Spring 2016

Global optimization using metadynamics and a polarizable force field: application to protein loops

Armin Avdic
University of Iowa

GLOBAL OPTIMIZATION USING METADYNAMICS AND A POLARIZABLE FORCE FIELD: APPLICATION TO PROTEIN LOOPS

by

Armin Avdic

A thesis submitted in partial fulfillment
of the requirements for the Master of Science
degree in Biomedical Engineering in the
Graduate College of
The University of Iowa

May 2016

Thesis Supervisor: Assistant Professor Michael J. Schnieders

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

MASTER'S THESIS

_____

This is to certify that the Master's thesis of

Armin Avdic

has been approved by the Examining Committee for
the thesis requirement for the Master of Science degree
in Biomedical Engineering at the May 2016 graduation.

Thesis Committee: _____
Michael J. Schnieders, Thesis Supervisor

_____
Michael Mackey

_____
Edwin L. Dove

To those who have inspired, supported and encouraged me.

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor, Dr. Michael J. Schnieders. He has been an amazing advisor, and has provided me with more opportunities than I could have ever imagined. In addition to enabling my dreams for the future, his influence throughout my last two years has instilled valuable scientific and life skills that will advance both my professional and academic careers. In regards to the thesis, I would like to thank Dr. Schnieders for developing the underlying simulation engines required for this work, and Dr. Tim Fenn for his work on the real space component of our hybrid target.

Thanks to Dr. Edwin L. Dove for teaching me about the beauty in the underlying mathematics behind engineering principles. I would also like to thank Dr. Dove, Dr. Schnieders, and Dr. Michael Mackey for providing me opportunities to discover a passion for teaching.

Furthermore, I could not have found better lab mates if I tried; Stephen LuCore and Jacob Litman have been instrumental in helping me gain background knowledge in the field and the lab's software package. Jill Hauer, Ian Nessler, Jarod Benowitz, and Hernan Bernabe were always willing to provide feedback and friendship, and I would like to give a special thank you to Mallory Tollefson for directly contributing to this thesis work by porting the kinetic closure code.

Finally, I would like to thank my parents, sister and Ellen Black for their support and always encouraging me to pursue academics.

ABSTRACT

Genetic sequences are being collected at an ever increasing rate due to rapid cost reductions; however, experimental approaches to determine the structure and function of the protein(s) each gene codes are not keeping pace. Therefore, computational methods to augment experimental structures with comparative (i.e. homology) models using physics-based methods for building residues, loops and domains are needed to thread new sequences onto homologous structures. In addition, even experimental structure determination relies on analogous first principles structure refinement and prediction algorithms to place structural elements that are not defined by the data alone.

Computational methods developed to find the global free energy minimum of an amino acid sequence (i.e. the protein folding problem) are increasingly successful, but limitations in accuracy and efficiency remain. Optimization efforts have focused on subsets of systems and environments by utilizing potential energy functions ranging from fixed charged force fields (Fiser, Do, & Sali, 2000; Jacobson et al., 2004), statistical or knowledge based potentials (Das & Baker, 2008) and/or potentials incorporating experimental data (Brunger, 2007; Trabuco, Villa, Mitra, Frank, & Schulten, 2008).

Although these methods are widely used, limitations include 1) a target function global minimum that does not correspond to the actual free energy minimum and/or 2) search protocols that are inefficient or not deterministic due to rough energy landscapes characterized by large energy barriers between multiple minima.

Our Global Optimization Using Metadynamics and a Polarizable Force Field (GONDOLA) approach tackles the first limitation by incorporating experimental data (i.e. from X-ray crystallography, CryoEM or NMR experiments) into a hybrid target

function that also includes information from a polarizable molecular mechanics force field (Lopes, Roux, & MacKerell, 2009; Ponder & Case, 2003). The second limitation is overcome by driving the sampling of conformational space by adding a time-dependent bias to the objective function, which pushes the search toward unexplored regions (Alessandro Barducci, Bonomi, & Parrinello, 2011; Zheng, Chen, & Yang, 2008).

The GONDOLA approach incorporates additional efficiency constructs for search space exploration that include Monte Carlo moves and fine grained minimization. Furthermore, the dimensionality of the search is reduced by fixing atomic coordinates of known structural regions while atoms of interest explore new coordinate positions. The overall approach can be used for optimization of side-chains (i.e. set side-chain atoms active while constraining backbone atoms), residues (i.e. side-chain atoms and backbone atoms active), ligand binding pose (i.e. set atoms along binding interface active), protein loops (i.e. set atoms connecting two terminating residues active) or even entire protein domains or complexes. Here we focus on using the GONDOLA general free energy driven optimization strategy to elucidate the structural details of missing protein loops, which are often missing from experimental structures due to conformational heterogeneity and/or limitations in the resolution of the data.

We first show that the correlation between experimental data and AMOEBA (i.e. a polarizable force field) structural minima is stronger than that for OPLS-AA (i.e. a fixed charge force field). This suggests that the higher order multipoles and polarization of the AMOEBA force field more accurately represented the true crystalline environment than the simpler OPLS-AA model. Thus, scoring and optimization of loops with AMOEBA is more accurate than with OPLS-AA, albeit at a slightly increased computational cost.

Next, missing PDZ domain protein loops and protein loops from a loop decoy data set were optimized for 5 ns using the GONDOLA approach (i.e. under the AMOEBA polarizable force field) as well as a commonly used global optimization procedure (i.e. simulated annealing under the OPLS-AA fixed charge force field). The GONDOLA procedure was shown to provide more accurate structures in terms of both experimental metrics (i.e. lower R_{free} values) and structural metrics (i.e. using the MolProbity structure validation tool). In terms of R_{free}, only one out of seven simulated annealing results was better than the Gondola global optimization. Similarly, one simulated anneal loop had a better MolProbity score, but none of the simulated annealing loops were better in both categories. On average, GONDOLA achieved an R_{free} value 19.48 and simulated annealing saw an average R_{free} value of 19.63, and the average MolProbity scores were 1.56 for GONDOLA and 1.75 for simulated annealing.

In addition to providing more accurate predictions, GONDOLA was shown to converge much faster than the simulated annealing protocol. Ten separate 5 ns optimizations of the 4 residue loop missing from one of the PDZ domains were conducted. Five were done using GONDOLA and five with the simulated annealing protocol. The fastest four converging results belonged to the GONDOLA approach. Thus, this work demonstrates that GONDOLA is well-suited to refine or predict the coordinates of missing residues and loops because it is both more accurate and converges more rapidly.

PUBLIC ABSTRACT

The human genome project sparked a revolution in the availability of low-cost genetic information and dramatically improved our understanding of human health and disease. Many approaches are being explored to assist clinical decision making in light of low-cost genetic information. For missense variants that have not been characterized biochemically, computational approaches capable of predicting the impact on protein structure, function and human phenotype are sorely needed. The starting point for such approaches are accurate protein structures for both wildtype and variant sequences. However, X-ray crystallography, a widely used method for the experimental determination of protein structure, is too time-consuming to be applied to all missense variants of clinical interest. Therefore, computational methods to augment experimental structures with comparative (i.e. homology) models based on physics-based methods for building missing residues, loops and domains of protein structures are imperative.

Here we propose an algorithm called GONDOLA to predict the atomic coordinates of protein residues, loops and domains. GONDOLA uses a state-of-the-art polarizable force field called AMOEBA to describe the interactions between atoms, and molecular dynamics with a time-dependent bias to drive efficient global optimization of the structure. The approach improves the quality of both experimental protein structures and those generated from homology modeling relative to existing methods. We demonstrate the power of GONDOLA by showing that it converges more rapidly than global optimization by simulated annealing, while also providing more accurate protein loops based on both experimental and physical criteria.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Pioneers in biochemistry, like Christian Anfinsen, proposed that a primary amino acid sequence contains all information needed to define a protein fold, which is equivalent to asserting that the folded or native structure of a sequence is at the free energy minimum (Anfinsen, 1973). Currently, fast and reliable methods for deterministically finding the global free energy minimum of an amino acid sequence have only been partially realized. Such optimization efforts have focused on subsets of systems and environments by optimizing potential energy functions ranging from fixed charged force fields (Fiser et al., 2000; Jacobson et al., 2004), statistical or knowledge based potentials (Das & Baker, 2008) and/or potentials that incorporate experimental data (Brunger, 2007; Trabuco et al., 2008).

While tremendously insightful, these methods often have one or more limitations including 1) target function minimum that do not correspond to the free energy minimum and 2) search protocols that are inefficient or not deterministic due to rough energy landscapes characterized by large energy barriers between multiple minima. To address the first limitation, it is possible to target experimental data, such as from X-ray crystallography, CryoEM or NMR experiments, and also to include information from next generation polarizable molecular mechanics force fields (Lopes et al., 2009; Ponder & Case, 2003). To address the second limitation, sampling of conformational space can be driven by addition of a time-dependent bias to the objective function, which pushes the search toward unexplored regions (Alessandro Barducci et al., 2011; Zheng et al., 2008).

Here we explore the synthesis of novel objectives functions and global optimization schemes to efficiently determine the coordinates of missing protein loops. We focus on a general free energy driven optimization strategy based on a target function that is a weighted combination of a polarizable force field and electron density maps (i.e. that can be defined by either X-ray crystallography or Cryo-EM experiments). Using this target, global optimization proceeds via molecular dynamics combined with addition of a time-dependent bias (generally known as metadynamics) to drive the sampling towards new regions of conformational space. To reduce the search space, atomic coordinates of known structural regions are fixed, while atoms of interest explore new coordinate positions. The overall approach can be used for optimization of side-chains (i.e. set side-chain atoms active while constraining backbone atoms), residues (i.e. side-chain atoms and backbone atoms active), ligand binding (i.e. set atoms along binding interface active), homology models (i.e. set atoms connecting two domains active) or even entire protein optimization.

In this work, we will demonstrate the approach by elucidating the structural details of mobile segments of proteins (i.e. loops), which are often difficult to resolve from experimental data. These protein loops, flexible regions of a polypeptide chain that anchor two contiguous secondary structures, play decisive roles in important biological functions such as protein kinase activation, molecular recognition (Ciarapica, Rosati, Cesareni, & Nasi, 2003) and for the catalytic and ligand binding sites of enzymes (Bernstein et al., 2004; Mol et

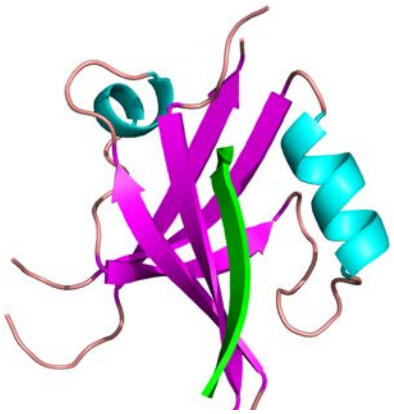

Figure 1: Protein structure of PDZ domain in complex with Syndecan1 peptide. (PDB ID: 4GVD)

al., 2003; Slesinger, Jan, & Jan, 1993; Steichen et al., 2012; Stuart et al., 1986). Accurate loop predictions are essential in structural refinement (R. Bruccoleri, 2000; Dmitriev & Fillingame, 2007) and are often needed to understand fundamental biophysics of dynamic processes in a variety of structural biology applications (Espadaler, Querol, Aviles, & Oliva, 2006; Martin, Cheetham, & Rees, 1989; Tasneem, Iyer, Jakobsson, & Aravind, 2005; Yarov-Yarovoy, Baker, & Catterall, 2006).

The most common method of determining protein structures, X-ray crystallography, often fails to define loop atom coordinates with a high degree of certainty due to their high flexibility and mobility. Thus, loops are often omitted from PDB submissions, but must be rebuilt to permit downstream analysis including molecular simulation studies of function or for molecular design applications. As an example, PDZ domains are often missing short loops. PDZ domains are common tertiary structures found in over 250 different cell types in the human body, and are found in cell signaling complexes of the cell membrane to act as protein-protein recognition sites binding specifically to C-termini. More specifically, the flexible region responsible for this interaction is known as the carboxylate-binding loop (Penkert, DiVittorio, & Prehoda, 2004). Specificity in ligand binding is partially contributed by this loop (Doyle et al., 1996); however, it is often omitted from structures deposited on the Protein Data Bank (Berman et al., 2000).

This thesis explores computational determination of protein loops using the combination of 1) a novel target function defined by a weighted sum of experimental data (i.e. an electron density map) and a polarizable atomic multipole force field (Fenn & Schnieders, 2011; M. J. Schnieders, Fenn, & Pande, 2011) and 2) a global optimization protocol characterized by addition of a time dependent bias to force exploration of new regions of

phase space (Park et al., 2014; Michael J. Schnieders et al., 2012). The advantages of the target function are quantified using both structural (i.e. backbone and side-chain conformation) and experimental metrics (i.e. R/R_{free}). The efficiency of the search protocol is compared to established global optimization schemes such as simulated annealing. Finally, all algorithms are publicly disseminated as part of the open source Force Field X (FFX) molecular biophysics software available from the University of Iowa (http://ffx.biochem.uiowa.edu).

CHAPTER 2: BACKGROUND

2.1: Force Fields

Classical modeling of molecular systems began with seminal hand calculations by Frank Westheimer in the 1940's (Westheimer & Mayer, 1946). Those molecular mechanics equations lead to the development of computer-aided simulations that describe atomic resolution interactions and molecular forces. As interest rose in studying larger systems, the quantum mechanical level of detail, i.e. electronic motion, was unsuitable due to prohibitive computational costs, so classical mechanics and hybrid systems were used to study nuclear motion (Warshel & Levitt, 1976). Thus, the use of classical mechanics in the field of computational biophysics began as a quest to understand and predict experimental results of various molecular properties; these approximations were necessary due to the computational expense of quantum mechanical calculations for the simulation of biological systems.

2.1.1: Fixed Charge Force Fields

By representing organic chemistry as mechanical systems of atoms that interact based on harmonic bonded terms and non-bonded through-space interactions, the potential energy as a function of coordinates quickly became useful in understanding emergent physical properties. More specifically, classical bonded terms (i.e. bond stretching and bond angle bending) and non-bonded terms (i.e. van der Waals and Coulomb interactions between atoms) are described by Hooke's Law (1676), Coulomb's Law (1785), and Mie (1903) or Lennard-Jones (1924) potentials (Jorgensen, 2013). The functional form of the mechanical system and the parameters needed to calculate the various molecular properties, are known as a force field. Furthermore, the development of these classical

mechanics models focused on electrostatic models that were limited to fixed atomic charges from around 1960 until the turn of the 21st century (Ponder & Case, 2003). Although fixed atomic partial charges can implicitly include electronic polarization of the electronic cloud in proportion to a defined environmental field, this approach lacks energetic transferability between high and low dielectric environments (Ponder & Case, 2003). While fixed charge force fields are efficient, their accuracy is often limited for cases that include transfer between vacuum and condensed phase states or between the surface of a protein and its hydrophobic core. Thus, most molecular dynamics calculations currently lack explicit inclusion of polarization or higher order fixed atomic multipoles.

2.1.2: Polarizable Force Fields

Fixed charged force fields are continually being improved based on modest modifications to their functional form and by more complete optimization of their parameters. For example, torsional potential energy terms that dictate secondary structure populations have recently been revisited (Hornak et al., 2006; MacKerell, Feig, & Brooks, 2004); however, the implicit treatment of polarization inherently limits the energetic transferability between environments in a manner that cannot be resolved without introduction of explicit atomic polarizablility into the electrostatics model. Consequently, a group of more sophisticated force fields has emerged (GEMM (Elking, Cisneros, Piquemal, Darden, & Pedersen, 2010) SIBFA (Gresh, Cisneros, Darden, & Piquemal, 2007), Charmm Drude Model (Lopes et al., 2013), Charmm Fluc-Q model (Patel & Brooks, 2006). These force fields include explicit polarization, which allows charges to redistribute in response to the total electric field of the environment and approximates the

quantum mechanical response of organic molecules (Böttcher, 1993). For example, negatively charged electron density moves towards the positive potential of cationic charges and away from the negative potential of anionic charges. Force fields can implement a polarization response via charge-on-springs (i.e. Drude) (Anisimov et al., 2005; Yu et al., 2010), fluctuating charges (Patel & Brooks, 2006), or induced point dipoles (Ren & Ponder, 2002). An example of the latter is the Atomic Multipole Optimized Energetics for Biomolecular Applications (AMOEBA), which treats polarization using induced dipoles and fixed electron distribution using permanent multipoles (Ren, Wu, & Ponder, 2011; Shi et al., 2013). In principle, explicit inclusion of polarization improves transferability between environments and offers advantages over electrostatic models used in previous loop building and refinement algorithms.

2.1.3: AMOEBA Force Field Functional Form

The AMOEBA protein force field is described by six bonded and three non-bonded terms (Shi et al., 2013)

$$U_{AMOEBA} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{b}\theta} + U_{\text{oop}} + U_{\text{torsion}} + U_{\text{tor-tor(GLY)}} + U_{\text{vdW}} + U_{\text{ele}}^{\text{perm}} + U_{\text{ele}}^{\text{ind}} \quad \textbf{Eqn. 1}$$

The energy contributions of bond and angle terms capture higher-order deviations from ideal bond lengths (i.e. b_0) and angles (i.e. θ_0) to account for anharmonicity. The bond stretching between two atoms is described by

$$U_{\text{bond}} = \text{K}_\text{b}(b - b_0)^2[1 - 2.55\,(b - b_0) + 3.793125\,(b - b_0)^2] \quad \textbf{Eqn. 2}$$

and the energy of a bond angle by

$$U_{\text{angle}} = \text{K}_\theta(\theta - \theta_0)^2[1 - 0.014\,(\theta - \theta_0) + 5.6\times10^{-5}\,(\theta - \theta_0)^2 - 7.0\times10^{-7}\,(\theta - \theta_0)^3 + 2.2\times10^{-8}\,(\theta - \theta_0)^4] \quad \textbf{Eqn. 3}$$

The bonded term for the coupling of bond stretching with bond angle bending (i.e. a bond-angle cross term) is given by

$$U_{\mathrm{b}\theta} = \mathrm{K}_{\mathrm{b}\theta}[(b - b_0) + (b' - b'_{\theta})](\theta - \theta_0) \qquad \textbf{Eqn. 4}$$

To restrain sp^2 hybridized trigonal centers to out-of-plane bending is described as a scaled value of the angle (i.e. χ) between *j1* vector and *ijk* plane for sequentially bonded atom centers (*i*,*j*,*k*,*1*) as given by

$$U_{\mathrm{oop}} = \mathrm{K}_{\chi}\chi^2 \qquad \textbf{Eqn. 5}$$

Torsional energies describe rotational barriers about the central bond of four linearly bonded atoms that define a dihedral angle ϕ using a Fourier expansion with *n* terms, where the n^{th} term is parameterized by its magnitude $\mathrm{K_n}$ and phase δ_n

$$U_{\mathrm{torsion}} = \sum_n \mathrm{K_n}[1 + \cos(n\phi \pm \delta_n)] \qquad \textbf{Eqn. 6}$$

A torsion-torsion energy coupling term was implemented in previous version of AMOEBA via a grid-based correction to ensure correct conformational energies for φ-ψ based on quantum mechanics target data. However, recent efforts have focused on achieving similar quality from traditional 3-term Fourier expansion torsional functions. An exception is the backbone φ-ψ coupling for glycine residues, which continue to be corrected by a two-dimensional bicubic spline (Shi et al., 2013).

Pairwise additive van der Waals (vdW) interactions are described using a buffered 14-7 potential, which has a general functional form described by

$$U_{\mathrm{vdW}} = \varepsilon_{ij}\left(\frac{1+\delta}{\rho_{ij}+\delta}\right)^{n-m}\left(\frac{1+\gamma}{\rho_{ij}^{m}+\gamma} - 2\right) \qquad \textbf{Eqn. 7}$$

The potential well depth is given by ε_{ij} and ρ_{ij} represents R_{ij}/R^0_{ij}, where R^0_{ij} is the minimum energy distance and R_{ij} is the separation between atoms *i* and *j*. Furthermore,

the fixed values of n = 14, m = 7, δ = 0.07, and γ = 0.12 were chosen (Halgren, 1992). Thus, van der Waals interactions are given by

$$U_{\mathrm{vdW}} = \varepsilon_{ij}\left(\frac{1.07}{\rho_{ij}+0.07}\right)^{7}\left(\frac{1.12}{\rho_{ij}^{7}+0.12}-2\right) \qquad \textbf{Eqn. 8}$$

where the combining rules for heterogeneous atom pairs are $R_{ij}^{0} = \frac{(R_{ii}^{0})^{3}+(R_{jj}^{0})^{3}}{(R_{ii}^{0})^{2}+(R_{jj}^{0})^{2}}$ for minimum energy distance and $\varepsilon_{ij} = \frac{4\varepsilon_{ii}\varepsilon_{jj}}{\left(\varepsilon_{ii}^{1/2}+\varepsilon_{jj}^{1/2}\right)^{2}}$ for well depth.

Furthermore, AMOEBA charges and van der Waals parameters are designed to reproduce properties of molecules in both vapor and condensed phase environments. This transferability depends on an electrostatic model defined by ideal induced dipoles and permanent multipoles (through quadrupole order) placed at each atomic center. The permanent electrostatic energy between atoms *i* and *j* separated by a distance r_{ij} is given by $U_{\mathrm{ele}}^{\mathrm{perm}}(r_{ij}) = M_i^T T_{ij} M_j$ where

$$T_{ij} = \begin{vmatrix} 1 & \frac{\partial}{\partial_{x_j}} & \frac{\partial}{\partial_{y_j}} & \frac{\partial}{\partial_{z_j}} & \cdots \\ \frac{\partial}{\partial_{x_i}} & \frac{\partial^2}{\partial_{x_i}\partial_{x_j}} & \frac{\partial^2}{\partial_{x_i}\partial_{y_j}} & \frac{\partial^2}{\partial_{x_i}\partial_{z_j}} & \cdots \\ \frac{\partial}{\partial_{y_i}} & \frac{\partial^2}{\partial_{y_i}\partial_{x_j}} & \frac{\partial^2}{\partial_{y_i}\partial_{y_j}} & \frac{\partial^2}{\partial_{y_i}\partial_{z_j}} & \cdots \\ \frac{\partial}{\partial_{z_i}} & \frac{\partial^2}{\partial_{z_i}\partial_{x_j}} & \frac{\partial^2}{\partial_{z_i}\partial_{y_j}} & \frac{\partial^2}{\partial_{z_i}\partial_{z_j}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{vmatrix} \frac{1}{r_{ij}} \qquad \textbf{Eqn. 9}$$

and the permanent multipole for atom *i* is given by a vector containing a partial charge, an ideal point dipole and an ideal traceless quadrupole

$$M_i = [q_i, \mu_{i,x}, \mu_{i,y}, \mu_{i,z}, Q_{i,xx}, Q_{i,xy}, Q_{i,xz}, \ldots Q_{i,zz}]^T \qquad \textbf{Eqn. 10}$$

Polarization is described by an induced dipole at each atomic center, which for atom *i* is given by

$$\mu_{i,\alpha}^{\text{ind}} = \alpha_i E_{i,\alpha} \qquad \textbf{Eqn. 11}$$

where $E_{i,\alpha}$ is the total electric field along the α-axis (i.e. where α is *x*, *y* or *z*) and α_i is the polarizability of atom *i*. The total electric field is generated by permanent multipoles and induced dipoles at all other atomic centers (neglecting masking rules) as summarized by

$$\mu_{i,\alpha}^{\text{ind}} = \alpha_i \left(\sum_{\{j\}} T_{\alpha}^{ij} M_j + \sum_{\{j'\}} T_{\alpha\beta}^{ij'} \mu_{j',\beta}^{\text{ind}} \right) \qquad \textbf{Eqn. 12}$$

The polarizable AMOEBA model has been defined for water (Ren & Ponder, 2003) and ions (Grossfield, Ren, & Ponder, 2003), small molecules (Ren et al., 2011) and proteins (Shi et al., 2013) and support for continuum electrostatics has been established (M. J. Schnieders, Baker, Ren, & Ponder, 2007; M. J. Schnieders & Ponder, 2007). Application of the AMOEBA model in the context of biomolecular X-ray crystallography refinement at both high and low resolution has demonstrated improvements to both MolProbity assessment (Chen et al., 2010) and agreement with experimental scattering data (Fenn & Schnieders, 2011; Fenn, Schnieders, Brunger, & Pande, 2010; M. J. Schnieders et al., 2011). For example, global optimization of amino acid side-chain conformations for PCNA structures was recently explored (LuCore et al., 2015). Additionally, a general automatic parameterization procedure using the *Poltype* tool has been developed. *Poltype* is used to parameterize arbitrary organic molecules for the AMOEBA force field (J. C. Wu, Chattree, & Ren, 2012), which broadens the scope of applicability of AMOEBA refinement approaches to most data sets found in the Protein Databank (Berman et al., 2000).

2.2: Local and Global Optimization

Both local and global optimizations are critical for protein structure optimization in the context of interpreting experiments and for ab initio predictions. The algorithms used to determine optimal loop conformations often depend on both. Here we introduce some of the most commonly used techniques for optimization, with a particular focus on schemes that have been applied to protein loop prediction.

2.2.1: Local Optimization: Steepest Decent, Newton and Quasi-Newton Algorithms

Relaxing biomolecules to local minima has been accomplished through methods with varying trade-offs. By exploiting derivatives of the energy function, line search algorithms are able to discover lower energy coordinates for a system. The simplest use of the derivative landscape is a greedy choice algorithm coined steepest descent; step directions are always chosen to coincide with the negative gradient direction. Alternatively, Newton's method makes use of the second derivative matrix (i.e. the Hessian) to locate the potential energy minimum using fewer steps, albeit at a greater cost per step. The Hessian information helps to reduce oscillations about the local minimum, but calculation and storage of n^2 elements is computationally intensive and memory expensive. Fortunately, the inverse Hessian can be estimated using a series of gradient evaluations. This leads to quasi-Newton algorithms, such as the limited memory-BFGS (Byrd, Lu, Nocedal, & Zhu, 1995) scheme, which converge in fewer iterations than steepest descent, but with a similar computational cost per step (Ponder & Richards, 1987).

2.2.2: Global Optimization Using Molecular Dynamics and Simulated Annealing (SA)

Protein loop optimization in 3-dimensional Euclidean space is simplified to some degree due to constraints defined by the loop end-points (i.e. peptide carboxyl or amino termini). As mentioned previously, the local optimization methods use slope information (i.e. the gradient) to step in the direction of a local minimum. On the other hand, global optimizers must avoid favorable local minimum and overcome barriers that would otherwise halt further progress. In other words, the potential energy functions for proteins describe a rough landscape with many barriers and local minima. To escape local minima and overcome barriers, the simulated annealing global optimization techniques use temperature and kinetic energy in a simulation method known as molecular dynamics (MD) (Alder & Wainwright, 1959) where MD simulations act on forces via discrete integration of mass and acceleration of the atoms in a molecular system. Higher temperature increases the velocity distribution of the atoms and the total kinetic energy of the system. The larger the kinetic energy, the greater the ability of the system is to escape local minima (i.e. higher barriers can be crossed). Furthermore, the procedure finished by following a slow cooling protocol, which allows system (i.e. the loop) to relax into the global minimum potential energy. The success of the algorithm is directly tied to the cooling protocol and duration (Brunger, Krukowski, & Erickson, 1990).

2.2.3: Potential Smoothing

Overarching strategy of potential smoothing is to flatten the original potential energy surface and reduce the number of local minima (Piela, Kostrowicki, & Scheraga, 1989). The use of potential smoothing for global optimization repeatedly transforms the potential energy surface until only one minimum remains. The minima are described by

the depth of their potential well, which is defined as the difference between the minimum value at the boundary and the value at the bottom of the well (i.e. the local minimum). The underlying assumption is that shallower potential wells will blur into the surrounding surface more easily than a deep well, so it is expected that the deformation of the potential energy surface will disappear shallow wells by absorbing them into growing deeper potential wells. The method avoids issues associated with the generation of Boltzmann distributions for each temperature gradient, which can be problematic for the aforementioned temperature dependent global optimization methods (Hart, Pappu, & Ponder, 2000). However, there is a drawback because the modified potential energy surface has potential wells with shallower depths, modified positions, altered gradients, and a different overall size. In response to these issues, once a single minimum is reached the algorithm is reversed and the minimum is traced back to its origins.

2.2.4: Metadynamics and Orthogonal Space Random Walk (OSRW)

Metadynamics is a powerful enhancement to MD sampling based on addition of a time dependent bias to the total potential energy, which drives the system to escape local free energy minima and more rapidly explore phase space (Kong & Brooks, 1996; Laio & Parrinello, 2002). The time-dependent bias is usually based on a summation of Gaussian functions whose locations are a function of a state variable (i.e. λ) (A. Barducci, Bussi, & Parrinello, 2008). Thus, metadynamics is an attractive alternative to potential smoothing because the time-dependent bias flattens the potential energy landscape without the smoothing transformation, which is cumbersome to derive and implement for advanced force fields. Furthermore, *a priori* definition of a cooling schedule and duration required for simulated annealing are avoided. For these reasons, the current work explores a

metadynamics optimization strategy, rather than potential smoothing or simulated annealing, to achieve efficient configurational search during optimization of loop coordinates.

As a metadynamics simulation converges, the λ-dependent biasing potential $f_m(t,\lambda)$ approaches the negative value of the λ-dependent AMOEBA free energy$-G_{AMOEBA}(\lambda)$. The time-dependent total potential energy equation is given by

$$U_m = U_{AMOEBA}(\lambda) + f_m(t,\lambda) \qquad \textbf{Eqn. 13}$$

where λ is a thermodynamic path variable. In the case of loop optimization, λ = 1 corresponds to the loop interacting with its condensed phase environment, while λ = 0 corresponds to the loop being uncoupled from the environment where it experiences a vapor state.

While metadynamics increases search efficiency along the reaction coordinate defined by λ, hidden barriers remain (Zheng et al., 2008). The Orthogonal Space Random Walk (OSRW) method expands the Gaussian-shaped repulsive potential to include bias along the derivative of the potential energy with respect to λ ($F_\lambda = \partial U/\partial\lambda$) to give a total potential energy defined by

$$U_m = U_{AMOEBA}(\lambda) + f_m(\lambda) + g_m(\lambda, F_\lambda) \qquad \textbf{Eqn. 14}$$

where $g_m(\lambda, F_\lambda)$ is the sum two-dimensional Gaussian-shaped repulsive potentials (i.e. hills) centered at states given by $[\lambda(t_i), F_\lambda(t_i)]$ (Michael J. Schnieders et al., 2012):

$$g_m(\lambda, F_\lambda) = \sum_{t_i} h\, e^{\left(\frac{|\lambda-\lambda(t_i)|^2}{2w_1^2} \times \frac{|F_\lambda - F_\lambda(t_i)|^2}{2w_2^2}\right)} \qquad \textbf{Eqn. 15}$$

The additional biasing dimension promotes crossing of hidden barriers relative to the simpler one-dimensional bias of original metadynamics approaches. This motivates the

choice of OSRW for protein loop optimization, where the goal is to broadly explore the conformational landscape to discover the minimum free energy configuration.

2.2.5: Previous Loop Optimization Efforts

In some respects, the challenge of modeling protein loops is a subset of the general protein folding problem and much has been learned from prior attempts at *ab initio* loop structure prediction. Conceptually, it is useful to subdivide loop determination into three interrelated pieces: 1) loop closure, 2) enumeration of conformations and 3) choice of target function or score. Loop closure bridges known peptide carboxyl and amino termini anchors with a viable starting conformation. Next, conformational enumeration proceeds using a sampling scheme such as MD or Monte Carlo. Both loop closure and conformational enumeration depend on a target function or score to access the relative probability of loop coordinates and ultimately choose the best conformation (or ensemble of conformations).

Loop closure is often done under geometric constraints that are consistent with peptide backbone geometry (i.e. steric overlaps must be avoided and low-energy ϕ-ψ angles achieved). Furthermore, loop closure of three residues or less must have a discrete number of possible loop conformations based on peptide ring closure work (Go^{-} & Scheraga, 1970) that showed the number of geometric constraints is consistent and equal to the number of degrees of freedom (i.e. six torsion angles for three residue loops). Because there are a discrete number of loop conformations for three residues, inverse kinematics methods have been developed for loop conformations (Canutescu & Dunbrack, 2003; Coutsias, Seok, Jacobson, & Dill, 2004). First, a greedy algorithm known as Cyclic Coordinate Descent (CCD) iterates over any number of loop dihedral

angles and adjusts the angle to minimize the sum of the squared distance between back bone atoms of the moving C-terminal anchor and the known fixed C-terminal anchor (Canutescu & Dunbrack, 2003). Another commonly used algorithm stemming from inverse kinematics, kinematic closure (KIC), reduces the tripeptide closure problem to a 16-degree polynomial and analytically solves for all possible loop conformations (Coutsias et al., 2004). However, the initial conformation of loop atoms may not ultimately limit the quality of conformations (Fiser et al., 2000), such that random placement or buildup of atoms is sufficient for closure when used in tandem with a robust global optimizer. Even so, it is reasonable to expect that initial conformations near the global minima of the target function will converge most efficiently. Other loop closure methods include random tweak (Fine, Wang, Shenkin, Yarmush, & Levinthal, 1986; Xiang, Soto, & Honig, 2002), direct tweak (Soto, Fasnacht, Zhu, Forrest, & Honig, 2008), a meet in the middle approach (Jacobson et al., 2004; Spassov, Flook, & Yan, 2008; Zhu, Pincus, Zhao, & Friesner, 2006), and polypeptide fragment mining based on closure gap distance (Deane & Blundell, 2000, 2001; Ko et al., 2011).

Protein loop conformations produced from loop closure are the starting coordinate seeds for conformational optimization techniques that locally or globally explore the search space of loop conformations. Methods to explore search space include MD simulations (R. E. Bruccoleri & Karplus, 1990), SA (Collura, Higo, & Garnier, 1993) or a variety of Monte Carlo (MC) move sets to locate low energy conformations (Li & Scheraga, 1987). In the case of MC, local moves propose new loop conformations followed by application of the Metropolis acceptance-rejection criterion (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) to decides whether to keep the original conformation or accept the

new placement. Several MC variations have explored both move sets and alternative acceptance criteria to optimize protein structure, including hierarchical MC (Jacobson et al., 2004), biased MC searches (Abagyan & Totrov, 1994; M. G. Wu & Deem, 1999) and a few others (Cui, Mezei, & Osman, 2008; Mandell, Coutsias, & Kortemme, 2009; Rohl, Strauss, Chivian, & Baker, 2004). Other optimization efforts considered use of fine-grained sampling in conjunction with conjugate gradients (Fiser et al., 2000), limited-memory BFGS (de Bakker, DePristo, Burke, & Blundell, 2003; DePristo, de Bakker, Lovell, & Blundell, 2003), Newton-Raphson minimization (Spassov et al., 2008), side chain optimization (Lee, Lee, Park, Coutsias, & Seok, 2010) and steepest descent minimization (Cui et al., 2008).

Finally, enumeration and evaluation of the optimization steps or loop conformations is guided by a variety energy functions ranging from fixed charged force fields (Fiser et al., 2000; Jacobson et al., 2004), statistical or knowledge based potentials (Das & Baker, 2008) and/or potentials that incorporate experimental data (Brunger, 2007; Trabuco et al., 2008). In this work, we focus on a target function that is a weighted combination of the polarizable AMOEBA force field and electron density maps that arise from either X-ray crystallography or Cryo-EM. Combination of this hybrid target rests on maximum-likelihood principles (Murshudov, Vagin, & Dodson, 1997) and is defined by

$$E_{\mathrm{Tot}} = E_{\mathrm{chem}} + w_A E_{\mathrm{x-ray}} \qquad \textbf{Eqn. 16}$$

where E_{tot} is the hybrid target, E_{chem} is the force-field energy, $E_{x\text{-}ray}$ is a metric of agreement between real-space map (Cowtan, 2005; Read, 1986) and the electron density map, and w_A is the weight given to that metric.

CHAPTER 3: GLOBAL LOOP OPTIMIZATION USING A POLARIZABLE FORCE FIELD

GONDOLA is flexible global optimization strategy for structural biophysics. Here we will focus on its relative merits for finding optimal protein loop conformations using the OSRW flavor of metadynamics and a target function that combines experimental data and a polarizable force field (i.e. AMOEBA). In this case, the experimental target function uses a real-space density map (i.e. a SigmaA weighted $2F_o$-F_c map) that was used successfully in previous PCNA structure refinement work (LuCore et al., 2015).

3.1: Formal Loop Definition

Given two consecutive secondary structure elements, a loop can be defined as residues in the range [X_1-X_{n-1}] where X_0 corresponds to the terminal residue of the secondary structure proximal to the N terminus of the defined loop, and X_n represents the origin of the second secondary structure. X_0 and X_n serve as anchoring points for loop closure. In practice, this definition can be relaxed to include upstream and/or downstream structural regions in any conformation, so long as they serve to anchor the beginning and end of intermediate sequence (i.e. loop) that will be optimized. Despite being included in our loop definition, the terminal coordinates are immobile during optimization procedures. Furthermore, the loop optimization occurs with all non-loop elements remaining stationary during molecular dynamics simulation; however, all atoms contribute to the total potential of the system.

3.2: Loop Build-Up

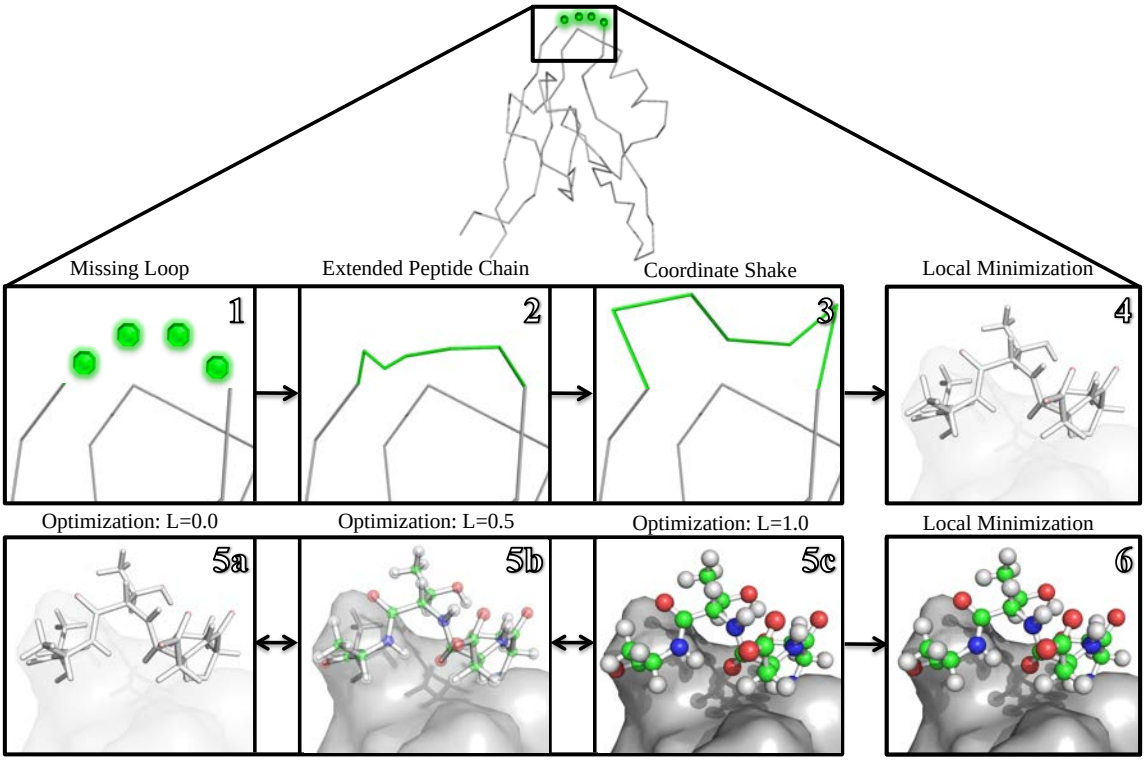

Figure 2. Shown are the initial steps of placing protein loops (1-3), the initial starting seed relaxation (4), back and forth metadynamic exploration of loop conformation space where L is λ state value (5a-5c), and the final loop local minimization (6).

The algorithm begins by building the missing residues along a vector from the X0 carbonyl carbon to the nitrogen of Xn as an extended polypeptide chain that connects the defined anchoring residues (step 2 of Figure 2). The initial coordinates of built loops are given a small, random coordinate bump (step 3 of Figure 2) because off-center starting coordinates avoid numeric instabilities (e.g. singular multipole rotation matrices) inherent to our target function. Next, the loop is subject to a local minimization using the Limited Memory-BFGS (L-BFGS) method without including non-bonded energy terms (step 4 of Figure 2). By removing van der Waals interactions while bonds and angles are relaxed, unreasonably large steric hindrance energies are avoided. The resulting conformation provides a relatively stable starting seed for the downstream metadynamics step.

3.3: Global Optimization Using Metadynamics and a Polarizable Force Field (GONDOLA)

3.3.1: Defining the λ Path

The third and most expensive step uses the OSRW variant of metadynamics to explore conformational space for the global loop minimum. For MD that updates positions and velocities using Beeman integration (Beeman, 1976), coupling to a heat bath at 300 K is performed using either the Berendsen velocity rescaling thermostat (Berendsen, Postma, Vangunsteren, Dinola, & Haak, 1984) or the Bussi thermostat (Bussi, Zykova-Timan, & Parrinello, 2009). Alternatively, stochastic dynamics (Allen, 1980) can also be used with a time step of 1 fs. As before, the coordinates of non-loop atoms are fixed while the loop is allowed to move according to the integration scheme. The non-loop portion of the protein contributes to the potential energy as a function of the state variable λ. This controls the strength of non-bonded energy terms (i.e. softcore van der Waals and electrostatic interactions), and as the simulation proceeds λ varies continuously between 0 and 1 (Michael J. Schnieders et al., 2012). Favorable conformations allow loop and non-loop atoms to attain their full non-bonded interactions (i.e. λ=1), which are consistent with the condensed phase environment (i.e. the crystal). When λ is zero, the loop atoms only experience bond stretching and angle bending terms, which can be thought of as an alchemical (i.e. unphysical) vapor-like phase. While the loop is in such an alchemical state, it can easily escape conformations that are confined by potential energy barriers found in the crystalline environment. In addition to smoothly eliminating non-bonded interactions as λ approaches 0, torsion, pi-torsion and torsion-torsion AMOEBA energy terms are also scaled to zero. Thus, in the unphysical λ=0 state, the potential has

effectively been smoothed to remove all barriers to rotations about dihedral angles, leaving only bond-stretching and angle-bending topological restraints.

3.3.2: Condensed Phase Quenching

Periodic local minimizations quench the system to assess the depth of the current potential energy well, based on the condensed phase potential energy (i.e. $\lambda=1$). Because the objective of the search is to find the global minimum of the physical end state and not the alchemical vapor state (i.e. $\lambda=0$), assessment of conformational energy and fit to the data is only done when $\lambda > 0.5$. Importantly, for λ greater than 0.5, the loop conformation is favorable enough that softcore non-bonded interactions are substantially contributing and scaled bonded terms terms are at least half of their physical value. A typical ensemble average of the partial derivative of the potential energy function (**Figure 3)** demonstrates that the state of the protein loop is driven towards condensed phase whenever the protein loop has discovered a favorable environment (i.e. $<\partial U/\partial\lambda>$ is less than 0 for λ greater than 0.5).

Assessment of λ state occurs at intervals of 1000 fs. Shorter intervals of simulation time between assessments would use unnecessary computational resources without allowing the loop time to explore a new potential well. Similarly, intervals that are too larger may miss minima due to the time-dependent bias driving the simulation to explore new regions of phase space.

Local minimizations are performed on an unbiased hybrid target. The energy and coordinates of the condensed phase optimized structure are saved if they are the energy value is the lowest the simulation has found. After minimization, the loop is reverted back to the coordinates and λ state to allow the OSRW search to continue.

3.3.3: The Biasing Potential

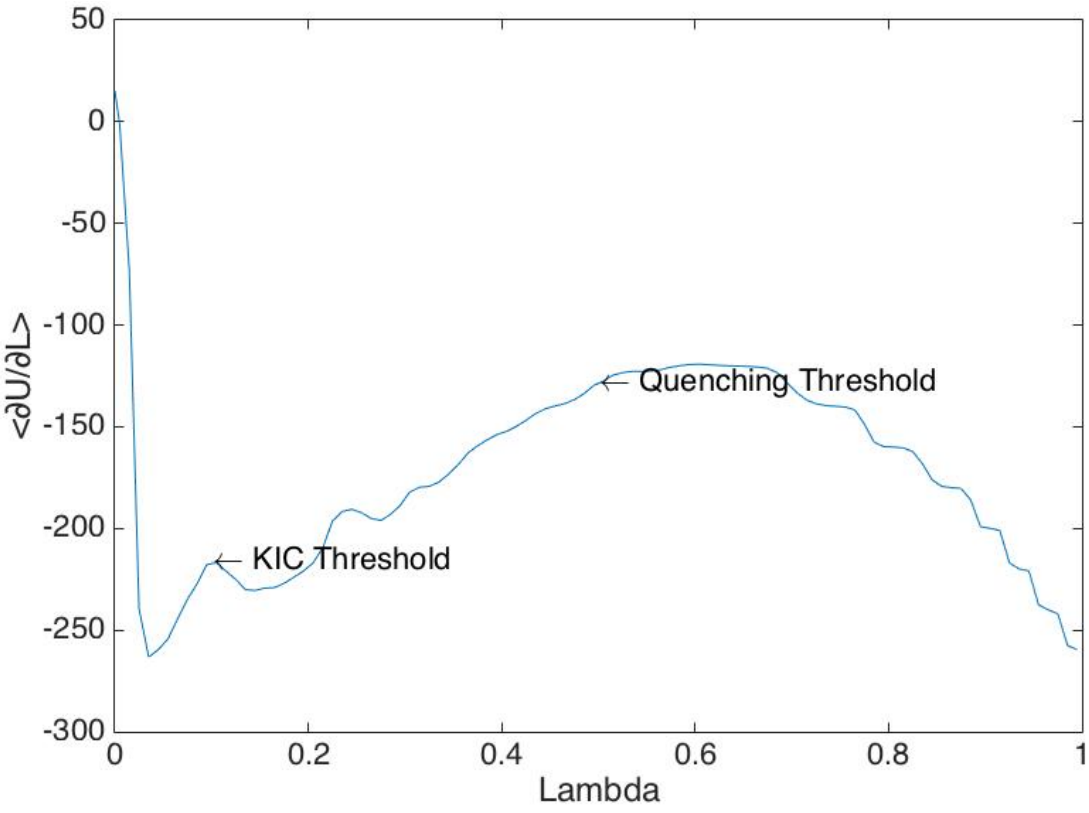

Figure 3. Shown is a typical ensemble average partial derivative of the potential energy with respect to lambda ($\langle \partial U/\partial \lambda \rangle$). Its value as a function of λ for loop optimization demonstrates that the loop is readily pushed toward the λ=1 state (i.e. condensed phase) except for λ near 0. The threshold for local minimization (quenching) as described in section 3.3.2 is marked at λ=0.5. The topological clues provided by this graph (i.e. barriers present for various λ states) support this decision because states of λ > 0.5 are closer to the true potential energy surface seen by λ=1. KIC moves can be attempted when λ < 0.1 (see section 3.3.4) to improve search space exploration for highly smoothed non-bonded interactions.

While molecular dynamics is running at 300K on the loop portion of the protein, the OSRW method works to flatten the energy landscape by applying Gaussian bias potentials that are a function of the state variable (λ) and the partial derivative of the potential energy in terms of λ (i.e. ∂U/∂λ). The bias enables the loop to sample coordinate search space efficiently by promoting escape over barriers via smoothing the potential energy surface. Thus, OSRW presents a computationally efficient global optimization

procedure, if run for long enough that will flatten the potential energy landscape and allow the loop to cross barriers that exist along both λ and ∂U/∂λ (i.e. hidden barriers). Optionally, it is possible to tune efficiency by utilizing a higher initial bias magnitude (i.e. bias > 0.002 Kcal/mol). This may allow the optimization to escape barriers more quickly; however, the buildup of large bias potentials becomes problematic for the ∂U/∂λ term. Unlike λ, ∂U/∂λ is not bounded, so accumulation of bias potential along ∂U/∂λ pushes the protein loop to explore increasingly high-energy states, which can eventually lead to unstable simulations. A somewhat analogous concern exists for high-temperature SA, where increasingly high atomic velocities require reductions in MD time step. The concerns associated with using large Gaussian magnitudes can be mitigated by the use of transition tempered OSRW (TT-OSRW) (Dama, Rotskoff, Parrinello, & Voth, 2014). TT-OSRW systematically decreases the size of new bias potentials by a depreciating scalar as the sum of bias potential and true potential reaches a flat potential energy surface. By using the current minimum bias height the TT-OSRW method scales by

$$\text{Bias}_{\text{current}} = \text{Bias}_{\text{orig}} e^{\left(-\frac{min-height}{\Delta T}\right)} \quad \textbf{Eqn. 17}$$

where ΔT tunes the decay rate.

3.3.4: Vapor Phase MC with KIC Based Move Set

Although the biasing potential eventually smoothes the overall potential energy surface along both λ and ∂U/∂λ, escaping deep energy wells may still consume a large amount of compute power. To promote discovery of substantially different loop configurations, it is possible to propose aggressive MC moves every 1000 fs if the loop is near the vapor state (i.e. non-bonded repulsion has been turned off). We define “near” vapor phase for the purposes of MC moves as λ < 0.1, although the success probability of a given MC move set can be further tuned by changing this limit. The proposed MC move is defined as a

chain move of five iteration calls to the tripeptide analytic solution of KIC. Each of the iterations picks the center residue to be a random residue from X_{n-1} to X_1. The criterion is applied only to the final solution such that the chained moves are aggregated to propose a drastically altered configuration. These large swings in the vapor phase allow the loop to explore new wells and cross barriers more efficiently. Additionally, because the move is an analytically derived correct answer to a sub-problem, the loop is guided towards minima containing this solution.

3.3.5: Parallelization

Parallelization of this algorithm was achieved through using a Message Passing Interface defined by the Parallel Java API (Kaminsky, 2007). The software is launched on a main compute node for scheduling. Then each compute node in use builds a loop via the random build-up, which allows each node to have a different starting seed. The nodes run their own GONDOLA, but all of the calculations contribute to the same global histogram (i.e. the same OSRW bias). In other words, each built loop shares its search findings (repulsive Gaussians) with the other loops. Finally, optimal loops are locally minimized using L-BFGS with a reciprocal space crystallography target to finalize coordinates and b-factors.

3.4: Finalizing Structure and Metrics

The final steps after **Figure 2** include 1) a full potential minimization of the loop atom coordinates, 2) a hybrid target local optimization for the entire protein (coordinates and b-factors), and 3) evaluation of the final loop conformation using experimental metrics (R/R_{free}) and structural metrics including clashes, poor rotamers, unfavorable dihedral angles, and other bond geometry using the MolProbity structure validation tool, which

assigns structural scores based on van der Waals contacts, hydrogen-bond distances, side-chain rotamers, and peptide backbone conformation.

CHAPTER 4: PROTEIN LOOP OPTIMIZATION RESULTS

The first goal of the protein loop applications presented below is to demonstrate that limitations of previous generation fixed charge force fields for scoring loops can be overcome using the polarizable atomic multipole AMOEBA force field. The second goal is to assess the relative efficiency of the novel metadynamics search protocol compared to previous loop optimization approaches such as simulated annealing.

4.1: Force Fields as Scoring Functions

The target functions explored here are a sum of force field and experimental energy terms. If the former recapitulates the crystalline environment to high degree, there will be concordance with the experimental data (i.e. the target function will be relatively insensitive to the weighting of the force field and experimental energy terms). To access the concordance, **Figure 4** shows the agreement between the R_{free} value for the experimental data and two different force field potential energies.

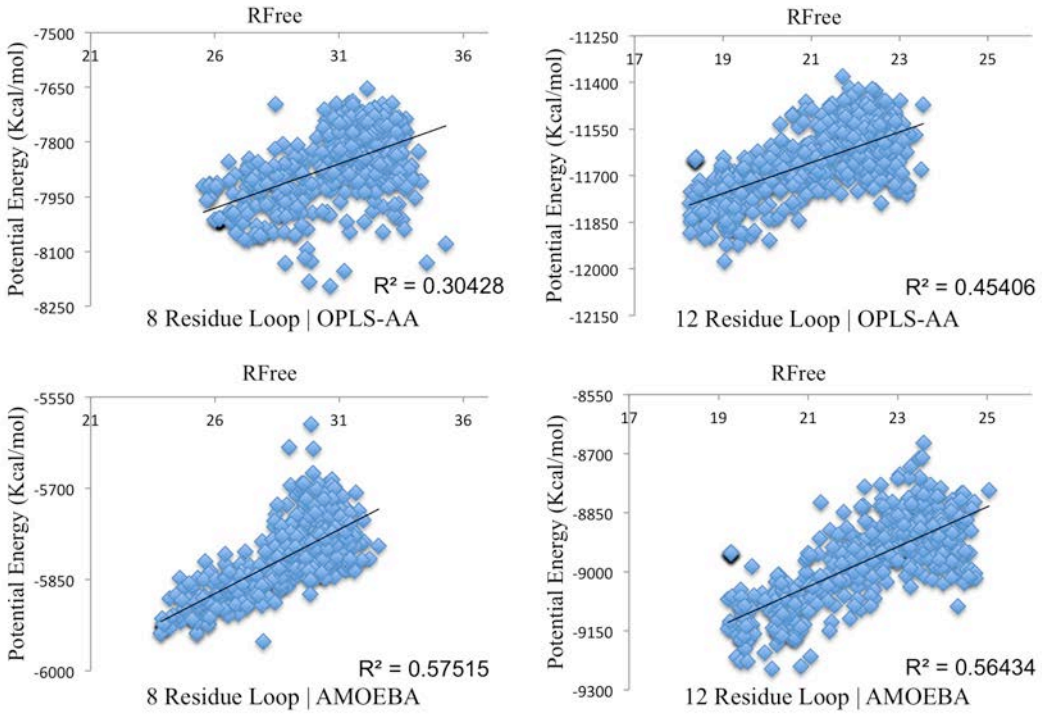

Figure 4. 574 eight residue long decoy loops corresponding to PDB ID 1CBS and 553 twelve residue long decoy loops corresponding to PDB ID 1AKZ from a commonly used loop decoy data set (Jacobson et al., 2004) were scored using a polarizable force field (AMOEBA) and a fixed charged force field (OPLS-AA) (shown on the top-left and bottom-left). Each loop received a local minimization in the respective force field, and then the force field potential energies were compared to the R_{free} value of each structure. AMOEBA is shown to correlate better to R_{free} values than OPLS-AA, which supports the claim that a polarizable force field serves as a more accurate scoring function.

The AMOEBA and OPLS-AA force fields showed some correlation (i.e. coefficient of determination > 0) to R_{free}, suggesting that both force fields provide reasonable forces for crystalline protein simulations. However, the higher order multipoles and polarization of the AMOEBA force field more accurately represented the true crystalline environment. Thus, scoring of loops with AMOEBA provides more realistic evaluations. **Figure 4** also shows that neither force field is perfectly correlated (i.e. coefficient of determination = 1) to experimental data, which supports the need for a hybrid target function. Note that evaluation of force field potential energy does not capture entropy (i.e. well depth is

critical and well width is ignored) such that its minima are not equivalent free energy minimum. Each of these points motivates use of the AMOEBA potential (i.e. better representation of the crystalline environment) as part of a hybrid target (i.e. experimental data allows us to bias the scoring function toward free energy minima) function for optimizing protein loops (Fenn & Schnieders, 2011).

4.2: Convergence Analysis: Metadynamics Compared to SA

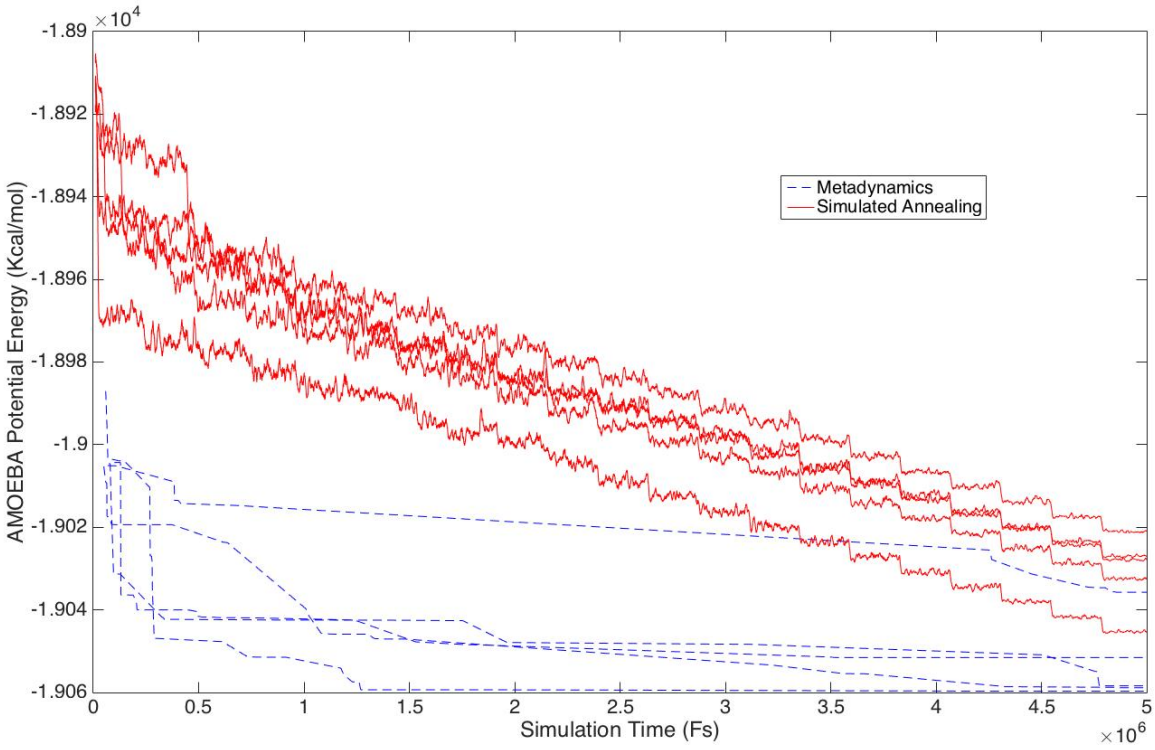

Figure 5. Shown are ten loop optimizations of the four residue loop in PDB ID 4GVD occurring over 5ns of simulation time (five simulated annealing simulations following a typical cooling protocol (Hart et al., 2000) and five simulations based on the metadynamics approach discussed in this thesis). The four lowest energy structures were outputs from the metadynamics approach. The sixth best loop also belonged to the metadynamic method.

Establishing a target function independently of loop optimization allowed analysis between current gold standards in loop optimization and the GONDOLA approach as seen in **Figure 5**. Out of five metadynamics (i.e. GONDOLA) and five simulated annealing trials, GONDOLA discovered four loops with energies lower than any discovered via SA.

Furthermore, three loop conformations with energies lower than any found by SA were achieved before 0.5 ns of simulation time. The GONDOLA approach allows the search to be restarted (i.e. continuation from any of the achieved loops), whereas SA requires the entire temperature schedule to be repeated (i.e. temperature must be raised to allow kinetic energy to overcome the highest barrier between the current well and that of the global minimum).

4.3: Building PDZ Domains and Rebuilding Known Loops

4.3.1: Initial Structure Evaluation

The performance of loop optimization was tested against two datasets. The first contains three PDZ Domains, which do not have coordinates determined for a section of the carboxylate binding loop, and two of these proteins are also missing a two residue loop slightly further down the chain. The proteins that are missing two loops had both loops built and optimized simultaneously. The second data set removes known loop coordinates from a loop decoy data set that has been used extensively for previous loop building assessment (Jacobson et al., 2004).

Table 1. The PDB ID, resolution, R/R_{free} values and MolProbity analyses for the deposited protein models are shown. Three PDZ Domains are listed, and each is missing one or more protein loops. Seven additional loops were assessed to measure the algorithms performance over increasing loop lengths. All seven of these latter loops have known conformations based on structures in the Protein Data Bank.

Data	Loop	Res.	Reported		FFX		MolProbity		Clash		Rama		Poor
Set	Size	(Å)	R	R_{free}	R	R_{free}	Score	%	Score	%	Out	Fav	Rot%
PDZ DOMAINS													
4NXP	5	2.30	25.8	21.7	24.0	25.3	2.00	90	5.28	99	0.0	97.6	6.8
4GVD	4, 2	1.85	24.3	19.6	22.9	25.0	1.94	70	8.99	82	0.0	98.3	4.1
3KZE	2, 2	1.80	21.2	19.6	19.1	21.0	1.56	91	5.69	94	0.0	99.3	2.2
LOOP DECOYS													
1CBS	4	1.80	20.0	23.7	19.4	20.0	1.40	97	4.01	98	0.0	97.8	1.6
2ARC	5	1.50	17.9	23.2	18.8	23.2	2.02	40	6.79	84	0.0	97.0	4.1
2ARC	6	1.50	17.9	23.2	18.8	23.2	2.02	40	6.79	84	0.0	97.0	4.1
1CBS	7	1.80	20.0	23.7	19.4	20.0	1.40	97	4.01	98	0.0	97.8	1.6
Mean		1.79	21.0	22.1	20.4	22.5	1.76	75	5.94	91	0.0	97.8	3.5

All of the deposited structures had a resolution better than 2.0 Å, except 4NXP, which was slightly worse at 2.30 Å. The R_{free} values were recomputed with the FFX software package to provide a baseline for a direct comparison with optimized structures. The difference between R and R_{free} in 1CBS, 3KZE, and 4NXP did not correspond to the deposited values, which could indicate that the correct R_{free} flags were not deposited. MolProbity assessment further indicates that there are minimal clashes, zero Ramachandran outliers, and some poor rotamers.

4.3.2: Evaluation of Optimized Structures

Table 2. The R/R_{free} values and MolProbity analyses for the PDZ Domain and loop decoy data sets are given for SA using the OPLS-AA fixed charged force field and GONDOLA under the AMOEBA force field refinement methods. All R/R_{free} values were calculated in FFX for consistency. The order which loops are presented is identical to **Table 1**.

Data	Optimization			MolProbity		Clash		Rama		Poor
Set	Method	R	R_{free}	Score	%	Score	%	Out %	Fav %	Rot %
PDZ	**DOMAINS**									
4NXP	SA | OPLS	21.03	29.06	1.84	95	0.7	100	2.0	94.4	9.1
	GONDOLA	20.43	26.24	1.26	100	0.0	100	2.0	96.6	5.2
4GVD	SA | OPLS	20.20	23.23	1.12	100	0.0	100	1.0	96.8	3.6
	GONDOLA	20.24	23.22	1.50	95	3.0	99	0.0	97.4	2.4
3KZE	SA | OPLS	21.45	27.04	1.42	96	1.3	100	2.0	97.9	4.9
	GONDOLA	21.05	25.80	1.40	96	2.7	99	2.0	99.0	2.9
LOOP	**DECOYS**									
1CBS	SA | OPLS	17.93	22.41	1.77	80	1.8	100	2.0	95.6	4.9
	GONDOLA	17.59	22.21	1.70	89	2.3	99	2.0	95.6	3.3
2ARC	SA | OPLS	18.85	23.25	2.11	32	7.2	81	3.0	96.6	4.4
	GONDOLA	18.36	22.38	1.88	53	3.8	96	4.0	96.0	3.7
2ARC	SA | OPLS	18.20	22.57	1.98	43	6.7	85	2.0	96.6	3.3
	GONDOLA	17.68	22.04	1.55	82	2.3	99	1.0	97.2	3.3
1CBS	SA | OPLS	19.73	24.16	2.01	62	2.7	99	3.0	91.9	4.1
	GONDOLA	21.04	25.69	1.61	100	1.4	100	2.0	94.8	3.3
Mean	SA | OPLS	19.63	24.53	1.75	73	2.9	95	2.1	95.7	4.9
	GONDOLA	19.48	23.94	1.56	88	2.2	99	1.9	96.6	3.4

Table 2 demonstrates that, on average, GONDOLA was able to achieve a lower R_{free} value (i.e. the models represent experimental data more accurately), a MolProbity score that was lower than both the SA annealing method and initial structure, fewer steric

clashes, fewer unfavorable Ramachandran values, and decreased the number of poor rotamers compared to the initial structure. As was shown in **Figure 5**, the convergence rate is dependent on the random build-up loop structure (i.e. the starting seed); therefore, it is not surprising the SA method was able to achieve a lower R_{free} value (i.e. 24.16 < 25.69) in one of the seven loops attempted (7 residue loop of 1CBS). SA was also able to yield a better MolProbity score for 4GVD despite having a slightly higher R_{free} value compared to the output of GONDOLA.

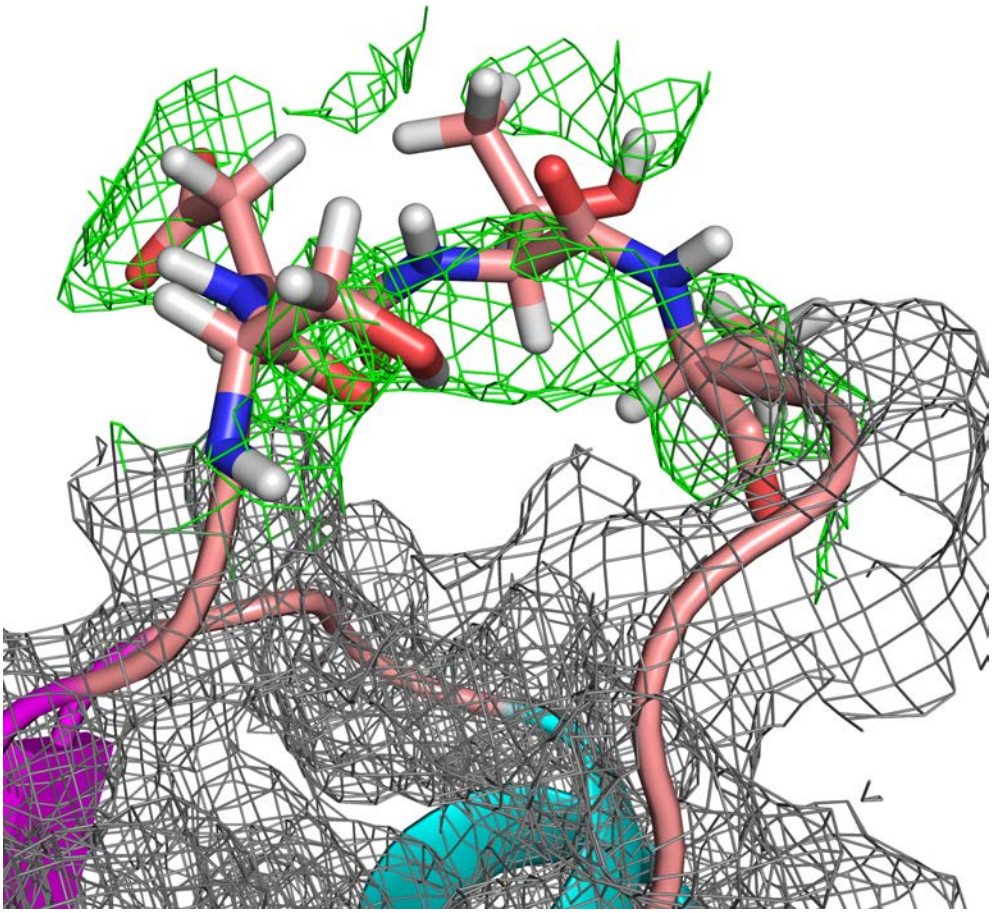

Figure 6. The protein structure of a PDZ domain in complex with Syndecan1 peptide (PDB ID: 4GVD) is shown with a 2Fo-Fc map contoured at 0.75σ. Residues predicted with GONDOLA are shown as stick models, and the corresponding map contours are highlighted in green.

CHAPTER 5: CONCLUSION

5.1: Summary of the GONDOLA Approach

The overarching goal of this work was to produce an efficient search method for computing ab initio protein structures and to evaluate the benefits gained by using a polarizable force field and a hybrid target function. Three carboxylate binding loops and two small two residue loops from PDZ Domains were constructed from PDB files missing coordinate data. For the well-behaved system, PDB ID 4GVD, the overall structure's R_{free} was decreased using both SA with the OPLS-AA fixed charge force field, and the GONDOLA approach. Known protein loops from a heavily used loop modeling data set were also reconstructed.

GONDOLA was shown to converge faster than simulated annealing in **Figure 5** while **Figure 4** demonstrated that scoring loop conformations with the AMOEBA potential more correlated to experimental metrics (i.e. R_{free}) than fixed partial charge force fields for crystalline environments. Comparisons made in **Table 2** confirmed that for every metric (i.e. R_{free}, vdW clash score, Ramachandran peptide backbone angles, and side-chain conformations) that, on average, the GONDOLA approach was able to construct the protein loop more accurately than SA with a fixed charge force field.

5.2: Future Direction and Alternative Applications

With genomic and protein sequence data being gathered at astonishing and increasing rates, experimental protein structure determination cannot keep pace. Therefore, it is imperative to augment experimental structures with predictive models. Furthermore, ab initio methods for building missing residues, loops and domains during refinement of experimental models are also critical. This work demonstrates that GONDOLA is well-

suited to refine or predict the coordinates of missing residues. The flexibility in the optimization protocol allows GONDOLA to be easily reconfigured for refinement of side-chains (i.e. set side-chain atoms active while constraining backbone atoms), residues (i.e. side-chain atoms and backbone atoms active), ligands (i.e. set atoms along binding interface active), or any desired portion of a protein.

Of particular future interest are the benefits of GONDOLA for homology modeling. Current theory that underlies the widely used SwissMod (Kiefer, Arnold, Kunzli, Bordoli, & Schwede, 2009) and Modbase (Pieper et al., 2014) homology modeling databases attempt to use multiple sequence alignments, coupled with assumptions of evolutionary conservation of protein folds to thread new protein sequences onto existing structures using fixed charge force fields. As one could imagine, the proteins are not identical, and this approach may fail for regions where the target sequence strays from the template sequence. These regions between highly conserved folds are fundamentally similar to the protein loops analyzed in this thesis, and therefore, may be amenable to GONDOLA's accurate ab initio structure prediction.

Furthermore, GONDOLA is not restricted to crystalline environments or the time-dependent biasing protocol described here. For example, alternative biasing potentials might focus on sampling backbone or residue torsional angles. The AMOEBA force field potential could also be augmented to include explicit or implicit solvent, which may further improve conformational preferences of surface exposed residues, the binding interface between a protein and its ligand, or any other application where solvent environment is critical.

REFERENCES

Abagyan, R., & Totrov, M. (1994). BIASED PROBABILITY MONTE-CARLO CONFORMATIONAL SEARCHES AND ELECTROSTATIC CALCULATIONS FOR PEPTIDES AND PROTEINS. *Journal of Molecular Biology, 235*(3), 983-1002. doi:10.1006/jmbi.1994.1052

Alder, B. J., & Wainwright, T. E. (1959). Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics, 31*(2), 459-466. doi:doi:http://dx.doi.org/10.1063/1.1730376

Allen, M. P. (1980). BROWNIAN DYNAMICS SIMULATION OF A CHEMICAL-REACTION IN SOLUTION. *Molecular Physics, 40*(5), 1073-1087. doi:10.1080/00268978000102141

Anfinsen, C. B. (1973). PRINCIPLES THAT GOVERN FOLDING OF PROTEIN CHAINS. *Science, 181*(4096), 223-230. doi:10.1126/science.181.4096.223

Anisimov, V. M., Lamoureux, G., Vorobyov, I. V., Huang, N., Roux, B., & MacKerell, A. D. (2005). Determination of electrostatic parameters for a polarizable force field based on the classical Drude oscillator. *Journal of Chemical Theory and Computation, 1*(1), 153-168. doi:10.1021/ct049930p

Barducci, A., Bonomi, M., & Parrinello, M. (2011). Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 1*(5), 826-843. doi:10.1002/wcms.31

Barducci, A., Bussi, G., & Parrinello, M. (2008). Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters, 100*(2). doi:10.1103/PhysRevLett.100.020603

Beeman, D. (1976). SOME MULTISTEP METHODS FOR USE IN MOLECULAR-DYNAMICS CALCULATIONS. *Journal of Computational Physics, 20*(2), 130-139. doi:10.1016/0021-9991(76)90059-0

Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A., & Haak, J. R. (1984). MOLECULAR-DYNAMICS WITH COUPLING TO AN EXTERNAL BATH. *Journal of Chemical Physics, 81*(8), 3684-3690. doi:10.1063/1.448118

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235-242. doi:10.1093/nar/28.1.235

Bernstein, L. S., Ramineni, S., Hague, C., Cladman, W., Chidiac, P., Levey, A. I., & Hepler, J. R. (2004). RGS2 binds directly and selectively to the M1 muscarinic acetylcholine receptor third intracellular loop to modulate G(q/11 alpha) signaling. *Journal of Biological Chemistry, 279*(20), 21248-21256. doi:10.1074/jbc.M312407200

Böttcher, C. J. F. (1993). Dielectrics in Static Fields *Theory of Electric Polarization* (2 ed.). Amsterdam: Elsevier Pub. Co.

Bruccoleri, R. (2000). Ab Initio Loop Modeling and Its Application to Homology Modeling. In D. Webster (Ed.), *Protein Structure Prediction* (Vol. 143, pp. 247-264): Humana Press.

Bruccoleri, R. E., & Karplus, M. (1990). CONFORMATIONAL SAMPLING USING HIGH-TEMPERATURE MOLECULAR-DYNAMICS. *Biopolymers, 29*(14), 1847-1862. doi:10.1002/bip.360291415

Brunger, A. T. (2007). Version 1.2 of the Crystallography and NMR system. *Nature Protocols, 2*(11), 2728-2733. doi:10.1038/nprot.2007.406

Brunger, A. T., Krukowski, A., & Erickson, J. W. (1990). SLOW-COOLING PROTOCOLS FOR CRYSTALLOGRAPHIC REFINEMENT BY SIMULATED ANNEALING. *Acta Crystallographica Section A, 46*, 585-593. doi:10.1107/s0108767390002355

Bussi, G., Zykova-Timan, T., & Parrinello, M. (2009). Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *Journal of Chemical Physics, 130*(7). doi:10.1063/1.3073889

Byrd, R. H., Lu, P. H., Nocedal, J., & Zhu, C. Y. (1995). A LIMITED MEMORY ALGORITHM FOR BOUND CONSTRAINED OPTIMIZATION. *Siam Journal on Scientific Computing, 16*(5), 1190-1208. doi:10.1137/0916069

Canutescu, A. A., & Dunbrack, R. L. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science, 12*(5), 963-972. doi:10.1110/ps.0242703

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., . . . Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D-Biological Crystallography, 66*, 12-21. doi:10.1107/s0907444909042073

Ciarapica, R., Rosati, J., Cesareni, G., & Nasi, S. (2003). Molecular recognition in helix-loop-helix and helix-loop-helix leucine zipper domains - Design of repertoires and selection of high affinity ligands for natural proteins. *Journal of Biological Chemistry, 278*(14), 12182-12190. doi:10.1074/jbc.M211991200

Collura, V., Higo, J., & Garnier, J. (1993). MODELING OF PROTEIN LOOPS BY SIMULATED ANNEALING. *Protein Science, 2*(9), 1502-1510. Retrieved from <Go to ISI>://WOS:A1993LU65000015

Coutsias, E. A., Seok, C., Jacobson, M. P., & Dill, K. A. (2004). A kinematic view of loop closure. *Journal of Computational Chemistry, 25*(4), 510-528. doi:10.1002/jcc.10416

Cowtan, K. (2005). Likelihood weighting of partial structure factors using spline coefficients. *Journal of Applied Crystallography, 38*, 193-198. doi:10.1107/s0021889804031474

Cui, M., Mezei, M., & Osman, R. (2008). Prediction of protein loop structures using a local move Monte Carlo approach and a grid-based force field. *Protein Engineering Design & Selection, 21*(12), 729-735. doi:10.1093/protein/gzn056

Dama, J. F., Rotskoff, G., Parrinello, M., & Voth, G. A. (2014). Transition-Tempered Metadynamics: Robust, Convergent Metadynamics via On-the-Fly Transition Barrier Estimation. *Journal of Chemical Theory and Computation, 10*(9), 3626-3633. doi:10.1021/ct500441q

Das, R., & Baker, D. (2008). Macromolecular modeling with Rosetta *Annual Review of Biochemistry* (Vol. 77, pp. 363-382).

de Bakker, P. I. W., DePristo, M. A., Burke, D. F., & Blundell, T. L. (2003). Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the generalized born solvation model. *Proteins-Structure Function and Genetics, 51*(1), 21-40. doi:10.1002/prot.10235

Deane, C. M., & Blundell, T. L. (2000). A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins-Structure Function and Genetics, 40*(1), 135-144. doi:10.1002/(sici)1097-0134(20000701)40:1<135::aid-prot150>3.3.co;2-t

Deane, C. M., & Blundell, T. L. (2001). CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Protein Science, 10*(3), 599-612. doi:10.1110/ps.37601

DePristo, M. A., de Bakker, P. I. W., Lovell, S. C., & Blundell, T. L. (2003). Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles. *Proteins-Structure Function and Genetics, 51*(1), 41-55. doi:10.1002/prot.10285

Dmitriev, O. Y., & Fillingame, R. H. (2007). The rigid connecting loop stabilizes hairpin folding of the two helices of the ATP synthase subunit c. *Protein Science, 16*(10), 2118-2122. doi:10.1110/ps.072776307

Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M., & MacKinnon, R. (1996). Crystal structures of a complexed and peptide-free membrane protein-binding domain: Molecular basis of peptide recognition by PDZ. *Cell, 85*(7), 1067-1076. doi:10.1016/s0092-8674(00)81307-0

Elking, D. M., Cisneros, G. A., Piquemal, J. P., Darden, T. A., & Pedersen, L. G. (2010). Gaussian Multipole Model (GMM). *Journal of Chemical Theory and Computation, 6*(1), 190-202. doi:10.1021/ct900348b

Espadaler, J., Querol, E., Aviles, F. X., & Oliva, B. (2006). Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics, 22*(18), 2237-2243. doi:10.1093/bioinformatics/btl382

Fenn, T. D., & Schnieders, M. J. (2011). Polarizable atomic multipole X-ray refinement: weighting schemes for macromolecular diffraction. *Acta Crystallographica Section D-Biological Crystallography, 67*, 957-965. doi:10.1107/s0907444911039060

Fenn, T. D., Schnieders, M. J., Brunger, A. T., & Pande, V. S. (2010). Polarizable Atomic Multipole X-Ray Refinement: Hydration Geometry and Application to Macromolecules. *Biophysical Journal, 98*(12), 2984-2992. doi:10.1016/j.bpj.2010.02.057

Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L., & Levinthal, C. (1986). PREDICTING ANTIBODY HYPERVARIABLE LOOP CONFORMATIONS II MINIMIZATION AND MOLECULAR DYNAMICS STUDIES OF MCPC603 FROM MANY RANDOMLY GENERATED LOOP CONFORMATIONS. *Proteins Structure Function and Genetics, 1*(4), 342-362. doi:10.1002/prot.340010408

Fiser, A., Do, R. K. G., & Sali, A. (2000). Modeling of loops in protein structures. *Protein Science, 9*(9), 1753-1773. Retrieved from <Go to ISI>://WOS:000089615200014

Go¯, N., & Scheraga, H. A. (1970). Ring Closure and Local Conformational Deformations of Chain Molecules. *Macromolecules, 3*(2), 178-187. doi:10.1021/ma60014a012

Gresh, N., Cisneros, G. A., Darden, T. A., & Piquemal, J. P. (2007). Anisotropic, polarizable molecular mechanics studies of inter- and intramoecular interactions and ligand-macromolecule complexes. A bottom-up strategy. *Journal of Chemical Theory and Computation, 3*(6), 1960-1986. doi:10.1021/ct700134r

Grossfield, A., Ren, P. Y., & Ponder, J. W. (2003). Ion solvation thermodynamics from simulation with a polarizable force field. *Journal of the American Chemical Society, 125*(50), 15671-15682. doi:10.1021/ja037005r

Halgren, T. A. (1992). The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *Journal of the American Chemical Society, 114*(20), 7827-7843. doi:10.1021/ja00046a032

Hart, R. K., Pappu, R. V., & Ponder, J. W. (2000). Exploring the similarities between potential smoothing and simulated annealing. *Journal of Computational Chemistry, 21*(7), 531-552. doi:10.1002/(sici)1096-987x(200005)21:7<531::aid-jcc3>3.0.co;2-c

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., & Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics, 65*(3), 712-725. doi:10.1002/prot.21123

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., & Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics, 55*(2), 351-367. doi:10.1002/prot.10613

Jorgensen, William L. (2013). Foundations of Biomolecular Modeling. *Cell, 155*(6), 1199-1202. doi:http://dx.doi.org/10.1016/j.cell.2013.11.023

Kaminsky, A. (2007). *Parallel Java: A unified API for shared memory and cluster parallel programming in 100% Java.* Paper presented at the Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International.

Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research, 37*, D387-D392. doi:10.1093/nar/gkn750

Ko, J., Lee, D., Park, H., Coutsias, E. A., Lee, J., & Seok, C. (2011). The FALC-Loop web server for protein loop modeling. *Nucleic Acids Research, 39*(Web Server issue), W210-W214. doi:10.1093/nar/gkr352

Kong, X. J., & Brooks, C. L. (1996). lambda-Dynamics: A new approach to free energy calculations. *Journal of Chemical Physics, 105*(6), 2414-2423. doi:10.1063/1.472109

Laio, A., & Parrinello, M. (2002). Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America, 99*(20), 12562-12566. doi:10.1073/pnas.202427399

Lee, J., Lee, D., Park, H., Coutsias, E. A., & Seok, C. (2010). Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins: Structure, Function, and Bioinformatics, 78*(16), 3428-3436. doi:10.1002/prot.22849

Li, Z. Q., & Scheraga, H. A. (1987). MONTE-CARLO-MINIMIZATION APPROACH TO THE MULTIPLE-MINIMA PROBLEM IN PROTEIN FOLDING. *Proceedings of the National Academy of Sciences of the United States of America, 84*(19), 6611-6615. doi:10.1073/pnas.84.19.6611

Lopes, P. E. M., Huang, J., Shim, J., Luo, Y., Li, H., Roux, B., & MacKerell, A. D. (2013). Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *Journal of Chemical Theory and Computation, 9*(12), 5430-5449. doi:10.1021/ct400781b

Lopes, P. E. M., Roux, B., & MacKerell, A. D. (2009). Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications. *Theoretical Chemistry Accounts, 124*(1-2), 11-28. doi:10.1007/s00214-009-0617-x

LuCore, S. D., Litman, J. M., Powers, K. T., Gao, S. B., Lynn, A. M., Tollefson, W. T. A., . . . Schnieders, M. J. (2015). Dead-End Elimination with a Polarizable Force Field Repacks PCNA Structures. *Biophysical Journal, 109*(4), 816-826. doi:10.1016/j.bpj.2015.06.062

MacKerell, A. D., Feig, M., & Brooks, C. L. (2004). Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society, 126*(3), 698-699. doi:10.1021/ja036959e

Mandell, D. J., Coutsias, E. A., & Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods, 6*(8), 551-552. doi:10.1038/nmeth0809-551

Martin, A. C., Cheetham, J. C., & Rees, A. R. (1989). Modeling antibody hypervariable loops: a combined algorithm. *Proceedings of the National Academy of Sciences, 86*(23), 9268-9272. Retrieved from http://www.pnas.org/content/86/23/9268.abstract

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). EQUATION OF STATE CALCULATIONS BY FAST COMPUTING MACHINES. *Journal of Chemical Physics, 21*(6), 1087-1092. doi:10.1063/1.1699114

Mol, C. D., Lim, K. B., Sridhar, V., Zou, H., Chien, E. Y. T., Sang, B. C., . . . McRee, D. E. (2003). Structure of a c-Kit product complex reveals the basis for kinase transactivation. *Journal of Biological Chemistry, 278*(34), 31461-31464. doi:10.1074/jbc.C300186200

Murshudov, G. N., Vagin, A. A., & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D-Biological Crystallography, 53*, 240-255. doi:10.1107/s0907444996012255

Park, J., Nessler, I., McClain, B., Macikenas, D., Baltrusaitis, J., & Schnieders, M. J. (2014). Absolute Organic Crystal Thermodynamics: Growth of the Asymmetric Unit into a Crystal via Alchemy. *Journal of Chemical Theory and Computation, 10*(7), 2781-2791. doi:10.1021/ct500180m

Patel, S., & Brooks, C. L. (2006). Fluctuating charge force fields: Recent developments and applications from small molecules to macromolecular biological systems. *Molecular Simulation, 32*(3-4), 231-249. doi:10.1080/08927020600726708

Penkert, R. R., DiVittorio, H. M., & Prehoda, K. E. (2004). Internal recognition through PDZ domain plasticity in the Par-6-Pals1 complex. *Nature Structural & Molecular Biology, 11*(11), 1122-1127. doi:10.1038/nsmb839

Piela, L., Kostrowicki, J., & Scheraga, H. A. (1989). THE MULTIPLE-MINIMA PROBLEM IN THE CONFORMATIONAL-ANALYSIS OF MOLECULES - DEFORMATION OF THE POTENTIAL-ENERGY HYPERSURFACE BY THE DIFFUSION EQUATION METHOD. *Journal of Physical Chemistry, 93*(8), 3339-3346. doi:10.1021/j100345a090

Pieper, U., Webb, B. M., Dong, G. Q., Schneidman-Duhovny, D., Fan, H., Kim, S. J., . . . Sali, A. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research, 42*(D1), D336-D346. doi:10.1093/nar/gkt1144

Ponder, J. W., & Case, D. A. (2003). Force fields for protein simulations. *Protein Simulations, 66*, 27-+. Retrieved from <Go to ISI>://WOS:000187012900002

Ponder, J. W., & Richards, F. M. (1987). AN EFFICIENT NEWTON-LIKE METHOD FOR MOLECULAR MECHANICS ENERGY MINIMIZATION OF LARGE MOLECULES. *Journal of Computational Chemistry, 8*(7), 1016-1024. doi:10.1002/jcc.540080710

Read, R. J. (1986). IMPROVED FOURIER COEFFICIENTS FOR MAPS USING PHASES FROM PARTIAL STRUCTURES WITH ERRORS. *Acta Crystallographica Section A, 42*, 140-149. doi:10.1107/s0108767386099622

Ren, P. Y., & Ponder, J. W. (2002). Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *Journal of Computational Chemistry, 23*(16), 1497-1506. doi:10.1002/jcc.10127

Ren, P. Y., & Ponder, J. W. (2003). Polarizable atomic multipole water model for molecular mechanics simulation. *Journal of Physical Chemistry B, 107*(24), 5933-5947. doi:10.1021/jp027815+

Ren, P. Y., Wu, C. J., & Ponder, J. W. (2011). Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *Journal of Chemical Theory and Computation, 7*(10), 3143-3161. doi:10.1021/ct200304d

Rohl, C. A., Strauss, C. E. M., Chivian, D., & Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins-Structure Function and Bioinformatics, 55*(3), 656-677. doi:10.1002/prot10629

Schnieders, M. J., Baker, N. A., Ren, P. Y., & Ponder, J. W. (2007). Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum. *Journal of Chemical Physics, 126*(12). doi:10.1063/1.2714528

Schnieders, M. J., Baltrusaitis, J., Shi, Y., Chattree, G., Zheng, L., Yang, W., & Ren, P. (2012). The Structure, Thermodynamics, and Solubility of Organic Crystals from Simulation with a Polarizable Force Field. *Journal of Chemical Theory and Computation, 8*(5), 1721-1736. doi:10.1021/ct300035u

Schnieders, M. J., Fenn, T. D., & Pande, V. S. (2011). Polarizable Atomic Multipole X-Ray Refinement: Particle Mesh Ewald Electrostatics for Macromolecular Crystals. *Journal of Chemical Theory and Computation, 7*(4), 1141-1156. doi:10.1021/ct100506d

Schnieders, M. J., & Ponder, J. W. (2007). Polarizable atomic multipole solutes in a generalized Kirkwood continuum. *Journal of Chemical Theory and Computation, 3*(6), 2083-2097. doi:10.1021/ct7001336

Shi, Y., Xia, Z., Zhang, J. J., Best, R., Wu, C. J., Ponder, J. W., & Ren, P. Y. (2013). Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of Chemical Theory and Computation, 9*(9), 4046-4063. doi:10.1021/ct4003702

Slesinger, P. A., Jan, Y. N., & Jan, L. Y. (1993). THE S4-S5 LOOP CONTRIBUTES TO THE ION-SELECTIVE PORE OF POTASSIUM CHANNELS. *Neuron, 11*(4), 739-749. doi:10.1016/0896-6273(93)90083-4

Soto, C. S., Fasnacht, M., Zhu, J., Forrest, L., & Honig, B. (2008). Loop modeling: Sampling, filtering, and scoring. *Proteins-Structure Function and Bioinformatics, 70*(3), 834-843. doi:10.1002/prot.21612

Spassov, V. Z., Flook, P. K., & Yan, L. (2008). LOOPER: a molecular mechanics-based algorithm for protein loop prediction. *Protein Engineering Design & Selection, 21*(2), 91-100. doi:10.1093/protein/gzm083

Steichen, J. M., Kuchinskas, M., Keshwani, M. M., Yang, J., Adams, J. A., & Taylor, S. S. (2012). Structural Basis for the Regulation of Protein Kinase A by Activation Loop Phosphorylation. *Journal of Biological Chemistry, 287*(18), 14672-14680. doi:10.1074/jbc.M111.335091

Stuart, D. I., Acharya, K. R., Walker, N. P. C., Smith, S. G., Lewis, M., & Phillips, D. C. (1986). ALPHA-LACTALBUMIN POSSESSES A NOVEL CALCIUM-BINDING LOOP. *Nature, 324*(6092), 84-87. doi:10.1038/324084a0

Tasneem, A., Iyer, L. M., Jakobsson, E., & Aravind, L. (2005). Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. *Genome Biology, 6*(1). Retrieved from <Go to ISI>://WOS:000226337200010

Trabuco, L. G., Villa, E., Mitra, K., Frank, J., & Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure, 16*(5), 673-683. doi:10.1016/j.str.2008.03.005

Warshel, A., & Levitt, M. (1976). THEORETICAL STUDIES OF ENZYMIC REACTIONS - DIELECTRIC, ELECTROSTATIC AND STERIC STABILIZATION OF CARBONIUM-ION IN REACTION OF LYSOZYME. *Journal of Molecular Biology, 103*(2), 227-249. doi:10.1016/0022-2836(76)90311-9

Westheimer, F. H., & Mayer, J. E. (1946). The Theory of the Racemization of Optically Active Derivatives of Diphenyl. *The Journal of Chemical Physics, 14*(12), 733-738. doi:doi:http://dx.doi.org/10.1063/1.1724095

Wu, J. C., Chattree, G., & Ren, P. Y. (2012). Automation of AMOEBA polarizable force field parameterization for small molecules. *Theoretical Chemistry Accounts, 131*(3). doi:10.1007/s00214-012-1138-6

Wu, M. G., & Deem, M. W. (1999). Efficient Monte Carlo methods for cyclic peptides. *Molecular Physics, 97*(4), 559-580. Retrieved from <Go to ISI>://WOS:000082358500009

Xiang, Z. X., Soto, C. S., & Honig, B. (2002). Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Sciences of the United States of America, 99*(11), 7432-7437. doi:10.1073/pnas.102179699

Yarov-Yarovoy, V., Baker, D., & Catterall, W. A. (2006). Voltage sensor conformations in the open and closed states in ROSETTA structural models of K+ channels. *Proceedings of the National Academy of Sciences of the United States of America, 103*(19), 7292-7297. doi:10.1073/pnas.0602350103

Yu, H., Whitfield, T. W., Harder, E., Lamoureux, G., Vorobyov, I., Anisimov, V. M., . . . Roux, B. (2010). Simulating Monovalent and Divalent Ions in Aqueous Solution Using a Drude Polarizable Force Field. *Journal of Chemical Theory and Computation, 6*(3), 774-786. doi:10.1021/ct900576a

Zheng, L., Chen, M., & Yang, W. (2008). Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. *Proceedings of the National Academy of Sciences of the United States of America, 105*(51), 20227-20232. doi:10.1073/pnas.0810631106

Zhu, K., Pincus, D. L., Zhao, S. W., & Friesner, R. A. (2006). Long loop prediction using the protein local optimization program. *Proteins-Structure Function and Bioinformatics, 65*(2), 438-452. doi:10.1002/prot.21040