2010

DEVELOPMENT AND APPLICATIONS OF THE HINT FORCEFIELD IN PREDICTION OF ANTIBIOTIC EFFLUX AND VIRTUAL SCREENING FOR ANTIVIRALS

Aurijit Sarkar
Virginia Commonwealth University

DEVELOPMENT AND APPLICATIONS OF THE HINT FORCEFIELD IN PREDICTION OF ANTIBIOTIC EFFLUX AND VIRTUAL SCREENING FOR ANTIVIRALS

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

by

AURIJIT SARKAR
M.Sc. Shri GS Institute of Technology & Science, Indore MP, India 2003
B.Sc. Devi Ahilya University, Indore MP, India 2000

Advisor: GLEN EUGENE KELLOGG, Ph.D.
ASSOCIATE PROFESSOR, DEPARTMENT OF MEDICINAL CHEMISTRY &
INSTITUTE FOR STRUCTURAL BIOLOGY AND DRUG DISCOVERY

Virginia Commonwealth University
Richmond VA

August 2010

“The woods are lovely dark and deep,

But I have promises to keep,

And miles to go before I sleep,

And miles to go before I sleep”

- Robert Frost

(1874-1963)

From *Stopping by Woods on a Snowy Evening*

Acknowledgements

This work is dedicated to all those people in my life who have shown unwavering faith in my capabilities and edged me on toward success.

My parents, Cdr. Biswajit Sarkar (Indian Navy, retd.) and Mrs. Sumita Sarkar have always been there to support me in times of need. More importantly, they have never failed to steer me in the right direction whenever I strayed. Without their guidance and support, I would never have reached this far. Their love and care are gratefully acknowledged. I hope I have proven myself worthy of being called their son.

Dr. Glen Kellogg, my doctoral advisor, has taught me much across the past five years. His contribution to my development as a scientist can never be matched. Without his constant support, I would be lost in this great and scary world of research. I will never forget his lessons, which will be a guiding light till my dying day.

I will always be indebted to Dr. Richard Westkaemper, Dr. Umesh Desai, Dr. H. Tonie Wright and Dr. W. Mike Holmes, who have served as members of my graduate student committee, for their patient efforts in transforming my many ineptitudes into (perhaps) some potential.

A lone man may find success in life, but never happiness. Friends play an important role in the development of an individual, often providing support and lending a sympathetic ear. However, Pinky Vinaykya’s role in my life stands out. I may not have liked the truths she made me face, but they transformed my life forever! Her role in my transformation into a mature adult remains unparalleled.

My life in Richmond and my successful completion of my doctoral degree would have been much harder, if not for my friends Nida, Maria and Tamara. Their acceptance of my eccentric ways and constant need for attention has been a boon. I gratefully acknowledge the role of their friendship in my life.

I would also like to thank School of Pharmacy and the Department of Medicinal Chemistry, Virginia Commonwealth University for the financial assistance I received during the initial phase of my PhD, without which I perhaps would not dare to venture so far away from home.

Table of Contents

List of Tables

List of Figures

List of Schemes

ABSTRACT

DEVELOPMENT AND APPLICATIONS OF THE HINT FORCEFIELD IN PREDICTION OF ANTIBIOTIC EFFLUX AND VIRTUAL SCREENING FOR ANTIVIRALS

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

by

AURIJIT SARKAR
M.Sc. Shri GS Institute of Technology & Science, Indore MP, India 2003
B.Sc. Devi Ahilya University, Indore MP, India 2000

VIRGINIA COMMONWEALTH UNIVERSITY, 2010

Advisor: GLEN EUGENE KELLOGG, Ph.D.
ASSOCIATE PROFESSOR, DEPARTMENT OF MEDICINAL CHEMISTRY &
INSTITUTE FOR STRUCTURAL BIOLOGY AND DRUG DISCOVERY

This work was aimed at developing novel tools that utilize HINT, an empirical forcefield capable of quantitating both hydrophobic and hydrophilic (hydropathic) interactions, for implementation in theoretical biology and drug discovery/design.

The role of hydrophobicity in determination of macromolecular structure and formation of complexes in biological molecules is undeniable and has been the subject of research across several decades. Hydrophobicity is introduced, with a review of its history and contemporary theories. This is followed by a description of various methods that quantify this all-pervading phenomenon and their use in protein folding and contemporary drug design projects – including a detailed overview of the HINT forcefield.

The specific aim of this dissertation is to introduce our attempts at developing new methods for use in the study of antibacterial drug resistance and antiviral drug discovery. Multidrug efflux is commonly regarded as a fast growing problem in the field of medicine. Several species of microbes are known to have developed resistance against almost all classes of antibiotics by various modes-of-action, which include multidrug transporters (a.k.a. efflux pumps). These proteins are present in both gram-positive and gram-negative bacteria and extrude molecules of various classes. They protect the efflux pump-expressing bacterium from harmful effects of exogenous agents by simply evacuating the latter. Perhaps the best characterized mechanism amongst these is that of the AcrA-AcrB-TolC efflux pump. Data is available in literature and perhaps also in proprietary databases available with pharmaceutical companies, characterizing this pump in terms of the minimum inhibitory concentration ratios (MIC ratios) for various antibiotics. We procured a curated dataset of 32 β-lactam and 12 antibiotics of other classes from this literature. Initial attempts at studying the MIC ratios of β-lactam antibiotics as a function of their three dimensional topology via 3D-

quantitative structure activity relationship (3D-QSAR) technology yielded seemingly good models. However, this methodology is essentially designed to address single receptor-ligand interactions. Molecules being transported by the efflux pump must undoubtedly be involved in multiple interactions with the same. Notably, such methods require a pharmacophoric overlap of ligands prior to the generation of models, thereby limiting their applicability to a set of structurally-related compounds. Thus, we designed a novel method that takes various interactions between antibiotic agents and the AcrA-AcrB-TolC pump into account in conjunction with certain properties of the drugs. This method yielded mathematical models that are capable of predicting high/low efflux with significant efficiency (>93% correct). The development of this method, along with the results from its validation, is presented herein.

A parallel aim being pursued by us is to discover inhibitors for hemagglutinin-neuraminidase (HN) of human parainfluenza virus type 3 (HPIV3) by *in silico* screening. The basis for targeting HN is explored, along with commentary on the methodology adopted during this effort. This project yielded a moderate success rate of 34%, perhaps due to problems in the computational methodology utilized. We highlight one particular problem – that of emulating target flexibility – and explore new avenues for overcoming this obstacle in the long run. As a starting point towards enhancing the tools available to us for virtual screening in general (and for discovering antiviral compounds in specific), we explored the compatibility between sidechain rotamer libraries and the HINT scoring function. A new algorithm was designed to optimize amino acid residue sidechains, if provided with the backbone coordinates, by generating sidechain

positions using the Dunbrack and Cohen backbone-dependent rotamer library and scoring them with the HINT scoring function. This rotamer library was previously used by its developers previously to design a very successful sidechain optimization algorithm called SCWRL. Output structures from our algorithm were compared with those from SCWRL and showed extraordinary similarities as well as significant differences, which are discussed herein. This successful implementation of HINT in our sidechain optimization algorithm establishes the compatibility between this forcefield and sidechain rotamer libraries. Future aims in this project include enhancement of our current algorithm and the design of a new algorithm to explore partial induced-fit in targets aimed at improving current docking methodology.

This work shows significant progress towards the implementation of our hydropathic force field in theoretical modeling of biological systems in order to enhance our ability to understand atomistic details of inter- and intramolecular interactions which must form the basis for a wide variety of biological phenomena. Such efforts are key to not only to understanding the said phenomena, but also towards a solid basis for efficient drug design in the future.

CHAPTER 1

HYDROPHOBICITY: THEORIES, ESTIMATION AND APPLICATIONS

1.1 HYDROPHOBICITY AND BIOLOGICAL PHENOMENA

Hydrophobicity (or lipophilicity) is a well-known and extensively studied phenomenon. It is commonly understood to be the tendency of non-polar molecules to form aggregates in order to reduce their surface of contact with polar molecules such as water [1]. Its manifestations include simple observable macroscopic phenomena such as the immiscibility of oil and water or modern techniques such as chromatographic separation. The importance of hydrophobic interactions at the atomic or molecular scale has long been recognized in various areas of science [1]. While the concepts have changed and the applications have expanded, the fact remains that hydrophobic interactions are often the driving force in a variety of physical and biological phenomena, although they are often complemented by hydrophilic interactions.

Our hypothesis is that a large majority of biological phenomena can be explained by explicitly addressing the hydrophobic and hydrophilic interactions between molecules. In fact, we suggest that predictive models can be developed in order to explain such phenomena by explicitly quantitating the extent of these interactions. This work is a compilation of results aimed at demonstrating the validity of this hypothesis.

To begin, the concept of hydrophobic interactions is presented here, along with a very short history and discussion of theoretical and experimental studies on the phenomenon. The same set of forces and interactions that partitions a molecule between polar and hydrophobic solvent phases, i.e. determines its hydrophobicity, is pervasive in all biological interactions including small molecule binding and protein folding. An in-depth perspective on computational studies involving hydrophobic interactions is presented. These studies include methods for estimation of the hydrophobic nature of small and large biological molecules and applications of this in drug discovery or design.

These are followed by an example of quantitative modeling in order to effectively address some complex biological phenomena, specifically antibiotic transport by efflux pumps, followed by an implementation of these interactions in drug discovery for antiviral agents. Finally, a weakness of current methodology used in the aforementioned discovery process is discussed, followed by laying grounds for development of new hydropathy-based tools to address the same.

1.2 A BRIEF HISTORICAL OVERVIEW OF HYDROPHOBICITY

Even before the turn of the 20th century, the importance of hydrophobic interactions in biological phenomena, particularly drug activity, was recognized by the work of Meyer and Overton [2,3,4]. In 1937, Butler showed a linear relationship between heat of hydration and entropy of hydration [5]. He estimated the energies of interaction of different functional groups with water and showed that the heats of

hydration are additive in nature. He also explained that the heats of hydration do not determine the free energy of interactions, but that there is a direct proportionality between them. The reasons were unclear at this point, but it was hypothesized that entropy might be dependent on the size of the "cavity" that contains the molecule. The importance of H-bonds was also briefly discussed as formation of H-bonds between polar parts of the molecule causes an increase in entropy, which favors dissolution of an otherwise non-polar molecule.

Frank and Evans, in the middle of the 20th century, described the formation of "icebergs" of water around non-polar parts of molecules [6]. Their findings were based on the deviation of entropy of vaporization for certain substances when dissolved in aqueous and non-aqueous solutions. The formation of a regularized lattice-like structure of water molecules surrounding non-polar moieties has been experimentally validated with crystallography [7] and is now more or less taken for granted. This theory was extended to proteins by Klotz, who explained the variation in pK_a, molecular volume, denaturation and the masking of expected behavior of protein functional groups in terms of this "iceberg" formation [8]. In fact, the association of two molecules can become energetically favorable due to the increase in entropy when these ordered water molecules are scattered or disordered (see Figure 1.1).

Kauzmann first coined the term "hydrophobic bond" in 1959, which caught the attention of many scientists at the time; this notion was supported by a number of research investigations of that era [9]. The work of Némethy, Scheraga and Steinberg also supported the use of this term [10]. Perhaps it was the tendency of the non-polar

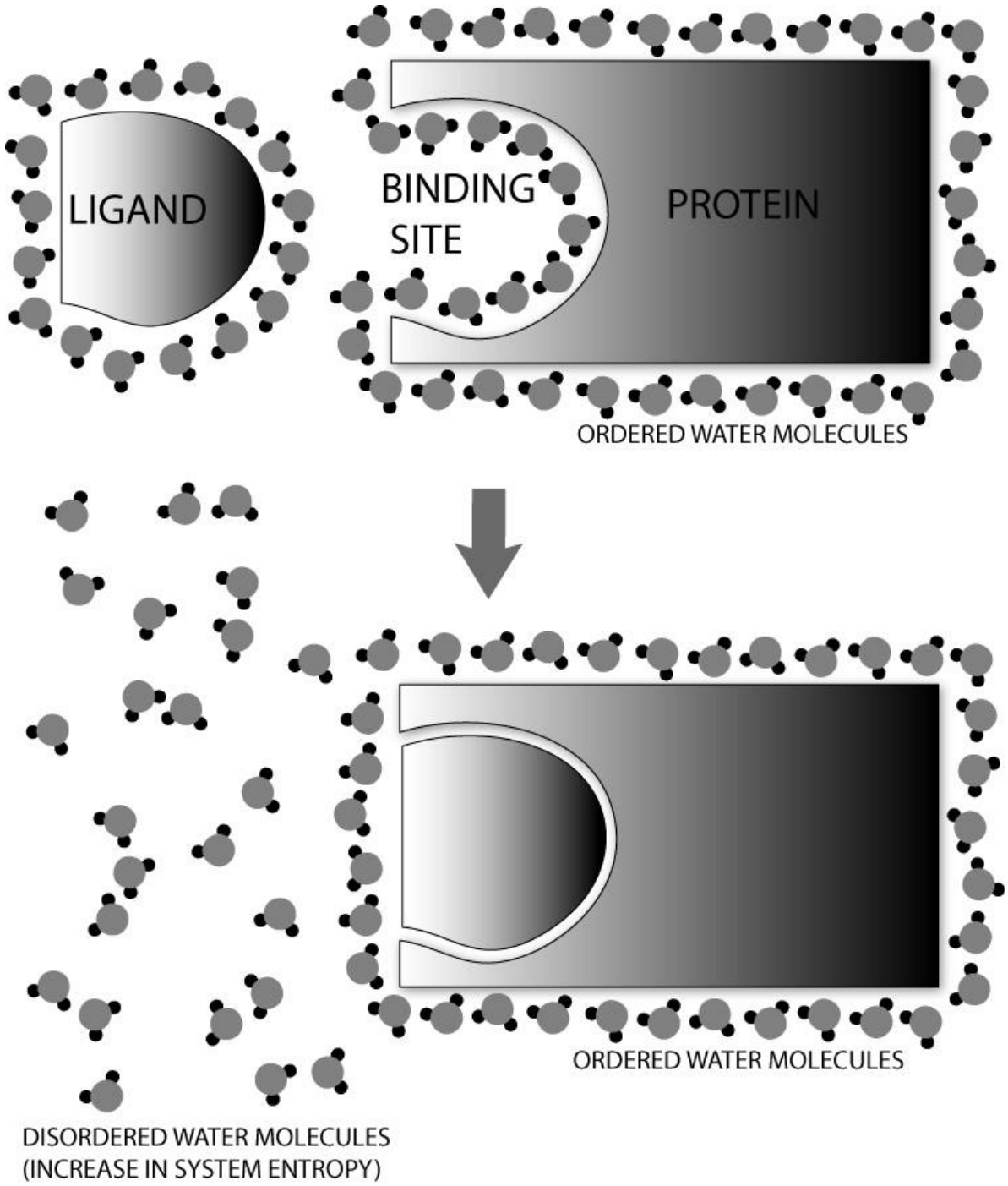


Figure 1.1 ***The Hydrophobic Effect.*** Hydrophobic molecules are surrounded by an ordered cage of water molecules. When two such molecules come together, they aggregate in order to reduce their surface area in contact with the polar water molecules. This causes a number of water molecules to be removed from their ordered formation, thus increasing disorder (increased entropy) and potentially making the process energetically favorable.

substances to form aggregates that caused scientists to draw parallelisms with the general definition of a bond – “the tendency of two atoms to stay together in space”. Hydrophobic bonds have been described as endothermic, i.e., as temperature increases their strength increases until a maximum value is reached at approximately 60°C [10]. However, the stability of proteins depends on not only these hydrophobic “bonds,” but also hydrogen bonds. These have an inverse behavior, i.e., they become *weaker* with increasing temperature. Thus, as temperature increases beyond 60°C, both H-bonds and hydrophobic interactions decrease in strength, causing proteins to unfold.

In the 1970s, Robert Hermann published a series of three papers on “the theory of hydrophobic bonding” [11-13] where the large negative entropy of partitioning a hydrophobic molecule into a non-polar solvent was explained by the loss of order in water molecules in direct contact with the hydrophobic surface. The ordered arrangement of water molecules on the surface of a molecule is due to dipole-dipole interactions with the immediate next layer of waters. In effect, this phenomenon is similar to surface-tension where the first layer arranges itself in order to reduce contact with the hydrophobic air, while less order exists in the second and succeeding layers. Order continues to decrease in layers away from the hydrophobic surface and there is a linear, but inverse, correlation between hydrophobic surface area and its solubility in water [11]. Hermann also determined that the free energy for hydration of a hydrophobic molecule is linearly related to the number of water molecules that can be packed around it. This first study did not take into account cavity curvature and was

restricted to small molecules. Later work [12] described a correlation between a molecule's hydrophobic surface area and its solubility in water. Hermann also addressed hydrophobic interactions at a distance [13] taking into account not only solubility, but also the distances between hydrophobic entities with the Lennard-Jones potential as has also been suggested by Reynolds *et al* [14]. Leo, Hansch and Jow established a relationship between hydrophobicity and two other factors – the nature of the solute surface and the molecular (CPK) volume [15]. The major innovation of this study is that they used the partition coefficient for 1-octanol/water ($LogP_{o/w}$) as a measure of hydrophobicity rather than solubility. This parameter has been used almost ubiquitously in studies thereafter. Most importantly, these observations could not be explained by the simple concept of a "hydrophobic bond", but rather as a complex phenomenon involving the interplay of flexible molecules and solvent under particular conditions.

The argument on semantics over the use of the term "hydrophobic bond" has continued ever since, but the fact that hydrophobic phenomena can explain a multitude of observations in science cannot be ignored. Here, we attempt to describe how naming and characterizing this effect has changed the realms of computational chemistry and drug design. A comprehensive review of the research on hydrophobicity is available elsewhere [1] for those interested in the intricacies of experimental approaches towards the phenomenon.

1.3 CALCULATIONS OF HYDROPHOBICITY AND THE HYDROPHOBIC EFFECT

1.3.1 Estimation of $LogP_{o/w}$

Hansch and Leo published their seminal paper on the determination and uses of partition coefficients in 1971 [16]. This paper was and perhaps continues to be, the most comprehensive article on the subject. It explains the fundamentals of partition phenomena and provides detailed descriptions of the history and theory of the same. It also contains a very comprehensive tabulation of LogP values for various substances. However, most interesting to theoreticians is the discussion of additive-constitutive properties wherein the utilization of the Hammett equation in calculations of partitioning free energy and the effects of various stereoelectronic effects on the partition coefficient are described. Also, various uses of partition coefficients for such diverse research topics as countercurrent distribution, measurement of equilibria, hydrophile-lipophile balance, drug dissolution and "hydrophobic bonding ability" are outlined. Of note, the partitioning of alcohols between water and red blood cells was compared to their partitioning between water and 1-octanol. The energy of partitioning per methylene group was the same for both cases, i.e., approximately -690 cal mol^{-1}. The repercussions of this quantification of hydrophobic interaction energies have been key to drug design projects as well as computational chemistry. The Hansch and Leo method for theoretical estimation of molecular LogP values, which is the basis of the C-LOGP method (*vide infra*), is also described in great detail.

A loose categorization of different methodologies for estimation of LogP is provided in Table 1, complete with a few typical examples of each. Here, the discussion

of these methods will be limited to a general overview highlighting the application of these methods in drug design and the relevance of accuracy for these prediction methods in that context. Several comprehensive reviews of the computational estimation of octanol-water partition coefficients are available [17-21].

To commence, it is a monumental understatement to say that a lot of good research has been done in this field. Many diverse empirical methods exist today that predict LogP of various molecules with different degrees of context-dependent certainty [17, 20]. Some of the major types are discussed below.

Table 1.1 Various types of methods for LogP calculations. This table shows a rough classification of methods used for theoretical prediction of LogP for compounds. Examples of all the different types are included.

Approach	**Methodology**	**Example(s)**
Substructure approaches	Fragment-based methods	Rekker's method [22], Leo's C-LOGP method [23-27], ACD/LogP method [28]
	Atom-based methods	XLOGP method [33-35], Ghose-Crippen method [29-32]
Whole molecule approaches	Molecular Lipophilicity Potential and related approaches	MLP [107-108]
	Topology descriptions	MS-WHIP [40]
	Molecular Property descriptions	Toulmin's ΔLogP method [41]

Fragment-based methods – Rekker's fragment based system was the first fragment-based computational method to estimate LogP [22]. Fragment-based methods implement and statistically deconvolve empirical data from experimental LogP values of compounds. Scheme 1.1 contains a short example of this approach. In order to explain the effect of inter-fragmental interactions, certain additive correction factors are introduced. Several other algorithms of this type exist including the C-LOGP [23-27] and ACD/LogP [28] methods. The criticism most often applied to this methodology is that the fragmentation of the target molecule is "arbitrary". This is not actually true for C-LOGP as there is a complete and unambiguous set of rules. However, they can be difficult to visualize and fragments can be much more complex than organic functional groups. Thus, fragments observed in new molecules can be missing from the C-LOGP database library, yielding poor predictions of LogP [17,20]. However, there are also advantages to these methods: significant and complex electronic interactions are automatically taken into account when they exist within a library fragment [20]; when the fragments coincide with real organic functional groups their interpretation is intuitive; the correction factors can be used to understand the relationship between functional groups or the effect of the observed feature on solubility, e.g., factors representing aliphatic chain branching explain the increased water solubility of branched hydrocarbons; and since fragment methods are based on empirical data, their associated algorithms are very fast and practical to implement in software.

Atom-based methods – These are similar to the fragment-based methods, but assume the hydrophobicity of a molecule to be the sum of the individual atomic

contributions. Scheme 1.2 provides an overview of the principle behind atom-based additive methodologies. Again, several methodologies of this type exist, including the well known Ghose-Crippen [29-32] and XLOGP [33-35] methods. Mostly these algorithms avoid correction factors by taking into account these sorts of contributions with a large set of atom types according to the individual environment it exists in within the molecule [20]. In order to somewhat reduce the atom type set the XLOGP algorithm implements a small number of correction factors. The reduced dependence on corrections is the major advantage of these methods. As described by Buchwald and Bodor, the major disadvantage of this method is that often the molecule is "more than a sum of its parts" [17]. Furthermore, human interpretability is reduced as the size of the atom database set grows and the correspondence with organic and medicinal chemistry principles is lost.

Molecular methods – Over the last two decades quantum mechanical calculations have been increasingly used in applied research including drug discovery, particularly with respect to estimations of interactions between solute and solvent molecules. A number of studies have used quantum chemical principles for estimation of molecular hydrophobicity [17]. Early work includes that of Rogers and Cammarata [36,37] and also that of Hopfinger and Battershell [38]. Klopman and Iroff used charge densities to calculate partition coefficients [39]. More recently, Bravi and Wikel described a method to predict LogP using a technology called Molecular Surface – Weighted Holistic Invariant Parameters (MS-WHIP) [40]. Unfortunately, a relatively large standard deviation between predicted and actual LogP was observed [18].

Toulmin *et al*. described another prediction method for octanol/water partition coefficients [41] that correlated minimized molecular electrostatic potentials with the H-bonding capability of molecules. In this method ΔLogP is defined as the difference between $LogP_{oct}$ (logarithm of the 1-octanol/water partition coefficient) and their predicted $LogP_{hxd}$ (logarithm of the hexadecane/water partition coefficient). H-bonding capability has a profound effect on partition coefficients with a strong correlation between ΔLogP and V_{min} (minimized molecular electrostatic potential). A strong correlation was also reported between ΔLogP or $LogP_{hxd}$ and CNS penetration of compounds, i.e., through the blood-brain barrier. This highlights the importance of H-bond donors and acceptors in normal partitioning phenomena.

Livingstone *et al.* described a method that uses neural networks (NN) to predict LogP values from a training set of electrotopological descriptors [42] of 900 drug and pesticide-like compounds [43]. Other studies involving artificial-intelligence utilize parameters calculated by various methods in unsupervised-learning processes to develop predictive models [44,45]. Taskinen and Yliruuski provide an in-depth analysis of such models in their review on NN modeling [46]. They note that, while NN methods are accurate in predicting LogP values of molecules within the size, functional group, etc. confines of the training set, they are less accurate in predictions for molecules outside the training set. However, this is the case for all LogP estimation methods.

Scheme 1.1 ***Fragmental methods for determination of LogP values***. Rekker's method is highlighted with an example adapted from Mannhold and van de Waterbeemd [18].

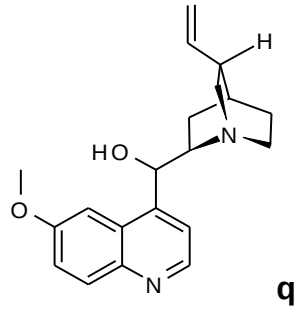
Fragmentation Methods

This approach breaks a molecule into fragments and assumes that the total LogP of a molecule is the sum total of all contributions of each fragment. However, the molecular environment affects the contributions by each fragment. Hence, correction factors are included in the calculation as shown by the following equation:

$$LogP = \sum_{i=1}^{n} a_i f_i + \sum_{i=1}^{m} b_j F_j$$

where,
LogP = log of the partition coefficient
a = the number of fragments,
f = fragmental constant
b_j = frequency of F_j
Fj = correction factor for the j^{th} fragment

A simple calculation by Rekker's fragmental method is illustrated in the panel at the right. The experimentally determined value of LogP for quinidine's is 3.44.



quinidine

Fragments:		
1 quinolinyl (-1H)		+1.617
1 O (aromatic)		-0.450
1 OH (aliphatic)		-1.448
1 N (aliphatic)		-2.074
	SUM	**-2.355**
CH residual:	$C_{11}H_{18}$	+4.893
	SUM	**2.538**
Corrections:		
Proximity effect ($+2C_M$),		+0.438
Electronegativity facing bulk ($-2C_M$),		-0.438
O-C-Ar ($+1C_M$)		+0.219
	LogP	**2.757**

Scheme 1.2 ***Atom contribution methods***; the calculation of LogP for quinidine by atom contributions is shown (adapted from Mannhold and van de Waterbeemd [18]).

Atom Contribution Methods

This is an extension of the fragmental contribution method. It is assumed that the total LogP of the molecule is a contribution by each individual atom comprising it (instead of a contribution by fragments). Calculation as shown by the following equation:

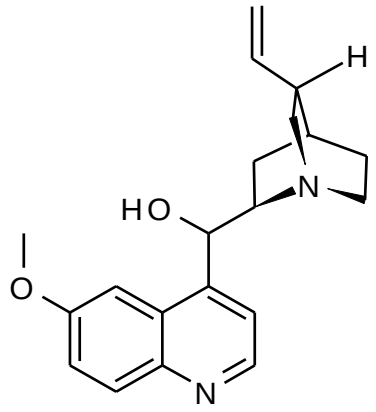
$$LogP = \sum n_i \, a_i$$

where,

n_i = the number of atoms of type i

a_i = fragmental constant

A simple calculation by the Ghose / Crippen method is illustrated in the panel at the right. The experimentally determined value of Log P for quinidine is 3.44.



Quinidine

Type	Description	Frequency	Contribution
2	C in CH_2R_2	2	-0.9748
3	C in CHR_3	2	-0.7266
5	C in CH_3X	1	-1.0824
6	C in CH_2RX	2	-1.6740
8	C in CHR_2X	2	-1.0420
15	C in $=CH_2$	1	-0.1053
16	C in =CHR	1	-0.0681
24	C in R--CH—R	4	+0.0272
25	C in R--CR—R	2	+0.3200
26	C in R--CX—R	2	-0.2066
27	C in R--CH—X	1	+0.0598
46	H attached to C^0_{sp3} with no X next to C	1	+0.4410
47	H attached to C^1_{sp3} or C^0_{sp3}	16	+5.3488
48	H attached to C^2_{sp3}, C^1_{sp3} or C^0_{sp3}	1	+0.3161
50	H attached to heteroatom	1	-0.3260
52	H attached to C^0_{sp3} with one X next to C	5	+1.8475
56	O in alcohol	1	+0.1402
60	O in Al - Al, Ar_2O, R:R or R-O-C=X	1	+0.2712
68	N in Al_3N	1	+0.3954
75	N in R--N--R or R--N—X	1	-0.1106
		LogP	**2.852**

Key symbols: R is group connected to C; X is heteroatom; "=" is double bond; ":" is an aromatic single bond such as the C-N bond in pyrrole; subscripts give the hybridization state and superscripts the formal oxidation number.

Hydrophobicity of amino acids and proteins – Understanding the hydrophobic behavior of amino acids, peptides and proteins has implications far beyond the seemingly simple task of calculating LogP for twenty or so small molecules (the amino acids). Abraham and Leo extended the Hansch and Leo fragment-based method of LogP calculation to amino acid zwitterions and side-chains [47]. Excellent agreement was reported for 19 out of the 20 natural amino acids. Proline, however, was calculated to be more hydrophilic than in reality, probably due to poor fragment parameterization for its secondary cyclic amine. With this method, hydrophobicity values for amino acid side chains were best predicted if a field effect was applied to the alpha-carbon. The field effect is the sum total of polar proximity effects of both the backbone amidic (peptide) bonds surrounding the alpha-carbon atom of any given residue. This field effect parameter accounts for the charge distribution on side-chain atoms and hence directly affects the hydrophobic/hydrophilic nature of the residue. Application of this effect allowed a higher correlation between predicted and calculated values of hydrophobicity for side-chains. In additional studies, Buchwald and Bodor reported a correlation between the van der Waal's volume of peptides with their LogP values [48]. Another approach was adopted by Steinmetz, where 3-D QSAR Comparative Molecular Field Analysis (CoMFA) studies were applied in a similar manner [49]. Experimentally determined LogP values of free and blocked di- and tripeptides were analyzed statistically to produce another set of parameters [50]. Akamatsu's work on the solvent partitioning of peptides using regression analysis of the experimental data to abstract

the hydrophobic parameters [51-54] is commonly regarded as the most convincing and accurate [55]. A comparison between software programs in predicting peptide LogPs was published recently [56]. In general, fragment-based methods are sensitive to composition but not to peptide sequence, which can be considered to be a major flaw of these programs. Also, it is important to note that most current programs are inefficient and ineffective in calculating LogP values for long peptides.

Summary of LogP estimation methods – A lot of effort has gone into devising methods for high prediction accuracy for LogP. However, most methods are accurate for members or close relatives of their own training sets but continue to be less accurate outside their training sets. It should also be pointed out that a considerable portion of the predictive inaccuracy may in fact lie with the data itself. Such data has often been obtained with experimental procedures whose accuracy varies with the method used [57]. One example brings this into focus: as many drugs and drug-like molecules contain ionizable functional groups, the conditions of measurement, particularly pH, are extremely relevant to measured LogP. Thus, if a user attempts to estimate LogP for a molecule, e.g., by specifying a carboxylic acid-containing species, what LogP value should be reported? The molecule in its acid form? The molecule in its ionized (nominally pH 7) form? Or the weighted average representing the equilibrium between the two forms? (This is what the experiment, as performed on the molecules in the training set, measures.) As other functional groups on the molecule can shift that equilibrium, how does this affect the contribution of the carboxylic acid/carboxylate fragment (or constituent atoms) to the predicted LogP?

While whole-molecule approaches are designed to estimate LogP values with great accuracy without extensive piecemeal (atom- or fragment-wise) empirical parameterization, their predictive nature in the end must also be compared to experimental data, limiting our ability to really judge the accuracy of predictions. This begs the question: do we need to emphasize accuracy of predictions so much? We suggest that when it comes to drug design, it is largely the ΔLogP changes between analogues that will drive the evolution in design with respect to physicochemical properties of the molecule. Virtually all methods of estimating LogP can accurately describe the replacement of a proton by a hydroxyl, the halogenation of an aromatic ring, substitution of an amine for a methyl, or nearly any of the chemical modifications that would be performed in fine-tuning a lead compound. The prediction of LogP for random organic compounds is probably not an important real world exercise. Also, representing such an important physicochemical property as a simple scalar value underutilizes the information content of the molecule's 3-D topology and, particularly, its hydropathic structure. The combination of topology and hydropathy provides us with structural details of immense importance, which play a direct role in intermolecular interactions, e.g., ligand binding, protein-protein associations, etc. However, we do recognize the importance of LogP in QSAR studies and also in assessing the drug likeness of a compound, both of which will be discussed below.

Can predictive methods for estimating the LogP of a peptide translate into a meaningful number for protein hydrophobicity? The idea that an additive atom-based or fragment-based algorithm (or even a whole molecule approach) could describe the

dissolution of a protein into water and/or 1-octanol is probably preposterous. To start, it is likely that a severe conformational change would occur if macromolecules pass from aqueous to organic solvents, e.g., hydrophobic residues would rearrange to the surface while the hydrophilic ones attempt to optimize hydrogen bonding and/or electrostatic interactions at the core. In other words, a protein would be an entirely different chemical species when interacting with solvents of different polarity – if it could actually be solubilized. However, the atomic, fragment or residue-level components of such a total LogP should be useful descriptors for understanding the forces and energetics of protein secondary, tertiary and quaternary structure and have been used in various schemes of describing and predicting protein folding for more than 20 years.

While proteins might change their conformation drastically on partitioning between aqueous and 1-octanol phases, similar conformational changes are also expected in small molecules. Such changes occur regularly for small molecules in both aqueous and organic phases due to the comparatively lower energy barriers which separate these conformations. However, it is not hard to imagine that organic phases would stabilize conformations where intramolecular hydrogen bonds, ionic salt bridges or dipole-dipole interactions exist because Coulomb interactions are strengthened in these circumstances, while aqueous phases would stabilize those conformations which show a higher degree of hydrophobic interactions. On the other hand, LogP is a self-contained parameter which accounts for all such conformational preferences because it is a bulk property and thus is the result of equilibrium between all such probable outcomes.

1.4 HYDROPHOBICITY SCALES AND PROTEIN FOLDING

There has long been evidence that protein secondary structure is dependent on the hydrophobic properties of the amino acid residue side chains. There is, in fact, a reproducible pattern of these properties in well-defined secondary structural elements such as α helices and β sheets. Thus, considerable effort has been expended in developing hydrophobicity scales that can aid in predictions of protein folding patterns. Some of these scales are based on water-ethanol transfer free energies [58,59], while others are based on partitioning between the bulk aqueous phase and the air-water interface [60], or on water-vapor partition free energies [61]. Kyte and Doolittle discussed the weaknesses of all three of these in a paper that also introduced their own hydrophobicity scale [62]. In their view, water-ethanol transfer free energy-based methods suffer because some amino acids are known to be insoluble in both water and ethanol and the latter may not be a truly inert solvent. Using partition data from transfer between the aqueous phase and air-water interfaces was also problematical because the hydrogen bonds that must be broken and the charges that must be neutralized to remove a residue from the aqueous phase during the formation of the native structure probably remain unchanged at an air-water interface. Thus, they would not be a factor in the overall reaction.

The “hydropathy” parameter of Kyte and Doolittle [62] is an amalgam of water-vapor transfer free energies and the interior-exterior distribution of amino acid side-chains determined by Chothia [63]. A moving-segment approach that continuously determines the average hydropathy while it advances through a sequence is used to

obtain a plot of hydropathy as a function of sequence. On this plot, any parts of the sequence that are above the average hydropathy for the sequence are termed hydrophobic and correspond well with experimentally determined “internal” regions of proteins. Conversely, sequence elements with hydropathy below the average are termed hydrophilic and correspond well with areas of the protein that are “exterior” and likely to be in contact with the polar solvent. The motivation is that analysis of these data may indicate the “folding” pattern of the sequence. To further exploit this, Wimley and White reported a new forcefield derived from partitioning two series of model peptides into the interface of neutral (zwitterionic) phospholipid membranes [64]. An alternative approach was introduced in 1986 by Eisenberg and McLachlan [65] for calculating the stability of protein structures in water based on atomic coordinates. The contribution of each protein atom to the solvation free energy is estimated as the product of the solvent accessibility of the atom and an atomic solvation parameter.

Li and Deber [66] used circular dichroism (CD) data to rank order helical propensity of proteins within membranes. Residues such as Ile, Val and Thr, which usually exist as β-sheets in an aqueous environment, prefer an α-helical conformation in lipid membranes. Thus, the helical propensity of amino acid residues correlates with the hydrophobic nature of the side chain. More recently, Dyson, Wright and Scheraga have explained [67] how strict classification of side chains as polar or non-polar has obscured certain facts about protein folding. For example, methylene groups present in large polar or charged amino acid side chains, like the four methylenes in the lysine side chain, can be considered non-polar. Interestingly, this fact was imbedded as one of the

factors in the Hansch and Leo system for estimating LogP [68] nearly 30 years earlier! These methylenes can aggregate with other non-polar groups and assist in hydrophobic collapse of the sequence.

Felitsky *et al.* introduced the use of a new parameter called “average area buried upon folding” (AABUF) [69] that explains both local contacts and long-range interactions. AABUF was used to study folding of apomyoglobin and provided additional insight into hydrophobic collapse and early folding events. Studies on polyalanine and polyleucine helices in water by MacCallum *et al.* [70] confirmed that in folding many unfavorable enthalpic events are counterbalanced by favorable entropic contributions by the solvent. This indicates a very small free energy barrier for folding. Thus, folding is mainly a desolvation phenomenon. Similarly, the Mardia and Nyirongo procedure for generating virtual protein Cα traces simulates the hydrophobic effect during folding [71] and produces models that are globular and compact.

Another related application of hydrophobicity is in the development of algorithms to simulate folding of hydrophobic-polar (HP) models in 2 and 3 dimensions [72]. The concept is to simplify the complex problem of folding by reducing it to representing residues by spheres with H (hydrophobic) and P (polar) character. The ensuing simulations are based on the observation that hydrophobic forces are the major forces determining native conformation of small globular proteins. These model simulations have been used to develop mathematical strategies for solving the combinatorial explosion problem, rather than actually simulating the hydrophobic effect [73-74].

As these studies have progressed over the past 20 years or so, the understanding of the hydrophobic effect and its impact on protein structure has matured. The early emphasis of using hydrophobicity scales to define folding patterns has shifted to algorithms that define protein folding in terms of mathematical approaches to reduce the calculational combinatorial explosion caused by exhaustive sampling of conformational space. However, it must be repeated that the same forces and energetics that drive solvent partitioning in the shake flask are at the core of protein folding. The difficulty is to unravel them and define algorithms that can simulate folding in these terms.

1.5 LogP IN DRUG DESIGN

Small molecule hydrophobicity has long been a consideration in drug discovery and design. The relationship between anesthetic effect of certain gases and their hydrophobicity has been extremely well established [4,75,76]. As described by Meyer in 1937 [3], chemically inert substances accumulate in “lipoids” and at a certain concentration, produce narcosis. The concentration itself is dependent on the animal, but independent of the narcotic itself. Hansch *et al.* also confirmed the Meyer-Overton hypothesis about a direct relation between hydrophobic nature of a compound and its anesthetic capabilities [75] through statistical correlations. However, Hansch suggested the additional involvement of a polar factor because molecules with polar hydrogens showed greater anesthetic action. Hansch *et al.* also introduced a similar theory for the hypnotic effect of barbiturates [76]. Other studies also have shown the important correlation of partition coefficients with binding affinities of drugs to receptors [77,78].

McFarland used a very simple probabilistic treatment of drug diffusion from the site of administration to the site of action via a collection of hydrophobic and hydrophilic barriers (Scheme 1.3, [79]). His relationship included a ratio for the rates of permeation for drugs between aqueous and organic phases (k/l) which relates to the partition coefficient at equilibrium. Inclusion of the Hammett equation into this study gave an intuitively satisfying parabolic relationship between drug potency and hydrophobicity (Scheme 3): higher doses of drugs with unfavorable partition coefficients (either too high or else too low) are required for them to reach the site of action. Recently, Kier has proposed a general theory of inhaled anesthetics [80].

1.5.1. The Lipinski "Rule of 5"

Hydrophobicity, of course, has also been a key factor in Lipinski's "rule of 5" [81,82]. In simple terms, Lipinski's rule can be stated as such: Poor absorption or permeation is more likely for a chemical entity when: a) there are more than 5 H-bond donors (sum of OHs and NHs); b) the molecular weight is over 500; c) the LogP is over 5; or d) there are more than 10 H-bond acceptors (sum of Ns and Os). The only exceptions to these rules were said to be substrates for biological transporters and natural products, which have a tendency to be highly complex molecules with multiple stereogenic centers and rarely contain nitrogen [83,84].

Scheme 1.3 ***A parabolic relationship exists between drug potency and hydrophobicity*****.** McFarland's equation relating probability and partition coefficient [79].

Model of *n* alternating aqueous phases and lipophilic membranes, in a hypothetical biological system is shown below:

Aq_0	Lip_1	Aq_2	Lip_3	Aq_4……Lip_{n-2}	Lip_{n-1}	Aq_n

Assume that the rate of passage of molecules from aqueous to lipophilic zones is *k*, while the rate of passage of molecules in the opposite direction is given by *l*. Then, the partition coefficient of the molecule will be given by *k/l*.

If $P_{x,y}$ is the probability of moving a molecule from layer x to layer y, the probability of moving a molecule from aq_0 to aq_n is given by:

$$P = P_{0,1}.P_{1,2}.P_{2,3}.P_{3,4} \dots P_{n-2,n-1}.P_{n-1,n}$$

Although passage of molecules would actually be affected by a number of factors, unbiased passage of the molecule is assumed here. This reduces the entire problem to one of pure probability. So, we will have:

$$P_{0,1} = P_{2,3} = P_{n-2,n-1}$$

Similarly, the following equation can also be obtained:

$$P_{1,2} = P_{3,4} = P_{n-1,n}$$

Combining all three above equations, we have:

$$P_{0,n} = (P_{0,1})^{n/2}.(P_{1,2})^{n/2}$$

Now, the number of molecules being transferred from aq_0 to lip_1 is proportional to *k*. The total number of molecules is proportional to the sum of *k* and *l*. The probability of a molecule moving from layer 0 to 1 is given by $P_{0.1}$, which is mathematically defined as:

$$P_{0,1} = \frac{k}{k+l}$$

Dividing both numerator and denominator on the right hand side by *l*, we get:

$$P_{0,1} = \frac{k/l}{k/l+1}$$

If $P_{1.0}$ is the probability of a molecule passing from lip_1 to aq_0,

$$P_{1,0} = 1 - P_{0,1}$$

Since we have assumed equal probabilities, we have:

$$P_{1,2} = 1 - P_{0,1}$$

Substituting this equation into the fourth equation described above, we have:

$$P_{0,n} = \left(P_{1,2}\right)^{n/2}.(1 - P_{1,2})^{n/2}$$

$$P_{0,n} = \left(\frac{k/l}{k/l+1}\right)^{n/2}.\left(1 - \frac{k/l}{k/l+1}\right)^{n/2}$$

$$P_{0,n} = \left(\frac{k/l}{k/l\ +1}\right)^{n/2}.\left(\frac{1}{k/l\ +1}\right)^{n/2}$$

$$P_{0,n} = \frac{(k/l)^{n/2}}{(k/l\ +1)^{n}}$$

This relationship will be parabolic, indicating that there is an optimum range of LogP with respect to drug potency.

In 2000 Lipinski introduced changes to address terms such as ‘drug-like’ because it was predicted that ADME (Absorption, Distribution, Metabolism and Excretion) screening of molecules (into drug-like or non-drug-like) would precede screening for activity at biological receptors [82]. The rule of 5 was further extended [85] to define a number of useful parameters: a) the presence of greater than 10 rotatable bonds reduces oral bioavailability; b) 0 < LogD < 3 enhances the probability of good intestinal permeability (LogD is logarithm of the distribution coefficient D, which is in turn defined as the ratio of the sum of concentrations of all forms, whether charged or neutral, or different functional conformations of the substance distributed between two mutually immiscible phases); c) a polar surface area (PSA) of less than 60-70 describes CNS active compounds; d) an N+O count of less than or equal to 5 enhances the probability of passing the blood-brain barrier; e) if LogP – (N + O) > 0, the molecule tends to be CNS active; f) orally-active drugs have lower molecular weight and fewer H-bond donors, acceptors and rotatable bonds; g) pulmonary drugs tend to have a larger PSA; and h) if the molecular weight < 300, LogP < 3, H-bond donors and acceptors < 3 and rotatable bonds < 3, the compound can be called “lead-like”.

This revolutionary work, which brilliantly summarized over 100 years of Medicinal Chemistry trial and error, made possible a number of rational filters and screens that, in principle, would improve the likelihood that a compound with promising “activity” could produce a “lead” and eventually yield a “drug”. Muegge described various methods for classification of drug-like compounds in his 2003 publication [86]. Similar publications addressing the terms ‘drug-like’ and ‘tool-like’ were also made [87,88]. Oprea *et al.*

reported the presence of a “medicinal chemistry lead-like space” and urged careful use of Lipinski’s rules [89]. A very interesting discussion [82] on how the properties of drug candidates from two pharmaceutical companies have varied across time pointed out that stress on rational methods of drug design in Merck laboratories caused no significant change in MLogP (Moriguchi LogP [90]) values across time. In contrast, there was a measurable increase in MLogP values for candidates from Pfizer since almost 50% of their hits were discovered with high-throughput screening (HTS) methods. Because the easiest method to increase in-vitro potency is to appropriately position a hydrophobic moiety onto a lead compound, HTS methods almost invariably select more hydrophobic candidates. Similar trends were observed [91] in that more than half of the molecules reported to have high-activity towards the end of the last century had a high LogP (> 4.25), high molecular weight (>425) and log of solubility in its neutral state (estimated from its molecular weight and LogP values), i.e., $LogS_w$ (< -4.25), only about 35% of the true lead compounds had these properties. It was also noted that as these molecules go through clinical trials, there is a distinct decrease in LogP values for compounds that make it to the market. One thing is clear from these studies and an analysis by Proudfoot of drugs currently on the market [92]: the lipophilicity of molecules that make it all the way to commercialization has remained in the same range for a number of years. In other words, there is a delicate balance between the hydrophobic and hydrophilic nature of a molecule that is absolutely essential for it to be transported to the site of action by diffusion across membranes.

1.5.2 Hydrophobicity in QSAR

Similarity between molecules is often perceived by chemists both qualitatively and quantitatively. A synthetic chemist would describe two molecules as similar if they have similar topologies, bond connectivities, functional groups or maybe synthetic strategies. Structure-Activity Relationships (SARs) are based on such comparisons in the context of physiological function, but are mostly limited to qualitative or semi-quantitative treatments of biological phenomena or activities. However, more stringent definitions of similarity have been formulated and can be used with chemical computing software to perceive (and even predict) chemical equivalence provided the likeness is scrutinized critically. Thus, a more mathematical and quantitative approach called the Quantitative Structure-Activity Relationship (QSAR), wherein affinities of ligands for their binding sites, inhibition constants, rate constants and other biological activities are correlated to molecular properties such as lipophilicity, polarizability, electronic and steric properties, was developed. Comprehensive reviews have been published on the subject in the past [93,94], which should be referred to by those wishing to learn about the QSAR concept in depth. Here, we will focus on the key role of hydrophobicity in these studies.

There are many different approaches used in classical QSAR studies, including establishment of relationships between activity and physicochemical properties such as steric properties (Hansch analysis, extrathermodynamic approach), structural features (Free Wilson analysis) [94], or topological descriptors (Kier-Hall indices) [42]. 3D QSAR methods, especially those such as CoMFA, consider three-dimensional ligand

structures and use those to propose the binding modes of those ligands at a common protein active site [94]. Data is often analyzed by statistical methods such as Multiple Linear Regression (MLR), Partial Least Squares (PLS) or by use of artificial intelligence (AI) methods such as Neural Networks (NN) or Support Vector Machines (SVM) [95] in order to detect correlations between a target activity and various descriptors (like LogP).

Hansch and Fujita first introduced their method and coined the term QSAR, for correlation of biological activity to chemical structure in the 1960s [78,96-98]. The method correlated, by the use of regression analysis, ligand structural variations with the biological activities of those ligands. In time, these studies would become a distinct scientific field and a mainstay of drug discovery and design research. Many applications have been reported across the past five decades. A review by Kubinyi has described, in great detail, the various subtleties of the science [99]. Indeed, in the absence of a detailed target or receptor structure, this ligand-based drug design method gives invaluable quantitative information to drug designers. It is important to note that the first publication on QSAR in 1962 [78] showed the importance of hydrophobicity through LogP. Scheme 1.4 explains the general concept behind the Hansch Analysis technique where the free energy-based substituent constant π is based on the Hammett function σ. π is dependent on the substituent's chemical nature and, since molecules must repeatedly partition between lipid membranes to be effective drugs, the constituting fragments of π should be such that their additive effect would allow easy partitioning into either membranous or aqueous phases.

Scheme 1.4 *Hansch Analysis*; a method to relate physicochemical parameters to drug potency. For details, refer to Hansch and Fujita [95].

The 'extrathermodynamic approach' relates various free energy-like descriptors in a model:

$$\log\frac{1}{C} = a(logP)^2 + blogP + c\sigma + \cdots + k$$

(C, LogP and σ are the inhibition constant, log of the partition coefficient and steric parameters respectively, while a,b,c and k are constants.) This model explains that drug transport from site of application to the site of action depends on the lipophilicity of the drug and is non-linear under typical conditions. Although special conditions could reduce this equation into simpler forms [73], this equation surmises the behavior of any molecule under normal diffusion conditions.
A novel parameter π defines the lipophilicity of substituent X:

$$\pi_X = logP_{RX} + logP_{RH}$$

where LogP is the log of the partition coefficient. This equation was a variant of the Hammett Equation,

$$\rho\sigma = logK_{RX} + logK_{RH}$$

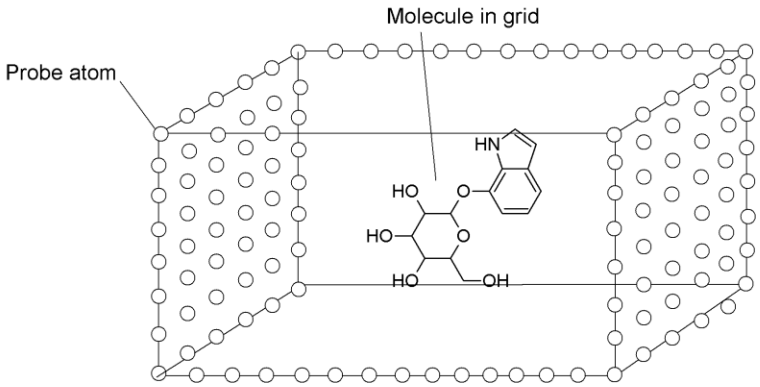
where the reaction equilibrium constants have been substituted with partition coefficients.

Values for LogP and σ of different molecules may be correlated with their IC_{50} or K_i values by statistical analysis such as multiple linear regressions (MLR) or partial least squares (PLS). Artificial intelligence methods such as Neural Networks (NN) and Support Vector Machines (SVM) have also been used. These analysis methods have pros and cons: While MLR and PLS do a good job of finding linear relationships between variables, they tend to oversimplify. On the other hand, artificial intelligence methods tend to pick up on minute non-linear trends and tend to over-fit models.

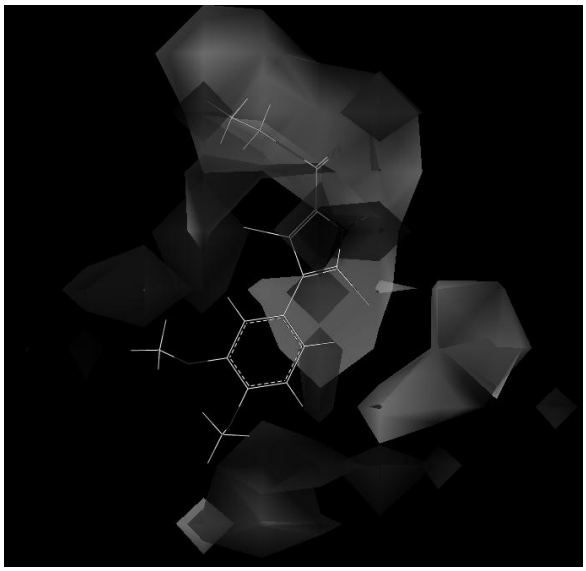
It must be noted, however, that hydrophobicity is not always the principal parameter determining activity [100]. For example, when DNA is the drug target, e.g., in binding to the major or minor groove, QSAR analyses often show negligible hydrophobic terms because the negatively charged phosphate groups of DNA are hydrophilic. On the other hand, DNA intercalation would likely be a hydrophobic effect. Radical reactions also typically lack hydrophobic terms in QSAR analyses, although these studies are mostly on small datasets and more thorough studies would be

desirable. Finally, it has been suggested from QSAR studies on Multiple Drug Resistance that this process might be accomplished without hydrophobic assistance, although this conflicts with the fact that efficiency of efflux pumps is often correlated with the hydrophobicity of their substrates [101].

3D QSAR methods like Comparative Molecular Field Analysis (CoMFA) [102] generate 3D field maps around aligned molecules to display zones of steric, electronic and lipophilic tolerance or intolerance. This gives a visual understanding of biological activity that contrasts well with the often messy collection of molecular descriptors in classical QSAR studies, thereby allowing easier interpretation of results. This, in turn, may lead to a better basis for designing novel scaffolds and/or chemical substituents to the existing scaffold. The basic idea behind this method is explained in Figure 2. Kellogg *et al.* introduced a method for hydrophobic field calculation for CoMFA [103] using an empirical force field (Hydropathic INTeractions or HINT, discussed in Section 1.7). This was one of the first attempts in 3D QSAR to modify the purely enthalpic treatment of ligand-receptor binding by inclusion of an implicit entropic term. References to the use of HINT-CoMFA in drug design are available [103-106]. Another attempt to include hydrophobicity into CoMFA was made by Gaillard, Testa and coworkers in their papers [107,108] describing the use Molecular Lipophilicity Potential (MLP) in 3D QSAR along with its applications. This alternative method of using hydrophobicity in CoMFA studies has found a number of applications in molecular modeling and drug design work [109-111].



Molecules are overlapped and placed in a grid, which is spread all around the overlapped molecules up to the extent of several Angstroms in all directions. Each grid point is treated as a probe; neutral Carbon atoms are used as probes for van der Waal's interactions, while charged atoms can be used as Coulombic interaction probes. Groups can also be used as probes, especially when trying to elucidate H-bond donors or acceptors. Simple physics equations for interactions of different varieties can calculate energy of each grid point, thereby extracting data for analysis. This data can then be checked for trends using PLS, MLR or AI algorithms.



The resultant map, shown above, is a map of regions where certain physicochemical parameters are tolerated (or not tolerated), which serves as an aid to chemists.

Figure 1.2 ***3D QSAR;*** 3 Dimensional Quantitative Structure Activity Relationships (3D QSAR) are models generated by taking into account the 3 dimensional positions of various physicochemical characteristics of a set of overlapped molecules and the effect they have on drug potency. Refer [71] for details.

1.6 QUANTIFICATION OF HYDROPHOBIC INTERACTIONS

Equations calculating energy from structure, a.k.a. force fields, have been in use for many years in computational chemistry and molecular modeling [112]. Generally, force fields have been restricted to enthalpic terms that are simple to correlate with bond formation or bond breaking and simple Newtonian physical phenomena like bond stretches and bends, electrostatics and dispersion. The hydrophobic effect is, in some measure, an entropic phenomenon and is not easily derivable from these first principles. Nevertheless, a few examples of quantifying lipophilicity and its effect on biomolecular energetics have been reported [12,113-119]. Hermann and Chothia [12,113], among others, proposed that hydrophobicity can be quantified by the calculation of hydrophobic surface area. Oobatake and Ooi present an excellent review of this approach [114]. Cramer and Truhlar introduced a solvation model [115] that included charge distributions on solute molecules, the energetic effects of cavity formation and restructuring of water around such cavities and even subtle variations in charge distribution due to interactions between solute particles and surrounding solvent molecules. Sharp and coworkers introduced a new solvation model illustrating the dependence of the hydrophobic effect on curvature of the site [116]. This was an attempt to explain the difference between the calculated energy for hydration of hydrocarbons (about 25 cal mol^{-1} $Å^{-2}$) and the surface tension at the water-hydrocarbon interface (about 75 cal mol^{-1} $Å^{-2}$). This altered surface area measurement suggested that the “macroscopic” hydration energy is 47 cal mol^{-1} $Å^{-2}$. Indeed, the assumption that the energy of hydrophobic interactions is dependent on the area of the hydrophobic-

water interface is the mainstay of much research in the area. However, alternative approaches have had some success. Cesari *et al.* presented a model describing the hydrophobic interactions within globular proteins based on analysis of X-ray data [117] where fold definitions were clearly shown to be a function of hydrophobicity. Hummer described the development of a hydrophobic force field as an alternative to surface-area models [118]. A highly developed model for quantitating hydrophobic interactions is the HINT (Hydropathic INTeractions) system that is discussed below.

1.7 THE HINT PARADIGM

A notably different approach was taken by Kellogg and Abraham [119,120] in designing the "natural" force field HINT (see Scheme 1.5). This non-covalent interaction force field is derived from partition coefficients based on the Hansch and Leo LogP estimation method. It is very empirical in nature and approximates all components of biomolecular interactions, including hydrogen bonding, Coulombic interactions along with entropy and solvation/desolvation effects in addition to hydrophobic interactions because all of these effects are inherent in the experiments that measure LogP [68]. Interestingly, the Hansch and Leo method encodes many interaction effects within the "correction" factors. For example, intramolecular hydrogen bonding within a small molecule, which would make the molecule less polar (and seemingly more hydrophobic) because the involved polar hydrogen and its partner acceptor are less able to interact with water solvent, is encoded with a factor that gives an internally calibrated indication of the energetics of hydrogen bonding (0.6 – 1.0 LogP units, i.e., 0.8 – 1.4 kcal mol^{-1}).

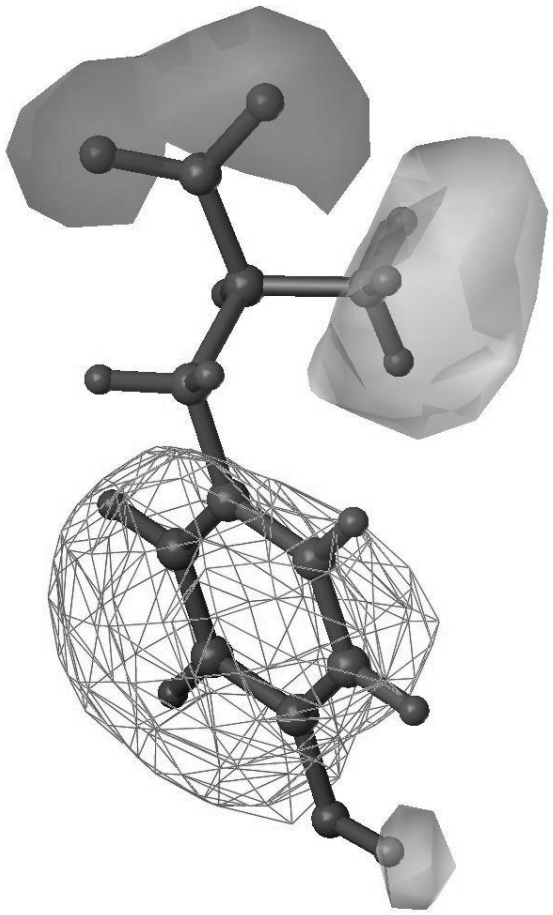
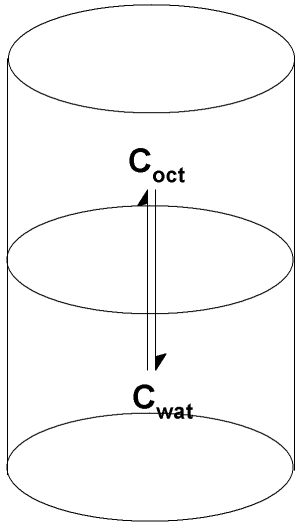


Figure 1.3 ***HINT map for the molecule of tyrosine.*** This map shows a hydrophobic area on the molecule represented as a cage around the benzene ring. The polar areas on the map are further depicted: acidic (light grey lobes) and basic (dark grey lobe).

Scheme 1.5 ***The HINT Paradigm***. A "natural" free energy force field based on LogP. It is available as a toolkit, allowing flexibility in development of applications. Refer to Kellogg and Abraham [118] for details.



A representation of a shake flask. Substances distribute themselves between the water and octanol layers with concentrations C_{wat} and C_{oct}, respectively, in a particular ratio called the partition coefficient.

By definition,

$$P = \frac{C_{oct}}{C_{wat}}$$

LogP can be considered the sum total of individual lipophilic propensities of each atom called hydrophobic atom constant (a_i), i.e.

$$LogP = \sum a_i$$

The values of a_i are readily available from various methods, as described earlier; HINT itself uses an adaptation of the Hansch and Leo C-LOGP approach [23-27]. The HINT method calculates scores (b_{ij}) of each atom against all other atoms according to the equation

$$b_{ij} = a_i S_i a_j S_j T_{ij} R_{ij} + r_{ij}$$

Where, a_i is the hydrophobic atom constant for the i^{th} atom and S_{ij} is the solvent accessible surface area. T_{ij} is a variable which takes on the values of +1 or -1, depending on the acid and base properties of the pair of atoms being considered. For example, if the atoms under consideration are both amino nitrogens, the interaction is unfavorable and T_{ij} is -1. In contrast, if one is a polar (amine) hydrogen and the other is a carboxyl oxygen, their interaction would be favorable and T_{ij} is +1. R_{ij} is the exponential term e^{-r}, where r is the distance between the i and j atoms. r_{ij} is a van der Waal's term.

The total HINT score would be the sum total of each atom-atom score thus calculated, i.e.

$$HINT\ Score = \sum b_{ij} = \sum\sum a_i S_i a_j S_j T_{ij} R_{ij} + r_{ij}$$

If ΔG is the change in Gibb's Free Energy, R is the gas constant and T is the absolute temperature,

we know that $LogP = \sum a_i$ and $LogP = -\Delta G/2.303RT$;

thus, $HINT\ score \propto f(a_i)$.

Hence, HINT scores reflect free energy by taking into account both enthalpic and entropic factors. Because the absolute value of HINT score may not be predictive of a discrete biological association event, these absolute scores are not as important or relevant as differences in HINT score values between analogous systems, much the same as differences in free energy, i.e., $\Delta\Delta G$. The difference in Gibb's Free Energy between two states is a very important parameter as it tells us about the spontaneity or likelihood of the change, whereas the value for any one state itself is often of less consequence.

A key principle behind HINT is that significant understanding of biological phenomena, particularly interactions, can be revealed by representing hydrophobicity as a 3D “field” property rather than as a simple scalar (number) [121]. For example, consider the tyrosine molecule shown in figure 3. Hydropathic properties of the molecule are mapped in three dimensions around the structure of tyrosine, creating a HINT map. The use of these maps creates a visual representation of properties that are often mentioned casually, such as hydrophobic or polar nature of functional groups. Not only does this methodology allow a chemist to form a qualitative understanding of the molecular topology, but also forms the basis for quantitative estimation of physicochemical properties by using the HINT force field. This may further be used in the depiction of molecular interactions, which has a direct repercussion in drug design.

1.7.1 Intermolecular Interaction Analysis

Perhaps the most important application of HINT is in the assessment of intermolecular interactions. HINT calculations derive an interaction score that in numerous studies [122-125] has been shown to correlate with free energy of interaction. Although it is data set dependent, i.e., for specific protein-ligand or polynucleotide-ligand systems, it is estimated that on average, 515 HINT score units correspond to 1 kcal mol^{-1} free energy of binding [119]. A recent report indicated the value of the HINT score in ligand docking studies by a comparison to the scoring functions within FlexX, AutoDock and GOLD [123]. The most important advantage of the HINT methodology is that it inherently estimates enthalpic as well as entropic contributions to binding (Scheme 5). It has been shown [122,126] that errors in prediction for very diverse sets of protein-

ligand complexes are approximately ± 2.6 kcal mol^{-1}, although within a family of ligands binding to the same protein this error can approach ± 1 kcal mol^{-1}. There often is an order of magnitude difference between values of K_i measured by different laboratories on the same protein-ligand complex, which corresponds to a possible 1.0-1.5 kcal mol^{-1} experimental uncertainty. Thus, the error value reported above between experimental results and HINT scores indicates that HINT is a robust method for binding affinity predictions. Further sources of error include uncertainties in positions of atoms in models, incorrectly assigned atom types, or (often) missing solvent molecules in the source crystallographic structure data. The HINT method has been used successfully in quite a number of projects [122,126-129]. In a recent example, Tripathi *et al.* generated a model capable of predicting antiproliferative activity of pyrrole derivatives against cancer cell lines. In this study, experimentally determined IC_{50}s of a number of compounds were correlated with HINT scores from docking these ligands to αβ-tubulin to generate molecular models that could be scored and yielded a significant correlation. This correlation could distinguish active molecules from inactive ones by the HINT score value and, thus, provides a basis for design of novel molecules with anticancer activity. In another interesting application the sequence specificity of anthracycline groove-binding intercalators was evaluated and predicted by HINT score [130]. This work illustrated that the HINT score could be parsed into relevant free energy subsets that can be ranked and compared for particular intercalator functional groups and/or nucleotide bases in DNA double helix strands.

1.7.2 Computational Titration

An extension of the HINT force field known as Computational Titration [124] is used to evaluate the ionization states of functional groups on ligands or residues at the binding site. It is well known that these variations can have a strong influence on binding affinities. The method models, in parallel, multiple ionization states for both ligand and protein creating a collection of ionization state ensembles. Each distinct protonation state ensemble is optimized for hydrogen bonding, including water positions and analyzed by HINT score. The best scoring complex indicates the optimum state for binding and suggests the corresponding pH for that optimum binding. However, the pH at which crystals are grown and analyzed can be different from this optimum pH. The resulting model can help reconcile the differences between *in silico* models and data. However, at room temperature, where binding data is measured, there are likely to be many protonation models of similar, accessible energy. Computational Titration analysis helps develop an understanding of the relationship between these states. There is now a computational titration server for public use at http://hinttools.isbdd.vcu.edu/CT [131].

1.7.3 Analysis of Bridging Waters

Another factor relating to the stability of biomolecular complexes is the contribution of water molecules within the binding site and bridging between the ligand and biomolecule [125,132]. The presence of these bridging water molecules can be a very important factor in binding of molecules, but water molecules can play a variety of

roles as they facilitate biomolecular interactions and stabilize structure. Often, due to a variety of experimental reasons, positions of water molecules in crystal structures are not well defined, even after x-ray crystallographic analysis. This mischaracterization and non-detection of water positions can be correlated with x-ray crystallographic resolution, with better resolution both locating a larger number of water molecules and placing their positions more accurately. Thus, to thoroughly evaluate structure, it is often necessary to verify water molecules systematically with tools such as the GRID program of Goodford [133]. Concomitantly, it is desirable to know which of these waters are subject to displacement by ligands and which are conserved. Using HINT score combined with a metric based on geometry, Amadasi *et al*. developed a robust method to calculate the relevance of binding site waters; those with particularly high relevance score would be expected to yield extra entropy if a ligand was designed to displace it, i.e., similar to the cyclic HIV-1 protease inhibitors [134]. In another study, the contribution of bridging water molecules to overall free energy of binding has been derived and quantitated [132].

1.7.4 3D QSAR with HINT

A very early application of the HINT force field was the introduction of field hydrophobicity parameters into 3D QSAR technology, to complement the original steric and electronic fields in CoMFA [103]. The steroid data set originally reported by Cramer *et al.* [102] was reexamined with the addition of a HINT-derived field. While this study provided little advantage in terms of statistical improvement due to a variety of reasons described previously [119], it provided a distinct advantage in chemical interpretability

for chemists aiming to design new molecules based on such a QSAR study. Quite a few reports of studies based on the HINT-CoMFA methodology have been reported since then [104-106,135-140] and some, particularly where the ligands or active sites are particularly non-polar, do show significant statistical improvement when hydrophobic fields are included. Fields in 3D QSAR are another class of descriptor that often needs to be optimized for the data set [141] in that each data set has forces and structures that may be best represented hydrophobically, sterically, electrostatically, or with other types of fields.

1.8 AIMS AND OVERVIEW OF WORK

The overall aim of computational chemistry is manifold: (1) to develop bioactive agents, (2) to help understand and quantify complex biological phenomena and (3) to develop tools that aid in computational exploration of biological interaction events. The major tool used by us in order to achieve all three objectives is the HINT forcefield. Herein, we employ our in-house computational tools (HINT and the HINT Toolkit), which are an amalgamation of experimental and theoretical methods to explore biological functions of molecules and discover biologically active agents in this work. We also outline strategies to enhance already available computational tools here.

As described earlier in the chapter, the HINT paradigm has been successfully applied in exploring biological phenomena, particularly in binding of macromolecules and also in drug design (*vide supra*). In most projects reported thus far in literature, the HINT forcefield has been applied to simple binding phenomena, i.e. interactions between macromolecules or those between a drug and its target. Given this fact, we

asked ourselves whether it is possible to explain complex biological phenomena (such as transporter-facilitated molecular efflux) in terms of hydropathic interactions using the HINT forcefield. In other words, our hypothesis is that HINT can explain protein-mediated molecular transport in terms of successive, but independent binding events?

One major field of interest for us is to explore new avenues and develop tools for exploration of chemical space in an attempt to simulate the motion of proteins during their interaction with ligands and its application in drug design. In accordance with this aim, we commenced a project to explore the compatibility between existing knowledge-based databases of amino acid residue sidechains and the HINT scoring function. The major aims of this project were to set up grounds for the development of a novel algorithm that will assist in simulation of partial active site flexibility. With this in mind, we hypothesize that HINT can address intramolecular interactions as well as intermolecular ones.

Herein, chapter 2 describes different approaches adopted in order to predict efflux of antibiotic substrates by the AcrA-AcrB-TolC efflux pump. A 3D-QSAR study of efflux yielded ostensibly predictive models, which were validated within a dataset obtained from literature. An alternative methodology was designed due to the inherent problems of 3D-QSAR, which have also been described. This alternative method led to interesting quantitative predictions of high/low efflux for substrates.

Chapter 3 describes a virtual screening approach towards identification of agents which inhibit hemagglutinin-neuraminidase (HN) of human parainfluenza virus type III

(HPIV3). The current status of this project is discussed therein, along with problems with sidechain placement during the docking process.

The following chapter expounds early attempts at development of a sidechain optimization algorithm aimed at creating a basis for the development of improved docking simulation tools in the future. The current status of the algorithm, along with future directions, is delineated.

1.9 REFERENCES

1. Meyer, E.E.; Rosenberg, K.J.; Israelachvili, J. Recent progress in understanding hydrophobic interactions. *Proc. Nat. Acad. Sci. USA,* **2006,** *103*, 15739-15746.
2. Meyer, H. Zur theorie der alkoholnarkose: I. Welche eigenschaft der anaesthetika bedingt ihre narkotische wirkung? *Arch. Exp. Pathol. Pharmakol.* **1899,** *42*, 109-118.
3. Meyer, K.H. Contributions to the theory of narcosis. *Trans. Faraday Soc.* **1937,** *33*, 1062-1064.
4. Overton, E. Über die allgemeinen osmotischen eigenschaften der zelle, ihre vermutlichen ursachen und ihre bedeutung für die physiologie. *Vierteljahrsschr. Naturforsch. Ges. Zürich,* **1899,** *44*, 87-136.
5. Butler, J.A.V. The energy and entropy of hydration of organic compounds. *Trans. Faraday Soc.* **1937,** *33*, 229-236.
6. Frank, H.S.; Evans, M.W. Free volume and entropy in condensed systems III. Entropy in binary liquid mixtures; partial molal entropy in dilute solutions; structure and thermodynamics in aqueous electrolytes. *J. Chem. Phys.* **1945,** *13*, 507-532.
7. Teeter, M.M. Water structure of a hydrophobic protein at atomic resolution: pentagon rings of water molecules in crystals of crambin. *Proc. Nat. Acad. Sci. USA,* **1984,** *81*, 6014-6018.
8. Klotz, I.M. Protein hydration and behavior. *Science,* **1958,** *128*(3328), 815-822.
9. Kauzmann, W. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **1959**, *14*, 1-63.
10. Scheraga, H.A.; Némethy, G.; Steinberg, I.Z. The contribution of hydrophobic bonds to the thermal stability of protein conformations. *J. Biol. Chem.* **1962,** *237*, 2506-2508.
11. Hermann, R.B. Theory of hydrophobic bonding. I. Solubility of hydrocarbons in water, within the context of the significant structure theory of liquids. *J. Phys. Chem.* **1971,** *75*, 363-368.
12. Hermann, R.B. Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.* **1972,** *76*, 2754-2759.
13. Hermann, R.B. Theory of hydrophobic bonding. III. Method for the calculation of the hydrophobic interaction based on liquid state perturbation theory and a simple liquid model. *J. Phys. Chem.* **1975,** *79*, 163-169.
14. Reynolds, J.A.; Gilbert, D.B.; Tanford, C. Empirical correlation between hydrophobic free energy and aqueous cavity surface area. *Proc. Nat. Acad. Sci. USA,* **1974,** *71*, 2925-2927.
15. Leo, A.; Hansch, C.; Jow, P.Y.C. Dependence of hydrophobicity of apolar molecules on their molecular volume. *J. Med. Chem.* **1976,** *19*, 611-615.
16. Leo, A.; Hansch, C.; Elkins, D. Partition coefficients and their uses. *Chem. Rev.* **1971,** *71*, 525-616.

17. Buchwald, P.; Bodor, N. Octanol-water partition: searching for predictive models. *Curr. Med. Chem.* **1998,** *5*, 353-380.
18. Mannhold, R.; van de Waterbeemd, H. Substructure and whole molecule based approaches for calculating LogP. *J. Comput.-Aid. Mol. Des.* **2001,** *15*, 337-354.
19. Mannhold, R.; Petrauskas, A. Substructure versus whole-molecule approaches for calculating LogP. *QSAR Comb. Sci.* **2003,** *22*, 466-475.
20. Manhold, R.; Poda, G.I.; Ostermann, C.; Tetko, I.V. Calculation of molecular lipophilicity: state-of-the-art and comparison of LogP methods on more than 96000 compounds. *J. Pharm. Sci.* **2009,** *98*, 861-893.
21. Livingstone, D.J. Theoretical property predictions. *Curr. Top. Med. Chem.* **2003,** *3*, 1171-1192.
22. Nys, G.G.; Rekker, R.F. The concept of hydrophobic fragmental constants (f-values). II. Extension of its applicability to the calculation of lipophilicities of aromatic and hetero-aromatic structures. *Chem. Ther.* **1974,** *9*, 361-374.
23. Leo, A.; Jow, P.Y.; Silipo, C.; Hansch, C. Calculation of hydrophobic constant (LogP) from Pi and f constants. *J. Med. Chem.* **1975,** *18*, 865-868.
24. Leo, A.J. Some advantages of calculating octanol-water partition coefficients. *J. Pharm. Sci.* **1987,** *76*, 166-168.
25. Leo, A.J. Hydrophobic parameter: measurement and calculation. *Methods Enzymol.* **1991,** *202*, 544-591.
26. Leo, A.J. Calculating $LogP_{oct}$ from structures. *Chem. Rev.* **1993,** *93*, 1281-1306.
27. Leo, A.J.; Hoekman, D. Calculating logP (oct) with no missing fragments; the problem of estimating new interaction parameters. *Perspect. Drug Discov. Des.* **2000,** *18*, 19-38.
28. Petrauskas, A.; Kolovanov, E.A. ACD/LogP method description. *Perspect. Drug Discov. Des.* **2000,** *19*, 99-116.
29. Ghose, A.K.; Crippen, G.M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity. *J. Comp. Chem.* **1986,** *7*, 565-677.
30. Ghose, A.K.; Crippen, G.M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. II. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987,** *27*, 21-35.
31. Ghose, A.K.; Pritchett, A.; Crippen, G.M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. III. Modeling hydrophobic interactions. *J. Comp. Chem.* **1988,** *9*, 80-90.
32. Ghose, A.K.; Vishwanadhan, V.N. Wendoloski, J.J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A,* **1998,** *102*, 3762-3772.
33. Wang, R.X.; Fu, Y.; Lai, L.H. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* **1997,** *37*, 615-621.

34. Wang, R.X.; Gao, Y.; Lai, L.H. Calculating partition coefficient by atom additive method. *Perspect. Drug. Discov. Des.* **2000,** *19*, 47-66.
35. Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.* **2007,** *47*, 2140-2148.
36. Rogers, K.S.; Cammarata, A. A molecular orbital description of the partitioning of aromatic compounds between polar and non-polar phases. *Biochim. Biophys. Acta.* **1969,** *193*, 22-29.
37. Rogers, K.S.; Cammarata, A. Superdelocalizability and charge density. a correlation with partition coefficients. *J. Med. Chem.* **1969,** *12*, 692-693.
38. Hopfinger, A.J.; Battershell, R.D. Application of SCAP to drug design. 1. Estimation of octanol-water partition coefficients using solvent-dependent conformational analyses. *J. Med. Chem.* **1976,** *19*, 569-573.
39. Klopman, G.; Iroff, L.D. Calculation of partition coefficients by the charge density method. *J. Comput. Chem.* **1980,** *2*, 157-160.
40. Bravi, G.; Wikel, J.H. Application of MS-WHIM parameters: 3. Prediction of molecular properties. *Quant. Struct.-Act. Relat.* **2000,** *19*(1), 39-49.
41. Toulmin, A.; Wood, J.M.; Kenny, P.W. Toward prediction of alkane/water partition coefficients. *J. Med. Chem.* **2008,** *51*, 3720-3730.
42. Kier, L.B.; Hall, L.H. *Molecular Structure Description*; Academic Press: New York, NY, **1999.**
43. Livingstone, D.J.; Ford, M.G.; Huuskonen, J.J.; Salt, D.W. Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J. Comput.-Aid. Mol. Des.* **2001,** *15*, 741-752.
44. Molnár, L.; Keserű, G.M.; Papp, Á.; Gulyás, Z.; Darvas, F. A neural network based prediction of octanol-water partition coefficients using atomic5 fragmental descriptors. *Bioorg. Med. Chem. Lett.* **2004,** *14*, 851-853.
45. Liao, Q.; Yao, J.; Yuan, S. SVM approach for predicting LogP. *Mol. Divers.* **2006,** *10*, 301-309.
46. Taskinen, J.; Yliruusi, J. Prediction of physicochemical properties based on neural network modelling. *Adv. Drug Del. Rev.* **2003,** *55*, 1163-1183.
47. Abraham, D.J.; Leo, A.J. Extension of the fragment method to calculate amino acid zwitterions and side chain partition coefficients. *Prot. Struct. Funct. Genet.* **1987,** *2*, 130-152.
48. Buchwald, P.; Bodor, N. Octanol-water partition of non-zwitterionic peptides: Predictive power of a molecular size-based model. *Prot. Struct. Funct. Genet.* **1998,** *30*, 86-89.
49. Steinmetz, W.E. A CoMFA analysis of selected physical properties of amino acids in water. *Quant. Struct.-Act. Relat.* **1995,** *14*, 19-23.
50. Sotomatsu-Niwa, T.; Ogino, A. Evaluation of the hydrophobic parameters of the amino acid side chains of peptides and their application in QSAR and conformational studies. *J. Mol. Struct. (Theochem),* **1997,** *392*, 43-54.

51. Akamatsu, M.; Yoshida, Y.; Nakamura, H.; Asao, M.; Iwamura, H.; Fujita, T. Hydrophobicities of di- and tri-peptides having unionizable side-chains and correlation with substituent and structural parameters. *Quant. Struct.-Act. Relat.* **1989,** *8*, 195-203.
52. Akamatsu, M.; Okutani, S.; Nakao, K.; Hong, N.J.; Fujita, T. Hydrophobicities of N-acetyl-di- and tripeptide amides having unionizable side chains and correlation with substituent and structural parameters. *Quant. Struct.-Act. Relat.* **1990,** *9*, 189-194.
53. Akamatsu, M.; Fujita, T. Quantitative analyses of hydrophobicity of di- to pentapeptides having unionizable side chains with substituent and structural parameters. *J. Pharm. Sci.* **1992,** *81*, 164-174.
54. Akamatsu, M.; Katayama, T.; Kishimoto, D.; Kurokawa, Y.; Shibata, H.; Ueno, T.; Fujita, T. Quantitative analyses of the structure-hydrophobicity relationship for N-acetyl di- and tripeptide amides. *J. Pharm. Sci.* **1994,** *83*, 1026-1033.
55. Tao, P.; Wang, R.; Lai, L. Calculating partition coefficients of peptides by the addition method. *J. Mol. Model.* **1999,** *5*, 189-195.
56. Thompson, S.; Hattotuwagama, C.K.; Holliday, J.D.; Flower, D.R. On the hydrophobicity of peptides: comparing empirical predictions of peptide logP values. *Bioinformation,* **2006,** *1*, 237-241.
57. Berthod, A.; Carda-Broch, S. Determination of liquid-liquid partition coefficients by separation method. *J. Chromatogr. A,* **2004,** *1037*, 3-14.
58. Nozaki, Y.; Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.* **1971,** *246*, 2211-2217.
59. Rose, G.D. Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature,* **1978,** *272*, 586-890.
60. Bull, H.B.; Breese, K. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys.* **1974,** *161*, 665-670.
61. Wolfenden, R.; Andersson, L.; Cullis, P.M.; Southgate, C.C.F. Water, protein folding and the genetic code. *Science,* **1979,** *206*, 575-577.
62. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982,** *157*, 105-132.
63. Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **1976,** *105*, 1-14.
64. Wimley, W.C.; White, S.H. Experimentally determined hydrophobicity scale or proteins at membrane interfaces. *Nat. Struct. Biol.* **1996,** *3*, 842-848.
65. Eisenberg, D.; McLachlan, A.D. Solvation energy in protein folding and binding. *Nature,* **1986,** *319*, 199-203.
66. Li, S.-C.; Deber, C. A measure of helical propensity for amino acids in membrane environments. *Nat. Struct. Biol.* **1994,** *1*, 368-373.
67. Dyson, H.J.; Wright, P.E.; Scheraga, H.A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc. Nat. Acad. Sci. USA,* **2006,** *103*, 13057-13061.

68. Hansch, C.; Leo, A.J. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; J. Wiley and Sons: New York, NY, **1979.**
69. Felitsky, D.J.; Lietzow, M.A.; Dyson, H.J.; Wright, P.E. Modeling transient collapsed states of an unfolded protein to provide insights into early folding events. *Proc. Nat. Acad. Sci. USA,* **2008,** *105*, 6278-6283.
70. MacCallum, J.L.; Moghaddam, M.S. Chan, H.S.; Tieleman, D.P. Hydrophobic association of α-helices, steric dewetting and enthalpic barriers to protein folding. *Proc. Nat. Acad. Sci. USA,* **2007,** *104*, 6206-6210.
71. Mardia, K.V.; Nyirongo, V.B. Simulating virtual protein Cα traces with applications. *J. Comput. Biol.* **2008,** *15*, 1209-1220.
72. Lau, K.F.; Dill, K.A. Lattice statistical mechanics model of the conformation and sequence space of proteins. *Macromol.* **1989,** *22*, 3986-3997.
73. Gupta, A.; Manŭch, J.; Stacho, L. Structure-approximating inverse protein folding problem in the 2D HP model. *J. Comput. Biol.* **2005,** *12*, 1328-1345.
74. Shmygelska, A.; Hoos, H.H. An ant colony optimization algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinformatics,* **2005,** *6*, 30.
75. Hansch, C.; Vittoria, A.; Silipo, C.; Jow, P.Y. Partition coefficients and the structure-activity relation of the anesthetic gases. *J. Med. Chem.* **1975,** *18*, 546-548.
76. Hansch, C.; Steward, A.R.; Anderson, S.M.; Bentley, D. Parabolic dependence of drug action upon lipophilic character as revealed by a study of hypnotics. *J. Med. Chem.* **1968,** *11*, 1-11.
77. Lobl, T.J.; Tindall, D.J.; Cunningham, G.R.; Kemp, P.L.; Campbell, J.A. The correlation of steroid partition coefficients with binding affinities to the rat cytoplasmic androgen receptor, rat androgen-binding protein and human testosterone estradiol-binding globulin. *Ann. N. Y. Acad. Sci.* **1982,** *383*, 477-478.
78. Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficient. *Nature,* **1962,** *194*, 178-180.
79. McFarland, J.W. Parabolic relation between drug potency and hydrophobicity. *J. Med. Chem.* **1970,** *13*, 1192-1196.
80. Kier, L.B. A theory of inhaled anesthetic action by disruption of ligand diffusion chreodes. *Am. Assoc. Nurse Anesthet. J.* **2003,** *71*, 422-428.
81. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997,** *46*, 3-26.
82. Lipinski, C.A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Meth.* **2000,** *44*, 235-249.
83. Clardy, J.; Walsh, C. Lessons from natural molecules. *Nature,* **2004,** *432*, 829-837.
84. Keller, T.H.; Pichota, A.; Yin, Z. A practical view of 'druggability'. *Curr. Op. Chem. Biol.* **2006,** *10*, 357-361.
85. Lipinski, C.A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **2004,** *1*, 337-341.

86. Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003,** *23*, 302-321.
87. Walters, W.P.; Murcko, A.A.; Murcko, M.A. Recognizing molecules with drug-like properties. *Curr. Op. Chem. Biol.* **1999,** *3*, 384-387.
88. Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature,* **2004,** *432*, 855-861.
89. Oprea, T.I.; Davis, A.M.; Teague, S.J.; Leeson, P.D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001,** *41*, 1308-1315.
90. Moriguchi, I. Quantitative-structure activity studies. I. parameters relating to hydrophobicity. *Chem. Pharm. Bull.* **1975,** *23*, 247-257.
91. Oprea, T.I. Current trends in lead discovery: are we looking for the appropriate properties? *Mol. Divers.* **2000,** *5*, 199-208.
92. Proudfoot, J.R. The evolution of synthetic oral drug properties. *Bioorg. Med. Chem. Lett.* **2005,** *15*, 1087-1090.
93. Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity – a review. *QSAR Comb. Sci.* **2003,** *22*, 1006-1026.
94. Kubinyi, H. QSAR and 3D QSAR in drug design. Part 1. Methodology. *Drug Discov. Today,* **1997,** *2*, 457-467.
95. Dudek, A.Z.; Arodz, Ť.; Gálvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb. Chem. High T. Scr.* **2006,** *9*, 213-228.
96. Hansch, C.; Fujita, T. ρ-σ-π analysis: a method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964,** *86*, 1616-1626.
97. Muir, R.M.; Fujita, T.; Hansch, T. Structure-activity relationship in the auxin activity of mono-substituted phenylacetic acids. *Plant Physiol.* **1967,** *42*, 1519-1526.
98. Fujita, T.; Hansch, C. Analysis of the structure-activity relationship of the sulfonamide drugs using substituent constants. *J. Med. Chem.* **1967,** *10*, 991-1000.
99. Kubinyi, H. QSAR and 3D QSAR in drug design. Part 2. Applications and problems. *Drug Discov. Today,* **1997,** *2*, 538-546.
100. Hansch, C.; Kurup, A.; Garg, R.; Gao, H. Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chem. Rev.* **2001,** *101*, 619-672.
101. Li, X.-Z.; Nikaido, H. Efflux-mediated drug resistance in bacteria. *Drugs,* **2004,** *64*, 159-204.
102. Cramer III, 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.
103. Kellogg, G.E.; Semus, S.F.; Abraham, D.J. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput.-Aid. Mol. Des.* **1991,** *5*, 545-552.

104. Debnath, A.K. Application of 3D-QSAR techniques in anti-HIV1 drug design – an overview. *Curr. Pharm. Des.* **2005,** *11*, 3091-3110.
105. Oprea, T.I.; Waller, C.L.; Marwill, G.R. 3D-QSAR of human immunodeficiency virus (I) protease inhibitors. III. Interpretation of CoMFA results. *Drug Des. Discov.* **1994,** *12*, 29-51.
106. Nayyar, A.; Malde, A.; Jain, R.; Coutinho, E. 3D-QSAR study of ring-substituted quinoline class of anti-tubercular agents. *Bioorg. Med. Chem.* **2006,** *14*, 847-856.
107. Testa, B.; Carrupt, P.-A.; Gaillard, P.; Billois, A.; Weber, P. Lipophilicity in molecular modeling. *Pharm. Res.* **1996,** *13*, 335-343.
108. Gaillard, P.; Carrupt, P.-A.; Testa, B.; Boudon, A. Molecular lipophilicity potential, a tool in 3D-QSAR: methods and applications. *J. Comput.-Aid. Mol. Des.* **1994,** *8*, 83-96.
109. Barreca, M.L.; Carotti, A.; Carrieri, A.; Chimirri, A.; Monforte, A.M.; Calace, M.P.; Rao, A. Comparative molecular field analysis (CoMFA) and docking studies of non-nucleoside HIV-1 RT inhibitors (NNIs). *Bioorg. Med. Chem.* **1999,** *7*, 2283-2292.
110. Carrieri, A.; Carotti, A.; Barreca, M.L.; Altomare, C. Binding models of reversible inhibitors to type-B monoamine oxidase. *J. Comput.-Aid. Mol. Des.* **2002,** *16*, 769.
111. Efremov, R.G.; Chugunov, A.O.; Pyrkov, T.V.; Priestle, J.P.; Arseniev, A.S.; Jocoby, E. Molecular lipophilicity in protein modeling and drug design. *Curr. Med. Chem.* **2007,** *14*, 393-415.
112. Ponder, J.W.; Case, D.A. Force fields for protein simulation. *Adv. Prot. Chem.* **2003**, *66*, 27-85.
113. Chothia, C. Hydrophobic bonding and accessible surface area in proteins. *Nature,* **1974,** *248*, 338-339.
114. Oobatake, M.; Ooi, T. Hydration and heat stability effects on protein unfolding. *Prog. Biophys. Mol. Biol.* **1993,** *59*, 237-284.
115. Cramer, C.J.; Truhlar, D.G. An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. *Science,* **1992,** *256*, 213-217.
116. Sharp, K.A.; Nicholls, A.; Fine, R.F.; Honig, B. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science,* **1991,** *252*, 106-109.
117. Cesari, G.; Sippl, M.J. Structure derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **1992,** *224*, 725-732.
118. Hummer, G. Hydrophobic force fields as a molecular alternative to surface-area models. *J. Am. Chem. Soc.* **1999,** *121*, 6299-6305.
119. 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.
120. Kellogg, G.E.; Burnett, J.C.; Abraham, D.J. Very empirical treatment of solvation and entropy: a force field derived from $LogP_{o/w}$. *J. Comput.-Aid. Mol. Des.* **2001,** *15*, 381-393.
121. Abraham, D.J.; Kellogg, G.E. The effect of physical organic properties on hydrophobic fields. *J. Comput.-Aid. Mol. Des.* **1994,** *8*, 41-49.

122. Cozzini, P.; Fonabaio, 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.
123. Spyrakis, F.; Amadasi, A.; Fornabaio, M.; Abraham, D.J.; Mozzarelli, A.; Kellogg, G.E. The consequences of scoring docked ligand conformations using free energy correlations. *Eur. J. Med. Chem.* **2007,** *42*, 921-933.
124. 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.
125. 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.
126. Tripathi, A.; Fornabaio, M.; Kellogg, G.E.; Gupton, J.T.; Gewirtz, D.A.; Yeudall, W.A.; Vega, N.E.; Mooberry, S. Docking and hydropathic scoring of polysubstituted pyrrole compounds with antitubulin activity. *Bioorg. Med. Chem.* **2008,** *16*, 2235-2242.
127. Simoni, D.; Invidiata, F.P.; Eleopra, M.; Marchetti, P.; Rondanin, R.; Baruchello, R.; Grisolia, G.; Tripathi, A.; Kellogg, G.E.; Durrant, D.; Lee, R.M. Design, synthesis and biological evaluation of novel stilbene-based antitumor agents. *Bioorg. Med. Chem.* **2009,** *17*, 512-522.
128. Porotto, M.; Fornabaio, M.; Kellogg, G.E.; Moscona, A. A second receptor binding site on human parainfluenza virus type 3 hemagglutinin-neuraminidase contributes to activation of fusion mechanism. *J. Virol.* **2007,** *81*, 3216-3228.
129. Burnett, J.C.; Botti, P.; Abraham, D.J.; Kellogg, G.E. Computationally accessible method for estimating free energy changes resulting from site-specific mutations of biomolecules: systematic model building and structural/hydropathic analysis of deoxy and oxy hemoglobins. *Prot. Struct. Funct. Genet.* **2001,** *42*, 355-377.
130. Cashman, D.J.; Kellogg, G.E. A computational model for anthracycline binding to DNA: Tuning groove-binding intercalators for specific sequences. *J. Med. Chem.* **2004,** *47*, 1360-1374.
131. 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.-Aid. Mol. Des.* **2009,** *23*, 621-632.
132. 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.
133. Goodford, P. Multivariate characterization of molecules for QSAR analysis. *J. Chemometrics,* **1996,** *10*, 107-117.

134. 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.
135. 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.
136. Tsakovska, I.M.; Pajeva, I.K. Molecular modeling of triazine type MDR modulators using CoMFA and CoMSIA approaches. *SAR & QSAR Environ. Res.* **2002,** *13*, 473-484.
137. Welch, W.; Ahmad, S.; Airey, J.A.; Geron, K.; Humerickhouse, R.A.; Besch, Jr. H.R.; Deslongchamps, P.; Sutko, J.L. Structural determinants of high-affinity binding of ryanoids to the vertebrate skeletal muscle ryanodine receptor: a comparative molecular field analysis. *Biochemistry,* **1994,** *33*, 6074-6085.
138. Grell, W.; Hurnaus, R.; Griss, G.; Sauter, R.; Rupprecht, E.; Mark, M.; Luger, P.; Nar, H.; Wittneben, H.; Muller, P. Repaglinide and related hypoglycemic benzoic acid derivatives. *J. Med. Chem.* **1998,** *41*, 5219-5246.
139. Bursi, R.; Grootenhuis, P.D.J. Comparative molecular field analysis and energy interaction studies of thrombin-inhibitor complexes. *J. Comput.-Aid. Mol. Des.* **1999,** *13*, 221-232.
140. Iskander, M.N.; Coupan, I.M.; Winkler, D.A. Investigation of 5-HT4 agonist activities using molecular field analysis. *J. Chem. Soc. Perkin Trans.* **1999,** (2), 153-158.
141. Kellogg, G.E. Finding optimum field models for 3D QSAR. *Med. Chem. Res.* **1997,** *7*, 417-427.

CHAPTER 2

PREDICTING EFFLUX OF ANTIBIOTICS BY ACRA-ACRB-TOLC PUMPS: A 'SYSTEMS HYDROPATHY' APPROACH.

2.1 Introduction

Antimicrobial drugs have been crucial tools of healthcare for decades due to their effectiveness in control of bacterial infections. However, soon after their discovery it was realized that some pathogens rapidly developed resistance to antibiotics [1,2]. Initially, this problem was overcome by discovery of new classes of antibiotics such as aminoglycosides, macrolides and glycopeptides, but it soon became clear that bacteria had an impressive array of defensive mechanisms that conferred on them, resistance to many modes of attack [1]. Organisms that cause pneumonia and cutaneous infections, *Streptococcus pneumonia*, *Streptococcus pyogenes* and *staphylococci*, are now resistant to almost all of the older, first generation, β-lactam antibiotics [2] like penicillin, which were discovered through screening of mold samples and act in general by mechanically weakening cell walls and making them susceptible to osmotic lysis [3-5]. Members of the *Enterobacteriaceae* and *Pseudomonas* families, which cause diarrhea, urinary infection and sepsis, are also resistant [2]. The development of bacterial resistance to antibiotics has mostly been attributed to their excessive use in the clinic as well as at home [1,2]. However, semisynthetic modifications of β-lactam antibiotics have given us second- and third-generation antibiotics that have prolonged the therapeutic usefulness of the drug class.

One primary mechanism of antibiotic resistance is extrusion of the foreign chemical, which is termed efflux. In 1980, it was reported that tetracycline could be actively effluxed from the bacterial cell [6]. Since then, many efflux-related mechanisms have been discovered. Efflux pumps are transporters involved in extrusion of toxic substances from cells, thereby limiting the detrimental effects of these substances [7]. They may be substrate-specific and responsible for transporting biological compounds such as bile salts, or may be promiscuous and transport structurally-diverse compounds such as various classes of antibiotic drugs [8]. Overexpression of these structurally complex and versatile proteins may thus lead to antibiotic resistance. While efflux pump proteins are present in both Gram-positive and Gram-negative bacteria and also in eukaryotes, antibiotic resistance due to efflux is more of a problem in Gram-negative bacteria than in Gram-positive bacteria [9]. This is due to the presence of an outer membrane in Gram-negative bacteria that demonstrates comparatively lower permeability and complements the efflux activity of these pumps.

Several such pump systems have been described: *Campylobacter jejuni* (CmeABC) [10-11], *Escherichia coli* (AcrAB-TolC, AcrEF-TolC, EmrB, EmrD) [12], *Pseudomonas aeruginosa* (MexAB-OprM, MexCD-OprJ, MexEF-OprN and MexXY-OprM) [12], *Streptococcus pneumonia* (PmrA) [13], *Salmonella typhimurium* (AcrAB) [14] and *Staphylococcus aureus* (NorA) [15]. These pumps basically fall into five major families, including the MF (major facilitator), MATE (multidrug and toxic efflux), SMR (small multi-drug resistance), ABC (ATP-binding cassette) and RND (resistance-nodulation-division)

families [16]. It has been shown that co-expression of multiple types of efflux pumps can cause an additive or even multiplicative effect on drug resistance [17].

AcrAB was first described as an efflux system in 1995 [18]. AcrB is responsible for efflux of bile salts, thus protecting enteric *E. coli* from the detrimental effects of these powerful detergents [19]. As is typical with other members of the RND-type efflux protein systems, AcrAB is also a proton antiporter. AcrA and AcrB homologues in *Haemophilius influenzae* HI0894 and HI0895 respectively, are also responsible for drug efflux [20]. The importance of AcrAB in multidrug resistance has been described in several publications, [21-27] where knock-out, knock-in and mutation studies were used to describe the extent to which the AcrA-AcrB-TolC transporter is responsible for expulsion of structurally diverse antibiotics from bacterial cells.

One important observation from all these studies was that efflux pumps seem to preferentially extrude hydrophobic ligands [24]. However, despite extensive studies on the efflux of antibiotics by the AcrA-AcrB-TolC efflux pump, reliable and generalizable predictions for efflux by this pump continue to be elusive. Such predictions would obviously be of tremendous interest to those engaged in design/discovery of antibiotics, as this would allow them to more efficiently channel effort and resources. Interestingly, despite the tremendous importance of efflux as an ancillary consideration in drug discovery, very few computational studies designed to predict the effect on proteins other than P-glycoprotein [28-30] have been reported. However, two-dimensional quantitative structure-activity relationship (QSAR) studies performed previously [31,32] yielded what appeared on the surface to be remarkable regression equations for efflux

of β-lactam substrates by the AcrA-AcrB-TolC pump from *Salmonella thypimurium* (16 compounds, target: minimal inhibitory concentration for three strains, reported r^2 from cross validation (q^2) > 0.9). No surprising conclusions were made in this study, i.e., molecules showing more hydrophilic character, including hydrogen bonding capability, were likely to be poor efflux substrates and that efflux correlated with properties like LogP (for partitioning the drug between 1-octanol and water), the y-axis component of electrostatic dipole moment, the surface area of hydrophobic carbons, the number fractions of carbons and heteroatoms and the number of charged groups and the number of nitrogen and sulfur atoms, all of which would supposedly influence interactions between pump and substrate [31]. However, the number of descriptors in the equations (up to 9) suggests serious overfitting of such a small data set and the inclusion of multiple methods of LogP prediction within the same QSAR equation is also a concern: while LogPs calculated by different methods do not necessarily encode exactly the same information about a molecule, they *must* be largely correlated and some of the other descriptors, e.g., surface area of hydrophobic carbons, also likely correlate with LogP. Most importantly, it does not appear that this type of model has been embraced by potential users of efflux prediction in drug design, possibly because of its poor chemical and physical interpretability.

In the present contribution, we describe a computational modeling method that allows us to successfully identify the extent of efflux of individual ligands by taking into account interplay between properties of the ligands as well as their molecular-level interactions with the AcrA-AcrB-TolC efflux pump. Although an initial 3D-QSAR study

produced seemingly predictive models for the β-lactams, we decided to “invent” a new, largely intuitive method based on what is known about efflux and found a successful prediction of efflux values for a structurally diverse dataset, composed of both β-lactam and non-β-lactam antibiotics. Our approach holds a superficial similarity to the ‘systems biology’ approach, where the effects of a factor on multiple pathways are taken into account by compartmentalization in order to explain observed gross phenomena and we are thus calling our method “systems hydropathy” (*vide infra*). It is also interesting to note that our results suggest certain novel mechanistic details of efflux, hitherto untested and unreported.

We further propose that this method may be extended to several complex transporter systems such as the ABC proteins, which are suspected to cause resistance to anticancer drugs [33].

2.2 EXPERIMENTAL SECTION

2.2.1 Crystal structures of AcrB and TolC.

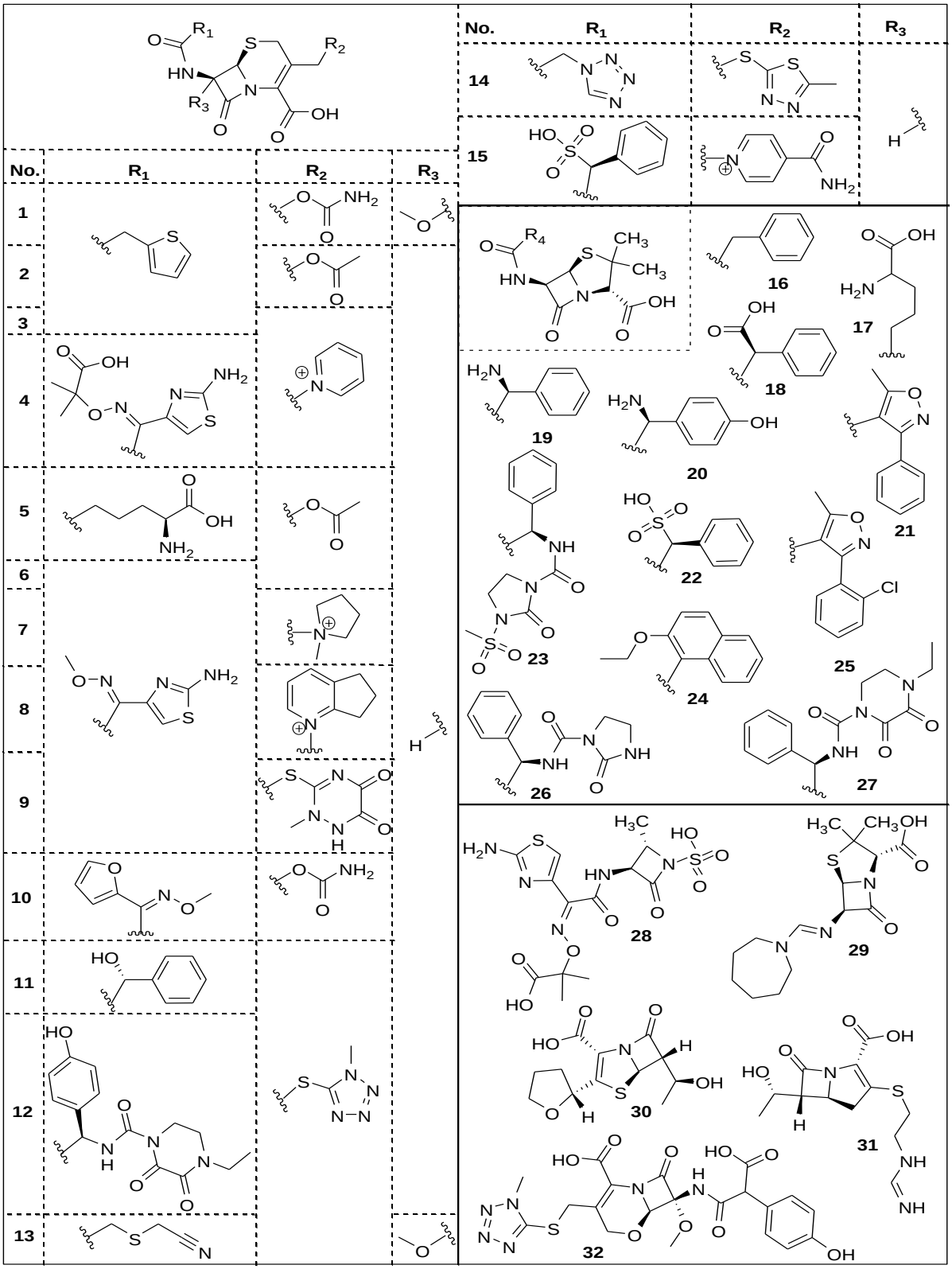
The crystal structures of AcrB and TolC (PDB codes 2drd and 1ek9 respectively) are available in the RSCB database [34,35]. The protein structures were modeled using Sybyl version 8.1 (Tripos International) [36]. Hydrogens were added, followed by 1000 steps of Powell minimization with Gasteiger-Hückel charges while keeping the heavy atoms still. Then the entire structure was minimized to a gradient of 0.005 kcal mol^{-1} $Å^{-1}$. Visual inspection revealed no unrealistic steric clashes between residues. Where

required, the ligand was removed from the protein binding state (from PDB code 2drd) to void the docking region.

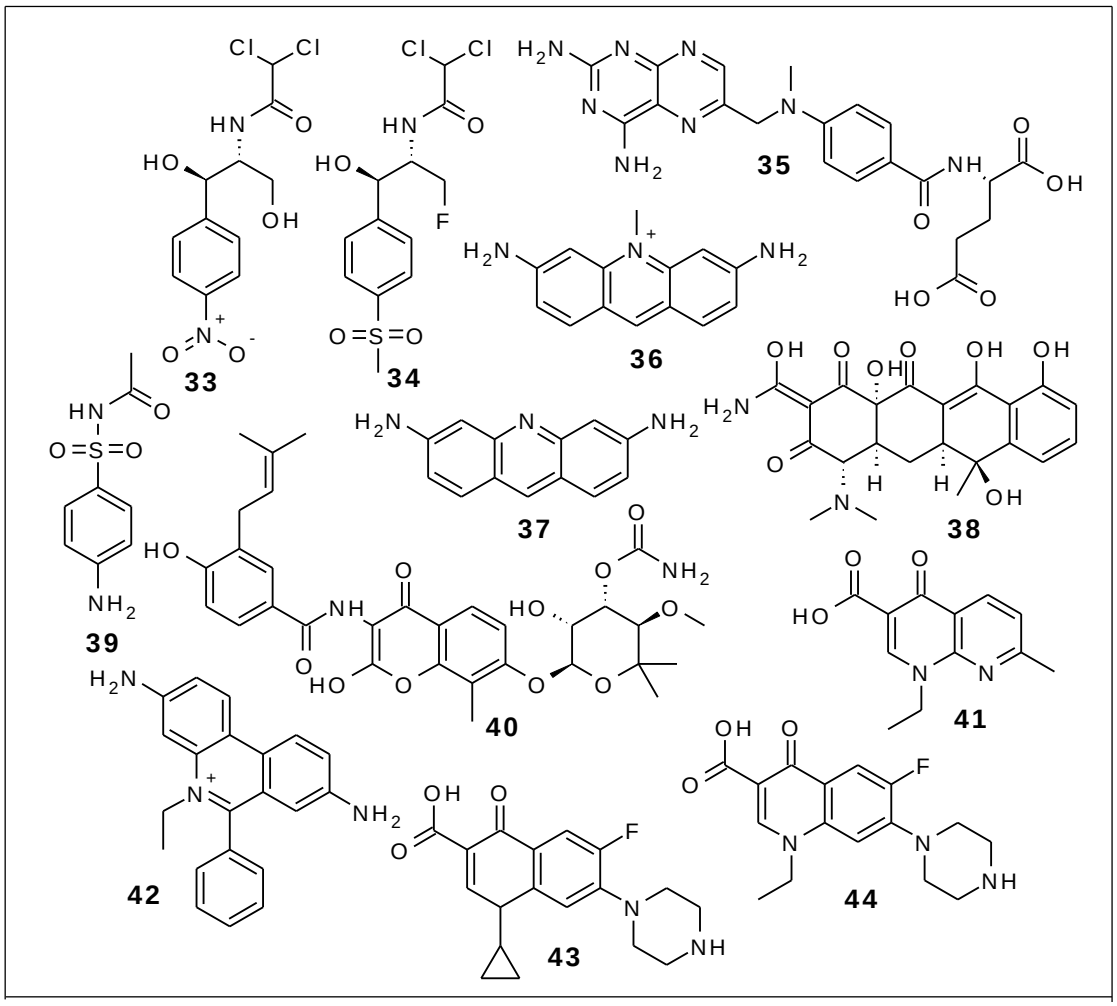
2.2.2 Efflux data and substrate molecules.

A review of literature produced a curated set of 32 β-lactam antibiotics (Scheme 2.1) and 12 non-β-lactam antibiotics (Scheme 2.2), for which efflux data has been reported [23-25,27,37]. This data was in the form of minimum inhibitory concentration (MIC) ratios (i.e., ratio of MIC in the presence of efflux pump to the MIC in the absence of efflux pump through knockout). All reported MIC ratios for these compounds were taken and an average MIC ratio was obtained. Because the experimental data are reported as powers of 2, i.e., 1, 2, 4, 8, 16, etc., logarithms (base 2) were calculated for these average MIC ratios and these logarithmic values were used as the dependent "efflux" parameter in our analyses (Table 2.1). The range of MIC values for wild type (WT) pumps across all sources is also reported here. The narrow range of WT MIC values for each antibiotic clearly indicates that similar if not identical pump strains were used during experimental procedures reported in these referenced works such that these MIC ratios are directly comparable.

As all of the substrate antibiotic molecules have acid and/or base functionalities, the structures were modeled in Sybyl in both their neutral (non-ionic) and charged (usually zwitterionic) forms and minimized to a gradient of 0.005 kcal mol^{-1} A^{-1}. An attempt was made to initially place each compound in its lowest energy conformation by manually selecting from available rotamers.



Scheme 2.1. Chemical structures for the β-lactam antibiotic compounds in the study.



Scheme 2.2. The chemical structures for the non-β-lactam antibiotic compounds used in the study.

2.2.3 Docking and scoring.

The initial methods development for this study was performed with only the β-lactam ligand set (Scheme 2.1) and then applied to the full, extended dataset. All compounds were docked in both their neutral and charged forms using GOLD version 3.0 [38]. When docked to different parts of the TolC cavity (see Figure 2.1 and Results and Discussion for further description), 100 positions for each compound in Scheme 2.1 were obtained at each of eleven different overlapping areas of the protein, giving us a total of 1100 solutions per compound. However, a total of 2000 positions were obtained per compound when docked into the binding state or extrusion state of AcrB. The antibiotics in Scheme 2.2 were docked at four different positions as selected by the model found for the β-lactam compounds (see Results and Discussion), using identical procedures and parameters. All docked positions were scored using the Hydropathic INTeractions (HINT) forcefield [39-41] as reported in previous work [42-43]. The HINT forcefield has been previously shown to not only estimate enthalpic contributions towards binding but also entropic and solvation contributions [39-43]. The pose corresponding to the highest HINT score for each ligand at each position was selected for further analysis. GOLD scores were ignored because the software is known to fail for hydrophobic systems [38]. The best docked positions were combined into protein-ligand structures that were minimized with 500 iterations. The interactions were re-scored at the minimized positions and these HINT scores were used as descriptors. Further explanation regarding application of the various HINT scores as utilized in this manuscript is given below.

Table 2.1. Efflux and molecular parameters for data set molecules (see text for further description).

No.	Antibiotic	Reported MIC ratios[a]					MIC range for WT pump (µg/ml)	Avg MIC ratio	Log_2(avg. MIC ratio)	LogP (AlogPs)[c]	Mol. Width (Å)[d]
		Ref 23[b]	Ref 24[b]	Ref 25[b]	Ref 27[b]	Ref 37[b]					
1	Cefoxitin	4	4				4	4	2	0.22	8.82
2	Cefalothin	1	4				2-8	2.5	1.322	0.63	7.87
3	Cephaloridine	2	2				2-8	2	1	1.67	8.47
4	Ceftazidime	1					0.12	1	0	0.78	8.52
5	Cephalosporin C		1				16	1	0	-2.18	8.16
6	Cefotaxime		4	2			0.12	3	1.585	0.14	7.62
7	Cefepime	1	1				0.0075	1	0	-1.54	8.73
8	Cefpirome	1,2					0.015	1.5	0.585	1.57	8.99
9	Ceftriaxone	1	2				0.015-0.12	1.5	0.585	-0.01	8.66
10	Cefuroxime	16		8			1.56-2	12	3.585	-0.24	8.82
11	Cefamandole	4, 8		4			4	5.33	2.415	-0.05	8.28
12	Cefoperazone	2					0.03	2	1	-0.11	11.03
13	Cefmetazole	1	1	2			0.05	1.33	0.415	-0.38	8.44
14	Cefazolin	1	1				0.39-0.5	1	0	-0.4	8.32
15	Cefsulodin	1	1				16-64	1	0	0.6	9.54
16	Penicillin G	2	32				8-16	17	4.087	1.92	6.99
17	Penicillin N		1				8	1	0	-1.43	7.82
18	Carbenicillin	1, 4	4	1			1.56-8	2.5	1.322	1.13	7.07
19	Ampicillin	2		2	4		0.78-12.5	2.67	1.415	0.88	7.07
20	Amoxicillin	1					4	1	0	0.75	7.79
21	Oxacillin	512		256			1024	384	8.585	2.05	8.77
22	Sulbenicillin		4	1			8	2.5	1.322	0.37	7.33
23	Mezlocillin	32					1	36	5.170	0.21	10.16
24	Nafcillin		128	128			200	128	7	3.21	8.84
25	Cloxacillin	128	256	128			256-512	171	7.415	2.61	8.56

Table 2.1 continued…

26	Azlocillin	4, 8					16	6	2.585	0.2	9.92
27	Piperacillin	16					4	16	4	0.67	10.92
28	Aztreonam			1			0.05	1	0	0.06	7.87
29	Mecillinam	2					0.5	2	1	1.41	6.95
30	Faropenem			4			0.39	4	2	0.24	6.98
31	Imipenem	1					0.12	1	0	-0.19	7.36
32	Latamoxef		1				0.12	1	0	0.22	8.43
33	Chloramphenicol				8	4	4-6.25	6	2.585	0.11	7.27
34	Florfenicol				8		6.25	8	3	0.98	8.45
35	Methotrexate				8		640	8	3	-0.91	9.48
36	Acriflavine				128		400	128	7	2.56	9.39
37	Proflavine				8		100	8	3	2.10	5.99
38	Tetracycline				8		1.25	8	3	-0.56	8.88
39	Sulfacetamide				1		2000	1	0	0.15	6.27
40	Novobiocin	32		64	64		32-100	53.3	5.736	3.07	8.68
41	Nalidixic acid				2		3.125	2	1	0.95	9.62
42	Ethidium bromide				256		800	256	8	4.33	9.01
43	Ciprofloxacin				4		0.01	4	2	-0.57	9.07
44	Norfloxacin				1		0.004	1	0	-0.47	9.46

[a]The MIC ratio is the ratio of minimum inhibitory concentration (MIC) in the presence of efflux pump to the MIC in the absence of the pump. All reported MIC ratios are for the E. coli AcrAB-TolC pump, or the closely related S. typhimurium homolog.
[b]All WT strains used include JC7623, SH5014, TG1, HS414 and ECM1194 while all pump knockout strains include JZM120, SH7616, KAM3/pHSG398 or pHSG299, HS832 and ECM1694.
[c]LogP was predicted by the ALogPs method [41-43].
[d]The molecular width of each efflux substrate was calculated by performing a molecular dynamics simulation of 1 ns duration, followed by aligning the farthest atoms along the Z axis and measuring the projection of all other atoms on the XY plane. Assuming that the molecule spins rapidly, the largest such projection gives us a rough measure of its molecular width.

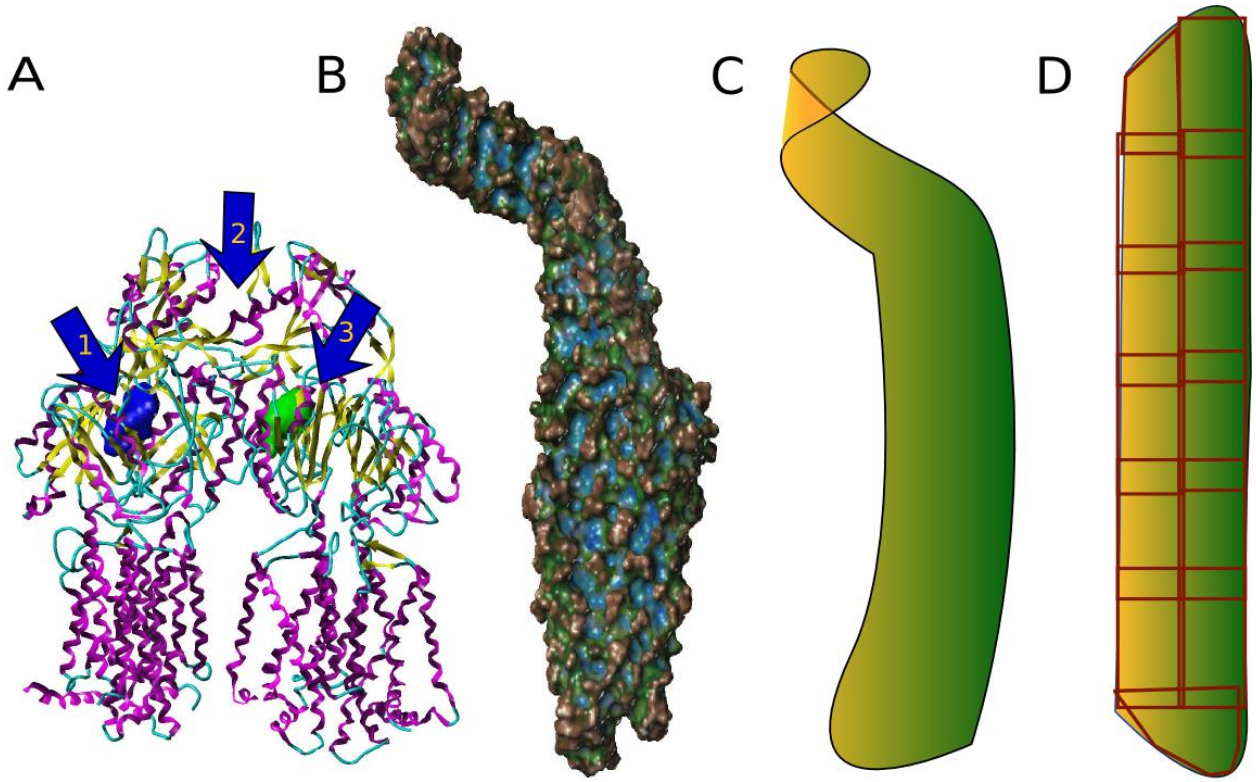


Figure 2.1. **Docking Efflux Substrates Into Different Regions of AcrB and TolC.** All 32 β-lactam and 12 non-β-lactam structures were docked into AcrB and several overlapping regions of TolC. **(A)** The hydrophobic pocket of AcrB is shown for both binding state (blue) and extrusion state (green) conformations. The numbered arrows indicate the specific locations of AcrB in which the substrate molecules were docked: 1) binding state, 2) intermonomer region, 3) the extrusion state. **(B)** Molecular surface of the TolC channel shown colored according to cavity depth. The peaks are shown in copper brown, while troughs are shown in blue. **(C)** Ribbon cartoon of a TolC monomer, depicting its twisted shape. **(D)** The straightened cartoon avatar of one TolC monomer, showing different zones where antibiotic structures were docked. All the zones overlapped sufficiently to ensure a thorough investigation of molecular interactions between the entire TolC lumen surface and each antibiotic.

2.2.4 LogP calculations.

LogP was calculated by using an online server [44], which gave values of multiple LogP prediction methods including ALogPs, ALogP, MLogP, XLogP2 and XLogP3 [44-46] that are based on different principles and yet are known to predict LogP values quite well. Our regressions, however, stipulated that only one set of LogP values could be used in any resulting equation in order to avoid using multiple highly correlated descriptors, which would have led to models with exaggerated statistical parameters. The best correlation between predicted LogP and efflux was shown by ALogPs values. For comparison and to evaluate the significance of differences between LogP calculated for charged and uncharged forms of the molecules, LogP values were also calculated for the β-lactams using HINT [39] ("calculate method", "essential" hydrogens only).

2.2.5 Prediction of molecular width by molecular dynamics simulations.

Considering an efflux pump to be a tubular passage, one descriptor that we expected to correlate with efflux was molecular width, i.e., it should have a negative correlation with efflux, akin to a very large ball not being able to pass through a tubular pipe. For this purpose, we used molecular dynamics simulations to calculate the effective width of the efflux substrates. Figure 2.2 describes the physical concept behind our calculations: we aligned each structure with the z-axis and calculated the projections of its atoms on the x- and y-axes. The largest projection for each structure was taken as its width. For completeness, we ran a 1 ns simulation for each molecule using the Sybyl molecular dynamics module at 300K, recording snapshots every fs and performed width calculations every 100 fs of simulation. The resulting average of

widths, calculated using 10,000 conformers for each ligand, was used as the molecular width descriptor.

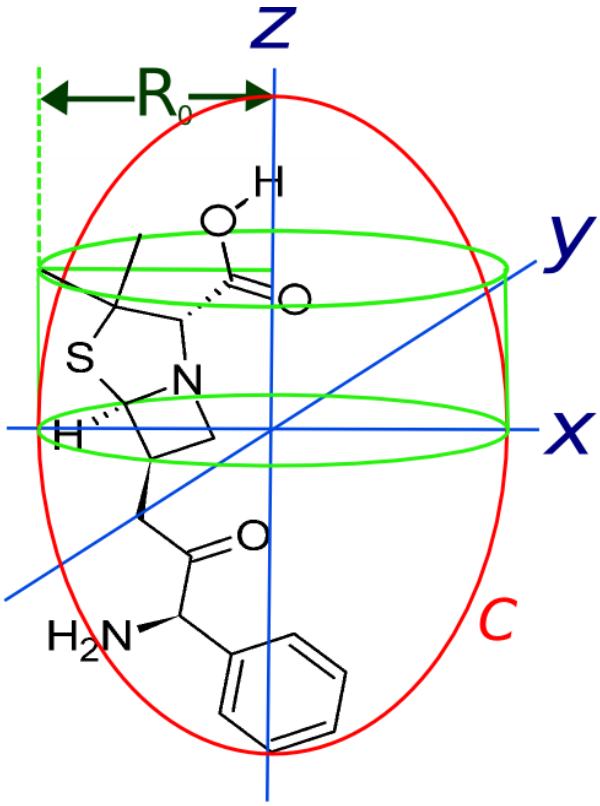


Figure 2.2. ***Calculation of Molecular Width.*** The figure shows ampicillin depicted in a 3D coordinate system, with the longest pair of atoms aligned with the Z axis. Projection of an atom on the XY plane is shown and its length is depicted by R_0. The largest X or Y component of all such projections was taken as the radius of a cylinder *C* that can hold the entire structure, which is an approximate measure of the width of the structure.

2.2.6 3D-QSAR methods.

The 3D-QSAR analyses were performed using the Comparative Molecular Fields Analysis fields (CoMFA) module of Sybyl. The β-lactam molecules were manually aligned based on the common β-lactam substructure and re-minimized to ensure no structures with high internal strain energy were present during analysis. The set of 32

β-lactams was divided into two sets, a training set of 24 compounds and a test set of 8 compounds using a random number generator to avoid any bias in selecting training and test sets. 3D-QSAR studies were performed for the charged and neutral forms separately. All compounds thus aligned were placed in a grid (spacing 1 Å, margin 4 Å). Both standard CoMFA fields (hydrogen bonding, electrostatic and steric parameters) and HINT fields (hydropathic) were calculated at each grid point. Partial Least Squares (PLS) and Sample-distance Partial Least Squares (SAMPLS) [47] analyses were performed to correlate positions/properties of various chemical substituents with efflux values. Various descriptor field combinations were comprehensively explored, creating several models to explore the predictability of each. The selected best models, i.e., those showing highest cross-validated r^2 (q^2) at a minimum (optimum) number of components, identified the descriptor field sets to be used for further analysis.

2.2.7 The systems hydropathy method.

Using solely the 32 β-lactams, a statistical analysis of docking/scoring scores and substrate physicochemical properties as descriptors was performed using PLS and SAMPLS. During this investigation, only one predicted LogP value was used in any given regression equation to avoid using multiple correlated descriptors. An exhaustive search through the various HINT score sets calculated, as described above, for substrate binding to the various pump zones, singly and in combination, was performed to find the best possible descriptor combination. This model (*vide infra*), i.e., the

combination of descriptors that best quantified β-lactam efflux values, was then extended to the non-β-lactam set. The complete set of 44 compounds was separated into training sets of 33 compounds and test sets of 11 compounds. Cross-validation was performed on these training sets and test sets to confirm internal stability of the models.

2.3 RESULTS AND DISCUSSION

The intrinsic function of efflux pumps is to expel extraneously acquired molecules that could harm the cell. These are promiscuous proteins that are largely responsible for antibiotic resistance in Gram-negative bacteria. In contrast to normal receptors and enzymes that bind or else catalyze reactions involving small molecules at a specific site, the function of these proteins is to *transport* these small molecules – usually over fairly large distances. Perhaps due to this difference in function and the multiple steps involved in what is clearly a dynamic rather than static process, their activity has been resistant to computational chemistry/biology attempts at prediction. In fact, for a long time, the only useful known trait of the AcrA-AcrB-TolC efflux pump was that it transports hydrophobic molecules more easily [24]. However, recently available crystallographic data for the AcrB and TolC components of this efflux pump has enabled a more systematic and structural evaluation of the efflux mechanism [34].

We believed that this emerging structural information for pump molecules, combined with modeling tools that effectively characterize hydrophobic interactions and related effects, could illuminate the process of efflux. Our key technology is the HINT model [39-41], which is an empirical modeling tool based on the free energy of solvent transfer

between two phases, 1-octanol and water, representing hydrophobic and polar biological environments, respectively. HINT has been used to evaluate ligand binding [42,43], protein-protein associations [48-50] and other phenomena involving biological molecules [51,52] and has been generally successful in quantitating the free energies of these interactions. In addition, qualitative representations of biological processes involving molecular associations have been developed within the HINT paradigm; e.g., we recently developed molecular models validating previously proposed dual site mechanisms for inhibition of paramyxovirus hemagluttanin-neuramindase [53,54].

In obtaining and curating a high quality and relevant data set for our analysis, we restricted this study to 32 β-lactam compounds (see scheme 2.1) and 12 non-β-lactams from several families (see scheme 2.2), whose efflux data was available in published reports [23-25,27,37]. Although similar data for them is available, the aminoglycoside and macrolide classes of antibiotics are not included in the curated set; the former because they are hydrophilic and therefore not effluxed as readily by the AcrAB pump - the AcrD pump is apparently more responsible for their efflux than the AcrB pump [55,56] and the latter because of their large size relative to that of the AcrB entrance. However, we cannot discount the possibility that these antibiotics might enter TolC via a different route [57], thus still involving parts of the AcrA-AcrB-TolC pump (and concomitantly being affected by the pump knockout mutants). We recognize that a much larger but proprietary set of data is very likely available within pharmaceutical companies, but wanted the entire data set to be available from the primary literature for this initial work. The experimental measure of efflux used in this study is Log_2(MIC

ratio), where the MIC (minimum inhibitory concentration) ratio is the MIC for a cell with an intact AcrA-AcrB-TolC pump normalized by the MIC for that cell with the pump knocked out. Using this ratio in lieu of MIC itself has benefits as it more clearly represents the change in effectiveness of an antibiotic. In contrast, MIC is a poorer target metric as it is dependent on a number of factors such as concentration of cells per unit volume of culture and is thus more laboratory and procedure-dependent.

2.3.1 3D-QSAR.

It seemed possible that binding in the AcrB pocket would affect the efflux of substrates more than binding elsewhere in the pump because it has been shown that this protein must undergo a conformational change in order to pass substrates into TolC [34]. One way to test this is to perform 3D-QSAR analyses where the interactions between substrates and a hypothetical but undefined receptor are simulated by the molecular fields of the substrates. This approach was applied to the β-lactam data set after conformationally aligning these molecules to simulate their putative binding modes within a binding site presumed to be AcrB. If such an analysis provided predictive results with respect to efflux, then we could at least partially address our goal of predicting high or low efflux. We performed this 3D-QSAR study of the β-lactam data set with molecules in both their charged and neutral forms using the Comparative Molecular Field Analysis (CoMFA) method of Cramer [58]. The results seem promising for neutral compounds, with a cross-validated r^2 (q^2) = 0.53 (4 components) on the training set of 24 compounds and a (predictive) r^2 = 0.80 for the test set of 8 compounds, as illustrated for a typical run in Figure 2.3. This model used the CoMFA

Standard Steric field, H-bond Steric field and H-bond Electrostatics field (0.35), with relative contributions of 27%, 38% and 35%, respectively. Although their internal statistical metrics were initially acceptable, CoMFA models for β-lactam substrates in their charged state could not be validated.

However, 3D-QSAR experiments are dependent upon the alignment of a *common substructure,* in this case the β-lactam. Thus, this model formalism would not be utilizable for the extended dataset including the non-β-lactams. Furthermore, these experiments are based on the assumption that substrates bind to a single site in the pump, whether in AcrB as we proposed or elsewhere, when in fact AcrB changes conformation between its binding and extrusion states [34]. Thus, there must be a dynamic change in its interactions with ligands as they are processed. In fact, the binding pocket of AcrB is lined with a number of hydrophobic phenylalanine residues [34], indicating a preference for hydrophobic or, at a minimum, less polar substrates – suggesting that ligands would more favorably bind in their uncharged forms.

At the same time, we cannot ignore the possibility that the phenylalanine rings could also be acting as receptors for π-cation interactions. The 3D-QSAR results above support the assertion of "neutral state" binding to AcrB and may be interpreted as evidence of at least transient binding at this site being a rate-limiting initial step. In contrast, the TolC lumen is exposed to the extracellular environment due to its position on the bacterial outer membrane and this lumen is very likely solvated – suggesting that ligands bound here would favor their charged form. The translocation of substrates from AcrB to TolC must expose them to water, providing a mechanism for transforming

them from their uncharged to charged forms. Substrates by necessity interact with various parts of AcrB and TolC during their efflux extrusion and the process of efflux is certainly affected by a multitude of different interactions and thus cannot be completely addressed by simple methods such as 2D or 3D-QSAR that are based on molecules in a single state, bound within a single well-behaved binding site/mode with the concomitant assumption of a pharmacophore recognition-driven process.

While the 3D-QSAR model is unfortunately not extensible to other substrate families because of the requirement that the molecules in the model have a common alignment, the results for the β-lactam data set are an indicator of key principles behind the pump mechanism.

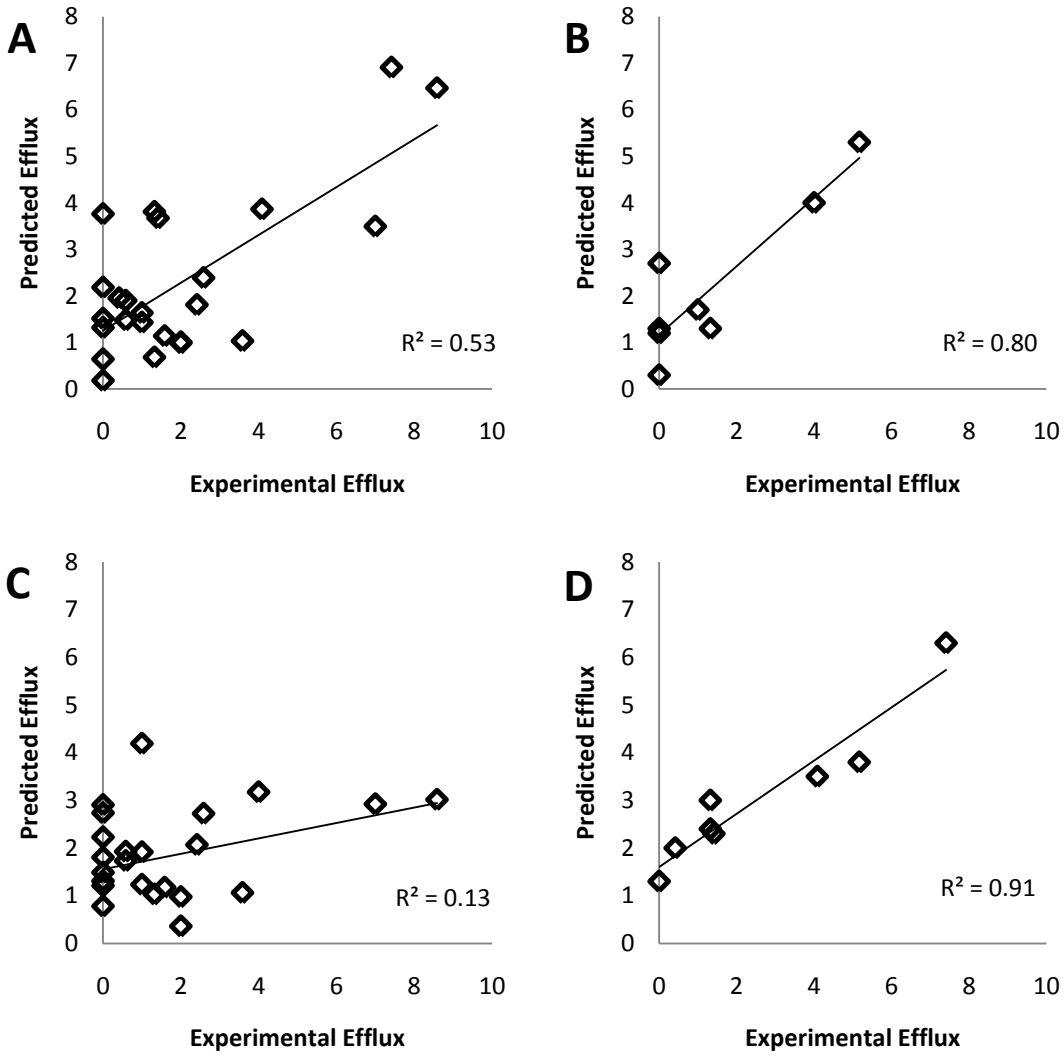


Figure 2.3 **Training set and test set validations for 3D-QSAR models.** Two runs show variable predictions of efflux values of substrates. The **(A)** training set and **(B)** test set for the first run are shown along with the **(C)** training set and **(D)** test set for the second run. The cross-validated r^2 (q^2) is indicated for the training sets and the predictive r^2 is indicated for the test sets. While the first run shows a useful model, the second signifies instability of the model.

2.3.2 What factors might affect efflux?

The extrusion of ligands through an efflux pump such as AcrA-AcrB-TolC must depend on a number of factors, including certain properties of the substrates themselves as well as their interactions with the protein complex as measured by docking/scoring calculations.

First, the size of the substrate molecules would seem to be one of their most critical features. Molecular widths were calculated by molecular dynamics simulations, as described above in the Experimental Section. These values for each molecule are listed in Table 2.1. In designing this parameter, we had presumed that there would be a negative correlation, i.e., molecules with larger cross sections would be effluxed with more difficulty. Also, the previous reports of a correlation between LogP and efflux [31] mandated the use of this descriptor. Although it would definitely be preferable to incorporate experimental LogP values in our study, they are not uniformly available. The LogP calculations were performed using several methods, but the best correlation between predicted LogP and efflux was shown by ALogPs (Table 2.1). LogP alone (see Figure 2.4) is clearly insufficient to describe efflux effectively. In addition, to account for the dynamic nature of efflux, intermolecular interactions between efflux substrates and various zones or compartments of the pump must be coordinated in order to transport the substrates through the pump. Thus, the substrate molecules were docked into various locations within AcrB and TolC. These docked positions are direct representations of interactions between substrate and the efflux pump subunits. Using

the HINT model to score these interactions gives us an additional advantage of taking desolvation energy and entropy into account at these loci [39].

2.3.3 Systems hydropathy.

Since efflux pumps conduct substrates from the periplasmic space to the extracellular medium, this process is affected in multiple ways by interactions of the substrates with the transporter protein, which can be simulated by docking the substrates in the AcrA-AcrB-TolC pump. However, as there are not specific and well-defined docking region(s) within the pump, the substrates were docked into multiple zones or compartments, in both their charged and uncharged forms, as illustrated in Figure 2.1. Important sites were thus initially surveyed with β-lactams within the binding and extrusion states of AcrB. Since the TolC lumen is open to the extracellular space, ligands present here should be solvated and exist mostly in their charged forms. Again, the β-lactams were used to survey the potential sites within TolC. Compartmentalization of individual events (albeit of a different variety and on a different size scale) is also seen in the 'systems biology' approach; each compartmentalized effect is recorded individually, but the effects are viewed on a holistic level in order to study trends that cannot be observed by the reductionist approach. Similarly, we interpret the larger efflux effect as partially being a result of compartmentalized hydropathic interactions between substrate and pump, leading us to call our method 'systems hydropathy'.

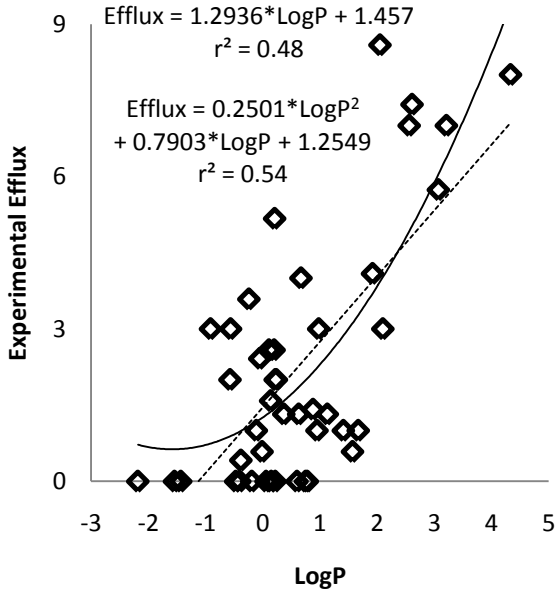


Figure 2.4 **Correlation Between ALogPs Predicted LogP Values and Efflux.** The plot of predicted LogP values versus efflux shows an r^2 value of 0.48, illustrating the fact that LogP alone does not allow prediction of antibiotic efflux. Also shown is the quadratic fit of Efflux with respect to LogP (Efflux = a + b*LogP + c*Logp^2), which shows an r^2 value of 0.55.

A partial least squares (PLS) analysis was conducted to explore the interplay between the substrate molecular properties and the interactions between the charged/uncharged β-lactam antibiotics and the AcrB and TolC proteins. Multiple equations of correlation were obtained through exhaustive exploration of the descriptor space. Only those descriptors that showed both interpretable trends and significant contributions to a model fitting the set of 32 β-lactam molecules were retained. With the important descriptors and binding sites thus identified, the 12 non-β-lactam substrates were docked and scored at those sites. The best combination of descriptors yielded the final multilinear model:

$$\text{Efflux} = -1.31 - (1.7\times10^{-4})*\text{HINT}_{nB} - (5.3\times10^{-4})*\text{HINT}_{cE} + (6.9\times10^{-4})*\text{HINT}_{AcrB(hole)} - (1.0\times10^{-3})*\text{HINT}_{Z3} + 1.10*\text{LogP} + 0.43*\text{MolWidth} \quad \text{(eq. 1)}$$

Here, HINT_{nB} is the HINT score of the neutral substrate docked to the AcrB binding state. Similarly, HINT_{cE} is the HINT score of the charged substrate in the AcrB extrusion state. $\text{HINT}_{AcrB(hole)}$ and HINT_{Z3} represent the HINT scores of the charged substrate at the intermonomeric space of AcrB and zone 3 of TolC, respectively. LogP values are as predicted by the ALogPs algorithm, chosen as described above due to their better correlation with efflux and MolWidth is the molecular width. A table with all descriptor values for each compound is given in appendix B.

Cross-validation with leave-one-out on the data set yielded a q^2 of 0.56 and an r^2 of 0.66 with 2 components for equation 1. Figure 2.5A displays the predictive model of equation 1 and table 2.2 sets out the predicted efflux and deviations calculated. To further evaluate the predictive ability of the model, the data set of 44 compounds was

randomly divided into training sets of 33 compounds and test sets of 11 compounds. New models were built with leave-one- out cross-validation over the training sets and used to predict the efflux of their corresponding test sets. The predicted test set efflux for a typical run of this nature is illustrated in Figure 2.6.

It is probably more important to classify substrates as being susceptible or not to efflux than to predict their numerical MIC ratio. Thus, using a definition of “high” efflux as ≥ 4, the equation 1 model was able to identify low/high efflux molecules with a 93.18% (41/44) success rate. Other results, i.e., with different high/low cutoffs, are summarized in Table 2.3. Note that even the predictions in error in terms of this binary classification scheme are often fairly close to the experimental efflux (Table 2.2). In summary, this method allows reliable predictions for whether a given antibiotic is a good substrate for efflux by the AcrA-AcrB-TolC pump.

Table 2.2. Efflux predictions for data set molecules.

Antibiotic	Average MIC ratio[a]	MIC ratio prediction (Equation 1)		MIC ratio prediction (Equation 2)	
		MIC ratio	Error	MIC ratio	Error
1	4	3.34	-0.66	2.78	-1.22
2	2.5	4.00	1.50	2.34	-0.16
3	2	10.13	8.13	6.48	4.48
4	1	3.31	2.31	2.73	1.73
5	1	0.42	-0.58	1.22	0.22
6	3	2.35	-0.65	1.89	-1.11
7	1	0.69	-0.31	2.18	1.18
8	1.5	6.21	4.71	6.39	4.89
9	1.5	2.37	0.87	2.57	1.07
10	12	3.64	-8.37	3.11	-8.89
11	5.33	1.19	-4.15	1.45	-3.88
12	2	3.46	1.46	4.97	2.97
13	1.33	1.50	0.16	1.44	0.11
14	1	1.61	0.61	1.28	0.28
15	1	2.32	1.32	2.43	1.43
16	17	13.70	-3.30	9.06	-7.93
17	1	1.21	0.21	1.39	0.39
18	2.5	8.77	6.27	5.16	2.65
19	2.67	1.72	-0.94	1.55	-1.12
20	1	2.88	1.88	1.48	0.48
21	384	59.84	-324.17	83.29	-300.72
22	2.5	2.09	-0.41	1.62	-0.88
23	36	4.88	-31.12	4.04	-31.97
24	128	64.80	-63.20	101.62	-26.38
25	171	133.07	-37.60	437.76	267.10
26	6	8.33	2.33	5.30	-0.70
27	16	14.55	-1.45	8.80	-7.20
28	1	1.80	0.80	1.33	0.33
29	2	12.73	10.73	10.01	8.01
30	4	4.49	0.49	4.25	0.25
31	1	0.41	-0.59	0.79	-0.21
32	1	3.49	2.49	2.46	1.46
33	6	0.76	-5.24	1.67	-4.33
34	8	11.73	3.73	5.61	-2.39
35	8	2.89	-5.11	3.26	-4.74
36	128	34.78	-93.22	30.80	-97.20
37	8	9.09	1.09	6.29	-1.71
38	8	4.35	-3.65	3.38	-4.62
39	1	1.78	0.78	1.28	0.28
40	53.3	43.32	-9.98	47.34	-5.96
41	2	9.12	7.12	6.78	4.78
42	256	100.92	-155.08	307.41	51.41
43	4	2.56	-1.44	3.33	-0.67
44	1	3.18	2.18	3.53	2.53

[a]The MIC ratio is the ratio of minimum inhibitory concentration (MIC) in the presence of efflux pump to the MIC in the absence of the pump.

Table 2.3 Classification accuracy of efflux predictive model.

Definition of "High" Efflux	% Correct Predictions (Eqn. 1)	% Correct Predictions (Eqn. 2)
≥4	93.18 (41/44)	93.18 (41/44)
≥3	79.55 (35/44)	84.09 (37/44)
≥2	72.73 (32/44)	70.45 (31/44)
≥1	77.27 (34/44)	72.73 (32/44)

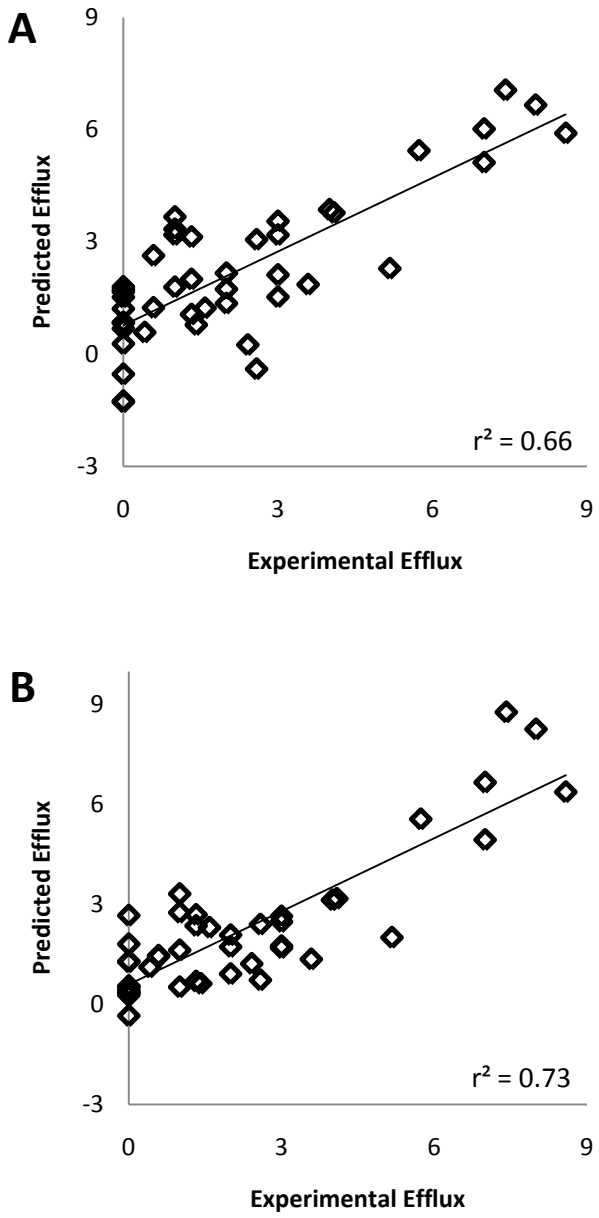


Figure 2.5 **Correlation Plots for Predicted vs. Experimental Efflux as Obtained with the Systems Hydropathy Approach.** Predicted versus experimental efflux values plotted for all 44 ligands based on **(A)** equation 1 and **(B)** equation 2.

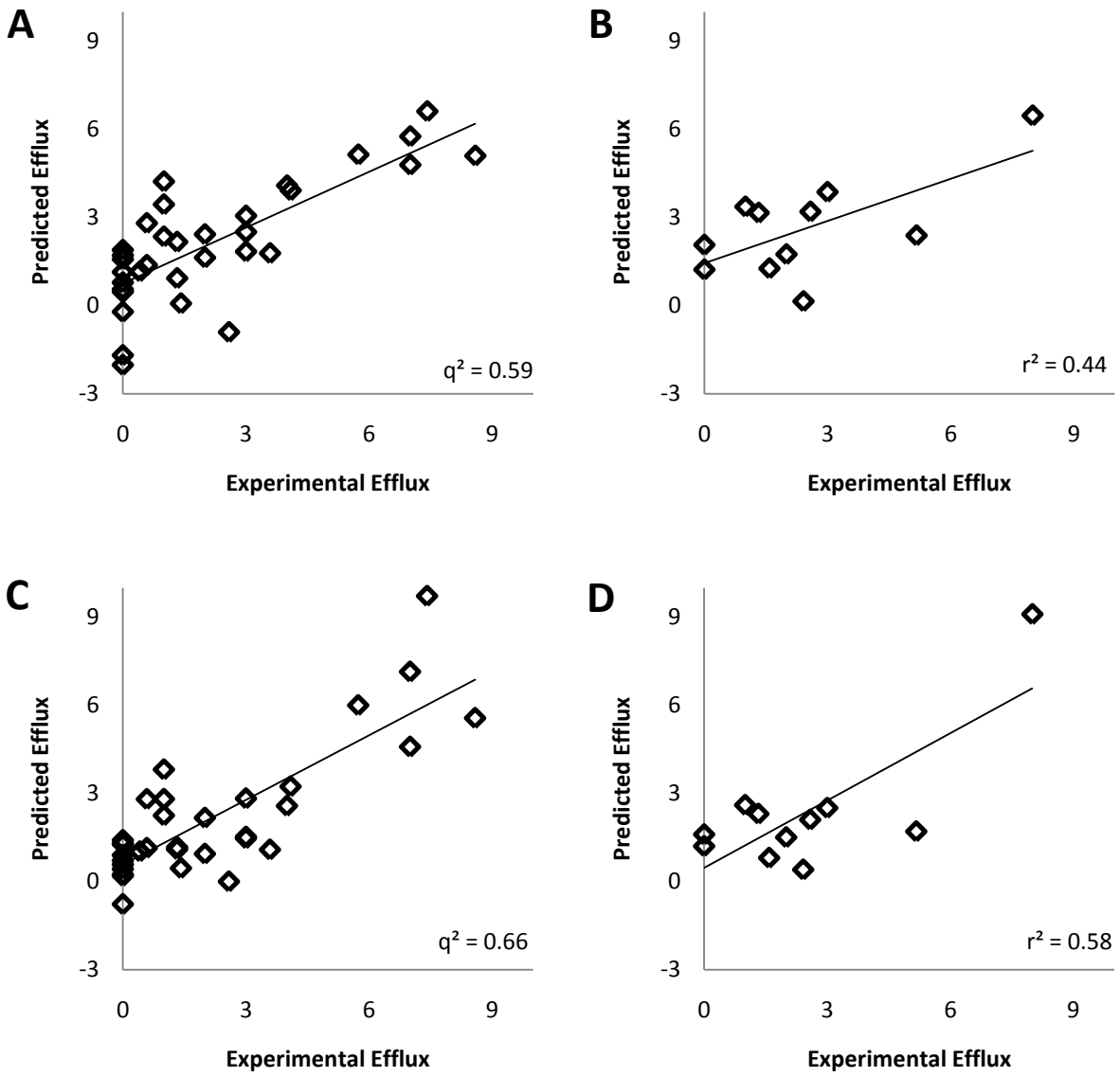


Figure 2.6 **Systems Hydropathy Validation.** Predicted Versus Experimental Efflux For Training and Test Set Substrates. **(A)** Correlation between the predicted and experimental efflux values for training set of 33 compounds using leave-one-out cross-validated model built with descriptors of eq. 1. **(B)** Correlation between the predicted and experimental efflux values for independent test set of 11 compounds using equation of (A). **(C)** Correlation between the predicted and experimental efflux values for training set of 33 compounds using leave-one-out cross-validated model built with descriptors of eq. 2 **(D)** Correlation between the predicted and experimental efflux values for independent test set of 11 compounds using equation of (C).

2.3.4 Model and descriptor interpretation.

A key requirement for a universally useful predictive model is that the physicochemical implications of the model's descriptors are interpretable and intuitive. The descriptors must yield not only statistical information but also chemical information that can be applied to fruitful drug design. However, when considered together, as in a regression equation, the model should have more value than the sum of its parts, i.e., the individual descriptors. Table 2.4 shows the fractional contribution of each descriptor for the model. All descriptors were found to have a significant contribution in the prediction of efflux values, with LogP having the largest, a nearly 41% contribution, to the model. In this section the descriptors and their qualitative and quantitative contributions to the overall model are described.

First, we should describe the roles that effects represented by LogP may play in the biological process. LogP represents more than solubility and related phenomena. This is especially true in the study of MIC ratios because the phenomenon is composed of two independent events: influx of the antibiotic through the outer membrane, followed by extrusion of the same by an efflux pump. LogP plays an important role during permeation of the antibiotic through the outer membrane, as has repeatedly been described by Lipinski's rules [59,60]. Also, since the AcrB binding pocket, which captures antibiotics and other substrates to commence efflux, is hydrophobic, it is not hard to imagine the importance of this descriptor in the process of efflux. Especially when the range of compounds is large, LogP can often be better represented in a quadratic form in QSAR equations, i.e., $a + b*LogP + c*LogP^2$ [61,62]. This form allows

for the likely scenario that both too high and too low LogP are detrimental to the biological effect being modeled and that there is an optimum range of LogP. In the case of efflux (i.e., $\log_2$[MIC ratio]), even though the linear relationship between LogP and efflux, i.e., Efflux = a + b*LogP, is fairly good (r^2 = 0.48), there is a modest improvement with the quadratic to r^2 = 0.55 (see Figure 2.4). Both forms of the regression support our expectation that more hydrophobic moieties are more easily effluxed by the pump. The quadratic probably provides a better fit due to the peculiarities of the data – in that there is no negative efflux – presumably, compounds with very negative LogP would have an efflux of zero. Overall, both support the model that ligands are initially captured by AcrB and then transported into the solvated TolC lumen and that those ligands with highly negative LogP values will be unable to enter the hydrophobic AcrB binding pocket. We hypothesize that ligands with highly positive LogP values (more positive than in this data set) would be able to easily enter AcrB but then resist deposition into the polar environment of the TolC lumen and thus "clog" the pump.

Table 2.4. Fractional contribution of descriptors to models.

Descriptor	Percent Contribution (Equation 1)	Percent Contribution (Equation 2)
$LogP^2$	-	20.5
LogP	40.9	22.0
$HINT_{Z3}{}^2$	-	12.7
$HINT_{Z3}$	18.2	10.9
$HINT_{AcrBhole}{}^2$	-	4.2
$HINT_{AcrBhole}$	11.9	3.5
$HINT_{nB}$	2.6	0.7
$HINT_{cE}$	13.0	13.3
MOLWIDTH	13.5	10.6

It is likely that $HINT_{nB}$ appears in the equation because these substrates should bind to the AcrB protein's binding state in their neutral form. It has previously been proposed that efflux pumps capture antibiotics from the periplasmic region [34]. Since this region is lined on either side by lipid bilayers, this environment is less polar than that of the cytosolic medium. This would, thus, partially shift the acid-base equilibria for ligands in the periplasmic space towards their non-ionic (more hydrophobic) forms. As there are several phenylalanine residues in the AcrB binding pocket, it is thus quite hydrophobic and this, the entrance to the pump, would preferentially bind less polar ligands, or those that are in a non-ionic form at the time of capture. This would suggest that a number of substrate molecules linger near the entrance and only pass within when their equilibrium-mediated ionization state matches the requirements of the AcrB binding pocket. Confusingly, this term has a negative correlation coefficient that suggests strong binding here disfavors efflux. This, in a sense, would appear to be true, as very strong binding to this site should cause the substrate to be "stuck" and not effluxed. Of course, negligible binding to this site should also be a negative factor, but presumably, intermediate binding should favor efflux. Much as above, this should be a classic case for using quadratic descriptors (e.g., $a + b*HINT_{nB} + c*HINT_{nB}^2$) in constructing regression models, but as the contribution of $HINT_{nB}$ in the model of equation 1 is quite small (see Table 2.3), we should not expect a strong correlation in models using only this independent variable in any case. Thus, as observed in figure 2.8A, the attempt to fit efflux with only the $HINT_{nB}$ quadratic descriptor yielded a poor regression ($r^2 = 0.030$) compared to the linear fit of efflux with $HINT_{nB}$ ($r^2 = 0.028$). Moreover, it must be noted

that while a high $HINT_{nB}$ score would favor the capture of antibiotics by AcrB, this phenomenon can also be facilitated, at least in part, by simply “partitioning” molecules into the AcrB binding pocket, a phenomenon that is likely encoded in the dominant LogP descriptor.

After AcrB captures a ligand, it undergoes a conformational change as observed in its crystal structure (PDB code 2drd) [34], where both ligand-bound and unbound states were observed in the same multimeric structure. The conformational change opens up the binding pocket towards the TolC lumen. The ligand would now likely be bathed in water entering the AcrB binding pocket from the extracellular environment through TolC. This (suddenly) now polar environment would shift the acid-base equilibria towards favoring charged forms of the ligands. Thus, we propose the significance of the $HINT_{cE}$ descriptor that represents binding in the AcrB extrusion state in our models. Similar to above for $HINT_{nB}$, the $HINT_{cE}$ descriptor also has a negative correlation coefficient in the models and thus also represents a complex effect. In this case the effect was not better modeled in this data set by using a quadratic representation: $a + b*HINT_{cE} + c*HINT_{cE}^2$ ($r^2 = 0.148$) cf. linear ($r^2 = 0.148$, figure 2.8B), so the simple explanation is that tight binding here is detrimental to efflux. Substrate ligands washed out of the hydrophobic pocket of AcrB in their charged forms would have a higher affinity towards the intermonomeric region that has a higher density of charged residue sidechains. The positive coefficient of $HINT_{AcrBhole}$ is indicative of the “pull” exerted on ligands by this region, enabling it to exit the extrusion state of AcrB towards the TolC channel. For this descriptor, the quadratic representation ($r^2 = 0.044$)

gives a slightly better correlation than that of the simple linear model (r^2 = 0.003), but both are poor, in concert with the small contribution of this descriptor to the model. This does suggest the possibility that both weak and tight binding can inhibit efflux. Linear and quadratic relationships between efflux and $HINT_{AcrBhole}$ are demonstrated in figure 2.8C.

On traversing through the TolC lumen, the substrate would successively interact with multiple positions on the protein. In accordance with our theory that stronger interactions slow down extrusion of ligands, the term $HINT_{Z3}$ correlates negatively with efflux in our regression equations. This is easily explained by looking at the inner surface map of TolC (figure 2.9A). Zone 3, which happens to be a deep pocket, is found in the center of the TolC lumen surface (figure 2.9B). Although a transient attraction between the substrate and the residues at this site may favor the substrate's passage, strong interaction with this deep pocket would slow or halt the passage of ligands through the TolC lumen, thereby reducing the extent of efflux by the pump. This effect is somewhat better modeled with a quadratic descriptor (r^2 = 0.219) rather than linear (r^2 = 0.192), which are delineated in figure 2.8D.

Molecular width (MolWidth) appears in equation 1 and possesses an unexpected *positive* correlation coefficient; i.e., larger substrates are more favorably extruded by the pump. We are proposing that this effect arises from 'induced fit' of larger substrates on AcrB, thus forcing AcrB to transform from the unbound and flaccid access state to the much larger binding state. Also, the substrate bulk could force a change in tertiary structure that causes AcrB to assume the extrusion state. There is experimental

evidence to support this hypothesis: the crystal structure (PDB ID 2drd) of AcrB clearly shows that the binding pocket is shrunken in the access state, while in the binding state it is wide open with an entrenched ligand [34]. After releasing the ligand, the pocket returns to its shrunken conformation in the extrusion state. It has previously been suggested that the proton pump mechanism provides energy for conformational change [63]. We propose that binding of larger ligands might supplement (or trigger?) this effect by providing steric "encouragement." However, in agreement with our initial assumption, we still believe that transitioning beyond a certain size should also be detrimental for efflux. In other words, this descriptor also should be better represented in quadratic form. However, as observed in figure 2.8E, the quadratic relationship (r^2 = 0.055) yielded only a small improvement when compared to the linear model (r^2 = 0.054). This is possibly due to the small range of molecular width possessed by the compounds in our dataset.

To consolidate the above information, we suggest equation 2, in which LogP, $HINT_{AcrBhole}$ and $HINT_{Z3}$ are modeled in their quadratic forms:

$$\text{Efflux} = -2.09 - (4.9\times10^{-5})*HINT_{nB} - (6.1\times10^{-4})*HINT_{cE} + (2.3\times10^{-4})*HINT_{AcrBhole} - (6.8\times10^{-4})*HINT_{Z3} + 0.66*\text{LogP} + 0.44*\text{MolWidth} + 0.22*\text{LogP}^2 + (2.1\times10^{-7})*HINT_{AcrB(hole)}^2 + (1.04\times10^{-7})*HINT_{Z3}^2 \quad \text{(eq. 2)}$$

This model has q^2 = 0.63 and r^2 = 0.73 with 2 components. Its results are presented in Figures 2.5B, 2.6C and 2.6D and Tables 2.2, 2.3 and 2.4. While a noteworthy improvement in statistical parameters was observed, no apparent change in the model's ability to correctly predict high/low efflux (Table 2.3). However, it must be

acknowledged that this model is at higher risk of being statistically invalid because of the addition of three more fitted parameters for the coefficients of the squared terms.

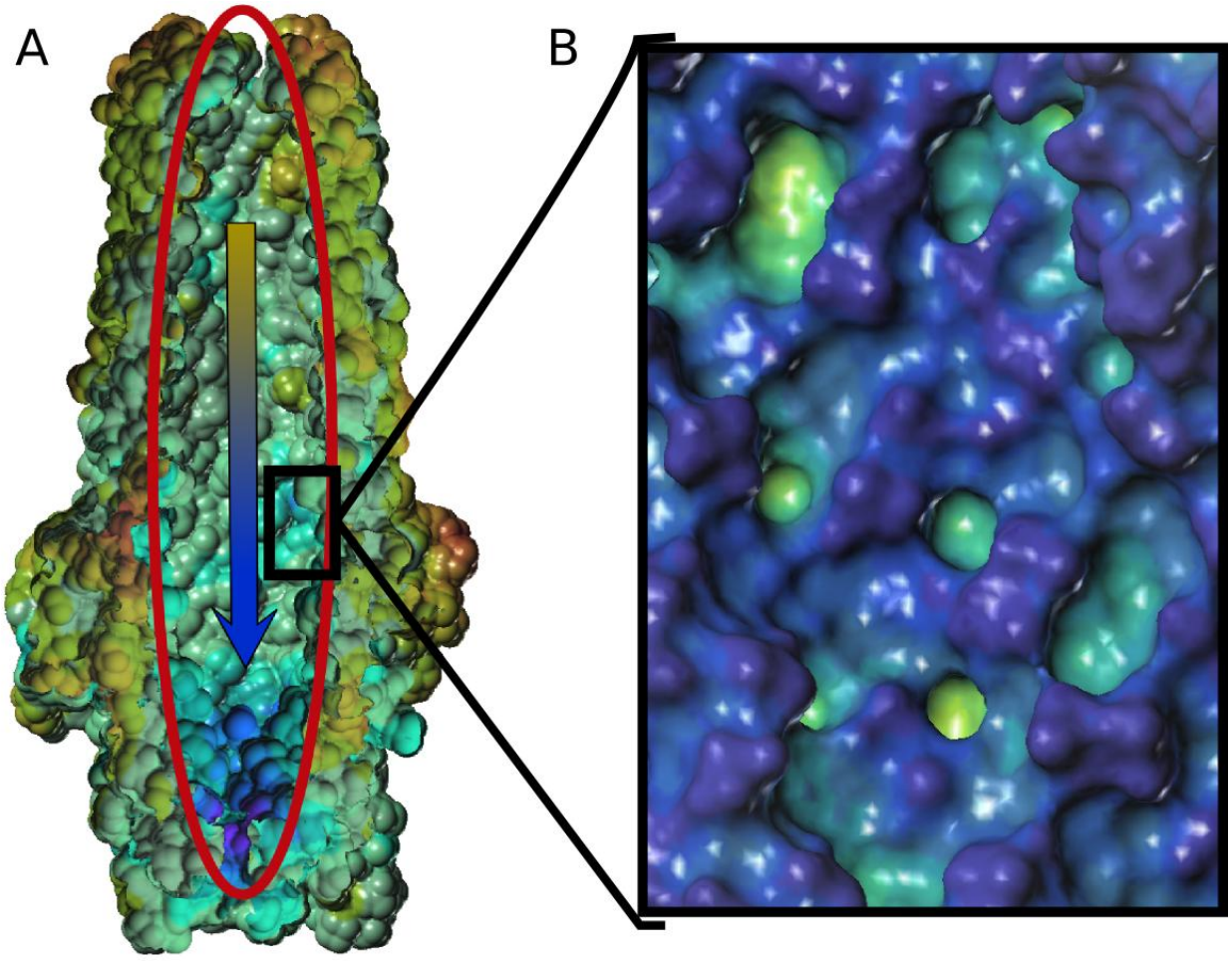


Figure 2.7 **Surface Maps for TolC. (A)** An electrostatic surface map of TolC shows the surface of the TolC lumen (enclosed by the red circle) **(B)** Zone 3 of the TolC efflux pump is a deep cavity on the wall of the lumen, with blue depicting peaks while blue-green depicts troughs.

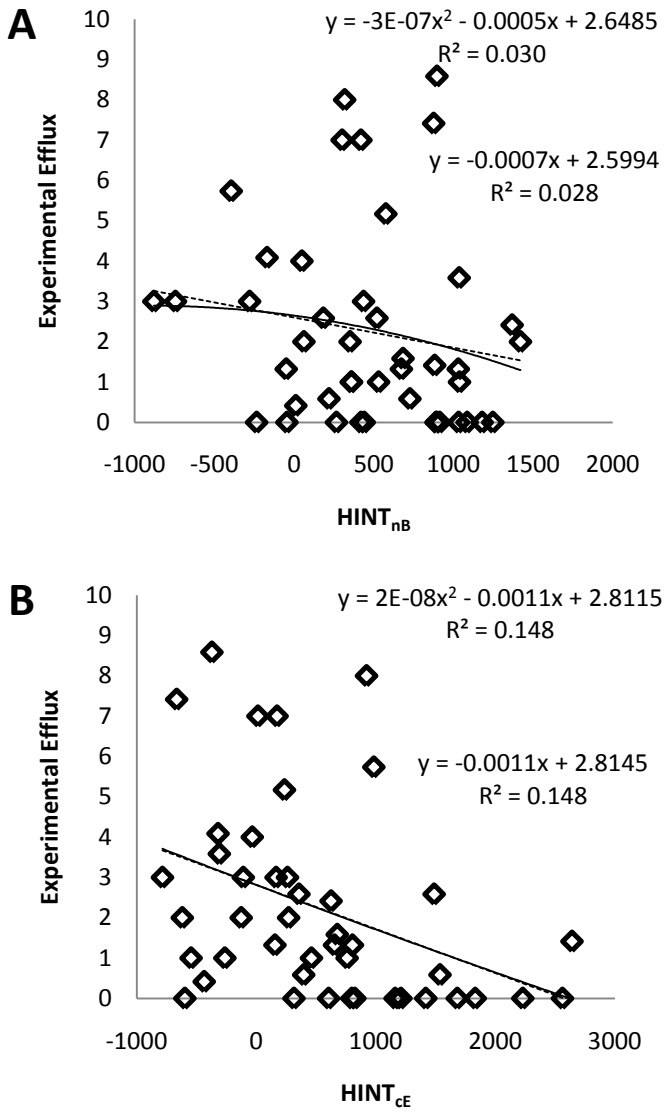


Figure 2.8 Relationship between Efflux and Individual Descriptors. Linear and quadratic models for the relationship between efflux and each descriptor used in this study are demonstrated. Efflux as a function of (A) $HINT_{nB}$ and (B) $HINT_{c\#}$ are demonstrated.

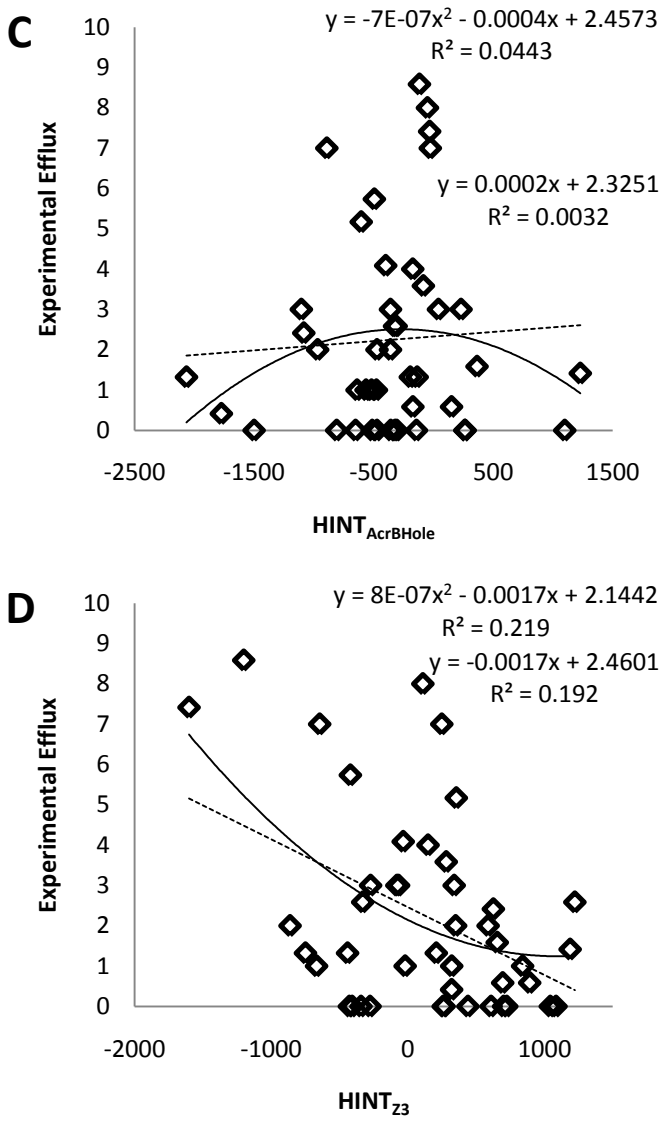


Figure 2.8 ***continued.*** Efflux as a function of (C) $HINT_{AcrBhole}$ and (D) $HINT_{Z3}$ are demonstrated.

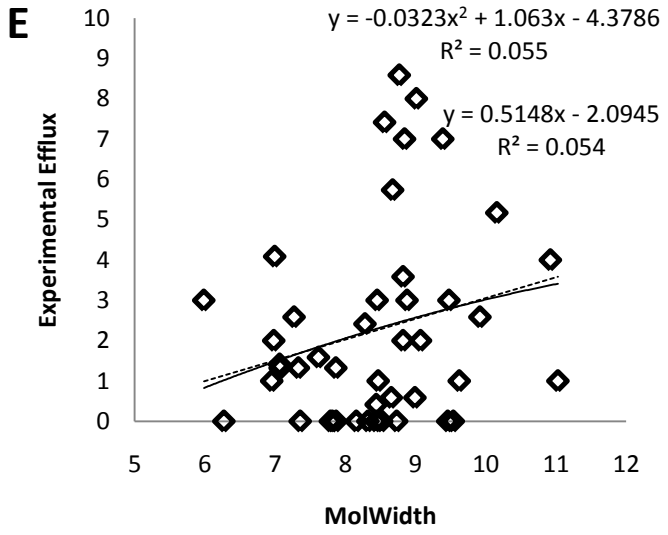


Figure 2.8 *continued.* Efflux as a function of (E) molecular width.

Figure 2.9 summarizes our proposed mechanism for the AcrA-AcrB-TolC efflux process. The dielectric environment within the periplasmic region is unknown, but it is likely that there is less water present between the two lipid bilayers than in either the extracellular or cytoplasmic regions (see Figure 2.9A that indicates the color scales used in the remaining panels of Figure 2.9). Furthermore, the periplasmic space is occupied by peptidoglycan chains and a gel containing a wide variety of enzymes, which should tend to reduce the polarity. Efflux substrates in the periplasm would exist in a reversible equilibrium between their charged and uncharged forms that slightly favors the uncharged forms (Figure 2.9B, inset) in this (slightly) more hydrophobic region. The uncharged forms will be more likely captured by the AcrB hydrophobic pocket (Figure 2.9B), upon which AcrB will assume the extrusion state (Figure 2.9C) partly due to the bulk of the substrate, as indicated by the positive correlation between efflux and molecular width. It should be pointed out that if a substrate molecule binds too tightly to the binding state form of AcrB, or if it cannot sterically trigger the extrusion state of AcrB, that substrate would appear to be immune from extrusion and may block the pump's function. The absence of favorable substrate binding at this state would also preclude efflux. Once the extrusion state is formed, the AcrB entrance is closed and thus isolated from the periplasmic space but now open towards TolC, exposing the substrate to water present in the TolC lumen (Figure 2.9C). Similarly, the ligand's ionization equilibrium is concomitantly shifting towards the charged form (Figure 2.9D,

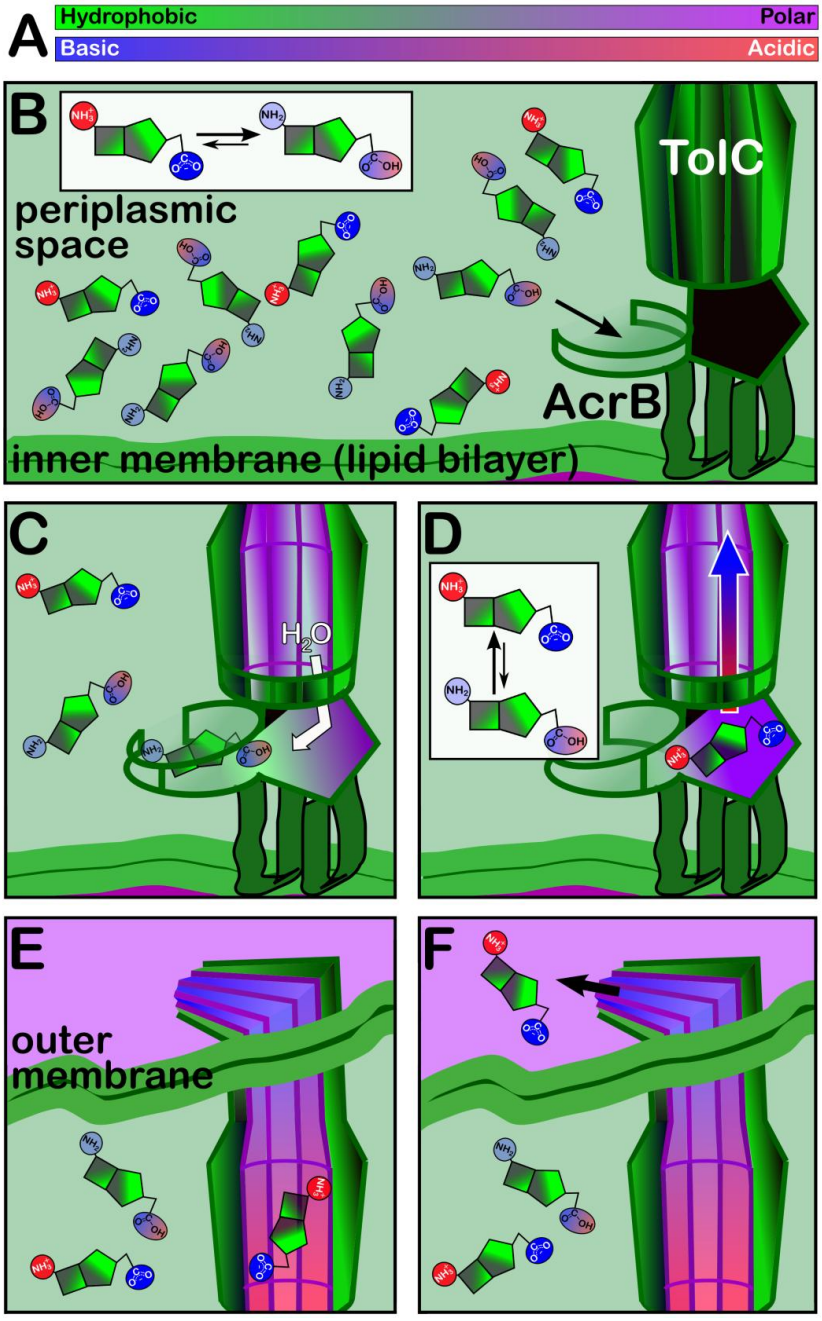


Figure 2.9 ***Proposed Efflux Mechanism.*** **(A)** Color ramps for hydrophobic-polar (green to purple) and acid-base (blue to red) spectra. **(B)** Antibiotics present in the periplasmic region exist in equilibrium between their charged and uncharged states (inset). The uncharged forms predominate due to the local environment and are captured by the AcrB hydrophobic pocket. **(C)** AcrB assumes extrusion state partly due to the bulk of the substrate and is isolated from the periplasmic space. The cavity is now open towards TolC, exposing the substrate to water present in the TolC lumen. **(D)** The ionization equilibrium shifts towards the charged form (inset) and the substrate is released into the intermonomeric space of AcrB. **(E)** The charged efflux substrate is now able to diffuse through the water present in the TolC lumen. The electric field present inside the TolC lumen causes orientation of the substrate such that negatively and positively charged groups point towards opposite ends of the protein. **(F)** The efflux substrate is released into the extracellular space.

inset), leading to release of the substrate into the intermonomeric space of AcrB (Figure 2.9D). Again, if the substrate ligand binds too weakly or too tightly to the AcrB extrusion state, it may not be effluxed and, in fact, may in the latter case block the pump. However, both these "clogging" events involve equilibria that may reverse to unclog the pump. The charged efflux substrate can now diffuse through water present in the TolC lumen (Figure 2.9E). Ultimately the substrate will be extruded into the comparatively polar extracellular medium (Figure 2.9F).

2.4 CONCLUSIONS

Despite the engineering of Nature to facilitate the extrusion of undesired molecules within a cell, there are likely to be multiple reasons why a particular substrate is resistant to efflux. In addition to the obvious descriptors of hydrophobicity (LogP) and size (although the correlation with molecular width we observed was initially counterintuitive), the ligand's ability to bind *and* release from various pockets within the pump machinery is at least as critical as the aforementioned descriptors that are not cognizant of its interactions with the pump. However, despite the relatively successful predictions of efflux by these models, there are a number of considerations inherent in the approach that should be discussed. Primary is the dataset itself. Unfortunately, the available data is both relatively small in quantity and skewed towards the lower efflux range, which corresponds to the range of molecules of more clinical interest. Our development of regression model equations with six descriptors on 44 substrates using PLS is not ideal, while our expanded model with three of these variables represented in

quadratic form is potentially bordering on overfitting. Although we have some comfort from the fact that these models were subjected to cross-validation, which yielded good statistical parameters, Wold and Dunn [64] state that, even when using PLS, regression studies are only valid when the number of independent variables is far less than the number of dependent target values. Clearly, we would like to have a larger data set, but restricted this analysis to the β-lactam class of antibiotics and some non-β-lactam antibiotics because of their loose chemical similarities that, in turn, suggest efflux extrusion by the same pump and mechanism. Data from a wider class of antibiotics are potentially available, but their use may be premature in testing a new computational method.

In conclusion, we have proposed the systems hydropathy approach, which has been used in this work to predict efflux values of the AcrA-AcrB-TolC efflux pump. The analogy to systems biology stems from our combining the various compartmentalized functions of the pump's protein components into a holistic model that has more value than a reductionist analysis of the pump. Nevertheless, the model suggested some interesting mechanistic details about the efflux process that seem intuitively true. On further development, this approach could be expanded to more non-β-lactam antibiotics, other efflux systems affecting antibiotic efflux and potentially mammalian efflux systems that have been shown to extrude, among other molecules, anticancer chemotherapeutic agents. The key puzzle piece is obtaining structural data for the protein components of additional pump molecules.

2.5 REFERENCES

1. Gold, H.S.; Moellering Jr., R.C. Antimicrobial-drug resistance *N. Engl. J. Med.* **1996,** *335,* 1445-1453.
2. Neu, H.C. The crisis in antibiotic resistance. *Science* **1992,** *257,* 1064-1073.
3. Walsh, C. Molecular mechanisms that confer antibacterial drug resistance. *Nature* **2000,** *406*, 775-781.
4. Spratt, B.G. Penicillin-binding proteins and the future of β-lactam antibiotics. *J. Gen. Microbiol.* **1983,** *129,* 1247-1260.
5. Broome-Smith, J.; Spratt, B.G. An amino acid substitution that blocks the deacylation step in the enzyme mechanism of penicillin-binding protein 5 of Eschiricia coli. *FEBS Lett.* **1984,** *165,* 185-189.
6. McMurry, L.; Petrucci, R.E., Jr.; Levy, S.B. Active efflux of tetracycline encoded by four genetically different tetracycline resistance determinants of Eschericia coli. *Proc. Natl. Acad. Sci. U.S.A.* **1980,** *77,* 3974-3977.
7. Webber, M.A.; Piddock, L.J.V. The importance of efflux pumps in bacterial antibiotic resistance. *J. Antimicr. Chemother.* **2003,** *51,* 9-11.
8. Bambeke, V.F.; Balzi, E.; Tulkens, P.M. Antibiotic efflux pumps. *Biochem. Pharmacol.* **2000,** *60,* 457-470.
9. Nikaido, H. Multidrug efflux pumps of gram-negative bacteria. *J. Bacteriol.* **1996,** *178,* 5853-5859.
10. Lin, J.; Michel, L.O.; Zhang, Q. Cme ABC functions as a multidrug efflux system in Campylobacter jejuni. *Antimicrob. Agents Chemother.* **2002,** *46,* 2124-2131.
11. Pumbwe, L.; Piddock, L.J.V. Identification and characterisation of CmeB, a Campylobacter jejuni multidrug efflux pump. *FEMS Microbiol. Lett.* **2002,** *206,* 185-189.
12. Poole, K. Efflux mediated resistance to fluoroquinolones in gram-negative bacteria. *Antimicrob. Agents Chemother.* **2000,** *44,* 2233-2241.
13. Gill, M.J.; Brenwald, M.P.; Wise, R. Identification of an efflux pump gene pmrA, associated with fluoroquinolone resistance in Streptococcus pneumoniae. *Antimicrob. Agents Chemother.* **1999,** *43,* 187-189.
14. Nikaido, H. Preventing drug access to targets: cell surface permeability barriers and active efflux in bacteria. *Seminars Cell. Developmental Biol.* **2000,** *12,* 215-233.
15. Kaatz, G.W.; Seo, S.M. Inducible NorA-mediated multidrug resistance in Staphylococcus aureus. *Antimicrob. Agents Chemother.* **1995,** *39,* 2650-2655.
16. Saier, M.H., Jr. Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya. *Adv. Microbiol. Physiol.* **1998,** *40,* 81-136.
17. Lee, A.; Mao, W.; Warren, M.; Mistry, A.; Hoshino, K.; Okumura, Y.; Ishida, H.; Lomovskaya, O. Interplay between efflux pumps may provide either additive or multiplicative effects on drug resistance. *J. Bacteriol.* **2000,** *182,* 3142-3150.
18. Ma, D.; Cook, D.N.; Alberti, M.; Pon, N.G.; Nikaido, H.; Hearst, J.E. Genes acrA and acrB encode a stress-induced efflux system of Escherichia coli. *Mol. Microbiol.* **1995,** *16,* 45-55.

19. Thanassi, D.G.; Cheng, L.W.; Nikaido, H. Active efflux of bile salts by Escherichia coli. *J. Bacteriol.* **1997,** *179,* 2512-2518.
20. Sánchez, L.; Wubin, P.; Viñas, M.; Nikaido, H. The acrAB homolog of Haemophilus influenza codes for a functional multidrug efflux pump. *J. Bacteriol.* **1997,** *179,* 6855-6857.
21. Gotoh, N.; Murata, T.; Ozaki, T.; Kimura, T.; Kondo, A.; Nishino, T. Intrinsic resistance of Escherichia coli to mureidomycin A and C due to expression of the multidrug efflux system AcrAB-TolC: comparison with the efflux systems of mureidomycin-susceptible Pseudomonas aeruginosa. *J. Infect. Chemother.* **2003,** *9,* 101-103.
22. Yang, S.; Clayton, S.R.; Zechiedrich, E.L. Relative contributions of the AcrAB, MdfA and NorE efflux pumps to quinolone resistance in Escherichia coli. *J. Antimicr. Chemother.* **2003,** *51,* 545-556.
23. Mazzariol, A.; Cornaglia, G.; Nikaido, H. Contributions of the AmpC b-lactamase and the AcrAB multidrug efflux system in intrinsic resistance of Escherichia coli K-12 to b-lactams. *Antimicr. Ag. Chemother.* **2000,** *44,* 1387-1390.
24. Nikaido, H.; Basina, M.; Nguyen, V.; Rosenberg, E. Multidrug efflux pump AcrAB of Salmonella typhimurium excretes only those β-lactam antibiotics containing lipophilic side chains. *J. Bacteriol.* **1998,** *180,* 4686-4692.
25. Nishino, K.; Yamada, J.; Hirakawa, H.; Hirata, T.; Yamaguchi, A. Roles of TolC-dependent multidrug transporters of Escherichia coli in resistance to β-lactams. *Antimicr. Ag. Chemother.* **2003,** *47,* 3030-3033.
26. McMurry, L.M.; Oethinger, M.; Levy, S.B. Overexpression of marA soxS or acrAB produces resistance to triclosan in laboratory and clinical strains of Escherichia coli. *Fed. Euro. Microbiol. Soc. Microbiol. Lett.* **1998,** *166,* 305-309.
27. Sulavik, M.C.; Houseweart, C.; Cramer, C.; Jiwani, N.; Murgolo, N.; Greene, J.; DiDomenico, B.; Shaw, K.J.; Miller, G.H. Hare, R.; Shimer, G. Antibiotic susceptibility profiles of Escherichia coli strains lacking multidrug efflux pump genes. *Antimicr. Ag. Chemother.* **2001,** *45,* 1126-1136.
28. Hou, T.; Wang, J.; Zhang, W.; Wang, W.; Xu, X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006,** *13,* 2653-2667.
29. Johnson, S.R.; Zheng, W. Recent progress in the computational prediction of aqueous solubility and absorption. *AAPS J.* **2006,** *8,* E27-E40.
30. Clark, D.E. Computational prediction of ADMET properties: recent developments and future challenges. *Annu. Re. Comput. Chem.* **2005,** *1,* 133-151
31. Ferreira, M.M.C.; Kiralj, R. QSAR study of b-lactam antibiotic efflux by the bacterial multidrug resistance pump AcrB. *J. Chemometrics,* **2004,** *18,* 242-252.
32. Kiralj, R.; Ferreira, M.M.C. Molecular graphics approach to bacterial AcrB protein-b-lactam antibiotic molecular recognition in drug efflux mechanism. *J. Mol. Graph. Model.* **2006,** *25,* 126-145.
33. Borst, P.; Oude Elferink, R. Mammalian ABC transporters in health and disease. *Annu. Rev. Biochem.* **2002,** *71,* 537-592.

34. Murakami, S.; Nakashima, R.; Yamashita, E.; Matsumoto, T.; Yamaguchi, A. Crystal structures of a multidrug transporter reveal a functionally rotating mechanism. *Nature* **2006,** *443,* 173-179.
35. Kononakis, V.; Sharff, A.; Koronakis, E.; Luisi, B.; Hughes, C. Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein transport. *Nature* **2000,** *405,* 414-419.
36. Tripos Sybyl molecular modeling suite, version 8.1. www.tripos.com/sybyl
37. Lee, A.; Mao, W.; Warren, M.S.; Mistry, A.; Hoshino, K.; Okumura, R.; Ishida, H.; Lomovskaya, O. Interplay between efflux pumps may provide either additive or multiplicative effects on drug resistance. *J. Bacteriol.* **2000**, *182*, 3142-3150.
38. Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997,** *267,* 727-748.
39. 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.
40. 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.
41. Spyrakis, F.; Amadasi, A.; Fornabaio, M.; Abraham, D.J.; Mozzarelli, A.; Kellogg, G.E. The consequences of scoring docked ligand conformations using free energy correlations. *Eur. J. Med. Chem.* **2007,** *42,* 921-933.
42. Tripathi, A.; Fornabaio, M.; Kellogg, G.E.; Gupton, J.T.; Gewirtz, D.A.; Yeudall, W.A.; Vega, N.E.; Mooberry, S. Docking and hydropathic scoring of polysubstituted pyrrole compounds with antitubulin activity. *Bioorg. Med. Chem.* **2008,** *16,* 2235-2242.
43. Simoni, D.; Invidata, F.P.; Eleopra, M.; Marchetti, P.; Rondanin, R.; Baruchello, R.; Grisolia, G.; Tripathi, A.; Kellogg, G.E.; Durrant, D.; Lee, R.M. Design, synthesis and biological evaluation of novel stilbene-based antitumor agents. *Bioorg. Med. Chem.* **2009,** *17,* 512-522.
44. Tetko, I.V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description, *J. Comput. Aid. Mol. Des.* **2005,** *19,* 453-463.
45. VCCLAB, Virtual Computational Chemistry Laboratory, http://www.vcclab.org, **2005.**
46. Tetko, I.V. Computing chemistry on the web, *Drug Discov. Today,* **2005,** *10,* 1497-1500.
47. Bush, B.L.; Nachbar, R.B. Jr. Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J. Comput.-Aided Mol. Des.* **1993,** *7,* 587-619.
48. Abraham, D.J.; Kellogg, G.E.; Holt, J.M.; Ackers, G.K. Hydropathic analysis of the non-covalent interactions between molecular subunits of structurally characterized hemoglobins. *J. Mol. Biol.* **1997,** *272,* 613-632.

49. 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 β37 mutant hemoglobins. *Biochemistry* **2000,** *39,* 1622-1633.
50. Burnett, J.C.; Botti, P.; Abraham, D.J.; Kellogg, G.E. Computationally accessible method for estimating free energy changes resulting from site-specific mutations of biomolecules: systematic model building and structural/hydropathic analysis of deoxy and oxy hemoglobins. *Prot. Struct. Funct. Gen.* **2001,** *42,* 355-377.
51. Cashman, D.J.; Kellogg, G.E. A computational model for anthracycline binding to DNA: tuning groove-binding intercalators for specific sequences. *J. Med. Chem.* **2004,** *47,* 1360-1374.
52. Cashman, D.J.; Rife, J.P.; Kellogg, G.E. Docking and hydropathic analysis of Hoechst 33258 with double-stranded RNA. *Med. Chem. Res.* **2003,** *12,* 445-455.
53. Porotto, M.; Fornabaio, M.; Greengard, O.; Murrell, M.T.; Kellogg, G.E.; Moscona, A. Paramyxovirus receptor-binding molecules: engagement of one site on the hemagglutinin-neuraminidase protein modulates activity at the second site. *J. Virol.* **2006,** *80,* 1204-1213.
54. Porotto, M.; Fornabaio, M.; Kellogg, G.E.; Moscona, A. A second receptor binding site on human parainfluenza virus type 3 hemagglutinin-neuraminidase contributes to activation of the fusion mechanism. *J. Virol.* **2007,** *81,* 3216-3228.
55. Rosenberg, E.Y.; Ma, D.; Nikaido, H. AcrD of *Escherichia coli* is an aminoglycoside efflux pump. *J. Bacteriol.* **2000**, *182*, 1754-1756.
56. Elkins, C.A.; Nikaido, H. Substrate specificity of the RND-type multidrug efflux pumps AcrB and AcrD of *Escherichia coli* is determined predominantly by two large periplasmic loops. *J. Bacteriol.* **2002**, *184*, 6490-6498.
57. Yu, E.W.; Aires, J.R.; McDermott, G.; Nikaido, H. A periplasmic drug-binding site of the AcrB multidrug efflux pump: a crystallographic and site-directed mutagenesis study. *J. Bacteriol.* **2005**, *187*, 6804-6815.
58. Cramer III, 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.
59. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, *46*, 3-26.
60. Lipinski, C.A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Meth.* **2000**, *44*, 235-249.
61. Hansch, C; Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616-1626.
62. Romanelli, G.P.; Cafferata, L.F.R.; Castro, E.A. An improved QSAR study of toxicity of saturated alcohols. *J. Mol. Struct. (Theochem)*. **2000**, *504*, 261-265.
63. Su, C.-C.; Li, M.; Gu, R.; Takatsuka, Y.; McDermott, G.; Nikaido, H.; Yu, E.W. Conformation of the AcrB multidrug efflux pump in mutants of the putative proton relay pathway. *J. Bacteriol.* **2006,** *188,* 7290-7296

64. Wold, S.; Dunn III, W.J. Multivariate quantitative-structure activity relationships (QSAR) – conditions for their applicability. *J. Chem. Inf. Comput. Sci.* **1983,** *23,* 6-13.

CHAPTER 3

TARGETING PARAINFLUENZA VIRUS TYPE 3 BY VIRTUAL SCREENING; THE NEED FOR NEW TOOLS

3.1 AN INTRODUCTION TO HUMAN PARAINFLUENZA VIRUSES

Among the four human serotypes of parainfluenza viruses, Human Parainfluenza Viruses (HPIV) 1, 2, 3 and 4, HPIV3 is mostly implicated in bronchial pneumonia [1]. All the serotypes are known to be causative agents of acute lower respiratory diseases in infants and children [1,2]. HPIV3, which belongs to the Paramyxoviridae family of negative-stranded RNA viruses, is responsible for approximately 11% of the hospitalizations of pediatric patients in the United States [2]. Cell mediated immunity is important for preventing parainfluenza related diseases. For example, HPIV3 infection in T-cell deficient children can cause fatal giant-cell pneumonia and HPIV pneumonia shows 30% mortality in bone-marrow transplant patients [2].

Unlike other viral diseases, HPIV primary infections are known to not confer permanent immunity [1]. For example, 30% of children with relatively high neutralizing antibody counts were reinfected with the virus [2]. In fact in the 1960s, clinical trials of the inactivated HPIV 1, 2 and 3 vaccines showed variable amounts of antibody responses in seropositive and seronegative individuals, but failed to produce immunity [1]. This inability of the human immune system to provide protection against these

versatile viruses is a worrying factor, especially due to fears raised by a recent flurry of related fatal avian and porcine influenza disease occurrences around the world.

The plausible impact of these viruses on healthcare (especially pediatric healthcare) clearly delineates the importance of drug discovery efforts against the human parainfluenza virus. Such an effort has been undertaken, the procedures and results of which are described herein.

3.2 HEMAGGLUTININ-NEURAMINIDASE IN HPIV3 REPLICATION

The life cycle of HPIV3 starts with the recognition of sialic acid containing receptors on the host cell by hemagglutinin-neuraminidase (HN), which then triggers another membrane protein (F) [3]. The F protein is responsible for fusion of the viral membrane with the host cell membrane [2]. Although sialic acid alone is sufficient to trigger HN mediated F protein activity, it must be noted that not all sialic acid-containing receptors are recognized equally effectively [4]. Moreover, the neuraminidase function of HN is responsible for the release of new virions from the host cell [5] and thus is responsible for persistent infectivity of the virus.

It has been shown that mutations on HN cause modulation of immune responses toward this pathogen [2]. However, it is amply clear that unlike the Influenza A virus, parainfluenza viruses do not evolve by mutation of this membrane glycoprotein [1]. In fact, these viruses show a high sequence homology of 75% between even the human and bovine variants. The hemagglutinin and neuraminidase epitopes of HN are conserved across both these strains of the virus [1], which suggests the importance of

this protein in its life cycle. It is thus logical to devise strategies in order to inhibit this protein.

3.3 INHIBITION OF HEMAGGLUTININ-NEURAMINIDASE STOPS VIRAL ACTIVITY

In order to understand the inhibition of HN and its mechanism, one must first understand the biology behind the assays used to study this phenomenon. Following is a description of assays described in literature, which are used to distinguish hemagglutination and neuraminidase functions of HN.

3.3.1 Neuraminidase Assays

Potier and coworkers introduced a simple fluorometric assay for the quantitative assessment of neuraminidase activity [6]. The basic principle of this assay is to spectroscopically measure the release of 4-methylumbelliferone from the sodium salt of 2'-(4-methylumbelliferyl)-α-D-*N*-neuraminic acid, when the substrate is exposed to a neuraminidase enzyme (figure 3.1).

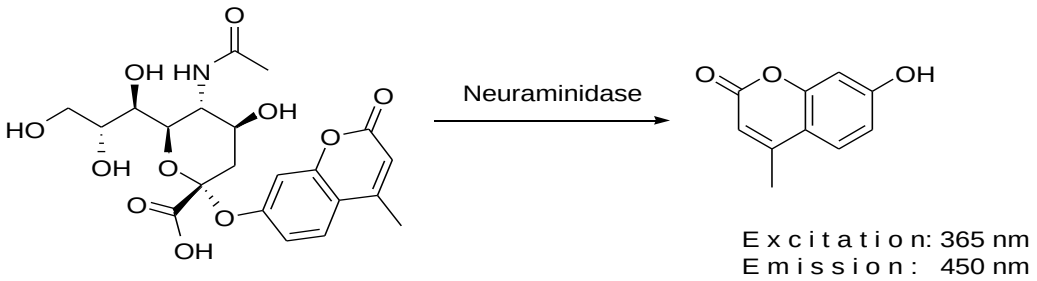


Figure 3.1 ***Principle of the Fusion Assay.*** Release of 4-methylumbelliferone as a result of hydrolysis of 2'-(4-methylumbelliferyl)-a-D-N-neuraminic acid by neuraminidase is measured by fluorescence spectroscopy

A similar assay was also developed by Warner and O'Brien in the same year [7]. This method is accurate, with up to a 3% variation [6] observed in the original literature by these authors.

This assay is used to assess the ability of the virus to cleave neuraminic acid from receptors of host cells, leading to the destruction of these proteins and enabling movement of new virions towards uninfected cells. Those viruses which lack neuraminidase activity (or else whose neuraminidase activity is reduced due to the presence of neutralizing agents such as antibodies and small molecule inhibitors) are unable to travel towards new plausible host cells in order to spread infection. Therefore, the most common use of this assay is to ascertain the persistent infective capabilities of the virus.

3.3.2 Fusion Assay

Horvath *et al.* have shown that the F protein can mediate membrane fusion of cells under the influence of HN [8]. However, in spite of the fact that both proteins are expressed by cells persistently infected with HPIV3, they do not fuse with each other, but readily fuse with non-infected cells [9]. This is because of the neuraminidase activity of HN, which cleaves neuraminic acid from receptors, thereby inhibiting cell fusion mediated by the protein. Further support for this theory comes from the fact that neuraminidase treated uninfected cells do not fuse with persistently infected cells [9].

This phenomenon has been utilized in the development of an effective and accurate assay in order to study the recognition of sialic acid (*N*-acetylneuraminic acid)

containing receptors by HN [10]. Cells containing the LacZ gene (genetic code for production of β-galactosidase) under the control of HIV LTR (HeLa-LTR-βgal cells) are infected with HPIV3 in order to cause a persistent infection, but as explained above, these infected cells do not fuse with one another. These persistently infected cells are then exposed to cells engineered to express the HIV Tat protein (HeLa-Tat cells). Only when these cells fuse does the Tat protein interact with the HIV LTR, thereby promoting the production of β-galactosidase by expression of the LacZ gene. The expression levels of this enzyme can be measured readily, thereby revealing the extent of cell fusion as an indirect measure of HN-receptor interactions.

3.3.3 Plaque Reductions Assay

When Bloom, Jimenez and Marcus first introduced the plaque assay [11], it was aimed at studying the effect of various antigens on antigen-sensitive cells. Observations clearly suggested that on antigen activation, such cells became more capable of supporting viral replication. A monolayer of such cells was grown in a petri dish and then exposed to live viral cultures. Any free virus was washed away post inoculation and the infected cell culture was incubated. Infectious centers were observed as plaques, which were directly related to the degree of activation of the antigen-sensitive cells. A similar technique was also reported in later years [12].

Across time, this technique has been converted to a purely virology technique, where the degree of plaque formation suggests the ability of the virus to replicate [10].

The number of plaques being formed is directly proportional to the degree to which the virus can replicate.

3.3.4 Hemadsorption Assay

When persistently infected cell lines are exposed to erythrocytes, whose cell membrane contains sialic acid containing receptors, they are adsorbed onto the surface of the infected monolayer [10]. An interesting fact is that the persistently infected cells fail to fuse with erythrocytes, perhaps because of a difference in membrane composition or cytoskeletal stiffness of the latter [13]. After incubation for a short time, washing removes any erythrocytes which are not adsorbed. Such adsorbed cells can then be visualized by phase contrast microscopy. The degree of adsorption is a direct representation of the HN expression levels of the infected cells and hence is a measure of hemagglutinin activity.

3.3.5 Neuraminic Acid Interaction with HN Mediates Membrane Fusion

It has been shown that cell fusion is mediated by the interaction between HN and sialic acid containing cell receptors [13]. When uninfected cells were treated with neuraminidase to destroy sialic acid containing receptors on their surface, membrane fusion was not observed in a fusion assay. Similarly, cells which do not produce sialic acid containing receptors did not show fusion with persistently infected cells in the fusion assay. Moreover, it has also been demonstrated that HN is specific in its selection of sialic acid containing receptors [4].

3.3.6 DANA and GANA inhibit Hemagglutinin Function of HN

2,3-dehydro-2-deoxy-n-acetyl neuraminic acid (DANA) is known to inhibit viral neuraminidase activity and its mode of action was investigated [10]. It was found that DANA blocked fusion as well as βgal production at 10 mM concentrations. The abolition of hemadsorption in the presence of DANA suggests that it blocks HN-receptor interactions (hemagglutinin function). DANA could inhibit 90% of plaque formation at 25mM concentrations.

Similar assays with 4-guanidino-DANA (GANA, a.k.a. Zanamivir) revealed mechanistic details of this compound as well [14]. It blocked hemadsorption in C28a, a variant of the HPIV3 virus that lacks neuraminidase activity, attesting to its HN-receptor interaction blocking abilities. The plaque forming capabilities of both WT and C28a viruses were blocked by DANA, demonstrating its ability to block fusogenic activity of HN. The lack of neuraminidase function of C28a was corrected by adding exogenous neuraminidase, which cleaves the host cell sialic acid-containing receptors. Thus, newly formed virions are able to avoid attachment to the host cell and were released into the environment. GANA also allowed release of new virions into the environment, palpably by blocking HN mediated recognition of host cell receptors. In contrast, when cell lines were exposed to WT HPIV3, virion release could not be blocked by addition of GANA. This proves GANA inhibits HN by preventing binding to host cell receptors.

3.4 VIRTUAL SCREENING FOR HN INHIBITORS

The above studies clearly showed the utility of sialic acid derivatives in inhibiting HN activity and also HPIV3 by consequence. Therefore, inhibition of HN is a plausible mode for antiviral activity. We therefore embarked on a search for inhibitors of HN. Following is a description of the methods adopted for the search, along with a discussion about problems encountered in the process.

3.4.1 Pharmacophore Identification

The crystal structures of HN in its unliganded form, bound to sialic acid, DANA and GANA were published by Lawrence et al. in 2004 (PDB ID: 1v3b, 1v3c, 1v3d and 1v3e respectively) [15]. The protein shows a six-blade β-propeller shape and was crystallized in a dimeric form.

The unliganded binding pocket of HPIV3 HN is similar to that of influenza virus neuraminidase, which is published elsewhere [16]. Figure 3.2 shows the interactions between HN and its ligands. In spite of differences between sidechain positions of all three HN-ligand complexes, the core sidechain positions of HN itself remain similar: Three arginine residues (R192, R424, R502) project into one side of the cavity, while the floor of the same side contains a tyrosine sidechain (Y530) hydrogen bonded to E409. The sidechain position for Y530 varies across different forms of HN crystals obtained, perhaps an effect caused by the variety of crystallization conditions used. However, it is interesting that in one of the two observed sidechain positions of Y530, it is “slotted in” so as to form interactions with two highly conserved residues –

hydrophobic pi-pi interactions with Y478 and hydrogen bonds with the P194 backbone. In the same conformation, Y530 also forms a hydrogen bond with a water molecule. R424 is also hydrogen bonded to E409. Another face of this cavity is partially hydrophobic due to the presence of Y319. All in all R192, R424, R502, Y530, E409, E549 and D216 form the active site residues.

When crystals of HN were soaked in 5-acetylneuraminic acid (sialic acid), the crystal structure thus obtained (figure 3.2A) reveals that the carboxylate group on the ligand interacts favorably with the three arginine cluster (R192, R424 and R502), although only two of these (R192 and R502) are hydrogen bonded with it. The hydroxyl group of Y530 and the glycosidic oxygen of sialic acid are less than 2.3 Å apart, while the latter is only 3.5 Å away from the carboxylate group of E409. This suggests hydrogen bond formation between these residues and the sialic acid glycosidic hydroxyl group. The C7 and C9 hydroxyl groups of sialic acid are hydrogen bonded to E276, while the *N*-acetyl group forms a hydrophobic interaction with Y319. The largest structural perturbation caused by binding of sialic acid to HN observed in these crystal structures is the movement of Y530 into the binding site cavity, which was observed in the "tucked in" conformation within the unliganded form in 30% of the crystals formed. A water molecule also exists near the triarginyl cluster and the ligand carboxylate group.

The triarginyl cluster remains almost in the same position in the structure of HN bound to GANA, except for a short movement of the guanidino group of R192 closer towards the substrate carboxylate group (figure 3.2B). This suggests that an additional

hydrogen bond might be formed between the protein and ligand due to the change in position of the carboxylate caused by introduction of unsaturation in the pyranose ring.

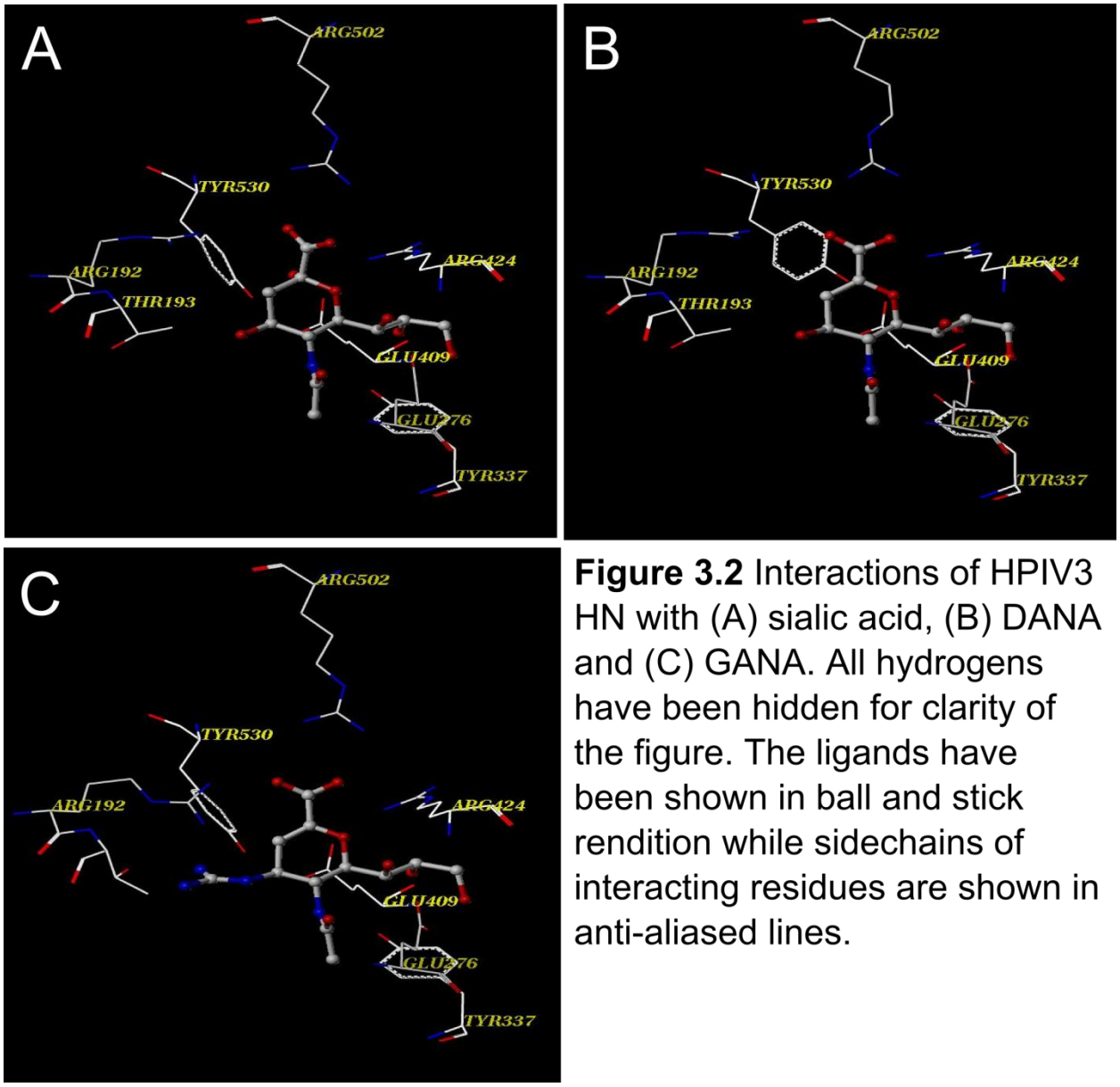


Figure 3.2 Interactions of HPIV3 HN with (A) sialic acid, (B) DANA and (C) GANA. All hydrogens have been hidden for clarity of the figure. The ligands have been shown in ball and stick rendition while sidechains of interacting residues are shown in anti-aliased lines.

The water molecule observed in the HN-sialic acid complex was not observed in this complex, which perhaps contributes towards increased ligand affinity. The hydrogen bonds formed between the glycosidic hydroxyl group of sialic acid and the two residues E409 and Y530 are now lost, but these sidechains are stabilized by formation of an intramolecular hydrogen bond. The hydrogen bonds between the sidechain hydroxyl groups of sialic acid and E276 remained with DANA.

The binding of GANA to HN causes a reversal of certain effects observed in the complex with DANA (figure 3.2C). The R192 sidechain guanidino group now moves away from the carboxylate of GANA, but forms a stacking interaction with the latter's guanidino group. The sidechain of Y530 now extends towards the guanidino group of GANA, forming a hydrogen bond with it. T193 is also hydrogen bonded with the guanidino group of GANA. All other interactions remain similar to those with DANA and sialic acid. The structures of sialic acid, DANA and GANA bound to HN are shown in figure 3.2.

Based on these interactions, it was clear that hydrogen bond acceptors were desirable in the triarginyl region of the HN binding pocket, while a hydrogen bond donor in the region where the GANA guanidino group is bound also increases affinity. Also, the hydrogen bonds between the C7 and C9 of sialic acid were maintained throughout the three structures of GANA and DANA with HN. These observations helped identify key features of the pharmacophore for the creation of queries.

3.4.2 Design of Queries

Two queries were designed based on the above pharmacophoric model. The pyranose ring of GANA was defined as a hydrophobic center, which was surrounded by three hydrogen bond donor features at the C7 and C9 hydroxyl groups and the guanidino group. A negative center was defined in the vicinity of the carboxylate group. The hydrophobic center was not defined for a query based on DANA, but was supplemented with an acceptor atom in the vicinity of the pyranose oxygen. Also, an acceptor atom and a steric feature were defined at the acetyl amino oxygen and methyl group respectively. Both queries are delineated in figure 3.3.

3D flex searches were run on the ZINC database, which contained 3,820,641 compounds at the time when this work was performed. These searches utilize a torsional minimizer in order to identify molecules which might adopt a conformation that fits the query being used. Approximately 3000 hits were identified from the queries, including a wide variety of chemical scaffolds.

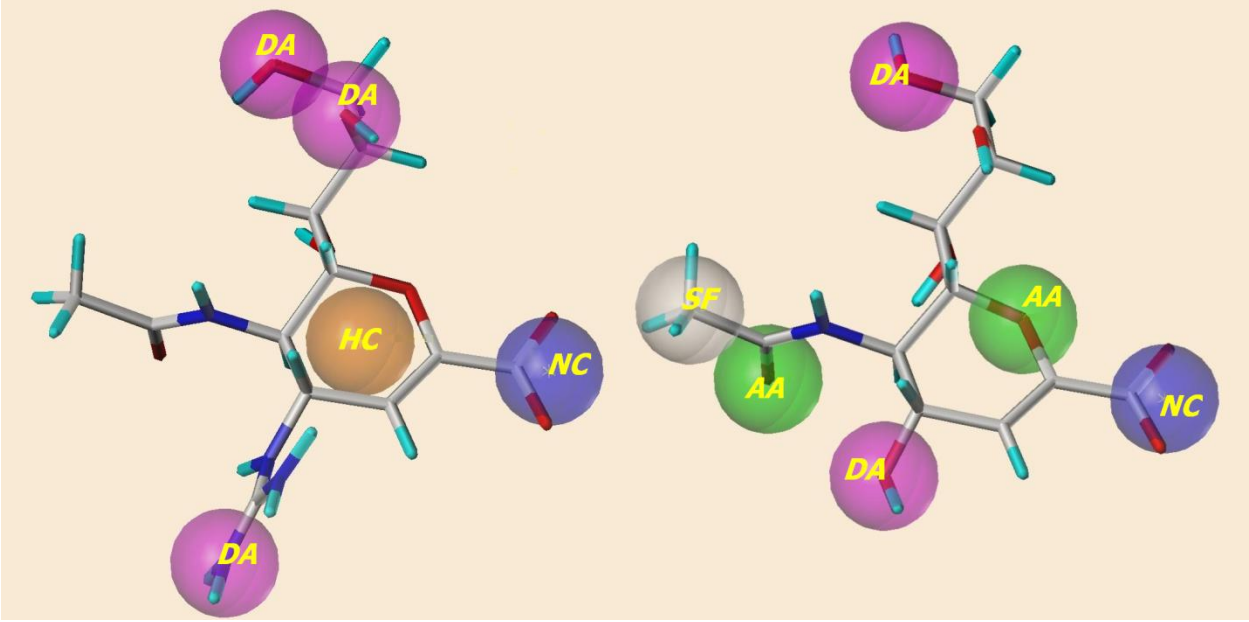


Figure 3.3 Queries on the ZINC Database. Orange spheres represent hydrophobic centers. The following features were defined: DA, donor atom; HC, hydrophobic center; NC, negative center; AA, acceptor atom; SF, steric feature.

3.4.3 Segregation of Drug-Like and Non Drug-Like Compounds

Not all the hits contained drug-like properties. Hence, these hits were carefully screened for non drug-like properties. The molecules were segregated based on three criteria: (1) drug-like versus non drug-like scaffolds and (2) Lipinski's rule of five.

3.4.3.1 Identification of drug-like and non drug-like scaffolds

Although the ZINC database is regularly used for virtual screening in order to identify drugs, our hits contained a number of molecules which do not seem drug-like. These molecules were identified by visual inspection and eliminated from further study. A few examples of such hits are shown in figure 3.4.

3.4.3.2 Lipinski's rule of five

Lipinski published a set of rules which increase the chances of finding molecules with favorable permeation and absorption characteristics which have been outlined in Chapter 1 (*vide supra*). This set of rules has been implemented in the UNITY module of Sybyl, which was used for datamining purposes in this project. The inbuilt molecular screen identifies agents with less than 5 hydrogen bond donors and 10 hydrogen bond acceptors. Moreover, it identifies molecules with a molecular weight less than 500 D and a CLogP of less than 5.

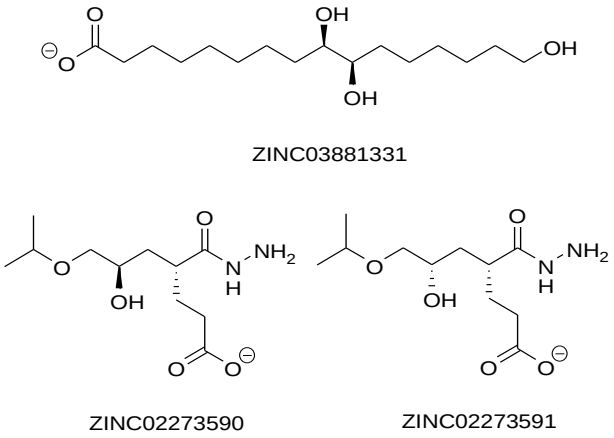


Figure 3.4 ***Structures Rejected by Visual Inspection.*** Some examples of molecules that were deemed nondrug-like and excluded from further analysis.

After trimming the list of ligands using these criteria, 1513 compounds remained.

3.4.4 Docking

The ability to theoretically predict interactions between molecules based on chemo-spatial considerations and the principles of physics has been of immense

interest to scientists across the decades. Research across several decades has given us multiple algorithms known as docking algorithms. Kuntz aptly described his docking algorithm as a method to explore geometrically feasible alignments of ligands and receptors of known structure [17]. The first examples of docking were perhaps those published by Levinthal et al. in 1975 and Salemme in 1976 [18,19], describing attempts at predicting structures of complexes in order to understand macromolecular interactions. Kuntz's algorithm [17] was perhaps the first description of a program that attempted to address docking of small molecules into proteins. This is a problem of immense complexity due to the large number of degrees of freedom and the resultant local minima an algorithm has to explore in order to obtain appropriate and accurate predictions.

Several strategies have evolved in order to address this complex and multifaceted problem, including incremental construction approaches such as FlexX [20], shape-based algorithms such as DOCK [21], genetic algorithms such as GOLD[22], systematic searches such as Glide [23], Monte Carlo simulations such as LigandFit [24] and surface-based molecular similarity methods such as Surflex [25]. In spite of the availability of several such algorithms, ligand docking does suffer from several problems [26]. However, its utility in computational ligand design is undeniable, especially due to the high benefit/cost ratio.

We utilized GOLD 3.0 to dock those hits which remained from visual inspection of virtual screening results into the HN binding pocket. One hundred genetic algorithm (GA) runs per compound were executed in order to obtain multiple poses of putative

ligands within the binding site. The total number of ligands remaining after visual inspection of hits obtained from UNITY was 1513, thus resulting in 151,300 protein-ligand complexes that needed to be scored.

3.4.5 Scoring of Docked Positions

The HINT forcefield [27,28] was used to score all protein-ligands complexes obtained via docking because scores obtained from this forcefield are known to correlate with binding free energy [29-32]. A list of top scoring compounds, along with their ZINC codes, is displayed in appendix A. The crystallographic structures of HN with Sialic Acid, GANA and DANA [15] showed HINT scores of 111, 1283 and 1673 respectively. These compounds were suggested for purchase and biochemical evaluation. The decision to test these 137 compounds was based on the comparison of HINT scores with those of the ligands bound to the crystal structures. An increase of 515 HINT units is associated with a 1kcal/mol increase in affinity; hence, many of these structures were expected to be strong binders of HN.

3.5 PROBING ANTIVIRAL MECHANISM

Anne Moscona's group at Weill-Cornell Medical School tested 50 of the 137 compounds (*vide supra*) and found 17 inhibitors of HN. In order to probe the mechanism of action of these inhibitors, a further docking study was conducted. Selected compounds were docked into site II (*vide infra*) of HN.

Previous studies by the Moscona and Kellogg research groups had already predicted the possibility of a second site on HN, which might interact with sialic acid

containing receptors [33-35]. It has been shown that mutations on this site of HN can cause premature triggering of the F-protein, thereby rendering the virus noninfectious [35]. In order to probe the possible effects of one of the 17 inhibitors of HN (ZINC02857325, Figure 3.5) identified by virtual screening, it was docked into site II and the HINT scores were compared to those of GANA at both sites (table 3.1).

Table 3.1 HINT analysis of inhibitors at site I and II of HN

Compound	HINT Score	
	HPIV3 HN Site I	HPIV3 HN Site II
GANA	2769	660
ZINC02857325	2260	1108

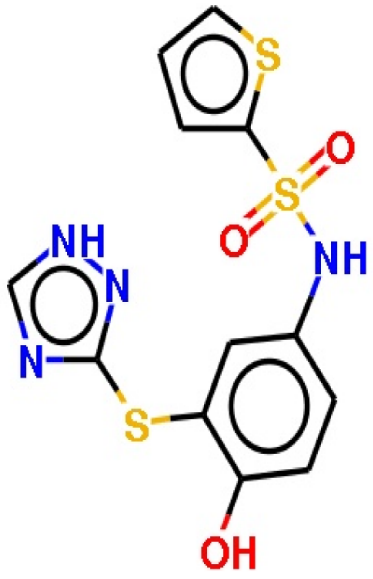


Figure 3.5 Structure of ZINC02857325, an inhibitor of HN identified by virtual screening

Based on these scores, it is possible that both GANA and ZINC02857325 will bind site I of HN. However, it is possible that this compound will bind site II with a higher affinity compared to GANA based on its HINT score at the same site. It has already been shown that site II is resistant to GANA [33] and therefore validates our docking results, therefore strengthening our belief that this compound might interact at site II of HN.

Furthermore, in the best docked position of GANA at site II of HN the guanidino group protrudes out of the protein into the surrounding medium and presumably interacts with water (figure 3.6A). It is highly possible that this causes the ligand to be solvated and thereafter vacates the binding site. It is possible that such an effect will not be observed with the much more hydrophobic ZINC02817325.

3.6 PROBLEMS WITH DOCKING

Several cases were identified, in which the rotation of sidechains could improve HINT scores. An example is delineated in figure 3.7.

This test case is a complex of ZINC02857325 docked into HN. The docked position itself showed a HINT score of ~ -10^6. This was basically due to a clash between the phenolic hydroxyl of Tyr337 and the sulfonamidic hydrogen. This situation could easily be remedied by a simple rotation of the chi_1 angle to alleviate the steric clash. When this was performed manually, an additional hydrogen bond was formed between the phenolic group of Tyr337 and the sulfonamidic hydrogen. This complex was then rescored using HINT and an improvement of over 10^6 HINT units was observed. Of

course, this was mostly due to the removal of the steric clash. An additional improvement was also visible when a different rotamer of Lys254 was placed, causing the formation of an additional H-bond between the triazole nitrogen and the protonated amino group of the residue.

This example clearly indicates the possibility that docking may not be able to accurately recreate the most probable binding pose for ligands without addressing target flexibility. Even after minimization, which would alleviate steric clashes, there is a chance that the most favorable binding pose may not be predicted because of the nature of minimization algorithms; they are designed to traverse downwards along the energy potential function and as a result will not overcome barriers in attaining the global minimum. In essence, it is probably not realistic to expect identification of the global minimum every time, but a reasonable investigation into target flexibility is required. We suggest rotating residue sidechain chi angles as a method for exploring induced fit.

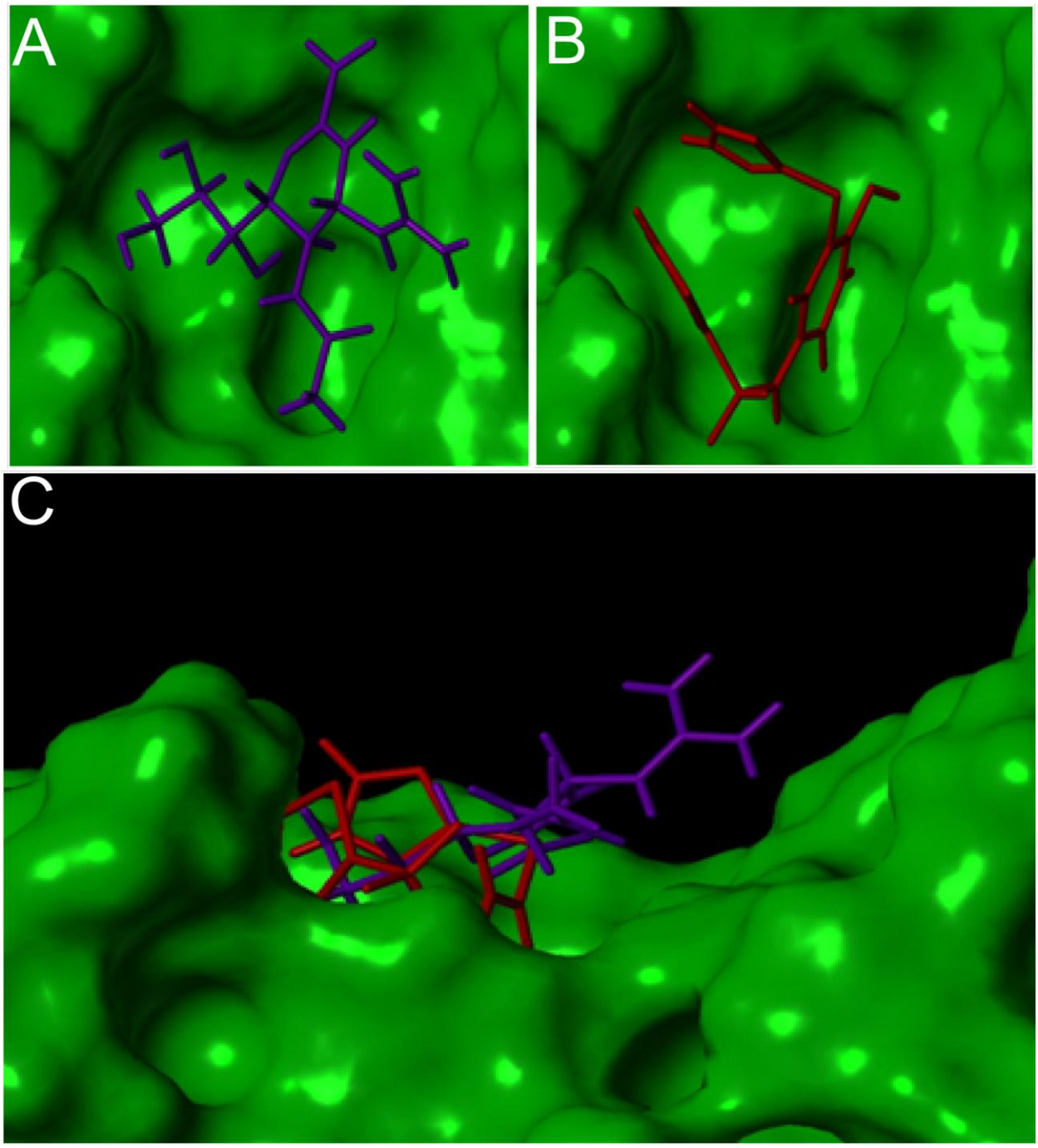


Figure 3.6 ***Interactions of GANA and ZINC02857325 with site II of HPIV3 HN.*** **(A)** and **(B)** show the best docked positions of GANA and ZINC02857325 at site II of HN. **(C)** shows the depth of both ligands; the guanidino group of GANA protrudes into the surrounding environment, while the more hydrophobic ZINC02857325 shows no such protrusion.

3.7 SUMMARY

A virtual screening for inhibitors of HPIV3 HN was conducted using datamining techniques, docking procedures and the HINT scoring function. This process identified several inhibitors of the protein, which are as of yet being tested for their mode of action in the Moscona laboratory at Weill Medical College of Cornell University. A 34% success rate has been observed so far (calculated as number of compounds found active for every 100 compounds tested experimentally).

It is common knowledge that virtual screening is, as of yet, only in its infancy; a lot of work has to be done in order to improve the tools and procedures which are currently in use. These problems [30] include a number of issues, such as protein and ligand flexibility, role of water in binding and solvation, the combinatorial issue of protonation/deprotonation of ligand and residues and perhaps the biggest problem of all – scoring functions. Serious and long-term research needs to be performed in order to address this multi-faceted problem.

In our quest to identify hits against HPIV3 HN, one problem we noticed was target flexibility; although some docking programs already take this into account, this led us to try and develop our own new tools. Chapter 4 addresses the current status of our attempts in this direction.

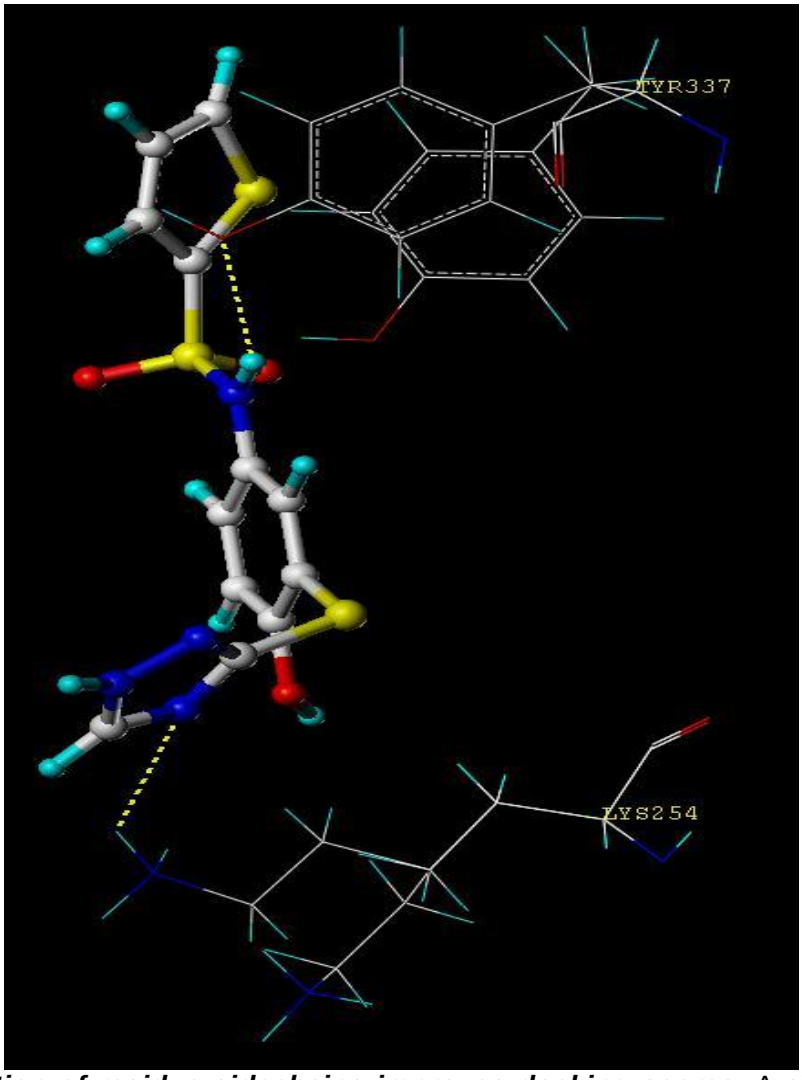


Figure 3.7 ***Rotation of residue sidechains improves docking scores.*** A situation is shown where simple rotation of amino acid sidechains increased HINT scores for ZINC02857325 docked into site I of HPIV3 HN. This clearly indicates the possibility that the "correct" binding pose for this ligand may not have been identified.

3.8 REFERENCES

1. Heilman, C.A. From the national institute of allergy and infectious diseases: respiratory syncytial and parainfluenza viruses. *J. Infect. Dis.* **1990,** *161,* 402-406.
2. Loughlin, G.M.; Moscona, A. The cell biology of acute childhood respiratory disease: therapeutic implications. *Pediatr. Clin. N. Am.* **2006,** *53,* 929-959.
3. Moscona, A.; Peluso, R.W. Relative affinity of the human parainfluenza virus type 3 hemagglutinin-neuraminidase for sialic acid correlates with virus-induced fusion activity. *J. Virol.* **1993,** *67,* 6463-6468.
4. Moscona, A.; Peluso, R.W. Analysis of human parainfluenza type 3 receptor binding variants: evidence for the use of a specific sialic acid-containing receptor. *Microb. Pathogen.* **1996,** *20,* 179-184.
5. Huberman, K.; Peluso, R.W.; Moscona, A. Hemagglutinin-neuraminidase of human parainfluenza 3: role of the neuraminidase in the viral life cycle. *Virol.* **1995,** *214,* 294-300.
6. Potier, M.; Mameli, L.; Bélisle, M.; Dallaire, L.; Melançon, S.B. Fluorometric assay of neuraminidase with a sodium (4-methylumbelliferyl-α-D-*N*-acetylneuraminate) substrate. *Anal. Biochem.* **1979,** *94,* 287-296.
7. Warner, T.G.; O'Brien, J.S. Synthesis of 2'-(4-methylumbellifuryl)-a-D-N-acetylneuraminic acid and detection of skin fibroblast neuraminidase in normal humans and in sialidosis. *Biochem.* **1979,** *18,* 2783-2787.
8. Horvath, C.M.; Paterson, R.G.; Shaughnessy, M.A.; Wood, R.; Lamb, R.A. Biological activity of paramyxovirus fusion proteins: factors influencing formation of syncitia. *J. Virol.* **1992,** *66,* 4564-4569.
9. Moscona, A.; Peluso, R.W. Fusion properties of cells infected with human parainfluenza virus type 3: receptor requirements for viral spread and virus-mediated membrane fusion. *J. Virol.* **1992,** *66,* 6280-6287.
10. Perlman, S.L.; Jordan, M.; Brossmer, R.; Greengard, O.; Moscona, A. The use of a quantitative fusion assay to evaluate HN-receptor interaction for human parainfluenza type 3. *Virol.* **1999,** *265,* 57-65.
11. Bloom, B.R.; Jimenez, L.; Marcus, P.I. A plaque assay for enumerating antigen-sensitive cells in delayed-type hypersensitivity. *J. Exp. Med.* **1970,** *132,* 16-30.
12. Kasahara, T.; Shioiri-Nakano, K.; Sugiura, A. Virus plaque assay: effective detection of virus plaque forming cells at the early stage of lymphocyte activation by mitogen and alloantigen. *Immunology* **1979,** *36,* 381-390.
13. Moscona, A.; Peluso, R.W. Fusion properties of cells persistently infected with human parainfluenza virus type 3: participation of hemagglutinin-neuraminidase in membrane fusion. *J. Virol.* **1991,** *65,* 2773-2777.

14. Porotto, M.; Greengard, O.; Poltoratskaia, N.; Horga, M.-A.; Moscona, A. Human parainfluenza virus type 3 HN-receptor interaction: effect of 4-guanidino-neu5ac2en on a neuraminidase deficient variant. *J. Virol.* **2001,** *75,* 7481-7488.
15. Lawrence, M.C.; Borg, N.A.; Streltsov, V.A.; Pilling, P.A.; Epa, V.C.; Varghese, J.N.; McKimm-Breschkin, J.L.; Colman, P.M. Structure of the hemagglutinin-neuraminidase from human parainfluenza virus type III. *J. Mol. Biol.* **2004,** *335,* 1343-1357.
16. Varghese, J.N.; Colman, P.M. Three-dimensional structure of the neuraminidase of influenza virus A/Tokyo/3/67 at 2.2 Å. *J. Mol. Biol.* **1991,** *221,* 473-486.
17. Kuntz, I.D.; Blaney, J.M.; Oatley, S.J.; Langridge, R.; Ferrin, T.E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982,** *161,* 269-288.
18. Levinthal, C.; Wodak, S.J.; Kahn, P.; Dadivanian, A.K. Hemoglobin interaction in sickle cell fibers I: theoretical approaches to the molecular contacts. *Proc. Nat. Acad. Sci.* **1975,** *72,* 1330-1334.
19. Salemme, F.R. An hypothetical structure for an intermolecular electron transfer complex of cytochromes c and b_5. *J. Mol. Biol.* **1976,** *102,* 563-568.
20. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, T. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996,** *261,* 470-489.
21. Ewing, T.J.A.; Makino, S.; Skillman, A.G.; Kuntz, I.D.; DOCK 4.0: search strategies for automated molecular docking of flexible molecular databases. *J. Comput.-Aided Mol. Des.* **2001,** *15,* 411-428.
22. Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997,** *267,* 727-748.
23. 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.
24. Venkatchalam, C.M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Modelling* **2003,** *21,* 289-307.
25. Jain, A.N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based engine. *J. Med. Chem.* **2003,** *46,* 499-511.
26. Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G.; Langer, T.; Wolber, G. Enhanced drug discovery through in silico screening: strategies to increase true positives retrieval rates. *Curr. Med. Chem.* **2008,** *15,* 2040-2053.
27. 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.

28. Kellogg, G.E.; Burnett, J.C.; Abraham, D.J. Very empirical treatment of solvation and entropy: a force field derived from $LogP_{o/w}$. *J. Comput.-Aid. Mol. Des.* **2001,** *15*, 381-393.
29. Cozzini, P.; Fonabaio, 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.
30. Spyrakis, F.; Amadasi, A.; Fornabaio, M.; Abraham, D.J.; Mozzarelli, A.; Kellogg, G.E. The consequences of scoring docked ligand conformations using free energy correlations. *Eur. J. Med. Chem.* **2007,** *42*, 921-933.
31. 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.
32. 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.
33. Porotto, M.; Fornabaio, M.; Kellogg, G.E.; Moscona, A. A second receptor binding site on human parainfluenza type 3 hemagglutinin-neuraminidase contributes to activation of fusion mechanism. *J. Virol.* **2007,** *81,* 3216-3228.
34. Porotto, M.; Fornabaio, M.; Greengard, O.; Murrell, M.T.; Kellogg, G.E.; Moscona, A. Paramyxovirus receptor-binding molecules: engagement of one site on the hemagglutinin-neuraminidase protein modulates activity at the second site. *J. Virol.* **2006,** *80,* 1204-1213.
35. Palermo, L.M.; Porotto, M.; Yokoyama, C.C.; Palmer, S.G.; Mungall, B.A.; Greengard, O.; Niewiesk, S.; Moscona, M. Human parainfluenza virus infection of the airway epithelium: viral hemagglutinin-neuraminidase regulates fusion protein activation and modulates infectivity. *J. Virol.* **2009,** *83,* 6900-6908.
36. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006,** *11,* 580-594.

CHAPTER 4

SIDECHAIN OPTIMIZATION USING BACKBONE-DEPENDENT ROTAMER LIBRARIES AND HINT

4.1 THE INDUCED FIT THEORY

In today's world of drug discovery/design that depends significantly on understanding intermolecular interactions, the contributions of Emil Fischer and Daniel Koshland are vital. Fischer first introduced the "lock and key mechanism" in the late 19th century [1-2] and was revolutionary in many respects; it was perhaps the most important description of enzymatic activity as a result of precise and specific intermolecular interactions. However, theories are created in order to explain all facts known at the time of their formulation and must be modified intermittently to include explanations for all future discoveries that bring about discrepancies in them. Koshland wrote a review on his modifications of Fischer's "key-lock mechanism" in the late 20th century [3], wherein he paid homage to the author while describing the then current status of his "induced fit theory." While Fischer's key-lock theory described enzyme activity in terms of the analogy of a key fitting into a lock, thereby allowing the enzyme to act on it, Koshland's theory specifically expounded on the importance of structural changes in the protein at the same time. In his own words, the induced fit theory can be stated as "*a) the precise orientation of catalytic groups is required for enzyme action, b) the substrate causes an appreciable change in the three-dimensional relationship of the amino acids*

at the active site and c) the changes in the protein structure caused by the substrate will bring the catalytic groups into the proper alignment, whereas a nonsubstrate will not."

4.2 EMULATING INDUCED FIT IN COMPUTATIONAL ALGORITHMS

The importance of induced-fit is not restricted to enzymatic activity, but also to binding of non-competitive small molecule inhibitors or binding of drugs to non-enzymatic proteins such as GPCRs, as well as between interacting macromolecules. Cozzini *et al.* describe the importance of the induced fit theory in drug discovery in their perspective [4]. They describe multiple theoretical and experimental methods used to explore induced fit in target-drug binding, laying special emphasis on the current status of applications that take target flexibility into consideration during the drug discovery/design procedure. The transition of molecular docking from its earliest incarnations, in which rigid molecules were docked into a rigid receptor, to the current algorithms and strategies employed to address this issue, has been described therein. For example, a popular docking program Autodock [5-7] has the ability to explore sidechain flexibility at the active site [8,9]. GOLD, FlexE, SLIDE and DOCK also use algorithms in order to explore sidechain flexibility in a variety of ways [10-13]. Incorporation of target flexibility in such popular docking programs is indicative of the importance of exploring induced fit in drug design endeavors.

4.2.1 Basic Structure of Algorithms for Emulating Target Flexibility

The degree of induced fit may vary between different protein-ligand interactions and the introduction of even partial flexibility at the protein binding site is a challenging

and computationally expensive task. This is especially true because exploring low energy conformations of the protein is a combinatorial problem even when sidechain flexibility alone is addressed. The overall algorithm for sidechain optimization can be broken down into independent but equally important components: 1) selecting a method for sidechain selection and positioning, 2) choice of a forcefield for evaluating sidechain positions and 3) deciding an approach to solve the combinatorial problem of sorting through the various permutations and combinations of sidechain positions in order to achieve reasonable results.

4.3 DESIGNING OUR OWN ALGORITHM - THE COGS AND WHEELS

In order to create our own sidechain optimization method, in this work we explore rotamer libraries as a source for sidechain coordinates. The choice of rotamer libraries is a conscious one, based on the following reasons: 1) These libraries contain coordinates for sidechain positions covering naturally occurring conformations and are thereby likely to place residues in reasonable positions and 2) having preordained positions for sidechains will probably translate into faster algorithms compared to randomized placement methods (e.g., Monte Carlo methods or molecular dynamics simulations) for achieving sidechain movement.

Several studies of sidechain rotamer distributions have been reported, including those by James and Sielecki [14], Ponder and Richards [15], Janin *et al* [16], Dunbrack and Karplus [17] and Dunbrack and Cohen [18]. These studies vary in scope and width; for example, while James and Sielecki [14] described their findings about rotameric

preferences of residue sidechains from a single protein crystal structure, Ponder and Richards [15] described the development of a complete rotamer library and an algorithm meant to place sidechains on a protein backbone.

4.3.1 Rotamer Libraries

The different amino acid rotamer libraries in existence can be categorized as backbone-dependent and backbone-independent.

4.3.1.1 Backbone-Independent Rotamer Libraries

The most important feature of backbone-independent rotamer libraries is that the rotamer position is not related to the backbone geometry in any way. The Ponder and Richards rotamer library [15] is an example of this category; the authors demonstrated that only 67 sidechain rotamers are adequate to place 15 of the 18 naturally occurring amino acid residues in which sidechain positioning is required (Ala and Gly do not fall under this category). Met, Arg and Lys residues were not addressed in this study due to their inherent flexibility. They used a rotamer library based on only 19 PDB files and a simple van der Waal's term in order to pack these atoms.

Another example of a backbone independent rotamer library was reported by Janin *et al.* [17] who, like Ponder and Richards, used a set of 19 PDB files as their source data for calculations. They described the preferred conformations for several protein sidechains. The fact that most of the χ_1 angles for a variety of amino acid residues coincided with two of three steric energy minima calculated for a blocked Lys

was interesting. A blocked Lys residue is defined as one that is allowed to rotate along the Cα-Cβ bond to translate through χ_1 angles, but all other χ angles are held constant.

It was clear from these studies that sidechain geometry is severely restricted for a given main chain geometry. This fact is exemplified by the fact that more than 60% of sidechains adopted only one or two rotameric positions [17]. Moreover, the number of configurations adopted also depends on the position of the residue relative to the protein surface. Also, the rotamers which are rare for a surface residue are even rarer for an internal residue, implying a strong preference for certain values of the χ_1 angle. Most importantly, steric energy was established as an important factor in sidechain rotamer placement.

4.3.1.2 Backbone-Dependent Rotamer Libraries

The major difference between backbone-dependent and backbone-independent rotamer libraries is the calculation of χ_1 and χ_2 probabilities as a function of Φ and Ψ angles.

Dunbrack and Karplus introduced their sidechain optimization method based on a backbone-dependent rotamer library [18] from a study based of 126 structures from the Brookhaven Protein Database. The library was generated by assessing the probability of χ angle values for a specific range of Φ and Ψ angles, i.e. the Φ and Ψ angles were divided into bins incremented by 20˚ (0˚ to 20˚, 20˚ to 40˚, 40˚ to 80˚, etc.). It was observed that several regions of the Φ, Ψ map were underpopulated due to geometric restrictions on the backbone and the small bin size. For all amino acid

residues except Ala, Pro and Gly, the χ_1 values were binned into the -120˚ to 0˚, 0˚ to 120˚ and the 120˚ to -120˚ bins. Amongst the residues not covered, no χ values exist for Ala and Gly due to the absence of the Cγ atom. The same bin values were used for χ_2 angle calculations for those residues where the χ_2 angle exists, except for Pro, Asn, Asp, Phe, Tyr and Trp. For Pro, the χ_1 angle was binned only into two bins – positive and negative values, corresponding to the two Pro conformations. The χ_2 angles for Phe, Tyr and His were found to be mostly concentrated around the ±90˚ region, which were treated as equivalent by adding 180˚ to any negative χ_2 angles. These residues show C2 symmetry across the Cβ-Cγ bond, which allows such equivalent treatment of these rotamers. Again, Asp and Asn χ_2 and χ_2+180˚values were treated as equivalent due to the symmetrical positions of the γ atoms, but were binned into -90˚ to -30˚, 30˚ to -30˚ and 30˚ to 90˚ divisions. Trp χ_2 angles were binned into 0˚ to 180˚ and -180˚ to 0˚ because its rigid aromatic rings does not allow for more sterically favorable positions. χ_3 and χ_4 angles were also binned for more flexible residues. Table 4.1 shows the χ_1 and χ_2 bin limits described above. The probability of finding χ_1 in each bin was calculated for each Φ and Ψ bin combination. Similarly, the probability of finding χ_2 in each of its bins was calculated, given that χ_1 was in one particular bin. This is usually designated as χ_{1+2}. Such binning and calculation of probabilities is the crux of backbone-dependent rotamer libraries.

The authors observed specific preferences of sidechain conformations not only for α-helices and β-sheets, but also for other regions. Mostly, residues preferred either one sidechain conformation or else two nearly-equally favored conformations.

Dunbrack and Karplus designed a sidechain optimization algorithm using the CHARMM forcefield and probabilities calculated by the above method. Overall, it was observed that using backbone-dependent rotamer libraries enhanced the number of correct sidechain predictions. When compared to an optimization method that utilizes a backbone-independent rotamer library [19] it was found that most of the differences in predictions were due to χ_1 allocation.

Table 4.1*. χ_1 and χ_2 bin limits* are shown. These limits define the bin size during probability calculations, which were made for each 120˚ bin of Φ and Ψ angles.

χ_1 **limits**	χ_2 **limits**	χ_1 **limits**	χ_2 **limits**
Lys, Arg, Met, Gln, Glu, Ile and Leu		***Ser, Thr, Cys, Val, Phe, His and Tyr***	
0˚ to 120˚	0˚ to 120˚	0˚ to 120˚	
0˚ to 120˚	120˚ to -120˚	120˚ to -120˚	
0˚ to 120˚	-120˚ to 0˚	-120˚ to 0˚	
120˚ to -120˚	0˚ to 120˚	***Pro***	
120˚ to -120˚	120˚ to -120˚	0˚ to 60˚	-60˚ to 0˚
120˚ to -120˚	-120˚ to 0˚	-60˚ to 0˚	0˚ to 60˚
-120˚ to 0˚	0˚ to 120˚	***Asp and Asn***	
-120˚ to 0˚	120˚ to -120˚	0˚ to 120˚	-90˚ to -30˚
-120˚ to 0˚	-120˚ to 0˚	0˚ to 120˚	-30˚ to 30˚
Trp		0˚ to 120˚	30˚ to 90˚
0˚ to 120˚	0˚ to 180˚	120˚ to -120˚	-90˚ to -30˚
0˚ to 120˚	-180˚ to 0˚	120˚ to -120˚	-30˚ to 30˚
120˚ to -120˚	0˚ to 180˚	120˚ to -120˚	30˚ to 90˚
120˚ to -120˚	-180˚ to 0˚	-120˚ to 0˚	-90˚ to -30˚
-120˚ to 0˚	0˚ to 180˚	-120˚ to 0˚	-30˚ to 30˚
-120˚ to 0˚	-180˚ to 0˚	-120˚ to 0˚	30˚ to 90˚

It was found that predicted sidechain positions were more accurate for buried residues. The predictions were accurate for hydrophobic residues except Leu and

aromatic residues except Trp. Cys residues were well predicted, perhaps because most of these were present in pairs and formed disulfide bonds, which were specifically identified before other sidechains were optimized. Thr sidechains were placed more accurately compared to Ser, perhaps because there is less room for maneuvering due to the presence of an extra $C\gamma$ atom. Asp and Glu were least well placed by this algorithm because these are mostly on the surface and have a lot more freedom to move. Moreover, contacts between successive units which exist in crystals were completely ignored in the optimization process.

Dunbrack and Cohen introduced an enhanced version of the Dunbrack and Karplus backbone-dependent rotamer library in 1997. They used Bayesian statistics in order to address the likelihood of obtaining one rotamer, given an original probability distribution similar to that of the Dunbrack and Karplus rotamer library. The analysis was simple in terms of assumptions; only one was employed – that the probability of one particular dihedral is dependent only on the value of the previous dihedral. For example, the probability of obtaining a χ_1 value depends on what the Φ and Ψ dihedrals are. Likewise, the probability of obtaining a particular χ_2 value depends only on what the χ_1 value is. The same assumption can be made for χ_3 and χ_4 as well. Using this rotamer library, Bower, Cohen and Dunbrack created a sidechain optimization algorithm - SCWRL [20].

4.3.2 Choice of Scoring Function

Traditionally, sidechain optimization algorithms have employed scoring functions that pack atoms together. This is especially advantageous because the function of these packing methods is quite complementary to the nature of crystal structures – in both cases, atoms are packed as tightly together as possible. However, such packing algorithms usually overlook phenomena such as electrostatic interactions, hydrophobic interactions, H-bond formation and pi-cation interactions.

One obvious improvement in such algorithms would be to explore the effect of force fields that can address these currently neglected interactions. We decided to use our in-house HINT scoring function [21,22]. Our hypothesis is that HINT will be complementary to steric-based potential functions for a number of reasons: (A) HINT employs a van der Waal's interaction function quite similar to the steric-based Lennard-Jones potential of SCWRL-like programs, (B) HINT addresses hydrophobic interactions, which have been shown to depend on surface area contacts [23,24], which is quite similar to traditional contact-based scoring functions and as explained above, (C) HINT takes several different types of interactions such as hydrogen bonding into account, while packing methods do not share this capability.

4.4 THE SCWRL ALGORITHM

Dunbrack's group introduced a popular program called SCWRL [20] (SideChain optimization With Rotamer Libraries) whose main aim was to predict crystal structure sidechain positions, given the backbone coordinates. This program was based on a

backbone-dependent rotamer library [18] and employed a modified Lennard-Jones potential equation for energy calculation purposes. In order to solve the combinatorial problem of parsing through rotamers, each sidechain is initially placed in its most favorable rotamer and clashes are identified as those residues which exceed a cutoff value for the steric potential. “Clusters” of clashing residues are identified as those that clash with each other and are solved by a combinatorial parsing of rotamer permutations and combinations.

4.4.1 SCWRL “Successfully” Identifies “Correct” Sidechain Positions

Dunbrack *et al.* have shown considerable success in identification of sidechain positions, if provided with the coordinates of backbone atoms [20]. The structure of this algorithm is highlighted below.

4.4.1.1 Initial Sidechain Rotamer Placement

The initial step used by SCWRL is very simple: after the backbone atoms have been read in, it places the sidechain atoms in their most probable form which does not clash with the backbone. Backbone clashes are detected by using a modified (linear) van der Waal’s equation, which calculates a minimum value of 0 kcal/mol for no steric interaction or 10,000 kcal/mol for steric clashes. Favorable van der Waal’s interactions were sacrificed to improve search speed.

4.4.1.2 "Cluster" Parsing Method

"Clusters" of clashing residues are identified and solved by a combinatorial parsing of rotamer permutations and combinations, intermittently employing a cluster-dividing technique if the number of residues within a "cluster" exceeds 15 residues. In a case where the number of residues does exceed this number, the residue whose sidechain has most clashes is identified and placed in its most probable position which does not clash with the backbone, thereby dividing the large cluster into smaller clusters that are more easily manageable. Each such division of the large cluster is then treated as a smaller cluster and a combinatorial search for sterically favorable sidechain rotamers is conducted. A favorable combination of rotamers is defined as one which has zero steric energy, as decided by the modified Lennard Jones potential equation. However, if no such combination is found, the set of rotamers showing lowest steric clashes is selected.

4.4.1.3 Criteria for "Success"

The authors of the SCWRL program compare χ_1 angles directly with the original PDB files in order to measure the success of their algorithm. However, their metric for comparison is whether the program correctly identifies this dihedral angle within ±40˚ of the actual χ_1 value from the original crystal structure. However, seeing that it uses backbone-dependent rotamer libraries, which are essentially programmed with the probabilities of χ angles existing given the Φ and Ψ angles of the backbone, it is hardly surprising that these angles were successfully predicted. Moreover, the χ_2 angles are

only compared for those cases in which χ_1 angles are "correctly" predicted (χ_{1+2} predictions).

Overall it was shown that the χ_1 angle was correctly predicted 77% of the time, while χ_{1+2} predictions were correct on 66% occasions. The percent correct predictions varied between different types of amino acid residues. For aromatic residues such as Phe and Tyr, the χ_1 determination was 90% accurate. SCWRL also accurately predicted the χ_1 values for residues with β- and γ- branched sidechains such as Val, Thr, Ile, Leu, His and Trp (>80% correct). SCWRL performed less well with Ser and those residues with long unbranched sidechains such as Met, Glu, Gln, Arg and Lys (60 to 72% correct). The χ_1 values for both Asp and Asn were predicted correctly over 73% of the time. An evaluation of χ_2 prediction showed that the degree of accuracy was lower than that for χ_1. However, this is expected because SCWRL is a method that depends on packing. As the distance of a sidechain atom from the backbone increases, the ability of the atom to move also increases due to lower steric interactions with the backbone. It can hardly be expected that the atoms responsible for the χ_2 dihedral angle will be packed as tightly as the Cγ atom unless these are completely buried within the bulk of the protein.

4.4.2 Can the HINT Scoring Function Complement the SCWRL Rotamer Library?

The ability of SCWRL to predict correct sidechain positions in crystal structures perhaps can, at least partially, be attributed to the inclusion of rotamer probabilities. When conducting molecular modeling or drug discovery studies, it becomes more

important to study proteins in their native state, e.g., the dominant state of a cytosolic protein in an aqueous environment. This cannot be achieved by using SCWRL's modified Lennard-Jones potential scoring function, because it completely ignores intra-protein interactions other than sterics. In accordance with our long term goal of creating an algorithm that can simulate sidechain flexibility in docked structures, we decided to test the compatibility between Dunbrack's backbone-dependent rotamer library [18] and the HINT scoring function, which should theoretically be able to address those interactions within the protein that are completely neglected by SCWRL. However, it must be noted here that the newest version of SCWRL (ver. 4.0 [25]) is capable of H-bond detection. The results of our pilot study aimed to analyze whether or not the SCWRL 1.0 rotamer library is compatible with the HINT scoring function are reported herein.

One way of approaching this question is to try to emulate SCWRL's original function of sidechain prediction for known crystal structures. If the HINT scoring function is at least as successful as SCWRL, or even if it comes close to doing so, we can perhaps safely conclude that HINT may be used to complement the SCWRL backbone-dependent rotamer library, or, in other words, that these two technologies are synergistic in molecular model-building applications.

4.5 THE HINTASCWRL ALGORITHM

The HINTaSCWRL (HINT assisted SCWRL/Hydropathic INTeractions Assisted SideChain optimization With Rotamer Libraries) Algorithm was designed using a

backbone-dependent rotamer library and a hydropathic forcefield, whose pseudocode is described below:

```
Read input PDB
Select protein from PDB
Calculate Φ and Ψ angles
Get χ values from sidechain data
Build multiple sidechain conformations for all residues
Check all rotamers for backbone clashes
Find best rotamer
{
        for all residues
        {
                Add Hydrogens
                For each rotamer with no backbone clashes
                {
                        Calculate HINTaSCWRL score
                        Keep rotamer with highest HINTaSCWRL score
                }
        }
}
Perform second round of optimization
Delete any Hydrogens present
Write output PDB
```

The program was written in C, using the HINT toolkit and the backbone-dependent rotamer library from SCWRL 1.0 (referred as SCWRL hereon). The algorithms are still in a very preliminary state and come with no additional features aimed at improving prediction capabilities; no mathematical strategies have been adopted in order to solve combinatorial issues, nor has any special modification been made in order to obtain optimum predictions. Indeed, this is the very first form of our protocol, which will need many cycles of refinement in order to improve its functioning.

4.5.1 The Backbone-Dependent Rotamer Library

HINTaSCWRL employs the same backbone-dependent rotamer library as SCWRL. The selection of the Cohen-Dunbrak backbone-dependent rotamer library was chosen in order to enable direct comparison with SCWRL generated sidechain positions.

4.5.2 The HINTaSCWRL Scoring Function

The HINTaSCWRL scoring function that was employed in this program can mathematically be denoted as follows:

$$Score = HINT\ score.\frac{LogP_{current\ rotamer}}{LogP_{most\ probable\ rotamer}}$$

The Log values of probabilities were taken because that linearizes the data. However, this caused a problem because probability values were all fractional and yielded negative values when converted to Log form, thereby reversing the sign (favorable/unfavorable) of the final score that should be supplied by HINT. This problem

is easily solved by normalizing against log(probability) of the most probable rotamer. This also allows us to measure the comparative likelihood of obtaining the current rotamer against the most probable rotamer.

A similar approach where the probability of the current rotamer was multiplied with its HINT score was also considered, but was not found to be as useful as the above scoring function because a few rotamers were frequently found to be several times more probable than all other rotamers combined. This caused probability to dominate the decision making process, thus making the HINT score virtually redundant (except in the case of a sign change facilitated by it). Additionally, it effectively reduced the conformational space explored because the lowest probability conformations (with probabilities ranging up to the 10^{-6} region) were virtually never considered.

4.5.3 Sorting Through Clashes and Bad Interactions

After the initial placement of sidechains, the residues with the worst scores were identified and optimized a second time. While the initial placement of sidechains took probability into account, this time only the HINT score was used.

4.6 THE TEST SET

A data set of 129 PDB files containing no ligands and with resolutions between 1-1.5 Å were downloaded from the RSCB Protein Data Bank and prepared by removal of all alternative conformations for sidechains. The PDB list is as follows:

2BCM, 2BN3, 2BOG, 2BZV, 2CG7, 2CIT, 2CL2, 2CYG, 2DF6, 2DPL, 2E0Q, 2E10, 2E3H, 2E3Z, 2ERF, 2ERW, 2FHZ, 2FQ3, 2FR2, 2FRG, 2FWG, 2G69, 2G7O, 2GBJ, 2GEC, 2GKG, 2GKT, 2GOM, 2GQV, 2GRC, 2GUV, 2GXG, 2GZV, 2H3L, 2H8E, 2HLR, 2I3F, 2IBL, 2IC6, 2ICC, 2IGD, 2IPR, 2IVY, 2IWN, 2IXM, 2J6B, 2J73, 2J8B, 2JCP, 2JIC, 2JLJ, 2LIS, 2NRR, 2NWD, 2O37, 2OCT, 2OEI, 2OHW, 2OKT, 2OLX, 2OVA, 2OZF, 2P4H, 2PMR, 2PND, 2PPO, 2PV2, 2QHT, 2QOL, 2QT4, 2QVK, 2R6Q, 2RB8, 2RK3, 2RK5, 2VC8, 2VIM, 2VY8, 2W1R, 2W2A, 2W6A, 2WJ5, 2WLV, 2YXF, 2YZ1, 2ZO6, 3A7L, 3BA1, 3BB7, 3BOI, 3BPV, 3BQS, 3BZT, 3BZZ, 3C8P, 3CA7, 3CJW, 3CKF, 3CT5, 3CTG, 3CX2, 3CZZ, 3D9X, 3DFG, 3DS4, 3DWV, 3EVP, 3EXV, 3EY6, 3EYE, 3FKE, 3FPO, 3FTD, 3FTK, 3FVA, 3HFO, 3HNX, 3HNY, 3HZ8, 3I4O, 3IVV, 3KB5, 3KGK, 3KJT, 3KTP, 3L32, 3L3E, 4EUG and 4PTI.

The algorithm was not trained on any of these structures, which contain up to 919 residues.

4.7 HINTASCWRL OUTPUT ANALYSIS

Since the main aim of this project was to test the ability to use the HINT forcefield with rotamer libraries and not necessarily to improve sidechain placement algorithms such as SCWRL, the most important aim of the HINTaSCWRL algorithm is to predict structures approximately as well as SCWRL. If this were true, it would demonstrate the ability to use HINT in conjunction with rotamer libraries. On the other hand, it would definitely be advantageous if we could improve sidechain predictions. With these aims

in mind, the output of HINTaSCWRL will be compared with the same from SCWRL through the rest of this chapter.

4.7.1 Analysis of Sidechain RMSD

The RMSD for each sidechain of the HINTaSCWRL and SCWRL output PDBs from the original PDB were calculated. Figure 4.1 shows line plots of RMSD for all residue sidechains across all 177 test structures. The first column consists of RMSD for each amino acid residue for HINTaSCWRL output PDB files, while the second column shows the same for SCWRL output files. The third column shows an RMSD difference between HINTaSCWRL and SCWRL output files. It must be noted at this point that these residues have been sorted in order of their solvent accessible surface area (SASA) before the plots were generated.

HINTaSCWRL results – Cys, Pro, Ser, Thr and Val residues showed the lowest deviation (up to 1.5 Å) from the original PDB, while Asn, Asp, Gln, Glu, Ile, Leu and Met showed a higher RMSD value (up to 2.5 Å) in comparison. Arg, His, Lys, Trp and Tyr showed the highest deviations (up to 5 Å) from the original crystal structures.

Comparison with SCWRL output files – All amino acid residues showed similar deviations from the original PDB, as is witnessed by the similarity between the first two columns (Figure 4.1). The third column shows the deviation of HINTaSCWRL output structures from SCWRL predicted ones. Most of the residues show very low to no deviation, which means that the output structures are extremely similar to each other.

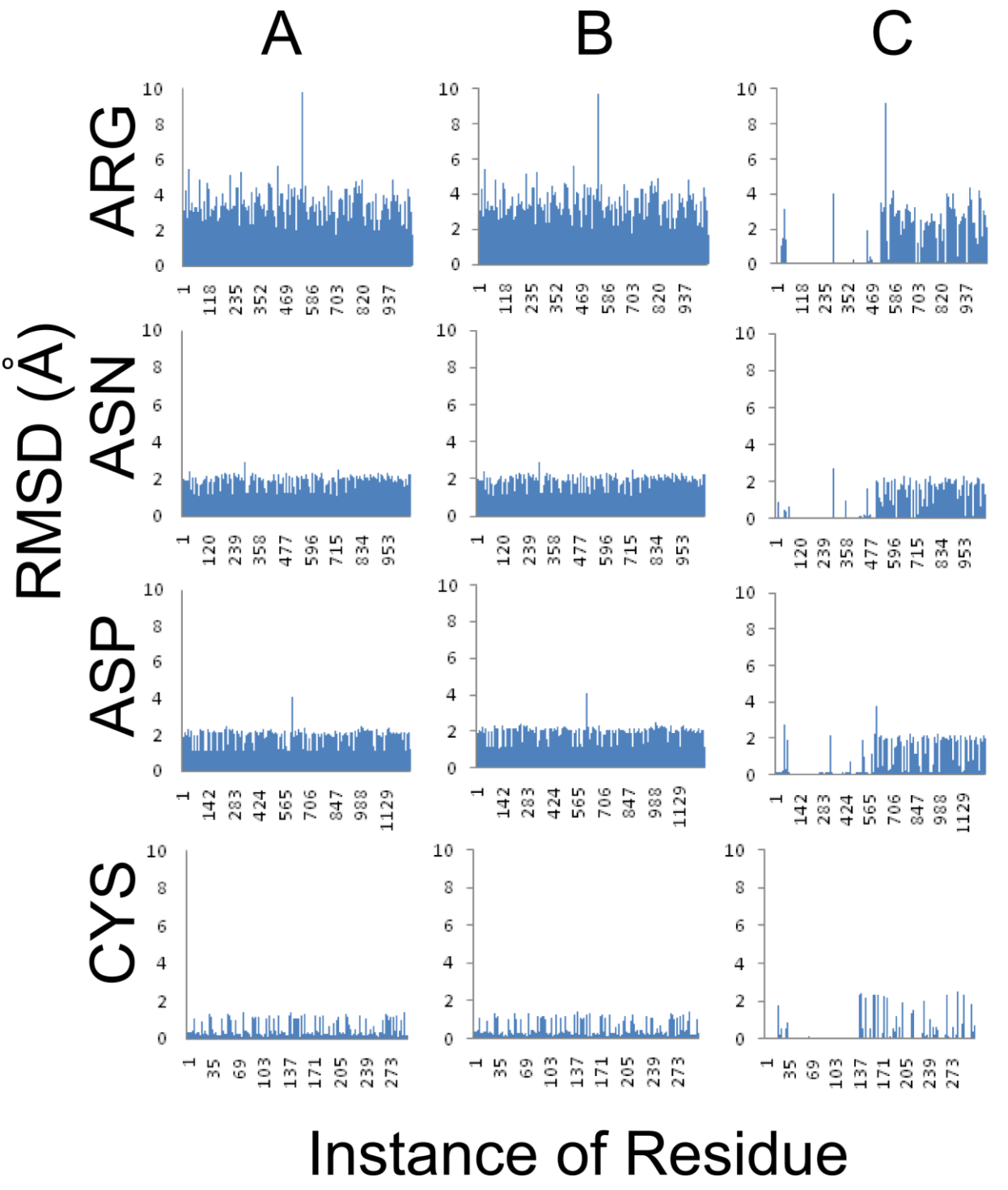


Figure 4.1 ***RMSD values for individual amino acid residue sidechains.*** (A) RMSD for HINTaSCWRL output PDB files (B) RMSD for SCWRL output. (C) RMSD between HINTaSCWRL and SCWRL output files.

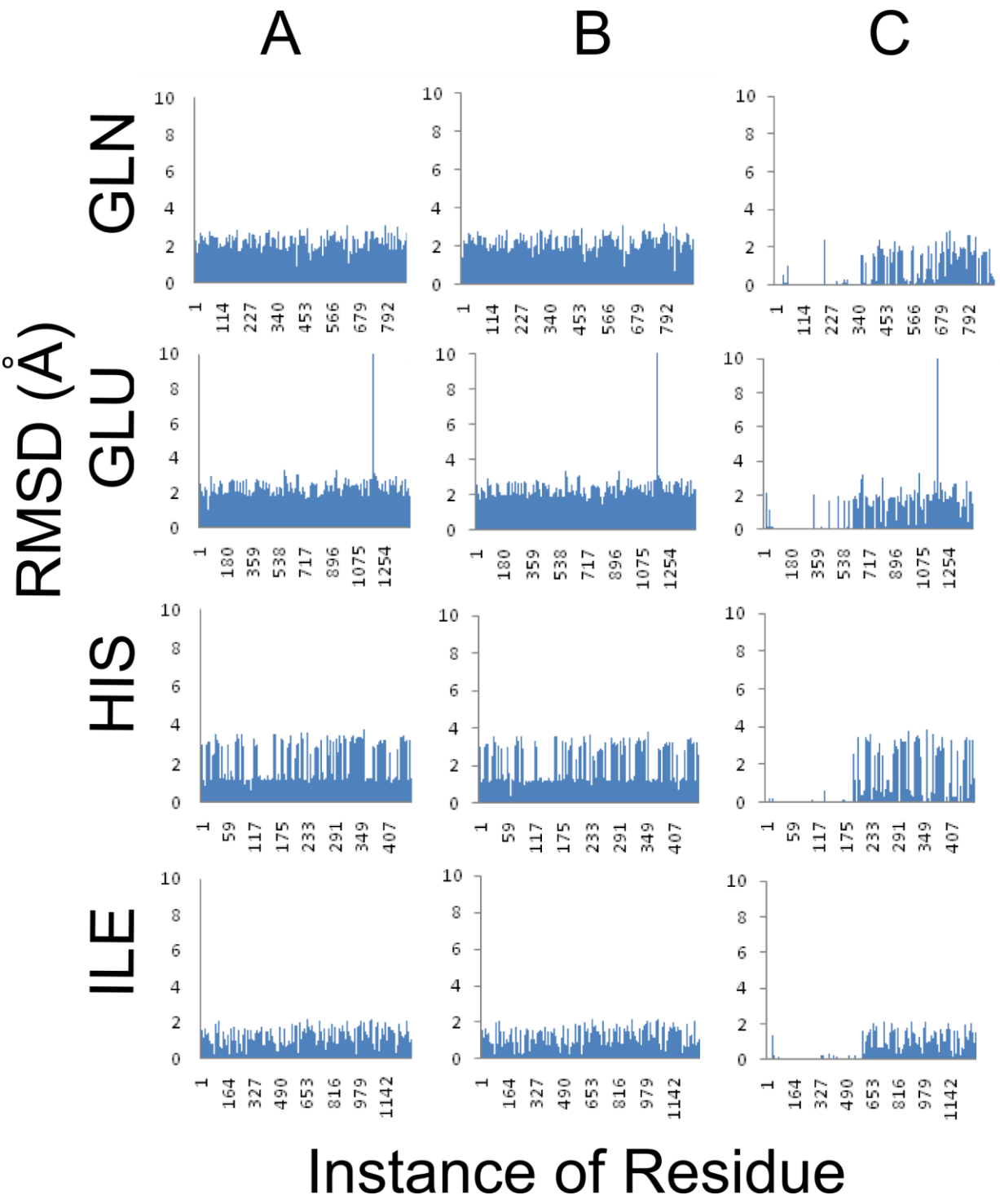


Figure 4.1 ***continued.***

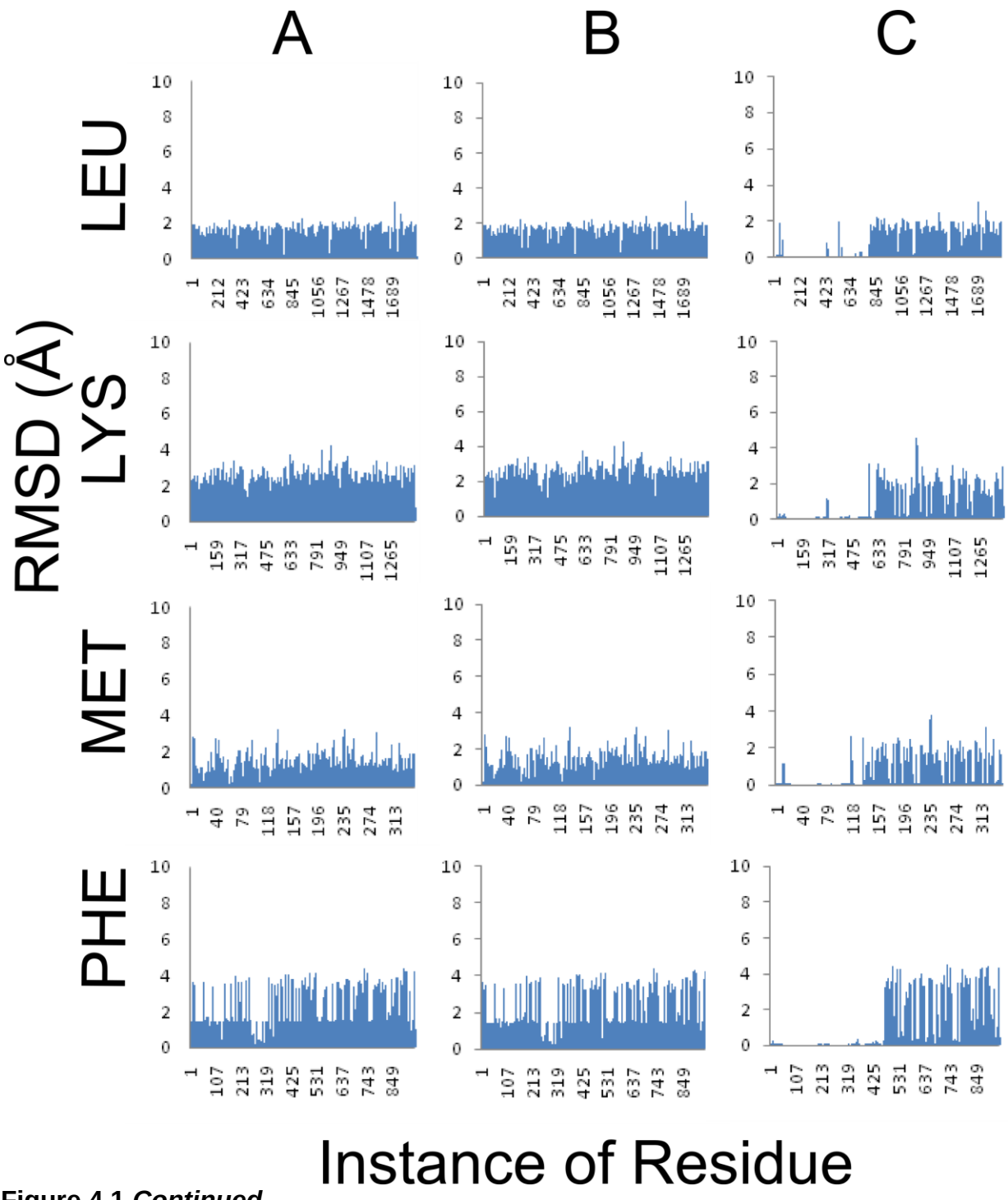


Figure 4.1 ***Continued***

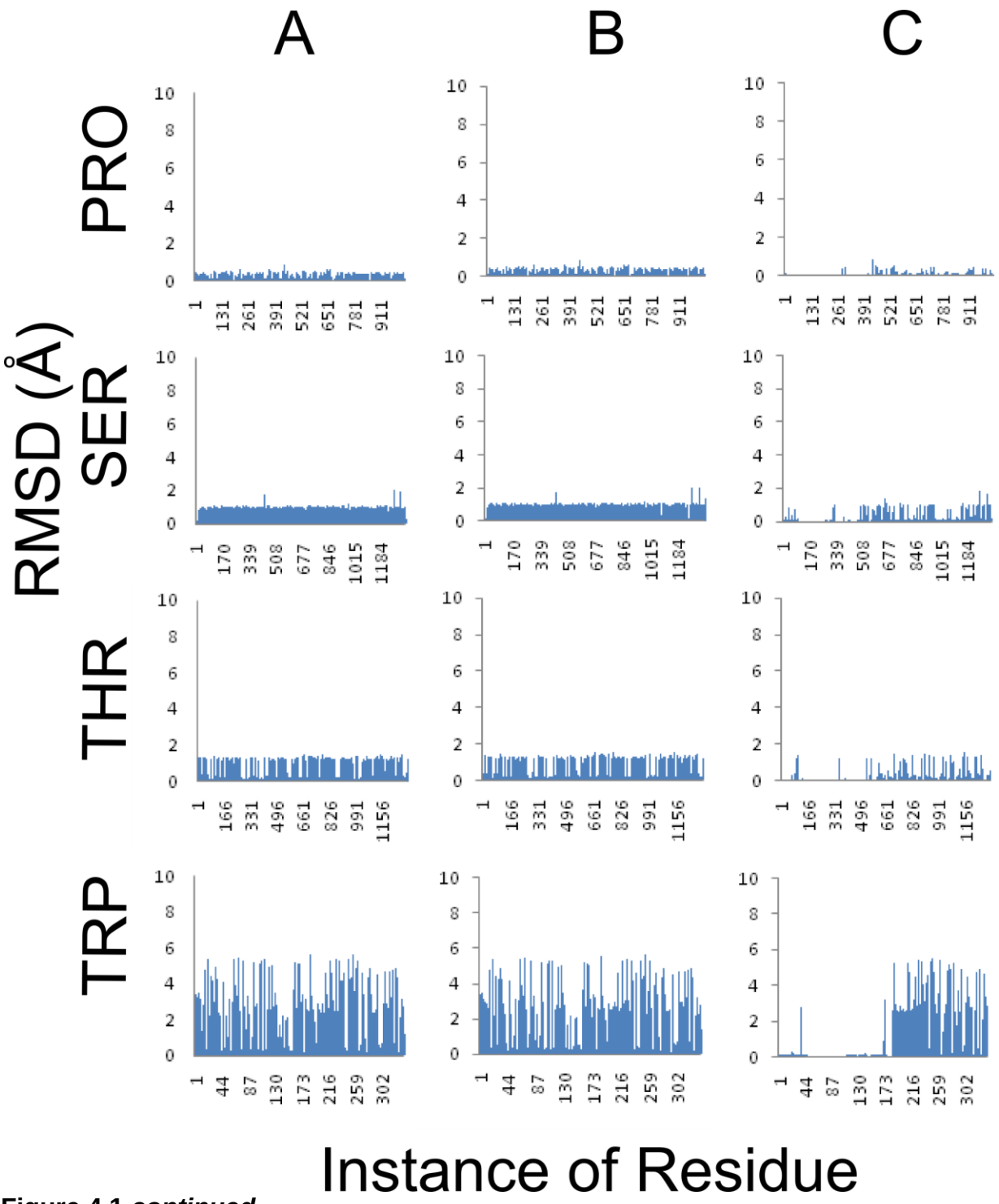


Figure 4.1 ***continued***

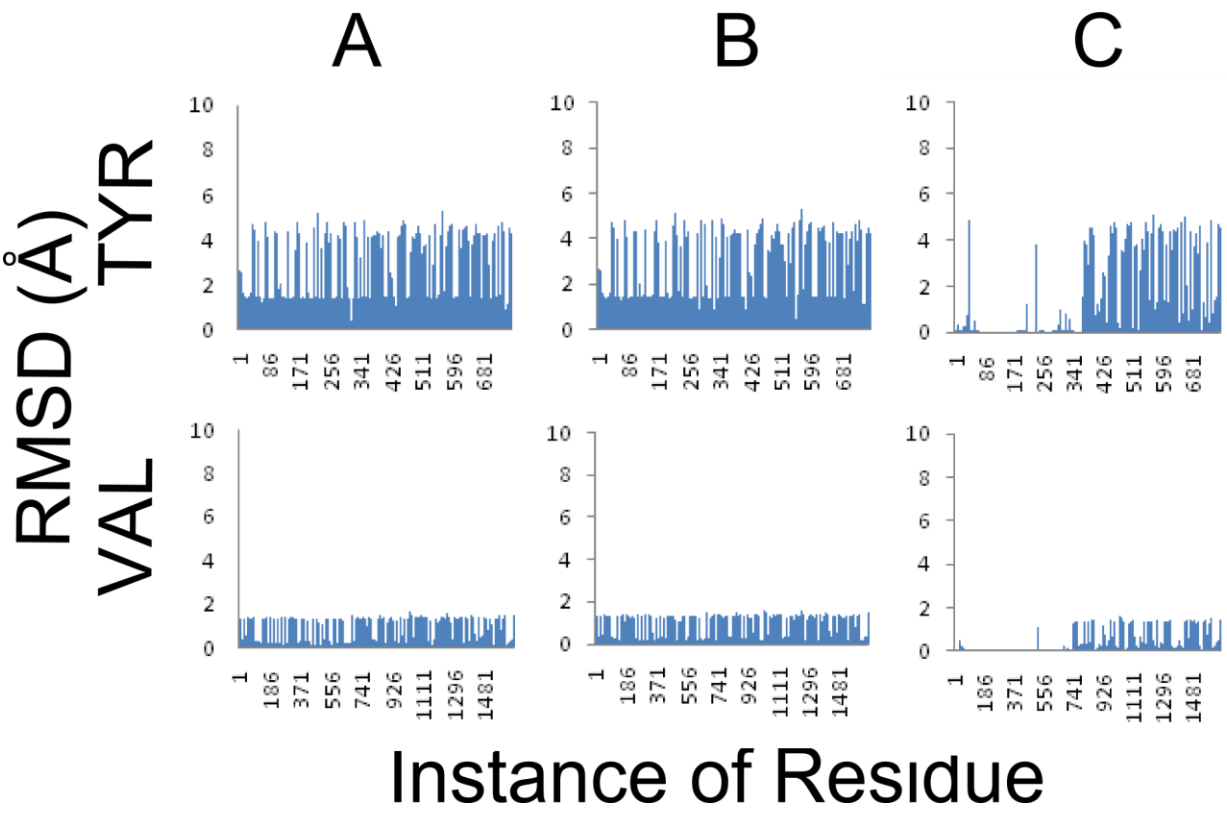


Figure 4.1 *continued*.

4.7.2 RMSD as a Function of Solvent Accessible Surface Area

Figure 4.2 shows RMSD of amino acid residue types as a function of Log(SASA) in HINTaSCWRL predicted structures. A prominent increase in the RMSD of residues is observed with increase in SASA. Therefore, residues closer to the outer surface of a protein have higher solvent accessibility. Such residues which are closer to the outer surface have a higher degree of steric freedom, resulting in poorer predictions.

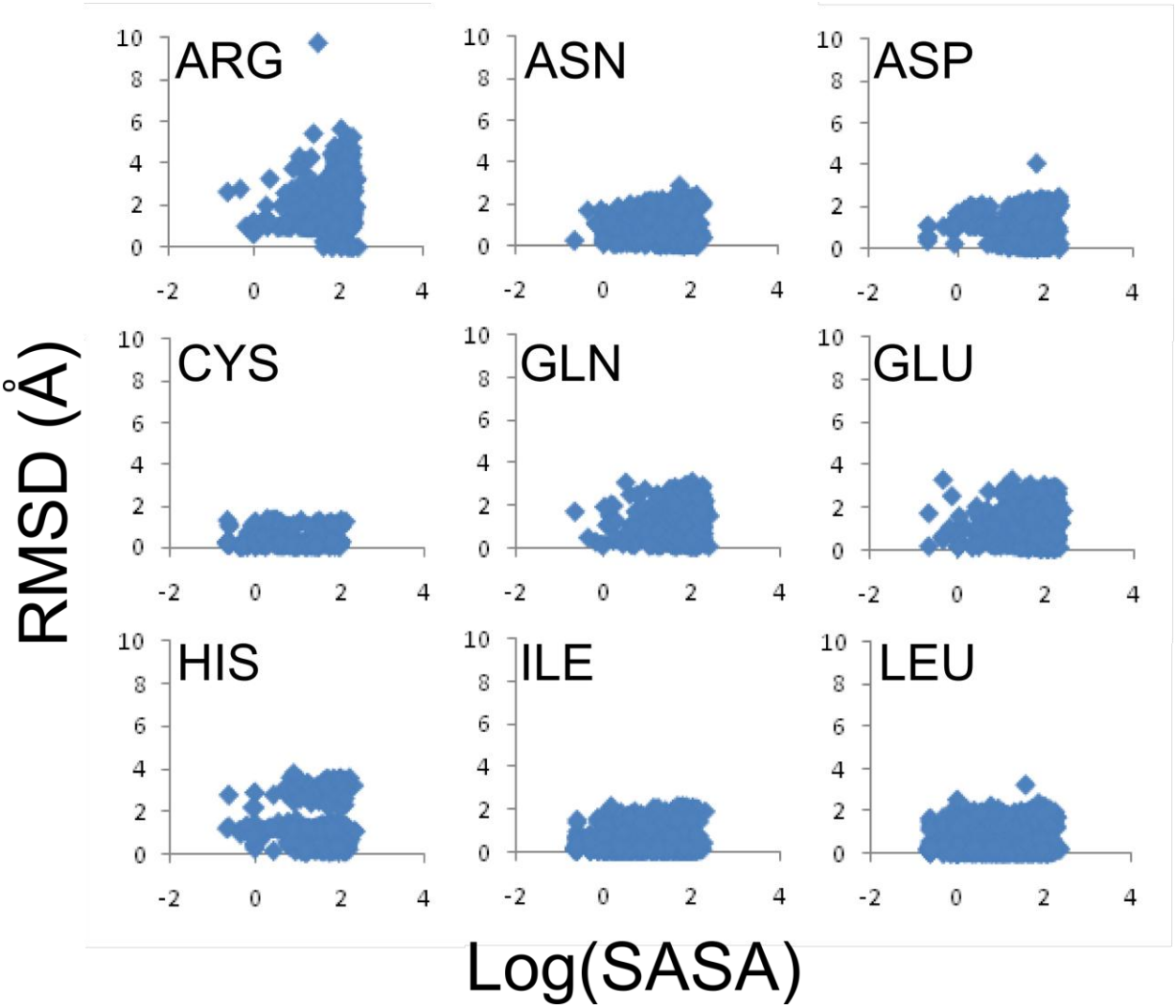


Figure 4.2 ***RMSD as a function of Log(SASA) in HINTaSCWRL output files.***

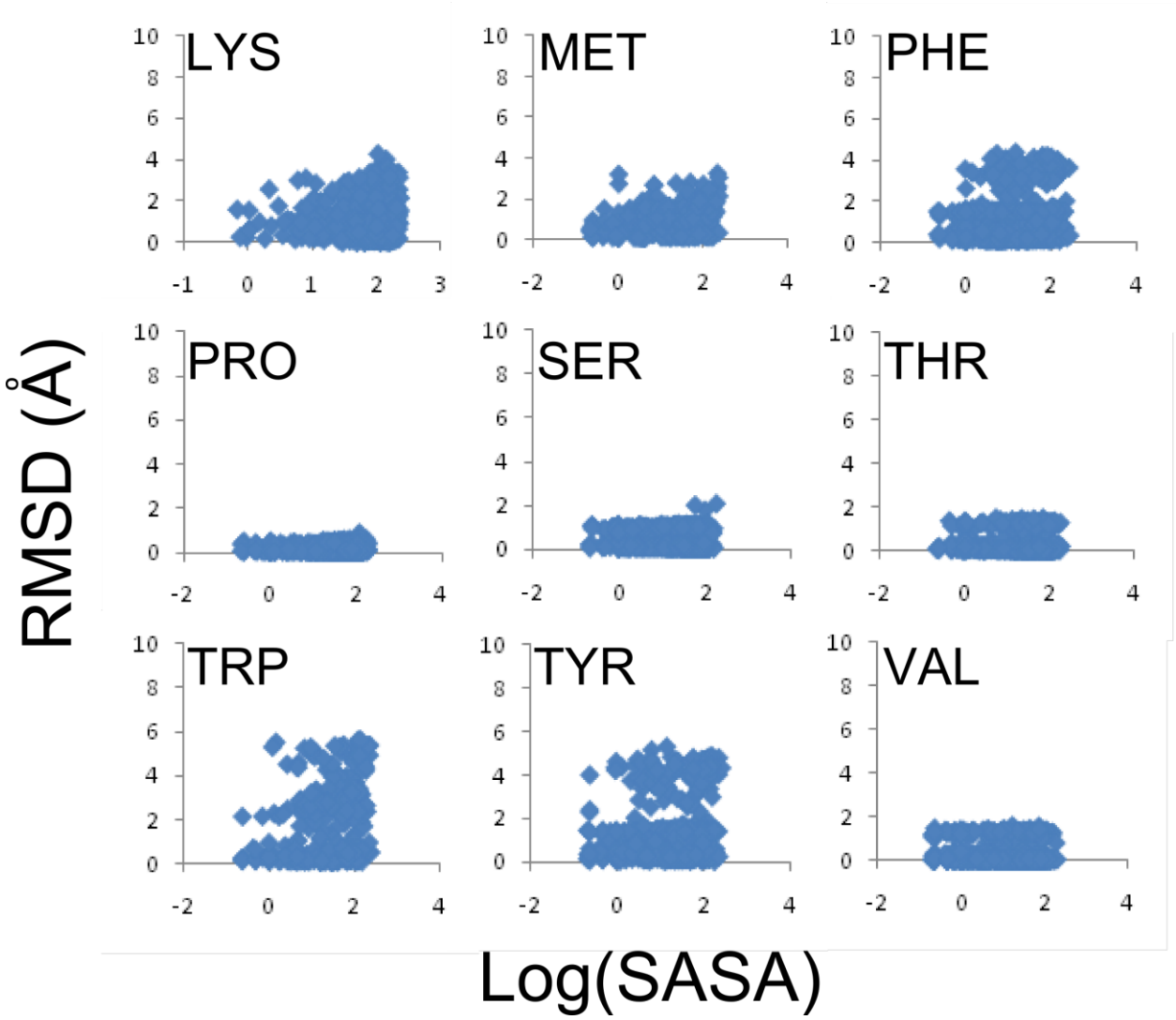


Figure 4.2 ***continued***

Poor predictability is expected in these cases; solvent accessible residues would have a higher degree of movement in their native environment. Hence, the inability of HINTaSCWRL to predict exact positions for these residues does not raise any concerns. However, Phe, Tyr and Trp show a regular trend of poor predictions irrespective of the SASA of any particular residue, but the increased incidence of poor predictions with increase in SASA is clearly visible in these residues.

The trend of increased RMSD compared with original PDBs is clearly visible in figure 4.3, where RMSD for all residues has been plotted against their SASA for both HINTaSCWRL and SCWRL predicted structures.

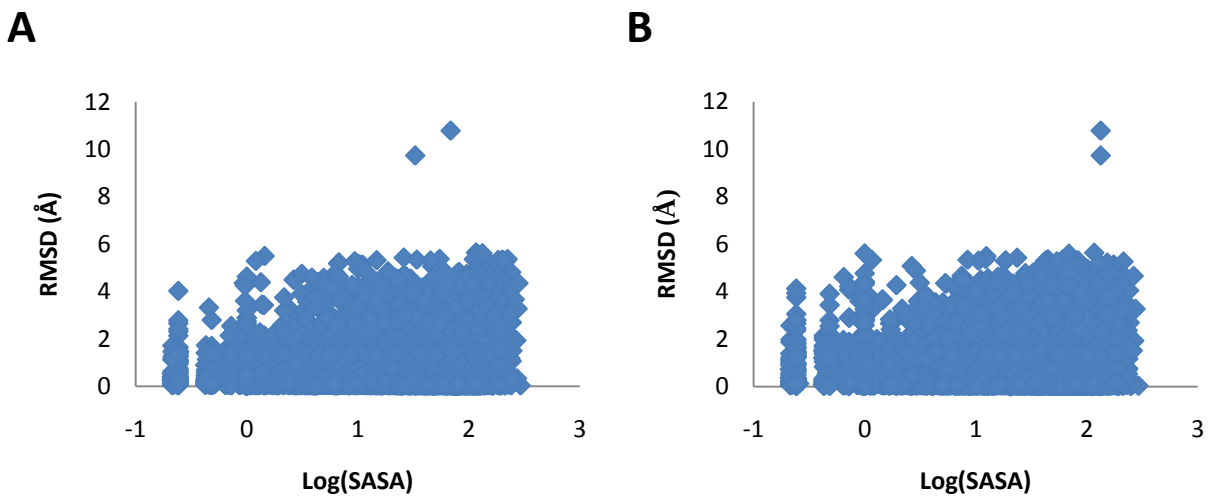


Figure 4.3 ***Overall RMSD across all residues as a function of Log(SASA).*** (A) is for the HINTaSCWRL output structures, while (B) shows the same data for SCWRL predicted structures.

4.7.3 Analysis of Average RMSD per Residue Type

The plot of average RMSD for each residue type in each structure is shown in figure 4.4. The red lines depict average RMSD for each residue type in HINTaSCWRL generated output structures, while the same for SCWRL output files are illustrated with a blue line. While smaller residues such as Cys, Ile, Leu, Ser, Thr and Val, along with sterically restricted residues like Pro, demonstrate lower average RMSD values, others such as Arg show a much higher value.

The higher average RMSD value for hydrophilic residues such as Arg, Lys, His, Asp, Glu, Asn and Gln is expected because they are usually found on the solvated surfaces of proteins. However, aromatic residues Phe, Tyr and Trp show a range of average RMSD values; from high to low. This could possibly be because of their equivocal distribution; they exist on the surface of proteins and also in their bulk. This would explain why these residues demonstrate a wide range of average RMSD values.

It is remarkable that all types of amino acid residues show very similar average RMSD value for both HINTaSCWRL as well as SCWRL.

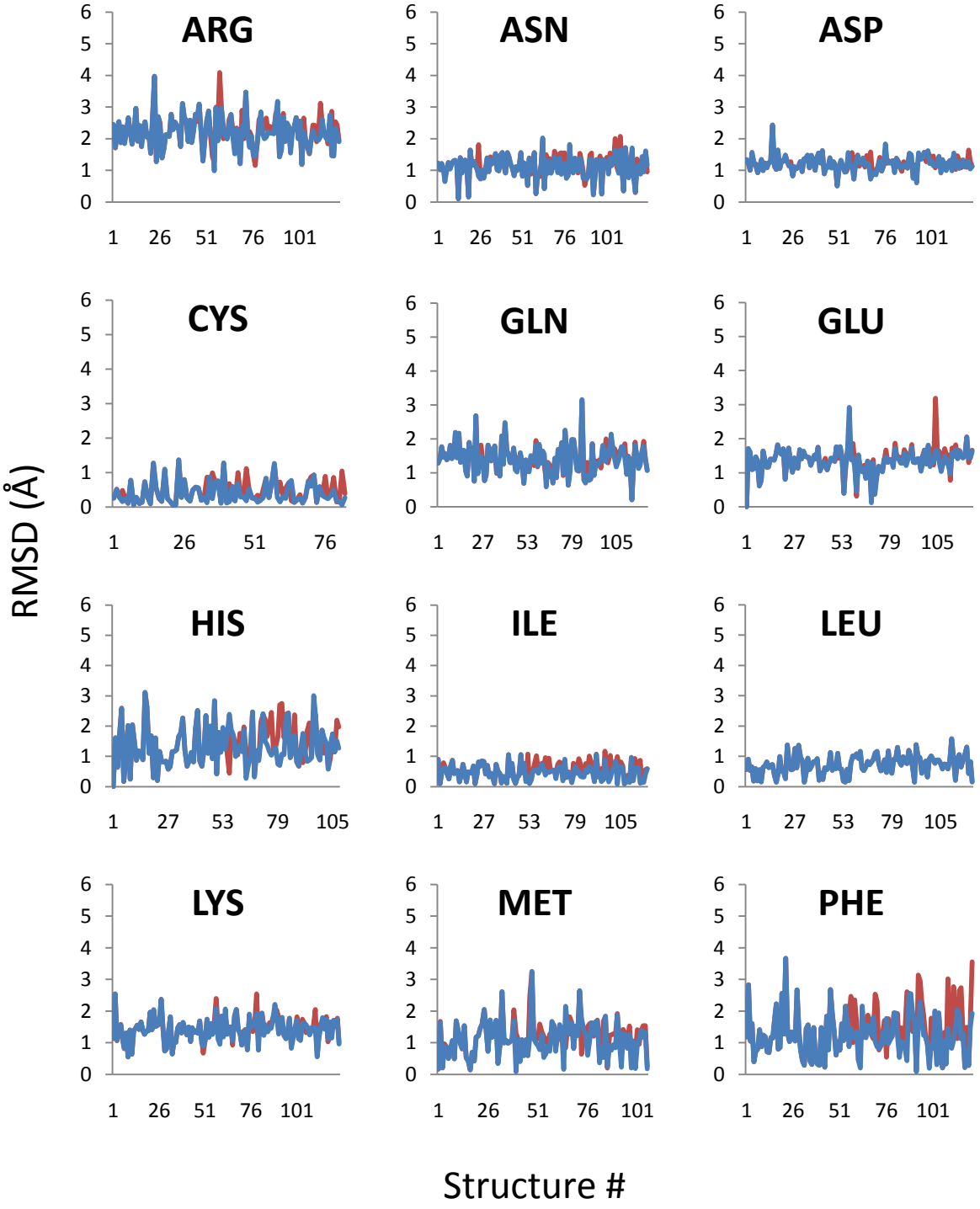


Figure 4.4 ***Average RMSD for each type of amino acid residue.*** HINTaSCWRL output is depicted by the red line, while SCWRL output is shown in blue.

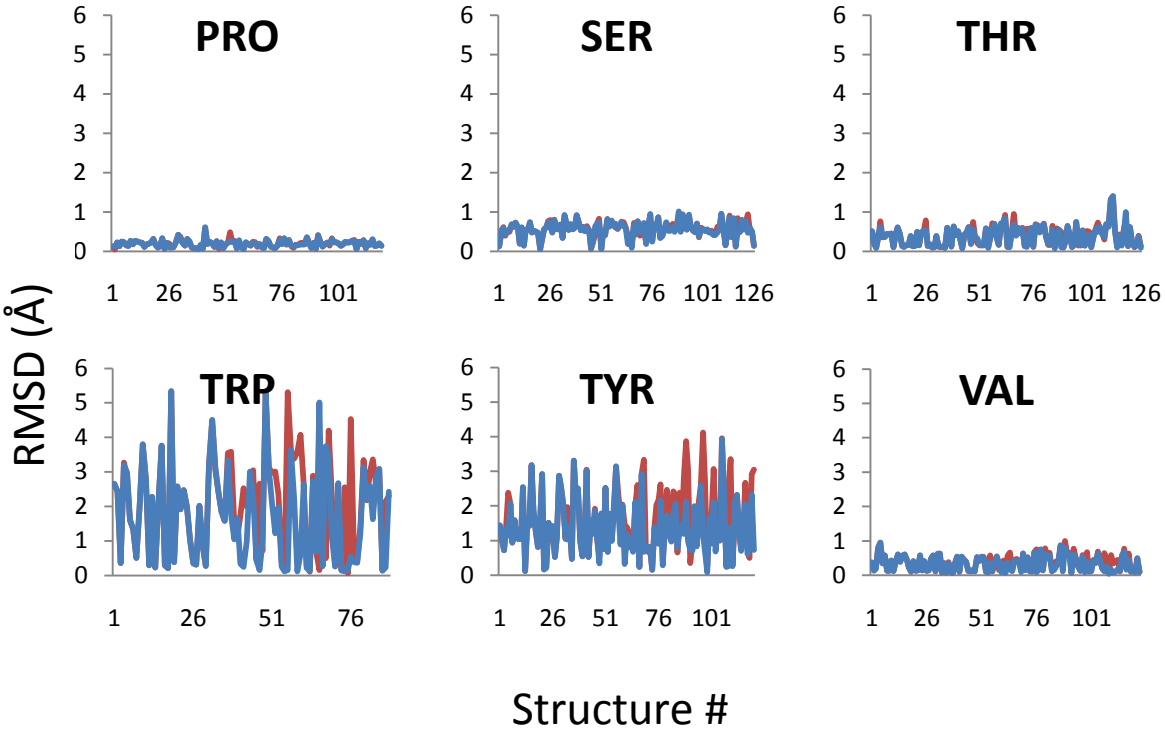


Figure 4.4 ***continued***

4.8 SELECTED HINTASCWRL OUTPUT STRUCTURES

4.8.1 Specific Case Studies

Considering the aims of this project, it is of immense interest to compare the output structures of HINTaSCWRL and SCWRL with the original PDB and then to compare both output structures to each other. This will tell us how successful either program is in predicting sidechain positions. More importantly, since we are mainly interested in assessing the compatibility between Dunbrack's rotamer library and the HINT scoring function, we would ideally like to see similar predictions by both algorithms. It would be even better if HINTaSCWRL is able to predict structures that are closer to the original PDB.

Thus, the minimum expectations from the structures predicted by our algorithm in order to claim success are: (A) the structures should be very close to the SCWRL predicted structures, unless they are closer to the original PDB and (B) most of the deviations should be localized near the surface of the structures, where residues enjoy a greater steric freedom that allows greater movement. Three randomly chosen protein structure predictions are presented with these aims in mind.

Only a few residues show a high RMSD in both HINTaSCWRL and SCWRL programs, designated in red, while most residues have a low RMSD. For 2CYG, the predicted residue positions are exactly the same for both programs, as visible by the entirely green color in figure 4.5 (C). In retrospect, the output created by both programs

were not always exactly the same and varied to a certain degree, as is visible from figure 4.5 (F) and (I). However, the similarity between the two programs is striking despite a small degree of differences.

The degree of structural deviation is described by the plots of RMSD for each structure in figure 4.6. While parts (A), (C) and (E) show plots of RMSD for each residue when compared with the original crystal structure, (B), (D) and (F) show the how much the two output structures deviate from each other.

Both algorithms show similar RMSD profiles when compared to the original PDB, as witnessed in figure 4.6 (A), (C) and (E). In two of the three test cases showed here, the structures produced by HINTaSCWRL and SCWRL are very similar to each other. However, these two algorithms predicted vastly different structures for 4EUG, as shown by figure 4.6 (F). A visual inspection of the structure of 4EUG demonstrated that most of the RMSD between the output structures predicted by both algorithms was due to residues on the surface of the protein, shown in figure 4.5 (I) and hence is quite acceptable.

While there was no difference between the HINTaSCWRL and SCWRL output structures for 2CYG, there were differences between the same for 2VC8 and 4EUG. The residues with the largest RMSD values for both 2VC8 and 4EUG were identified: Asn13 for 2VC8 and Lys171, which were both found to be on the surface. A close inspection of these two residues showed differences in interactions. While Lys171 for the HINTaSCWRL output structure showed formation of an extra H-bond, as observed

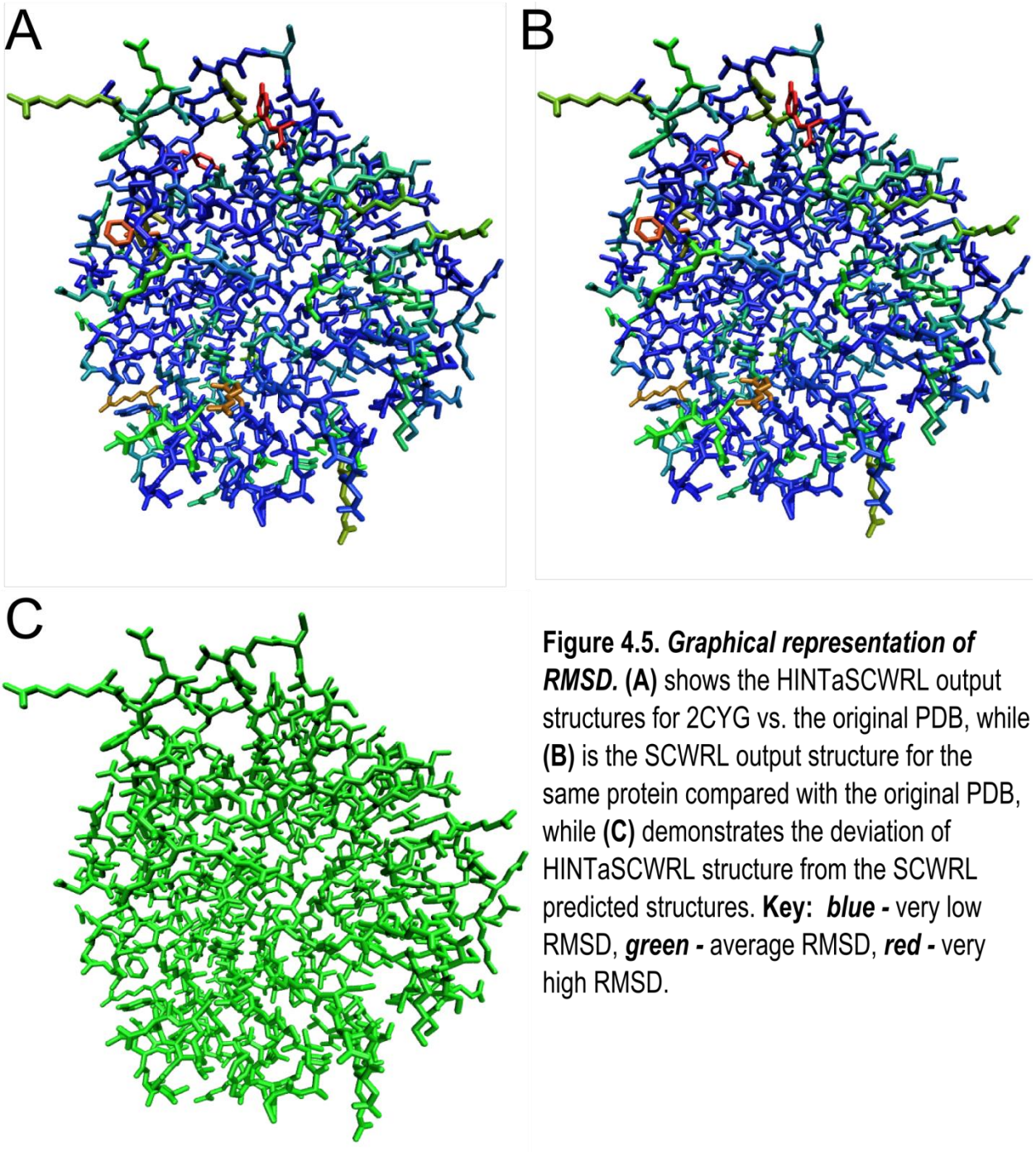
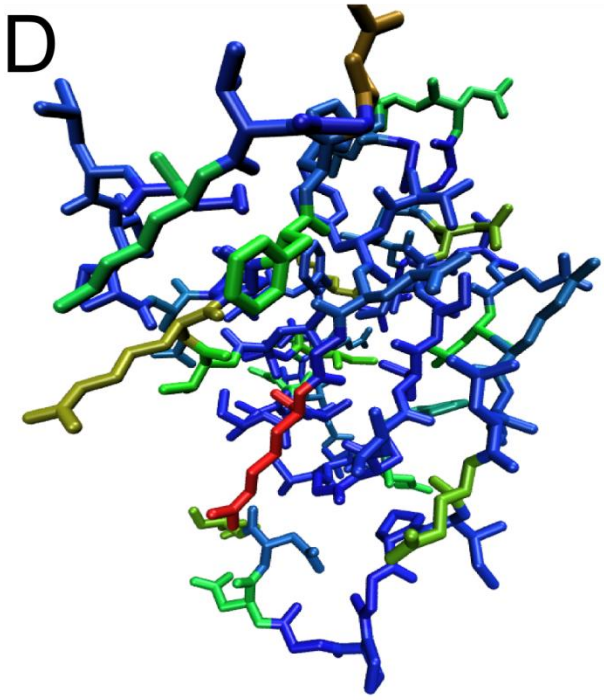
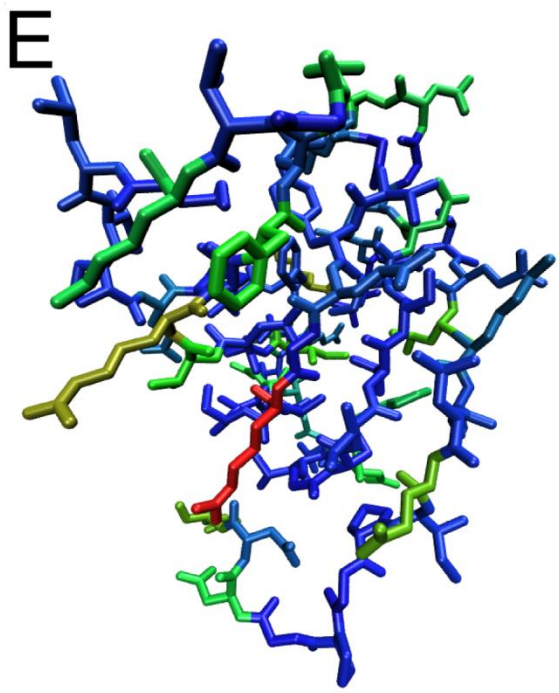


Figure 4.5. ***Graphical representation of RMSD.*** **(A)** shows the HINTaSCWRL output structures for 2CYG vs. the original PDB, while **(B)** is the SCWRL output structure for the same protein compared with the original PDB, while **(C)** demonstrates the deviation of HINTaSCWRL structure from the SCWRL predicted structures. **Key:** ***blue -*** very low RMSD, ***green -*** average RMSD, ***red -*** very high RMSD.





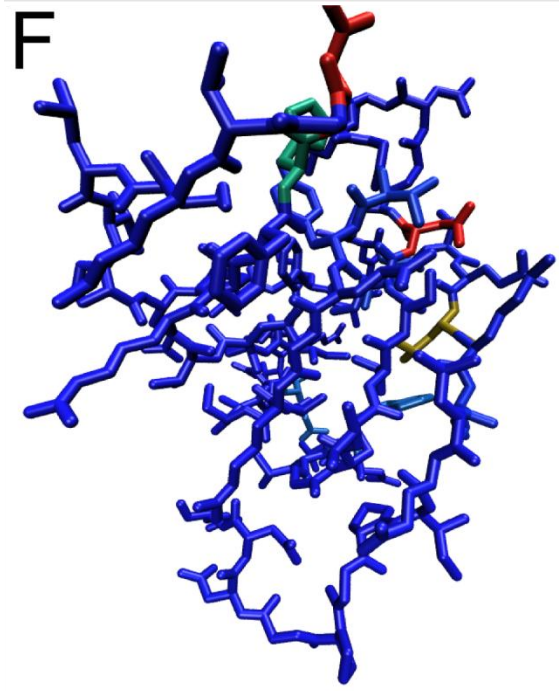


Figure 4.5. ***Graphical representation of RMSD continued..*** **(D)** shows the HINTaSCWRL output structures for 2CV8 vs. the original PDB, while **(E)** is the SCWRL output structure for the same protein compared with the original PDB, while **(F)** demonstrates the deviation of HINTaSCWRL structure from the SCWRL predicted structures. **Key:** ***blue -*** very low RMSD, ***green -*** average RMSD, ***red -*** very high RMSD.

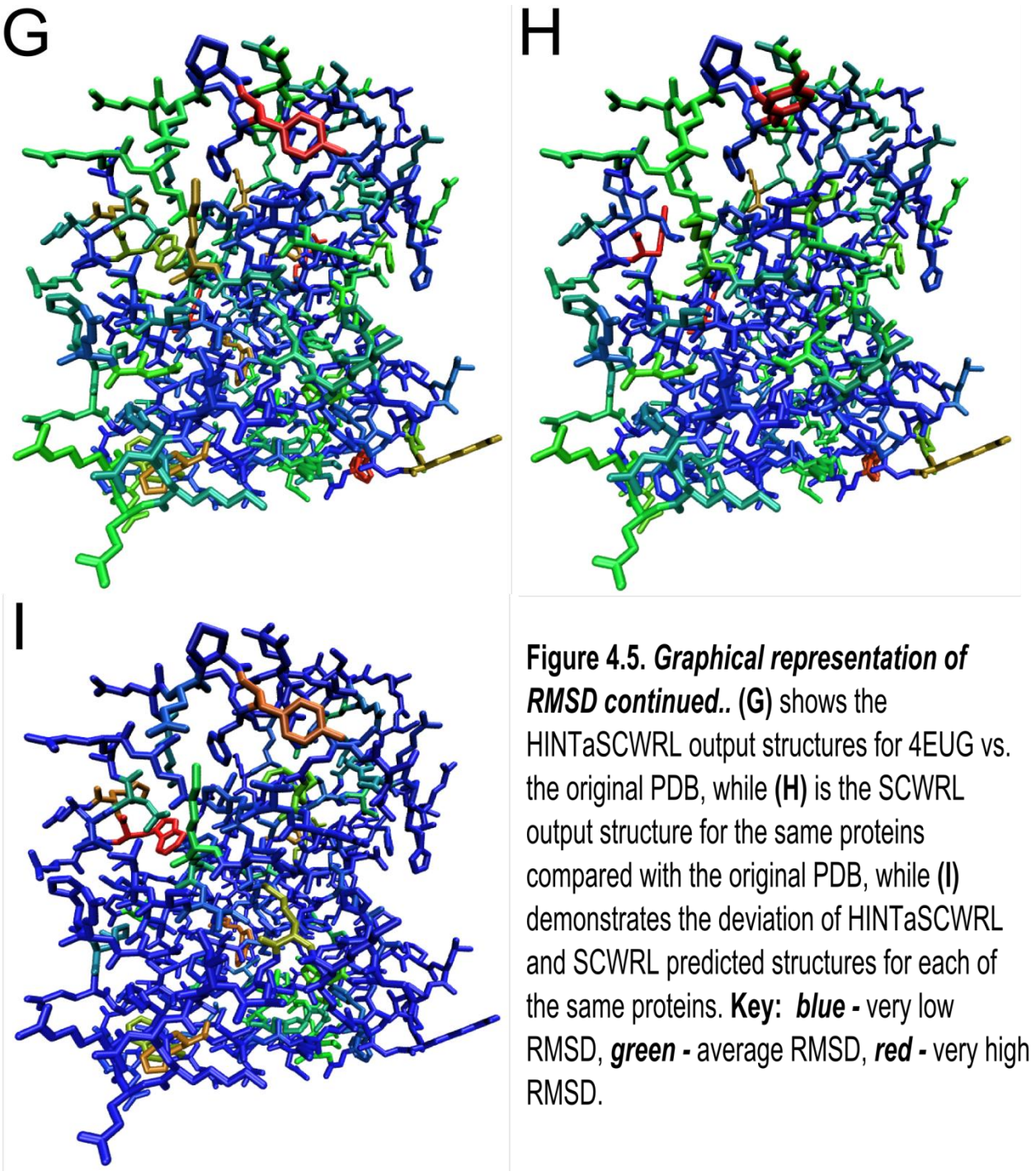


Figure 4.5. ***Graphical representation of RMSD continued..*** **(G)** shows the HINTaSCWRL output structures for 4EUG vs. the original PDB, while **(H)** is the SCWRL output structure for the same proteins compared with the original PDB, while **(I)** demonstrates the deviation of HINTaSCWRL and SCWRL predicted structures for each of the same proteins. **Key:** ***blue -*** very low RMSD, ***green -*** average RMSD, ***red -*** very high RMSD.

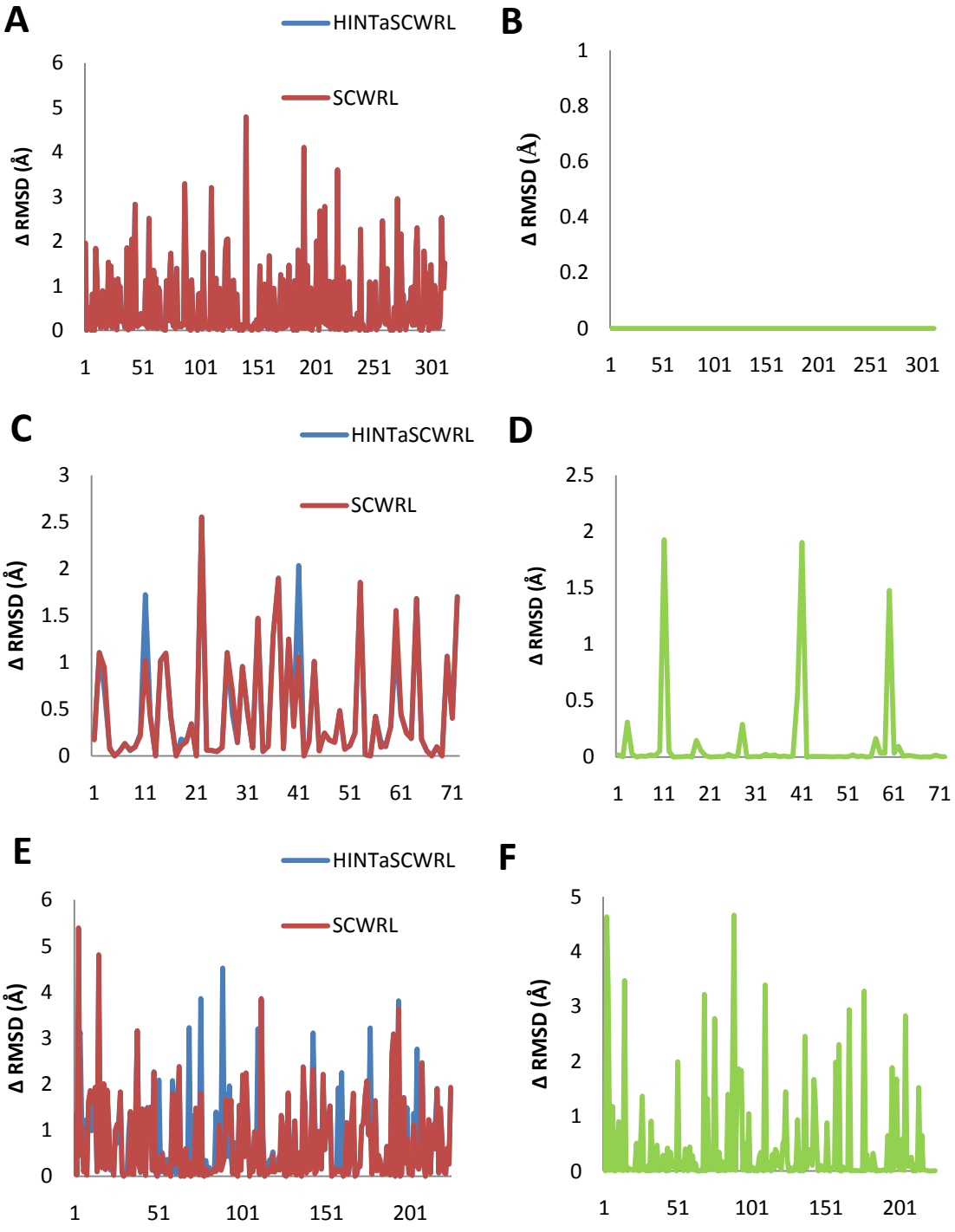


Figure 4.6 ***RMSD values plotted for each residue*** of 2CYG, 2VC8 and 4EUG in the first, second and third rows respectively. (A) (C) and (E) show RMSD values for HINTaSCWRL and SCWRL predicted structures in red and blue respectively. (B) (D) and (F) show deviations between the structures predicted by both programs in green.

in figure 4.7 (B). Interestingly, as is visible in figure 4.7 (A), Asn13 in the SCWRL output structure showed formation of an extra H-bond, which was unexpected. In contrast, the HINTaSCWRL output structure avoided bad hydrophobic-hydrophilic interactions with a nearby Ile. The formation of an additional H-bond in Asn13 within the SCWRL output structure must be a coincidence because this program only considers van der Waal's interactions.

4.9 CONCLUSIONS

The performance of our algorithm was comparable to that of SCWRL. The trend of RMSD distribution across 129 high resolution structures was similar for both programs, as was the dependence of RMSD values on solvent accessibility of the residue involved. It is remarkable that a number of residues were predicted very near their native conformations in the original PDB itself, as was shown by the random test cases (*vide supra*). It was observed that our algorithm would select different conformations of residue sidechains when it detected energetically favorable interactions that were not detected by SCWRL. The RMSD profiles of SCWRL and HINTaSCWRL predicted structures were similar to each other. In two of the three cases, both algorithms predicted very similar positions for all sidechains. However, there was one case in which the predictions were notably different. On the other hand, it was clearly shown that the all the residues which were predicted differently by the two algorithms existed on the surface of the proteins. Thus, this was not a concern because the higher steric freedom afforded to the surface residues allows movement and thus deviation is expected.

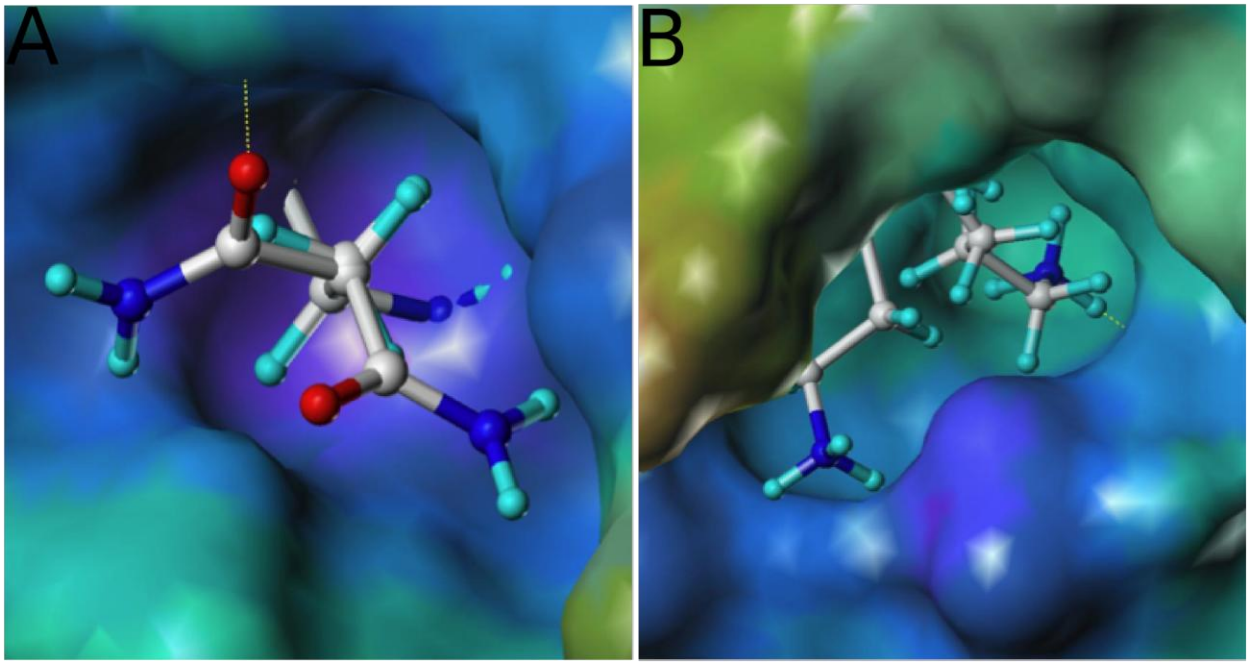


Figure 4.7 ***Positions of Sidechain Showing Highest Deviation.*** (A) Asn13 of 2VC8 in the SCWRL structure shows an extra H-bond while the same in the HINTaSCWRL structure shows none. (B) Lys171 of the HINTaSCWRL structure shows an additional H-bond compared with the same in the SCWRL predicted structure.

With these facts in mind, we may again ask ourselves: *Are we able to emulate SCWRL and its ability to optimize sidechains?* Since the structures predicted by both algorithms are very close to each other, with the major deviations isolated on the surface of the protein, we can assuredly say yes! We cannot, however, make a claim about improving sidechain prediction capabilities of SCWRL, at least with regard to emulating crystal structures.

However, the major aim of this project was not to emulate or improve sidechain optimization already provided by SCWRL, but to ascertain the ability to use HINT in conjunction with backbone-dependent rotamer libraries. Since the output structures obtained from our algorithm were extremely close to those produced by SCWRL and the differences were ascertained in the sample test cases to be caused by improved interactions, we have successfully fulfilled this aim.

At the same time, this project is still in its infancy. Therefore, we must stress the possibility that further experimentation with the sidechain placement algorithm and scoring function could possibly provide us with much improved sidechain placement.

4.10 FUTURE DIRECTIONS

With the implementation of the HINTaSCWRL algorithm, the compatibility between HINT and sidechain rotamer libraries has been established. The next stage of this project will be to create an algorithm that optimizes residue sidechains in the immediate vicinity of docked ligands. This new algorithm will be somewhat different from the current state of HINTaSCWRL because at present the latter only considers

interactions within the protein; the sidechain positioning in the next stage will have to balance intra-protein interactions with protein-ligand interactions. Moreover, the role of backbone-dependent vs. backbone-independent rotamer libraries in such an algorithm will have to be investigated.

Simultaneously, modifications in the sidechain placement strategy and scoring function must be explored in order to attempt improvement of the algorithm itself. The scaling of steric and hydropathic components of the HINT score is one possible avenue for exploration. This will be an interesting avenue to explore because packing methods have traditionally been adequate to place sidechains in algorithms such as SCWRL. Packing methods (i.e., using steric potentials such as the Lennard-Jones potential function and its variants) have proven to be especially useful due to the ease of implementation and enhanced speed of execution, thereby providing reasonable results within shorter runtimes. In contrast, a scoring function such as HINT takes longer to execute. However, it can account for several other kinds of interactions other than (and including) sterics. This should, at least in theory, allow better sidechain placement compared to algorithms that employ simple Lennard-Jones potentials.

Furthermore, weighting the probability factor could possibly enhance the quality of sidechain optimization. However, it is unlikely that simple modulation of the HINT score and probability factors alone will allow better emulation of crystal structures (compared to SCWRL), especially if the sidechain positions are being provided by a rotamer library. Unless such rotamer libraries are exhaustive, it might be difficult to cover sufficient conformational space to achieve highly accurate predictions. Thus,

sidechains might have to be rotated, in which case a strategy will have to be devised to overcome the combinatorial explosion which is imminent. This strategy can perhaps also be implemented during resolution of sidechain clashes.

4.11 REFERENCES

1. Fischer, E. Ber. Dtsch. Chem. Ges. 1890, 23, 2611.
2. Fischer, E. Ber. Dtsch. Chem. Ges. 1894, 27, 2985.
3. Koshland, D.E. Jr. The key-lock theory and the induced fit theory. *Angew. Chem. Int. Ed. Engl.* **1994,** *33,* 2375-2378.
4. Cozzini, P.; Kellogg, G.E.; Spyrakis, F.; Abraham, D.J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L.A.; Morris, G.M.; Modesto, O.; Pertinhez, T.A.; Rizzi, M.; Sotriffer, C.A. Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008,** *51,* 6237-6255.
5. Morris, G.M.; Goodsell, D.S.; Huey, R.; Hart, W.; Belew, R.; et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998,** *19,* 1639-1662.
6. Goodsell, D.S.; Olson, A.J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990,** *8,* 195-202.
7. Morris, G.M.; Goodsell, D.S.; Huey, R.; Olson, A.J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996,** *10,* 293-304.
8. Huey, R.; Morris, G.M.; Olson, A.J.; Goodsell, D.S. A Semiempirical free energy forcefield with charge-based desolvation. *J. Comput. Chem.* **2007,** *28,* 1145-1152.
9. Rosenfield, R.J.; Goodsell, D.S.; Musah, R.; Morris, G.M.; Goodin, D.B.; et al. Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. *J. Comput.-Aided Mol. Des.* **2003,** *17,* 525-536.
10. Verdonk, M.L.; Cole, J.C.; Harshorn, M.J.; Murray, C.W.; Taylor, R.D. Improved protein-ligand docking algorithm in a model binding site. *J. Mol. Biol.* **2004,** *337,* 1161-1182.
11. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001,** *308,* 377-395.
12. Zadovsky, M.I.; Kuhn, L.A. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci.* **2005,** *14,* 1104-1114.
13. Wei, B.Q.; Weaver, L.H.; Ferrari, A.M.; Matthews, B.W.; Shoichet, B.K. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.* **2004,** *337,* 1161-1182.
14. James, M.N.G.; Sielecki, A.R. Structure and refinement of penicillopepsin at 1.8 Å resolution. *J. Mol. Biol.* **1983,** *163,* 299-361.

15. Ponder, J.W.; Richards, F.M. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **1987,** *193,* 775-792.
16. Janin, J.; Wodak, S.; Levitt, M.; Maigret, B. Conformations of amino acid side chains in proteins. *J. Mol. Biol.* **1978,** *125,* 357-386.
17. Dunbrack Jr., R.L.; Karplus, M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **1993,** *230,* 543-574.
18. Dunbrack Jr., R.L.; Cohen, F.E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Prot. Sci.* **1997,** *6,* 1661-1681.
19. Desmet, J.; DeMaeyer, M.; Hazes, B.; Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **1992,** *356,* 539-542.
20. Bower, M.J.; Cohen, F.E.; Dunbrack Jr., R.L. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new tool for homology modeling. *J. Mol. Biol.* **1997,** *267,* 1268-1282.
21. 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.
22. Kellogg, G.E.; Burnett, J.C.; Abraham, D.J. Very empirical treatment of solvation and entropy: a force field derived from $LogP_{o/w}$. *J. Comput.-Aid. Mol. Des.* **2001,** *15*, 381-393.
23. Hermann, R.B. Theory of hydrophobic bonding. I. Solubility of hydrocarbons in water, within the context of the significant structure theory of liquids. *J. Phys. Chem.* **1971,** *75*, 363-368.
24. Hermann, R.B. Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.* **1972,** *76*, 2754-2759.
25. Shapovalov, M.V.; Dunbrack, R.L. Jr. Statistical and conformational analysis of the electron density of protein side chains. *Proteins,* **2007,** *66,* 279-303.

CHAPTER 5

CONCLUSIONS

Hydrophobicity impacts every aspect of drug design and even delivery, as has been repeatedly pointed out over the past century and within this dissertation. Studies of this phenomenon have resulted in multiple theories, algorithms and tools for applying the concept. A large amount of effort has been put forth into studying the partition coefficient both experimentally, especially in terms of its prediction because of its importance in “druggability” of compounds. Many theoretical methods are robust in estimating LogP for molecules similar to their training set, but large errors are fairly common for compounds with large chemical and structural differences from that set.

Since Hansch and Fujita introduced the QSAR method, drug design projects have repeatedly found use for hydrophobic parameters. This dependence of drug design on lipophilicity is intuitive, arising from drugs and proteins coming together, or proteins folding, in order to reduce the surface area in contact with polar water molecules. Quantification of this phenomenon has taken many forms, such as calculating of hydrophobic surface contact area to represent hydrophobic interactions, supplementing 3D QSAR with hydrophobic fields (HINT and MLP) [1-3] and direct quantification of intermolecular interactions with HINT [4-8]. While there are numerous force fields available, most are Newtonian in origin and concentrate on H-bonding, Coulombic interactions, van der Waal’s interactions and London forces for estimating the strength of molecular interactions, all of which are mostly if not entirely enthalpic.

HINT is different in that it accounts for both hydrophobic and hydrophilic interactions and is therefore representative of enthalpic, as well as that of entropic contributions towards biological interactions, being derived from a free energy experiment. The availability of the HINT toolkit [9] makes it possible to develop application programs for computer-aided drug discovery and design.

In this work, we presented current state of the projects aimed at exploring our hypothesis that most biological phenomena can be explained by addressing hydrophobic and hydrophilic interactions. The efflux pump project best epitomized the validity of this hypothesis. While the HINT force field has been successfully implemented in characterizing binding of small molecules to macromolecular targets and also intermacromolecular interactions in the past [4-8], these multidrug transporters posed a significantly different and complex challenge. The very fact that these huge proteins do not just bind small molecules, but transport them, was the root of this challenge. In theory, any such transport mechanism should be addressable by treating it as a series of consecutive and independent binding events. In accordance with this, we devised a method where HINT scores (representing these consecutive binding events) were used to successfully predict MIC ratios of multiple antibiotics of various classes, in conjunction with certain properties of the antibiotics themselves. It was found that LogP was a major contributor in the statistical models generated therein. However, the same descriptor alone was insufficient to achieve the same degree of predictability. More importantly, LogP itself is a measure of the efflux substrates hydrophobic nature and its

contribution towards the final model represents another way in which hydrophobic interactions might affect biomolecular phenomena and thereby supports our hypothesis.

The third chapter described our attempts at discovery of antiviral agents, which resulted in identification of several interesting compounds, of which 34% were found to inhibit hemagglutinin-neuraminidase. This project produced only moderate success rates, partly due to some problems with the tools employed therein – particularly the inability of docking methods to effectively address induced-fit during binding of small molecules. We chose to design new in-house tools in order to address these problems in the long run and have presented our preliminary investigation towards establishing feasibility of the project herein. Our studies showed that even at the simplest level of implementation, the HINT scoring function successfully placed sidechains for residues, given the backbone coordinates, which is another example where our hypothesis appears to be true. It must be admitted though, that work on this project has only just begun and a lot more needs to be done before we can claim that our hypothesis is true beyond doubt in the context of this project. With the results of this project in mind, we have chosen rotamer libraries and the HINT scoring function as our basis for design of new tools to simulate target flexibility. Further attempts at optimizing the sidechain optimization algorithm will also continue. It is generally accepted that proteins fold in such a way that hydrophobic groups are largely shielded from water by hydrophilic groups. While evaluating the factors that affect prediction of protein folds, Park *et al.* noted that hydrophobicity of residues is the largest force defining protein structure, but that other factors were involved as well [10]. Accurate hydrophobicity measurements

and estimation of hydrophobic interactions could therefore have a tremendous impact on the modeling of not only protein folding, but also side chain orientation. Better modeling and representation of both protein folding as well as side chain positioning, will also contribute to the understanding of biological processes which are significantly altered by macromolecular flexibility.

Overall, good progress was made towards the implementation of our hydropathic force field in predictive model building and the design of new tools. There is no doubt in our mind that application of this methodology in computational life sciences and computer-aided drug design will lead to accurate theoretical prediction of biological phenomena. The complex phenomena of hydrophobicity and hydrophobic interactions are still only poorly understood and remain quite difficult to simulate. However, understanding and exploiting the hydrophobic effect in drug design, e.g., docking, target structure prediction, etc., will undoubtedly be increasingly important in the future. Hydrophobicity may not be the “Holy Grail” of biomolecular phenomena, but it is definitely the one of the “Commandments”.

5.1 REFERENCES

1. Kellogg, G.E.; Semus, S.F.; Abraham, D.J. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput.-Aid. Mol. Des.* **1991,** *5*, 545-552.
2. Testa, B.; Carrupt, P.-A.; Gaillard, P.; Billois, A.; Weber, P. Lipophilicity in molecular modeling. *Pharm. Res.* **1996,** *13*, 335-343.
3. Gaillard, P.; Carrupt, P.-A.; Testa, B.; Boudon, A. Molecular lipophilicity potential, a tool in 3D-QSAR: methods and applications. *J. Comput.-Aid. Mol. Des.* **1994,** *8*, 83-96.
4. Spyrakis, F.; Amadasi, A.; Fornabaio, M.; Abraham, D.J.; Mozzarelli, A.; Kellogg, G.E. The consequences of scoring docked ligand conformations using free energy correlations. *Eur. J. Med. Chem.* **2007,** *42*, 921-933.
5. Tripathi, A.; Fornabaio, M.; Kellogg, G.E.; Gupton, J.T.; Gewirtz, D.A.; Yeudall, W.A.; Vega, N.E.; Mooberry, S. Docking and hydropathic scoring of polysubstituted pyrrole compounds with antitubulin activity. *Bioorg. Med. Chem.* **2008,** *16*, 2235-2242.
6. Simoni, D.; Invidiata, F.P.; Eleopra, M.; Marchetti, P.; Rondanin, R.; Baruchello, R.; Grisolia, G.; Tripathi, A.; Kellogg, G.E.; Durrant, D.; Lee, R.M. Design, synthesis and biological evaluation of novel stilbene-based antitumor agents. *Bioorg. Med. Chem.* **2009,** *17*, 512-522.
7. Porotto, M.; Fornabaio, M.; Kellogg, G.E.; Moscona, A. A second receptor binding site on human parainfluenza virus type 3 hemagglutinin-neuraminidase contributes to activation of fusion mechanism. *J. Virol.* **2007,** *81*, 3216-3228.
8. Burnett, J.C.; Botti, P.; Abraham, D.J.; Kellogg, G.E. Computationally accessible method for estimating free energy changes resulting from site-specific mutations of biomolecules: systematic model building and structural/hydropathic analysis of deoxy and oxy hemoglobins. *Prot. Struct. Funct. Genet.* **2001,** *42*, 355-377.
9. The HINT toolkit at the Edu-Soft, LLC. Home page. http://www.edusoft-lc.com/toolkits/
10. Park, B.H.; Huang, E.S.; Levitt, M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **1997,** *266*, 831-846.

APPENDICES

Appendix A

List of Hits from Virtual Screening for Hemagglutinin-Neuraminidase Inhibitors

ZINC ID	HINT Score	Database
ZINC04552407	6783.014	**Sigma Aldrich**
ZINC04533949	6772.188	**Sigma Aldrich**
ZINC01530138	6701.919	**Sigma Aldrich**
ZINC04825403	6504.188	**Sigma Aldrich**
ZINC01737956	6410.739	**Sigma Aldrich**
ZINC02384787	6387.623	**Sigma Aldrich**
ZINC03873854	6317.374	**Sigma Aldrich**
ZINC03873852	6260.759	**Sigma Aldrich**
ZINC03873853	6125.606	**Sigma Aldrich**
ZINC03873185	6120.066	**Sigma Aldrich**
ZINC03873855	5931.146	**Sigma Aldrich**
ZINC04544949	5902.57	**Sigma Aldrich**
ZINC04545884	5895.988	**Sigma Aldrich**
ZINC02390911	5886.638	**Sigma Aldrich**
ZINC03871260	5865.923	**Sigma Aldrich**
ZINC03830892	5855	**Sigma Aldrich**
ZINC04899504	5842.17	**Sigma Aldrich**
ZINC01529261	5760.611	**Sigma Aldrich**
ZINC03873184	5660.111	**Sigma Aldrich**
ZINC05274030	5658.885	**Sigma Aldrich**
ZINC04544668	5558.37	**Sigma Aldrich**
ZINC04556739	5543.093	**Sigma Aldrich**
ZINC05295094	5479.984	**Sigma Aldrich**
ZINC03873183	5478.967	**Sigma Aldrich**
ZINC04556499	5465.758	**Sigma Aldrich**
ZINC03014483	5445.177	**Sigma Aldrich**
ZINC04899413	5437.815	**Sigma Aldrich**
ZINC02508221	5420.523	**Sigma Aldrich**
ZINC05274031	5374.107	**Sigma Aldrich**
ZINC03870127	5365.563	**Sigma Aldrich**
ZINC04533780	5351.754	**Sigma Aldrich**
ZINC05274029	5350.48	**Sigma Aldrich**
ZINC04533783	5330.799	**Sigma Aldrich**
ZINC04214182	5325.671	**Sigma Aldrich**
ZINC03873186	5316.4	**Sigma Aldrich**
ZINC03830893	5265.247	**Sigma Aldrich**
ZINC04556495	5250.292	**Sigma Aldrich**
ZINC04533725	5241.01	**Sigma Aldrich**
ZINC02390912	5217.136	**Sigma Aldrich**
ZINC03871275	5209.737	**Sigma Aldrich**
ZINC03873187	5182.773	**Sigma Aldrich**

ZINC05273655	5153.94	**Sigma Aldrich**
ZINC04533726	5144.388	**Sigma Aldrich**
ZINC03830452	5120.193	**Sigma Aldrich**
ZINC01569744	5117.354	**Sigma Aldrich**
ZINC03873182	5082.202	**Sigma Aldrich**
ZINC03871276	4990.011	**Sigma Aldrich**
ZINC03873188	4983.702	**Sigma Aldrich**
ZINC01575534	4962.166	**Sigma Aldrich**
ZINC02384673	4912.327	**Sigma Aldrich**
ZINC00236772	4869.534	**Vitas**
ZINC04556815	4853.672	**Sigma Aldrich**
ZINC04533731	4853.62	**Sigma Aldrich**
ZINC04552406	4840.231	**Sigma Aldrich**
ZINC02707649	4832.24	**LifeChemicals**
ZINC05274037	4823.205	**Sigma Aldrich**
ZINC03869424	4815.127	**Sigma Aldrich**
ZINC03014482	4809.821	**Sigma Aldrich**
ZINC04531662	4791.451	**Sigma Aldrich**
ZINC05257890	4786.626	**LifeChemicals**
ZINC04820544	4744.787	**Sigma Aldrich**
ZINC04106683	4699.967	**Keyorganics**
ZINC05274006	4696.594	**Sigma Aldrich**
ZINC01607828	4689.038	**Sigma Aldrich**
ZINC05257957	4688.907	**LifeChemicals**
ZINC05257889	4684.735	**LifeChemicals**
ZINC04556493	4681.904	**Sigma Aldrich**
ZINC04533734	4643.627	**Sigma Aldrich**
ZINC03872461	4643.069	**Sigma Aldrich**
ZINC04533843	4641.547	**Sigma Aldrich**
ZINC03873181	4634.583	**Sigma Aldrich**
ZINC01210754	4615.563	**Vitas**
ZINC05273546	4585.158	**Sigma Aldrich**
ZINC03871263	4572.668	**Sigma Aldrich**
ZINC02575474	4554.932	**Sigma Aldrich**
ZINC04533732	4552.692	**Sigma Aldrich**
ZINC04939701	4546.525	**LifeChemicals**
ZINC04533735	4496.596	**Sigma Aldrich**
ZINC03870005	4465.234	**Sigma Aldrich**
ZINC05274008	4386.57	**Sigma Aldrich**
ZINC03872463	4378.871	**Sigma Aldrich**
ZINC04760528	4363.481	**Sigma Aldrich**
ZINC04544665	4358.829	**Sigma Aldrich**
ZINC04511392	4318.523	**Asinex**
ZINC04514036	4302.86	**Sigma Aldrich**
ZINC04513860	4294.716	**Sigma Aldrich**
ZINC04556886	4291.104	**Sigma Aldrich**
ZINC04543872	4283.146	**Sigma Aldrich**

ZINC05273678	4276.885	**Sigma Aldrich**
ZINC04566466	4269.913	**Sigma Aldrich**
ZINC06142968	4261.341	**Otava**
ZINC01893413	4243.928	**Otava**
ZINC04739725	4236.693	**Vitas**
ZINC04566465	4205.4	**Sigma Aldrich**
ZINC05273548	4201.266	**Sigma Aldrich**
ZINC05511105	4191.064	**Otava**
ZINC01638013	4186.778	**Sigma Aldrich**
ZINC04533969	4184.555	**Sigma Aldrich**
ZINC04024388	4177.399	**Keyorganics**
ZINC04291876	4159.401	**LifeChemicals**
ZINC04291877	4154.487	**LifeChemicals**
ZINC03872462	4152.638	**Sigma Aldrich**
ZINC04099087	4145.463	**Sigma Aldrich**
ZINC03999322	4136.824	**Sigma Aldrich**
ZINC01607692	4134.645	**Sigma Aldrich**
ZINC05260452	4134.227	**Sigma Aldrich**
ZINC04557073	4121.642	**Sigma Aldrich**
ZINC03869383	4119.558	**Sigma Aldrich**
ZINC02575489	4091.542	**Sigma Aldrich**
ZINC04514038	4079.696	**Sigma Aldrich**
ZINC04556887	4048.887	**Sigma Aldrich**
ZINC04545848	4048.283	**Sigma Aldrich**
ZINC04544666	4045.601	**Sigma Aldrich**
ZINC02685419	4044.966	**LifeChemicals**
ZINC04159200	4032.111	**LifeChemicals**
ZINC04533971	4030.039	**Sigma Aldrich**
ZINC05511098	4024.855	**Otava**
ZINC02685397	4023.99	**LifeChemicals**
ZINC02685329	4003.465	**LifeChemicals**
ZINC02685344	3997.909	**LifeChemicals**
ZINC04535978	3996.699	**Sigma Aldrich**
ZINC03871262	3987.519	**Sigma Aldrich**
ZINC02685352	3980.843	**LifeChemicals**
ZINC05545049	3971.142	**Otava**
ZINC05273579	3967.291	**Sigma Aldrich**
ZINC04514042	3961.863	**Sigma Aldrich**
ZINC01893410	3958.873	**Otava**
ZINC00236768	3958.214	**Vitas**
ZINC01805621	3951.552	**LifeChemicals**
ZINC02685380	3927.787	**LifeChemicals**
ZINC00574780	3927.28	**Vitas**
ZINC04040894	3923.655	**Keyorganics**
ZINC05511073	3918.173	**Otava**
ZINC05260460	3916.985	**Sigma Aldrich**
ZINC04660887	3913.51	**Otava**

ZINC05273547	3908.913	**Sigma Aldrich**
ZINC01893416	3874.391	**Otava**
ZINC00816404	3862.188	**Otava**
ZINC01576098	3857.324	**Sigma Aldrich**
ZINC02685391	3856.261	**LifeChemicals**
ZINC04507507	3842.097	**Sigma Aldrich**
ZINC05274009	3832.105	**Sigma Aldrich**
ZINC03394329	3823.13	**Enamine**
ZINC04556744	3821.942	**Sigma Aldrich**
ZINC04513863	3818.507	**Sigma Aldrich**
ZINC04533809	3811.679	**Sigma Aldrich**
ZINC04521828	3811.322	**Sigma Aldrich**
ZINC06143684	3808.482	**Otava**
ZINC01120022	3806.062	**Otava**
ZINC02043137	3802.661	**Sigma Aldrich**
ZINC04552284	3801.73	**Sigma Aldrich**
ZINC01805618	3798.232	**Otava**
ZINC02685333	3785.887	**LifeChemicals**
ZINC05260454	3780.27	**Sigma Aldrich**
ZINC04545932	3778.958	**Sigma Aldrich**
ZINC05511099	3776.205	**Otava**
ZINC05235949	3771.423	**Otava**
ZINC02685401	3754.075	**LifeChemicals**
ZINC03870800	3743.268	**Sigma Aldrich**
ZINC02685375	3740.393	**LifeChemicals**
ZINC04556500	3733.955	**Sigma Aldrich**
ZINC05258510	3732.108	**LifeChemicals**
ZINC02685370	3731.883	**LifeChemicals**
ZINC02252499	3729.479	**Vitas**
ZINC04552272	3716.438	**Sigma Aldrich**
ZINC03871402	3712.854	**Sigma Aldrich**
ZINC03870109	3712.259	**Sigma Aldrich**
ZINC04735783	3706.273	**Otava**
ZINC04523248	3701.213	**Sigma Aldrich**
ZINC04533781	3700.264	**Sigma Aldrich**
ZINC04534321	3685.927	**Sigma Aldrich**
ZINC04543589	3664.898	**Sigma Aldrich**
ZINC04660890	3648.144	**Otava**
ZINC04552275	3638.23	**Sigma Aldrich**
ZINC05511076	3623.76	**Otava**
ZINC05545035	3621.934	**Otava**
ZINC02522613	3618.054	**Sigma Aldrich**
ZINC04735797	3616.808	**Otava**
ZINC04534319	3616.223	**Sigma Aldrich**
ZINC04672996	3613.007	**Vitas**
ZINC04523366	3605.412	**Sigma Aldrich**
ZINC04065004	3598.52	**Vitas**

ZINC05885060	3586.05	**Enamine**
ZINC05260451	3577.245	**Sigma Aldrich**
ZINC05273572	3568.69	**Sigma Aldrich**
ZINC04735821	3558.965	**Otava**
ZINC01893395	3554.209	**Otava**
ZINC06590276	3528.214	**Enamine**
ZINC02047153	3528.206	**Sigma Aldrich**
ZINC04514046	3514.745	**Sigma Aldrich**
ZINC04677137	3492.472	**Vitas**
ZINC04083848	3490.832	**Vitas**
ZINC04187766	3476.841	**Vitas**
ZINC05511070	3470.568	**Otava**
ZINC05511066	3463.047	**Otava**
ZINC05511089	3443.054	**Otava**
ZINC04735769	3406.209	**Otava**
ZINC03268222	3401.479	**Enamine**
ZINC04304712	3385.03	**Otava**
ZINC04739051	3373.027	**Vitas**
ZINC01122862	3367.326	**Vitas**
ZINC04167058	3300.651	**LifeChemicals**
ZINC00969636	3290.01	**LifeChemicals**
ZINC05511112	3261.722	**Otava**
ZINC01206008	3237.435	**Otava**
ZINC00038207	3231.201	**TimTec**
ZINC05235951	3228.151	**Otava**
ZINC04939716	3220.544	**LifeChemicals**
ZINC00653397	3201.981	**TimTec**
ZINC03305588	3164.451	**Enamine**
ZINC02700719	3131.295	**LifeChemicals**
ZINC04373351	3124.746	**Asinex**
ZINC00556918	3109.576	**Otava**
ZINC02699820	3093.368	**LifeChemicals**
ZINC05516169	3086.172	**Otava**
ZINC00783224	3050.043	**Asinex**
ZINC03274602	3048.249	**Enamine**
ZINC03218782	3035.004	**Enamine**
ZINC03248836	2934.242	**Enamine**
ZINC01782161	2869.992	**Otava**
ZINC02710650	2829.9	**LifeChemicals**
ZINC04106684	2818.322	**Keyorganics**
ZINC03217938	2801.973	**Enamine**
ZINC04513866	2798.426	**TimTec**
ZINC00839413	2732.374	**TimTec**
ZINC04373354	2729.865	**Asinex**
ZINC06590275	2702.556	**Enamine**
ZINC00783223	2678.131	**Asinex**
ZINC03358847	2656.639	**Enamine**

ZINC00074476	2629.257	**Asinex**
ZINC04482492	2608.585	**Enamine**
ZINC02685062	2599.128	**Otava**
ZINC05827290	2584.551	**Enamine**
ZINC07157728	2554.532	**Enamine**
ZINC04993151	2544.796	**Asinex**
ZINC03269810	2523.898	**Enamine**
ZINC04660889	2522.86	**Otava**
ZINC04037814	2517.244	**Asinex**

APPENDIX B

Descriptor Values for All Antibiotics

Antibiotic	Efflux	LogP	MolWidth	$HINT_{nB}$	$HINT_{Z3}$	$HINT_{cE}$	$HINT_{AcrBHole}$
1	2	0.22	8.824	1421.4049	349.9535	269.3283	-358.842
2	1.322	0.63	7.8673	-48.9615	208.3675	803.6375	-192.874
3	1	1.67	8.4704	359.8735	-19.4607	759.0005	-563.7534
4	0	0.78	8.5204	895.124	609.4122	1161	-141.4805
5	0	-2.18	8.1583	1180.493	-407.6976	1681.1809	-511.7244
6	1.585	0.14	7.6195	684.2153	654.4504	676.4913	365.8505
7	0	-1.54	8.7335	-234.6077	1042.9979	823.7745	265.72
8	0.585	1.57	8.9935	219.3652	892.0615	1536.8369	152.3447
9	0.585	-0.01	8.6566	727.4464	694.5172	395.6584	-171.2888
10	3.585	-0.24	8.8245	1036.7676	281.1794	-310.9403	-83.5398
11	2.415	-0.05	8.2824	1369.0051	626.1984	624.5814	-1084.8589
12	1	-0.11	11.03	531	840.4	460.5	-476.7
13	0.415	-0.38	8.4421	11.3957	320.7032	-437.5324	-1774.9761
14	0	-0.4	8.3236	1086.635	-342.8607	1831.3839	-471.9541
15	0	0.6	9.5402	420.3004	701.2516	798.1672	-1500.9824
16	4.087	1.92	6.9961	-169.3248	-33.8065	-321.4585	-398.8438
17	0	-1.43	7.8187	266.208	-427.9128	1417.8339	264.2717
18	1.322	1.13	7.0743	673.8554	-441.0727	154.8593	-135.5974
19	1.415	0.88	7.066	883	1188.2214	2641.8313	1230.6249
20	0	0.75	7.7892	1033	724.1004	2229.8894	1100.3538
21	8.585	2.05	8.7708	897.123	-1201.1604	-372.9877	-116.3972
22	1.322	0.37	7.3337	1031.1724	-749.2806	653.3875	-2067.1899
23	5.170	0.21	10.16	576.3	354.8	234.4	-604
24	7	3.21	8.8458	302.0901	-645.2233	12.4986	-891.8931
25	7.415	2.61	8.5613	876.5013	-1601.5297	-667.899	-30.9436
26	2.585	0.2	9.9196	183.9795	-330.1823	355.9706	-324.9861
27	4	0.67	10.9224	49.3987	147.8679	-35.149	-172.376
28	0	0.06	7.87	891.5	265.4	1209.4	-321.2
29	1	1.41	6.9541	1038.7643	-668.0494	-547.1687	-517.5837
30	2	0.24	6.9829	351.8297	-862.1061	-127.8168	-968.7239
31	0	-0.19	7.3593	1248.1177	1084.7677	2562.1274	-356.574
32	0	0.22	8.4364	914.1481	-276.4886	604.3444	-809.584
33	2.585	0.11	7.2725	520.443	1222.7812	1487.6881	-311.5596
34	3	0.98	8.4555	-744.2012	-80.3125	164.3973	42.7482

35	3	-0.91	9.4782	-278.7676	-67.7812	-786.8657	-1105.425
36	7	2.56	9.3903	419.9081	248.6562	172.3196	-25.4482
37	3	2.1	5.9864	436.4709	339.5	259.7588	231.7992
38	3	-0.56	8.881	-879.8488	-272.5312	-110.5498	-358.8451
39	0	0.15	6.2727	440.6515	259.0312	314.0884	-305.9413
40	5.736	3.07	8.6761	-394.235	-420.5938	982.9062	-493.9512
41	1	0.95	9.6246	360.6436	321.3438	-264.627	-638.6501
42	8	4.33	9.0142	318.0577	110.0312	921.7446	-49.2852
43	2	-0.57	9.0729	62.2588	587.7812	-621.0016	-471.9078
44	0	-0.47	9.4612	-44.4861	440.125	-599.4806	-651.1274

VITA

Aurijit Sarkar was born on October 19, 1979, in Indore, M.P., India and is an Indian citizen. He graduated in 2000 A.D. with a Bachelor's degree in Science, specializing in pharmaceutical chemistry, chemistry and zoology from G.M.C. Holkar Science College and Devi Ahilya University, Indore. In 2003 A.D. he received his Master's degree in Science, specializing in applied chemistry and specifically in fine chemicals and drugs from Shri G.S. Institute for Technology & Science, Indore and Rajiv Gandhi Technical University, Bhopal, where he conducted research in chemical kinetics. He subsequently went on to work as a faculty member at two different engineering colleges – Shri G.S. Institute for Technology & Science, Indore and Central India Institute of Technology, Indore – for approximately 18 months before joining the Department of Medicinal Chemistry at Virginia Commonwealth University for his PhD. He has received several honors during his graduate career, including the 2010 J. Doyle Smith Award for exemplary performance as a graduate student in the department and the Charles T. Rector & Thomas W. Rorrer, Jr. Dean's Award during the same year for distinguished achievements in the areas of scholarship, research, teaching and service. He has been actively involved in student organizations, having served as the President of the Department of Medicinal Chemistry Graduate Student Association and the Alpha Student Chapter of the American Chemical Society's Medicinal Chemistry Division during the year 2009.