UCSF
UC San Francisco Electronic Theses and Dissertations

Title
Computational analysis of molecular recognition

Permalink
https://escholarship.org/uc/item/39w5h4p2

Author
Masukawa, Kevin M

Publication Date
2005

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Computational Analysis of Molecular Recognition:
Molecular Dynamics and Free Energy Calculations

by

Kevin M. Masukawa

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry and Chemical Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

.....

Date University Librarian

Degree Conferred:.....

Copyright 2004

By

Kevin M. Masukawa

Acknowledgements

This thesis would not be possible without the professional and personal support from numerous individuals that I have interacted with over the course of my graduate career. To my colleagues at UCSF who contributed to a pleasurable graduate experience, thank you for the interesting discussions, both about science and on topics completely unrelated.

My deepest gratitude to my advisor, Irwin D. Kuntz, who willingly took me under his wing upon the unexpected death of my first advisor, Peter Kollman, and treated me as his own. He provided an enriching research environment, allowing me the scientific freedom to investigate issues in computational biology that appealed to me while all the time providing support and direction.

I would also like to thank David A. Case, who graciously acted as an extra source of guidance and support in Peter Kollman's absence.

To my first advisor, Peter A. Kollman, who succumbed to cancer on May 25th, 2001. His enthusiasm and thirst for knowledge was unparalleled and inspired me to investigate the issues that are addressed in this thesis. It was an honor to have worked with him.

Finally, this thesis would not have been possible without my friends and family. I extend special thanks to them for all the unwavering support they provided throughout these years.

Abstract

Computational Analysis of Molecular Recognition:

Molecular Dynamics and Free Energy Calculations

by

Kevin Masukawa

Therapeutics, whether antibodies, proteins, small molecules, or nucleic acids, rely on recognizing their targets with high affinity and selectivity in order to be effective. This ability to discriminate one species from another is central in drug design and its various mechanisms have yet to be fully understood. Researchers are now looking into computational methods to gain an increased understanding of the processes involved in molecular recognition.

While there exists a vast array of approaches, this work focuses primarily on a rigorous, generalizable method in structure-based drug design, namely molecular dynamics and MMPBSA. This method enables detailed descriptions of receptor-ligand complexes at the atomic level, providing a quantitative insight into the molecular mechanisms of interaction.

Chapter 2 begins our investigation into molecular recognition and describes work predicting charge states of essential residues in a nucleic acid system. This study successfully predicted pKa shifts of residues in the intricate 50S ribosome active site and provided support for a proposed mechanism of peptidyl transfer. Chapter 3 furthers our comprehension of molecular association through energetic studies on a carbohydrate-protein complex. Successful estimates of the free energies of binding and successive

decomposition of the binding determinants offered a detailed analysis of recognition and further enhanced our understanding of the importance of water molecules in association events.

Chapter 4 describes work investigating drug design from an alternative point of view. Instead of considering the receptor topology directly, we examined the databases that are commonly screened. In particular, we explored the development of a method that characterizes a small molecule database.

In addition to steric and chemical complementarity, protein motion also guides recognition. In our fifth chapter, we investigated the commonly used computational method to simulate motion, molecular dynamics. Using a high-resolution crystal structure, we followed atomic forces from an MD simulation which yielded characteristic frequencies for our protein in various simulation environments. To further analyze these methods, we also quantitated the atomic forces on particular atoms. We saw that instantaneous forces were chaotic and that, in general, the complexity of the simulation environment has a negligible effect on molecular motion.

Table of Contents:

Title Page ..... i.

Acknowledgements ..... iii.

Abstract ..... iv.

Table of Contents ..... vi.

List of Tables ..... ix.

List of Figures ..... x.

Chapter 1. Introduction 1

References 12

Chapter 2. Calculation of Adenosine pKa Shifts Inside the 50S Ribosome Active Site 18

Abstract 19

Introduction 20

Methods 24

Results and Discussion 28

Conclusions 32

References 34

Chapter 3. Investigation of Neuraminidase-Substrate Recognition Using Molecular Dynamics and Free Energy Calculations 38

Abstract 39

Introduction 40

Methods 43

Results and Discussion 53

Conclusions 69

References 72

Supporting Information 81

Chapter 4. Estimating Chemical Similarity Expectation Values for Novel Compounds 84

Abstract 85

Introduction 86

Methods 88

Results 93

Discussion and Conclusions 100

References 105

Chapter 5. Obtaining Low Frequency Protein Modes by Observing Atomic Force Behavior in All-Atom Simulations 107

Abstract 108

Introduction 109

Methods 112

Results and Discussion 118

Conclusions 134

References 138

Chapter 6. Conclusions and Future Outlook 141

References 146

List of Tables:

Chapter 2:

Table 1. Proton Affinity Values for 9-methyladenine in Gas Phase and in Solution 29

Table 2. Relative pKa Shifts of Adenosine N3 Relative to N1 31

Chapter 3:

Table 1. Average Energy Contributions to the Free Energy of Binding 55

Table 2. Internal Strain Energy Change of Ligands Upon Binding 57

Table 3. Experimental Resistance Inhibition Data for an E119A Mutant 66

Chapter 4:

Table 1. Estimates of $n(0)$ for Varying Bit-String Lengths and Function Extrapolations 96

Chapter 5:

Table 1. Force Magnitude Breakdown into individual components..... 122

List of Figures:

Chapter 2:

Figure 1. Proposed mechanism of peptidyl transfer 22

Figure 2. Ribosome active site showing suggested active site residues 23

Figure 3. Protonation sites on adenosine available for base catalysis. 23

Figure 4. Thermodynamic cycle used to calculate pKa shifts. 26

Figure 5. Cross section of ribosome active site showing 15Å sphere around A2486 27

Chapter 3:

Figure 1. Neuraminidase ligands studied in this investigation 42

Figure 2. Sialic Acid boat/chair conformations 47

Figure 3. Torsional potential for the C-C-O-C bond in pyranose 48

Figure 4. All-atom RMSD of sialic acid simulation 49

Figure 5. All atom RMSD plots of ligand positions 61

Figure 6. C_α Overlay of snapshots from Relenza MD trajectory taken with and without crystal waters 62

Figure 7. C_α Overlay of snapshots from Tamiflu MD trajectory taken with and without crystal waters 63

Figure 8. Computational alanine-scanning results 68

Chapter 4:

Figure 1. *v(r)* and *n(r)* for 40,000 points distributed uniformly random in a 5 dimensional box and a 15 dimensional box 95

Figure 2. *v(r)*, *n(r)*, and *n(v(r))* are plotted for the 1199 diverse subset of the ACD 98

Figure 3. Curve fittings to numerically derived *n(r)*..... 101

Figure 4. *n(r)*, and *v(r)* for the 1199 diverse subset of the ACD using Daylight fingerprints for *l* equal to 512 using the Soergel distance 103

Chapter 5:

Figure 1. Trigonal unit cell containing six alpha lytic protease proteins..... 113

Figure 2. Spectral density of the Fourier transform obtained from solvated MD of two bonded carbons in a box of water 117

Figure 3. Potential Energy RMSD over the course of the production MD simulations 119

Figure 4. Ribbon diagram of the asymmetric unit of αLP 120

Figure 5. Bonded two carbon simulations at varying temperatures 124

Figure 6. Force histograms for atom 787 CG1 on residue valine 58 (exterior). 126

Figure 7. Autocorrelation analysis of valine 58 atom 787 CG1 (exterior)..... 129

Figure 8. Autocorrelation analysis of valine 23 atom 310 CG1 (interior)..... 130

Figure 9. Spectral density of αLP in various environment 133

Figure 10. Averaged forces during the unit cell trajectory 135

Chapter 1: Introduction

by

Kevin M. Masukawa

Department of Pharmaceutical Chemistry

Chemistry and Chemical Biology Program

University of California, San Francisco

600 – 16th Street, Box 2240

San Francisco, CA 94143-2240

Computers in Drug Design. With the recent completion of sequencing the human genome, science entered a new era of research, providing researchers with vast amounts of information[1,2]. This significant achievement greatly increased our ability to understand debilitating diseases at the molecular level by providing a platform for the identification of genes and pathways that cause them. To date, in addition to the human genome, numerous other eukaryotic organisms have now been sequenced and the identification of essential genes that can serve as therapeutic targets has begun[3,4]. Probing the biological significance and the subsequent inhibition of these potential biological targets with small compounds is nontrivial and poses one of the next grand challenges in molecular and structural biology. To complement the vast amounts of structural information that will be available, researchers are now in need of complementary tools that aid in the transition of information from essential genes and targets to small molecule inhibitors. Ultimately, having the ability to sift through this information in an efficient yet meaningful manner enables the development of new therapeutics that satisfy unmet medical needs and are superior to existing treatments.

In this new era of research, computational methods have come to the forefront due to their ability to analyze data in short periods of time. With recent applications to biological problems, researchers have realized that computers provide a platform to rapidly yield information on how two highly specific molecular species recognize each other. In addition to understanding how computers can be used in biomedical research, an increase in computing power and better understanding of the properties that govern molecular recognition has helped make computers a common tool on the desktop of researchers today. Drug discovery is a long and expensive process; the reliable and rapid

identification of novel lead compounds can greatly speed up or enhance our ability to get new therapeutics to the general public.

Structure-based drug design, or rational drug design, is predicated on the ability of theorists to predict binding properties of systems based on known three-dimensional structures. There exists numerous factors required for the success of a compound to become a therapeutic, namely strong affinity, a good toxicity profile, and the ability to get to the site of action. In early drug design efforts, high affinity and selectivity are of utmost importance; the methods in structure-based drug design presented throughout this thesis are directly used to optimize them. Already the industry has seen successes in compounds developed through the use of structure-based design. Some of these include Viracept, a modulator of HIV protease; Tamiflu and Relenza, anti-influenza compounds; Capoten, for hypertension [5]. In all those instances, computers were used to estimate and elucidate molecular recognition for the development of novel therapeutics.

Molecular recognition. Molecular association of two species can be classified as a covalent or non-covalent process. Covalent interactions involve the transfer of electrons from one species to another; therefore, theoretical studies of these events require quantum mechanics calculations and are limited to small systems. Non-covalent interactions are governed by the chemical and structural complementarity of two molecular species through space; the techniques used to estimate their interactions being much more amenable to large protein systems. While multi-body non-covalent interactions can be complex, simple two body non-covalent interactions can be described as below [6]:

$$R_{eq} + L_{eq} \xleftrightarrow{K_{eq}} RL_{eq} \quad (1)$$

$$K_a = K_d^{-1} = \frac{[RL]}{[R][L]} \quad (2)$$

$$\Delta G = -RT \ln K_a = \Delta H - T\Delta S \quad (3)$$

Equation 1 expresses the complex, *RL*, in equilibrium with a free receptor, *R*, that has one binding site for a small compound or protein, *L*. The association constant in equation 2, K_a, describes the strength of the interaction. The commonly used dissociation constant, K_d, is shown as the reciprocal of K_a. A smaller K_d value represents a tighter binding complex. Due to its dependence on association constants and therefore description of molecular association, free energy, ΔG, shown in equation 3 is the most important thermodynamic quantity to computational chemists.

At the atomic level, the association between two partners (e.g. a ligand and its target) is highly complex with numerous intricate interactions affecting the free energy. These intra- and inter-molecular recognition elements are dependent on chemical and steric complementarity, molecular motion, and charge states of important residues involved in binding (pKas). As will be discussed further below, the estimation of conformational and binding free energies for a large solvated protein system remains nontrivial, and the computational methods used to estimate it vary in their complexity.

Structure-Based Design. Over the years, structure-based design has looked to quantitate the elements in equation 3, namely ΔH and ΔS. As a result, numerous methods have been developed that estimate these properties, taking into consideration chemical and structural complementarity of the partners[7]. Some of the more advanced techniques combining scoring functions with sampling methods to obtain configurations of the

system around which to score. Scoring functions calculate the energy of the system based on the spatial relationship of the atoms to each other. These scoring functions make use of a molecular mechanics force field that is parameterized against experimental or high level computational data[8-10]. A typical additive energy function will treat the system in a pairwise manner:

$$E_{total} = \sum_{Bonds} K_r (r - r_{eq})^2 + \tag{4}$$

$$\sum_{Angles} K_\theta (\theta - \theta_{eq})^2 +$$

$$\sum_{Dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] +$$

$$\sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right]$$

The potential energy of the system E_{total} is a function of all the coordinates within the system. The function above represents the bonds stretching and bond bending harmonically. The torsional potential is represented as a fourier series. Nonbonded interactions are treated with a 6-12 potential describing the van der Waals forces, and the electrostatic terms are treated using coulombs law. For protein systems, the number of pairwise interactions can reach into the millions of interactions that need to be evaluated, further emphasizing the need for computers in this arena.

Entropic contributions, ΔS, to the free energy are broken into solvent and solute components. Solvent entropy is usually represented as a function of the amount of

surface area buried upon binding (hydrophobic effect) and is tied into the enthalpic contributions[11,12]. Solute entropic contributions to binding are significantly more difficult to estimate and require time consuming computational techniques[13].

Various packages will score systems with variations on the energy function above, but will try to capture the same type of interactions, with the more elaborate methods taking into account desolvation penalties and solute entropy estimates. Some techniques used in structure-based design include: DOCK[14], Linear Interaction Energy (LIE)[15], Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA)[16], Free Energy Perturbation (FEP)[17], pharmacophore models[18], and other methods with less rigorous scoring functions[19,20]. Ultimately, all these methods yield a theoretical binding energy for a system, but with varying levels of accuracy and speed.

Pharmacophore models, usually implemented in cases where structural information is sparse or unreliable, perform a three-dimensional substructure search to find molecules that satisfy certain chemical and spatial requirements, that are predicted to have an effect on the biological activity of a target. Pharmacophore models can be developed using information from the receptor or from information centered around a known inhibitor[21,22]. While databases of compounds can be screened rapidly with pharmacophore models, these models inherently lack the structural information that has helped define success in structure-based design and are only useful in initial hit identification and are limited in lead optimization capabilities.

Today, numerous research groups have followed in the success of early DOCK versions and developed similar algorithms for structure-based design[23-25]. Much like the initial version, these structure-based docking methods have the capability to screen

hundred of thousands of compounds in a matter of days. Virtual screens of libraries are possible due to the stripped down versions of scoring functions, which generally just contain the nonbonded components, allowing the rapid estimation of binding free energies. Of course, this speed is provided with a trade off in accuracy.

Linear response or linear interaction energy methods use a regression type analysis to fit coefficients in front of the commonly used nonbonded terms, electrostatics and van der Waals, to describe association. While these method have shown to be predictive for a system with a fair amount of experimental data[26,27], by the nature of its parameterization, it remains non-generalizable to diverse systems.

One of the more robust molecular mechanics techniques available today is free energy perturbation (FEP) or thermodynamic integration (TI). FEP methods are often used to predict changes in binding energy due to small changes in structure. It is based on a statistical approach for sampling and utilizes a thermodynamic cycle to calculate free energies. While free energy perturbation yields results that are extremely accurate, they are currently too time consuming for practical use in drug design [28].

A recent method in structure-based drug design that will be discussed in detail and used in this body of work is Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA)[16]. MMPBSA, developed in 1998 by Peter Kollman and David Case, utilizes a combination of a high level scoring function with molecular dynamics (MD) to obtain an average estimate of the free energy of a system. MMPBSA allows a rapid, reliable, and accurate estimate of the free energy while maintaining its ability to be generalizable to a wide range of biological problems. Unlike many of its counterparts in structure-based design, MMPBSA incorporates molecular motion to obtain an average energy

around a set of configurations for a system and also uses molecular motion as another means to further understand binding mechanisms. Free energies obtained using MMPBSA account for desolvation penalties by using a continuum solvent approach in addition to its gas phase molecular mechanics energies. Using equation 5, MMPBSA calculates the free energy of binding by taking energy differences between the reactants in the system and the product (i.e. complex).

$$\Delta G_{Bind} = \Delta G_{Complex} - (\Delta G_{Receptor} + \Delta G_{Ligand}) \qquad (5)$$

As will be discussed in detail in chapter 3, neuraminidase studies utilizing MD and MMPBSA were able to shed light on the necessary recognition elements for a tight binding complex. In addition to the work contained in this thesis, much success in the past has been achieved using this method to estimate conformational energies of biomolecules[29-31] and to estimate free energies of binding for small molecules to protein or nucleic acid targets [32-35].

Our first study begins an investigation into one of the numerous factors that influence recognition, molecular pKa's. Pka's yield information on the tendency for a residue to be charged or uncharged and because electrostatics plays such a big role in binding, it is important that the research community look into our abilities to properly represent them.

The material outlined in chapter 2 of this thesis examines pKa changes within the active site of the recently solved ribosome crystal structure[36] in attempts to understand peptidyl transfer at the atomic level. Active site residues commonly rely on altering pKa

values of their functionalities prior to binding substrates or performing catalysis. Having the ability to predict pKa shifts in active sites would allow a better representation of active site conformations of targets and increase the chances of identifying potent lead compounds. Due to the high charge density, predicting pKa shifts of an intricate nucleic acid system is a challenge and presents one of the major hurdles in computational biology. Our investigation inside the ribosome active site was able to successfully uncover which of the proposed nucleotide sites experienced the largest pKa shift, suggesting its role in peptidyl transfer.

In chapter 3 we complement and improve current understandings of receptor ligand interactions, particularly in carbohydrate binding proteins. Of the three major classes of biopolymers, namely proteins, nucleic acids, and carbohydrates, carbohydrates have been the least studied. In the past, the lack of carbohydrate research has been associated with multiple factors, such as synthetic difficulties, low binding affinities, and the ease in which they are degraded by the body. Today, a better understanding of the role of carbohydrates in disease states, along with recent advances in synthetic methods has spawned a new interest in carbohydrate research[37-39]. We investigated the necessary binding determinants needed to establish a strong carbohydrate-protein interaction, thus helping to overcome one of the barriers to drug development in this field. An in-depth study involving neuraminidase-inhibitor complexes was used to increase our knowledge of carbohydrate binding proteins and aid in uncovering the molecular basis for their recognition events.

Solvated systems remain of most interest in structural biology because they best mimic real life processes. Water interactions with itself as well as with proteins and

ísco
Y
ITY O
C
ITY O
L
Sa
OF CA
ísco
Y
ITY OF
S
ITY OF
L
Sa
r OF CA
co
Y OF CA
ísco
Y
ITY OF C
S
I
ITY OF C
L

ligands remain an area yet to be fully explored in structural biology. While much success has been achieved mimicking water's intrinsic properties[40-42], limited studies have been performed that attempt to answer its role in molecular recognition. Carbohydrates are amphipathic by nature, consisting of hydrophobic and hydrophilic functionalities, therefore interactions with molecular waters is apparent. The role water plays in molecular recognition can be significant and should not be neglected. Our neuraminidase study addresses this issue common in many receptor-ligand complexes, and shed light on the important role water and conformational flexibility play in mediating interactions.

Sampling. In addition to electrostatic and steric complementarity, motion plays an integral role in how molecules recognize each other. Protein dynamics can lead to altering of an active site to accommodate ligands or even altering of a protein interface to modulate interactions with other proteins. Molecular dynamics is commonly employed throughout studies in this thesis to sample conformations of a system. Molecular dynamics utilizes the pairwise potential function in equation 4 and simulates the movement of each atom based on Newton's equations of motion. As we will illustrate in the work contained in this thesis, motion of small molecules and their receptors plays a large part in determining what they recognize and the strength of that interaction.

Structure-based drug design methods, particularly MMPBSA, are highly dependent on three-dimensional structures and are limited by our ability to accurately simulate movements. X-ray and NMR yield a single atomic structural from an ensemble of accessible structures, this single structure is time averaged and yields no information regarding the frequencies of motion. The fifth chapter of this thesis begins our investigation into the development of a method that further characterizes protein motion

and the forces associated with molecular mechanics force fields. All-atom molecular dynamics is used to study a high-resolution crystal structure of alpha lytic protease [43]. Use of the high-resolution structures provides a starting point that is mostly empirical and not processed too heavily by refinement. This fairly "untouched" system enables us to follow forces with the least amount of theoretical error due to system setup. General force behavior, consisting of magnitudes and direction, during a simulation is tracked and characterized for atoms in various environments.

2D Database Searching. In the absence of information about the structure of a receptor, drug design efforts focus on identifying ligands within a large database that are two-dimensionally chemically and structurally similar to known inhibitors in hopes that they will have similar biological activity. In chapter 4 we begin an examination into methods that characterize large databases of small molecules. The work in this chapter addresses drug design from the point of view of looking at the information contained in the databases as opposed to the structures of the targets. Databases of compounds contain hundreds of thousands of various potential hits. To help mine this information we implement a hyper-sphere model that was originally developed for protein folding studies[44] and attempt to characterize databases to yield the significance of a hit from a virtual screen. This work lays the groundwork for future advances in database characterization.

References:

(1) Austin, C. P. The impact of the completed human genome sequence on the development of novel therapeutics for human disease [Review]. *Annual Review of Medicine* **2004**, *55*, 1-13.

(2) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J. et al. The sequence of the human genome. *Science* **2001**, *291*, 1304-+.

(3) Thomas, J. W.; Touchman, J. W.; Blakesley, R. W.; Bouffard, G. G.; Beckstrom-Sternberg, S. M. et al. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **2003**, *424*, 788-793.

(4) Galperin, M. Y.; Koonin, E. V. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res* **2004**, *32*, 5452-5463.

(5) Hardy, L. W.; Malikayil, A. The Impact of Structure-Guided Drug Design on Clinical Agents. *Current Drug Discovery* **2003**, 15-20.

(6) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors [Review]. *Angewandte Chemie-International Edition* **2002**, *41*, 2645-2676.

(7) Joseph-McCarthy, D. Computational approaches to structure-based ligand design [Review]. *Pharmacology & Therapeutics* **1999**, *84*, 179-191.

(8) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179-5197.

(9) Jorgensen, W. L.; Maxwell, D. S.; Tiradorives, J. Development and Testing of the Opls All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225-11236.

(10) Mackerell, A. D.; Wiorkiewiczkuczera, J.; Karplus, M. An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids. *Journal of the American Chemical Society* **1995**, *117*, 11946-11975.

(11) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy. *Journal of the American Chemical Society* **2003**, *125*, 9523-9530.

(12) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci U S A* **1987**, *84*, 3086-3090.

(13) Hayward, S.; Kitao, A.; Go, N. Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins* **1995**, *23*, 177-186.

(14) Kuntz, I. D.; Blaney, J. M.; Oarley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *Journal of Molecular Biology* **1982**, *161*, 269-288.

(15) Aqvist, J.; Medina, C.; Samuelsson, J. E. New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Engineering* **1994**, *7*, 385-391.

(16) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, Rna, and Phosphoramidate - DNA Helices. *Journal of the American Chemical Society* **1998**, *120*, 9401-9409.

(17) Zwanzig, R. W. High-temperature Equation of State by a Perturbation Method. *The Journal of Chemical Physics* **1954**, *22*, 1420-1426.

(18) Carlson, H. A.; Masukawa, K. M.; McCammon, J. A. Method for including the dynamic fluctuations of a protein in computer-aided drug design. *Journal of Physical Chemistry* **1999**, *103*, 10213-10219.

(19) Spolar, R. S.; Livingstone, J. R.; Record, M. T., Jr. Use of liquid hydrocarbon and amide transfer data to estimate contributions to thermodynamic functions of protein folding from the removal of nonpolar and polar surface from water. *Biochemistry* **1992**, *31*, 3947-3955.

(20) Murphy, K. P.; Xie, D.; Garcia, K. C.; Amzel, L. M.; Freire, E. Structural energetics of peptide recognition: angiotensin II/antibody binding. *Proteins* **1993**, *15*, 113-120.

(21) Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L. et al. Developing a dynamic pharmacophore model for HIV-1 integrase. *J Med Chem* **2000**, *43*, 2100-2114.

(22) Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications to drug design. *Curr Med Chem* **2004**, *11*, 71-90.

(23) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J Mol Recognit* **1996**, *9*, 1-5.

(24) Kramer, B.; Rarey, M.; Lengauer, T. CASP2 experiences with docking flexible ligands using FlexX. *Proteins* **1997**, *Suppl 1*, 221-225.

(25) Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. E. Hook - a Program for Finding Novel Molecular Architectures That Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site. *Proteins* **1994**, *19*, 199-221.

(26) Zhou, R. H.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L. et al. New linear interaction method for binding affinity calculations using a continuum solvent model. *Journal of Physical Chemistry B* **2001**, *105*, 10388-10397.

(27) Wall, I. D.; Leach, A. R.; Salt, D. W.; Ford, M. G.; Essex, J. W. Binding constants of neuraminidase inhibitors: An investigation of the linear interaction energy method. *Journal of Medicinal Chemistry* **1999**, *42*, 5142-5152.

(28) Kollman, P. Free Energy Calculations - Applications To Chemical and Biochemical Phenomena. *Chemical Reviews* **1993**, *93*, 2395-2417.

(29) Duan, Y.; Kollman, P. A. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science* **1998**, *282*, 740-744.

(30) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. Molecular Dynamics in the Endgame of Protein Structure Prediction. *Journal of Molecular Biology* **2001**, *313*, 417-430.

(31) Reyes, C. M.; Kollman, P. A. Structure and thermodynamics of RNA-protein binding: Using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change. *Journal of Molecular Biology* **2000**, *297*, 1145-1158.

(32) Massova, I.; Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding [Review]. *Perspectives in Drug Discovery & Design* **2000**, *18*, 113-135.

(33) Gouda, H.; Kuntz, I. D.; Case, D. A.; Kollman, P. A. Free energy calculations for theophylline binding to an RNA aptamer: MM-PBSA and comparison of thermodynamic integration methods. *Biopolymers* **2003**, *68*, 16-34.

(34) Huo, S. H.; Wang, J. M.; Cieplak, P.; Kollman, P. A.; Kuntz, I. D. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: Insight into structure-based ligand design. *Journal of Medicinal Chemistry* **2002**, *45*, 1412-1419.

(35) Rizzo, R. C.; Toba, S.; Kuntz, I. D. A molecular basis for the selectivity of thiadiazole urea inhibitors with stromelysin-1 and gelatinase-A from generalized born molecular dynamics simulations. *Journal of Medicinal Chemistry* **2004**, *47*, 3065-3074.

(36) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science* **2000**, *289*, 905-920.

(37) Koeller, K. M.; Wong, C. H. Emerging themes in medicinal glycoscience. *Nature Biotechnology* **2000**, *18*, 835-841.

(38) Plante, O. J.; Palmacci, E. R.; Seeberger, P. H. Automated solid-phase synthesis of oligosaccharides. *Science* **2001**, *291*, 1523-1527.

(39) Sears, P.; Wong, C. H. Toward automated synthesis of oligosaccharides and glycoproteins. *Science* **2001**, *291*, 2344-2350.

(40) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **1983**, *79*, 926-935.

(41) Silverstein, K. A. T.; Haymet, A. D. J.; Dill, K. A. The strength of hydrogen bonds in liquid water and around nonpolar solutes. *Journal of the American Chemical Society* **2000**, *122*, 8037-8041.

(42) Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology* **1996**, *3*, 973-980.

(43) Fuhrmann, C. N.; Kelch, B. A.; Ota, N.; Agard, D. A. The 0.83 angstrom resolution crystal structure of alpha-lytic protease reveals the detailed structure of the active site and identifies a source of conformational strain. *Journal of Molecular Biology* **2004**, *338*, 999-1013.

(44) Sullivan, D. C.; Kuntz, I. D. Conformation spaces of proteins. *Proteins: Structure, Function, & Genetics* **2001**, *42*, 495-511.

isco
Y
SITY O
C
SITY O
L
Sa
OF CA
isco
Y
SITY OF
S
ITY OF
L
Sa
, OF CA
co
Y OF CA
isco
Y
SITY OF
S
I
SITY OF

Chapter 2: Calculation of Adenosine pKa Shifts Inside the 50S Ribosome Active Site

Kevin M Masukawa, Peter A Kollman[†], and Ray Luo[‡*]

Department of Pharmaceutical Chemistry

Chemistry and Chemical Biology Program

University of California, San Francisco

600 – 16th Street, Box 2240

San Francisco, California 94143-2240

[†] Deceased

[‡] Current Address: Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697.

Abstract:

We present theoretical pKa calculations on a conserved adenosine, 2486 (2451 *E.* Coli. numbering), located inside the 50S ribosome active site. Recent biochemical and structural evidence suggests this adenosine is involved in peptidyl transfer[1]. PKa calculations were used to evaluate the effect the electrostatic environment created by the ribosome active site has on this adenosine's pKa. Despite calculated intrinsic proton affinities showing that the N1 position on adenosine is more basic than the N3 position, pKa calculations in the ribosome active site showed that the N3 pKa shift is sufficient to not only overcome this differential, but to also make it the most likely candidate for proton addition, in accord with recent biochemical results. We also present evidence that suggests that this shift is not due to a particular hydrogen bond with G2102 (2067 E.Coli. numbering) and not due solely to the phosphodiester backbone of the ribosome. These calculations imply that the ribosome uses its three-dimensional organization of bases and sugars in the immediate vicinity to generate an electrostatic environment to cause this large pKa shift instead of this pKa shift being due to a specific hydrogen bond network or backbone charge. The theoretical work presented here lends support to A2486 playing a role in peptidyl transfer and to the notion that the high pKa shift seen in the ribosome stems from the N3 site on A2486 instead of the N1 site.

Introduction:

Proteins, essential for all living organism, vary significantly in structure and function despite being composed of a combination of the same 20 amino acids. The integrity of each protein is vital to maintain the intricate interactions it has with substrates within the cell. For that reason, the tightly regulated cellular machinery that is involved in protein synthesis must maintain a high level of fidelity and efficiency in order for the cell to function properly and survive. Protein synthesis, which involves a chemical step leading to the formation of a peptide bond, utilizes a complex network of cofactors to turn messenger RNA (mRNA) into polypeptides. Of the numerous cellular components involved in protein synthesis, the ribosome is integral because it contains the catalytic functionality.

The ribosome, acting as a scaffold, mediates interactions between tRNA, mRNA, elongation factors, and GTP, to catalyze the peptidyl transfer of one amino acid to a growing polypeptide chain[2-4]. Biochemical evidence suggests an acid-base, serving a role analogous to histidines in serine proteases, deprotonates incoming amino-acyl tRNA fragments in order to facilitate nucleophilic attack on the growing peptide chain (Figure 1)[5].

Recently, the atomic structure of the *Haloarcula Marismortui* 50S subunit was solved to 2.4Å resolution[6]. With the elucidation of the 50S subunit structure showing that the ribosome catalytic core is comprised purely of RNA, researchers now agree that the ribosome is a ribozyme and that a nucleotide likely serves as the catalytic acid-base[6]. In order for a residue to serve this role, its pKa must be near physiological values. Adenosine and cytidine have intrinsic pKa values far from neutrality, yet are the only two

RNA nucleotides that have titratable sites remotely capable of participating in acid-base catalysis[7,8]. Therefore, in order for the ribosome to use nucleotides to catalyze peptidyl transfer, it must utilize its local environment to help perturb pKa values to neutrality.

Biochemical and structural studies on the 50S ribosome have revealed that a conserved adenosine, 2486, in domain 5 experiences an unusually large pKa shift and is also well positioned to abstract a proton from incoming tRNA fragments (Figure 2)[6,9]. This aberrant pKa shift on A2486 (A2451 E. coli numbering) has been measured to be 7.6 ± 0.2 in E.coli by observing reverse transcriptase disruption through adenosine methylation over a pH range using dimethylsulphate (DMS)[9]. Questions regarding which site on adenosine is the most basic arise because while N3 is best positioned to act as a base, DMS is more reactive to the N1 site[10]. Thus, despite this evidence, there remains a discussion of whether adenosine 2486 is the catalytic nucleotide and if it is, which of its two titratable sites, N1 or N3, is acting as the catalytic base to initiate the transfer (Figure 3)[7]. To address this ambiguity, we report theoretical pKa calculations using the current 50S atomic structure to provide further evidence for adenosine 2486 being the recipient of the high pKa shift and go further to show that N3 on this adenosine, and not N1, is the preferred protonation site (Figure 3).

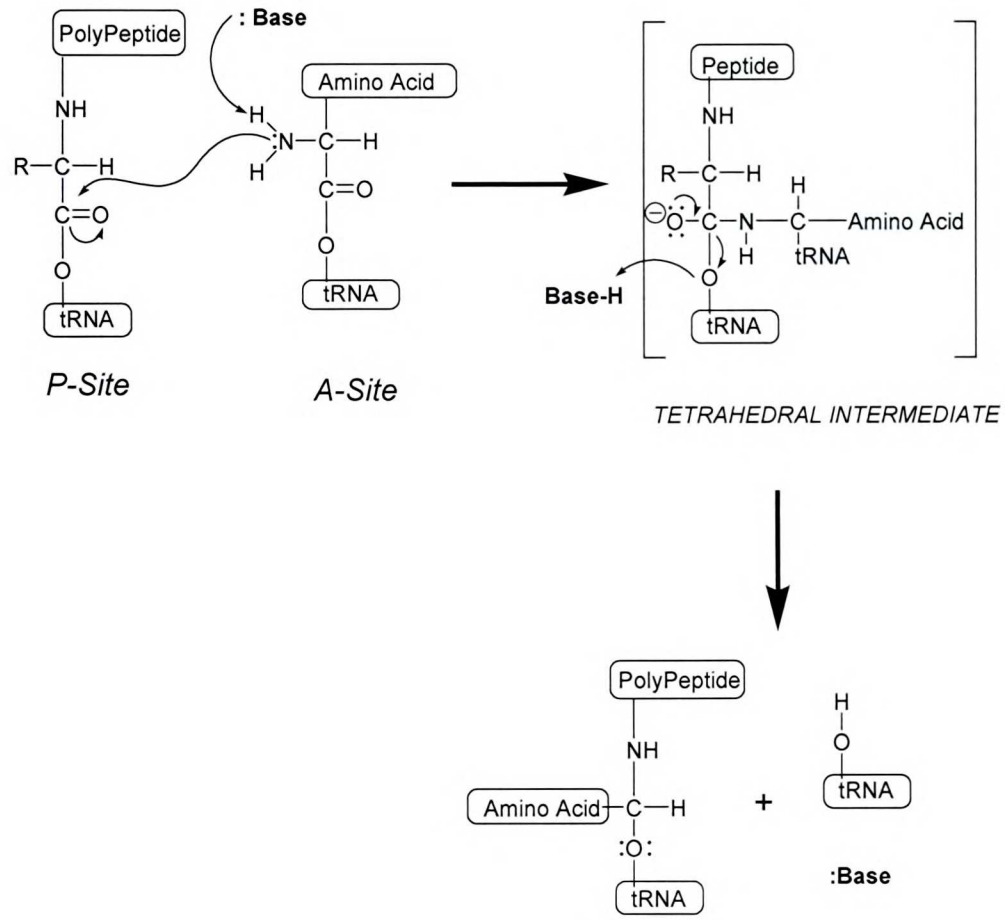


Figure 1. Proposed mechanism of peptidyl transfer within the context of the ribosome.

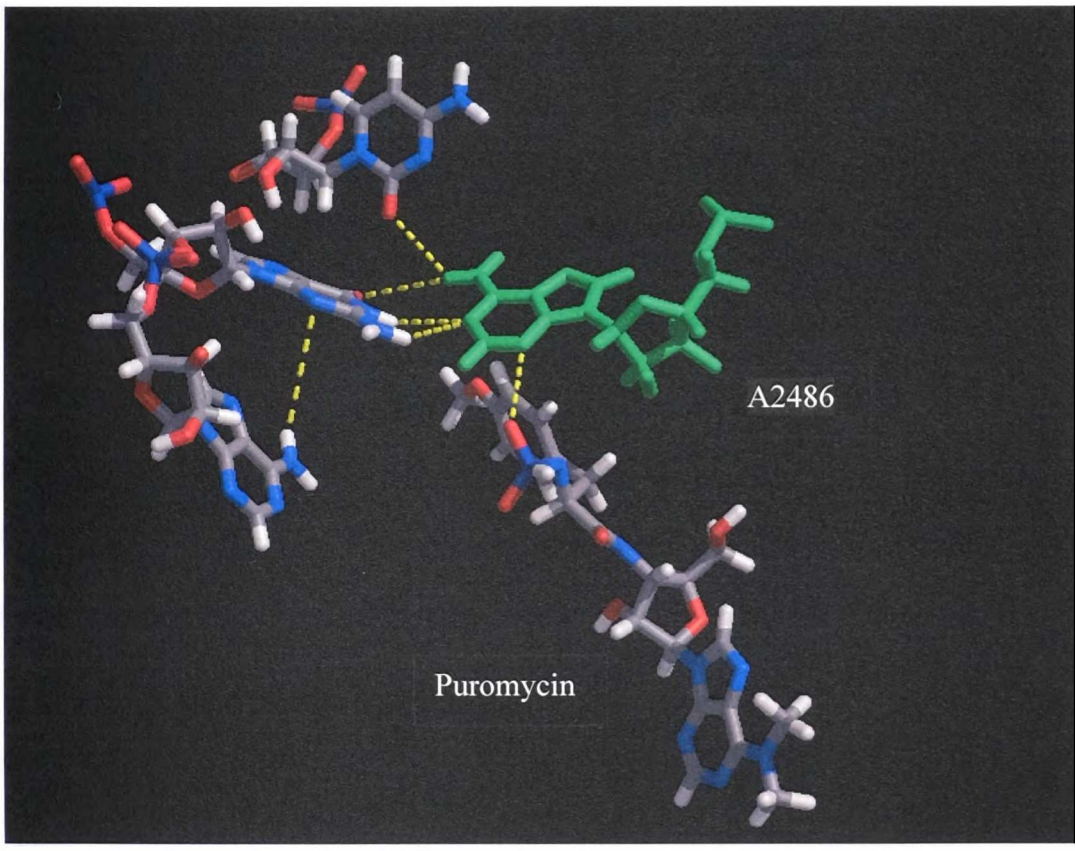
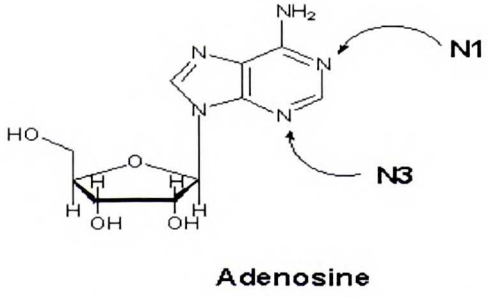


Figure 2. Ribosome active site showing suggested active site residues. Adenosine 2486 is highlighted in Green. Puromycin, a known peptidyl transfer inhibitor is also shown.



Adenosine

Figure 3: protonation sites on adenosine available for base catalysis.

Computational Methods:

Calculation of Proton Affinities. Due to the lack of experimental data on the individual pKas of N1 and N3 on adenosine, we initially performed quantum mechanical proton affinity calculations using 9-methyladenine to help establish the relative intrinsic energy differences between the two sites. Calculations were performed using Gaussian 98[11]. Hartree Fock (HF) with a 6-31G* basis set and second-order perturbation theory (MP2) with a cc-tzvp basis set, were both performed in gas phase and then in solution with the polarizable continuum model (PCM)[12] in Gaussian98[13]. The use of the increased level of theory with a larger basis set was to show the adequacy of the HF/6-31G* basis set for further calculations. For each set of calculations, protonated and unprotonated, geometry optimizations and subsequent single point calculations with zero point energy corrections were performed using the same level of theory and basis set.

Determination of pKas. Upon establishment of the relative intrinsic proton affinity differences, pKa calculations were devised so that the pKa shift on adenosine 2486 induced by the electrostatic potential of the ribosome could determine which site, N1 or N3, is the most likely candidate for proton addition within the ribosome. This work involved studying the single protonation state of adenosine at one of the ring nitrogens. PKa values were obtained using the cycle illustrated in figure 4 and further described by equation 1[14,15].

$$\text{50S-AdeN1-H} \xrightarrow{\Delta G_{(N1)}} \text{50S-AdeN1} + \text{H}$$

$$\text{50S-AdeN3-H} \xrightarrow{\Delta G_{(N3)}} \text{50S-AdeN3} + \text{H}$$

$$\Delta G_{(N1)} - \Delta G_{(N3)} = \Delta\Delta G$$

$$\Delta\Delta G = 2.3RT(pKa) \qquad (1)$$

Free energies from the adenosine in solution and in the ribosome were calculated to determine the effect the ribosome has on the electrostatic potential felt by the adenosine. Bases that pKa values would be calculated for were represented using the same charge and conformation inside the ribosome and outside in solution. The pKa values obtained for N1 and N3 on adenosine 2486 were then compared to other adenosines in the vicinity to establish that 2486 experiences an anomalous pKa shift. To prevent any undesired edge effects when comparing other adenosines, individual 15Å truncations were done for each adenosine being studied, 2485, 2488, and 2486, all of which are in the vicinity of the transition state analog puromycin[6].

In order to make this study tractable, a 15Å residue based sphere centered on A2486 was created from the newly determined bacterial *Haloarcula Marismortui* 50S crystal structure[16]. The introduced sphere was a residue-based truncation (Figure 5). Due to the nature of the truncation, fragments of RNA not connected in a chain were not capped at the 5' and 3' ends. This was done in order to mimic the effects of those particular residues not being at the end of a chain, therefore all nucleic acids, whether at chain ends or connected at the 3' or 5' ends had the same partial charge distribution. Protons were added to the ribosome crystal structure coordinates using Sybyl. Both the sheer magnitude of the structure and the lack of structural data for protein atoms within the ribonucleoprotein structure, contributed to necessitating truncation of the system. Additional studies utilizing a larger portion of the ribosome beyond 15 angstroms proved to have little effect on our calculated results (Data not shown).

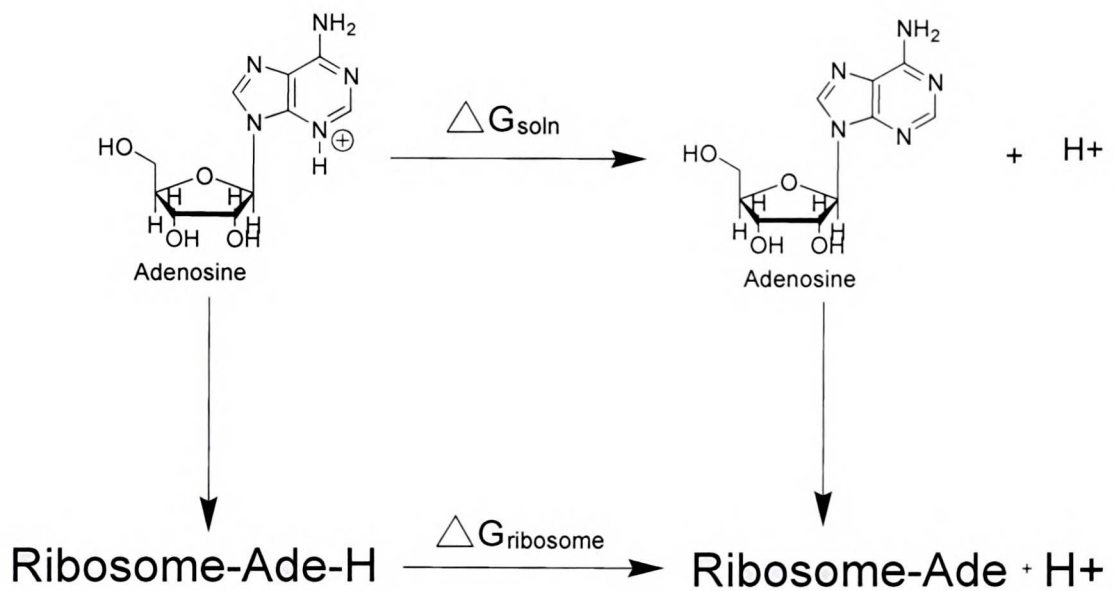


Figure 4. Thermodynamic cycle used to calculate pKa shifts. Top of equation represents adenosine protonation is solution and bottom represents protonation in the context of the ribosome.

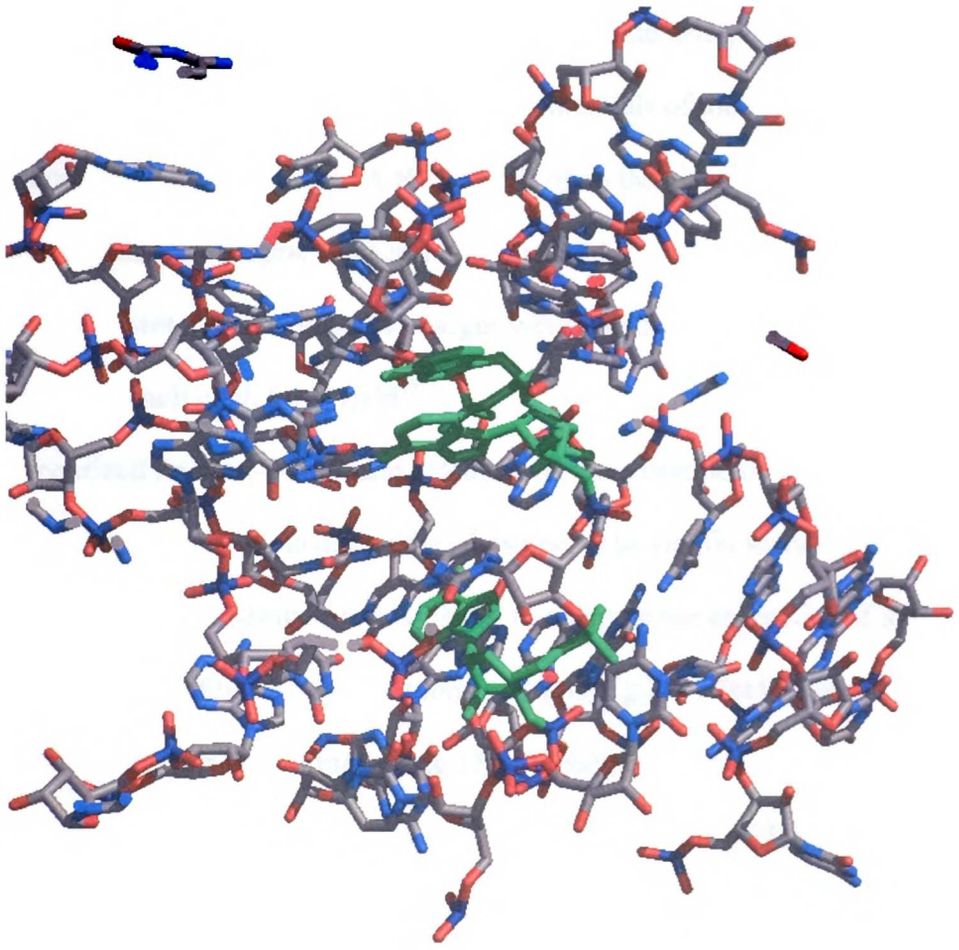


Figure 5. Cross section of ribosome active site showing 15Å sphere around A2486. For clarity, shown here is a zoomed in view showing only a portion of the 15Å sphere. Cutoff was residue based. Catalytic residues are shown in green.

Charges were derived for the protonated adenosine using Gaussian 98 software. Geometry optimized structures and electrostatic potentials of the atoms of adenosine were generated using HF/6-31G*. A RESP[17] fitting to the electrostatic potential was then performed to generate charges. This scheme was used for both N1 and N3 protonated adenosines. Unprotonated adenosine charges were obtained from a database of charges present in the Cornell et.al. force field[18].

Linearized Poisson-Boltzmann calculations were used to determine the energies of the adenosine inside and outside of the ribosome. The Delphi software package was used to solve the Poisson-Boltzmann equation to compare the energetics of ionization among states[19]. Use of non-linear iterations had a negligible effect on our calculations. Delphi calculations were performed using 150mM salt, an interior dielectric of 2, outer of dielectric of 80, and grid spacing of 0.5 angstrom. Alternative grid spacing of 1.0 and 0.25 were also seen to also have little effect on the results of these calculations (data not shown).

Results and Discussion:

Proton affinity calculations provide information into the amount of energy released when a proton is added to the system. It is computed as the energy difference for the species with an additional proton versus the energy of the system without[13]. Table 1 reveals the intrinsic proton affinity differences found on adenine. Of adenine's two protonation sites, N1 and N3, we determined that N1 is the first site to get protonated, in accord with biochemical data[9]. It has been known that adenosine titrates at 3.5 due to N1 protonation, while that of N3 has been roughly estimated to be 2 units lower[7]. Our proton

affinity calculations in gas and in solution using the HF/6-31G*, as well as the higher MP2, level of theory were able to further establish that the N3 position had a lower proton affinity by 2 kcal/mol versus N1 (Table 1), and that N7 was lower than both (data not shown). The lower proton affinity of 2 kcal/mol of N3 correlates to a pKa that is 1.5 units lower, thus roughly giving N3 an intrinsic pKa of 2.0. Being able to estimate the intrinsic pKa difference from proton affinity calculations, N3 has to encounter a pKa shift of at least 5 units higher in order for it to serve a similar role of histidines in serine proteases.

Table 1: Proton affinity values in kcal/mol for 9-methyladenine in gas phase and in solution. Values shown are relative to N1 sites. A larger value translates to a lower affinity for proton addition to that site. Calculations show that intrinsically N1 is the favored protonation site over N3 on adenine.

		HF/ 6-31G*	MP2/ cc-pvtz
Gas phase	N1	0.0	0.0
	N3	2.0	2.4
PCM	N1	0.0	0.0
	N3	1.8	1.5

Despite intrinsic proton affinity differences between N1 and N3 shown in table 1, our pKa calculations suggest that N3 on A2486 experiences a large enough pKa shift in the ribosome, making it the most likely candidate for proton addition. PKa calculations on adenosine 2486 showed that in the ribosome the N1 site experienced a weaker upward pKa shift than compared to N3. N3 experienced a pKa shift that, not only overcomes the intrinsic difference we saw via the proton affinity calculations, but also makes it a better

base for catalysis. The relative pKa shifts of nearby bases 2485 and 2488 are not to the same degree that A2486 experiences (Table 2). This unusually large pKa seen on A2486 is consistent with biochemical results[9] and supports the notion of it being the catalytic acid-base necessary for peptidyl transfer.

Recent mutagenic and biochemical studies have put into question the role and importance of this conserved adenosine in peptidyl transfer[20-22]. While in vivo studies have shown that mutations at A2486 produce a dominant lethal phenotype, Muth et.al. saw that in vitro, mutations to this conserved adenosine did not abolish catalytic activity. Recent studies have also hypothesized that the pKa shift seen in the ribosome active site is due to a pH dependent conformational change[1]. As seen from those studies, caution must be taken in assigning A2486 as the catalytic nucleotide in peptidyl transfer and deducing catalytic activity from the current conformation. The pKa values presented here are inferred from the current co-crystal structure of the 50S ribosome with puromycin and show that there exists an unusual electrostatic environment surrounding the N3 of this A2486. If the current crystal structure represents an active conformation, then our results suggest that this adenosine experiences an aberrant pKa shift making it a likely candidate for base catalysis or providing it the means to stabilize oxyanion formation during transition state structure development.

It has been suggested that a specific hydrogen bond network is responsible for this pKa shift on A2486[1]. Mutagenic data has shown that G2482 (2447 E.Coli numbering) is not essential for pKa perturbation of A2486[20-22]. This network includes Guanosine 2102 (2067 E.Coli numbering), which is within hydrogen bond distance to A2486:N1 in both apo forms and co-crystal structures with peptidyl transferase inhibitors[6]. Because of

G2102's close proximity to N1 on A2486, there is a question of how much of an effect its electrostatic potential has on the pKa of A2486. PKa shift calculations on A2486 without G2102 present showed that there still exhibited a substantial pKa shift of 5.5 units of N3 over N1 and suggests that the possible hydrogen bond to A2486:N1 by G2102 is not the sole reason for this anomalous pKa. This provides further evidence that this pKa shift is not due entirely to a specific hydrogen bond network with nearby bases.

Table 2: Relative pKa shifts of adenosine N3 relative to N1. A large value refers to N3 getting shifted more than N1 inside the ribosome. The larger the shift, the better N3 becomes for acid-base catalysis. Data obtained from FDPB calculations using Delphi as described in computational methods.

Adenosine	Relative pKa shift
2486	7.3
2485	-0.5
2488	2.8

We further investigated whether the negatively charged phosphate backbone was the main contributor to this pKa shift differential. According to our calculations, the backbone phosphates alone did not differentiate A2486 from other bases (data not shown), only when the entire charge distribution of the ribosome was included, was a larger pKa shift on A2486 observed.

ísco
SITY O
C
SITY O
L
Sa
OF CA
ísco
SITY OF
S
ITY OF
L
Sa
OF CAL
CO
OF CAL
ísco
SITY OF
S
SITY OF
L

Conclusions:

In Summary, proton affinity calculations have provided evidence that the intrinsic pKa of adenosine is due to the N1 site, and that N3 has a pKa roughly 1.5 units lower. Our preliminary pKa work on the adenosine then was able to establish that the electrostatic environment of the ribosome alters the N3 pKa on adenosine 2486 more than N1, thus providing evidence that proton abstraction to initiate catalysis originates from the N3 site and not the N1 site. Due to the necessity to truncate the large system, absolute pKa values of individual nucleic acids remains difficult to interpret, but relative pKa shifts can readily be inferred from the data.

Herein we have provided further evidence toward adenosine 2486 being a catalytic residue necessary for peptidyl transfer. In addition to showing that adenosine 2486 serves a role in peptidyl transfer, this work helps to clarify which nitrogen on the adenosine is the most likely site for proton addition. Ideally, conformational effects would aid in the interpretation of pKa values. Further work including molecular dynamics or Monte Carlo simulation methods is necessary to obtain an ensemble of structures to average pKa values. These additional studies that will look to support the findings in this paper can only be done when further structure work regarding proteins in the ribosome can be established.

The recent atomic structure of the *Haloarcula Marismortui* ribosome allows further mechanistic studies of ribosome function as well as establishes a starting point for structure-based drug design. With an improved understanding of how these prokaryotic ribosomes function, researchers can begin to further target these as ways to inhibit bacterial growth. Ultimately, by adding insight into the nature and mechanism of

bacterial protein synthesis, allows for the generation of better human therapeutics in the fight against microbial resistance to current antibiotics.

Acknowledgments:

KMM would like to thank Dr. Ray Luo for his involvement with this project. This work was supported by an NIH grants GM29072 (PAK) and by an NIH training grant through the University of California, San Francisco. I would also like to thank Dr. Bernd Kuhn and Dr. Matthew Lee for their helpful discussions.

References:

(1) Nissen, P.; Hansen, J.; Ban, N.; Moore, P. B.; Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **2000**, *289*, 920-930.

(2) Noller, H. F. Ribosomal Rna and Translation. *Annual Review of Biochemistry* **1991**, *60*, 191-227.

(3) Kurland, C. G. Structure and function of the bacterial ribosome. *Annual Review of Biochemistry* **1972**, *41*, 377-408.

(4) Moore, P. B. The ribosome returns. *Nature* **1988**, *331*, 223-227.

(5) Nissen, P.; Ippolito, J. A.; Ban, N.; Moore, P. B.; Steitz, T. A. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 4899-4903.

(6) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science* **2000**, *289*, 905-920.

(7) Saenger, W. *Principles of nucleic acid structure*; Springer-Verlag: New York, 1984; xx, 556.

(8) Nakano, S.; Chadalavada, D. M.; Bevilacqua, P. C. General acid-base catalysis in the mechanism of a hepatitis delta virus ribozyme. *Science* **2000**, *287*, 1493-1497.

(9) Muth, G. W.; Ortoleva-Donnelly, L.; Strobel, S. A. A single adenosine with a neutral pK(a) in the ribosomal peptidyl transferase center. *Science* **2000**, *289*, 947-950.

(10) Lawley, P. D.; Brookes, P. Further Studies on the Alkylation of Nucleic Acids and their Constituent Nucleotides. *Journal of Biochemistry* **1963**, *89*, 127-138.

(11) Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A. Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, g.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, Revision A7*: Pittsburgh, PA, 1998.

(12) Miertus, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chemical Physics* **1981**, *55*, 117-129.

(13) Foresman, J. B.; Frisch, A. *Exploring Chemistry with Electronic Structure Methods*; 2 ed.; Gaussian, Inc.: Pittsburgh, PA, 1993.

(14) Yang, A. S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. On the Calculation of Pk(a)S in Proteins. *Proteins-Structure Function and Genetics* **1993**, *15*, 252-265.

(15) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. The Determinants of Pk(a)S in Proteins. *Biochemistry* **1996**, *35*, 7819-7833.

(16) (1FFK), C. s. s. i. t. P. D. B.

(17) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. *Journal of Physical Chemistry* **1993**, *97*, 10269-10280.

(18) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179-5197.

(19) Sharp, K. A.; Honig, B. Electrostatic Interactions in Macromolecules - Theory and Applications. *Annual Review of Biophysics and Biophysical Chemistry* **1990**, *19*, 301-332.

(20) Polacek, N.; Gaynor, M.; Yassin, A.; Mankin, A. S. Ribosomal peptidyl transferase can withstand mutations at the putative catalytic nucleotide. *Nature* **2001**, *411*, 498-501.

(21) Thompson, J.; Kim, D. F.; O'Connor, M.; Lieberman, K. R.; Bayfield, M. A. et al. Analysis of mutations at residues A2451 and G2447 of 23S rRNA in the peptidyltransferase active site of the 50S ribosomal subunit. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 9002-9007.

(22) Bayfield, M. A.; Dahlberg, A. E.; Schulmeister, U.; Dorner, S.; Barta, A. A conformational change in the ribosomal peptidyl transferase center upon active/inactive transition. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 10096-10101.

Chapter 3: Investigation of Neuraminidase-Substrate Recognition Using Molecular Dynamics and Free Energy Calculations

Kevin M. Masukawa, Peter A. Kollman[†], Irwin D. Kuntz[*]

Department of Pharmaceutical Chemistry

Chemistry and Chemical Biology Program

University of California, San Francisco

600 - 16th Street, Box 2240

San Francisco, CA 94143-2240

[†] Deceased.

[*] To whom correspondence should be addressed. Tel: (415) 476-1937. Fax: (415) 502-1411. Email: kuntz@cgl.ucsf.edu.

Reproduced with permission from Masukawa, K. M., Kollman, P. A., et al. (2003). J Med Chem **46**(26): 5628-37. Copyright 2003 American Chemical Society.

Abstract:

Development of the new generation of therapeutics against the influenza viral coat protein neuraminidase is a response to the continuing threat of influenza epidemics. A variety of structurally similar compounds have been reported that vary greatly in their ability to inhibit neuraminidase, a critical enzyme that cleaves sialic acid and promotes virion release. To determine how neuraminidase exhibits this wide range of affinities with structurally similar compounds, molecular dynamics (MD) simulations, coupled with free energy calculations, were used to determine the binding components of a series of neuraminidase inhibitors. Using four co-crystal structures of neuraminidase-inhibitor complexes, we examined the structural and energetic components of ligand potency and selectivity. An in-depth energetic analysis, including internal energy, entropy, and non-bonded interactions, reveals that potency of ligand binding is governed by non-polar contacts. Electrostatic components generally oppose binding, although two of the best inhibitors use electrostatic interactions to orient the ligand. This investigation suggests that the enhanced selectivity and potency of the better ligands may arise from an improved positioning of their ligand atoms in the active site due to polar and hydrophobic functionalities. Simulations that included crystal water molecules in the active site indicate that the more potent ligands make less use of water-mediated interactions.

Introduction:

The viral disease Influenza afflicts millions of individuals each year. Current therapeutics, including both vaccines and drugs, have curtailed the deaths once associated with the virus, but despite these advances in treatment, the threat of resistance from antigenic shifts or drifts provokes the need for new treatment modalities. The influenza virus uses a pair of carbohydrate binding proteins, hemagglutinin and neuraminidase, to initiate viral fusion and subsequent budding of new virions from the infected cell.[1,2] Both of these glycoproteins are present on the surface of the influenza virus and are essential for virion propagation. Hemagglutinin recognizes target cells via sialic acid binding sites and then promotes viral fusion.[3] Neuraminidase cleaves terminal sialic acid moieties of virus progeny to promote the release and subsequent spreading of new virus particles.[4] Both glycoproteins have been suggested as therapeutic targets to prevent the spread of the influenza virus in the host.[5-7] Unlike hemagglutinin, neuraminidase has enzymatic constraints that reduce the frequency of resistant strains. To date, more success has been achieved designing compounds against neuraminidase than hemagglutinin.[8-14] Due to its role in the life cycle of the influenza virus, and its potential to prevent the spread of infection, neuraminidase is an appealing target.

Neuraminidase exhibits a high degree of selectivity among structurally similar compounds. Despite binding its natural ligand, sialic acid, with only millimolar affinity,[15,16] designed inhibitors of neuraminidase bind with a much higher affinity, sometimes reaching the low nanomolar range.[8,14,17,18] Figure 1 shows four inhibitors of neuraminidase exhibiting a broad range of binding affinities; the difference between the best binding ligand, Tamiflu, and the worst binding ligand, sialic acid, is over 12

kcal/mol. Currently, crystal structures of neuraminidase with these four inhibitors bound are available. Upon looking at these structures, their different recognition elements are not apparent; all four inhibitors bind in the same pocket and interact with the same set of residues. In order to exploit further the binding properties of neuraminidase for inhibitor design, a detailed analysis of how neuraminidase recognizes these structurally similar ligands with such large variation in binding affinity needs to be performed. In the past, the lack of structural data has forced researchers to use reduced representations of the neuraminidase binding site in attempts to elucidate the foundation of this binding affinity differential. These theoretical efforts, using pharmacophore[9] or protein structure-based models[19,20] used simplified scoring functions that neglected the contributions from internal conformational strain, entropy, or water interactions to the binding free energy. Considering the dynamic and amphipathic nature of carbohydrates, these missed contributions might be significant. Given the availability of four high-resolution X-ray co-crystal structures of neuraminidase with bound inhibitor, we decided to use advanced methods to quantify these interactions.

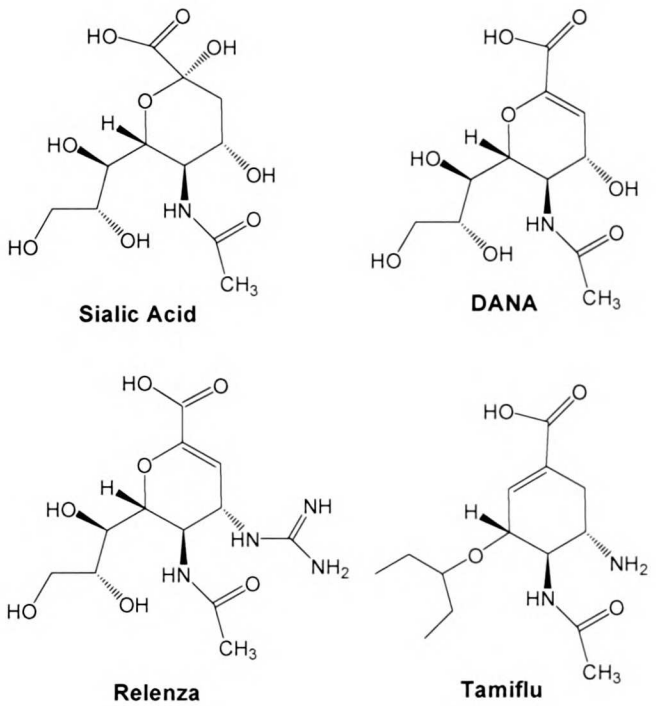


Figure 1. Neuraminidase ligands studied in this investigation. Experimental values given as IC_{50}'s, except for sialic acid.[8,16] Sialic Acid =6.1mM, DANA=10μM, Relenza=0.3-2.3nM, Tamiflu=0.01-2.2nM .

Here we report a detailed analysis of neuraminidase-inhibitor interactions using the computational technique developed in its current form in the Case and Kollman laboratories known as Molecular Mechanic Poisson-Boltzmann Surface Area (MM/PBSA). This method, like others,[21,22] utilizes a combination of molecular dynamics simulations and free energy calculations to predict free energy of a system based on its conformation. MM/PBSA has had great success explaining receptor-ligand interactions at the atomic level.[23-26] Molecular dynamics permits probing structural questions beyond what is possible from direct observation of the static crystal structure model alone. In

this study, molecular dynamics and MM/PBSA calculations were performed on a series of sialic acid analogs bound to neuraminidase (Figure1). The purpose of this investigation is to determine which thermodynamic quantities drive complexation and how these contributions differ among different ligands. The information gained from this analysis could lead to the discovery of new inhibitors with improved binding properties. Using this computational technique we also investigate the role structural waters play in mediating these interactions.

Computational Methodology

Abbreviations. DANA, 2-deoxy-2,3-didehydro-N-acetylneuraminic acid (Neu5Ac2en); MM/PBSA, Molecular Mechanics Poisson-Boltzmann Surface Area; Relenza, 2,4-dideoxy-2,3-didehydro-4-guanidino sialic acid (zanamivir); Tamiflu, ethyl-4-acetamido-5-amino-3-(1-ethylpropoxy)-1-cyclohexene-1-carboxylate (oseltamivir)[a]

System Setup. Atomic coordinates for all co-crystal complexes were obtained from the Protein Data Bank (PDB). Raw crystal structures were modified for dynamics using the LEAP module in Assisted Model Building with Energy Refinement (AMBER 6.0).[27] Four different crystal structures of neuraminidase bound to different inhibitors were used as starting structures in this study. The inhibitors were sialic acid (1MWE[28]), DANA (1F8B[29]), Relenza (1NNC[30]), and Tamiflu (2QWK[31]). All structures are co-crystal

[a] Note Tamiflu is the ethylester of ethyl 4-acetamido-5-amino-3-(1-ethylpropoxy)-1-cyclohexene-1-carboxylate, our study focuses on the active form of Tamiflu, the carboxylate, interacting with neuraminidase.

complexes of the avian influenza virus A, subtype N9 at pH 7.0. In the LEAP module of AMBER, protons were added to the system. In accord with crystallographic conditions of the complexes, all ionizable side chains were configured in their characteristic ionized states at pH 7.0. All histidines were singly protonated at the delta nitrogen. The binding site of neuraminidase has been seen to accommodate crystal waters when binding different ligands. Active site structural waters present in the crystal structure were not removed from the coordinate file if they lie between the ligand and protein. This treatment was justified by observing that solvent exposed waters diffuse into bulk solvent on our simulation time scale while buried waters do not. No restraints were placed on the crystallographic water molecules during the simulations. Sequestered water molecules were free to move, but did not diffuse out of the pocket over the length of the simulation. Directionality of the hydrogen bonds was random during the initial setup but was allowed to optimize during the equilibration. Structures were then solvated with a water cap (TIP3P)[32] out to 30 Angstroms from the inhibitor while maintaining crystal water positions. The dielectric constant in all simulations was set to 1 because explicit water molecules were added. To test the structural effects of these crystal water molecules, simulations were also performed with no water molecules within the protein-ligand interface. In those cases, all crystal waters were removed before the addition of cap water molecules, which were themselves excluded from being between the protein and ligand.

Atomic partial charges for the four ligands were derived for this study using the RESP method.[33] To obtain minimized geometries for electrostatic potential calculations, ligand geometries were first optimized with Gaussian98 using the Hartree-Fock (HF)/6-

31G* level of theory.[34] Single-point calculations with Gaussian98 were then performed to obtain the electrostatic potential around each compound using the same basis set and level of theory as in the optimization step. Fitting charges to the electrostatic potential was then performed with RESP. Sialic acid and DANA were both given a formal charge of -1 while Relenza and Tamiflu had 0. Equivalent atoms were given equal partial charges. Amino acid charges came from the RESP derived AMBER94 database.

Simulations. Molecular dynamics (MD) produces a time trajectory of the system by solving Newton's equations of motion for each atom. All MD simulations were performed using the SANDER module in the AMBER 6.0 suite of programs. An all-atom representation of the system was used employing the Cornell et.al.[35] force field to assign parameters for the standard amino acids. A "belly" region of 17Å was defined around the ligand. During the simulations these "belly" atoms, which are less than 17 Å from the ligand, were allowed to move, while all atoms beyond 17Å were held rigid. Structures were first allowed to relax under steepest decent minimization for 10 cycles, followed by a conjugant gradient minimization for 4990 cycles. Upon minimization, the water cap was equilibrated for 10 picoseconds (ps) to allow cavities in the protein to become solvated. The entire Belly region was then equilibrated for 100ps with the system gradually warmed to 300K over the initial 50ps. Production MD was performed for a nanosecond (ns) with a 2fs time step and the SHAKE algorithm[36] holding all bond lengths fixed. During the simulation, non-bonded cutoffs were set to 12Å, the temperature was maintained at 300K.

Due to the flexible nature of some of the ligands, it is incorrect to assume that the ligand conformations are the same in the bound state versus free in solution. The solution

conformations of the ligands were determined with separate simulations. Initial ligand conformations were taken directly from the co-crystal structures with neuraminidase. The ligands were then minimized and equilibrated using the same number of cycles used to optimize the complexes. Upon equilibration, ligands were simulated for 1ns at 300K in a box with 8.0Å of water surrounding each ligand. The conformations generated from this set of simulations were used to calculate energies associated with ligands free in solution.

Conformational Analysis of Sialic Acid. The conformational space inherent to carbohydrates is large, and theoretical studies on carbohydrates have been rare. Thus, a necessary first step is to establish that the available force fields are adequate to evaluate and discern distinct conformers of flexible carbohydrate rings. To ensure proper modeling, we continued parameterization and validation of the Cornell et.al. force field for monosaccharide-protein systems by comparing theoretical molecular mechanics and ab initio energy differences for sialic acid boat to chair conformations (Figure 2). Ab initio calculations of the boat/chair conformational states for a derivatized sialic acid, one of the more flexible neuraminidase binding compounds, yielded an energy difference of 6.8 kcal/mol favoring the chair conformation. A molecular mechanics score using the AMBER potential energy function yielded an energy difference of 8.9 kcal/mol for these two conformers. This discrepancy exists mainly because parameters for the C-C-O-C torsion were derived in the past from work on acyclic ether compounds. After optimization of torsional parameters specifically for this system, this initial energy difference between quantum and molecular mechanics energies was minimized to 0.86

kcal/mol. Figure 3 shows the analytical fit of our new torsion parameters to the actual sampled in our torsion drive simulation.

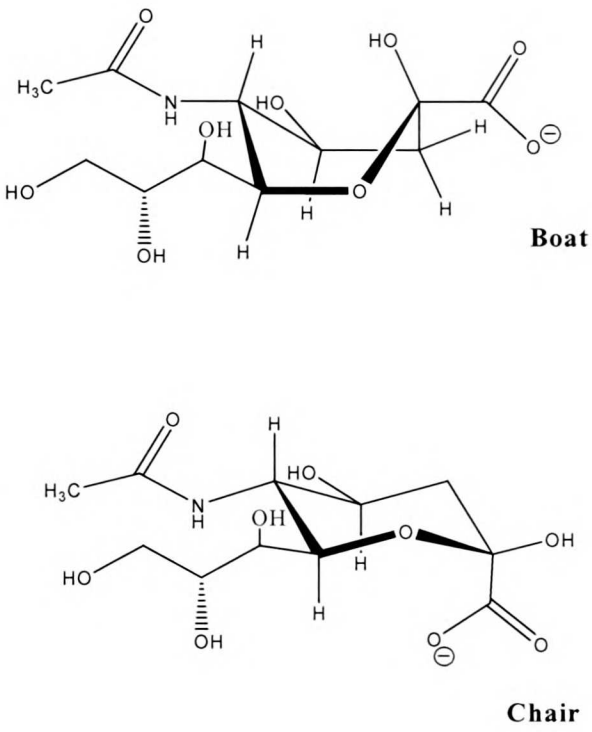


Figure 2 : Sialic Acid boat/chair conformations

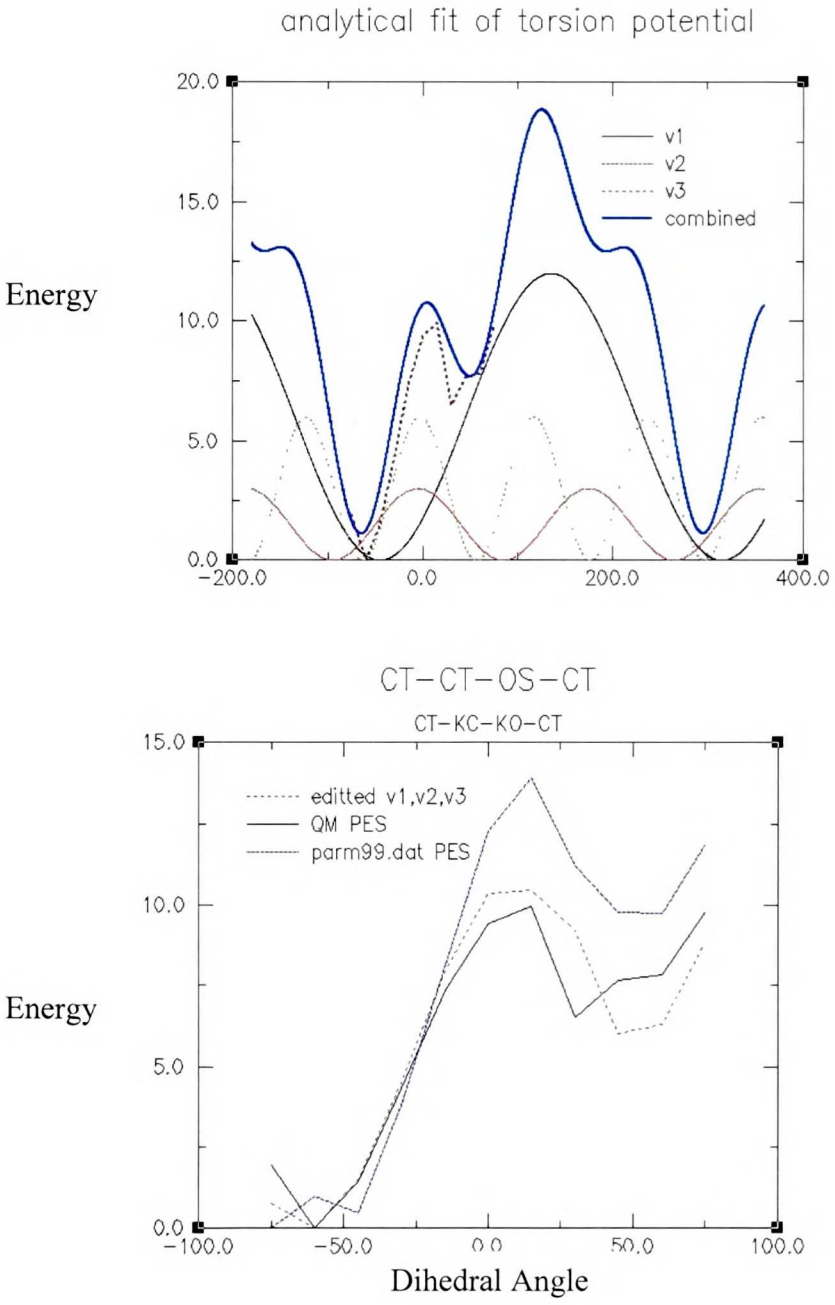


Figure 3. Top - Analytical fit of the torsional potential for the C-C-O-C bond in pyranose. Blue represents the potential resulting from a linear combination of the v1, v2, and v3 torsion potentials shown in black, red, and green, respectively. Energy is in kcal/mol. Bottom – new torsion potential versus old PARM99.

Along with energy differences between boat/chair conformations we also were interested in being able to properly sample both conformations in solution with an appropriate distribution of the two forms. Initial MD studies on sialic acid for 150ps show that this carbohydrate exhibits proper behavior in solution using our newly developed parameters (Figure 4). Starting from a twist-boat conformation, sialic acid over the course of the simulation switches to the chair conformer, which is the most stable energy conformation, consistent with ab initio results. Nonetheless, these preliminary simulations provide early evidence that the force field developed by Cornell et al.[35] with additional glucose parameters from Woods et al. will be able to mimic carbohydrate tendencies in solution. All ligands in this study containing a pyranose ring, sialic acid, DANA, relenza, were subject to the same parameters discussed above.

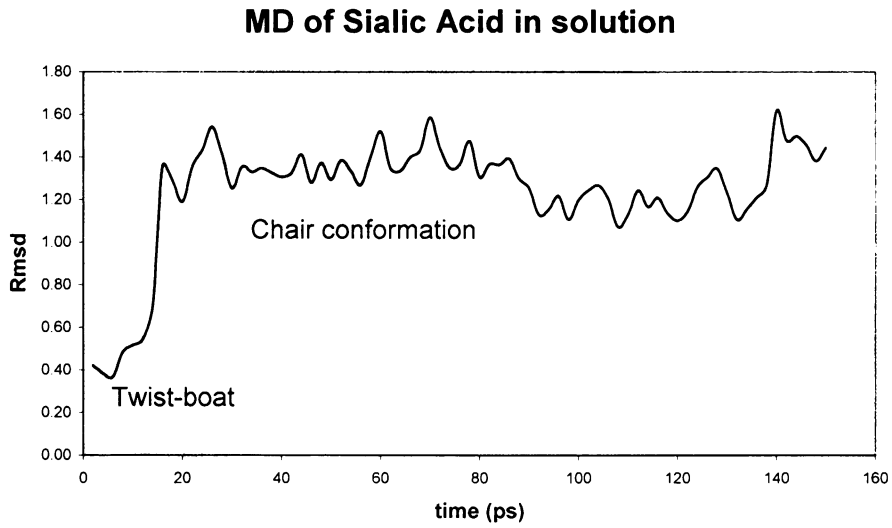


Figure 4: All-atom RMSD over 150ps referenced to minimized structure. Sialic acid starting conformation is twist-boat taken from the co-crystal structure with neuraminidase.

Energetic Analysis. Utilizing the trajectory generated by molecular dynamics, free energy calculations were performed using a Molecular Mechanics/ Poisson-Boltzmann Surface Area (MM/PBSA) scoring function to determine binding affinities as well as component energy contributions. Free energies of binding were determined by calculating internal energy changes upon binding using the AMBER program with the Cornell et.al. force field coupled with a continuum solvent model to evaluate free energies of solvation. An estimate of the absolute free energy of binding, ΔG_{Bind}, is obtained by taking the difference of the complex energy minus the separate reactants' energies.

$$\Delta G_{Bind} = \overline{G}_{complex} - (\overline{G}_{receptor} + \overline{G}_{ligand}) \qquad (1)$$

The average free energy of the complex, receptor, or ligand used in the above equation is composed of the molecular mechanic, solvation, and entropic energies of the system over the trajectory and is represented by the following equation.

$$\overline{G} = \overline{E}_{MM} + \overline{G}_{Solv} - T\overline{S} \qquad (2)$$

$$\overline{E}_{MM} = \overline{E}_{Bond} + \overline{E}_{Angle} + \overline{E}_{Dihedral} + \overline{E}_{VDW} + \overline{E}_{Elec} \qquad (3)$$

E_{MM} is the molecular mechanical energy obtained from bonded and non-bonded interactions within the system. G_{solv} represents the electrostatic and non-polar free

energy of solvation and TS is the solute entropic contribution to the system at temperature T(Kelvin). All energies represented in the above equations are averaged over the course of the molecular dynamics trajectories.

The continuum model for the free energy of solvation uses a finite-difference Poisson-Boltzmann approach to calculate the electrostatic energy, coupled with a surface area dependent term for the non-polar contribution to solvation. Ultimately, the molecular dynamics trajectory is split into "snapshots", stripped of all non-crystal water molecules, and evaluated individually using this scoring scheme. In the energetic analysis, crystallographic waters were considered as an extension of the protein. Using snapshots from an extended simulation to generate binding energies will allow us not only to sample the flexibility of the binding site, but also to obtain a more reliable free energy estimate of binding than compared to a single snapshot calculation.

Entropic contributions to binding were determined using a normal mode approximation.[37-39] The change in solute entropy upon association was estimated by calculating normal mode frequencies using the NMODE module within AMBER6.0. For each complex five snapshots from the MD trajectories were taken at 100ps intervals and a region defined by a 17Å sphere from the inhibitor was used for the entropy analysis. Each snapshot was minimized using a distant-dependent dielectric, 4r, before normal mode analysis. Solvent entropies of the system are implicitly included in the non-polar and polar solvation free energy terms.

The binding free energy was calculated by taking the difference in MM/PBSA energy for the ligand-protein complex and uncomplexed reactants. From the 1 ns simulations, snapshots were collected every 10ps. Ligated and apo protein conformations

were obtained from the simulations of the complex, while the unbound ligand conformations were obtained from a separate trajectory, as described above. MM/PBSA scores for each of these snapshots were averaged to obtain a free energy of association. The molecular mechanic (MM) energy was obtained using the Cornell et.al. force field as implemented in AMBER with no cutoffs. The Poisson-Boltzmann solvation energy was calculated with the DELPHI program[40] using a grid spacing of 0.5Å, 80% fill of the grid box, and an exterior dielectric of 80. While the exact value of the interior dielectric constant remains to be unequivocally established, a recent MM/PBSA study in protein structure prediction showed an improved correlation with experiment when the protein interior was modeled with a dielectric of 4 versus 1.[41] This result coupled with the solvent accessibility of the neuraminidase active site and the calculations of Jedrezjas et. al.[15] support the use of an interior dielectric constant of 4 for our study. Non-polar contributions to solvation were estimated as a function of the solvent-accessible surface area (SASA). The solvent-accessible surface area is calculated using the MSMS program[42] and the non-polar contribution to the free energy of binding is given by

$$\Delta G_{solv}^{non-polar} = \gamma(SASA) + b \qquad (4)$$

SASA is the solvent-accessible surface area and γ and b are constants. The values for γ and b were derived experimentally from the transfer of small hydrocarbons by Sitkoff et.al.[43] and are 0.00542 kcal/mol·A^2 and 0.92 kcal/mol, respectively.

Computational Alanine-Scanning. From the wild-type trajectory, snapshots were generated every 10ps for alanine-scanning. Mutations to alanine were performed

only on selected residues in the active site. Alanine mutations were generated by truncation of residues after the Cβ and adding a hydrogen in the same direction as the Cγ. Partial charges for the mutated residue were then changed to those of alanine. None of the residues mutated in this study were glycines.

Results and Discussion

Energetic Analysis. Using a combined MD/MMPBSA approach, our calculations show an excellent correlation with experimental results (Table 1). Table 1 shows the binding free energy components averaged over the MD trajectory for each co-crystal complex. Along with discriminating the worst ligand, our calculations also agree with the experimental finding that Tamiflu is the best binding ligand. All ligands are correctly rank-ordered and the magnitudes of the ΔGs are in good agreement with experiment.

If we examine the contributions to binding across the ligand series, there is an increase in the van der Waals contribution as the interaction becomes more favorable. This appears to explain the affinity progression because while all other non-bonded binding components remain relatively constant across the series, the tightly bound ligands have added non-polar packing in the active site. The better binding Tamiflu and Relenza gain over 6 kcal/mol more van der Waals energy over their less potent counterparts, DANA and sialic acid. Interestingly, the presence of the large ether moiety in Tamiflu does not increase its overall van der Waals interaction energy compared to that of Relenza. As will be discussed below, our simulations with crystal waters suggest that the ether moiety may facilitate better positioning of Tamiflu in the active site. The importance of non-polar interactions in targeting neuraminidase was also identified in previous

investigations.[10,12,44] Further analysis of the binding components reveals that while the van der Waals and non-polar solvation energies drive binding for all four ligands, sterics alone do not completely explain the affinity differential.

We also find that the overall electrostatic component to binding either provides no thermodynamic benefit or even opposes association despite the large number of polar groups in the active site. Of the four compounds studied, the only ligand that has a favorable electrostatic component to binding is Relenza (ΔE_{Elec} = -0.26kcal/mol). It appears that Relenza gains this electrostatic energy by having a large positive guanidinium group that interacts favorably with E227 at the bottom of the active site. Favorable internal electrostatic energy is gained as one proceeds from a hydroxyl (DANA) off the C4 position to an ammonium (Tamiflu), and guanidinium group (Relenza). However, the inclusion of desolvation effects leads to a net loss of binding except for Relenza. The importance of including desolvation penalties when calculating binding affinities to this receptor is reinforced by work done by Smith et.al. who reported that the addition of charge to various DANA analogs was deleterious to binding.[29] For example, it costs more for DANA in electrostatic energy to bind to neuraminidase than sialic acid, 3.31 kcal/mol versus 0.49 kcal/mol (Table 1). This is most likely due to DANA's lack of flexibility that precludes optimal charge complementarity in the active site to offset desolvation penalties, as evident by DANA's coulombic interaction energy being less than that for sialic acid, -35.15 versus -44.91 (Table 1). Where others have

Table 1. Average Energy Contributions (kcal/mol) to the Free Energy of Binding for Neuraminidase and the Set of Four Ligands.[a]

Contribution	Tamiflu	Relenza	DANA	Sialic Acid
ΔE internal[b]	-6.07 (0.51)	1.90 (0.68)	-5.36 (0.53)	5.16 (0.48)
ΔE VDW[c]	-29.16 (0.48)	-29.99 (0.51)	-23.13 (0.56)	-23.32 (0.47)
ΔE elect, internal[d]	-44.49 (0.52)	-58.49 (0.39)	-35.15 (0.53)	-44.91(0.28)
ΔG elect,solv[e]	46.81 (0.42)	58.22 (0.27)	38.46 (0.39)	45.4(0.21)
ΔG nonpolar, solv[f]	-4.46 (0.01)	-4.65 (0.01)	-4.12 (0.01)	-4.13 (0.01)
ΔE Elec, int+solv[g]	2.31 (0.17)	-0.26 (0.22)	3.31 (0.24)	0.49 (0.21)
-TΔS[h]	21.6 (1.35)	19.6 (1.56)	22.4 (1.64)	20.11 (1.47)
ΔG bind[i]	-15.78 (1.49)	-13.4 (1.63)	-6.90 (1.70)	-1.69 (1.58)
bind experiment[j]	**-15.2 to -12.0**	**-13.2 to -11.9**	**-6.91**	**-3.06**

[a] Obtained from trajectories that included crystal waters. Errors are given as Standard Errors of the Mean.
[b] Internal contributions from bond, angle, and dihedral terms
[c] Non-bonded van der Waals
[d] Non-bonded electrostatics
[e] Non-polar and electrostatic components to solvation
[f] Electrostatic components to solvation
[g] Total electrostatic change upon binding
[h] Entropic contributions to binding
[i] Total change of free energy in binding
[j] Reference 8,16. Experimental data given as IC_{50}s for all ligands except sialic acid. For direct comparison to calculated affinities, conversion to ΔG estimated by $\Delta G = -RT \ln IC_{50}$.

noted the importance of a positively charged group off the C4 position,[20,45] our work suggests that the benefit from adding these groups might not be solely from electrostatic interactions. The fact that Tamiflu is the most potent binder, despite Relenza having more favorable non-bonded interactions, suggests that internal strain energy is also an important factor in increased affinity and selectivity, as shown below.

Energetic Analysis of Free Ligands. Our studies of the free ligands in separate trajectories provide an estimate for the local adaptation energy upon complex formation. Along with internal changes in van der Waals, coulombic, and solvation energy, free ligand trajectories enable an estimation of the ligand strain upon binding. Ligand strain is defined as the difference in bond, angle, and dihedral terms between the conformation in the bound state and the conformation free in solution. Sialic acid contains a pyranose ring, which due to its flexible nature can populate many conformations in solution that vary greatly in energy. In the neuraminidase-sialic acid co-crystal structure, sialic acid is bound in a boat/twist-boat conformation, which is higher in energy than its chair form. Upon binding, the internal energy associated with the conformational change in the pyranose ring of sialic acid is calculated to cost roughly 5.0 kcal/mol in binding free energy. This penalty is much larger than compared to the other three inhibitors. The unfavorable strain energy upon binding is a major component of why sialic acid has a lower affinity for neuraminidase than DANA; the difference experimentally is roughly 2.5 kcal/mol for the pyranose ring only.[46] With the inclusion of the double bond in the pyranose ring, DANA does not experience a large conformational strain increase upon binding compared to its conformation in the free unbound form. Disregarding energetic terms associated with the ligands free in solution, both DANA and sialic acid have

similar calculated interaction energies, –26.95 and –27.17 kcal/mol, respectively (data not shown).

Just as DANA and sialic acid have strain energies to discern the two interactions, the same type of thermodynamic component also differentiates Tamiflu and Relenza. Tamiflu has a reduction in strain energy of –6.07 kcal/mol upon binding where Relenza pays 1.90 kcal/mol upon association. A further investigation reveals that much of the costs in internal energy is due to Relenza straining the acetyl group conformation upon binding. Tamiflu and the other two weaker binders do not experience a large strain on the acetyl group like Relenza does (Table 2). As will be discussed further, the strain on Relenza can be attributed to its dependence on waters for positioning in the active site.

Table 2. Internal Strain Energy Change (kcal/mol) of Ligands Upon Binding Neuraminidase.

	Acetyl group	Central ring
Sialic Acid	0.12	1.69
DANA	-0.66	-2.57
Relenza	4.59	-0.84
Tamiflu	-0.81	-1.85

Negative changes in strain energy upon complex formation are a result of the ligand binding in a conformation that is more favorable in the protein than outside. One explanation is that the enzyme site is preorganized to favor the bound geometry. Alternatively, Tamiflu's ammonium group may contribute to its strained geometry in solution. The high charge density maintains a close interatomic distance to the carbonyl

of the acetyl group in solution. Upon binding this distance is increased by 0.2Å on the average, lowering the strain on the ligand. Relenza, with the larger positively charged guanidinium group, does not induce a strained conformation in solution. Once bound, this guanidinium group shifts to charge complement E227, reducing the distance between the guanidinium and acetyl group adding strain on the ligand. DANA which also experiences a negative strain energy upon binding, undergoes a similar conformational strain reduction to Tamiflu. Interestingly, sialic acid encounters the same type of shift in distance, but perhaps its flexible ring prevents a large strain from being incurred in solution. As mentioned above, sialic acid's overall unfavorable binding strain is mainly contributed from the large ring flip induced upon binding. Due to the charged nature and large number of hydrogen bond donors and acceptors of these ligands, strained conformations in solution are not unexpected.

Given the large dimensionality of conformational space available to each ligand, a true estimate for the adaptation energy requires a higher sampling technique. The energetic contributions from the free ligand conformations generated through MD simulations are most likely overestimates of the underlying adaptation energy. Although local, inclusion of this ligand conformational energy using MD simulations allows an estimate for the adaptation energy each ligand experiences upon binding, regardless of the exact molecular mechanism. Clearly, these ligands in solution must be treated as an average over an ensemble of the conformations.

Role of Crystal Water Molecules. Carbohydrate binding proteins commonly use water molecules to mediate their interactions with ligands; the role these waters play has yet to be fully elucidated.[47-50] In each of the four co-crystal structures, at least three

crystallographic water molecules are seen trapped between neuraminidase and its inhibitor. To investigate their role in these recognition events, we performed a series of simulations with and without these crystal waters present in both the MD simulations and post-processing. By performing a set of simulations without water molecules between the protein and ligand, we can isolate structural perturbations in the active site that are directly related to the absence of these waters. This investigation revealed that with three of the four ligands, crystal waters are necessary to stabilize the ligand orientation in the active site (Figure 5). The RMSD plots of the ligand positions during MD simulations with and without crystal waters suggest that the weaker binders are more dependent on waters to mediate their interactions than the best binder, Tamiflu (Figure 5). In both sialic acid and DANA, the central ring shifts position in the absence of crystal waters, which alters the placement of some functional groups relative to their positions when crystal waters are included.

Relenza also experiences a minor displacement with crystal waters removed by undergoing a different type of movement compared to sialic acid or DANA. Upon inspection, it is apparent that much of the displacement of Relenza arises from the glycerol group shifting to compensate for water not being present (Figure 6). The guanidinium group, on the other hand, is situated deep in the active site and occupies the same space irrespective of water molecules being present or absent. Thus, Relenza appears to partially overcome crystal water dependence by having the positively charged guanidinium group anchor the entire ligand in the active site. Comparing the bound ligand conformations between simulations performed with and without crystal waters, an added strain of 1.02, 1.04, and 2.94 kcal/mol on the acetyl group, central pyranose ring, and glycerol group, respectively was seen when

including crystal waters. This conformational strain is necessary to establish favorable contacts with the protein. When simulations of Relenza were performed without crystal waters, the non-bonded van der Waals and electrostatic interaction energies are not as favorable (data not shown). The better binding Tamiflu rectifies this problem by not having a large conformational dependence on crystal water molecules, thus preventing a large unfavorable internal energy change upon binding.

Tamiflu, unlike the other ligands, samples the same binding position regardless of crystal water molecules after about 400ps of simulation. The ammonium and ether groups of Tamiflu act as anchors to secure the ligand in the active site reducing its dependence on water (Figure 7). While Relenza's glycerol group shifts significantly, Tamiflu's additional ether group lends stronger packing of Tamiflu into the active site. Our results suggest that Tamiflu's functional groups improve binding by orienting the ligand to take advantage of contacts within the active site and more importantly help to avoid the added internal strain upon binding. It appears that a low-energy conformation of Tamiflu complements the active site better than the highly strained conformation of Relenza.

Recognizing that the free energy of association is a combination of many terms, the loss in electrostatic energy by having charged groups is more than compensated for by gains in other molecular mechanic terms. This is evident by the fact that Relenza and Tamiflu, which both have an additional positively charged group, have negligible or slightly positive electrostatic terms that oppose binding, yet are the two best binders. Work by Chervenak and Toone has shown that the binding of carbohydrates in aqueous

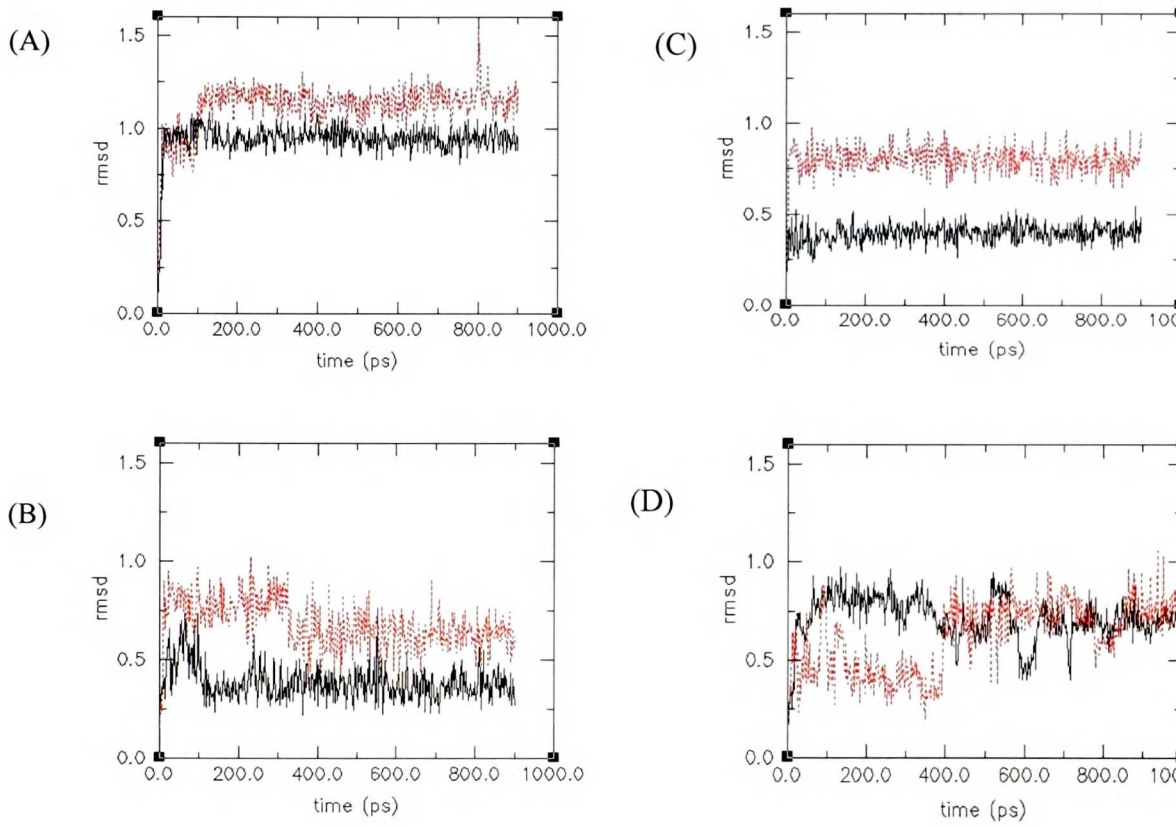


Figure 5. All atom RMSD plots of ligand positions as compared to the initial minimized crystal structures. From top to bottom, (A)sialic acid, (B)DANA, (C)Relenza, (D)Tamiflu. Dashed lines are simulation without crystal waters and solid lines are from simulations that included crystal waters.

Figure 6. C_α Overlay of snapshots from Relenza MD trajectory taken with and without crystal waters. White is conformation obtained from simulation with crystal waters, and grey is from simulation without crystal waters. Both snapshots are taken from 550ps into the simulation.

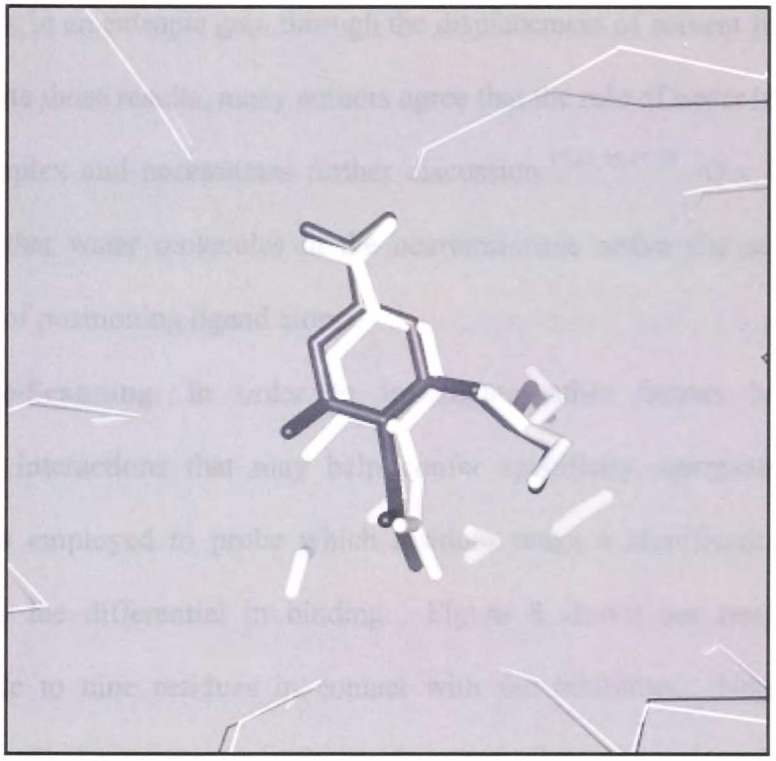


Figure 7. C_α Overlay of snapshots from Tamiflu MD trajectory taken with and without crystal waters. White is conformation obtained from simulation with crystal waters, and grey is from simulation without crystal waters. Both snapshots are taken from 550ps into the simulation.

solutions results in an entropic gain through the displacement of solvent from the bindin pocket.[51] Despite those results, many authors agree that the role of water in protein-ligan binding is complex and necessitates further discussion.[19,44,45,47,52] Our work presente here suggests that water molecules in the neuraminidase active site serve an equall important role of positioning ligand atoms.

Alanine-Scanning. In order to investigate other factors besides interna intramolecular interactions that may help confer specificity, computational alanine scanning[53] was employed to probe which residues make a significant intermolecula contribution to the differential in binding. Figure 8 shows our results for alanin mutations made to nine residues in contact with the inhibitors. Negative energeti changes (ΔG_{wt}-ΔG_{ala}) represent an unfavorable substitution. As expected, we see that i general, mutations of active site residues are highly unfavorable with all four inhibitor However, three residues in the active site are predicted to contribute differentially i binding to these ligands.

Alanine mutations to E119, D151, and E277 show the largest variance in fre energy of binding suggesting that they play a role in substrate selectivity. As seen fro Figure 8, while E119 contributes only slightly to binding sialic acid and DANA, Tamifl and Relenza are able to form much more favorable interactions. Mutagenic studies hav supported the notion of E119 playing a role in substrate selectivity, where inserting a alanine at this position has shown a 1.17 kcal/mol reduction in binding to DANA whil experiencing a 3.84 kcal/mol loss in affinity to Relenza[54](Table 3). This is in accord wit our calculations that show an E119A mutant only alters DANA binding by 0.83 kcal/mo compared to wildtype, while Relenza's affinity is weakened by 5.0 kcal/mol (Figure 8

Our calculations further suggest that this differential is due to the positive electrostatic potential of the extra functional groups on Tamiflu and Relenza acting as an anchor for the negative E119 buried in the active site pocket. Sialic acid and DANA do not have the extra positive charge off the C4 position. In similar fashion, D151 also helps to confer specificity by establishing stronger interactions with the better binding ligands. The data shows that D151 is not used in binding sialic acid or the structurally similar DANA, but can be exploited with the more potent inhibitors. Ghate found that the K_i for DANA to an D149E (B/Lee/40 numbering) mutant versus wild-type is of the same magnitude, 35uM and 53uM, respectively.[55] A deeper investigation into the energetic components reveals that the gain in free energy from D151 is from an increased coulombic energy associated with the positively charged groups on Tamiflu and Relenza interacting with the negative electrostatic potential of the aspartic acid. While there remains a slight electrostatic penalty when binding Relenza, it is still lower than DANA or sialic acid. E277 contributes almost 4 kcal/mol more free energy when binding Tamiflu or Relenza than it does with the other ligands. This glutamate is located on the floor of the binding pocket not in the vicinity of the C4 substituents off the central ring. One possible explanation for the increased interaction, despite being distant from the ammonium or guanidinium groups, could be that the added charge on Tamiflu and Relenza helps to provide their other functional groups better contacts with this residue.

Table 3. Experimental Resistance Inhibition Data for an E119A Mutant.

		IC_{50} (fold)[a]	
	Neu type	DANA	Relenza
Wildtype	A/Minnesota/80	1	1
E119A	A/Minnesota/80	7	600
$\Delta\Delta G_{E119A}$		-1.17 kcal/mol	-3.84 kcal/mol

[a] Reference 54.

Although this study suggests E119, D151, and E277 help confer specificity, th
are other residues that are important for establishing potent interactions with all
ligands. R118, R224, and R292 contribute equally to all four ligands, provic
significant free energy to these complexation events. The participation of R118
R292, both of which surround the common carboxylate moiety on the ligands
expected. R224 is positioned between the acetyl and glycerol functional groups of
ligands and provides electrostatic and van der Waals free energy to the comple
Conversely, Y406 provides a marginal free energy gain upon binding for all four liga
This residue is implicated to be essential for activity of neuraminidase becaus
phenylalanine mutation at this position shows no activity in fluorescence assays.[55,56]
agreement, results from this study suggest that Y406 may play more of a kinetic
substrate positioning role in catalysis as opposed to a thermodynamic role in bind
This result implies that despite its necessity, targeting Y406 in attempts to incre
affinity of potential drug candidates may be futile.

As mentioned earlier, our studies suggested that the tighter binding of Tam
and Relenza is explained by stronger van der Waals interactions. From the indivic

non-polar contributions listed in Figure 8B, it is difficult to discern which set of resi is responsible for this large van der Waals increase. While E276, which is below ether moiety, appears to take advantage of the increased non-polar character of Tam no set of residues explains the over 6 kcal/mol van der Waals interaction en difference between strong and weak ligands. E276 experiences a reorganization c side chain upon binding Tamiflu. Through our alanine-scanning, it appears that rotation allows neuraminidase more van der Waals interactions with Tamiflu compare the other three ligands that do not induce a rotation of E276. The fact that this rotati already present in our starting structures suggests that we have captured some of energetics associated with E276's rotation, but note that without complete samplin the free receptor, a true quantitative explanation for the energetic effects of E276 rota is difficult to draw. As evident in figure 8, the dispersed nature of van der Waals con suggests that targeting a specific hydrophobic residue or pocket to gain selectivity ma difficult. In contrast, while electrostatics, as a whole, fails to segregate ligands, alan scanning has shown that individual interactions from residues may be differentiated as a result can be further targeted to give rise to differences in binding energy (Fi 8C). Alanine-scanning of the neuraminidase active site with the four ligands uncovered that the main residues contributing to selectivity upon binding are E D151, and E277, and suggests that additional mutations at either D151 or E277 w mitigate binding of Tamiflu or Relenza.

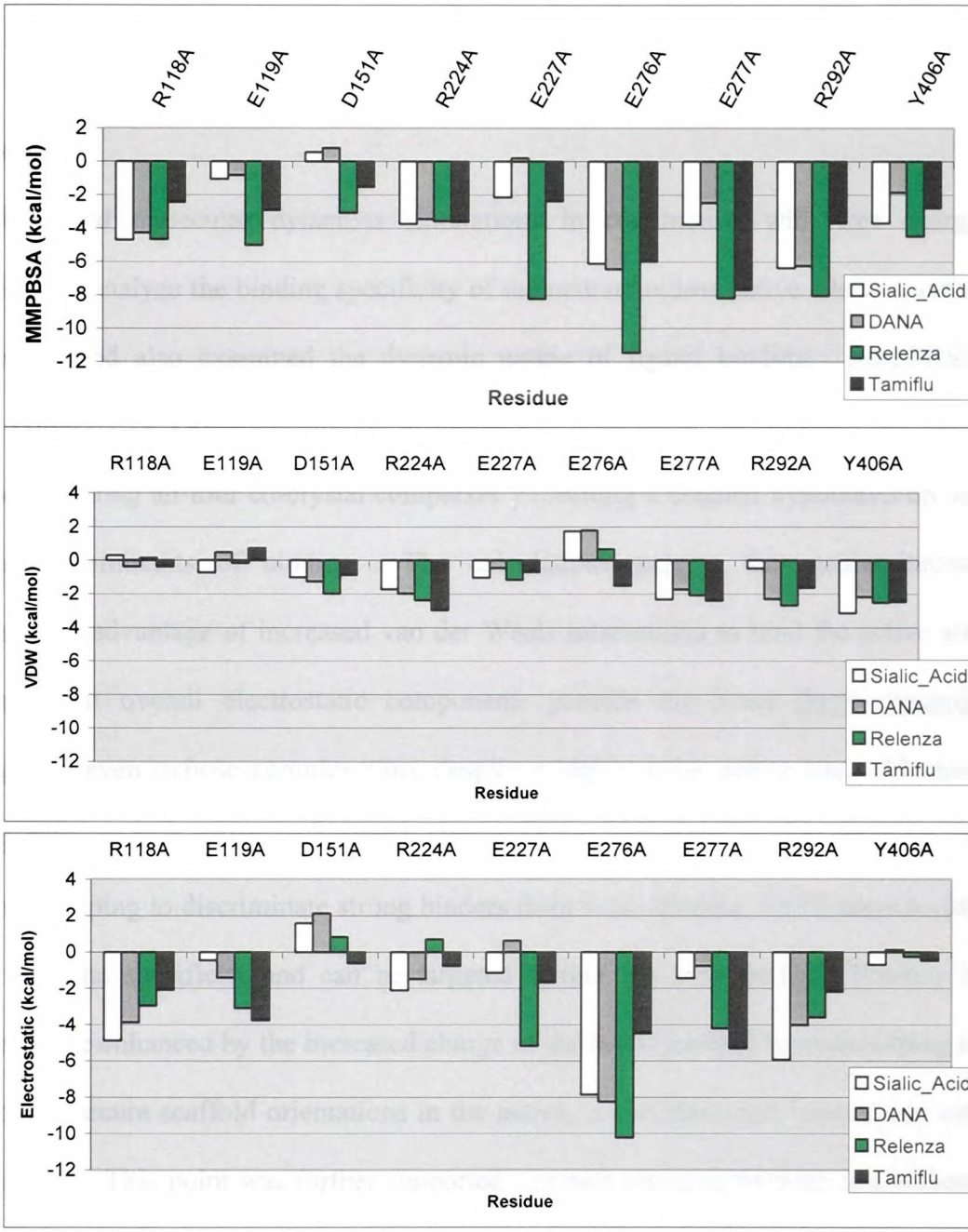


Figure 8. Computational alanine-scanning results for neuraminidase with th four ligands. Shown are the relative free energy (kcal/mol) changes obtaine from the mutated snapshots averaged over the MD trajectories. Negative fre energies ($\Delta\Delta=\Delta_{wildtype}-\Delta_{mutant}$) represent an alanine mutation that is unfavorable t binding. (A)$\Delta\Delta G_{binding}$; (B)$\Delta\Delta G_{vdw}$; (C)$\Delta\Delta G_{electrostatic}$.

Conclusions

We used molecular dynamics simulations in conjunction with free energy calculations to analyze the binding specificity of the neuraminidase active site to a series of inhibitors and also examined the dynamic nature of ligand binding. MM/PBSA analysis correctly rank-ordered binding affinities and segregated the different energetic components among all four co-crystal complexes permitting a detailed hypothesis on the structural determinants of binding. The calculations suggest that neuraminidase inhibitors take advantage of increased van der Waals interactions to bind the active site better and that overall electrostatic components provide no direct thermodynamic advantage, or even oppose complexation, despite a highly polar active site. Alanine-scanning suggests that while electrostatic effects do not contribute to strengthening the affinity nor helping to discriminate strong binders from weak binders, it still plays a vital role in substrate specificity and can be targeted further for drug design. Potency is suggested to be enhanced by the increased charge of the better binding ligands helping to establish and secure scaffold orientations in the active site so that other interactions can be maximized. This point was further supported through simulations with and without crystal water molecules in the active site, that showed that the lack of extra charged or hydrophobic groups caused a larger dependence on water for scaffold positioning in the active site. Recently, others have also noted the importance of a consistent scaffold binding mode for compounds that target neuraminidase.[10] Although electrostatics as a whole provides no net gain to binding, alanine-scanning revealed that an energy

differential was present when considering interactions of each ligand with individual charged residues.

This study exemplifies and stresses the importance of considering crystal water molecules in a recognition process. Tamiflu's high affinity is largely explained by its direct binding-site complementarity without the need for mediating water molecules, in contrast to its weaker binding counterparts. Given that many other carbohydrate-binding protein crystal structures have been solved with crystallographic waters present, inclusion of these waters is both feasible and necessary for understanding the interactions in order to develop potent inhibitors.

While this study was able to suggest ways neuraminidase recognizes a series of inhibitors, the power of this technique is its use as an analytical tool and cannot be used in a predictive manner without the implementation of higher sampling techniques. Higher sampling techniques would afford the broader search of the energy landscape and a closer estimate to the energy of each binding partner. The location of water molecules remains an issue in computational biology and hinders predictive studies of neuraminidase recognition.

Acknowledgments. This work is in memory of Peter A. Kollman who succumbed to cancer on May 25, 2001. K.M.M. appreciates the extra guidance he has received from Prof. David Case. K.M.M. also thanks both Dr. David Sullivan and Prof. David Case for helpful comments on the manuscript. This work was supported through NIH grant (GM-29072) to P.A.K. and I.D.K.

Supporting Information Available: Structures, partial charges, AMBER atom types, and additional force field parameters used for each ligand. This material is available free of charge via the Internet at http://pubs.acs.org.

References:

(1) Sears, P.; Wong, C. H. Carbohydrate mimetics: A new strategy for tackling the problem of carbohydrate-mediated biological recognition. *Angewandte Chemie-International Edition* **1999**, *38*, 2301-2324.

(2) Paulson, J. C. The Receptors. In *The Receptors*; Conn, P. M. Ed.; Academic Press: Orlando, Fla., 1984; pp 131-219.

(3) Huang, R. T.; Rott, R.; Klenk, H. D. Influenza viruses cause hemolysis and fusion of cells. *Virology* **1981**, *110*, 243-247.

(4) Colman, P. M. The Influenza viruses. In *The Viruses.*; Krug, R. M. Ed.; Plenum Press: New York, 1989; pp 175-218.

(5) Bodian, D. L.; Yamasaki, R. B.; Buswell, R. L.; Stearns, J. F.; White, J. M. et al. Inhibition of the fusion-inducing conformational change of influenza hemagglutinin by benzoquinones and hydroquinones. *Biochemistry* **1993**, *32*, 2967-2978.

(6) Bamford, M. J. Neuraminidase inhibitors as potential anti-influenza drugs. *Journal of Enzyme Inhibition* **1995**, *10*, 1-16.

(7) Weinhold, E. G. Design and Evaluation of a Tightly Binding Fluorescent Ligand for Influenza A Hemagglutinin. *Journal of the American Chemical Society* **1992**, *114*, 9270-9275.

(8) Babu, Y. S.; Chand, P.; Bantia, S.; Kotian, P.; Dehghani, A. et al. BCX-1812 (RWJ-270201): Discovery of a novel, highly potent, orally active, and selective influenza neuraminidase inhibitor through structure-based drug design. *Journal of Medicinal Chemistry* **2000**, *43*, 3482-3486.

(9) Wang, G. T.; Chen, Y. W.; Wang, S.; Gentles, R.; Sowin, T. et al. Design, synthesis, and structural analysis of influenza neuraminidase inhibitors containing pyrrolidine cores. *Journal of Medicinal Chemistry* **2001**, *44*, 1192-1201.

(10) Stoll, V.; Stewart, K. D.; Maring, C. J.; Muchmore, S.; Giranda, V. et al. Influenza neuraminidase inhibitors: structure-based design of a novel inhibitor series. *Biochemistry* **2003**, *42*, 718-727.

(11) Gubareva, L. V.; Kaiser, L.; Hayden, F. G. Influenza virus neuraminidase inhibitors. *Lancet* **2000**, *355*, 827-835.

(12) Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Cobley, K. N. et al. Novel inhibitors of influenza sialidases related to GG167 - Structure-activity, crystallographic and molecular dynamic studies with 4H-pyran-2-carboxylic acid 6-carboxamides. *Bioorganic & Medicinal Chemistry Letters* **1996**, *6*, 2931-2936.

(13) White, C. L.; Janakiraman, M. N.; Laver, W. G.; Philippon, C.; Vasella, A. et al. A sialic acid-derived phosphonate analog inhibits different strains of influenza virus neuraminidase with different efficiencies. *J Mol Biol* **1995**, *245*, 623-634.

(14) Vonitzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C. et al. Rational Design of Potent Sialidase-Based Inhibitors of Influenza Virus Replication. *Nature* **1993**, *363*, 418-423.

(15) Jedrzejas, M. J.; Singh, S.; Brouillette, W. J.; Air, G. M.; Luo, M. A Strategy for Theoretical Binding Constant, K-I, Calculations for Neuraminidase Aromatic Inhibitors Designed on the Basis of the Active Site Structure of Influenza Virus Neuraminidase. *Proteins* **1995**, *23*, 264-277.

(16) Potier, M.; Mameli, L.; Belisle, M.; Dallaire, L.; Melancon, S. B. Fluorometric assay of neuraminidase with a sodium (4-methylumbelliferyl- alpha-D-N-acetylneuraminate) substrate. *Anal Biochem* **1979**, *94*, 287-296.

(17) Finley, J. B.; Atigadda, V. R.; Duarte, F.; Zhao, J. J.; Brouillette, W. J. et al. Novel aromatic inhibitors of influenza virus neuraminidase make selective interactions with conserved residues and water molecules in the active site. *Journal of Molecular Biology* **1999**, *293*, 1107-1119.

(18) Chand, P.; Kotian, P. L.; Dehghani, A.; El-Kattan, Y.; Lin, T. H. et al. Systematic structure-based design and stereoselective synthesis of novel multisubstituted cyclopentane derivatives with potent antiinfluenza activity. *Journal of Medicinal Chemistry* **2001**, *44*, 4379-4392.

(19) Wang, T.; Wade, R. C. Comparative binding energy (COMBINE) analysis of influenza neuraminidase-inhibitor complexes. *Journal of Medicinal Chemistry* **2001**, *44*, 961-971.

(20) Vonitzstein, M.; Dyason, J. C.; Oliver, S. W.; White, H. F.; Wu, W. Y. et al. A Study of the Active Site of Influenza Virus Sialidase - an Approach to the Rational Design of Novel Anti-Influenza Drugs. *Journal of Medicinal Chemistry* **1996**, *39*, 388-391.

(21) Jayaram, B.; Sprous, D.; Young, M. A.; Beveridge, D. L. Free Energy Analysis of the Conformational Preferences of a and B Forms of DNA in Solution. *Journal of the American Chemical Society* **1998**, *120*, 10629-10633.

(22) Vorobjev, Y. N.; Almagro, J. C.; Hermans, J. Discrimination Between Native and Intentionally Misfolded Conformations of Proteins - ES/IS, A New Method for

Calculating Conformational Free Energy That Uses Both Dynamics Simulations With an Explicit Solvent and an Implicit Solvent Continuum Model [Review]. *Proteins* **1998**, *32*, 399-413.

(23) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H. et al. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research* **2000**, *33*, 889-897.

(24) Chong, L. T.; Duan, Y.; Wang, L.; Massova, I.; Kollman, P. A. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 14330-14335.

(25) Huo, S. H.; Wang, J. M.; Cieplak, P.; Kollman, P. A.; Kuntz, I. D. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: Insight into structure-based ligand design. *Journal of Medicinal Chemistry* **2002**, *45*, 1412-1419.

(26) Gouda, H.; Kuntz, I. D.; Case, D. A.; Kollman, P. A. Free energy calculations for theophylline binding to an RNA aptamer: MM-PBSA and comparison of thermodynamic integration methods. *Biopolymers* **2003**, *68*, 16-34.

(27) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E. et al. Amber, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Computer Physics Communications* **1995**, *91*, 1-41.

(28) Varghese, J. N.; Colman, P. M.; van Donkelaar, A.; Blick, T. J.; Sahasrabudhe, A. et al. Structural evidence for a second sialic acid binding site in avian influenza virus neuraminidases. *Proc Natl Acad Sci U S A* **1997**, *94*, 11808-11812.

(29) Smith, B. J.; Colman, P. M.; Von Itzstein, M.; Danylec, B.; Varghese, J. N. Analysis of inhibitor binding in influenza virus neuraminidase. *Protein Science* **2001**, *10*, 689-696.

(30) Varghese, J. N.; Epa, V. C.; Colman, P. M. Three-dimensional structure of the complex of 4-guanidino-Neu5Ac2en and influenza virus neuraminidase. *Protein Sci* **1995**, *4*, 1081-1087.

(31) Varghese, J. N.; Smith, P. W.; Sollis, S. L.; Blick, T. J.; Sahasrabudhe, A. et al. Drug design against a shifting target: a structural basis for resistance to inhibitors in a variant of influenza virus neuraminidase. *Structure* **1998**, *6*, 735-746.

(32) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **1983**, *79*, 926-935.

(33) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. *Journal of Physical Chemistry* **1993**, *97*, 10269-10280.

(34) Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A. Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.;

Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, g.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, Revision A7*: Pittsburgh, PA, 1998.

(35) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179-5197.

(36) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977**, *23*, 327-341.

(37) Levy, R. M.; Karplus, M.; Kushick, J.; Perahia, D. Evaluation of the Configurational Entropy for Proteins: Application to Molecular Dynamics Simulations of an a-Helix. *Macromolecules* **1984**, *17*, 1370-1374.

(38) Hayward, S.; Kitao, A.; Go, N. Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins* **1995**, *23*, 177-186.

(39) Case, D. A. Normal Mode Analysis of Protein Dynamics. *Current Opinion in Structural Biology* **1994**, *4*, 285-290.

(40) Sharp, K. A.; Honig, B. Electrostatic Interactions in Macromolecules - Theory and Applications. *Annual Review of Biophysics and Biophysical Chemistry* **1990**, *19*, 301-332.

(41) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. Molecular Dynamics in the Endgame of Protein Structure Prediction. *Journal of Molecular Biology* **2001**, *313*, 417-430.

(42) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced Surface - an Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38*, 305-320.

(43) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *Journal of Physical Chemistry* **1994**, *98*, 1978-1988.

(44) Wall, I. D.; Leach, A. R.; Salt, D. W.; Ford, M. G.; Essex, J. W. Binding constants of neuraminidase inhibitors: An investigation of the linear interaction energy method. *Journal of Medicinal Chemistry* **1999**, *42*, 5142-5152.

(45) Taylor, N. R.; Vonitzstein, M. A Structural and Energetics Analysis of the Binding of a Series of N-Acetylneuraminic-Acid-Based Inhibitors to Influenza Virus Sialidase. *Journal of Computer-Aided Molecular Design* **1996**, *10*, 233-246.

(46) Smith, B. J. A Conformational Study of 2-Oxanol - Insight into the Role of Ring Distortion on Enzyme-Catalyzed Glycosidic Bond Cleavage. *Journal of the American Chemical Society* **1997**, *119*, 2699-2706.

(47) Clarke, C.; Woods, R. J.; Gluska, J.; Cooper, A.; Nutley, M. A. et al. Involvement of water in carbohydrate-protein binding. *Journal of the American Chemical Society* **2001**, *123*, 12238-12247.

(48) Swaminathan, C. P.; Surolia, N.; Surolia, A. Role of water in the specific binding of mannose and mannooligosaccharides to concanavalin A. *Journal of the American Chemical Society* **1998**, *120*, 5153-5159.

(49) Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology* **1996**, *3*, 973-980.

(50) Toone, E. J. Structure and Energetics of Protein Carbohydrate Complexes. In *Current Opinion in Structural Biology*, 1994; pp 719-728.

(51) Chervenak, M. C.; Toone, E. J. A Direct Measure of the Contribution of Solvent Reorganization to the Enthalpy of Ligand Binding. *Journal of the American Chemical Society* **1994**, *116*, 10533-10539.

(52) Smith, B. J.; McKimm-Breshkin, J. L.; McDonald, M.; Fernley, R. T.; Varghese, J. N. et al. Structural studies of the resistance of influenza virus neuramindase to inhibitors. *Journal of Medicinal Chemistry* **2002**, *45*, 2207-2212.

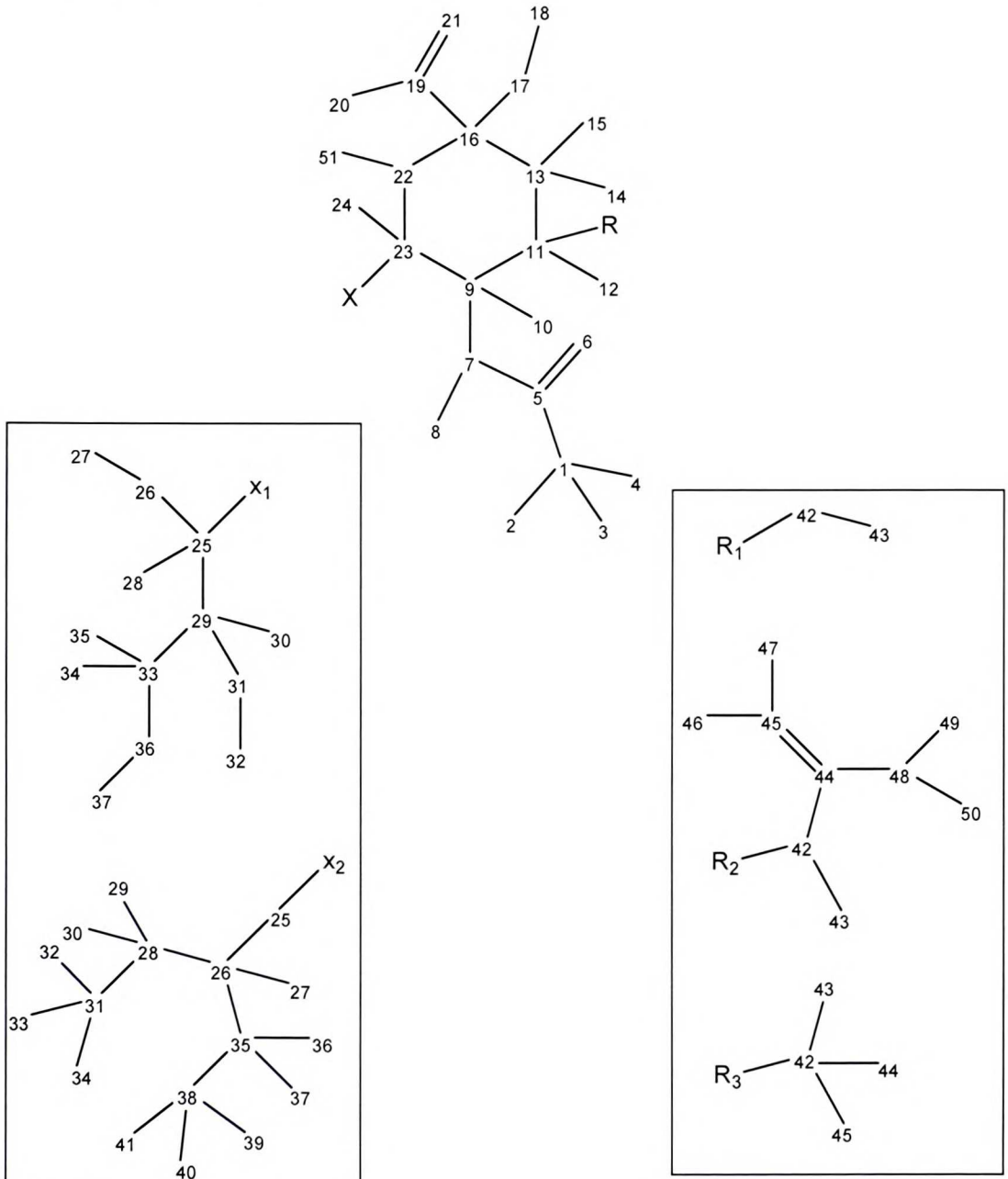
(53) Massova, I.; Kollman, P. A. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *Journal of the American Chemical Society* **1999**, *121*, 8133-8143.

(54) Gubareva, L. V.; Robinson, M. J.; Bethell, R. C.; Webster, R. G. Catalytic and framework mutations in the neuraminidase active site of influenza viruses that are resistant to 4-guanidino-Neu5Ac2en. *Journal of Virology* **1997**, *71*, 3385-3390.

(55) Ghate, A. A.; Air, G. M. Site-directed mutagenesis of catalytic residues of influenza virus neuraminidase as an aid to drug design. *European Journal of Biochemistry* **1998**, *258*, 320-331.

(56) Janakiraman, M. N.; White, C. L.; Laver, W. G.; Air, G. M.; Luo, M. Structure of Influenza Virus Neuraminidase B/Lee/40 Complexed with Sialic Acid and a Dehydro Analog at 1.8-Angstrom Resolution - Implications for the Catalytic Mechanism. *Biochemistry* **1994**, *33*, 8172-8179.

Supporting Information:



X1,R1 = Sialic Acid, DANA
X1,R2 = Relenza
X2,R3 = Tamiflu

atom	AMBER atom type	Sialic Acid charge	type	DANA charge	type	RELENZA charge	type	TAMIFLU charge
1	CT	-0.25312	CT	-0.31833	CT	-0.36963	CT	-0.20521
2	HC	0.06765	HC	0.08863	HC	0.12225	HC	0.06445
3	HC	0.06765	HC	0.08863	HC	0.12225	HC	0.06445
4	HC	0.06765	HC	0.08863	HC	0.12225	HC	0.06445
5	C	0.73315	C	0.71079	C	0.62663	C	0.74367
6	O	-0.63246	O	-0.6306	O	-0.60843	O	-0.62742
7	N	-0.60946	N	-0.5599	N	-0.50502	N	-0.68184
8	H	0.31186	H	0.31646	H	0.32574	H	0.36644
9	CT	0.04986	CT	-0.01309	CT	0.1131	CT	0.03689
10	H1	0.13491	H1	0.15287	H1	0.13643	H1	0.18504
11	CT	0.2002	CT	0.44969	CT	0.07542	CT	0.01438
12	H1	0.13554	H1	0.04349	H1	0.15024	HP	0.1144
13	CT	-0.10803	CM	-0.48333	CM	-0.4035	CT	0.01076
14	HC	0.07547	HA	0.16881	HA	0.18313	HC	0.05347
15	HC	0.07547					HC	0.05347
16	CT	0.18686	CM	0.08705	CM	0.11663	CM	-0.09128
17	OH	-0.64919						
18	HO	0.41434						
19	C	0.79383	C	0.84614	C	0.78287	C	0.80611
20	O	-0.767	O	-0.77246	O	-0.72391	O	-0.75597
21	O	-0.767	O	-0.77246	O	-0.72391	O	-0.75597
22	OS	-0.31529	OS	-0.22994	OS	-0.19604	CM	-0.27017
23	CT	-0.0178	CT	-0.05511	CT	0.0021	CT	0.05423
24	H1	0.12558	H1	0.08879	H1	0.11829	H1	0.10053
25	CT	0.0909	CT	0.21925	CT	0.1083	OS	-0.32214
26	OH	-0.59216	OH	-0.71642	OH	-0.7452	CT	0.03809
27	HO	0.37491	HO	0.44939	HO	0.48246	H1	0.04956
28	H1	0.12848	H1	0.05331	H1	0.06974	CT	-0.07787
29	CT	0.15077	CT	0.36366	CT	0.15276	HC	0.04897
30	H1	0.00556	H1	-0.00755	H1	0.03091	HC	0.04897
31	OH	-0.66892	OH	-0.68101	OH	-0.64323	CT	-0.06435
32	HO	0.42549	HO	0.42682	HO	0.42228	HC	0.03413
33	CT	0.2588	CT	0.15133	CT	0.2321	HC	0.03413
34	H1	0.04714	H1	0.00998	H1	0.0214	HC	0.03413
35	H1	0.04714	H1	0.00998	H1	0.0214	CT	-0.07787
36	OH	-0.73941	OH	-0.68853	OH	-0.66875	HC	0.04897
37	HO	0.43982	HO	0.42162	HO	0.42071	HC	0.04897
38							CT	-0.06435
39							HC	0.03413
40							HC	0.03413
41							HC	0.03413
42	OH	-0.71405	OH	-0.74358	N2	-0.39742	N3	-0.29807
43	HO	0.42485	HO	0.43692	H	0.31581	H	0.2943
44					CA	0.77445	H	0.2943
45					N2	-0.91105	H	0.2943
46					H	0.43938		
47					H	0.43938		
48					N2	-0.91105		
49					H	0.43938		
50					H	0.43938		
51							HA	0.1885

additional force field parameters

MASS

BOND

ANGLE

OH – CT - C	50.0	109.50
OS – CT - OH	50.0	109.50
CM - CT - OH	60.0	109.5
OS - CM - C	80.0	125.00
CM - CT - N2	50.0	123.20

DIHEDRAL

CT-CT-OS-CT	1	0.065	-10.0	-3.
CT-CT-OS-CT	1	1.750	-10.0	-2
CT-CT-OS-CT	1	3.0	135.0	1

Chapter 4: Estimating Chemical Similarity Expectation Values for Novel Compounds

David C. Sullivan, Kevin M. Masukawa, and Irwin D. Kuntz

Department of Pharmaceutical Chemistry

Chemistry and Chemical Biology Program

University of California, San Francisco

600 -16th Street, Box 2240

San Francisco, CA 94143-2240

Abstract:

Chemical "similarity" is a topic of considerable interest to pharmaceutical chemists in their effort to understand the size and complexity of the universe of chemical compounds. As a result, methods to help characterize public or proprietary databases of small molecules are pressing. We present a novel method for estimating the significance of intermolecular differences (similarities) using a cumulative probability density function motivated by a hyper-sphere model. In addition to estimating chemical similarity significance, our formulation of the problem facilitates extracting the intrinsic number of chemical descriptors, effective at a given similarity value, from many strongly inter-correlated molecular descriptors. We find that small molecule chemical space, as represented by a diverse set of compounds from the Available Chemicals Directory (ACD) is composed of ~11-24 orthogonal dimensions in the limit of high similarity. On average, about 10^{-17} - 10^{-14} of chemical space is at least 0.85 similar (by the Tanimoto measure) to any given compound, while only 10^{-28} - 10^{-23} lies within a 0.95 similarity radius. We discuss our chemical space model in the context of quantifying the relevance of activity neighborhood behavior and validating chemical descriptor sets.

Introduction:

Over the past decade, the use of virtual screening has risen dramatically. Virtual screening, or in-silico screening, encompasses numerous techniques in computational biology, but all are based on identifying compounds from a large database that fit a model. In the context of the work in this chapter, screening is defined as the pairwise comparison of thousands of compounds in attempts to identify molecules exhibiting similar biological activity. Virtual high-throughput screening (HTS) enables a large number of compounds to be evaluated in a matter of days. Whether it is to seed an initial drug discovery effort with a lead compound, expand a set of starting compounds for increasing synthetic options, improving affinity, or evading intellectual property constraints, virtual screening is viewed as an invaluable tool. Virtual screening provides information about the underlying chemical and geometrical makeup of a compound via two dimensional descriptors and compares it to a known active compound. The "neighborhood principle", which states that structurally similar compounds have similar biological activity[1], motivates computational screening based on chemical similarity to an existing lead compound. The measure of chemical similarity involves two steps. First, a chemical descriptor set must be chosen. There are several, including atom-connectivity based fingerprints[2] and substructure 'keys'[3]. Second, given two compounds appropriately projected into a descriptor space, a similarity measure must be chosen. Again, there are several,[4] including the Tanimoto coefficient and Euclidean distance.

More important than the raw scores from similarity analysis is the significance of the scores. This paper directly addresses that question. We do so by the standard approach of mapping a similarity score to the probability of two random compounds having an equivalent or better (more similar) score. The cumulative distribution function, which

can be constructed from a set of random compounds with calculated (dis)similarities, expresses this expectation probability. At least two complications require consideration. First, the high-similarity distribution is most interesting, however random compounds may not sample chemical space sufficiently fine to define this region of the distribution. Hence, assumptions must be introduced concerning distribution behavior in order to model and thus extrapolate the distribution function. To approach this issue, we use a hyper-sphere motivated statistical description that we previously applied to protein conformational space.[5] The second significant obstacle is constructing an appropriate random compound set. From the perspective of molecular design, we are most interested in properties of potential (i.e. synthesizable) small molecules as opposed to existing (i.e. purchasable) compounds. By this reasoning, we pose the question: If all 10^{60} to 10^{400} potential small, drug-like compounds[6] could be enumerated, would a small random subset (e.g. $<10^6$) have a chemical-property distribution similar to current compound databases? Our intuition tells us that real compound databases are significantly more clustered than this idealized random compound set, owing to historical biases; namely, existing compounds form the synthetic building blocks for new compounds. For our calculations, we work around this second obstacle to generating an appropriate cumulative distribution function by using a diverse set of compounds from the Available Chemical Directory (ACD)[7].

Validating a chemical descriptor set is generally performed by testing its ability to judge compounds with known similar biological activity as structurally 'similar' with a low false positive rate. In principle, performance statistics could be extracted directly from data sets with known activities. In practice, probing narrow definitions of activity (e.g. strong binding affinity) may require seeding a database with known active

compounds (i.e. strong binding inhibitors) which were developed by a directed optimization process which imparts an unknown amount of information. Thus, these seeded data points confound any statistical interpretations because the true-positive rate has no general meaning. Our work recognizes that the significance of similarities between known-active compounds can be determined directly by using an externally derived distribution function.

In this paper, we present a novel method for assessing the significance of chemical similarity values by considering the radial distribution of compounds in chemical space, where chemical space is defined by the compound set considered, the chemical descriptor set, and the distance metric used. For a diverse compound set, we extrapolate the distribution to high, unobserved similarity values, providing any similarity value a corresponding expectation value. Further, similarity values calculated with different descriptor sets can be discussed directly with respect to expectation. Ultimately, the information imparted from this analysis can help to evaluate which chemical descriptor sets best segregate known active compounds from random compounds.

Computational Methods:

Chemical descriptors

We use the Daylight fingerprints, which are based on the two-dimensional atom-connectivity molecular representation and describe atom connectivity patterns.[2] The bit-string length, l, limits the number of patterns that can be encoded. These molecular descriptors characterize a compound in a binary fashion with a value of 1 given to indicate the presence of some feature and 0 in its absence. The effect of the bit-string length on the apparent dimensionality is explored in this study by varying the descriptor

set from 128 through 2048 bits. Daylight fingerprints can be "folded" such that individual bits correspond to multiple patterns, but this is not done here.

Chemical similarity and difference measures

We investigate the behavior of several distance functions commonly used in the chemical diversity literature,[4] using the formulae for dichotomous variables as appropriate:

The normalized Hamming distance:

$$D_{A,B}^{Hamming} = [a + b - 2c] / l \quad (1)$$

The normalized Euclidean distance:

$$D_{A,B}^{Euclidean} = [(a + b - 2c) / l]^{1/2} \quad (2)$$

The Soergel distance (which is equivalent to the complement of the Tanimoto coefficient):

$$D_{A,B}^{Soergel} = 1 - c / [a + b - c] \quad (3)$$

The complement of the Dice coefficient:

$$D_{A,B}^{cDice} = 1 - 2c / [a + b] \quad (4)$$

The complement of the Cosine coefficient:

$$D_{A,B}^{cCosine} = 1 - c / [ab]^{1/2} \quad (5)$$

For two bit-strings, A and B, a is the number of 'on' (value 1) bits in A, b is the number of 'on' bits in B, c is the number of 'on' bits common to both A and B, and l is the

number of descriptors (the bit-string length). For the commonly used Tanimoto coefficient, a value of 1 indicates maximum similarity.

Compound set

We generated a diverse set of compounds from the 1998 release of the ACD by the following method: (1) Randomize the order of the compounds. (2) Generate fingerprints (l=128) for the database. (3) Relative to the first compound, remove all compounds below on the list less than 35 bits different (i.e. $D^{hamming}$ = 35/128). (4) Repeat on all remaining compounds to the end of the list. This diversification of the ACD reduced the 222,258 compounds to a set of 1199 compounds. The mean molecular weight of the diverse set, 340.7 g/mol, is only slightly larger than for the entire ACD (337.5 g/mol).

Distribution function properties

From the fingerprints for a set of N compounds we calculate all pair-wise chemical differences by a given distance function. A cumulative distribution function, $v(r)$, is calculated from this distance distribution. We define $v(r)$ as the probability that a randomly chosen compound, A, lies within a distance r from a second randomly chosen compound, B. The variable, r, is the distance by the chosen distance measure. The model is outlined in terms of distance, which can often be trivially transformed to similarity (e.g. Tanimoto similarity = $1 - D^{Soergel}$). We calculate $v(r)$ by considering all pairs of the N compounds, indexed by i,j:

$$v(r) = \frac{1}{N(N-1)/2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \begin{cases} 1 D_{i,j} < r \\ 0 D_{i,j} \geq r \end{cases} \quad (6)$$

If we assume the compounds have a random distribution in a Euclidean behaved space, then in the limit of small distances, *v(r)* approaches the function for the volume of a hyper-dimensional sphere,

$$v(r) = C(r^n) \quad (7)$$

where C is the content of an n-dimensional sphere of unit radius.[8] The compounds, being points in a descriptor space, act as surrogates for volume.

Equating the logarithms of both sides:

$$\log(v) = (n)\log(r) + \log(C) \quad (8)$$

In a plot of $\log(r)$ vs. $\log(v(r))$ the slope of the curve is the "apparent dimensionality" as a function of r, $n(r)$. This method has been used to analyze the "conformational space" of an ensemble of randomly folded 36-mer polyalanine chains, where $n(r)$ was extrapolated to r=0 and found to be close in value to the number of true mechanical degrees of freedom of the system[5].

Test Case

We test how well our method in conjunction with various distance metrics reports the dimensionality of delimited spaces represented by bit-strings. The spaces considered

are five dimensional and fifteen dimensional hyper-cubes with unit edge lengths. To simulate a diverse library, points are distributed throughout the hyper-cubes with a uniformly random distribution. To generate an *m*-dimensional point in an *m*-dimensional hyper-cube, *m* random numbers on [0,1] are chosen. Points are converted to bit-strings by representing each dimension by 100 bits, where the fraction of 'on' bits is taken from the corresponding random number. For example, if the first dimension value is 0.5923..., then the first 59 bits are '1' followed by 41 '0' bits. For *m* equal to five and fifteen, this is repeated five and fifteen times to give 500 and 1500 bit-length strings, respectively. This is repeated until we calculate a total of 40,000 points for each value of *m*. The *v(r)* and *n(r)* functions are constructed from these bit-string sets by calculating all pair-wise distances among the bit-strings. Using this tractable space, we compare the behavior of several distance functions. For comparison we calculate *v(r)* and *n(r)* on the original point distribution in the five and fifteen dimensional boxes, using the Euclidean distance for continuous variables,

$$D_{A,B}^{Continuous \quad Euclidean} = \{[\sum_{i=1}^{i=m} (x_{iA} - x_{iB})^2]^{1/2}\} / m^{1/2} \tag{9}$$

where *i* indexes the dimensions and the box-diagonal distance, $m^{1/2}$, normalizes the distances to the range [0,1].

Results:

Test Case Results

Figure 1 gives the *v(r)* distribution functions, both on an absolute scale (Figure 1a,d) and a logarithmic scale (Figure 1b,e), for the five dimensional box and 15 dimensional box. The curves differ depending on the distance metric used, which simply reflects different properties of the metrics. The *n(r)* functions (Figure 1c,f), in contrast, are expected to depend on the box dimensionality, *m*. With the exception of the dichotomous Euclidean distance measure, *n(r)* converges on the number of dimensions, *m*, with more complete convergence for the five dimensional case compared to fifteen dimensions. This is expected and illustrates the larger sampling space of higher dimensional spaces. The drop in apparent dimensionality, *n(r),* with increasing *r* reflects box boundaries constraining growth in *v(r)*. Interestingly, the dichotomous Euclidean distance is the least well-behaved distance metric by our criterion of reporting (the known) cube dimensionality correctly.

Diverse ACD results

Our test case shows that *n(r)*, using most of the distance measures, accurately estimates the underlying dimensionality of the data at $r=0$. While any distance measure would be suitable, with the exception of dichotomous Euclidean distance, our points can be made with only a couple measures and so we use the normalized Hamming distance and Tanimoto similarity (Soergel distance) for the remainder of the text.

Comparing descriptor sets: Daylight bit-string length dependence

v(r) and *n(r)* were calculated for the diverse ACD set using Daylight fingerprints of bit string lengths from 128 to 2048. While for all bit-string lengths, each bit is used (i.e. both on and off populated) across the compound set, the bit-strings become more sparse at large l (Table 1). As a result, the mean Hamming distance decreases with increasing l and the *v(r)* and *n(r)* curves shift to lower r for larger l (Figure 2a and 2b).

The most interesting apparent dimensionality value is at r=0, given our assumption that *n(0)* reports the underlying dimensionality of chemical space, as seen in figure 1c,f. For our diverse ACD, estimating this value requires an extrapolation of *n(r)* to r=0. No best functional form is immediately obvious to us for describing *n(r)* (Figure 2b-c). We give results for two functions. First, we arbitrarily use a linear fit to *n(r)* over *n(r)*>0.1 and interpret the y-intercept ($n_{y\text{-}int}$) as the chemical space dimensionality of the compound set (Table 1). Our second functional fitting is motivated by observing that n as a function of *v(r)*, *n(v(r))*, (Figure 2c) shows less dependence on bit-string length, l. Except for l=128, *n(v(r))* appears to be converging on a number between 10 and 20. We fit the *n(v(r))* plots to a functional form with two free parameters, n_0 and h:

$$n(v(r)) = n_0(1-h^{\log(v(r))}) \qquad (10)$$

This function approaches n_0 as *v(r)* (and therefore r) goes to zero. We report the n_0 parameter for this functional form fitting in Table 1, providing a second estimate for *n(0)* for each bit-string length. While the bit-string length, l, varies over more than an order of magnitude, the *n(0)* estimates show less variation, and are surprisingly anti-correlated with l. Excluding the l=128 results, all *n(0)* estimates lie in the range 11-19.

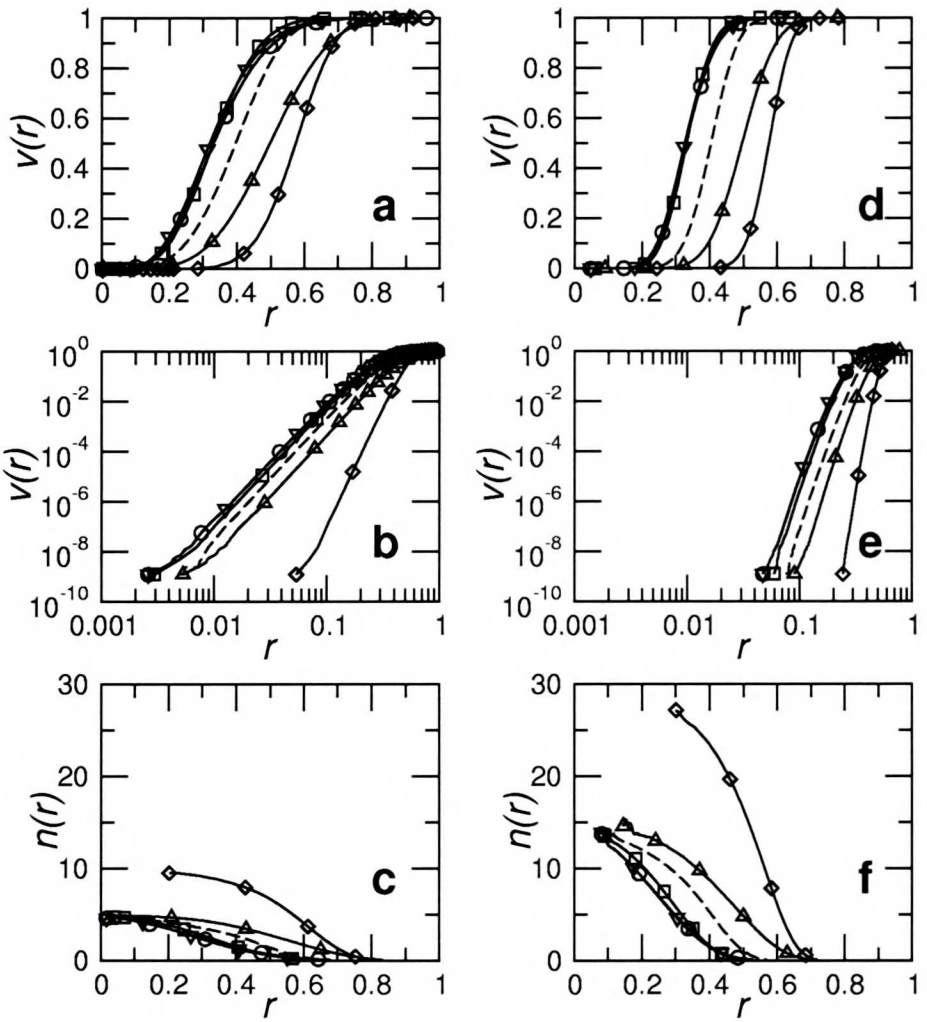


Figure 1: *v(r)* (**a-b**, **d-e**) and *n(r)* (**c,f**) for 40,000 points distributed uniformly random in a 5 dimensional box (**a-c**) and a 15 dimensional box (**d-f**). The dashed line plots *v(r)* and *n(r)* for continuous Euclidean distances calculated on points represented by 5 and 15 length vectors of random floating point real numbers on [0,1]. *v(r)* and *n(r)* are also constructed using points represented by 500 and 1500 length bit-strings (see text), using the dichotomous forms of the distance functions: Hamming (square), Euclidean (diamond), Soergel (up-triangle), Dice complement (circle), and Cosine complement (down-triangle). (**a**) and (**d**) give the *v(r)* functions on an absolute scale, while (**b**) and (**e**) report the same data on a

logarithmic scale from which the *n(r)* functions in (**c**) and (**f**) were calculated. *n(r)* functions were calculated from their respective *v(r)* functions starting at *v(r)* = 10^{-7} in order to reduce noise from low sampling.

Table 1: Estimates of $n(0)$ for Varying Bit-String Lengths and Function Extrapolations.

Bits (l)	Informative bits[a]	$\langle a \rangle$[b]	$n_{y\text{-}int}$ $[n(r)]$[c]	n_0 $[v(r)]$[d]
128	128	70.6	27.8	16.6
256	256	85.7	18.9	13.3
512	512	95.1	14.1	13.0
1024	1024	100.0	12.1	11.7
2048	2048	102.5	11.2	11.3

[a] Informative bits are those which are both on and off across the compound set.
[b] The mean number of on-bits for a compound, averaged across the compound set.
[c] The y-intercept, $n_{y\text{-}int}$, for linear fits to *n(r)*.
[d] The value of the free parameter n_0 in the two parameter fit to *n(v(r))* (see text).

The steep rise in the apparent dimensionality, $n(r)$, for l=128 at small r ($r \approx 0.3$) may be an artifact of the diversity filtering, which was performed in the l=128 space, and simply indicate an edge effect in the distance distribution. Because, by construction, no pair of compounds is closer than $D^{hamming}$ =35/128, $v(r)$ rises drastically at this distance resulting in a larger apparent dimensionality. The smaller jump in $n(r)$ at small r for the longer bit-string lengths suggests that this edge artifact softens upon calculating new molecular descriptors using larger bit-strings.

Extrapolating v(r)

From the perspective of lead optimization in drug discovery, the behavior of the distribution function, $v(r)$, is most interesting in the high similarity regime. Because of the high-dimensional nature of chemical space, only relatively large inter-chemical distances will be observed among compounds randomly sampled from chemical space. This limits the numerical coverage of $v(r)$ to the large r regime. We extrapolate $v(r)$ to small r based on extrapolations of $n(r)$, since $n(r)$ appears to exhibit some trend (Figures 2b-c) and because the reasonable value for $n(0)$ is limited by its physical interpretation - the number of ways to permute an average compound.

While $n(r)$ exhibits some trend, its exact functional form is not immediately obvious. We explore four different extrapolations of $n(r)$ (Figure 3a) in the l=512 fingerprint space and give the corresponding $v(r)$ extrapolations (Figure 3b). Extrapolation #1 is a linear fit to r versus $n(r)$ (Figure 2b), $n(r) = -37.81r + 14.14$.

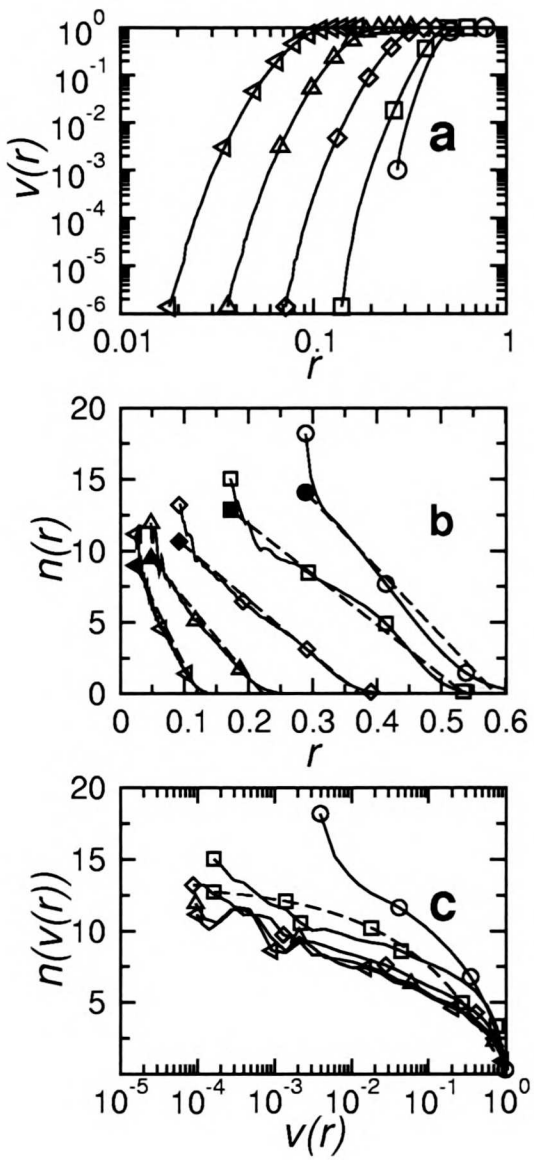


Figure 2: *v(r)* (**a**), *n(r)* (**b**), and *n(v(r))* (**c**) are plotted for the 1199 compound diverse subset of the ACD using Daylight fingerprints for *l* equal to 128 (circle), 256 (square), 512 (diamond), 1024 (up-triangle), and 2048 (left-triangle). *n(r)* (**b**), and *n(v(r))* (**c**) were calculated over $v(r) > 5\times10^{-5}$ in order to reduce noise. (**c**) also plots the equation $n(v(r)) = n_0(1\text{-}h^{log(v(r))})$, where $n_0 = 13.3$ and $h = 2.32$, which is fit to *n(v(r))* for $l = 256$ (dashed line, square).

Extrapolation #2 is a quadratic fit to r versus $n(r)$ over $0.092 \leq r \leq 0.221$, $n(r) = 317.2r^2 - 152.3r + 23.87$. Extrapolations #3 and #4 are fits to $n(v(r))$ (Figure 2c). Extrapolation #3 is as summarized in Table 1, $n(v(r)) = 13.03(1 - 1.836^{\log(v(r))})$. Extrapolation #4 is a linear fit to $\log(v(r))$ versus $n(v(r))$ over $8.8\times10^{-5} \leq v(r) \leq 0.068$, $n(v(r)) = -2.27\log(v(r)) + 3.82$. Given $n(r)$ [or $n(v(r))$], $v(r)$ is extrapolated from $r = 0.094$, $v(r) = 1.22\times10^{-4}$, by a simple Euler-type extrapolation (Figure 3b). Specifically, $n(r)$ [or $n(v(r))$] defines the slope for extending $v(r)$ to $v(r-\Delta r)$ in a $\log(r)$, $\log(v(r))$ plot. Here Δr=0.0001 base-10 logarithmic units. Testing smaller Δr values indicates convergence in the extrapolation (data not shown). With the exception of Extrapolation #4, the extrapolations behave somewhat similarly. $n(0)$ for Extrapolations #1-3 are 14.1, 23.9, and 13.0, respectively. Extrapolation #4 is not defined at $r = 0$ and its $n(1/512)$ value (~572) already exceeds the nominal dimensionality of the bit-string. Therefore, Extrapolation #4's behavior at small r appears unreasonable.

For comparison we construct $v(r)$ and $n(r)$ for the 512 bit-string space using the Soergel distance (Figure 4). The more widely used Tanimoto similarity measure (equal to $1 - D^{Soergel}$) defines the x-axis in Figure 4 for convenience. In contrast to behavior with the Hamming distance, $n(r)$ for the Soergel distance appears to plateau at ~18 with a spurious jump to larger values at the smallest r (Figure 4a). We use both the plateau value (18) and the largest observed $n(r)$ value (24) for extrapolating $v(r)$ (Figure 4b). Using these models for $n(r)$, the probability of obtaining a Tanimoto similarity score of 0.85 or better is 10^{-17} - 10^{-14}. A similarity of 0.95 corresponds to an expectation orders of magnitude smaller at 10^{-28} - 10^{-23}.

Discussion and Conclusion:

As drug discovery becomes increasingly dominated by chemical libraries and databases, interpreting results of chemical classification schemes will prove more important. This analysis on the distribution properties of random compounds raises several issues in this respect. We will focus on how to assess the validity of our results, what the general implications for medicinal chemistry might be, and the utility of the methods for the field of drug design.

Judging the reasonableness of our results requires careful consideration. The most limiting link in this type of analysis is construction of an appropriate and representative chemical library that captures the wealth of degrees of freedom available to the clever medicinal chemist. A rules-based scheme for generating random compounds may yield a huge potential chemical library [9], however any rule set limits an enumeration scheme to a very small fraction of chemical space unless the rule set is complete. "Drug space" biases[10] also require consideration, however even million-fold reductions in chemical space do not change the numbers here significantly.

The dimensionality values of chemical space, as reported by $n(0)$, give one handle on validity. Martin et al[11] show how principal components analysis and multidimensional scaling can reduce thousands of chemical descriptors to 16-20 dimensions which capture most of the chemical variance in combinatorial libraries. Our dimensionality values of 11-24 are in rough agreement and illustrate the strong correlation within a large descriptor

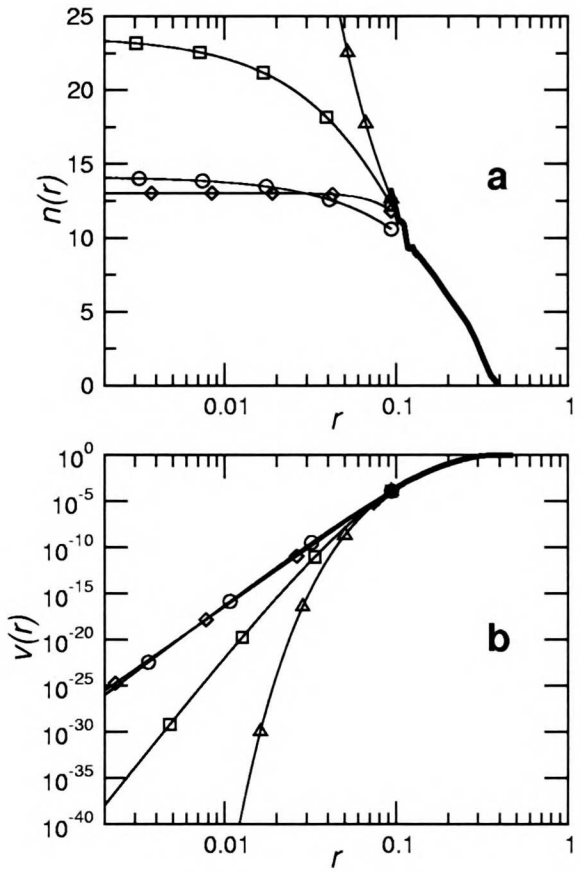


Figure 3: (**a**) Curve fittings to numerically derived *n(r)*. (#1) (circle) Linear fit to *r* versus *n(r)*, (#2) (square) hyperbolic fit to *r* versus *n(r)*, (#3) (diamond) exponential fit to log(*v(r)*) versus *n(v(r))*, (#4) (triangle) linear fit to log(*v(r)*) versus *n(v(r))* [see text for fitting details]. Solid line represents observed n(r). (**b**) Euler extrapolations of *v(r)* using the slope functions #1-4 (notation same as in (a)). Solid line represents observed v(r). All plots begin at $r = 1/512$ (≈ 0.00195).

set. Although a direct extension of dimensionality by our model, the expectation values calculated here are inherently more difficult to evaluate because of the extreme numbers involved. Common sense might deem the expectation values quite small. For example, only one in 10^{28} compounds may lie within a 0.95 Tanimoto radius about an arbitrary compound (Figure 4), far outside the resolution of *in vitro* high-throughput screening and even computational screening. However, given a low estimate for chemical space of 10^{60}, approximately 10^{32} compounds should meet this criterion. From this perspective, an expectation of 10^{-28} is quite large in absolute terms.

The challenge in creating molecular descriptors useful for drug-design is to include enough information capacity to segregate active compounds from inactive compounds for many disease targets. However, for any given target, the chemical variables relevant to drug activity may be an even smaller number than the 11-24 dimensions reported here.[12] Including information from irrelevant dimensions in calculating molecular distances effectively adds random noise that disperses a chemical neighborhood of true-active compounds. The noise-driven dispersal of active compounds would increase $v(r_{active})$ where r_{active} is the distance between pairs of active compounds. Thus, the distribution function, $v(r)$, provides a direct method for validating any chemical descriptor set provided that $v(r)$ has been constructed using the descriptor set and a intermolecular distances among sets of actives have been calculated with the same descriptor set. The distance measure is another variable that can likewise be evaluated.

Our formulation of the problem points to broader implications for drug design. From information theory, we know that $\log_2(W)$ bits are required to decode W distinct

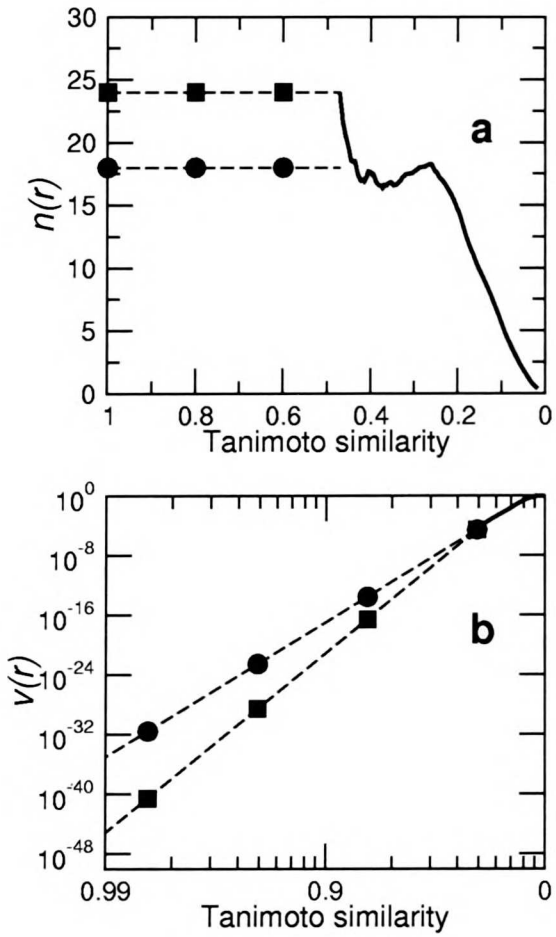


Figure 4. *n(r)* (**a**), and *v(r)* (**b**) are plotted for the 1199 diverse subset of the ACD using Daylight fingerprints for *l* equal to 512 using the Soergel distance. The functions have been transformed as $(1-D^{Soergel})$ in order to present the data using the widely used Tanimoto similarity measure. Two extrapolations are presented. The filled-circle corresponds to the low-*n(r)* estimate, and the filled square corresponds to the high-*n(r)* estimate.

outcomes [13]. If we make the approximation that $1/v(r)$ gives the number of distinct chemical neighborhoods of resolution r, we can evaluate the information required to sample any particular neighborhood. Decreasing the distance-threshold defines ever more exclusive chemical neighborhoods. Specifically, taking the negative base-2 logarithm converts the effective sampling resolution, $v(r)$, to the information measure of bits. To enter the 0.85 Tanimoto similarity radius about an arbitrary compound requires ~56 bits of information (i.e. $-\log_2(10^{-17})$, while 93 bits are required for the 0.95 radius. Using high-throughput screening of a million-compound library to choose a single best compound provides a maximum of $\log_2(10^6)$, equal to ~20, bits of information. Analysis of all the screening data can yield significantly more information per compound, especially with well-designed "informative" libraries[14]. In order to frame properly these numbers requires an estimate of the probability of a random compound being a drug (or high affinity ligand) for a particular drug-target/disease. If the information requirements of designing a drug can be estimated, the limits of any method at a step along the pathway can be better established.

References:

(1) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry* **2002**, *45*, 4350-4358.

(2) Daylight *Daylight Chemical Information Systems, Inc., Mission Viejo, CA 92691.*

(3) MACCS Molecular Design Ltd. Information Systems, San Leandro, CA 94577.

(4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching [Review]. *Journal of Chemical Information & Computer Sciences* **1998**, *38*, 983-996.

(5) Sullivan, D. C.; Kuntz, I. D. Conformation spaces of proteins. *Proteins: Structure, Function, & Genetics* **2001**, *42*, 495-511.

(6) Blaney, J. M.; Martin, E. J. Computational Approaches for Combinatorial Library Design and Molecular Diversity Analysis. *Current Opinion in Chemical Biology* **1997**, *1*, 54-59.

(7) ACD Molecular Design Ltd. Information Systems, San Leandro, CA 94577.

(8) Conway, J. H.; Sloane, N. J. A. *Sphere packings, lattices, and groups*; Springer-Verlag: New York, 1988; xxvii, 663.

(9) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design - A Molecular Modeling Perspective [Review]. *Medicinal Research Reviews* **1996**, *16*, 3-50.

(10) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug

Discovery and Developmental Settings [Review]. *Advanced Drug Delivery Reviews* **1997**, *23*, 3-25.

(11) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K. et al. Measuring Diversity - Experimental Design of Combinatorial Libraries for Drug Discovery. *Journal of Medicinal Chemistry* **1995**, *38*, 1431-1436.

(12) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information & Computer Sciences* **1999**, *39*, 28-35.

(13) Shannon, C.; Weaver, W. *The mathematical theory of communication*; University of Illinois Press, 1949.

(14) Miller, J. L.; Bradley, E. K.; Teig, S. L. Luddite: An information-theoretic library design tool. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 47-54.

Chapter 5: Obtaining Low Frequency Protein Modes by Observing Atomic Force Behavior in All-Atom Simulations

Kevin M. Masukawa, Irwin D. Kuntz

Department of Pharmaceutical Chemistry

Chemistry and Chemical Biology Program

University of California, San Francisco

600 - 16th Street, Box 2240

San Francisco, CA 94143-2240

Abstract:

As researchers continue structural studies on biological macromolecules, it is becoming clear that motion/dynamics is playing a sizeable role in molecular recognition. Slight changes in a protein's active site or protein surface can have profound effects on the binding landscape for a substrate. In this preliminary study, we utilized all-atom simulations of a high-resolution crystal structure to understand and quantitate molecular motion. We studied this system in environments of varying complexity; vacuum, a single solvated asymmetric unit, and an asymmetric unit in the context of the unit cell. In addition, we explored the nature of atomic forces that direct molecular motion in all these environments. Our studies showed that in all instances, forces fluctuated greatly, equating to roughly 50% of a force's magnitude. We also saw that nonbonded forces, while providing the means for low frequency motion, contributed very little to an atom's total force magnitude, roughly 10-15%. Explicit solvent simulations provided mean solvent collision times and mean forces of impact, $0.003 fs^{-1}$ and 2.26 kcal/mol Å, respectively.

Frequency analysis by Fourier transformation of force autocorrelations illustrated that molecular motion is not greatly affected with increasing system complexity. Solvated protein taken out of the context of the unit cell did not see a significant liberation of molecular motion despite the density of the system being less. Interestingly enough, we also saw that proteins in vacuum simulations also experienced similar frequencies of motion to our two more dense environments.

Introduction:

For years, structural biologists have acknowledged that molecular recognition is highly governed by shape complementarity of a substrate to its target, evolving a lock and key paradigm [1,2]. With advancements in biology and chemistry, structural biologists have now begun to investigate the role motion has in molecular recognition [3,4]. As with the advancements in biochemical techniques, increased computing power and efficiency has enabled computational investigations of the dynamic properties of these once static "locks" to become more tractable. Many experimental techniques provide structures that are time averaged; computational investigations, on the other hand, have the ability to give investigators structural information of a system as a function of time. It seems timely to build on the advances in experimental and computational methods to develop computational tools that further extend our understanding of protein motion at the atomic level.

Large scale motion facilitates or even governs binding of proteins to others, while motion in a protein's active site can aid in small molecule recognition. The dynamic behavior of a system can be classified into two types, harmonic and anharmonic. Harmonic forces dominate the local changes seen in a macromolecule with motions contributed mainly from high frequency bond stretching and angle bending. The anharmonic motions, on the other hand, are comprised of low frequency collective motions ($<300cm^{-1}$) contributed by nonbonded and dihedral components. These low frequency motions dominate the large structural changes and transitions seen in biological macromolecules. Since many of the structural changes that alter a recognition

site or the activity of an enzyme are large changes, it is the anharmonic motions that remain of most interest.

To date, mainly two theoretical techniques have been utilized to obtain the low frequency motions of macromolecules, normal mode and principal component analysis[5-7]. Normal modes account for the vibrational contributions to the internal energy at a given temperature. They correspond to the collective motions of the atoms in a system that can be excited [8]. By nature, normal mode analysis neglects solvent effects and assumes that the energy surface is quadratic (harmonic) within the well that is being sampled. As systems become more complex and malleable, this harmonic approximation breaks down. Principal component analysis (PCA) is one way to capture this low frequency anharmonic motion, but does so with a reduction of data[6,9]. Therefore, extraction of the harmonic and anharmonic frequencies of protein motion remains nontrivial.

Herein, we evaluate the use of molecular dynamics (MD) as a method to uncover the low frequency motions for macromolecules without a reduction in data that previous techniques may rely on. Molecular dynamics, unlike normal modes, has the ability to sample conformational space accessible to both harmonic and anharmonic motions. Previous studies have shown that for simple small systems, normal modes and those taken directly from MD simulations are well correlated [10], but as the motions get more complex, normal mode approximations begin to breakdown [9]. We look to extend our abilities of following low frequency modes by using molecular dynamics of macromolecules, sampling over longer time periods to capture anharmonic modes not normally accessible to normal mode analysis. This method allows for conformations of

the system to be characterized while maintaining a full atomistic representation of the system, including solvent effects.

In this study, we initially use a simplified test case to analyze the ability of molecular dynamics to properly capture the frequencies of motion in a simulation. We then apply this approach to a protein crystallized to ultra-high resolution, alpha lytic protease (αLP)[11], and investigate the effects the crystalline environment has on collective motion. Given the large number of biomolecular simulations that start from a structure that was obtained in a crystalline environment, yet simulated in solution, it is important to understand the degree of motion liberation as one moves from a crystalline to non-crystalline surrounding. Alpha lytic protease is a 198 residue protein and because of its evolution to be structurally stable under harsh environments, it has been a model to study protein folding[12]. This protein was chosen due to the large amount of structural detail available. The 0.83 Angstrom structure allows the greatest opportunity for a seamless transition from an experimental to a computational study by reducing the chances for setup/equilibration artifacts that may unnaturally affect motion. In this work we aim to uncover the environmental effects on protein collective motion, therefore, we also examined the frictional effects provided by the solvent by performing simulations in vacuum and comparing back to explicit solvent frequencies. Along with generating an understanding of protein collective motion, this study will provide insight into the force magnitudes within complex and simplified systems, thus yielding information on the general force behavior in various environments.

Computational Methods:

Simulation setup. All simulations were carried out using SANDER within the AMBER software package.[13] SANDER takes the derivative of the pairwise effective potential energy function shown in equation 1 to obtain forces on each atom to simulate molecular motion.

$$V(r) = \sum_{Bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{Angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{Dihedrals} \frac{V_n}{2}\left[1 + \cos\left(n\phi - \gamma\right)\right] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \quad (1)$$

Parameters for equation 1 were obtained from Cornell et.al[14]. All simulations were performed with the SHAKE algorithm [15] and TIP3 waters [16]. In order to track atomic forces, SANDER was modified to output the forces on each atom prior to each new integration step. A detailed description of each term in the above equation can be found in various references[14,17].

Alpha lytic protease. For this computational study, the 0.83 angstrom resolution crystal structure of αLP was used as the initial structure [11]. The crystal structure was of space group $P3_221$, with one molecule per asymmetric unit. Six copies of the alpha lytic protease were included in the simulation to reproduce the unit cell (Figure 1). Each copy was translated and rotated by their symmetry elements to the central αLP using the modeling package 'O'. Hydrogen atoms not included in the experimentally determined structure were added using the LEAP module within AMBER. All experimental glycerols, sulphates, and crystallographic waters were maintained during the setup. Glycerol charges and parameters were obtained using PARM99 and gaff [18,19]. Areas not containing crystallographic waters were filled in with TIP3P waters. Consistent with

experiment, a trigonal box was constructed encompassing the unit cell so that the simulation could be performed using PME[20].

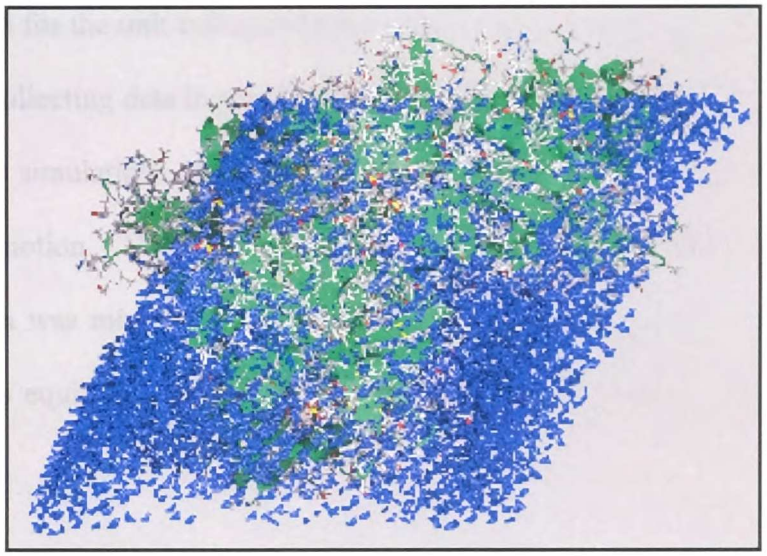


Figure 1. Trigonal unit cell containing six alpha lytic protease proteins. Ribbon diagrams show αLP and blue represents molecular waters.

The entire system was first minimized for 2000 steps. After minimization, all the waters were allowed to equilibrate for 10ps, initial velocities were assigned from a Boltzmann distribution. After water equilibration, the entire system was heated in NVT to 300K over 70ps. The system was then simulated in NPT for over 150ps to equilibrate the density before production MD. Upon reaching a stable density in NPT, the system was then equilibrated in the NVE ensemble to ensure the most natural trajectory (100ps). The NVE ensemble was simulated for 700ps after the equilibration for production MD. The 700ps simulation was split into equal 100ps segments and used in the analysis. Forces were only recorded for atoms of the central alpha lytic protease in the unit cell.

As opposed to the unit cell, simulations of a single asymmetric unit (ASU) of the alpha lytic protease in a box of water were also performed. A single alpha lytic protease was surrounded by a 10Å buffer of TIP3P waters in all directions. Similar equilibration procedures used for the unit cell calculation were repeated for the single αLP in a box of water prior to collecting data in production MD. 700ps of production MD was run.

Vacuum simulations of the αLP were performed to examine the effects of solvent on molecular motion. Vacuum calculations only involved 1 ASU of the αLP. The Vacuum system was minimized for 2000 steps then heated to 300K. The system was then allowed to equilibrate for another 150ps before collecting data in production MD (600ps).

Bonded carbons. Because we had little prior experience in examining forces in a solvated environment, we designed a simplified system consisting of two carbon atoms without hydrogens bonded together. Simulating just two bonded carbons required defining a new atom type. For this system, a modified atom type was generated using the same parameters for a carbon-carbon bond in the parm99 force field. Creating a new atom type enabled a simplified system that contained just two bonded atoms and not eight as for a saturated ethane. The two carbons were surrounded by 10Å of TIP3 water molecules. This system was initially minimized for 1000 steps, then heated in NVT for 45ps to 300K. It was then allowed to equilibrate in NPT for 50ps before production MD was performed in NVE for 500ps. This system was also averaged in a similar fashion to the proteins. The 500ps data was split into 10 X 50ps segments for analysis.

Force Analysis. SANDER was modified to output atomic forces in their x,y,z components for each MD step. Force vectors were outputted from the trajectory for all atoms over the course of the simulation. Instantaneous force magnitudes were determined for a particular atom, or in the protein case for a group of atoms, using the following equation:

$$|f| = \sqrt{f_x^2 + f_y^2 + f_z^2}$$

Where f_x, f_y, and f_z are the *x, y, and z* components of the force vector, respectively. To take advantage of the pairwise additivity of the force field given in equation 1, individual contributions to the entire force on an atom were separated out. This enabled force contributions from bond terms, angle terms, dihedral terms, and non-bonded forces to be isolated and characterized.

Time series analyses were performed on the discrete force vectors obtained over time from the simulations. Autocorrelation calculations yield the correlation between forces in successive steps of a trajectory as a time lag increases. They are used to detect non-randomness in the data and are obtained by the following equation:

$$c_{ff}(t) = \sum_{i=1}^{N} \frac{\langle f_i(t) \cdot f_i(0) \rangle}{\langle f_i(0) \cdot f_i(0) \rangle}$$

The normalized autocorrelation coefficient c_{ff}(t) calculates the dot product of each force vector to yield the degree of similarity between forces. The force on a particular atom at

a given time is represented as $f_i(t)$ and N represents the total number of atoms. Increasing lag times (t) were sampled until t = ½ total sampling time. The averaged dot products for a specific log time were then used to calculate the autocorrelation value. An autocorrelation value of 1 represents complete conservation of the force between successive steps, while a value of 0 represents no correlation for a force from a previous step. An autocorrelation value of -1 represents a force vector equal in magnitude, but opposite in direction from the reference vector.

To uncover the frequencies of motion, Fourier transformations (FT) were performed to yield a spectral density describing the proton motion. The autocorrelation of the force data from the simulations yielded plots that were highly convoluted with a linear combination of many periodic functions. Fourier transforms can deconvolute this signal and yield discrete bands for the frequencies underlying the long autocorrelation times. The Origin scientific graphing and analysis software was used to compute Fourier transformations of the autocorrelation data[21]. Origin utilizes the following equation to perform a discrete Fourier transform:

$$F(x) = \sum_{n=0}^{N-1} f_n(n) e^{(-i 2\pi k/N n)} \qquad F_k = \frac{k}{N}$$

N is the number of data points, F_n is the amplitude of the waves, and k is equal to the frequency. The spectrum yields the occurrence of a motion at a given frequency over the course of the trajectory. If the frequency of a motion is purely oscillatory and uninterrupted, a Fourier transform of that signal would be an infinitely narrow line. As the motion becomes more chaotic, but an underlying frequency remains, the FT gets

broadened out. Spectral plots were averaged using a 41 point Savitzy-Golay smoothing within the Origin software, as shown in figure 2.

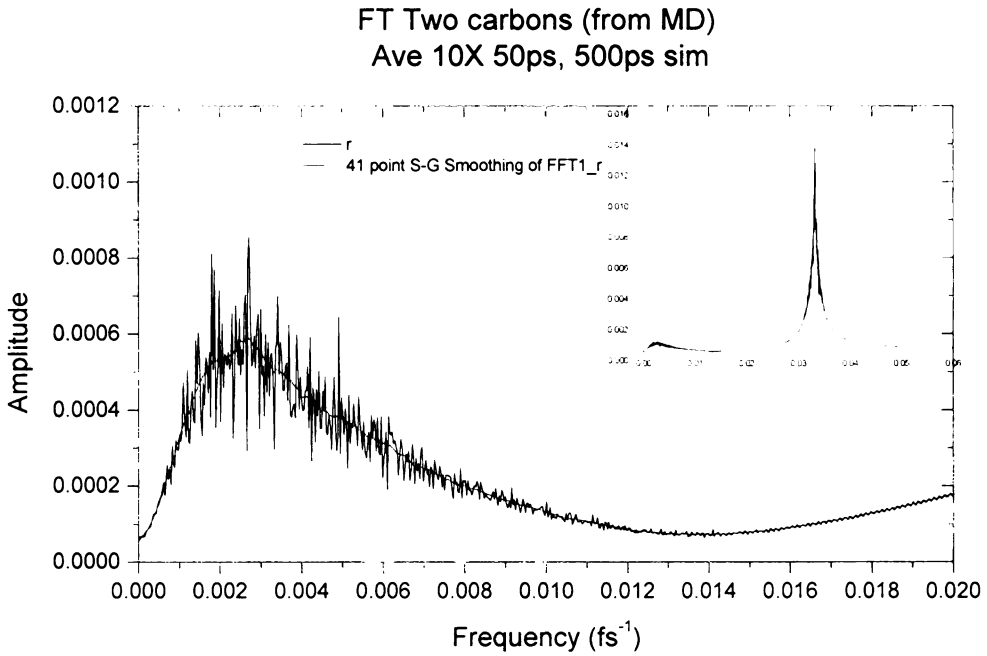


Figure 2. Spectral density of the Fourier transform obtained from solvated MD of two bonded carbons in a box of water. Inset shows the higher frequency discrete peak describing the carbon-carbon bond stretching. Frequencies below $0.02fs^{-1}$ are interactions of solvent with the carbon system. The 500ps simulation was split into 10 X 50ps runs. Force autocorrelations were then calculated for each run and an averaged force autocorrelation plot was Fourier transformed. 41 point Savitzy-Golay averaging shown in red.

Results/ Discussion:

Molecular dynamics simulations and analysis were performed for each of the following systems; two carbons in a box of water, αLP in vacuum, αLP in the unit cell, and a single solvated αLP. Over the data collection windows, figure 3 shows that the energy of most of the systems are well equilibrated. The exception is the potential energy of the unit cell simulation which appears to drift slowly to a lower energy suggesting that it may not be as well equilibrated the others. If this drift to a lower free energy is accompanied by a significant conformational change in the protein structure, our FT analysis should be able to observe it. Over the course of each trajectory, atomic forces were extracted and monitored. Initially, force magnitudes were investigated for their behavior in different environments. Time series analysis and subsequent Fourier analysis were then performed to uncover the frequencies of motion.

Force magnitudes. Molecular dynamics simulations were performed on two main systems, a bonded atom system and the protein, αLP, in varying environments. Raw forces for each atom were outputted at equal time intervals during the MD simulations. For the bonded atom system, forces on one of the carbon atoms were tracked over the course of the simulation, 1fs timesteps and data collected every 2fs. The protein simulations used 1fs timesteps and force data was collected every 4fs. Sensitivity of the results to varying timesteps and data collection windows was explored with the bonded carbon system. Altering either parameter had no effect on the qualitative autocorrelation or Fourier transform spectra (data not shown).

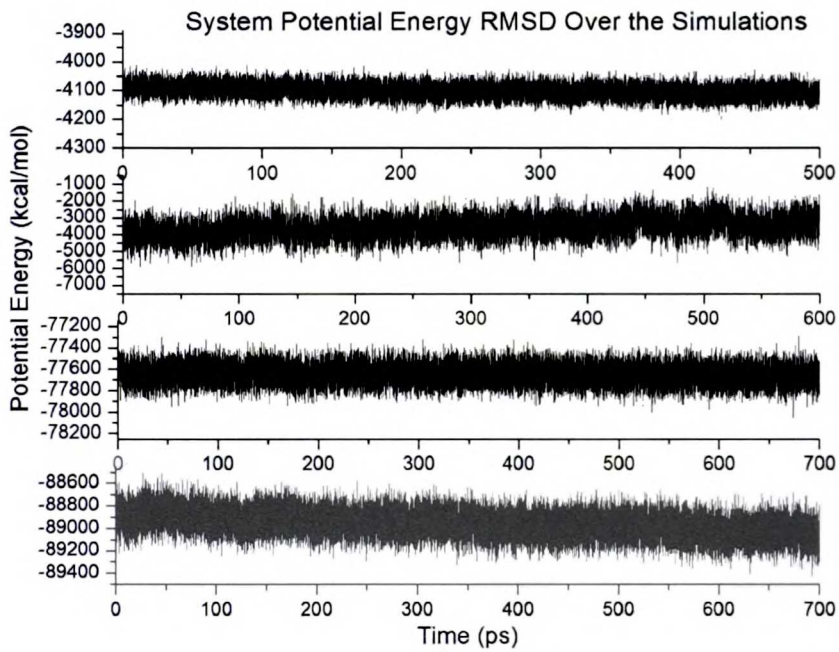


Figure 3. Potential Energy RMSD over the course of the production MD simulations. Top – two carbon simulation, Second – ASU in vacuum, Third – Asymmetric unit αLP, Bottom – αLP in unit cell.

For direct comparison with simplified systems, forces on a single carbon atom in the protein were monitored. Valines were chosen due to their simplicity and side chain resemblance to the bonded atom system. In particular, atom 787 of valine 58 and atom 310 of valine 23 were monitored. These particular residues were chosen so that analyses pertaining to the effects of the immediate environment can be made. As illustrated in figure 4, valine 58 has its side chain exposed to solvent, while valine 23 is buried in the protein.

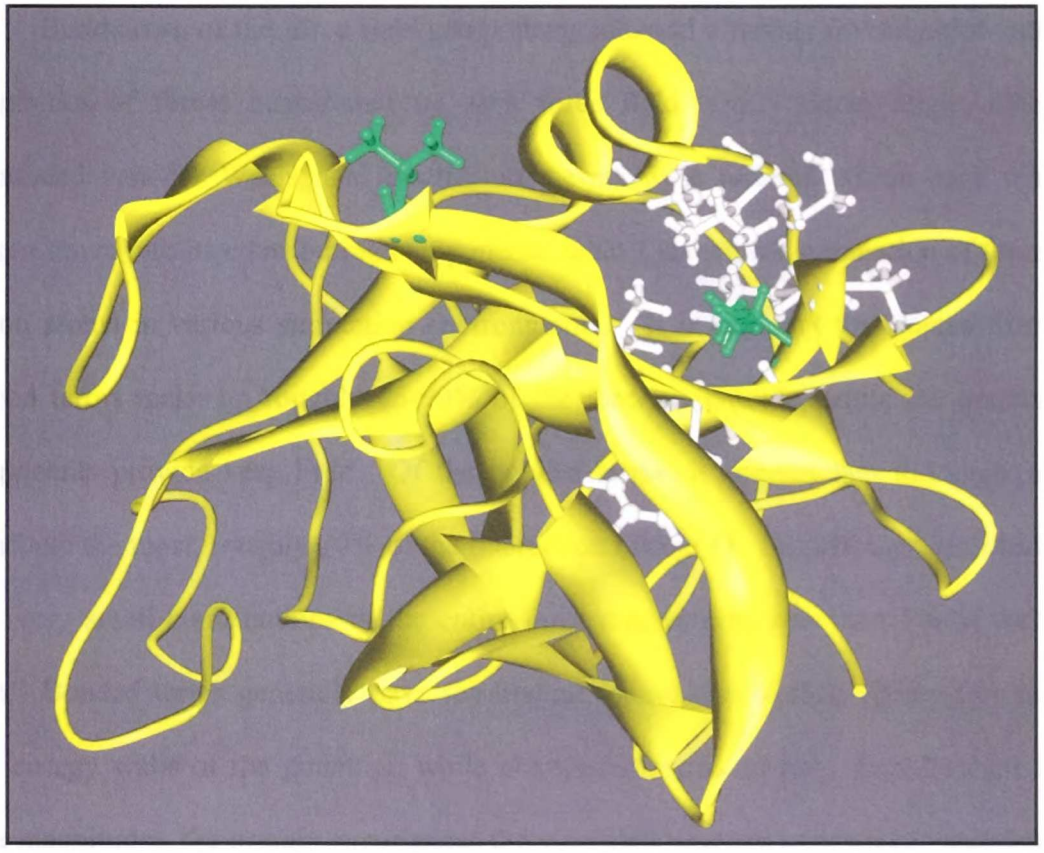


Figure 4. Ribbon diagram of the asymmetric unit of αLP. Green residues are the valines followed for the magnitudes study. Val 58 atom 787 CG1 is solvent exposed, Val 23 atom 310 CG1 is buried.

In general, we observe that the mean atomic forces increase as more interactions are considered. Mean forces on the carbon of the two carbon system are 10.47 kcal/mol Å, while forces on the carbon in the protein increases to roughly three times that. Total mean force for each atom studied is the same regardless of the location in the protein or type of simulation environment. As we will discuss in more detail, a discrepancy becomes apparent as we examine only the nonbonded components.

Breakdown of the force field components allowed a further investigation into the magnitudes of forces contributed by each force field term. Bond, angle, dihedral, nonbonded van der Waals, and electrostatic terms were isolated within each system. Broken down into its contributing components, table 1 shows the summation of forces on carbon atoms in various simulation environments. It is apparent that forces from the bonded terms make up roughly 85-90% of the force magnitude, while the nonbonded components provide very little. Of the bonded terms, it appears that the angle terms contribute the most, roughly 70% of the interaction (Table 1). Surprisingly, the dihedrals have very small contributions to the entire force magnitudes, less than 1% of the total force. Bonded terms generally have the highest values due to their ability to visit the high energy walls of the potential, while nonbonded terms do not. Despite their large force magnitudes, the protein experiences these bonded terms in an average sense leading to harmonic motion. As mentioned earlier, this harmonic motion results in stationary movements that do not alter conformation.

In contrast to the bonded terms, nonbonded components to the force are designed to prevent contact, thus their interactions remain small. On average, the nonbonded components make up about 10-15% of the total force interaction acting on an atom. These nonbonded terms, while small, contribute the most to the large low frequency motions of proteins ($<300 cm^{-1}$).

Table 1. Force Magnitude Breakdown. Forces remaining on atom after removal of specific force components are given. Parenthesis give standard deviations. Force tracked for valine 58 atom 787 CG1 (exterior), and valine 23 atom 310 CG1 (interior).

	C1	αLP Val 58	αLP Val 23	UC Val 58	UC Val 23
No Bonds with Hydrogen	10.472 (4.9)	29.41 (13.49)	29.60 (13.54)	29.78 (13.52)	29.40 (13.35)
No Bonds	N/A	24.29 (10.19)	24.27 (10.25)	24.57 (10.35)	24.07 (10.13)
No Bonds/Angles	N/A	2.93 (1.56)	4.76 (2.02)	2.83 (1.46)	4.93 (2.09)
No Bonds/Angles/Dihedrals	2.257 (2.1)	2.77 (1.51)	4.72 (1.96)	2.67 (1.40)	4.78 (1.98)

	Vac Val 58	Vac Val 23
No Bonds with Hydrogen	30.73 (13.92)	31.38 (14.50)
No Bonds	25.21 (10.56)	25.46 (10.85)
No Bonds/Angles	4.97 (2.68)	4.12 (1.99)
No Bonds/Angles/Dihedrals	4.82 (2.09)	4.02 (1.90)

As further illustrated in table 1, the nonbonded interactions depend on the immediate surrounding environment. Comparison of the forces on atom CG1 of valine 58 to the forces on atom valine 23, we see that forces in the interior of the protein are almost twice as high compared to residues primarily subjected to solvent interactions. This increase in force with the complexity of its surrounding is seen in both the αLP in a box of water and the αLP in the unit cell. Vacuum simulations of the αLP show slightly different behavior. Oddly, the forces on the exposed valine CG1 are larger than those on its counterpart in the interior. Performing the simulation in vacuum makes the system

void of molecular waters and prevents proper screening of nearby residues. This lack of solvent increases the interactions between neighboring residues and becomes apparent in the higher than expected forces on valine 58. Upon closer visual inspection, valine 58, which is solvent exposed has a neighboring arginine 57 that could explain the large nonbonded force component in vacuum. Valine 58 also has another arginine, residue 64, that is in the vicinity. In a solvated system it is feasible that the electrostatic potential from these two residues would be screened by solvent, but because vacuum simulations are void of this effect, these two arginines may have a larger effect leading to an artificially high force on the valine we studied.

A system of two bonded carbons in a box of water, which represents a simplified system, was used to isolate the water contributions to the forces. Forces from the nonbonded components represent the intermolecular interactions of one of the carbon atoms in the system with the surrounding solvent. This system enabled an estimation of the mean force of solvent collision, particularly TIP3P, and yielded an estimate for the frequency of solvent interactions with the solute, which will be discussed further below. Vacuum simulations of the bonded atom system saw a disappearance of nonbonded forces, providing further evidence that the solvent is represented by these interactions (data not shown). Additional simulations were also performed at varying temperatures. Altering the temperature of each system should result in the nonbonded interaction frequencies (i.e. waters) being shifted and the bonded frequencies remaining the same. As shown in figure 5, we indeed see a shift lower in frequency of this initial peak as the temperature decreases from 300K to 150K to 50K. It becomes apparent that these peaks vary completely with the square root of the temperature of the system. The fact that this

peak correlates with changes in temperature provides further support that the peak we are looking at in this frequency range is attributable to solvent interactions. No noticeable shift occurred for the high frequency bonded term.

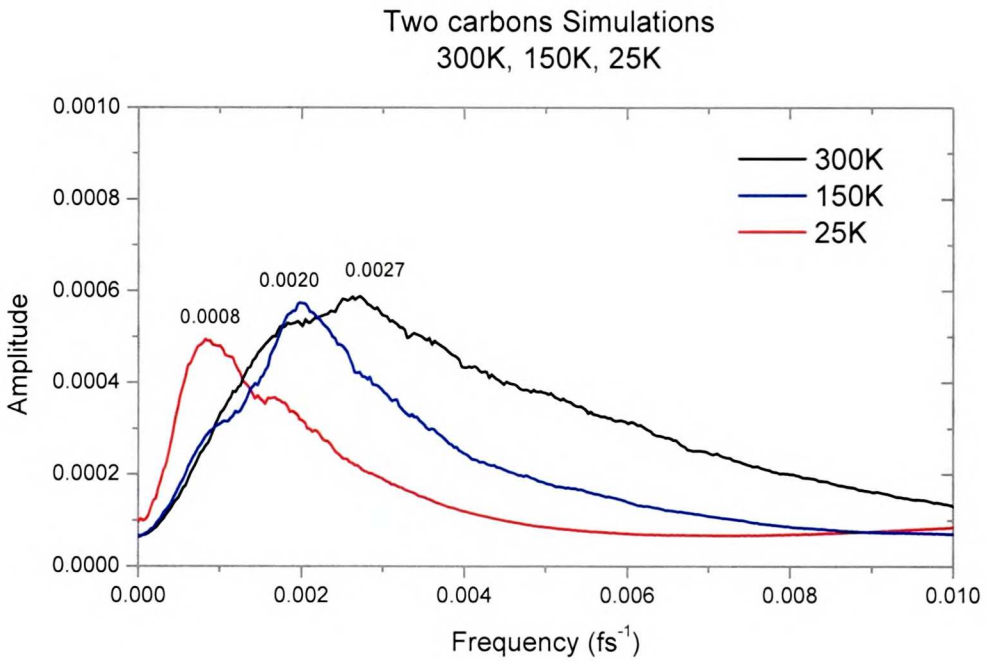


Figure 5. Bonded two carbon simulations at varying temperatures, 300K, 150K, 25K. Sensitivity to temperature provides support that frequencies in this range are composed of nonbonded intermolecular interactions.

Investigation of the force magnitudes demonstrates that the waters collide with very little force, on average 2.26 kcal/mol Å (Table 1). As we see with the protein, the finding that the nonbonded components contribute very little to the overall interaction is indicative of all nonbonded interactions, not just free floating waters. Upon further investigation in the protein system, we uncover that the protein atoms are influenced

more by their neighboring residues than they are by solvent. Vacuum simulations show that the exterior valine maintains a large nonbonded force component despite no solvent being present. The notion of solute atoms having a large effect on nonbonded forces is supported by that fact that force magnitudes generally increase as you bury the residue inside the protein, mentioned above. Subsequent simulations of the αLP in solvated environments show a slight increase in the force acting on this valine, but it is apparent that the solvent adds minimal interaction.

Upon looking at the atomic force behavior, we observe that the instantaneous forces can be extremely sensitive to atomic position. As evident by the large variance in magnitude distributions, recorded forces from snapshot to snapshot were markedly different despite the systems being well equilibrated (Figure 6). The distribution in forces span a 5 kcal/mol Å window in the simplified system, while the protein atoms experience a 15 kcal/mol Å force magnitude spread (Table 1). These large magnitudes in force fluctuations are apparent in essentially all simulation environments and on various locations within the same simulation (exterior vs. interior). This finding provides evidence that the force fluctuations are inherent in the force field and simulation method, not necessarily dependent on the complexity of the environment (i.e. solvent or large self-interacting protein). In all instances, the width of the force distribution remains roughly half of the mean force on that atom (Figure 6). Force fluctuations for the simplified two carbon system consist of approximately the same percentage of the total mean force as those seen in the complex unit cell system. The same can be seen when comparing vacuum simulations to those with solvent (Table 1).

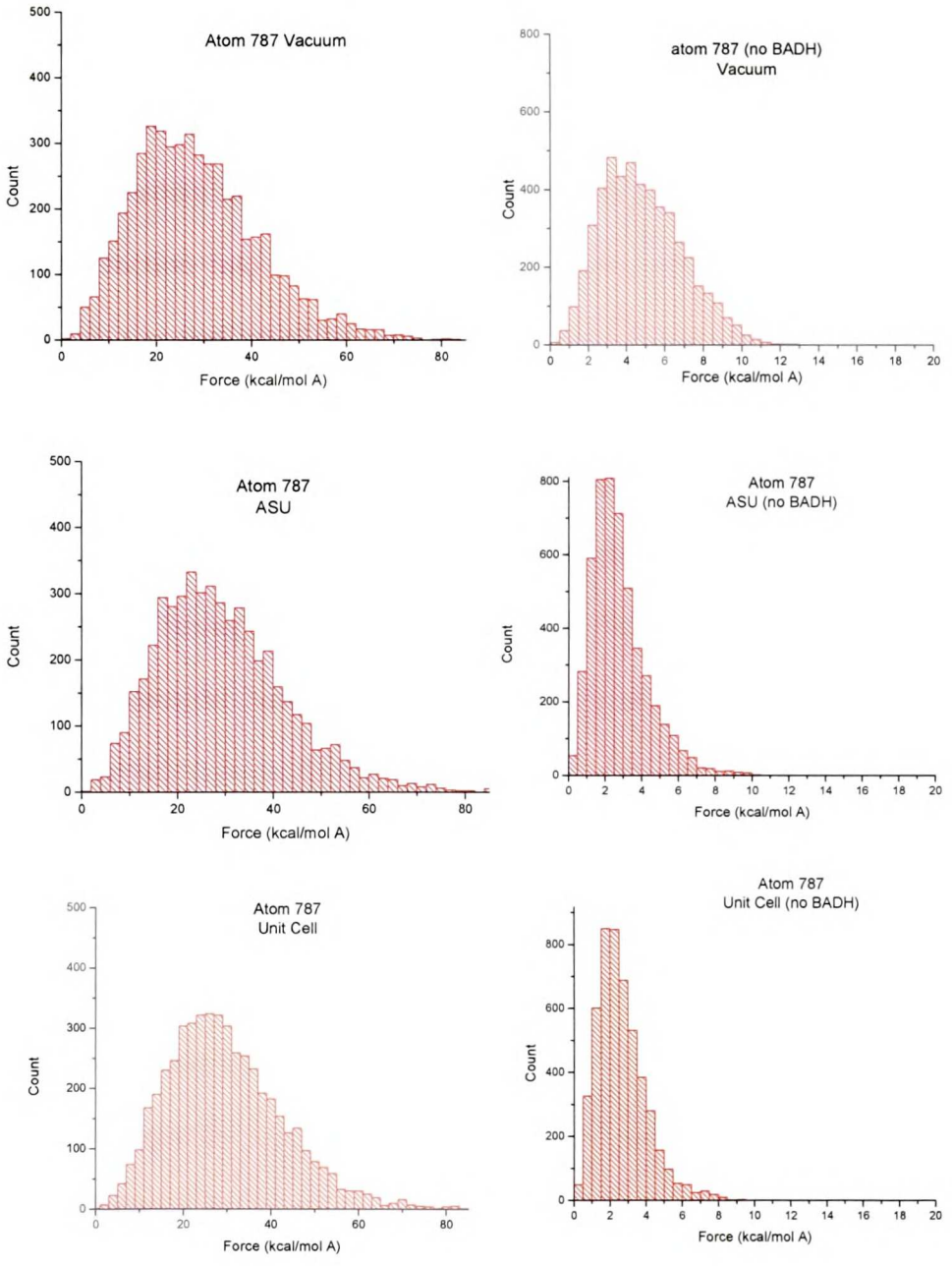


Figure 6. Histograms showing force distribution for atom 787 CG1 on residue valine 58 (exterior) over time. Histogram data for residue 23 is very similar and not shown. Left column represent forces when all interactions are present, right column are forces from the nonbonded components only. Top – vacuum simulation; Middle – ASU; Bottom – Unit Cell.

Autocorrelation analysis. Interested in the time dependent behavior, we performed an autocorrelation analysis utilizing the force vectors at each time step. As before, the pairwise additivity of the force field components allowed a breakdown of each autocorrelation function into separable force field components. To uncover atomic behavior, analysis was performed for atoms within the αLP, atom CG1 on valine 58 and atom CG1 on valine 23. As we see from plot 7 and 8, the autocorrelation values of the forces over the trajectory were highly oscillatory with rather long relaxation times. The long relaxation times suggest that the forces remain highly correlated throughout the simulation, but upon further investigation, we see that this high degree of correlation is mainly due to the harmonic bonded terms. Decomposition of force field terms from the autocorrelations reveals that, unlike the high oscillatory terms, the nonbonded terms have a rather quick relaxation time (Figure 7 and 8). Notice that an offset is seen for the nonbonded spectra. Ideally, the autocorrelation would be centered around zero. This is mainly due to the simulation data being calculated for an all interacting system (i.e. bond/angles/dihedrals/nonbonded components were all evaluated). Although we only follow forces from specific components (i.e. nonbonded terms), the coordinates of the system were influenced by all the interactions and remain correlated to some extent. Similar behavior was seen for other atoms on various residues throughout the protein (data not shown).

Dephasing times are found to be longer in vacuum as opposed to simulations with explicit solvent. Atom 787 on the exterior experiences a longer dephasing time compared to its counterpart in the solvated ASU or unit cell simulations (Figure 7). This information lends support to the notion that interactions on the surface of the protein are

more chaotic when you add solvent or are in the context of a unit cell. Interestingly, notice that the autocorrelation plots look similar for the exposed valine in both the ASU and unit cell conditions. This would lead one to believe that the surface of the protein looks the same whether in a crystalline or fully solvated environment. That finding is further supported by comparing force magnitudes on the surface and observing that there is no difference for the ASU or unit cell simulations (Table 1). As expected for an atom in the interior, relaxation times look very similar in all simulation conditions.

To capture the force behavior for an entire protein, force autocorrelations were performed for the entire αLP in multiple environments. Resultant vectors were summed for all atoms on the central ASU, 1-2755. Forces for atoms contained in this segment were summed up and treated as a whole when doing the autocorrelation analysis. These autocorrelation values were then used for a Fourier analysis to uncover the frequencies of motion. As expected, the autocorrelation data from the protein is highly complex and meaningful interpretation without deconvolution remains difficult.

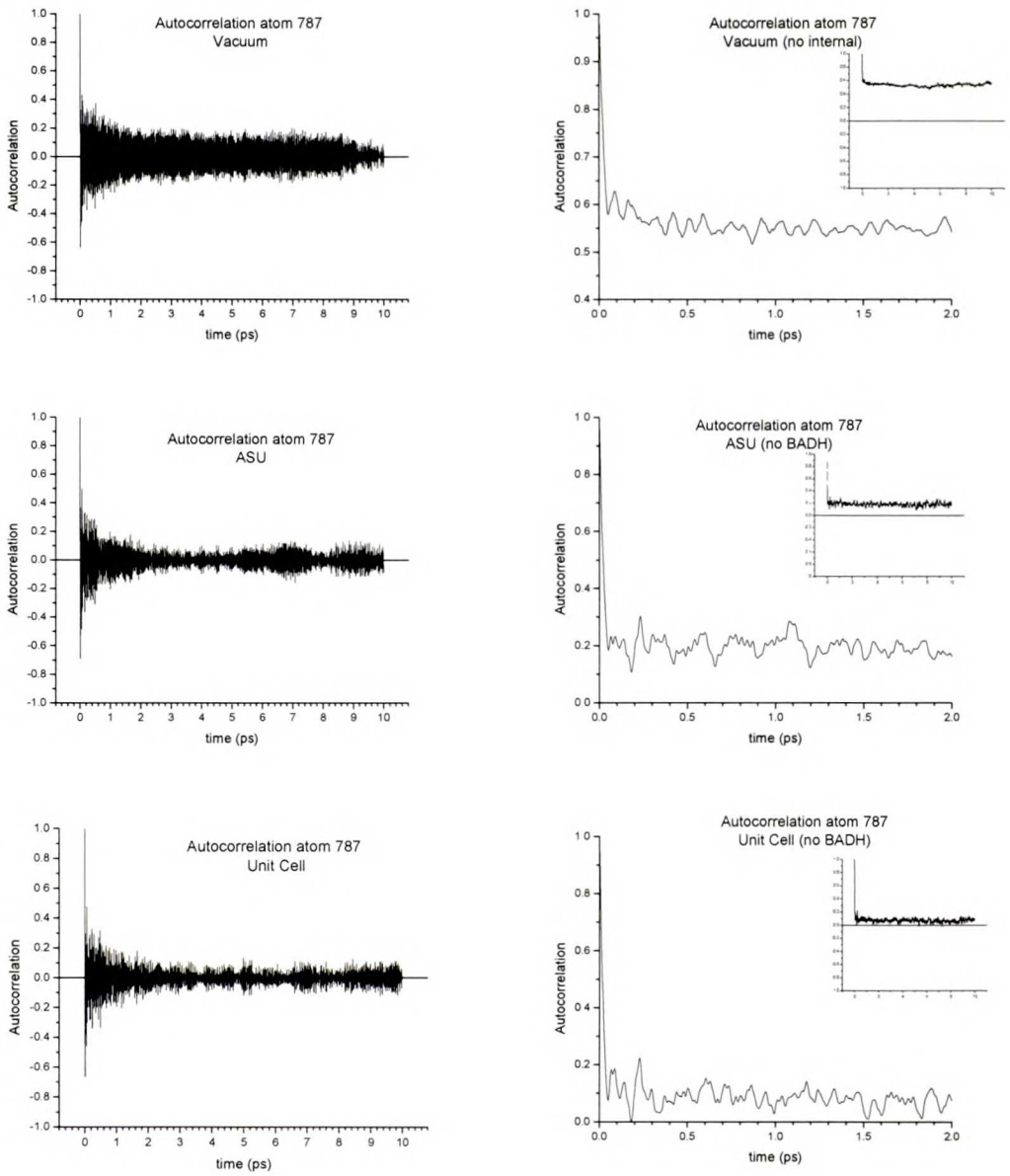


Figure 7. Autocorrelation analysis of valine 58 atom 787 CG1 (exterior). Top – vacuum simulation; Middle – ASU; Bottom – unit cell simulation. Left column represents autocorrelations with all force field terms present. Right column shows autocorrelation of nonbonded terms only.

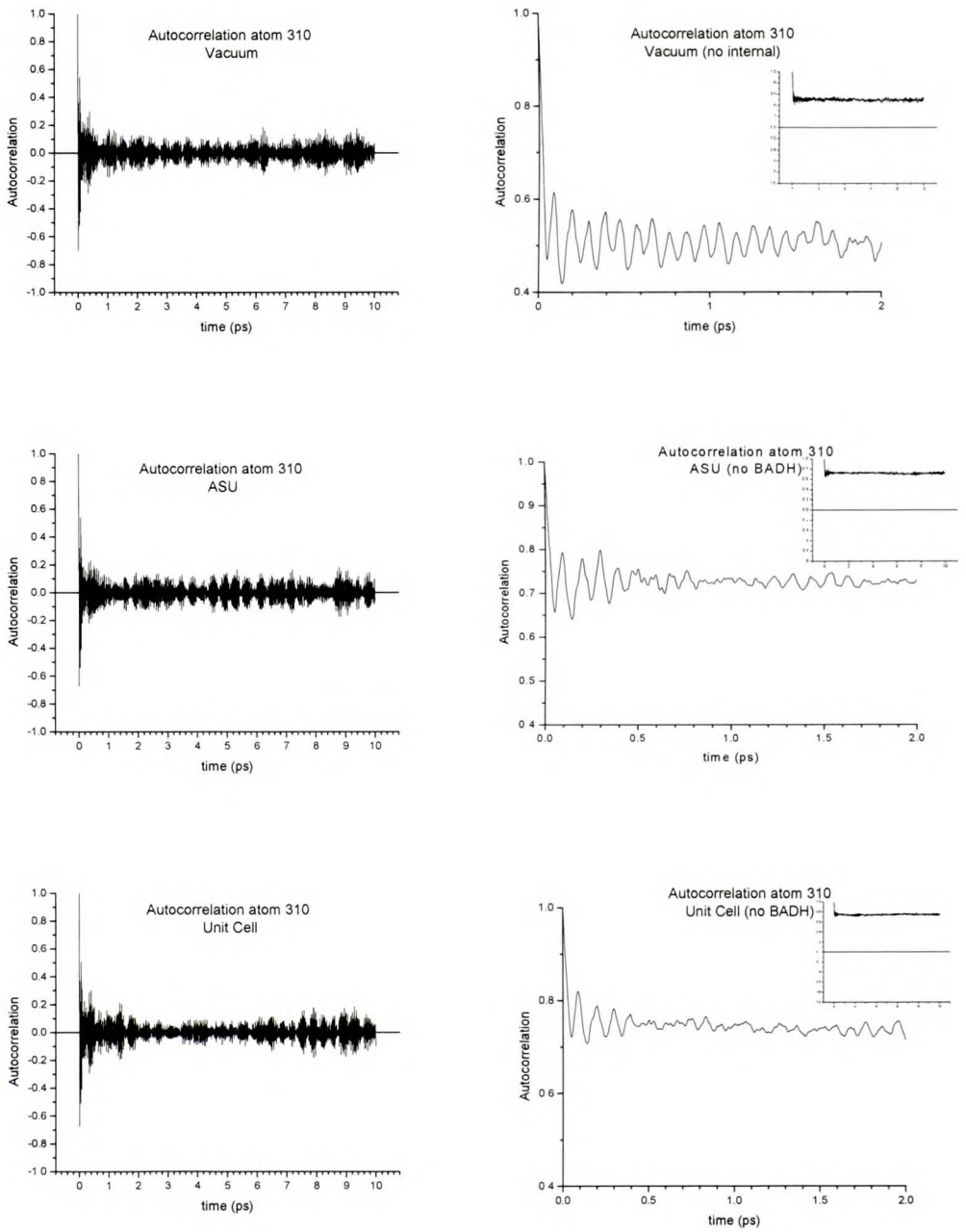


Figure 8. Autocorrelation analysis of valine 23 atom 310 CG1 (interior). Top – vacuum simulation; Middle – ASU; Bottom – unit cell simulation. Left column represents autocorrelations with all force field terms present. Right column shows autocorrelation of nonbonded terms only.

Fourier analysis. To uncover the frequencies underlying the simulation data, Fourier analyses were performed to yield spectral density for each system. Fourier plots describe the motion of the system in terms of amplitudes and frequencies. All protein plots are shown in figure 9, with only the frequencies that yield large motions plotted (<0.02 fs^{-1}). In this range, frequencies are broad due to the complexity of the interactions. Determination of frequencies of motion from protein autocorrelation data that was obtained without bonded terms resulted in retention of the low frequency protein fingerprints revealing that the low frequency modes (<0.02 fs^{-1}) correspond to those contributed from the nonbonded components in the force field (van der Waals and electrostatics).

To better understand solvent-solute interactions, we initially set forth to obtain an estimate for the mean collision time a molecule experiences in a simulation with TIP3P waters. Utilizing the simulations of the solvated bonded carbon system, we see that the water collisions are disperse with a mean collision time of 0.003 fs^{-1} (Figure 1). Vacuum simulations of the simplified bonded carbons saw a loss of peaks in this low frequency region supporting the notion that frequencies of the solvent are captured in this low frequency space. This solvent peak is also observed in solvated protein simulations, as evident by a leading low frequency edge within this same low frequency space (Figure 9).

In order to separate the solvent contributions from the low frequency protein motions, a protein vacuum simulation was performed. It is apparent that the majority of frequencies below $0.003 fs^{-1}$ are mainly due to solvent collisions and that frequencies

contributed from protein motion are mainly in the 0.003-0.012 region of frequency space (Figure 9).

Using this method, we next explored the effects various simulation environments have on protein motion. To estimate the effects of solvent friction on protein dynamics, comparisons of the frequencies of motion for αLP in the vacuum to motions for a solvated ASU were performed. We observed that the presence of solvent only slightly alters the potential energy surface of our αLP. Figure 9 illustrates in vacuum, motions are more pronounced. Peaks are a little more apparent suggesting that motion is not as chaotic as in the solvated ASU. This less interrupted motion is attributed to the lack of solvent collisions that would lead to more stochastic movement. Overall, while we do see this slight artifact from the inclusion of solvent, our studies suggest that the solvent has little effect on the overall shape of the spectrum. Frequencies and amplitudes are not altered significantly.

Similar to studies of the αLP in the vacuum versus solvated ASU, to study the effects of crystal packing, the crystalline environment protein was compared to an ASU protein simulated in a box of water. As seen in figure 9, simulations in the crystalline environment are very similar, showing roughly the same amount of conformational freedom in either environment despite the density of the system being less than that in the unit cell, 1.03 and 1.17, respectively. As apparent by the similar fingerprints in the low frequency space for the unit cell and ASU calculations, friction due to crystal packing does not seem to have an effect on the low frequency motions. The result is not surprising given the low temperature factors seen in this protein.[11].

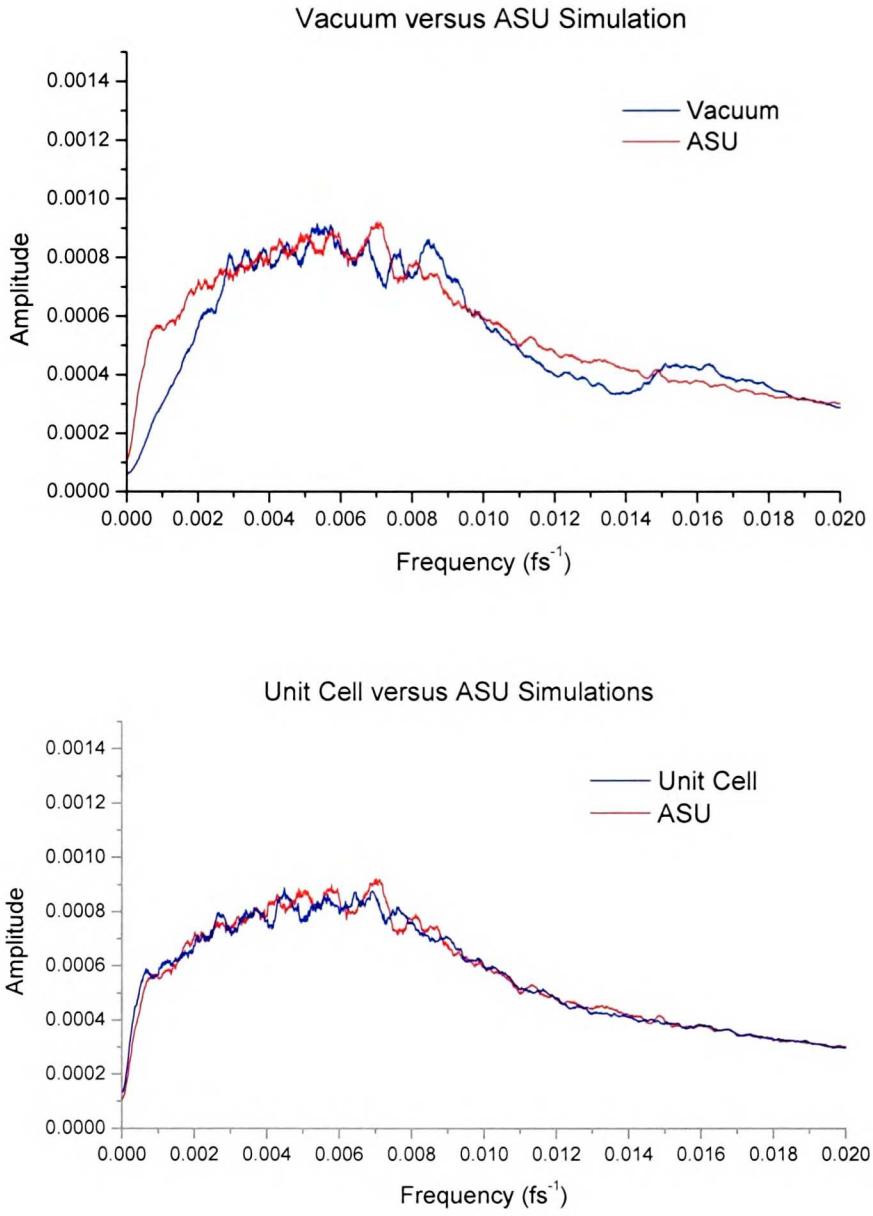


Figure 9. Spectral density of αLP in various environments. Data obtained from averages over multiple MD runs as described in the computational setup. Plots are 41 point Savitky-Golay. Top - Vacuum (red) versus ASU (blue). Bottom – Unit cell (red) versus ASU (blue).

In addition, when the atomic forces for each residue over the course of the unit cell trajectory were mapped onto the alpha carbons of the αLP (Figure 10), a uniform distribution of forces is observed suggesting that the force field does not put undue strain on particular secondary elements that would lead to movement. Similar studies with the ASU resulted in an equivalent distribution as the unit cell (Data not shown). Further providing evidence is the fact that force magnitudes for valine 787 exposed to the surface in the αLP and in the αLP in the unit cell experienced extremely similar force environments, 2.77 versus 2.67 kcal/mol A, respectfully.

Conclusion:

Computational studies generally rely on experimentally determined structures as starting points for analysis. Current experimental techniques, such as X-ray crystallography and NMR, elucidate structures of biological macromolecules and capture the system in a time-averaged state. While these techniques yield a detailed view of the system, they are unable to provide a time dependent measure of the frequency of protein motion.

We utilized all-atom molecular dynamics simulations to uncover general force behavior in simple and complex systems. Our observation of the force magnitudes showed that the nonbonded components to the force tended to be much lower in magnitude than compared to their bonded counterparts. Further investigations into nonbonded interactions revealed solvent interactions provided a small contribution to the force magnitudes on each atom exposed to solvent compared to nonbonded interactions

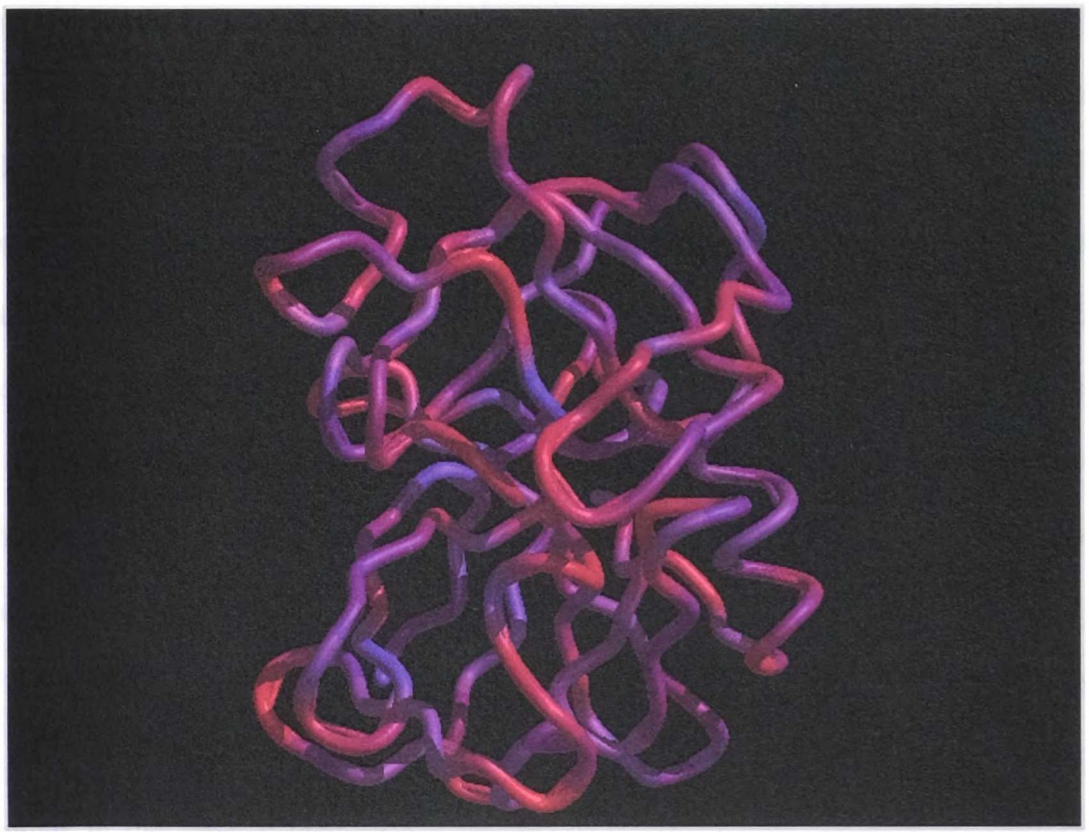


Figure 10. Averaged forces during the unit cell trajectory projected onto the C alpha for each residue. αLP exhibits a uniform distribution of forces throughout the protein

from the protein; suggesting that the intermolecular forces within the protein had more of an influence than intermolecular forces with solvent. Evident in all systems, force behavior studies also showed that forces experienced large amplitudes with high degrees of fluctuation suggesting that this behavior is inherent in the force field.

We compared forces for a protein simulated in various environments and investigated each's role in affecting protein motion. To investigate the dampening of

dynamic motions in a protein due to solvent effects or crystal packing, alpha lytic protease was simulated in multiple environments. Altering the environment around alpha lytic protease presented very similar spectra. Crystalline simulations versus ASU simulations in periodic boundary conditions showed that the low frequency motions are not librated or hindered with the change in environment. This shows that underestimating the dampening of motion is not an issue in explicit solvent simulations of an asymmetric unit. In vacuum, the low frequencies motions are slightly more pronounced due to a lack of solvent bombardment. In general, vacuum simulations maintained a spectrum similar to its solvated counterparts.

The development of a generalizable method to extract frequencies of motion from an all-atom simulation would be valuable to investigate protein movement in response to changes in the local environment. The results presented here demonstrate the use of extracting forces based on MD simulations to capture low frequency motions for an ultra-high resolution crystal structure of alpha lytic protease. Looking at the forces directly has the advantage of allowing the removal of certain forces in the post-processing, as opposed to seeing all movement with atomic RMSDs. Removal of the high frequency contributions to the forces yielded collision times and low frequency motions characteristic of large proteins.

In conclusion, the use of molecular dynamics has proven to be a reliable method to extract low frequency motions from macromolecules and this study hopes to provide the groundwork for further studies of this nature. While our preliminary survey has provided some interesting insights, much ore work needs to be done before a quantitative picture of molecular motion can be drawn. In the future, this type of analysis can be

useful in attempting to understand how biological macromolecules use motion to facilitate function and recognition. The generality of this method allows investigation of motion that is not restricted to large proteins. This method could also be used to help differentiate small molecules that bind to proteins. Changes in the low and high frequencies of motion of small molecules upon binding could allow researchers to gain a better understanding of the strain induced via induced fitting of small molecules to protein active sites.

References:

(1) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078-1082.

(2) Jorgensen, W. L. The many roles of computation in drug discovery [Review]. *Science* **2004**, *303*, 1813-1818.

(3) Carlson, H. A.; McCammon, J. A. Accommodating protein flexibility in computational drug design. *Mol Pharmacol* **2000**, *57*, 213-218.

(4) Freire, E. Statistical thermodynamic linkage between conformational and binding equilibria. *Adv Protein Chem* **1998**, *51*, 255-279.

(5) Case, D. A. Normal Mode Analysis of Protein Dynamics. *Current Opinion in Structural Biology* **1994**, *4*, 285-290.

(6) Hayward, S.; Kitao, A.; Go, N. Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins* **1995**, *23*, 177-186.

(7) Ichiye, T.; Karplus, M. Collective Motions in Proteins; A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins* **1991**, *11*, 205-217.

(8) Leach, A. R. *Molecular Modelling: Principles and Applications*; Addison Wesley Longman: Edinburgh Gate, Harlow, 1996.

(9) Levy, R. M.; Karplus, M.; Kushick, J.; Perahia, D. Evaluation of the Configurational Entropy for Proteins: Application to Molecular Dynamics Simulations of an α-Helix. *Macromolecules* **1984**, *17*, 1370-1374.

(10) Dauberosguthorpe, P.; Maunder, C. M.; Osguthorpe, D. J. Molecular Dynamics - Deciphering the Data. *Journal of Computer-Aided Molecular Design* **1996**, *10*, 177-185.

(11) Fuhrmann, C. N.; Kelch, B. A.; Ota, N.; Agard, D. A. The 0.83 angstrom resolution crystal structure of alpha-lytic protease reveals the detailed structure of the active site and identifies a source of conformational strain. *Journal of Molecular Biology* **2004**, *338*, 999-1013.

(12) Cunningham, E. L.; Jaswal, S. S.; Sohl, J. L.; Agard, D. A. Kinetic stability as a mechanism for protease longevity. *Proc Natl Acad Sci U S A* **1999**, *96*, 11008-11014.

(13) D.A. Case, D. A. P., J.W. Caldwell, T.E. Cheatham III, J. Wang, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R.J. Radmer, Y. Duan, J. Pitera, I. Massova, G.L. Seibel, U.C. Singh, P.K. Weiner and P.A. Kollman *AMBER 7*; University of California, San Francisco: San Francisco.

(14) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179-5197.

(15) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977**, *23*, 327-341.

(16) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **1983**, *79*, 926-935.

(17) Leach, A. *Molecular Modelling, Principles and Applications*; Essex: Edinburgh Gate, Harlow, 1996.

(18) Wang, J. M.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **2000**, *21*, 1049-1074.

(19) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157-1174.

(20) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *Journal of Chemical Physics* **1993**, *98*, 10089-10092.

(21) Software, M. *Origin: Scientific Graphing and Analysis Software*; 6 ed.: Northampton, MA.

Chapter 6: Conclusions and Future Outlook

by

Kevin M. Masukawa

The overarching goal of this body of work has been to obtain a firmer grasp on the determinants of molecular association. We utilized molecular dynamics and a high level scoring function to perform a detailed analysis on a variety of biological problems. Due to the complexity of interactions, there are numerous aspects that must be considered when describing molecular recognition. We approach many of these in this thesis, including charge states, individual thermodynamic components, dynamics, and the role of water molecules.

Chapter 2 began our study into molecular recognition with an analysis of pKa shifts in the 50S ribosome active site. Being able to estimate pKa values is nontrivial and in a complex nucleic acid system, it becomes even more daunting. Work in this chapter was able to identify a particular site on a nucleotide that is significantly perturbed and provided evidence for it being involved in catalysis. In Chapters 3, we analyzed our ability to successfully estimate free energies of binding for a carbohydrate-protein system and drew inferences from the data that deepened our understanding of selectivity within this system. Both chapters 2 and 3 pressed our abilities to model biological systems by approaching issues that had yet been fully addressed by the broader scientific community. These early chapters exhibited the power of computational methods, extending our knowledge of molecular recognition.

In Chapter 4, we explored an alternative technique in drug design. We considered the databases that researchers commonly use to screen for novel compounds and developed a model to characterize the information in it. Using a hyper-sphere model previously used to characterize protein conformational space, we illustrated this model's

potential to also characterize a small molecule database. Ultimately, this method would be useful to evaluate the significance of the chemical similarity between a top scoring compound from an HTS screen and the rest of the database.

Finally, chapter 5 involved an exploration of a common method used to simulate motion, molecular dynamics. Molecular dynamics provides sampling of a system to capture a detailed description of the system as it steps through time. Due to the time dependent nature of the information, this technique provides a platform to consider frequencies of motion. All-atom simulations of a high-resolution crystal structure were used to estimate the frequencies of motion in a complex protein system. We observed that instantaneous atomic forces varied widely regardless of the complexity of the simulation environment or your location within the protein. In addition, molecular dynamics studies showed that, in general, the frequencies of motion are unaffected when the protein is allowed to move in different environments (i.e. vacuum, fully solvated, and unit cell).

Despite much success being achieved over the years, computational structural biology is faced with a multitude of challenges as the field continues to progress[1,2]. While many will agree that free energy perturbation or thermodynamic integration techniques are extremely accurate, their capabilities in a broad class of problems are limited and thus not amenable to many problems in drug design. As a result, researchers have turned to less rigorous techniques for drug design, but problems facing these methods are more pronounced. The two main challenges that hindered current methods in structure-based design are errors in scoring and inadequate sampling.

The amount of conformational space available to biomolecules is vast and current all-atom techniques can sample only a limited subsection of this area due to the rugged terrain of the energy landscape and limits in computing power. Advances in sampling techniques such as torsion angle molecular dynamics[3], Monte Carlo[4], or parallel tempering[5], help to overcome these barriers or smooth the potential surface, but many obstacles remain, particularly in the parameterization of the variables used in these techniques. Improved sampling techniques will enable a more thorough search of the energy landscape, thus estimates of binding free energies will better match global minima rather than those in local minima.

As shortcomings in sampling limit our ability to generate true estimates of binding free energies, scoring functions are also being evolved to better accurately quantify association events. Scoring functions are the bridge between experimentally derived binding affinities and computational estimates; therefore they are at the forefront of scrutiny. Scoring functions are very complex and sensitivity to the parameters it is fed remains high. To further complicate scoring function issues, choosing interior dielectric constants which to model solvent behavior as well as appropriate charge models is crucial and an industry standard has yet to be unequivocally established. As researchers continue studies like those contained in this thesis, our understanding of molecular recognition deepens, and inevitably better scoring functions will evolved.

Both, sampling and scoring, have profound effects on the quantitative estimate of binding free energy for a study and remain at the center of discussion among many in the field[6,7]. More efficient computers and a better understanding of recognition at the atomic

level will help to optimize current scoring functions and help push forward the next generation techniques.

References:

(1) Kuntz, I. D.; Agard, D. A. Assessment of the role of computations in structural biology. *Adv. Protein Chem* **2003**, *66 PG. 1-25. 2003. Adv.Protein Chem. 66 PG. 1-25. 2003. ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495 USA. URL*, http://www.apnet.com ADVANCES IN PROTEIN CHEMISTRY ACADEMIC PRESS INC.

(2) Kollman, P. A. Advances and Continuing Challenges In Achieving Realistic and Predictive Simulations of the Properties of Organics and Biological Molecules. *Accounts of Chemical Research* **1996**, *29*, 461-469.

(3) Stein, E. G.; Rice, L. M.; Brunger, A. T. Torsion-Angle Molecular Dynamics as a New Efficient Tool for Nmr Structure Calculation. *Journal of Magnetic Resonance* **1997**, *124*, 154-164.

(4) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **1953**, *21*, 1087-1092.

(5) Mitsutake, A.; Okamoto, Y. Replica-exchange extensions of simulated tempering method. *Journal of Chemical Physics* **2004**, *121*, 2491-2504.

(6) Sheinerman, F. B.; Norel, R.; Honig, B. Electrostatic aspects of protein-protein interactions. *Current Opinion in Structural Biology* **2000**, *10*, 153-159.

(7) Rizzo, R. C.; Aynechi, T.; Kuntz, I. D.; Case, D. A. Estimation of Absolute Free Energies of Hydration using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions. **In Preparation**.

UNIVERSITY OF CALIFORNIA
San Francisco
LIBRARY

7375028
3 1378 00737 5028