure. Two clusters/poses for the ligand protein (within 2.0 Å of each other) were seen within the top 10 models, as indicated by the number. Asterisks in the model quality column correspond to CAPRI quality criteria – high accuracy (***), medium accuracy (**), acceptable (*) and incorrect (-). *Hit* – a prediction with medium accuracy or better.

These improvements in the solvated docking results can be attributed to the presence of water molecules that change the physical and chemical properties of the receptor protein interface. For instance, water HOH143 is involved in bridging interactions between B/Tyr58 of HyHEL-63 (receptor protein) and C/Val99 and C/Asp101 of HEL (ligand protein), as seen in the crystal structure (Figure 4.11). In unsolvated docking, because of the absence of an explicit water molecule near B/Tyr58, the region is occupied by residues of ligand protein – C/Gly22 in cluster 1 and C/Gln57 in cluster 2, forming direct polar interactions with the Tyr –OH group and resulting in non-native like predictions. For cluster 3, a native like prediction with medium accuracy was generated, although the interaction between B/Tyr58 and C/Val99 is not observed (Figure 4.12). But when docking is performed using the solvated docking protocol, the presence of water HOH143 molecule results in better shape and hydropathic complementarity with the ligand surface and thereby leads to more accurate native-like predictions – cluster 2 retaining the water-mediated interaction between B/Tyr58 and C/Val99, as seen in the crystal structure. (Figure 4.13)

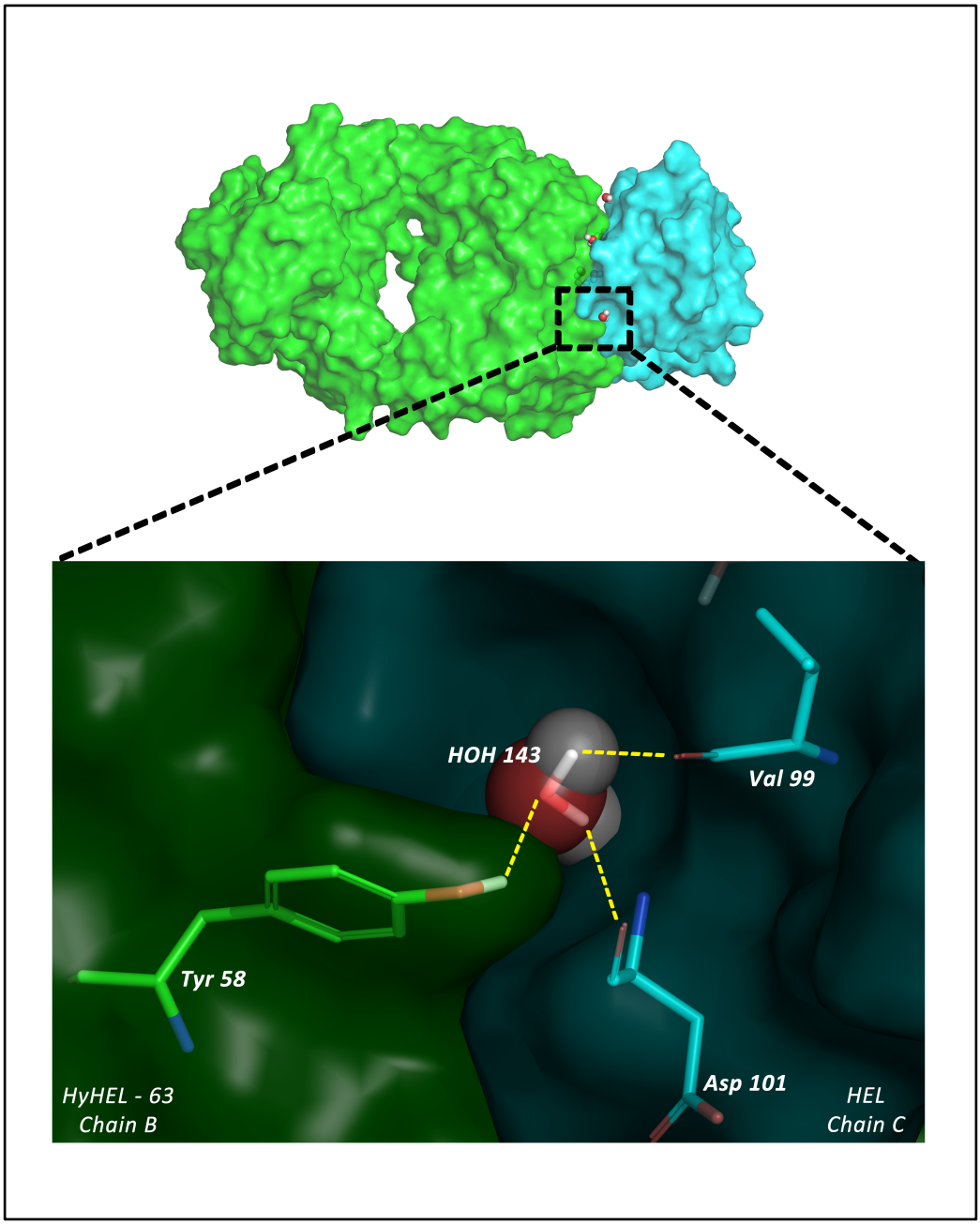


Figure 4.11 – Bridging interactions formed by *Relevant* interfacial water HOH 143 with Tyr58 of HyHEL-63 (receptor protein, green surface) and Val99 and Asp101 of HEL (ligand protein, cyan surface) observed in the crystal structure of the complex (PDB ID: 1dqj). Image prepared using PyMOL.

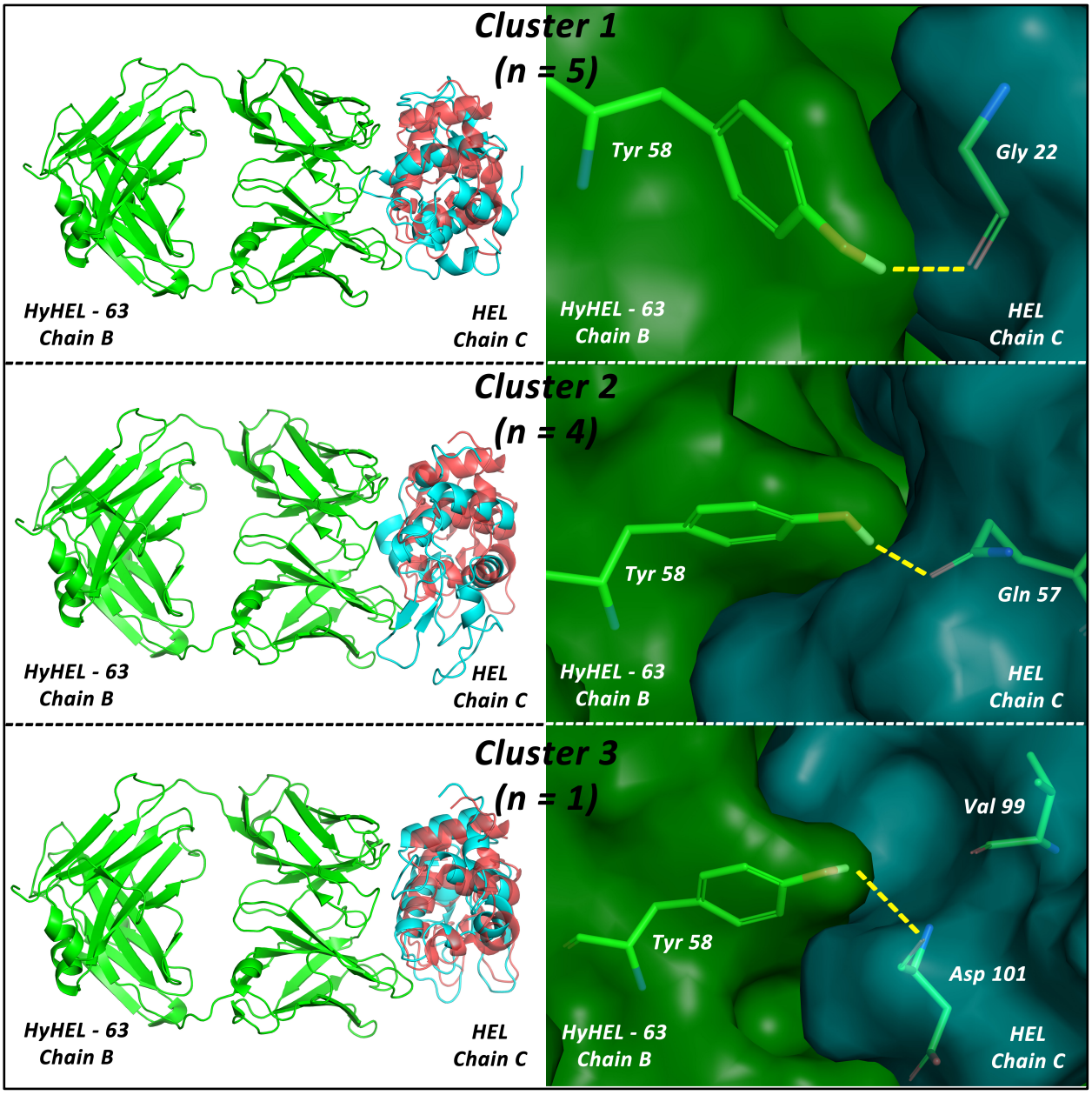


Figure 4.12 – Unsolvated docking results for HyHEL-63 – HEL complex.
The left panel for each cluster shows the overlay of predicted ligand pose (cyan cartoon) with the crystal structure (red cartoon). Only cluster 3 (n = 1 prediction) was a medium-accuracy native-like prediction. The right panel focuses on the interactions of B/Tyr58 with ligand residues. For cluster 1 and cluster 2, native residue-residue contacts are not retained, corresponding to incorrect predictions.
Image prepared using PyMOL.

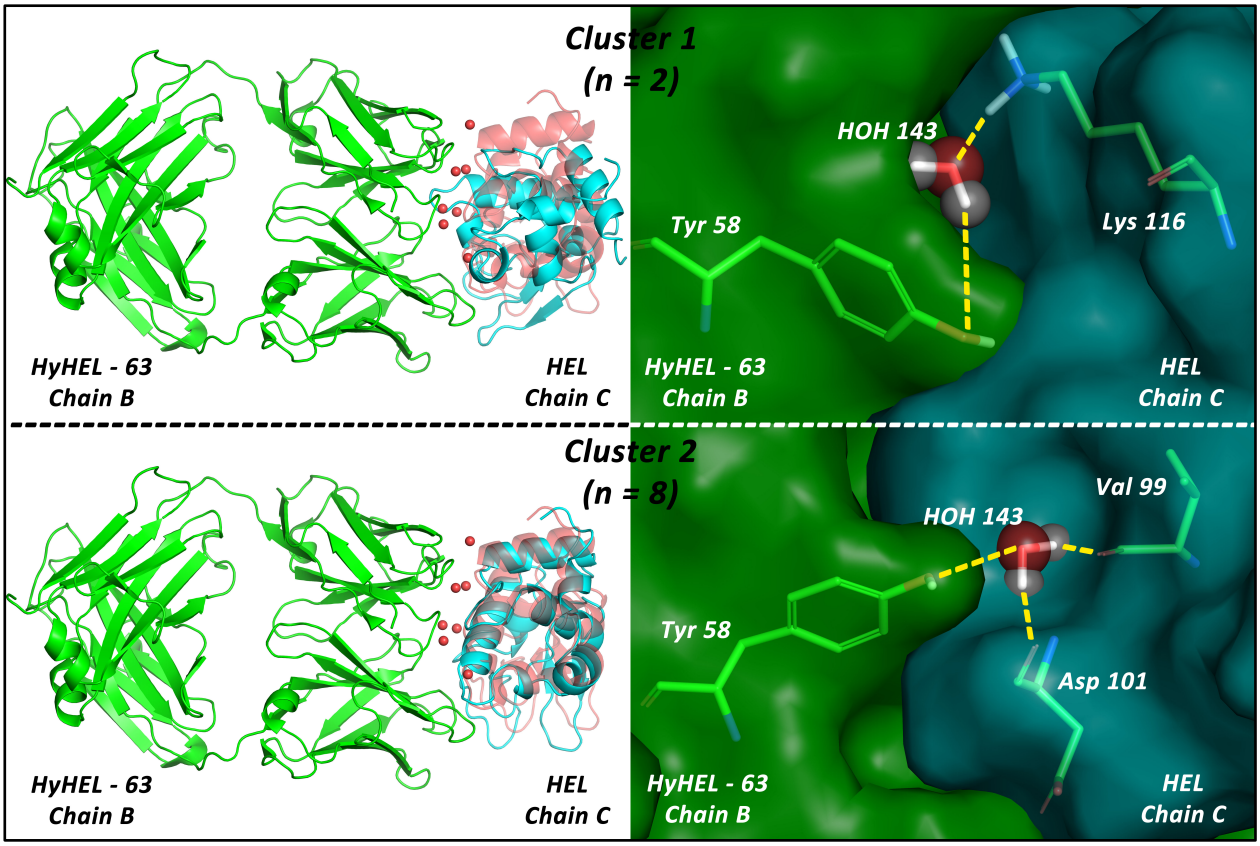


Figure 4.13 – Solvated docking results for HyHEL-63 – HEL complex.
The left panel for each cluster shows the overlay of predicted ligand pose (cyan cartoon) with the crystal structure (red cartoon). Majority of top 10 predictions (n = 8) were native-like predictions.
The right panel focuses on the bridging interactions of HOH143 with receptor and ligand residues. For cluster 1, non-native like prediction was generated showing water-mediated interaction between B/Tyr58 and C/Lys116. For cluster 2 (that consists of majority of top 10 predictions, n = 8), native water-mediated residue-residue contacts are retained, with B/Tyr58 showing water-mediated hydrogen-bonding network with C/Val99 and C/Asp101.
Image prepared using PyMOL.

Such results indicate that our solvated docking protocol, which utilizes HINT-based tools, can be successfully used for improving protein-protein docking results. A significant improvement in the quality of top predictions indicates that HINT scoring function can effectively discriminate between poses and suggests that scoring can be improved when waters are explicitly accounted for.

Even so, this approach is woefully crude because water is represented as only a single atom during the search stage and does not really reproduce its chemical properties. It is only during the scoring stage, when protons are added and water molecules are optimized, that the complete set of properties for the waters are incorporated. Since the HINT scoring function has shown to effectively identify native-like poses from incorrect ones, even greater improvements can be expected if we can introduce HINT scoring in the conformational space search to ascertain the viability of a particular pose. This will result in more accurate predictions to proceed to the subsequent refinement and scoring stages, overall improving the success rate of the docking algorithm.

Our current study focused on understanding the direct influence of interfacial water on the quality of structure prediction for protein-protein complexes. For this purpose, we performed a bound-bound docking, which means the starting structures of the proteins were obtained from the crystal structure of the bound complex. This eliminates two major issues that might result in incorrect predictions – protein flexibility associated with unbound docking and positions of Relevant water molecules. Ideally, we would like to like to start with unbound structures of the interacting partners and try

to predict the bound form, but this is a more difficult problem as it adds many more degrees of freedom to an already spectacularly under determined problem.

The HINT forcefield and ancillary tools like the HINT water relevance metric, HINT-based computational titration and a novel HINT map based 3D refinement algorithm can be successfully used to model the hydropathic complementarity, the positions of Relevant interfacial waters, the correct ionization states of interfacial residues and interfacial protein flexibility, respectively. With appropriate refinements, a novel HINT-based docking approach can be designed that can accurately model protein-protein complexes.

4.5 Conclusion

Characterizing the nature of interactions between proteins that have not been experimentally co-crystallized requires a docking approach that can successfully predict the spatial conformation adopted in the complex. Interfacial waters contribute immensely to the kinetics and thermodynamics underlying protein-protein interactions. In this chapter, we have demonstrated the utility of HINT scoring and other computational tools based off it towards structural prediction of protein-protein complexes, by explicitly accounting for interfacial waters that are generally ignored in the current docking programs. We have shown that using hydropathic complementarity and *not* ignoring these *Relevant* waters in the modeling of protein complexes does show an statistically significant improvement in the quality of predictions generated by the docking algorithm. The analysis of illustrative example of anti-lysozyme – lysozyme complex revealed that certain binding modes that would otherwise be ranked higher can be eliminated by the steric presence of water molecules. Also, the explicit presence of interfacial waters may result in additional hydrogen bond interactions, improving the energy scores, and thereby ranking the correct binding modes higher.

References

1. Peri, S.; Navarro, J. D.; Amanchy, R.; Kristiansen, T. Z.; Jonnalagadda, C. K.; Surendranath, V.; Niranjan, V.; Muthusamy, B.; Gandhi, T. K.; Gronborg, M.; Ibarrola, N.; Deshpande, N.; Shanker, K.; Shivashankar, H. N.; Rashmi, B. P.; Ramya, M. A.; Zhao, Z.; Chandrika, K. N.; Padma, N.; Harsha, H. C.; Yatish, A. J.; Kavitha, M. P.; Menezes, M.; Choudhury, D. R.; Suresh, S.; Ghosh, N.; Saravana, R.; Chandran, S.; Krishna, S.; Joy, M.; Anand, S. K.; Madavan, V.; Joseph, A.; Wong, G. W.; Schiemann, W. P.; Constantinescu, S. N.; Huang, L.; Khosravi-Far, R.; Steen, H.; Tewari, M.; Ghaffari, S.; Blobe, G. C.; Dang, C. V.; Garcia, J. G.; Pevsner, J.; Jensen, O. N.; Roepstorff, P.; Deshpande, K. S.; Chinnaiyan, A. M.; Hamosh, A.; Chakravarti, A.; Pandey, A. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **2003,** *13,* 2363-2371.
2. Isserlin, R.; El-Badrawi, R. A.; Bader, G. D. The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database (Oxford)* **2011,** *2011,* baq037.
3. Licata, L.; Briganti, L.; Peluso, D.; Perfetto, L.; Iannuccelli, M.; Galeota, E.; Sacco, F.; Palma, A.; Nardozza, A. P.; Santonico, E.; Castagnoli, L.; Cesareni, G. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **2012,** *40,* D857-861.
4. Gandhi, T. K.; Zhong, J.; Mathivanan, S.; Karthick, L.; Chandrika, K. N.; Mohan, S. S.; Sharma, S.; Pinkert, S.; Nagaraju, S.; Periaswamy, B.; Mishra, G.; Nandakumar, K.; Shen, B.; Deshpande, N.; Nayak, R.; Sarker, M.; Boeke, J. D.; Parmigiani, G.; Schultz, J.; Bader, J. S.; Pandey, A. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* **2006,** *38,* 285-293.
5. Dutta, S.; Berman, H. M. Large macromolecular complexes in the Protein Data Bank: a status report. *Structure* **2005,** *13,* 381-388.
6. Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **2003,** *52,* 2-9.
7. Wodak, S. J.; Janin, J. Computer analysis of protein-protein interaction. *J. Mol. Biol.* **1978,** *124,* 323-342.
8. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002,** *47,* 409-443.
9. Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983,** *221,* 709-713.
10. Norel, R.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Shape complementarity at protein-protein interfaces. *Biopolymers* **1994,** *34,* 933-940.
11. Norel, R.; Petrey, D.; Wolfson, H. J.; Nussinov, R. Examination of shape complementarity in docking of unbound proteins. *Proteins* **1999,** *36,* 307-317.
12. Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Protein-protein docking dealing with the unknown. *J. Comput. Chem.* **2010,** *31,* 317-342.
13. Shoichet, B. K.; Kuntz, I. D. Protein docking and complementarity. *J. Mol. Biol.* **1991,** *221,* 327-346.

14. Palma, P. N.; Krippahl, L.; Wampler, J. E.; Moura, J. J. BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins* **2000,** *39,* 372-384.
15. Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A.* **1992,** *89,* 2195-2199.
16. Helmer-Citterich, M.; Tramontano, A. PUZZLE: a new method for automated protein docking based on surface shape complementarity. *J. Mol. Biol.* **1994,** *235,* 1021-1031.
17. Fischer, D.; Lin, S. L.; Wolfson, H. L.; Nussinov, R. A geometry-based suite of molecular docking processes. *J. Mol. Biol.* **1995,** *248,* 459-477.
18. Ausiello, G.; Cesareni, G.; Helmer-Citterich, M. ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure. *Proteins* **1997,** *28,* 556-567.
19. Gardiner, E. J.; Willett, P.; Artymiuk, P. J. Protein docking using a genetic algorithm. *Proteins* **2001,** *44,* 44-56.
20. Gabdoulline, R. R.; Wade, R. C. Protein-protein association: Investigation of factors influencing association rates by brownian dynamics simulations. *J. Mol. Biol.* **2001,** *306,* 1139-1155.
21. Fernandez-Recio, J.; Totrov, M.; Abagyan, R. Soft protein-protein docking in internal coordinates. *Protein Sci.* **2002,** *11,* 280-291.
22. Koshland, D. E. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **1958,** *44,* 98-104.
23. Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. J. Principles of flexible protein-protein docking. *Proteins* **2008,** *73,* 271-289.
24. Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem.* **1994,** *15,* 488-506.
25. Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L. Docking: successes and challenges. *Curr. Pharm. Des.* **2005,** *11,* 323-333.
26. Ritchie, D. W.; Kemp, G. J. Protein docking using spherical polar Fourier correlations. *Proteins* **2000,** *39,* 178-194.
27. Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **2003,** *331,* 281-299.
28. Pierce, B.; Weng, Z. A combination of rescoring and refinement significantly improves protein docking performance. *Proteins* **2008,** *72,* 270-279.
29. Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins* **2005,** *60,* 150-169.
30. Fernandez-Recio, J.; Sternberg, M. J. The 4th meeting on the critical assessment of predicted interaction (CAPRI) held at the Mare Nostrum, Barcelona. *Proteins* **2010,** *78,* 3065-3066.
31. Lensink, M. F.; Mendez, R.; Wodak, S. J. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* **2007,** *69,* 704-718.

32. Fernandez-Recio, J.; Totrov, M.; Abagyan, R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins* **2003,** *52,* 113-117.
33. Dominguez, C.; Boelens, R.; Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **2003,** *125,* 1731-1737.
34. de Vries, S. J.; van Dijk, A. D.; Krzeminski, M.; van Dijk, M.; Thureau, A.; Hsu, V.; Wassenaar, T.; Bonvin, A. M. HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* **2007,** *69,* 726-733.
35. Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.* **2004,** *32,* W96-99.
36. Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Geometry-based flexible and symmetric protein docking. *Proteins* **2005,** *60,* 224-231.
37. Levy, Y.; Onuchic, J. N. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.* **2006,** *35,* 389-415.
38. Zhang, L.; Yang, Y.; Kao, Y. T.; Wang, L.; Zhong, D. Protein hydration dynamics and molecular mechanism of coupled water-protein fluctuations. *J. Am. Chem. Soc.* **2009,** *131,* 10677-10691.
39. Bogan, A. A.; Thorn, K. S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **1998,** *280,* 1-9.
40. Janin, J. Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure* **1999,** *7,* R277-279.
41. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985,** *28,* 849-857.
42. Pitt, W. R.; Goodfellow, J. M. Modelling of solvent positions around polar groups in proteins. *Protein Eng.* **1991,** *4,* 531-537.
43. Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L., 3rd; Vajda, S. Algorithms for computational solvent mapping of proteins. *Proteins* **2003,** *51,* 340-351.
44. Schymkowitz, J. W.; Rousseau, F.; Martins, I. C.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl. Acad. Sci. U.S.A.* **2005,** *102,* 10147-10152.
45. Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. U.S.A* **2007,** *104,* 808-813.
46. Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **2008,** *130,* 2817-2831.
47. Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins* **1999,** *34,* 17-28.
48. Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002,** *46,* 34-40.

49. Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem.* **2005,** *48,* 6504-6515.
50. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004,** *47,* 1739-1749.
51. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004,** *47,* 1750-1759.
52. Li, Z.; Lazaridis, T. Water at biomolecular binding interfaces. *Phys. Chem. Chem. Phys.* **2007,** *9,* 573-581.
53. Chen, J.; Brooks, C. L., 3rd; Khandogin, J. Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.* **2008,** *18,* 140-148.
54. van Dijk, A. D.; Bonvin, A. M. Solvated docking: introducing water into the modelling of biomolecular complexes. *Bioinformatics* **2006,** *22,* 2340-2347.
55. Burnett, J. C.; Kellogg, G. E.; Abraham, D. J. Computational methodology for estimating changes in free energies of biomolecular association upon mutation. The importance of bound water in dimer-tetramer assembly for beta 37 mutant hemoglobins. *Biochemistry* **2000,** *39,* 1622-1633.
56. Fornabaio, M.; Spyrakis, F.; Mozzarelli, A.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. *J. Med. Chem.* **2004,** *47,* 4507-4516.
57. Amadasi, A.; Spyrakis, F.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Mapping the energetics of water-protein and water-ligand interactions with the "natural" HINT forcefield: predictive tools for characterizing the roles of water in biomolecules. *J. Mol. Biol.* **2006,** *358,* 289-309.
58. Chen, D. L.; Kellogg, G. E. A computational tool to optimize ligand selectivity between two similar biomacromolecular targets. *J. Comput. Aided Mol. Des.* **2005,** *19,* 69-82.
59. Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. Robust classification of "relevant" water molecules in putative protein binding sites. *J. Med. Chem.* **2008,** *51,* 1063-1067.
60. Spyrakis, F.; Cozzini, P.; Bertoli, C.; Marabotti, A.; Kellogg, G. E.; Mozzarelli, A. Energetics of the protein-DNA-water interaction. *BMC Struct. Biol.* **2007,** *7,* 4.
61. Ahmed, M. H.; Spyrakis, F.; Cozzini, P.; Tripathi, P. K.; Mozzarelli, A.; Scarsdale, J. N.; Safo, M. A.; Kellogg, G. E. Bound water at protein-protein interfaces: partners, roles and hydrophobic bubbles as a conserved motif. *PLoS One* **2011,** *6,* e24712.
62. Lensink, M. F.; Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. *Proteins* **2010,** *78,* 3073-3084.

63. Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. *J. Med. Chem.* **2002,** *45,* 2469-2483.
64. Fornabaio, M.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 2. Computational titration and pH effects in molecular models of neuraminidase-inhibitor complexes. *J. Med. Chem.* **2003,** *46,* 4487-4500.
65. Marabotti, A.; Spyrakis, F.; Facchiano, A.; Cozzini, P.; Alberti, S.; Kellogg, G. E.; Mozzarelli, A. Energy-based prediction of amino acid-nucleotide base recognition. *J. Comput. Chem.* **2008,** *29,* 1955-1969.
66. Kellogg, G. E.; Fornabaio, M.; Chen, D. L.; Abraham, D. J.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A. Tools for building a comprehensive modeling system for virtual screening under real biological conditions: The Computational Titration algorithm. *J. Mol. Graph. Model.* **2006,** *24,* 434-439.
67. Bayden, A. S.; Fornabaio, M.; Scarsdale, J. N.; Kellogg, G. E. Web application for studying the free energy of binding and protonation states of protein-ligand complexes based on HINT. *J. Comput. Aided Mol. Des.* **2009,** *23,* 621-632.
68. Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein-protein docking benchmark version 4.0. *Proteins* **2010,** *78,* 3111-3114.
69. Kellogg, G. E.; Abraham, D. J. Hydrophobicity: is LogP(o/w) more than the sum of its parts? *Eur. J. Med. Chem.* **2000,** *35,* 651-661.
70. Levitt, M. Molecular dynamics of native protein. I. Computer simulation of trajectories. *J. Mol. Biol.* **1983,** *168,* 595-617.
71. Levitt, M.; Perutz, M. F. Aromatic rings act as hydrogen bond acceptors. *J. Mol. Biol.* **1988,** *201,* 751-754.
72. Kellogg, G. E.; Chen, D. L. The importance of being exhaustive. Optimization of bridging structural water molecules and water networks in models of biological systems. *Chem. Biodivers.* **2004,** *1,* 98-105.
73. McLachlan, A. D. Rapid comparison of protein structures. *Acta. Cryst.* **1982,** *A38*, 871-873.
74. *JMP v10*, SAS Institute Inc.: Cary, NC, 1989-2007.
75. Lo Conte, L.; Chothia, C.; Janin, J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **1999,** *285,* 2177-2198.
76. Chen, R.; Weng, Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins* **2002,** *47,* 281-294.
77. Zhang, C.; Vasmatzis, G.; Cornette, J. L.; DeLisi, C. Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.* **1997,** *267,* 707-726.
78. Rodier, F.; Bahadur, R. P.; Chakrabarti, P.; Janin, J. Hydration of protein-protein interfaces. *Proteins* **2005,** *60,* 36-45.
79. Kastritis, P. L.; van Dijk, A. D.; Bonvin, A. M. Explicit treatment of water molecules in data-driven protein-protein docking: the solvated HADDOCKing approach. *Methods Mol. Biol.* **2012,** *819,* 355-374.

80. Li, Y.; Li, H.; Smith-Gill, S. J.; Mariuzza, R. A. Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63. *Biochemistry* **2000,** *39,* 6296-6309.
82. Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.5.0.4.

CHAPTER 5

CONCLUSIONS

Substantive progress is being made in the atomic and residue level understanding of biomolecular interactions. Computational tools that provide a quantitative assessment of the energetic contributions of individual interactions involved in molecular recognition have been extensively used in drug design and development process. An empirical HINT force field for quantifying non-covalent interactions, correlating with the $\Delta \boldsymbol{G}_{\boldsymbol{interaction}}$, has been developed by *Kellogg* and *Abraham*, based on experimental partition coefficient $LogP_{o/w}$ data.[1] Being derived from an experimental thermodynamic parameter ($LogP_{o/w}$), HINT is advantageous over other force fields/scoring functions because it accounts not only for the electrostatic, Coulombic, van der Waal's and hydrophobic interactions, but also implicitly takes into consideration the contributions of entropy and solvation/desolvation processes towards the binding event. The implementation of HINT force field and various tools based off it has led to better understanding of the protein-ligand, protein-protein and protein-DNA associations.[2-4] In this dissertation, we have discussed the use of HINT force field in

molecular modeling studies for novel anticancer and antiviral targets, as well as in developing a novel solvated protein-protein docking protocol.

In our efforts to gain structural insights into the isoform selectivity of a thiazolidine-2,4-dione analog (K145) towards Sphingosine kinase 2, an anti-cancer target, we successfully built structural models of both isoforms of sphingosine kinase (SphK1 and SphK2) using the crystal structure of a kinase of bacterial origin as the template. This was followed by molecular docking of the ligand in the sphingosine-binding domains of both kinases. The analysis of proposed binding modes using HINT force field suggested that K145 binds more favorably to SphK2, but not SphK1, consistent with the biochemical assay results. We understand that the protein models and ligand binding modes have not been experimentally validated, and therefore should be used with caution. However, in absence of a crystallographic structure, the models can be used as hypothesis generator for future lead optimization and compound design efforts towards developing thiazolidine-based inhibitors as anti-cancer agents.[5]

The structural modeling of Human Cytomegalovirus (HCMV) Alkaline Nuclease (AN) UL98 has also been discussed in the dissertation. A similar homology-based modeling of UL98 AN was performed, using the crystal structure of a homologous exonuclease from another herpesvirus as the template, in order to identify active-site residues involved in its nucleolytic activity that are very important for viral replication. Again, hydropathic analysis of the protein-DNA complex using HINT force field identified the key residues: R164, S252, D254, E278 and K280 that showed significant interactions with 5’ end of DNA, to be important for its exo- and/or endo- nuclease activity. Mutagenesis studies were performed to validate the model, with alanine

substitutions abolishing activity. The results confirm the utility of the model in representing the active site region of UL98 AN.[6] The protein-DNA complex model was then used to build a pharmacophore model complimentary to the protein active site, to virtually screen the NCI database of ~250,000 compounds. Molecular docking and subsequent scoring (using HINT force field) of the hits, yielded a number of novel interesting compounds, which have the potential to bind at UL98 AN active site and inhibit it. Top 15 compounds have been obtained from the vendor and are currently being evaluated experimentally. The use of an experimentally validated protein model and HINT scoring for identifying top hits, gives us confidence that our approach might result in identification of novel antiviral agents.

Finally, we have addressed the issue of explicit solvent accounting in protein-protein docking algorithms. Several studies, using HINT based tools, have shown the importance of water molecules at interfaces, in mediating and stabilizing biomacromolecular associations.[7-10] Majority of the current protein-protein docking approaches only implicitly incorporate the effects of solvent by introducing the desolvations terms in their scoring functions, which might be one of the reason for considerable lack in prediction accuracy. We hypothesize that docking algorithms that explicitly account for water at interface with yield more native-like models. On a data-set of 15 protein-protein complex crystal structures, we identified those waters that showed bridging interactions between both interacting partners, using the HINT Water Relevance Metric. We developed a solvated docking protocol – wherein the relevant waters were forced to be a part of one of the two proteins, followed by rigid-body docking using ZDOCK,[11] and ranking the predictions using HINT energy scores. The

predictions were checked for accuracy using the standard CAPRI criteria.[12] Comparing the results of solvated docking approach with a traditional unsolvated one, we have shown considerable improvement in the quality of predictions obtained. Our implementation of HINT based tools to the docking protocol is still at the simplest level, with complete chemical properties of waters being utilized only at the scoring stage. Even so, our study shows promise. It is safe to say that we can expect considerably higher prediction accuracy with further attempts to optimize the protein-protein docking algorithms by incorporating not just hydropathic complementarity, but also explicit solvent accounting, correct interfacial residue ionization states, and interfacial side-chain flexibility. This study is just a beginning of the ultimate goal of our lab – to develop a protein-protein docking algorithm based on HINT tools.

To summarize, the overall goal of this multidisciplinary work was the application of HINT force field and HINT-based tools to different aspects of molecular modeling – from structural modeling of novel anticancer and antiviral targets, to explicit solvent accounting in a protein-protein docking approach. The results of this research will provide the scientific community with additional knowledge to better predict and model biomacromolecular structures.

References

1. Kellogg, G. E.; Abraham, D. J. Hydrophobicity: is LogP(o/w) more than the sum of its parts? *Eur. J. Med. Chem.* **2000,** *35,* 651-661.
2. Burnett, J. C.; Kellogg, G. E.; Abraham, D. J. Computational methodology for estimating changes in free energies of biomolecular association upon mutation. The importance of bound water in dimer-tetramer assembly for beta 37 mutant hemoglobins. *Biochemistry* **2000,** *39,* 1622-1633.
3. Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. *J. Med. Chem.* **2002,** *45,* 2469-2483.
4. Marabotti, A.; Spyrakis, F.; Facchiano, A.; Cozzini, P.; Alberti, S.; Kellogg, G. E.; Mozzarelli, A. Energy-based prediction of amino acid-nucleotide base recognition. *J. Comput. Chem.* **2008,** *29,* 1955-1969.
5. Liu, K.; Guo, T. L.; Hait, N. C.; Allegood, J.; Parikh, H. I.; Xu, W.; Kellogg, G. E.; Grant, S.; Spiegel, S.; Zhang, S. Biological Characterization of 3-(2-amino-ethyl)-5-[3-(4-butoxyl-phenyl)-propylidene]-thiazolidine-2,4-dione (K145) as a Selective Sphingosine Kinase-2 Inhibitor and Anticancer Agent. *PLOS ONE* **2013,** *8,* e56471.
6. Kuchta, A. L.; Parikh, H.; Zhu, Y.; Kellogg, G. E.; Parris, D. S.; McVoy, M. A. Structural modelling and mutagenesis of human cytomegalovirus alkaline nuclease UL98. *J. Gen. Virol.* **2012,** *93,* 130-138.
7. Fornabaio, M.; Spyrakis, F.; Mozzarelli, A.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. *J. Med. Chem.* **2004,** *47,* 4507-4516.
8. Chen, D. L.; Kellogg, G. E. A computational tool to optimize ligand selectivity between two similar biomacromolecular targets. *J. Comput. Aided Mol. Des.* **2005,** *19,* 69-82.
9. Spyrakis, F.; Cozzini, P.; Bertoli, C.; Marabotti, A.; Kellogg, G. E.; Mozzarelli, A. Energetics of the protein-DNA-water interaction. *BMC Struct. Biol.* **2007,** *7,* 4.
10. Ahmed, M. H.; Spyrakis, F.; Cozzini, P.; Tripathi, P. K.; Mozzarelli, A.; Scarsdale, J. N.; Safo, M. A.; Kellogg, G. E. Bound water at protein-protein interfaces: partners, roles and hydrophobic bubbles as a conserved motif. *PLOS ONE* **2011,** *6,* e24712.
11. Pierce, B.; Weng, Z. A combination of rescoring and refinement significantly improves protein docking performance. *Proteins* **2008,** *72,* 270-279.
12. Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins* **2005,** *60,* 150-169.

APPENDIX A.1

Clustal X default color scheme

Residue at position	Color	{Threshold, Residue group}
A,I,L,M,F,M,V	Blue	{+60%, WLVIMAFCHP)
R,K	Red	{+60%, KR}, {+80%, KRQ}
N	Green	{+50%, N}, {+85%, NY}
C	Blue	{+60%, WLVIMAFCHP)
C	Pink	{100%, C}
Q	Green	{+60%, KR}, {+50%, QE}, {+85%, QEKR}
E	Magenta	{+60%, KR}, {+50%, QE}, {+85%, EQD}
D	Magenta	{+60%, KR}, {+85%, KRQ}, {+50%, ED}
G	Orange	{+0%, G}
H,Y	Cyan	{+60%, WLVIMAFCHP}, {+85%, WYACPQFHILMV}
P	Yellow	{+0%, P}
S,T	Green	{+60%, WLVIMAFCHP}, {+50%, TS}, {+85%, ST}

Each residue in the alignment is assigned a color if the amino acid profile of the alignment at that position meets the minimum criteria specific for the residue type.

VITA

Hardik Ishwar Parikh was born on September 12th, 1985, in Vadodara, Gujarat, INDIA, and is an Indian citizen. He graduated from Tejas Vidyalaya, Vadodara, Gujarat, INDIA in July 2003, with a major in science stream. He received his Bachelors of Pharmacy from the Sinhgad College of Pharmacy, University of Pune, Pune, INDIA in July 2007. He went on to work as a Research Trainee Executive in the Analytical Development Lab in Formulations and Development Department, Pharma Research, Alembic Ltd., Vadodara, INDIA until July 2008. In August 2008, he joined the School of Pharmacy, Virginia Commomwealth University, Richmond, VA for his graduate doctoral studies, with a concentration in Medicinal Chemistry.