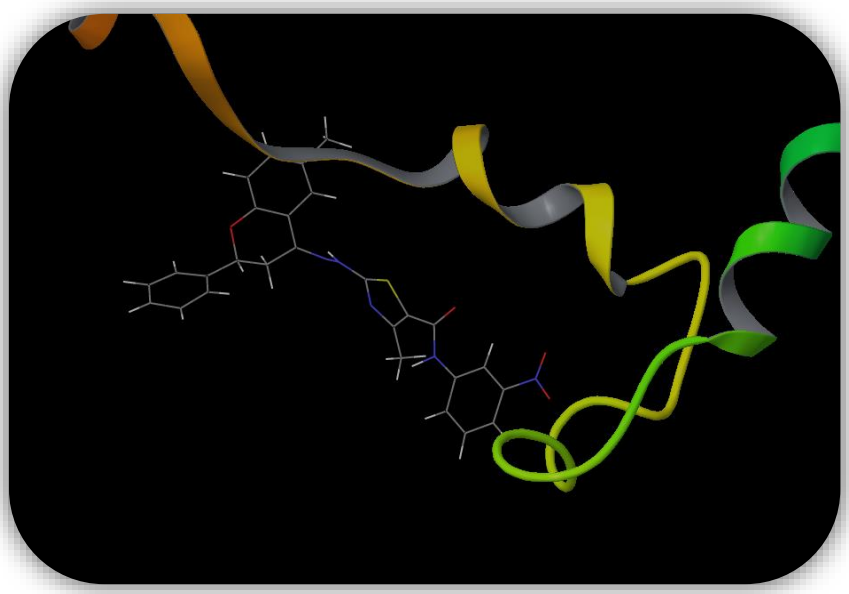
The POSSIM Force Field: Developing Parameters for Organic Molecules and Simulating Biochemical Systems

Daniel Sigalovsky

A Major Qualifying Project Report
Submitted to the Faculty of the
Worcester Polytechnic Institute
In partial fulfillment of the requirements for the
Bachelor of Science Degree



Approved By:

Professor George A. Kaminski

Contents

Acknowledgements

The research group would like to most sincerely thank Dr. George Kaminski for guiding the investigations conducted here as well as providing invaluable expertise and supervision without which this study could not have been completed. Additionally, I would like to thank John Cvitkovic, Xinbi Li, and Ity Sharma for numerous informative discussions that greatly helped the formation of this research.

Abstract

Computational modeling of chemical systems has provided a fast and cost-effective method of approximating molecular properties and interactions, especially those that cannot be determined experimentally. Here, the POSSIM (POlarizable Simulations with Second order Interaction Model) force field is used in order to demonstrate the efficacy of polarizability in accurately reproducing experimentally and quantum mechanically derived results for gas phase dimers and geometries, liquid properties, and energies of binding. Specifically, parameters for small-molecule analogs of tryptophan and arginine are fitted and then transferred to their full size peptides for incorporation into proteins. Pyrrole and methyl guanidinium ion were used as analogs for tryptophan and arginine, respectively, and the largest energy error after fitting was only 3.7%, demonstrating good agreement. The conformational energy error for tryptophan was 0.06% and for arginine 14.9%, although in comparison to other results this is reasonably accurate. Additionally, it was investigated how well a non-polarizable force field, Optimized Potentials for Liquid Simulations (OPLS), models the binding energy of the oncoprotein c-Myc for a library of known anti-cancer inhibitors. While this force field provided an average error of binding energy of about 8.9 kcal/mol with Molecular Dynamics and 4.2 kcal/mol with one inhibitor with Monte Carlo without fitting, it is anticipated that the POSSIM model will achieve a higher accuracy when implemented, with the hope that parameters can be developed as tools for investigating protein-ligand binding.

Introduction

Molecular modeling has been an integral part of chemical and physical research ever since the advent of the computer. While wet laboratory techniques have been invaluable in elucidating many of the daunting questions that these fields present, practical considerations such as time and cost can hinder the progress of this form of research. Molecular modeling seeks to prepare scientists for experiments in the laboratory by providing an initial guess that can be used as a heading, much like that which a compass provides when the desired direction is not known. At the same time, *in silico* techniques cut down on costs and save researchers time. Using the known properties of atoms and their behavior, one can perform calculations beforehand in order to inform about experiments needed to be run further on in the investigation process. Furthermore, theoretical calculations can be used to characterize atomic-scale properties that cannot be realistically investigated in the wet laboratory. Such properties ought to follow mathematically-derived physical principles, so molecular modeling can be used to test their validity.

However, molecular modeling is just a model like any other. Models have been used widely in any area of research because they simplify systems to forms that are not necessarily accurate but provide approximate answers to questions in order to guide researchers in the right direction. For example, model aircraft are often used to investigate how well the target airplane will perform. The observer understands that while this model will not operate on its own nor bear the same weight as the aircraft being represented, the model can still be used to assess the shape of the wings, the positioning of the rotors, and the distribution of cargo, just to name a few properties. Using such a model is much more trivial and practical than building a real size airplane and then performing these assessments, because doing this would require too much time and resources than is realistically feasible. In this same way, molecular modeling and computational chemistry aim to

represent real intermolecular interactions and molecular properties using knowledge derived from previous experiments.

Using the computational power of molecular modeling software, the research presented here sought to develop parameters for two of the twenty side chains of proteins commonly found in nature, namely tryptophan and arginine. Additionally, the group investigated the energetics of the interaction between a known oncoprotein (one that causes cancer, and here, specifically, Burkitt's lymphoma), c-Myc, and a host of inhibitors in an effort to model these interactions and develop new and more effective inhibitors. Additionally, an overarching aim of this research was to include a novel approximation of polarizability, one that requires fewer parameters than earlier polarizable models, to molecular modeling and thus more accurately reproduce ligand-protein binding energies.

Background

Here the group presents information pertinent to the understanding of how the field of computational chemistry arose and how it works to accurately represent the laws of nature. Additionally, information about force fields and how they are a crucial part of molecular modeling is given, as well as an explanation and analysis of previous attempts at developing these force fields and how the polarizable force field intends to fix some of the issues present in these earlier versions. Details of the goals of the research project are also provided, but generally these goals were as follows:

1. Develop parameters for the small molecule analogues of the residues tryptophan and arginine, which are called pyrrole and methylguanidinium, respectively,
2. Transfer these parameters to the whole dipeptides of tryptophan and arginine,

3. Characterize the energetic interactions between the oncoprotein c-Myc and individual inhibitors initially determined by Yin et. al.

In general, the aim is to use developed parameters for each atom type as a tool for ligand-protein applications, with c-Myc investigations as a first attempt in doing so.

Computational Chemistry and Molecular Modeling

Quantum mechanics essentially and specifically describes how molecules are put together and provides the most accurate data in molecular modeling. The way in which two atoms form a bond is the basis for building molecules and quantum mechanics provides mathematical reasons for this event to occur. Quantum mathematics can also be used to describe bulk processes in which the system is much larger than a simple pair of atoms and contains many different molecules moving around each other. Some positions and other properties of these molecules are more favorable than others. Thus arises the concept that the way in which molecules behave can be characterized by a set of probabilities. But how can one determine how probable some molecule is to move from one place to another and how likely this molecule is to adopt certain geometrical conformations? The answer lies within the basis for all computational chemistry, the Schrodinger equation, which seeks to predict the probabilistic behavior of particles in a given system:

$$\hat{H}\Psi = E\Psi,$$

where $\hat{H}$ is the Hamiltonian operator acting on the wave function Ψ, and E is the energy of the system. Essentially, the wave function describes the position of all particles within the system and equation seeks to describe the probability of finding these particles in that system in consideration of their energy. Additionally, it is important to note that this is time-independent Schrodinger equation, and thus only describes the probabilistic positions of particles in an instantaneous state of a given system. However, the most important idea to understand here is that particles move

according to their energy and that everything in nature seeks to lower its potential energy. The Schrodinger is powerful because it follows this law and states that energy informs position.

The ideal molecular modeling program will seek to find the solution of the Schrodinger equation and output the most probable position of the input atoms. The advantage of such quantum mechanical methods is that they are always accurate because these methods predict the behavior of electrons exactly. It is somewhat trivial to solve this equation for single electron systems and this can be even be done out by hand, but the complexity of the system exponentially increases with each electron added. Thus, eventually it becomes impractical to seek exact solutions to the Schrodinger equation for chemical systems in which many electrons are present. If a molecular modeling software is given such a chemical system and then asked to solve the Schrodinger equation exactly, it will take an enormous amount of time and CPU power to arrive at the solution, so much so that such a calculation would be impractical for research purposes. In addition, different quantum mechanical approaches may arrive at different solutions. For this reason, empirical and semi-empirical methods have been developed and are used in this research.

These methods seek, much like models, to approximate the solutions to the Schrodinger equation in various ways. Some of these methods are more accurate than others and some utilize more CPU time as well, but all of them require parameterization. This means that numerical parameters need to be developed and inputted into the method in order to maximize its accuracy. These parameters are derived from exact results that come either from quantum mechanical or experimental methods. Understandably, it would seem that performing these preliminary calculations and then importing the results into an empirical method is redundant. However, once these parameters are developed, they are transferrable from one computational experiment to another. Transferability is invaluable in computational chemistry research since parameterization

need only occur once and can then be used to run many different experiments. This technique is used in this research, for example, when simulating peptides. Once parameters have been developed for a certain side chain, they can then be used when in the context of a protein that contains that side chain. This approach is especially useful in polarizable models, since the polarizability parameter describes responses to changing electrostatic environments, while fixed-charge models do not contain this feature.

Force Fields

While empirical and semi-empirical methods are certainly effective for smaller organic systems, for biochemical systems that include proteins, lipids, ribonucleic polymers, and other many-atom molecules, the cost using these methods is far too great to be practical any longer. For situations like these, molecular mechanics methods are most useful. These methods are unique in that they do not employ wave functions or electron density functions, like quantum mechanical methods do. Instead, the energy of a molecule is described by an arithmetic combination of energy terms resulting from certain electrostatic and geometric aspects of the molecule in question, depending on the particular method, of course. Force fields constitute these equations. There are many different types that have been developed and some are more accurate than others. The goal of a researcher developing a force field is to maximize the accuracy so that the results are as close to quantum or experimental results as possible and to minimize the CPU time it takes in order to complete the calculations.

Generally, all force fields take the functional form

$$E_{total} = E_{bonded} + E_{non-bonded}.$$

The bonded energy term pertains to the geometry of a molecule, a factor that greatly influences the energy of that molecule. These terms are collectively named valence terms. These can include expressions for the energy due to bond stretching, angle bending, and dihedral angles. The difference between these terms is the number of atoms that define each quantity. Bond stretching is determined by two atoms and is described by their distance from each other, angle bending by three atoms and the angle defined by these atoms, and dihedral angle by four atoms and the angle

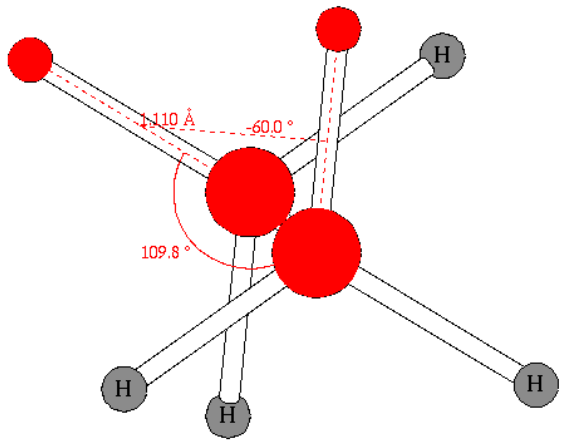


Figure 1 Ethane molecule depicting bond length, angle bending, and dihedral angle (shown here as -60.0°). This is the staggered conformation and is most stable.

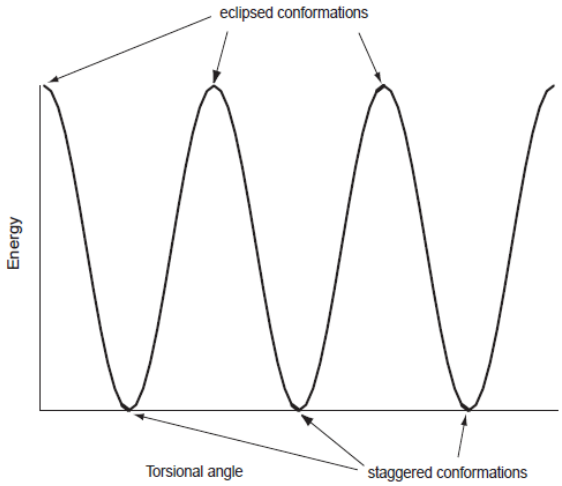


Figure 2 Energy as a function of dihedral angle in ethane, as pictured in Figure 1. Note the cosine behavior of the function.

between the two planes formed by them. Should a molecule choose to adopt a conformation that strains any of these properties, the energy of the molecule will increase and that conformation will be deemed unfavorable. For example, Figures 1 and 2 depict the dihedral angle formed in ethane and the energy of ethane as a function of the magnitude of this angle.

Note that the eclipsed conformation, in which the atoms overlap when viewing the molecule down the rotationary bond, results in a higher energy due to the proximity of these atoms. In this way, valence terms in the force field have a significant affect on the stability of a given molecule and parameters for each of these terms must be fit in order to maximize this stability. In

this particular case, dihedral energy terms contain the cosine function to reflect the energy curve given in Figure 2 and are generally in the form

$$E_{torsion} = \sum_i k_i[1 \pm \cos(n\theta)],$$

where *k* is a parameter that needs to be fitted based on quantum mechanical or experimental results and will change based on the types of atoms involved, steric hindrance, and other geometrical factors. Parameterizing a force field in this manner is a major part of its development, so there are many different ways in which this can be achieved. Parameters must be fit for multiple energy terms and parameterization protocols can become quite complicated as a result. The protocol for parameterizing the force field used in this research will be presented in a later section, but generally the technique involves mapping energy terms to geometric states and then developing coding routines that will use the potential energy surface of a molecule to output an optimal parameter set.

Force fields also contain energy terms that reflect non-bonded interactions. While valence terms result from geometries determined by the position of atoms that are bonded together, non-bonded terms describe the energy resulting from the interactions of charged particles, in which nuclei are treated as positive charges and electrons as negative charges. Non-bonded terms often include electrostatic and Van der Waals interactions that are treated separately. The Coulombic term adopts the point charge-point charge interaction formalism of

$$E_{Coulombic} = \frac{q_i q_j}{r_{ij}},$$

where i and j are particles, q is the charge for each respective particle, and r_{ij} is the distance between the two particles. Figure 3 depicts this energy between two such particles as a function of the distance between them.

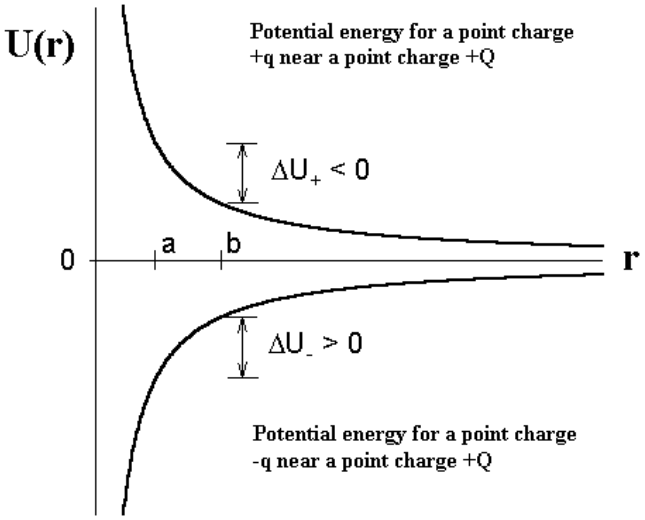


*Figure 3 Electrostatic potential as a function of distance from a charged particle. (*Note on electric, *1999)*

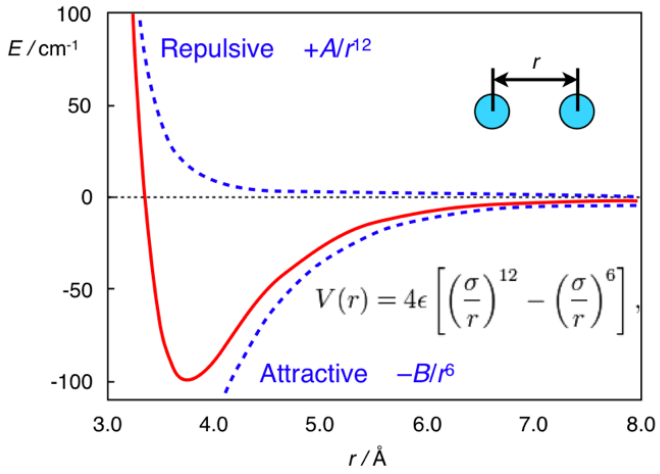


Figure 4 Lennard-Jones potential function. (Hansson & Jans, 2012)

On the other hand, Van der Waals interactions among atoms is included in the Lennard-Jones potential energy term, which describes all electrostatic interactions other than those due to point charges. Figure 4 depicts the energy trends for the Lennard-Jones term also in terms of the distance between two atoms. The equation shown in the figure is the form that many force fields take and is known as the 12-6 Lennard-Jones potential. There is also the 12-10 potential, in which the attractive term is treated more significantly. In the expression, ε denotes the depth of the well of the graph or the vertical distance from the x-axis to the function minima and σ is the value of r at which the potential is zero. σ is an important parameter because it defines the distance at which the attractive and repulsive forces cancel. In Figure 4, the potential function is asymptotic with the x-axis so the interaction energy is zero as the distance between the particles approaches infinity. Given two atoms that are at an infinite distance from one another, the Lennard-Jones potential is

zero. Upon bringing the two atoms into closer proximity, an attractive force will ensue from the dipole moments that are induced and the energy will decrease because the distance between the atoms is favorable. However, if the two atoms are brought further into proximity, orbital effects will need to be considered. According to the Pauli Exclusion Principle, two electrons of the same directional spin cannot occupy the same orbital. Thus, as orbitals come closer together and begin to overlap unfavorably, the energy will begin to increase, ultimately overcoming the initial attractive force and approaching infinity as the distance shrinks. The expression describing the Lennard-Jones energy is an example of an implicit treatment of electrostatic polarization. Some force fields will even combine this energy term with the Coulombic term into one that describes the total electrostatic energy. As we will see, such implicit treatment often leads to inaccurate results, especially when reproducing systems that contain atoms with large polarizable electron clouds and/or ions.

In summary, there are many different aspects of molecular properties that contribute to their stability and reactivity. In simulating these molecules, it is important to include as many of these aspects as possible in order to reflect the reality of nature. The principal aim of this study was to develop a force field that expanded the non-bonded term to include polarizability, in which electron charges around nuclei could move in response to an external electrical field and induce a dipole moment. However, describing polarization mathematically is complex and explicit and thorough treatment of electrostatic polarization can lead to the large CPU time that good molecular modeling practices seek to avoid. Thus, the goal was to introduce an expression that approximates the polarization energetic effect enough so that the computational expense is not too large but also to minimize the loss of accuracy resulting from this simplification. The following sections will discuss some of the force fields that are employed in computational laboratories today, both

polarizable and non-polarizable, and how well they reflect the quantum mechanical or experimental results upon which they are based.

Non-Polarizable Force Fields

These force fields employ electrostatic energy terms that do not attempt to explicitly define polarization. Here, we show what particular expressions such a force field uses and discuss some of its reported advantages and drawbacks. In a later section, we also present data on how these force fields compare in accuracy and robustness with respect to the specific systems we investigate in this study.

OPLS-AA

Jorgensen et al. initially developed the Optimized Potentials for Liquid Simulations – United Atom (OPLS-UA) force field in order to better describe intermolecular interactions in the liquid state. This force field, in contrast to that of OPLS-All Atom (OPLS-AA), implicitly treated hydrogens by only considering interaction sites that included non-hydrogen atoms. This was done in an attempt to minimize the CPU time required for computation, as it has been generally found that this amount of time is approximately the square of the number of interaction sites. Thus, for example, the number of interaction sites on ethane can be reduced from the full 8 to only 2, dramatically reducing computational time. However, it became clear that an explicit treatment of hydrogen sites was required in order to correctly account for charge distribution and torsional energy, ultimately boosting accuracy. This came in the form of the OPLS-AA force field, one of the most successful and chemically complete fixed-charge models. For example, the average error for the energy hydration of alkanes was reduced from 0.9 kcal/mol to 0.3 kcal/mol:

TABLE 8: Free Energies of Hydration Calculated with the OPLS Force Field; Intermediates in Parentheses Are for the All-Atom Simulations Only[a]

process	ΔG(all-atom)	ΔG(united-atom)	ΔG(exptl)[b]
$CH_4 \rightarrow (CH_4^{V_{DW}}) \rightarrow 0$	-2.20 ± 0.22	-2.46 ± 0.36	-2.00
$C_2H_6 \rightarrow CH_4$	0.25 ± 0.06	0.13 ± 0.20	0.17
$C_2H_6 \rightarrow (C_2H_6^{VDW}) \rightarrow 0$	-1.67 ± 0.18	-2.26 ± 0.33	-1.83
$C_3H_8 \rightarrow C_2H_6$	-0.22 ± 0.08	-0.98 ± 0.10	-0.13
$C_4H_{10} \rightarrow C_3H_8$	-0.29 ± 0.11	-0.27 ± 0.17	-0.12

Table 1 Comparison of OPLS-UA and -AA for calculation of hydrocarbon hydration. (Kaminski, 1994)

Additionally, results from liquid simulations have also been shown to yield good accuracy:

TABLE 4: Computed and Experimental Thermodynamic Properties of Liquid Ethane[a]

system	V	ϱ	ΔH_{vap}	C_p
all-atom				
2M	93.4 ± 0.5	0.535 ± 0.003	3.41 ± 0.02	19.7 ± 2.5
4M	92.2 ± 0.4	0.542 ± 0.002	3.45 ± 0.02	22.1 ± 2.1
6M	92.5 ± 0.3	0.540 ± 0.002	3.44 ± 0.01	2.15 ± 1.5
8M	92.5 ± 0.2	0.540 ± 0.001	3.44 ± 0.01	20.6 ± 1.5
united-atom				
2M	91.0 ± 0.7	0.549 ± 0.004	3.54 ± 0.03	27.4 ± 4.6
4M	91.2 ± 0.3	0.548 ± 0.002	3.53 ± 0.01	10.9 ± 1.7
6M	91.2 ± 0.2	0.548 ± 0.001	3.53 ± 0.01	19.0 ± 1.0
8M	91.2 ± 0.2	0.548 ± 0.001	3.53 ± 0.01	18.0 ± 0.8
experiment[b]	91.5	0.546	3.62	17.6

Table 2 Liquid ethane simulation results."2M" refers to 2 x 10^6configurations in the MC simulation. V (volume) is given in $Å^3$, ϱ (density) in g/cm^3, ΔH_{vap} (heat of vaporization) in kcal/mol, and C_p (heat capacity) in cal/(mol K). (Jorgensen, 1996)

The energy terms for OPLS-AA are as follows. The non-bonded interactions between molecules a and b are described by an energy term combining Coulombic and Lennard-Jones terms:

$$E_{ab} = \sum_i^{on\ a} \sum_j^{on\ b} [\frac{q_i q_j}{r_{ij}} + 4\varepsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{12}}{r_{ij}^{12}}\right)] f_{ij},$$

where i and j are atoms, q is charge, r is the distance between atoms i and j, ε and σ are Lennard-Jones parameters to be fitted, and f_{ij} is a coefficient. f_{ij} assumes the value 0 if atoms i and j are bonded (a 1-2 interaction) or are on the same bond angle (a 1-3 interaction). 1-4 interactions between these two atoms make the coefficient value 0.5 and 1.0 in all other cases. The Lennard-Jones parameters are combined between atoms i and j in a geometric fashion, such that

$$\sigma_{ij} = \sqrt{\sigma_{ii}\sigma_{jj}}$$

and

$$\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}}.$$

The bond stretch and angle bend terms are

$$E_{bond} = \sum_{bonds} K_r(r - r_{eq})^2$$

and

$$E_{angle} = \sum_{bonds} K_\theta(\theta - \theta_{eq})^2,$$

respectively. Here, K_r and K_θ are coefficients, r is the distance between two atoms and θ is the angle between three atoms. These energy terms are reminiscent of Hooke's law,

$$E_{stretch} = K(r - r_{eq})^2$$

and thus treats bond stretching and angles as spring-like vibrations. This formalism is generally conserved among molecular mechanics force fields and is included in the polarizable force field. Finally, the energy due to torsion is given by

$$E_{torsion} = \sum_i \frac{V_1^i}{2}[1 + \cos(\phi_i)] + \frac{V_2^i}{2}[1 - \cos(2\phi_i)] + \frac{V_3^i}{2}[1 + \cos(3\phi_i)],$$

where V^i is a Fourier coefficient that requires parameterization and ϕ_i is the dihedral angle.

Polarizable Force Fields

These force fields seek to explicitly treat electrostatic polarization. The two force fields presented here take different approaches in this approximation but each are a step forward in trying to accurately reproduce polarizable results. However, introducing polarizability is a non-trivial task. First, this involves adding additional parameters, which can exponentially increase the difficulty in parameterization. Simply put, imagine trying to solve a single variable equation, which is a fairly simple task. Suppose that another variable is introduced, so that now one is left to solve a system of equations in order to find two variables. Thus, for every additional variable and equation that is added, the more time consuming and challenging it is to determine each

variable. This is the challenge with polarization. A good approximation has to be implemented so that the benefit of increase in accuracy outweighs the additional CPU time required for calculations. For this reason, good error determinations are used to thoroughly assess the improved accuracy.

Generally, there have been two great limitations determined for force fields that do not explicitly treat electrostatic polarization, i.e. fixed-charge models. One is that, while these force fields perform well in calculating properties of pure liquids, doing so in non-condensed phases and non-uniform environments such as those found in biochemistry have resulted in large errors.

TABLE 3: Heats of Vaporization in kcal/mol, Molecular Volumes in Å^3, and Densities in g/cm^3 for the Pure Liquids

	ΔH_{vap}			V_{mol}			d		
liquid	PFF	OPLS	exp	PFF	OPLS	exp	PFF	OPLS	exp
CH_4	1.89	2.19[a]	1.96[a]	62.2	57.2[a]	62.8[a]	0.428	0.466[a]	0.424[a]
C_2H_6	3.32	3.44[a]	3.62[a]	94.4	92.5[a]	91.5[a]	0.529	0.540[a]	0.546[a]
C_3H_8	4.79	4.55[a]	4.49[a]	123.9	125.2[a]	126.0[a]	0.591	0.585[a]	0.581[a]
C_4H_{10}	5.62	5.43[a]	5.35[a]	157.2	161.3[a]	160.3[a]	0.614	0.598[a]	0.602[a]
CH_3OH	8.84	8.95[b]	8.95[b]	67.0	68.3[b]	67.7	0.794	0.779[b]	0.786
CH_3OCH_3	5.68	5.15	5.14	105.8	106.5[b]	104.1[b]	0.723	0.717	0.735[b]
CH_3COCH_3	7.92	7.24	7.48	127.5	121.2	123.0	0.756	0.795	0.784
NH_2COH	15.5	14.8[c]	15.5[d]	63.76	66.8[c]	66.3	1.175	1.120[c]	1.129[c]
NH_2COCH_3	13.2	13.7	13.4	112.1	109.3		0.876	0.897	
NMA	13.9	13.6[b]	13.3[b]	128.3	133.9	135.9	0.947	0.907[b]	0.894[b]
C_6H_6	8.20	8.05[e]	8.09[e]	153.7	148.6[e]	148.4[e]	0.844	0.873[e]	0.874[e]
C_6H_5OH	13.8	14.1[b]	13.8[b]	146.8	148.8[b]	147.8	1.065	1.050[b]	1.058[b]
CH_3SH	5.96	6.05	5.87[b]	92.5	89.5	90.0[b]	0.864	0.892[b]	0.888[b]
C_2H_5SH	6.73	6.79[b]	6.58[b]	121.9	120.7[b]	123.8[b]	0.846	0.855[b]	0.833[b]

Table 3 Pure liquid properties comparing OPLS and PFF with experimental values.(Kaminski 2004)

Above is a table comparing Polarizable Force Field (PFF) and OPLS in calculating the properties of various organic liquids. One can observe that OPLS and PFF display an approximately equal accuracy in reproducing experimental results. For example, OPLS had an error of 0.0 for the heat of vaporization of methanol, 0.6 for volume, and 0.007 for density, in comparison to PFF, which had errors of 0.11, 0.7, and 0.008, respectively. However, OPLS failed to provide good results for dimer calculations in the gas phase:

TABLE 4: Gas-Phase Dimerization Energies in kcal/Mol and Distances between Heavy Atoms in Å

	E_{dimer}			R		
molecule	PFF	OPLS	ab initio	PFF	OPLS	ab initio
CH_4	−0.44	−0.48	−0.50	3.86	3.77	4.06
CH_3OH	−5.63	−6.41	−5.59	2.81	2.78	2.80
CH_3OCH_3	−1.45	−1.69	−1.46	3.46	3.09	3.09
CH_3COCH_3	−5.74	−4.65	−5.76	3.38	3.45	3.28
NH_2COH	−14.0	−14.1[22]	−14.0	2.82	2.81[22]	2.84
NH_2COCH_3	−12.9	−13.9	−12.8	3.01	2.78	2.94
NMA	−14.1	−11.3	−14.4	2.90	2.83	2.92
C_6H_6	−3.14	−2.11	*b*	3.57		*b*
C_6H_5OH	−5.87	−7.99	−5.68	3.10	−2.74	2.98
CH_3SH	−2.77	−2.31	*b*	−3.77	3.74	*b*

Table 4 Dimerization properties for various compounds, comparing OPLS and PFF with ab initio quantum mechanical results.(Kaminski 2004)

Here, for methanol, OPLS had an error of 0.82 for energy of the dimer and 0.02 for dimer distance, versus 0.04 and .01 with PFF. OPLS does perform well with nonpolar aliphatic compounds such as methane, however. Dimerization energies and distances calculated by this point-charge force field with polar compounds have proven to be unreliable and thus the need for explicit treatment of polarization is apparent, especially for biochemical systems containing peptides with carbonyl-amino backbones as well as polar side chains.

A second issue with point-charge models is that, in agreement with the dimerization results presented above, hydrogen bonding interactions are not adequately represented. It has been shown that such intermolecular interactions, including hydrogen bonding between carbonyl and amine groups as well as hydrocarbon groups, require explicit polarization in order to reflect quantum chemical results (Stern, Kaminski 1999). Such interactions are critically important to the stability of proteins, as hydrogen bonding routinely takes place between residues and certain areas of the backbone. Energetic analyses of the conformations of several different polypeptides was performed by Stern et al. in order to confirm this concept. One of these included an alanine tetrapeptide, whose results are given in the following table:

TABLE 5: Alanine Tetrapeptide Conformational Energies (kcal/mol)

	ab initio[a]	OPLS	this work	this work
ϕ/ψ fitting[b]		no[c]	no[c]	yes
1[d]	2.71	2.56	2.19	2.88
2	2.84	2.20	1.61	1.84
3	0.00	−1.57	−0.35	0.22
4	4.13	3.33	3.13	3.69
5	3.88	4.32	3.14	3.70
6	2.20	2.94	2.64	1.45
7	5.77	3.85	5.64	5.48
8	4.16	6.79	6.22	5.38
9	6.92	5.82	5.54	6.74
10	6.99	9.35	9.86	8.21
RMS error[e]		1.47	1.34	0.71

Table 5 Energies of the various conformers of the alanine tetrapeptide, comparing the results from ab initio quantum mechanics to those of OPLS and a polarizable force field that incorporated fluctuating charge in order to represent polarizability. The results for this polarizable force field are shown here both with and without dihedral ϕ/ψ refitting from the alanine dipeptide, from which torsional parameters were transferred. (Stern et al., 1999)

From the above results it is clear that the fix-charge model failed to correctly reflect quantum mechanical data, which has been verified for excellent accuracy by Vargas et al. (Vargas 2001). The inaccuracy of point charge models in representing these charge-charge interactions can be explained by the fact that a given point charge will induce localization of an opposite charge in the immediate vicinity, and thus lower the overall energy. This is a result of polarization. If, however, such polarization were not permitted to take place, as in OPLS, such energy changes would not be observed and the energy would be overestimated. This is particularly important in conformational studies, in which hydrogen bonding is participant. Polarizable models correctly reflect these energy changes.

While explicit polarization has been shown to solve some of the problems present in fixed-charge models, polarizable force fields have some problems of their own, especially those force fields that adopt the Fluctuating Charge (FQ) formalism. This model is based on the electronegativity equalization method, which describes how electron clouds centered about a nucleus seek to equalize their electrochemical potential across the entirety of the cloud. Electrons are treated as dynamic variables and move according to Newtonian mechanics from areas of low

to high electronegativity (χ_i), which depends on the chemical potential (μ_i) in the following manner:

$$\mu_i = \frac{\partial E}{\partial N} = -\chi_i = -e\frac{\partial E}{\partial Q_i},$$

where E is the ground state energy of the electron gas, N is the number of electrons, Q is the atomic charge, and e is the fundamental electron charge. The model takes into account the fact that electronegativities of a given atomic site depends on its charge and the electronegativities of its environment. Thus, electrons in this model will fluctuate in an effort to equalize the atomic electronegativity, redistributing partial charges with changing electrostatic potential. (Rick et al., 1994)

One of these issues with FQ is replicating accurate results for bifurcated hydrogen bonding. These occur when either a hydrogen bond donor is interacting with two hydrogen bond acceptors, or when an acceptor is interacting with two donors. An example of this is the $C7_{ax}$ conformer of the alanine dipeptide, pictured below.

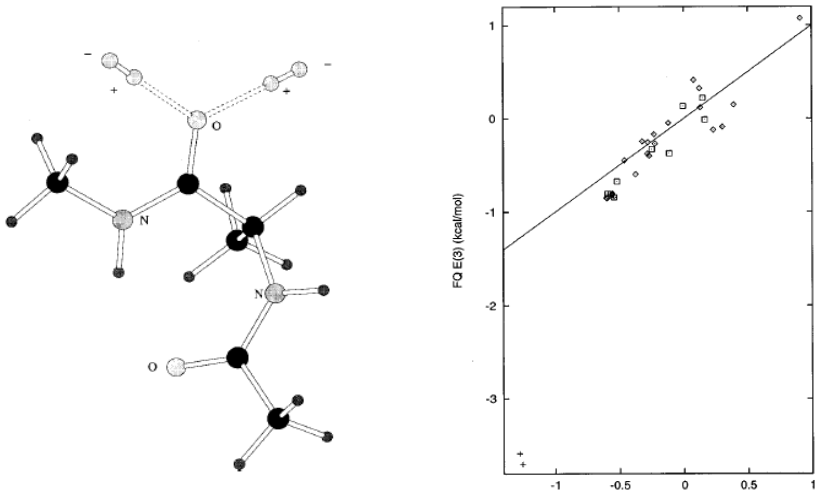


Figure 5 The alanine dipeptide C7ax conformation is shown to the left. Note the carbonyl accepting electrostatic interactions from the two fixed-charge dipole probes. A linear comparison of the Fluctuating Charge polarizable model with quantum mechanical results is shown to the right. The two plus signs indicate trimer results. (Banks et al 1999)

The two dipole-dipeptide trimer yielded poor agreement with *ab initio* quantum mechanical calculations, as can be observed by the graph in the above figure, in which the linear line represents perfect agreement between the two calculation methods. Such discrepancy in three-body energies has been noted by other groups, who have studied the three-body energies of water trimers and found differences in results between the FQ model and ab initio data in the case of bifurcated hydrogen bonding (Liu et al 1998). Liu et al. recognized three different scenarios in which bifurcated hydrogen bonding occurred between three water molecules if randomly taken out of a bulk solution of water. It was in these cases that the Fluctuating Charge model failed to adequately reflect quantum mechanical results.

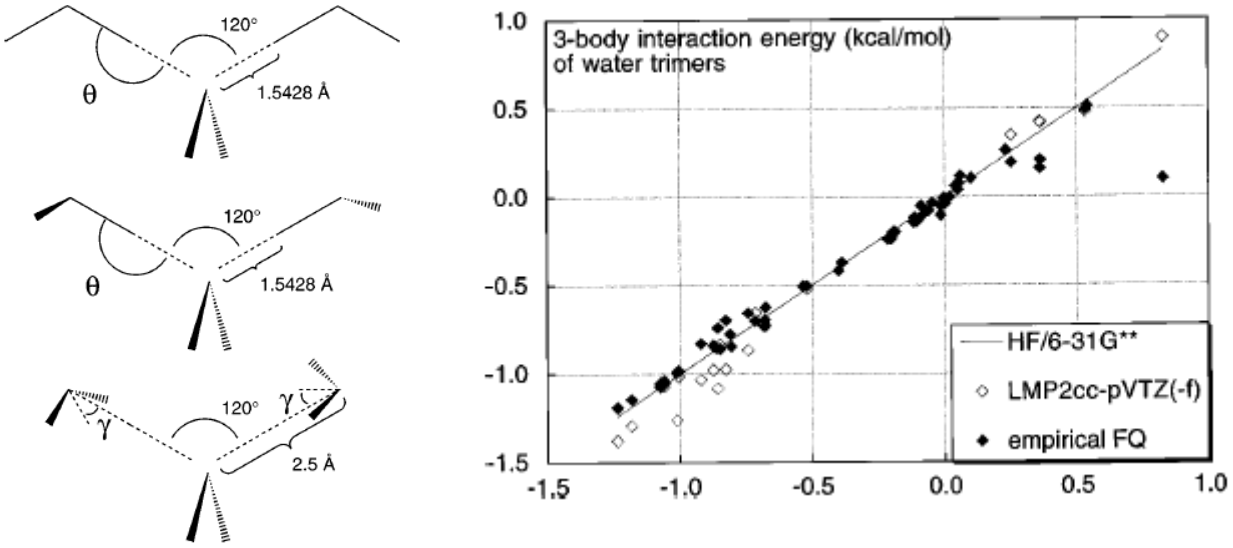


Figure 6 Three different bifurcated hydrogen bonds in water trimers is pictured to the left. The graph to the right compares FQ results to those of ab initio. The solid line represents perfect agreement with quantum results. (Rick et al., 1994)

Another problem encountered with the Fluctuating Charge model is the representation of out-of-plane polarization, which is actually quite similar to the problem with bifurcated hydrogen bonds. This stems mainly from the fact that FQ will be insensitive to the vectors of an induced electric field. Suppose there is an atomic site somewhere in space, and that two dipole probes are place adjacent to each other next to this site and oriented in the same way. Let us state that the

angle formed by these three points is zero. Further suppose that these two probes were to separate from each other so that the angle determined by the three particles were to increase and approach 180°, so that the probes are now opposite each other, with the atomic site in the middle. The electrochemical potential the atomic site will be equal in these two cases according to FQ. To illustrate this point, out-of-plane polarization cases in which an aromatic ring has been subjected to an external electric field due to two dipole probes have been tested with different force fields in comparison to *ab initio* data:

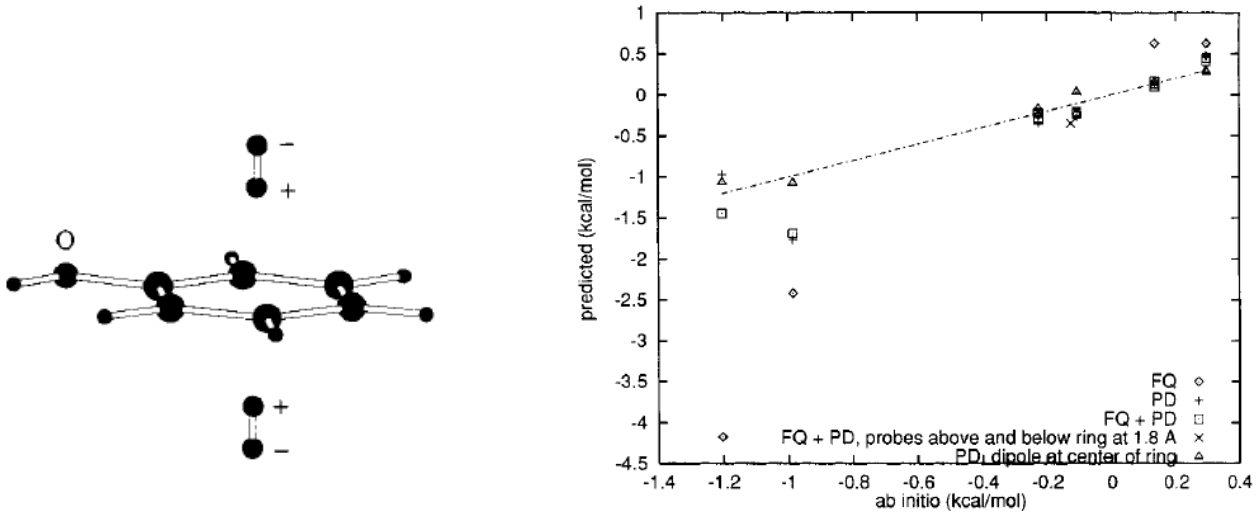
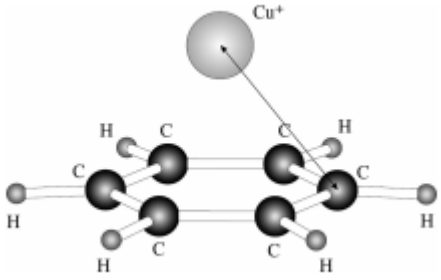


Figure 7 A linear comparison of FQ, PD (Polarizable Dipoles), and combinations thereof. The dotted line represents perfect agreement. The lowest energy trimer corresponds to the figure at left, demonstrating out-of-plane polarization. In addition, the second lowest energy trimer corresponds to a bifurcated hydrogen bond at the oxygen site. (Stern et al., 1999)

Having described some of the advantages and disadvantages of polarizable force fields, it is also important to consider how to best develop one. One of the biggest issues with this development is the functional form that the force field should take, which is essentially the mathematical description of polarization. One method, whose results have been analyzed here, is the fluctuating charge method, or FQ, that was developed by Berne et al. in 1994 in applications to liquid water (Rick et al., 1994). However, as has been shown, this model fails to perform accurately in several key cases that are required for biochemical modeling, which actively involves

multi-body interactions and contains many instances of hydrogen bonding. This may be due to the same reason that fixed-charge models such as OPLS fail, since the development of FQ was based on a pure liquid. Since the basis for the force field was homogeneous, it may not respond very well to heterogeneous solutions such as those present in biochemical systems. The functional form that is taken by the force field used here, POSSIM, as well as its parent force field, PFF, is the induced dipole and has shown great promise in replicating both experimental and quantum mechanical results for systems that include peptides and peptide-ligand interactions. These results will be discussed in a later section titled "Peptide Modeling."

The out-of-plane polarization problem has been resolved using an inducible dipole formalism with copper-benzene complex calculations. Geometry optimization was performed on a copper (I) ion and a benzene ring, with the ion sitting directly above the ring so as to interact with the electron clouds electrostatically. The results are shown here:



model	energy	$Cu^+ \cdots C(C_6H_6)$ distance
OPLS2005	−14.0	2.77
OPLS, refitted for TIP3P	−25.2	2.14
OPLS, refitted for TIP4P	−26.0	2.11
PFF	−54.4	2.30
reference	−56.9 to −61.3	2.31

Figure 8 A visual model of the Cu (I) – benzene ring complex is shown at left, while calculation results for each model is shown at right. Parameters were refit for OPLS in the hopes that energy agreement with QM would improve. (Ponomarev et al., 2011)

Here, the inducible dipole model PFF gave an average error of 4.7 kcal/mol and 0.01 Å, while OPLS gave best resultant error of 33.1 kcal/mol and 0.2 Å. The fixed-charge model significantly underestimated the interaction energy, probably due to the fact that it did not take into account the favorable energy drop due to induce dipole affinity.

The success of the inducible dipole functional form has already been shown to improve upon the OPLS results for gas-phase dimers as well as the FQ results for out-of-plane and bifurcated hydrogen bonding. POSSIM uses a second-order approximation of the full inducible dipole formalism adopted by PFF, whose mathematical descriptors are described in the next section. This approximation has been made in an attempt to reduce the computational cost required for calculating explicit polarization. Therefore, it has been tested to ensure that no sacrifice in accuracy is present in the model in exchange for the reduced CPU time. Kaminski et al. calculated the dimerization energies of several biochemically important aliphatic hydrocarbons in the gas phase and compared the results to those of the full-dipole PFF model, OPLS, and QM:

	energy				distance			
system	QM[a]	OPLS[a]	PFF0[b]	PFF[c]	QM[a]	OPLS[a]	PFF0[b]	PFF[c]
H_2O-H_2O O···O	−5.02	−6.78	−5.54	−4.52	2.91	2.68	2.88	2.91
CH_4-CH_4 C···C	−0.5[d]	−0.48	−0.44	−0.48	3.7[d]	3.77	3.86	3.76
MeOH−MeOH O···O	−5.59	−6.41	−5.63	−5.59	2.80	2.78	2.81	2.81
MeOH−OH_2 O···O	−4.90[e]			−5.12	2.86[e]			2.80
MeHO−HOH O···O	−4.77[f]			−4.93	2.91[f]			2.90

Table 6 Gas-phase dimerization energies as computed by QM, OPLS, the full dipole PFF (PFF0), and POSSIM (PFF). (Kaminski et al., 2009)

The approximation model predicted all energies to within 0.5 kcal/mol, which shows good agreement with quantum mechanical results. Additionally, dimers of hydrocarbons with water have been analyzed as a test of realistic solvation, which had not been previously implemented in the development phase. Liquid phase calculations of the heats of vaporization as well as molecular volumes were performed with pure liquids of water and several hydrocarbons:

	ΔH_{vap}				V			
system	exptl[a]	OPLS[a]	PFF0[b]	PFF[c]	exptl[a]	OPLS[a]	PFF0[b]	PFF[c]
H_2O	10.51	10.46	10.54	10.577 ± 0.055	30.0	30.0	30.15	29.891 ± 0.226
CH_4	1.96	2.19	1.89	2.209 ± 0.026	62.8	57.2	62.2	56.863 ±0.381
C_2H_6	3.62	3.44	3.32	3.456 ± 0.063	91.5	92.5	94.4	92.140 ± 0.571
C_3H_8	4.49	4.55	4.79	4.499 ± 0.129	126.0	125.2	123.9	126.096 ± 0.903
C_4H_{10}	5.35	5.43	5.62	5.335 ± 0.164	160.3	161.3	157.2	163.433 ± 1.416
NMA	13.3	13.6	13.9	13.319 ± 0.372	135.9	133.9	128.3	135.670 ± 0.566
CH_3OH	8.95	8.95	8.84	9.005 ± 0.140	67.7	68.3	67.0	67.447 ± 0.535
average error		0.129	0.240	0.083		1.574	2.450	1.485

Table 7 Liquid phase calculations with the POSSIM (PFF), full-dipole model (PFF0), and OPLS force fields in comparison to experimental results. Simulations were performed using the Monte Carlo method. (Kaminski et al., 2009)

System	Second-order model		Full Dipole Model, Tolerance = 0.01		Full Dipole Model, Tolerance = 10^{-6}	
	Time	E_{pol}	Time	E_{pol}	Time	E_{pol}
H_2O	0.169	−545.6	0.673	−580.9	1.878	−581.1
CH_3OH	0.057	−147.7	0.212	−160.1	0.550	−160.3
CH_3NH_2	0.165	−81.24	0.514	−86.12	1.470	−86.15
CH_3SH	0.028	−41.67	0.278	−51.32	0.832	−51.33
NH_2COCH_3	0.062	−200.3	0.183	−206.2	0.422	−205.9

Table 8 Monte Carlo simulation time comparisons between the POSSIM second-order model and the full PFF model. Times are those required for each simulation step. Tolerances displayed here are convergence criteria and are in units of charge units •Å. (Kaminski et al., 2003)

The average POSSIM error for heats of vaporization were 0.083 kcal/mol, as opposed to PFF0 and OPLS which displayed errors of 0.129 and 0.240 kcal/mol, respectively. As for molecular volume, POSSIM had 1.485 Å^3, which is significantly better than that of PFF0 and slightly so for OPLS. Additionally, Table 8 demonstrates the differences in the computational time that the second-order approximation and full model required for the same length Monte Carlo simulation. The average time required to simulate each of the five molecules listed were .0942 s, .372 s, and 1.03 s for POSSIM, PFF (tolerance = 0.01), and PFF (tolerance = 10^{-6}). Thus, it has been shown that the second order approximation model is not only appropriate for gas and liquid phase calculations, but also that this model achieves roughly the same level of accuracy as the full dipole model with a substantial decrease in computational cost. It should also be noted that N-methylacetamide (NMA) is included in this and many other biochemical computation works because it represents the most fundamental unit of a protein. This molecule is pictured at left. The peptide bond characteristic of the backbone is present, connecting the carbonyl carbon to the nitrogen. The methyl groups two either side of the molecule normally bear amino acids. Thus, this molecule is often used as an initial stepping stone in force

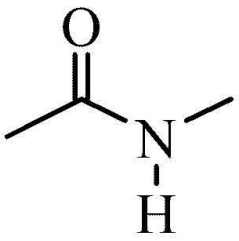


Figure 9 NMA

field development so that the researcher can be sure that larger peptides will be modelled accurately.

POSSIM

The POlarizable Simulations with Second order Interaction Model (POSSIM) force field has shown great promise so far in being able to replicate organic and biochemical systems. This model adopts the traditional formalism of OPLS for bonded terms and the Lennard-Jones potential, but adds on an explicit electrostatic energy term. As opposed to FQ, POSSIM treats polarization as an induced dipole μ:

$$E_{pol} = -\frac{1}{2}\sum_i \boldsymbol{\mu}_i \boldsymbol{E}_i^0,$$

where

$$\boldsymbol{\mu_i} = \alpha_i \boldsymbol{E}_i^{total}$$

and $\mathbf{E}^{total}$ is the total electric field, including that due to induced dipoles. Furthermore, α_i is the scalar polarizability coefficient, which is a crucial parameter that requires adjusting when developing atom types for the force field. The smaller α is, the less polarizable the atom. However, the polarizability coefficient is typically entered as a parameter in reciprocal form. Thus, hydrogen atoms are given an α^{-1} of 9999.99 and oxygen atoms some value less than 1 in order to reflect their relative electron cloud sizes. $\mathbf{E}^{total}$ is given by the expression

$$\boldsymbol{E}_i^{total} = \boldsymbol{E}_i^0 + \sum_{j \neq i} \boldsymbol{T}_{ij} \boldsymbol{\mu}_j,$$

where the dipole-dipole interaction tensor $\mathbf{T}_{ij}$ is expressed as

$$\boldsymbol{T}_{ij} = \frac{1}{R_{ij}^3}\left(\frac{3\boldsymbol{R}_{ij}\boldsymbol{R}_{ij}}{R_{ij}^2} - \boldsymbol{I}\right)$$

and $\mathbf{E}^0$ is the electric field matrix in the absence of induced dipoles. **I** is the unit tensor and R_{ij} is the distance between atoms i and j. Thus,

$$\boldsymbol{\mu_i} = \alpha_i \boldsymbol{E}_i^0 + \alpha_i \sum_{j \neq i} \boldsymbol{T}_{ij} \boldsymbol{\mu}_j$$

The above equation contains a summation that is taken iteratively. This model uses a second-order approximation of this equation, so that it becomes

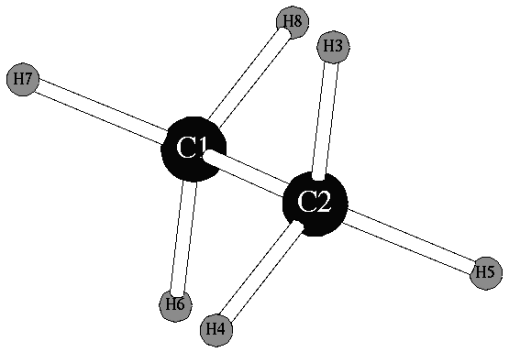
$$\mu_i^{II} = \alpha_i \boldsymbol{E}_i^0 + \alpha_i \sum_{j \neq i} \boldsymbol{T}_{ij} \boldsymbol{\mu}_j^0$$

$$= \alpha_i \boldsymbol{E}_i^0 + \alpha_i \textstyle\sum_{j \neq i} \boldsymbol{T}_{ij} \alpha_j \boldsymbol{E}_i^0 .$$

Geometry Optimization

One of the methods used in this research to test the accuracy and functionality of the force field was to perform geometry optimizations on the molecules in question. These protocols seek to search the potential energy surface of the molecule, which is a function of its shape and conformation, in order to find a minimum and output the corresponding geometry. The geometry of a single molecule is composed of the corresponding bonded terms in the force field: bond stretches, angle bends, and dihedral angles. However, here we perform geometry optimizations of dimers, and so the non-bonded terms are also included. Using geometry optimizations in developing a force field is critical because the physical parameters describing the atoms in a system must be optimized to reflect reality. Often, quantum mechanical geometry optimizations are performed first as a basis and then geometrical measurements are taken so as to benchmark what the force field under development ought to output. If a force field can be shown to recreate the same potential energy surface as a quantum mechanical or experimental method, then the experimenter can be sure that the force field can be used for further geometry-energy predictions.

Geometry optimizations begin with defining the initial input structure of the molecule in question. Since the geometry is going to be perturbed during the potential surface search process, it is most convenient to define a molecule in the form of a z-matrix:




```
C
C  1  CClength
H  1  CHlength   2  CCHangle
H  1  CHlength   2  CCHangle   3  120.0
H  1  CHlength   2  CCHangle   4  120.0
H  2  CHlength   1  CCHangle   3   60.0
H  2  CHlength   1  CCHangle   4   60.0
H  2  CHlength   1  CCHangle   5   60.0

CClength   1.5
CHlength   1.0
CCHangle  109.5
```

Figure 10 An example Z-matrix of ethane. Note how convenient it is to define all lengths and angles initially, so that it is easier for the program to modulate these values.

Returning to the simple example of ethane, its z-matrix begins by defining the first atom, C1. This is usually positioned at the coordinate (0,0,0). Then, the second atom, C2, is defined through a distance from C1 and is positioned at (CClength, 0, 0). The third atom is then defined through the angle it forms with C1 and C2, and the fourth atom is defined through the dihedral angle it forms with the first three atoms. Every atom thereafter is defined in the same way as this fourth atom, except that its definition may not be based on the immediately previous atoms. For example, in this Z-matrix H4 is defined through its angle with C2 and C1, rather than H3 and C2. This is because the latter set of atoms are not bonded to each other and the geometrical values cannot be defined this way. An important facet to keep in mind is that there are many different permutations of a Z-matrix, that is, there are multiple ways of defining the matrix for one molecule. Generally, especially with larger molecules such as peptides, it is crucial that atom definitions follow the backbone of the molecule as much as possible. Otherwise, the optimization algorithm will perturb side chain atoms that do not have very much impact on the overall energy of the molecule, and the result will not have searched enough of the potential energy surface to find a good minimum.

Another method of defining input geometry is to simply list the coordinates of each atom:

C	0.000000	0.000000	0.750000
C	0.000000	0.000000	-0.750000
H	0.000000	0.942641	1.083807
H	-0.816351	-0.471321	1.083807
H	0.816351	-0.471321	1.083807
H	0.816351	0.471321	-1.083807
H	-0.816351	0.471321	-1.083807
H	0.000000	-0.942641	-1.083807

Figure 11 Coortesian input for ethane.

While this format is convenient for many-molecule calculations and simulations, it is not as good for geometry optimizations. The program will try to determine which atoms are bonded to each other by plotting each of the atom coordinates given and then seeing which interatomic distances are reasonable enough for bonding. However, adjusting geometrical parameters will not be as smooth as with a Z-matrix, and so the trajectory of the optimization may take the following form:

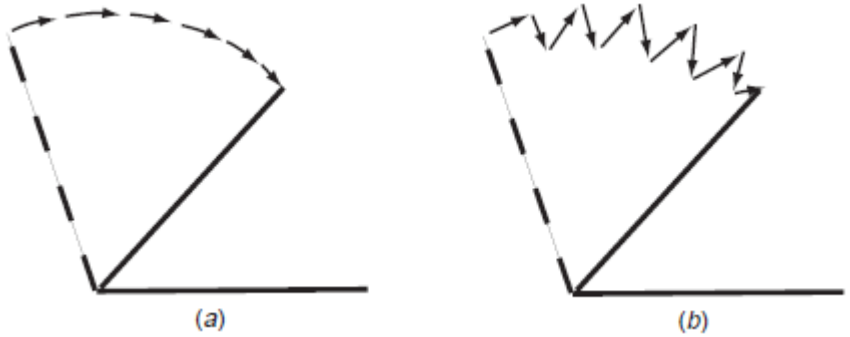


Figure 12 The paths taken to adjust an angle bend when the molecule is defined through (a) a Z-matrix and (b) Coortesian coordinates.

There are numerous algorithms for finding the energy minimum. The most basic method is simply perturbing each of the parameters in the Z-matrix and calculating the resultant energy, which is ultimately time consuming and inefficient. More accurate methods employ second derivative or Hessian matrices in order to monitor the gradient of the energy surface as a function of each parameter, and will thus follow the gradient downwards until a minimum is reached. Regardless of the particular method being used, the output geometry is always visually investigated to ensure that it is reasonable and no extraneous strains or steric hindrances present.

Molecular Dynamics and Monte Carlo Simulations

While force fields provide the energy gradient for the movement of molecules by mathematically describing which interactions are favorable and which are not, the simulation of a chemical system still requires that an algorithm be used to predict how molecules will behave once a force field is applied. Such an algorithm uses the energetic results from the force field expressions and try to determine where molecules will move. After all, particles in nature, at least in the liquid and gas phases, are subject to motion. There are two classes of simulation algorithms that seek to make these predictions: Molecular Dynamics (MD) and Monte Carlo (MC). Both of these techniques have been shown to be reliable in predicting molecular behaviors but rely on vastly different approaches in their forecasts. However, simulations in general are unique in that they take temperature into account with thermodynamic sampling, whereas geometry optimizations do not.

Molecular Dynamics

Unlike Monte Carlo simulations, MD simulations are time-dependent and averages structures over real time. In addition, each particle in an MD simulation is propelled by a force in accordance with Newton's second law of motion, the basis of this simulation technique. It is important to note that minutes viewing a trajectory of an MD simulation are often only on the order of milliseconds in real time. An MD trajectory with frames picoseconds apart may seem to be moving at a reasonable pace but in fact these molecules move extremely quickly around in solution in real time. So what algorithm does MD use?

1. The user usually builds the initial configurations of molecules in the preferred system with the software provided. In biochemical research, proteins and associated complexes are

provided by their PDB structures. Additionally, MD software often allows explicit solvation.

2. Given the starting setup of the input structures, the program will begin the simulation by choosing an initial velocity for each atom from a Boltzmann distribution, depending on the target temperature set by the user, which is usually room temperature at 298 K. This distribution is pictured in Figure 13.

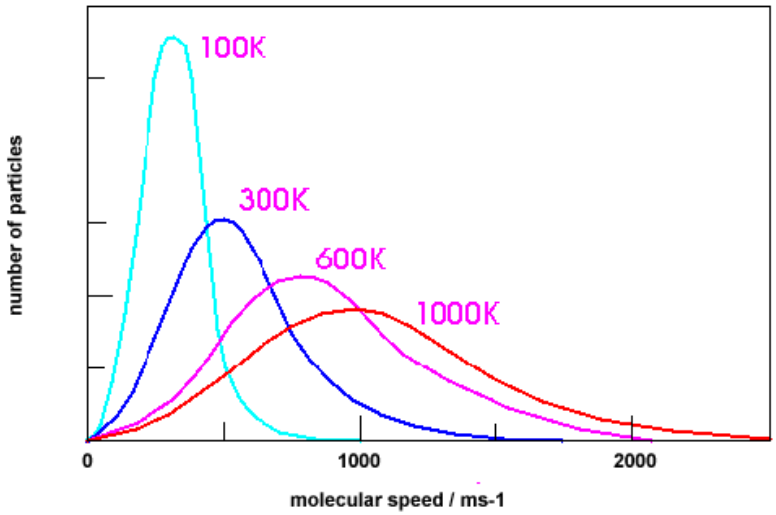


Figure 13 Various Boltzmann distributions based on the temperature of the system. As the temperature increases, there is a wider range of molecular speeds and thus, having the number of particles conserved, fewer particles having a given molecular speed. http://ibchem.com/IB/ibnotes/full/sta_htm/Maxwell_Boltzmann.htm

3. The momentum of each atom is computed using the velocity chosen and the mass of the atom.
4. The force acting on each atom is calculated based on the molecular mechanics force field being implemented in the simulation.
5. A new location for each atom is determined for the next period in time following the time step period set by the user (usually a fraction of a picosecond; if it is any larger the calculations will not converge). The program integrates Newton's law of motion $F_{net} = ma$ over the duration of the time step based on the integration algorithm being used.

6. Given the new location, a new velocity and acceleration is calculated from the force field equations.
7. Steps 3-6 are repeated for the duration of the simulation that is set by the user.

The ultimate goal in simulating chemical systems with MD is to run it for a sufficient number of iterations so that the system reaches equilibrium, that is, the total energy becomes relatively constant. It is important to note that equilibrium is not necessarily reached when the potential energy surface meets a minimum, but we will see this later with some of the simulations run here. Typical MD simulations with rather small systems require about 10^6 steps in order to reach equilibrium, however, with larger systems that contain macromolecules such as proteins, equilibrium may not be reached for 5×10^6 steps or more. Once equilibrium is achieved, the energy is stable enough for analysis and conclusions can be drawn about how certain molecules in the system interact. The total time in an MD simulation is simply the product of the magnitude of the time-step and the number of steps. So, for example, if the time-step is 0.001 ps and the simulation has been run for 10^6 steps, the total time is 1000 ps, or 1 ns.

There are several different algorithms used in MD to integrate the net force over the time step. The most popular is the Verlet method due to its low computational cost and time required. Simply put, the algorithm seeks to solve the equation

$$\overrightarrow{x_n} = \vec{x}(t_n),$$

where

$$t_n = t_0 + n\Delta t$$

and Δt is the size of the time step and n is the current time step number, by defining the initial position of a given atom as

$$\overrightarrow{x_0} = \vec{x}(t_0).$$

and iterating the following standard equation

$$\overrightarrow{x_{n+1}} = 2\overrightarrow{x_n} - \overrightarrow{x_{n-1}} + a(\overrightarrow{x_n})\Delta t^2$$

to determine the next position after the end of the time step.

The velocities used for the above algorithm will vary depending on the type of ensemble being used for the simulation. An ensemble is a collective term describing the macromolecular properties of a system containing many particles with differing micromolecular properties. For example, there is the NVT ensemble, which is the ensemble used for the simulations run in this study, which states that the number of molecules, the volume, and temperature of the system ought to be the same throughout the simulation. There is also the NPT and NVE ensembles, which seek to keep the pressure and energy of the system constant, respectively. What is different between this ensembles is the goal of the researcher. Should one be interested in simulating a system that is in the gas phase, for example, the NPT ensemble should be used and the size of the system will change accordingly to stay at the pressure set by the user. Since most biochemical systems remain in the liquid phase with relatively constant density, we use the NVT ensemble in the work presented here.

Periodic Boundary Conditions

In keeping with the idea that a good molecular modeling protocol minimizes the computational cost required, it is essential to work with as few molecules as possible in order to reduce the CPU time. However, the protocol must do so sparingly in order to preserve the accuracy of the representation of the natural system. One problem that must be overcome is that these natural systems often contain many millions of molecules. It is impractical to simulate all of these molecules at once in a computer program because the computational cost would simply be too

great. One approach at simplifying this is through the use of periodic boundary conditions. Figure 14 is a simple depiction of this concept.

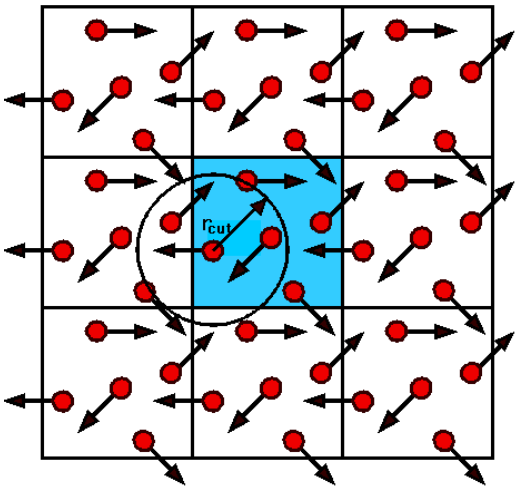


Figure 14 Simple demonstration of periodic boundary conditions, with red dots indicating atoms. The r_{cut} is the cutoff radius for force calculation about each atom.http://www.compsoc.man.ac.uk/~lucky/Democritus/Theory/pbc-mi.html

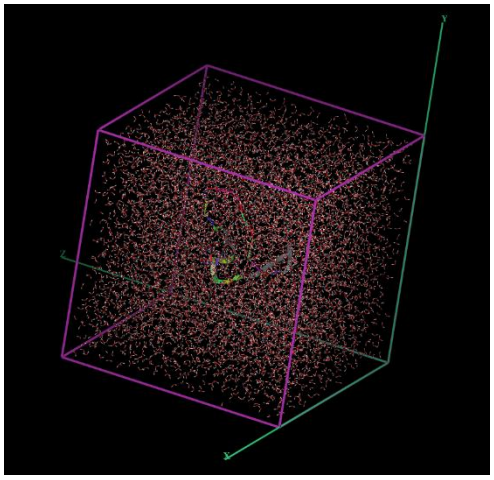


Figure 15 An example of periodic boundary conditions used in this research. The lightly colored protein can be seen with water molecules surrounding it. The purple and green lines delineate the boundaries.

Essentially, this approach takes advantage of the fact that often a molecular system is composed of some solute(s) and a solvent composed of the same molecules. In biochemical systems, this solvent is almost always water. A solute in question is placed in some solvent in a cubic box of a certain size, which is usually twice the size of the solute system so that there is enough space for solvent-solute interactions. While calculations are only performed in this one box, the box is iterated infinitely around itself, as is shown in Figure 14. Thus, should a molecule drift to the edge of the box and out of it, the molecule will appear on the opposite side of the box with the same velocity and acceleration vectors as before. Periodic boundary conditions severely cut down the CPU time required to perform calculations and is used in the context of this research, as is demonstrated in Figure 15, which shows a protein and ligand (not seen here) as well as the water solvent in a defined box.

An error can be encountered with the periodic boundary approach, however. If all intermolecular interactions are considered, then long range interactions will be inaccurately represented. Take for example an atom this is positioned in the corner of the box and one is interested in calculating the energy due to its interaction with an atom that is on an opposite corner of the same box. The researcher will notice that this same atom is also positioned in the same corner of the adjacent simulation box, so a short range and long range interaction energy will be calculated that is inaccurate. For this reason, a cutoff radius is applied so that all interactions beyond this radius will not be calculated, as is shown in Figure 14. This cutoff must be less than half of the length of an edge of the box so as to avoid double counting interaction energies. In addition, Ewald summation can be used to estimate long-range electrostatic effects, but the Lennard-Jones potential is generally short-range so it does not require a designated cut-off.

Additionally, certain constraints can be applied to an MD simulation if it is especially large in order to decrease computational cost. One possible constraint is to eliminate all bond rotations, so that the simulation will only proceed with the input conformations of the molecules. Additionally, water molecules can be constrained or frozen in their structure but not position, so as to not have to calculate the bond stretch and angle bend energy terms for the thousands of water molecules present in the system.

Monte Carlo

Monte Carlo (MC) simulations take a more probabilistic approach than the integration method used by MD. Generally, an MC will rely more heavily on the energy of a system than the forces acting on each individual molecule. The algorithm for an MC abides by the following steps:

1. An initial configuration of atoms in the system is chosen, usually inputted by the user as in MD.

2. The current energy of the system is computed.
3. Attempt a random trial move for an atom. The program will try to make such a move follows the Boltzmann distribution.
4. Compute the hypothetical energy of the system after having moved each atom.
5. Determine if this new energy is reasonable. If it is, accept the move; otherwise, reject the move and revert all coordinates to their previous values and begin from step 3 again. Acceptance or rejection of a move is based on the Metropolis-Hastings algorithm, which follows a probability distribution of random sampling. In essence, the new move will be accepted if the resultant energy is lower than the previously, or if it is not too much higher than what is allowable within the Boltzmann distribution. Ultimately, each new set of coordinates is termed a configuration.
6. Repeat steps 2-5 for the number of configurations set by the user.

One difference between an MC and an MD is that the real time equivalent of the simulation duration cannot be determined in an MC. However, the goal of reaching equilibrium is still the same, even if MD averages structures over time and MC over ensembles; in any case, as long as convergence has been reached, the properties of a system using either method ought to be the same at the end of the simulation. Each system is different, so there is no standard for how many configurations to run in order to reach equilibrium. At the end of the run, an output file will be created that includes the number of accepted and rejected moves, so the user can determine how favorable the simulation was. If the size of the moves attempted in step 3 is too large or a Boltzmann distribution of movement probabilities is not reached, the number of rejected moves will be very large. A good acceptance ratio, defined as the number of accepted moves divided by the number of total attempted moves, is usually anywhere between 0.5 and 0.7. Ensembles such

as those used in MD can also be applied here, but periodic boundary conditions cannot. One advantage of the MC is that it takes less CPU time to run a simulation of the same size, since it does not need to calculate force vectors for each particle in the system. Simulations used here, for example, in an MC take only a few hours, while the same system in MD may take several days. Additionally, larger noise in the progress of the system energy has been observed with MD simulations.

Peptide Modeling

The POSSIM model has shown to be reliable in full peptide contexts. Having developed parameters for the protein backbone analog, NMA, Ponomarev et al. (2011) confirmed the transferability of these parameters to the alanine dipeptide molecule. Below are results describing the conformational errors of this peptide in comparison to those of OPLS and QM:

Conformer	Energy			φ			ψ		
	QM	OPLS	POSSIM	QM	OPLS	POSSIM	QM	OPLS	POSSIM
$C7_{eq}$	0.00	0.00	0.00	−81.4	−79.5	−77.2	85.6	61.8	34.4
C5	1.00	0.91	1.37	−160.5	−149.8	−160.3	165.9	159.9	159.2
$C7_{ax}$	2.71	2.40	2.17	70.3	77.5	78.1	−76.8	−46.6	−36.2
β_2	2.56	2.82	2.77	−105.1	−105.1	−105.1	10.6	10.6	10.6
α_L	4.21	5.96	5.79	68.3	68.3	68.3	22.4	22.4	22.4
α'	5.47	5.96	5.98	−162.0	−156.5	−162.9	−33.2	−48.5	−38.0
PII	2.78	2.18	3.52	−85.0	−85.0	−85.0	160.0	160	160.0
α_R	2.71	2.39	0.99	−83.7	−83.7	−83.7	−3.9	−3.9	−3.9
error	–	0.73	0.97	–	3.2	1.6	–	9.4	12.9

Table 9 Comparison of the energies and φ and ψ angles for each alanine dipeptide conformer as calculated via QM, OPLS, and POSSIM. Average errors are shown in the bottom row. (Ponomarev et al., 2011)

Given the relatively good error for our polarization model, the data demonstrates that bonded parameters can be transferred. In testing the transfer of non-bonded parameters for those including electrostatic and Van der Waals forces, the group investigated the geometries and binding energies associated with an alanine dipeptide-water dimer:

property	Structure A			Structure B		
	QM	OPLS	POSSIM	QM	OPLS	POSSIM
Binding energy	−10.71	−10.04	−9.75	−11.68	−11.79	−12.24
R(O…O)	2.83	2.81	2.82	2.83	2.75	2.74
R(O…N)	3.10	2.94	3.06	3.05	2.94	2.98
φ, dimer	−83.4	−83.4	−83.4	−84.3	−84.3	−84.3
ψ, dimer	90.3	90.3	90.3	131.2	131.2	131.2
φ, monomer	−83.4	−83.4	−83.4	−84.3	−84.3	−84.3
ψ, monomer	90.3	90.3	90.3	131.2	131.2	131.2

Table 10 Properties of alanine dipeptide-water dimer. Structures A and B contain different positions of the water molecule, which hydrogen bonds twice. (Ponomarev et al., 2011)

While OPLS seems to demonstrate better binding energy agreement with QM, having an error of 0.89 kcal/mol vs 1.12 kcal/mol for POSSIM, this difference was also observed with NMA-water dimers, in which the POSSIM model underestimated the interaction energy by 0.89 kcal/mol on average. Since these parameters were transferred, this is not unexpected. However, the hydrogen bonding lengths determined by POSSIM were 0.09 Å closer to QM results than OPLS on average, and the average conformational angle error for POSSIM was 5.3º vs. 12.8 for OPLSº. Generally, both models are in good agreement with quantum mechanical results, which indicates two things: (1) parameters can be transferred without modification between model molecules and full peptides and result in good property calculations and (2) such parameters are reliable enough to then be transferred to larger polypeptides.

Thus, parameters were then transferred to a tridecaalanine peptide (Ala-13) and simulated in water in order to see if the POSSIM force field is robust enough to maintain this peptide's alpha-helix conformation.

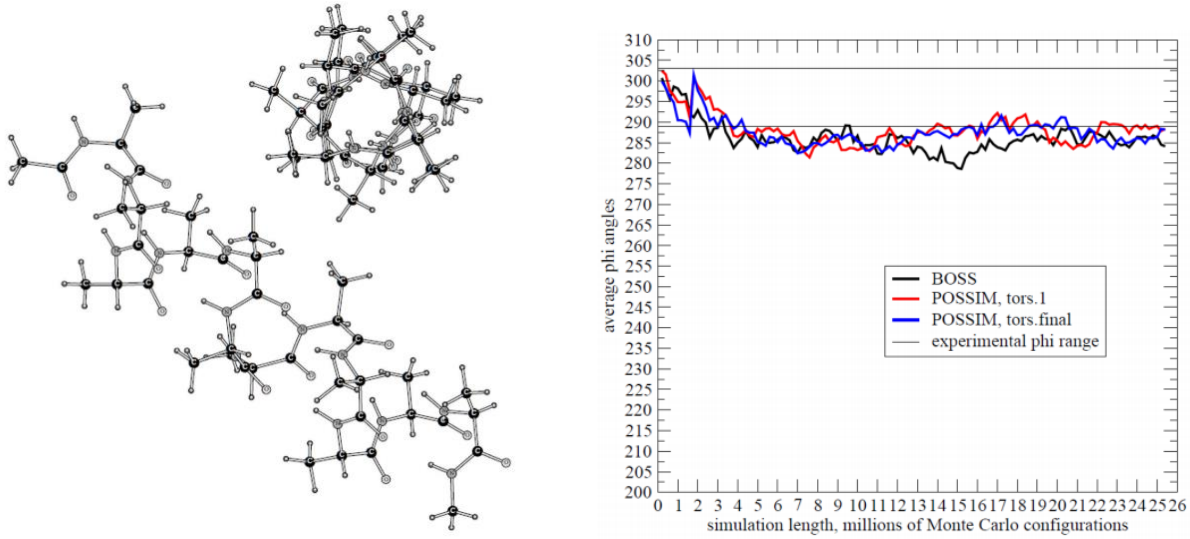


Figure 16 Side and down-the-barrel views of the tridecaalanine peptide are given to the left, while the right graph shows a recording of the average φ angles throughout the 40×10^6 *configuration simulation as compared to the experimental measures. (Ponomarev et al., 2011)*

The above results show that not only did the alpha-helix form in aqueous solution with the polarizable model but also that this conformation was stable over the course of the simulation. (Kaminski et al., 2011) This is a large step in the development of parameters for full scale proteins,

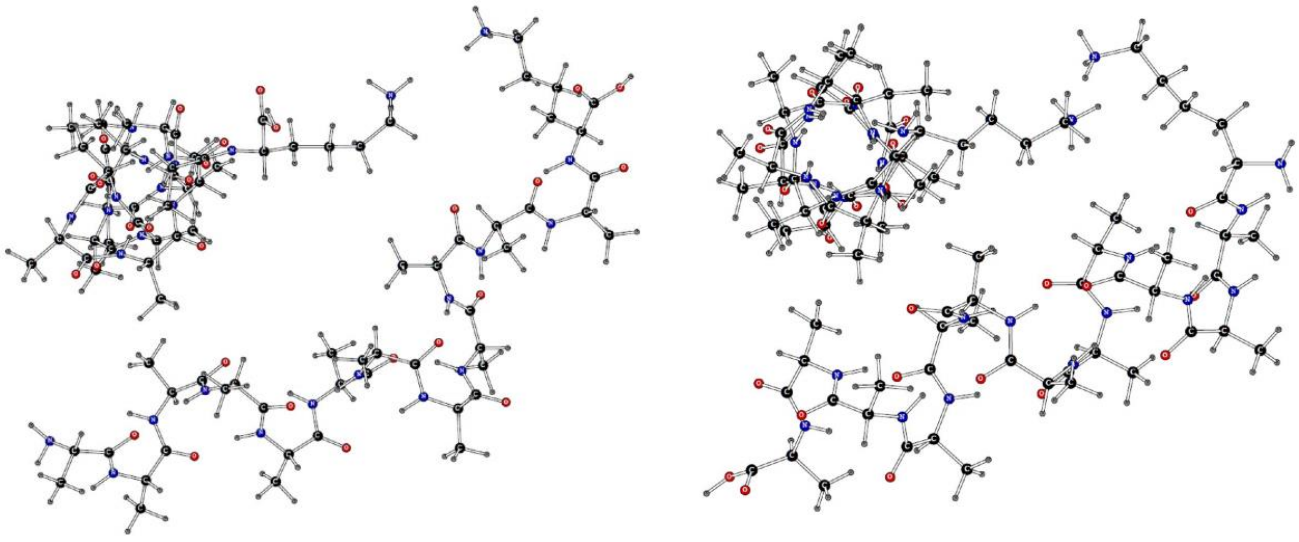


Figure 17 Tridecaalanine peptide with lysine added to the C-terminus is shown at the left, while that with lysine added to the N-terminus is shown at the right. Both appear to maintain the alpha helix, with the C-terminus addition showing some distortion of the secondary structure. (Ponomarev et al., 2012)

however only for one type of side chain. Ponomarev et al. (2012) investigated whether a lysine substitution with one of the alanine side chains in Ala-13 would perturb the alpha helix or maintain it, as seen in Figure 17. Having performed Monte Carle simulations with such a substition on both the N- and C-termini, the group further demonstrated the capacity of POSSSIM to be robust. Lysine substitution was not random. It was important to show that the model could be upheld in strong electrostic environments, as this side chain is usually positively charged and contains polarizable atoms. This is yet another large step in the full-scale protein direction. All that remains is development of parameters for the rest of the side-chain ensemble, and we can begin to construct proteins with any amino acid sequence. (Ponomarev et al., 2012).

Modeling Protein-Ligand Interactions

Lamb and Jorgensen (1997) reviewed some of the ways in which the interaction energies between proteins and ligands can be approximated. One such method, developed by Åqvist et al. (1994), used linear response in order to model binding energy as a function of the Coulombic and Lennard-Jones interactions:

$$\Delta G_{bind} = \beta\left(\Delta E^{Coulombic}\right) + \alpha(\Delta E^{Lennard-Jones})$$

ΔE in this case the difference between the ligand-solvent and ligand-protein interaction energies for the bound and unbound ligand states, as depicted by:

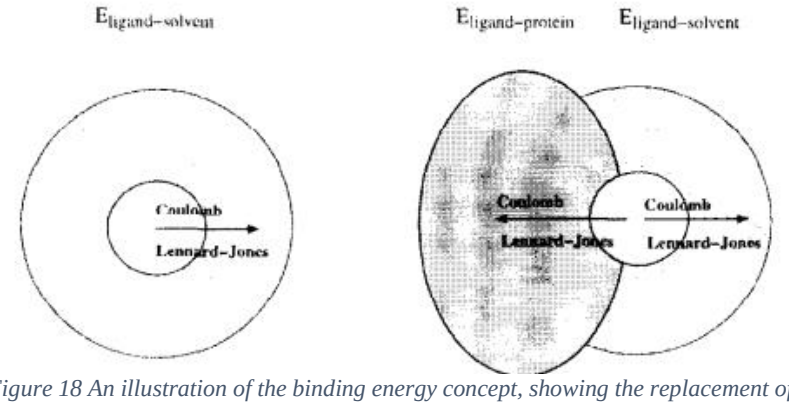


Figure 18 An illustration of the binding energy concept, showing the replacement of proximal solvent with a protein in the bound state. (Lamb & Jorgensen, 1997)

Essentially what this relationship is describing is the favorability of replacing part of the solvent area surrounding the ligand with the protein, that is, does the total energy decrease if this replacement takes place? Generally, the group showed that the β parameter is usually 0.5, such that the total binding energy between the ligand and the protein is usually about a half of the Coulombic energy, and the Lennard-Jones coefficient α is relatively small, about 0.1 in many cases, and serves more as a correction to the electrostatic term. Åqvist et al. performed several MD simulations with HIV protease (Hansson & Åqvist, 1995), trypsin (Åqvist, 1996), and glucose/galactose receptor (Åqvist & Mowbray, 1995) and found that the 0.5 value for β and 0.161 for α upheld in those cases. This technique is employed in the protein-inhibitor interactions investigated here, in which we assume 0.5 for β and then fitted α for optimal binding energy agreement with experimental results.

Burkitt's lymphoma and the Leucine Zipper Formed by Myc and Max Proteins

As it has been shown that POSSIM is reliable for some smaller peptides, the next step after developing parameters for all of the commonly found side chains is to model inhibitor-protein interactions in therapeutic contexts. In this study in particular we examine the binding energies of six inhibitors identified by Yin et al. (2003) for the oncoprotein c-Myc. This protein typically dimerizes with Max to form the leucine zipper, a transcription factor complex that has been shown to be overexpressed in Burkitt's lymphoma and other cancers (Grandori et al., 2005; Felton-Edkins et al., 2003; Dang et al., 2006). It appears that the overexpression of the MYCC gene leads to deregulation of 10-15% of Polymerase II-regulated genes and rRNA and tRNA genes regulated by RNA pol I and III, respectively. The Myc protein together has two domains: an N-terminus domain that contains the trans-activation region, which is not present in the Max protein, and a C-

terminus basic-Helix-Loop-Helix leucine zipper (bHLH-zip) domain that contains the dimerization and DNA-binding regions. It is this c-Myc peptide that we seek to investigate.

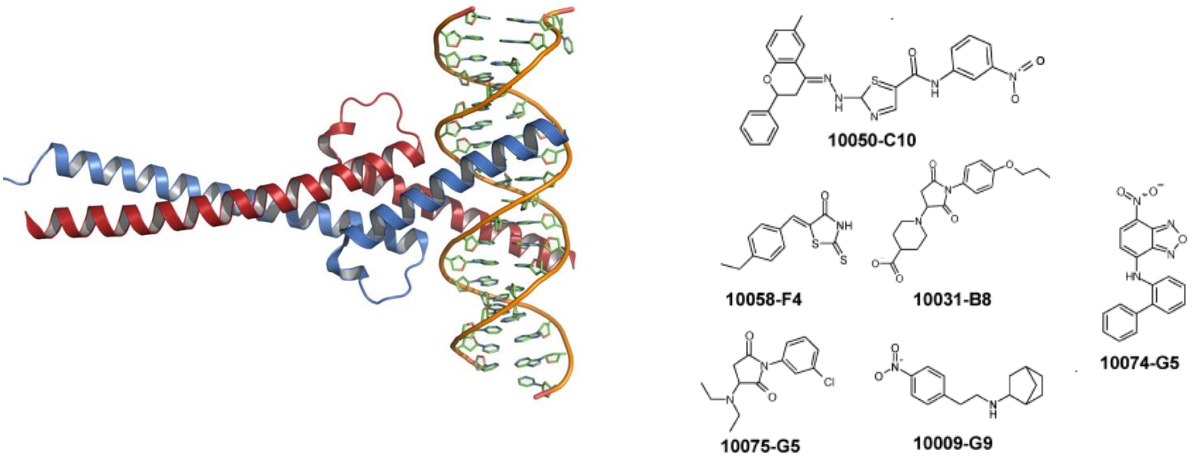


Figure 19 The crystal structure of the leucine zipper is shown at left, with Myc red and Max blue. Inhibitors identified by Yin et al. are shown at right, with their corresponding codes. (Hammoudeh et al., 2009) (Nair et al., 2003)

What is most interesting about the c-Myc protein is that it is an Intrinsically Disordered (ID) protein. Such proteins have a great degree of backbone flexibility in aqueous solution and tend to lack any tertiary structure, as well as having an unstable secondary structure that unravels in physiological context. However, such proteins are still biologically active in this disordered form and can actually make use of their unusual flexibility to adhere to many partners. ID proteins have recently received great attention in the drug discovery field since they appear to violate the protein structure-function dogma and their tendency to undergo a disordered to ordered geometry upon partner-binding. In the approach taken by Hammoudeh et al. (2009) in order to investigate the binding constants for each inhibitor as well as their respective specific binding regions, the inhibitors are used to stabilize a local part of the protein in order to restrict its flexibility. Such rigidity weakens the affinity of this region of the protein for the Max obligate heterodimer, and so simultaneous binding of these inhibitors should prevent dimerization and thus halt cancer cell

development (Hammoudeh et al., 2009). The binding constants found by this group via fluorescence polarization assay were as follows:

compound	$Myc_{353-437}$affinity (μM)	K_{obs} (μM) from 10058-F4 competition	K_{obs} (μM) from 10074-G5 competition
10058-F4	5.3 ± 0.7^a		NB
10074-G5	2.8 ± 0.7^a	NB	
10009-G9		40 ± 10	NB
10075-G5		24 ± 4	NB
10031-B8		16 ± 4	NB
10050-C10		NB	0.9 ± 0.3

Table 11 Summary of the experimentally determined binding constants (K_{obs}) via fluorescence polarization competition assay. (Hammoudeh et al., 2009)

These values were converted to $\Delta G_{binding}$ using the standard equation

$$\Delta G = RTlnK_D$$

in order to yield the following target experimental results for our simulations and force field approximations:

Inhibitor	ΔG (Hammoudeh et al.) (kcal/mol)
10009-G9	-5.99
10031-B8	-6.54
10050-C10	-8.24
10058-F4	-7.2
10074-G5	-7.6
10075-G5	-6.3

Table 12 Experimental binding constants converted to ΔG values, to which calculated values will be compared.

Methods

The methods employed in this study can be essentially broken down into two overarching goals: (1) having had backbone parameters already developed and the rest of the side chain parameter sets being constructed simultaneously by other members of the laboratory, to develop parameters for the amino acids tryptophan and arginine, and (2) to simulate each inhibitor in complex with c-Myc according to the binding sites located by Hammoudeh et al. Development of parameters follows a protocol that has been implemented in other force fields such as OPLS and

has been shown to be robust with respect to transferability. The goal of such a protocol is to test the atom parameters in as many different contexts as possible and see if good accuracy can be upheld in each. As for the source of data for fitting, it is preferable to have solely experimental data to work off of but this is simply not realistic, as many of the features explored here contain micro-scale descriptors that cannot be determined in the laboratory with any reliable accuracy. Thus, experimental data is used whenever available and quantum mechanical calculations are performed in order to fill in the gaps.

Developing parameters for tryptophan and arginine

Since backbone parameters have already been developed, molecules that encapsulate the structure of the side chain itself are investigated and are dubbed *small molecule analogs* of their respective amino acids. Here, pyrrole is used as an analog for tryptophan and the methyl guanidinium ion for arginine. Since there are far fewer atoms in these analogs than in the full size peptides, parameterization will be easier and then a simple transfer of parameters can be done in conjunction with those for the backbone.

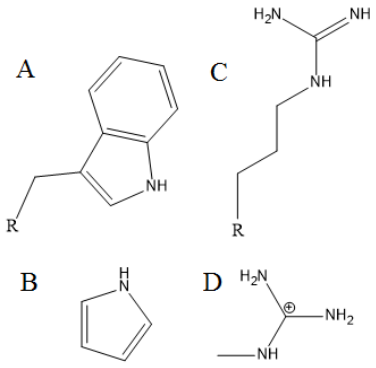


Figure 20 (A) Tryptophan (B) Pyrrole (C) Arginine (D) Methyl guanidinium

First, non-bonded parameters, including the polarizability α, partial charge q, and σ and ε from the Lennard-Jones potential were fitted via 3-body energies. These 3-body interactions involved the molecule and two dipole probes placed at plausible hydrogen bonding sites in close proximity. The configurations for pyrrole and methyl guanidinium are depicted here:

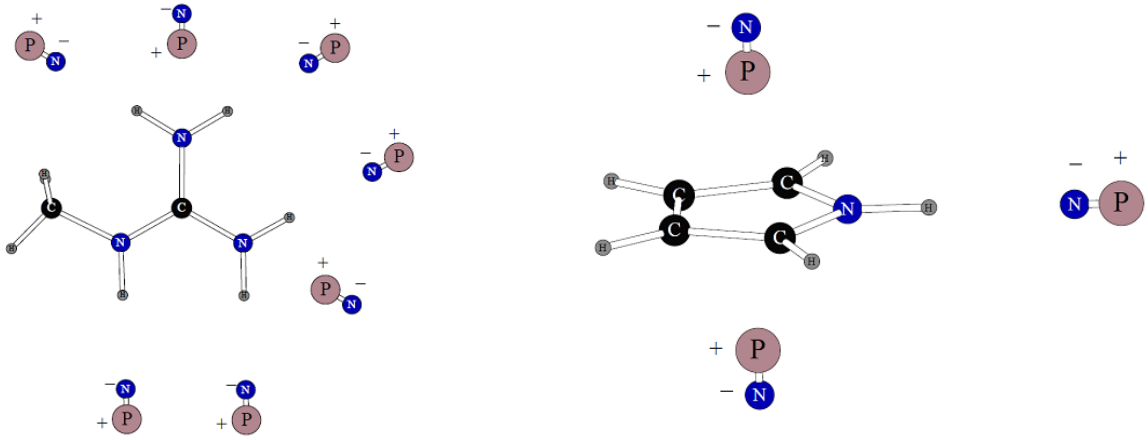


Figure 21 Arrangements of dipole probes meant to perturb the surrounding electrostatic environment. Each probe carries only partial charge, and the nitrogen-phosphorous assignments are arbitrary.

These dipole probes may be appear to contain a nitrogen and phosphorous group, but in reality they are atom types approximated by the software. These probes only carry fixed charges and are oriented in such a way as to mimic favorable electrostatic interaction. For example, positive ends face the electronegative nitrogen atoms and negative ends near the hydrogen atoms. Such a probe will induce an electric field around the molecule and the non-bonded terms can be scaled to appropriately represent the resultant energy. There were 21 three-body configurations with methyl guanidinium and 3 with pyrrole. Three body energies were calculated with the following formula:

$$E_{3-body} = E(1+2+3) - E(1+2) - E(1+3) - E(2+3) + E(1) + E(2) + E(3),$$

where E(i) is the energy of molecule i. Quantum mechanical calculations were performed as a basis for comparison with the Jaguar Suite, using Density Functional Theory (DFT) with the B3LYP method and the cc-pVTZ(-f) basis set (Becke, 1988).

Then, the two small molecules were subjected to dimerization with a single water molecule at a likely hydrogen bonding location, with non-bonded parameters being transferred without change, using the following configurations:

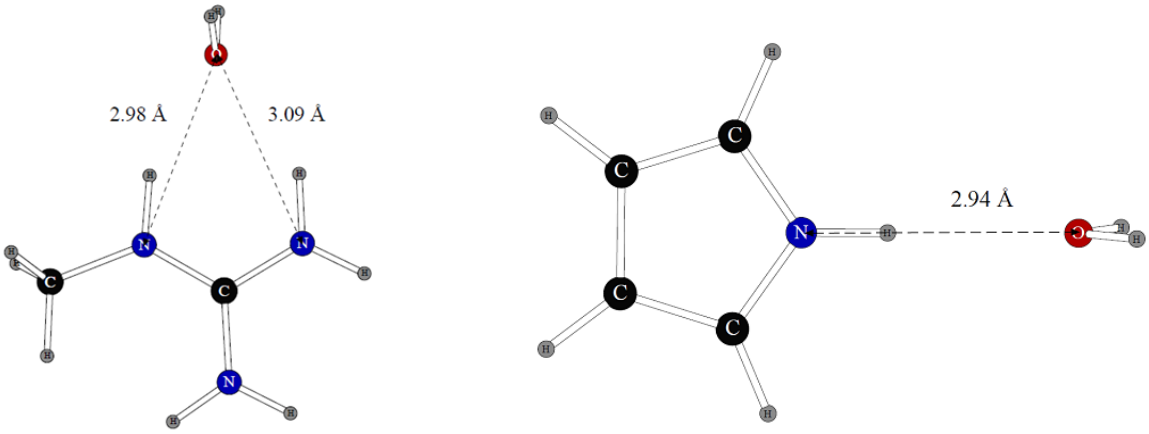


Figure 22 Water dimer structures. The distances shown here are those determined by QM. Note the perpendicularity of the water plane to that of the analyte molecule.

The dimer distances and energies were calculated quantum mechanically using the LMP2/cc-pVTZ(-f) – LMP2/cc-pVQZ method, and LMP2/cc-pVTZ(-f) was used to construct torsional profiles for each molecule. Thus, bonded terms for the Fourier coefficients in the torsional expression were fitted as well. Terms for the bond stretch and angle bending were taken from a previously determined database of atom types. Additionally, a pure liquid MC simulation was performed with 216 pyrrole molecules, using an NPT ensemble and a target temperature and pressure of 25 ºC and 1 atm, respectively. This was performed in both the liquid and gas phases in order to determine the heat of vaporization using the following thermodynamic relationship:

$$\Delta H_{vap} = \Delta U_{vap} + RT = U_{gas} - U_{liquid} + RT$$

The density of the pure liquid was also determined and compared to experimental data presented by Jorgensen and McDonald (1998).

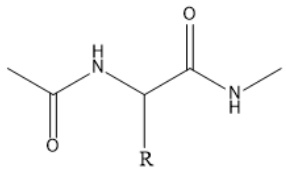


Figure 23 The dipeptide format used here, where R is the side chain under study.

Parameters were then transferred from analog to side chain in the form of a dipeptide, such that the side chain structure can be easily inserted into a protein, as depicted to the left. Adjustment of torsional parameters was done by fitting conformational states to those derived from LMP2/cc-pVTZ(-f)//HF-6-31G** quantum mechanics. A torsional subspace was set up so as to correlate conformation with energy. A single point was mapped to the quantum mechanical energy minima, corresponding to certain φ, ψ, and χ angle values, and then four points were placed in all four directions outward from this first point in increments of 20º. Thus, torsional parameters were fit to these 17 points in order to recreate the potential energy surface. In arginine, all torsional angles were fixed during torsional scanning, while these angles were allowed to change in the case of tryptophan. Initial estimations of these torsional parameters was determined by applying non-Boltzmann distribution weights to each dihedral angle, based on their influence on the molecular energy:

$$W_i = Ae^{-b \cdot G_i},$$

where W_i is the weight for the point i in the subspace, A is a coefficient that is fitted for optimal weight application, and G_i is the magnitude of the torsional surface gradient at point i. Parameters were then optimized for agreement with QM results.

Simulating c-Myc:inhibitor complexes

The c-Myc structure was imported into the Schrödinger Maestro Suite using the 1NKP PDB and isolated. Docking for each individual inhibitor was performed using the docking sequences identified by Hammoudeh et al. The c-Myc protein was truncated according to the extent of each of these binding sites in order to minimize any potential energy reading noise from the movement of the rest of the protein. Side-by-side simulations were done in both the truncated

and full length cases in order to ensure that conformations remained essentially the same and such truncation was legitimate. Soaking of the complex with water was done with the Impact feature and the box size made large enough in each individual case based on the system size, such that the edge of the box was at least 12 Å (the cutoff radius) from any atom within either the protein or ligand. Impact Molecular Dynamics with periodic boundary conditions with the same size as the water box was performed for 5×10^6 steps with a time step of 0.001 ps, resulting in a total real time of 5 ns.

A Monte Carlo simulation performed with the MCPRO program was done only with 10075-G5 to test how these results compare to those with MD. The same truncated protein was used and a water solvent sphere with a radius of 30 Å was applied, as well as an 8.5 Å cutoff radius. The MC was run for 1×10^6 configurations. The force field used in both the MC and the MD simulations was OPLS, in an effort to see how well the fixed-charge model upheld experimental observations.

Results

The three-body energy calculations yielded an average error of 0.118 kcal/mol for methyl guanidinium and 0.184 kcal/mol for pyrrole. Dimerization optimization yielded a qualitatively good result for methyl guanidinium, as the plane of the water molecule was exactly perpendicular to that of the ion, matching the quantum mechanical geometry. Quantitatively, the average error in O··N distances was 0.055 Å and for energy 0.67 kcal/mol. For pyrrole, the distance and dimerization energy error was 0.1 Å and 0.58 kcal/mol, respectively. Below are the torsional results after fitting:

Molecule	Dihedral	Angle Values	Energy, QM	Energy, POSSIM
$C_2N_3H_8^+$	H–N(H_2)–C–N	0º	0	–0.333
		30º	0.175	0.482
		60º	2.615	2.641
	C–N–C–N(H_2)	0º	0	–0.097
		15º	0.14	0.217
		30º	1.115	1.136
	H–C(sp^3)–N–C	0º	1.191	1.176
		60º	0	0.008
		180º	0	0.008
C_4NH_5	H–N–C–C	180°	0	–0.001
		165°	0.424	0.438
		150°	1.345	1.432

Table 13 Characteristic dihedral angles and energy values for methyl guanidinium, compared to calculated results.

The average energy error for methyl guanidinium was 0.099 kcal/mol and 0.009 kcal/mol for pyrrole. For the MC simulations with pyrrole, the calculated ΔH_{vap} was 10.60 kcal/mol and volume 117.4 $Å^3$, vs. 10.80 kcal/mol and 115.3 $Å^3$ experimentally, resulting in 1.9% error for heat of vaporization and 1.8% for molecular volume.

Torsional results for the full peptides of arginine and tryptophan were as follows:

	Energy	
Conformer	QM	POSSIM
1	0	0
2	10.76	10.73
3	3.29	3.28
4	13.87	19.97
5	8.58	8.65
6	4.25	4.19
Error		1.05

Table 14 Arginine conformer energy errors.

	Energy		ϕ		ψ		χ_1		χ_2	
Conf.	QM	P	QM	P	QM	P	QM	P	QM	P
1	0	0.03	–154.5	–158.7	148.4	160.2	–171.8	–171.2	–112.6	–122.8
2	0.15	0.31	–156.0	–159.6	145.8	157.9	–175.5	–170.4	87.6	85.1
3	1.3	2.57	–87.8	–81.7	77.3	38.3	–53.8	–65.8	115.3	131
4	1.65	2.27	–160.1	–164.3	165.2	163.1	52.8	65.9	84.4	79.3
5	2.18	2.36	–89.9	–82.7	76.6	47.8	–62.9	–68.1	–23.7	–11.1
6	2.22	2.52	–152.8	–160.3	164.7	161.4	58.5	68.3	–89.8	–92.9
7	3.26	2.37	–126.9	–82.7	140	48.9	–59.8	–68.2	–89.1	–12.0
8	2.91	2.36	–118.8	–82.7	146.7	47.8	–70.2	–68.1	–7.6	–11.0
9	3.41	2.28	–155.9	–164.2	171.7	162.8	68.5	65.9	–6.1	79.1
Error		0.75		13.5		32.9		6.5		23.9

Table 15 Tryptophan conformer energies as well as dihedral angles for both the backbone and side chain. Average errors are given in the last row.

With regard to the MD simulations, the following graph compares the reference ΔG values to those averaged throughout the entire simulation:

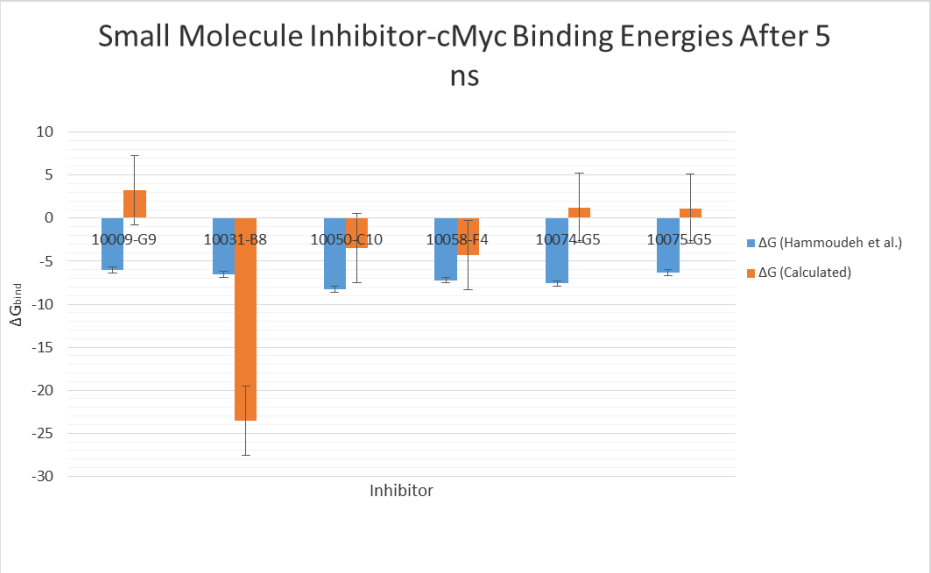


Figure 24 Graphical summary of the calculated $\Delta G_{binding}$ (orange) vs. experimental values (blue).

The average error in predicted binding energy was 8.9 kcal/mol, or 42.0%. In order to observe the stability of each inhibitor's interactive Coulombic energy with c-Myc over the course of the simulation, this energy was recorded and a running average produced for the last nanosecond of the five, where it should be closest to equilibrium:

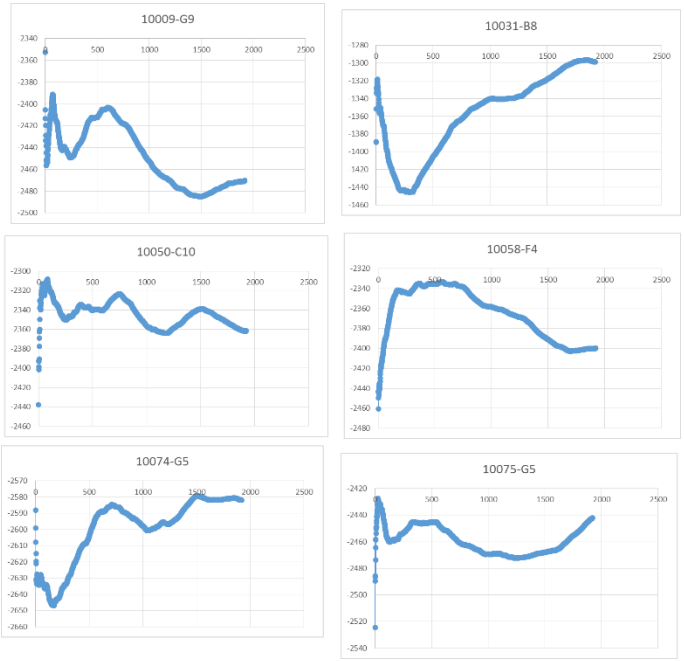


Figure 25 Running average Coulombic inhibitor-protein energy over the course of 1 ns.

Inhibitor	Standard Deviation (kcal/mol)
10009-G9	27.84
10031-B8	45.76
10050-C10	12.8
10058-F4	25.81
10074-G5	19.63
10075-G5	10.45

Table 16 Standard deviations of the data presented in Figure 25.

Since a truncated c-Myc protein was used, two MD simulations were run for a total of 1 ns with the 10050-C10 inhibitor: one with the truncated protein and one with the full protein. Phi and psi angles on the backbone in the overlapping residue sequence were calculated at the end of the simulation. The average unsigned difference in φ was 24.87° and in ψ 33.57°, demonstrating good homology and justifying the truncation method. Below are graphs showing phi and psi angles for each residue between the two proteins as a snapshot after 1 ns simulation:

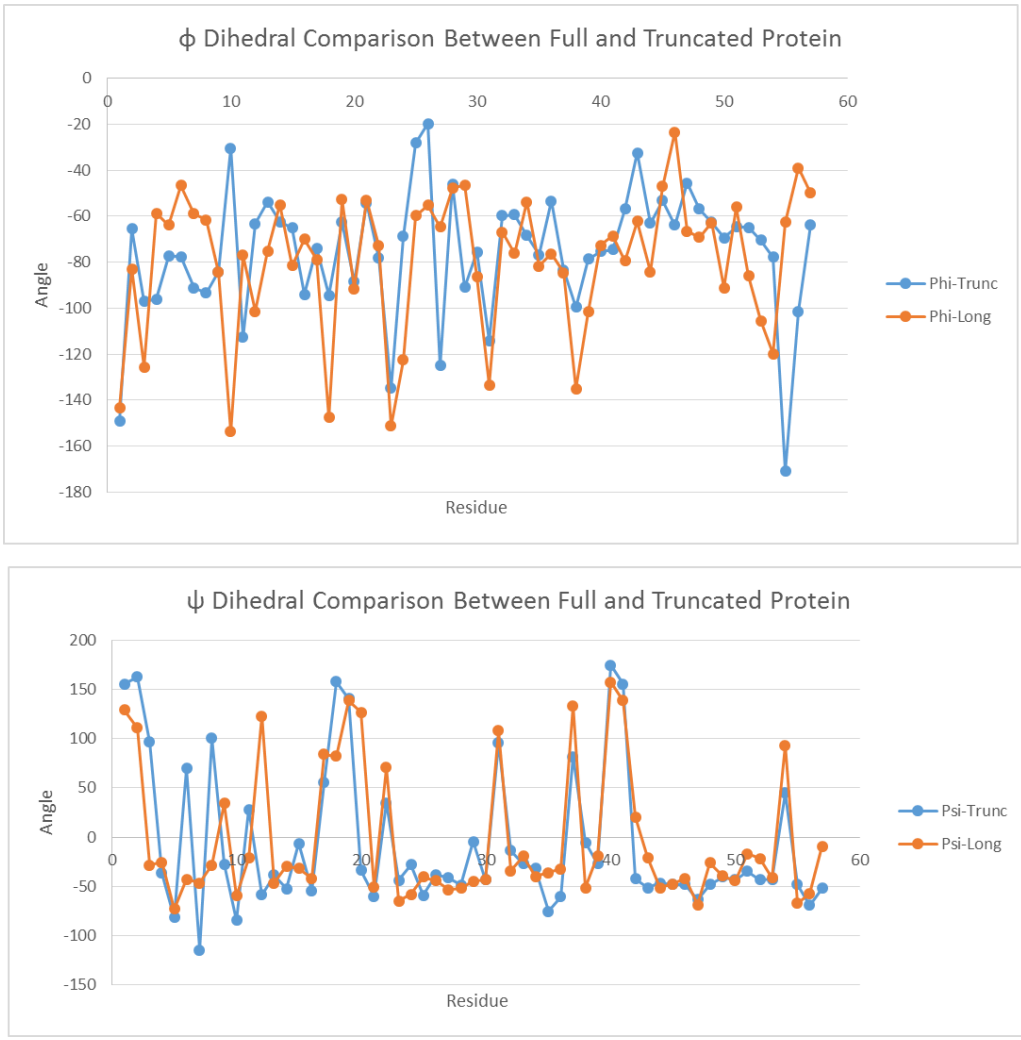


Figure 26 Comparison of instantaneous average backbone φ and ψ angles for both the truncated and full protein for each residue after a 1 ns simulation.

The Monte Carlo simulation run with 100075-G5 resulted in an error of 4.2 kcal/mol, vs. 7.16 kcal/mol from MD. Running average interaction energy was also recorded throughout the extent of the simulation:

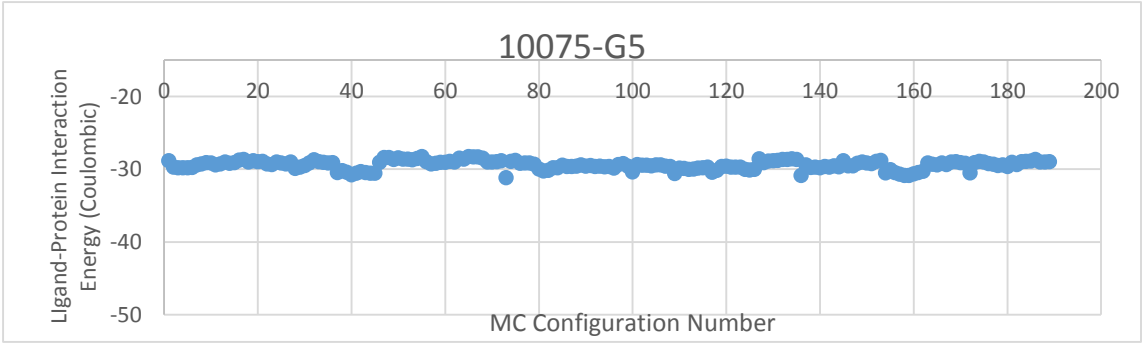


Figure 27 Running average Coulombic 10075-G5-protein energy for 10^6 configurations, with data points recorded every 500 configurations.

The standard deviation observed here was 0.61 kcal/mol.

Discussion

Having produced an average energy error of less than a kcal/mol using the parameters developed for the POSSIM force field in the context of the peptides examined here, it can be said that these deviations are small enough to imply good agreement with both quantum mechanical and experimental observations. Considering the disadvantages seen with non-polarizable models such as OPLS, it seems that explicit treatment of polarization is indeed necessary for accurate representation of biochemical molecules. Perhaps the improvement in representing hydrogen bonding among and between peptides as well as multiple-body interactions has led to this increase in accuracy and reliability. In addition, it has been shown in this work that the small to large molecule transfer of parameters method is a robust way of saving time in such development and paves the way for further transfer to large polypeptides dependably. In a sense, this method represents development of a model within a model, such that a minimal amount of CPU time can be used and the results amplified towards larger applications.

With regards to the simulations performed, it appears that the Molecular Dynamics simulations performed with OPLS did not achieve a satisfactory level of accuracy nor consistency. Considering that the experimental results indicate that the binding energy between each inhibitor and c-Myc is only about -7.0 kcal/mol on average and that the predicted error was 8.9 kcal/mol, such deviation discounts credibility in energetic predictions of future inhibitors. Further, the instability observed within the ligand-protein interaction energy exacerbates this problem. However, better results were observed with the Monte Carlo simulation technique. Here, binding energy error was reduced by 3 kcal/mol and average standard deviation was reduced from 23.7 kcal/mol to only 0.61 kcal/mol, which is a marked improvement. This stability can probably be

attributed to the fact that MC uses statistical sampling to judge if an atom move is favorable or not, and thus large movements, such as those in the backbone of the protein, will often be rejected. Thus, molecules are much more constrained in MC than in MD, and greater stability in energy can be observed. It is our goal to examine the rest of the inhibitors' binding energy using this method and see if the stability and accuracy can be increased. Furthermore, since these simulations were run using the OPLS fixed-charge model, it will be a next step to implement POSSIM and see how well the results turn out, given the accuracy of the polarizable parameters of the peptides developed here. However, a caveat may be that the simulations will take much longer, since the program will have to take polarization into account for each pair-wise interaction.

In any case, it is imperative, now that good parameters have been determined for all constituents of a protein, to start to construct these larger macromolecules. The applications thereof would be tremendous, allowing researchers to theoretically investigate the consequences of potential missense and frameshift mutations as well as the effects of protein-bound therapeutic drugs on a much faster time scale than in the wet laboratory. Further testing of the POSSIM force field will have to be done, but this framework has been shown here to serve as a good model, providing both suitable accuracy and requiring a practical amount of CPU time.

Literature Cited

Banks, J. L., Kaminski, G. A., Zhou, R., Mainz, D. T., Berne, B. J., & Friesner, R. A. (1999). Parametrizing a polarizable force field from ab initio data. I. The fluctuating point charge model. *The Journal of chemical physics*, *110*(2), 741-754.

Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, *38*(6), 3098.

Hammoudeh, D. I., Follis, A. V., Prochownik, E. V., & Metallo, S. J. (2009). Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. *Journal of the American Chemical Society*, *131*(21), 7390-7401.

Hansson, B., & Jans, S. (2012). *Redock: Docking redone*. Informally published manuscript, . Retrieved from http://www.cmbi.ru.nl/redock/verslag.pdf

Jaguar, v7.6, Schrödinger, LLC, New York, NY, 2009.

Kaminski, G. A., Friesner, R. A., & Zhou, R. (2003). A computationally inexpensive modification of the point dipole electrostatic polarization model for molecular simulations. *Journal of computational chemistry*, *24*(3), 267-276.

Kaminski, G. A., Ponomarev, S. Y., & Liu, A. B. (2009). Polarizable Simulations with Second-Order Interaction Model Force Field and Software for Fast Polarizable Calculations: Parameters for Small Model Systems and Free Energy Calculations. *Journal of chemical theory and computation*, *5*(11), 2935-2943.

Kaminski, G. A., Stern, H. A., Berne, B. J., & Friesner, R. A. (2004). Development of an accurate and robust polarizable molecular mechanics force field from ab initio quantum chemistry. *The Journal of Physical Chemistry A,108*(4), 621-627.

Lamb, M. L., & Jorgensen, W. L. (1997). Computational approaches to molecular recognition. *Current opinion in chemical biology*, *1*(4), 449-457.

Nair, S. K., & Burley, S. K. (2003). X-ray structures of Myc-Max and Mad-Max recognizing DNA: molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, *112*(2), 193-205.

Note on electric potential . (1999). Unpublished raw data, Haverford College, . Retrieved from http://www.haverford.edu/physics-astro/course_materials/phys102b/phys102_S99/notes/potnote.html

Ponomarev, S. Y., & Kaminski, G. A. (2011). Polarizable Simulations with Second-Order Interaction Model (POSSIM) Force Field: Developing Parameters for Alanine Peptides and Protein Backbone. *Journal of chemical theory and computation*, *7*(5), 1415-1427.

Ponomarev, S. Y., Click, T. H., & Kaminski, G. A. (2011). Electrostatic polarization is crucial in reproducing Cu (I) interaction energies and hydration.*The Journal of Physical Chemistry B*, *115*(33), 10079-10085.

Ponomarev, S. Y., Sa, Q., & Kaminski, G. A. (2012). Effects of Lysine Substitution on Stability of Polyalanine α Helix. *Journal of chemical theory and computation*, *8*(11), 4691-4706.

Stern, H. A., Kaminski, G. A., Banks, J. L., Zhou, R., Berne, B. J., & Friesner, R. A. (1999). Fluctuating charge, polarizable dipole, and combined models: parameterization from ab initio quantum chemistry. *The Journal of Physical Chemistry B*, *103*(22), 4730-4737.

Yin, X., Giap, C., Lazo, J. S., & Prochownik, E. V. (2003). Low molecular weight inhibitors of Myc–Max interaction and function. *Oncogene*, *22*(40), 6151-6159.