DEVELOPMENT OF NOVEL COMPUTATIONAL TECHNIQUES FOR THE STUDY OF BIOMOLECULAR SYSTEMS USING MOLECULAR DYNAMICS SIMULATION

By

Vahid Mirjalili

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Mechanical Engineering - Doctor of Philosophy

2014

ABSTRACT

DEVELOPMENT OF NOVEL COMPUTATIONAL TECHNIQUES FOR THE STUDY OF BIOMOLECULAR SYSTEMS USING MOLECULAR DYNAMICS SIMULATION

By

Vahid Mirjalili

In this dissertation, we have developed novel computational techniques that have been effectively utilized to extend our knowledge of proteins and lipid membrane systems. Application of molecular dynamics combined with newly developed techniques and protocols to study protein structure refinement and interactions of amino-acid analog pairs within lipid membranes are studied. A robust protocol for structure refinement of proteins from a given homologous model is designed and optimized that uses restrained molecular dynamics followed by optimal subset selection and structure averaging. This protocol is tested on CASP8 and CASP9 targets, and later successfully applied to CASP10 in blind prediction manner.

In order to understand physical characteristics of peptide interactions embedded in bilayer membrane, we have used umbrella sampling technique with model amino acid side-chain analog pairs to study their association free energy while placed in membrane bilayer. As a result of convergence issues observed in such simulations due to bilayer deformation, a novel enhanced sampling technique is developed which biases the density of water in a cylinder, thereby effectively imposing bilayer deformation. Applying this method to a DPPC bilayer, we were able to study free energy of pore formation in membrane bilayers, and showed that while the undergone mechanism is different from currently existing methods, the mechanism by our proposed method is closer to the natural pore formation mechanism.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Background

The use of computational techniques for study of biological systems is rapidly increasing. Molecular dynamics (MD) simulation is among the most influential computational methods that has given insights into physical behavior of complex biological systems. The increasing use of MD simulation in biological science is indicated by the rapid increase of number of research article published in recent years. In 2014, more than 35,000 research articles are published in scientific journals that have used MD in biological science (Google Scholar, see Figure 1-1).

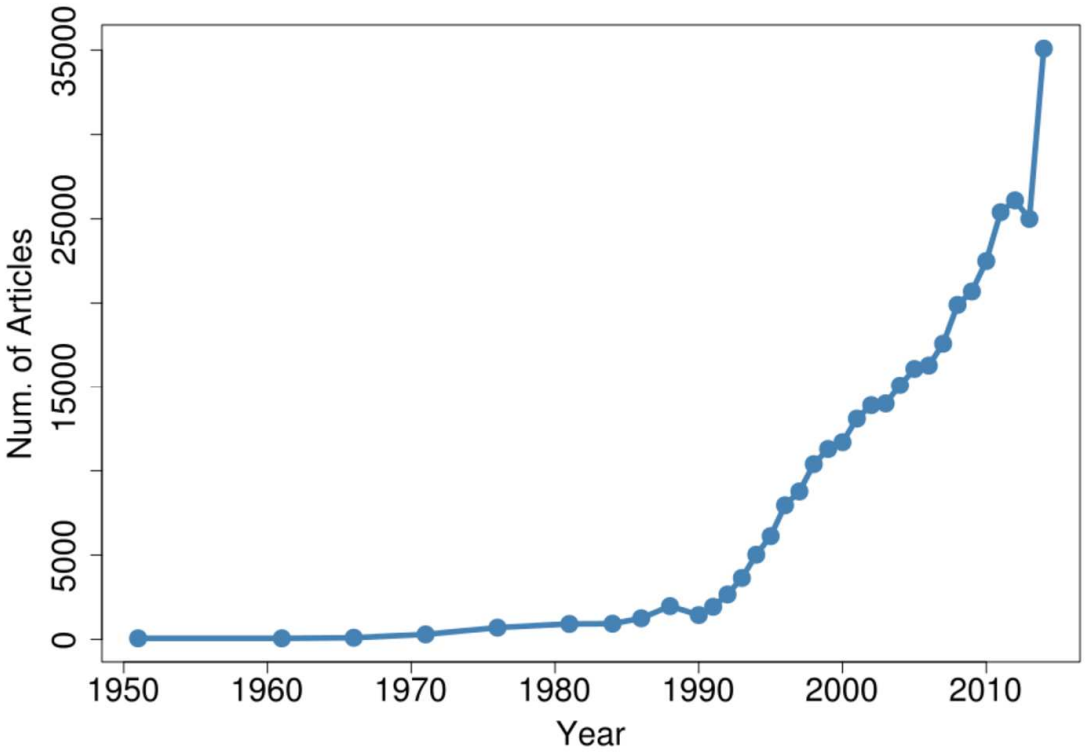

Figure 1-1 Number of research articles that have used molecular dynamics in biological science.

MD simulation can be used to analyze conformational dynamics and the kinetic and thermodynamic properties of proteins, nucleic acids and lipid membranes. While experimental techniques have limitations in extracting fine details of such systems, MD simulations have been

widely applied to analyze properties of these systems.[1-3] In recent years, MD simulations are considered a necessary stage prior/posterior to performing advanced experimental studies, which indicates the significance of results obtained from MD simulations.[4-6] Yet, computer simulations have gone beyond the limitations of experimental research. Scientists have used computer simulations to study how certain enzymes react to antibiotics.[7] The molecular-level insight obtained from computer simulations can be effectively used in future medicine to understand how bacteria become resistance to antibiotics.

In this dissertation, we have developed novel computational techniques that have been effectively utilized to extend our knowledge of proteins and lipid membrane systems. This chapter provides an introduction of MD and other computational techniques that are utilized. Then, later chapters focus in more details on the application of such techniques combined with newly developed techniques and protocols to study protein structure refinement and interactions of amino-acid analog pairs within lipid membranes.

1.2 Molecular Dynamics Simulation

Molecular dynamics (MD) simulation solves time evolution of a set of discrete particles by solving the Newton's equation of motion in classical mechanics. A molecular force-field defines the level of interactions among particles and their environment. The notion of discrete particles determines the resolution at which the physical system is described. With advancement of modern computers, study of biological systems in atomistic details has been made feasible.[8] Given a set of atomic coordinates ($\vec{r}_i$), atomic forces are determined from the derivatives of a

given set of potential energy functions, which is based on pairwise interactions of particles. From these forces, acceleration of each atom can be computed according to equation 1-1

$$F_i = -\nabla_i U = m_i a_i \qquad 1\text{-}1$$

where m_i is the mass of atom *i* and a_i is its acceleration. By numerical integration over equation 1-1, atomic velocities and their new coordinates can be determined.

1.3 Molecular Force-Field

A molecular force-field is a set parameters used in different potential energy terms. There are a number of force-fields for biological systems, and understanding their differences and application is essential. CHARMM[9-11], AMBER[12], GROMOS[13], and OPLS[14, 15] are among the most widely used force-fields for biological systems, however, a number of force-fields also exist at different resolutions.[16-18] While the set of potential energy functions used in different force-fields are different, they are generally categorized as bonded interactions and non-bonded interactions. We used CHARMM force-field[9] throughout this dissertation. The latest version of CHARMM force field[9] has the following energy terms as given in equation 1-2

$$V(\vec{R}) = \sum_{bonds} K_b (b - b_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi (1 + \cos(n\chi - \delta)) + \sum_{improper} K_\omega (\omega - \omega_0)^2 + \sum_{Urey-Bradley} k_{UB} (b^{1-3} - b^{1-3,0})^2 + \sum_{residues} u_{CMAP}(\Phi, \Psi) \qquad 1\text{-}2$$

where K_b, K_θ, K_χ, K_ω, and K_{UB} are the force constants for bonds, valence angles, dihedral angles, improper angles and Urey-Bradley term, respectively. The CMAP[19] term is a two dimensional spline-based energy function that was introduced to improve backbone treatments of proteins in MD simulations.

The non-bonded term contains the Lennard-Jones (LJ) and Coulomb terms in the following form as given in equation 1-3

$$V_{non-bonded} = \sum_{nonb.\,pairs} \varepsilon_{ij} \left[\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{6} \right] + \sum_{nonb.\,pairs} \frac{q_i q_j}{\varepsilon r_{ij}} \qquad 1\text{-}3$$

where $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$ and $R_{ij}^{\min} = \frac{R_i^{\min} + R_j^{\min}}{2}$, and (ε_i , $R_i^{\min}$) are atomic LJ parameters. In the Coulomb term, q_i is the charge of atom *i*, and ε is the permittivity of free space. r_{ij} in both terms represent the distance between atoms *i* and *j*.

1.4 Treatment of solvent molecules

Biomolecules are naturally embedded in a solvent environment and they interact directly with water molecules. Therefore, accurate treatment of solvents is necessary to derive physical conformations of such systems. Approaches for treatment of solvent molecules have two categories; one in which solvent molecules are explicitly present in the system, and they interact with the solute through non-bonded interactions. While this approach stands as the most accurate representation of a biological system, one of the main computational bottlenecks in inclusion of solvent molecules in the non-bonded calculations is that the cost of non-bonded calculations can be dominated by solvent-solvent interactions. As a result, a second approach has been developed that treats the solvents as a continuum environment, and utilizes the net effect of solvent molecules on the solute.

1.4.1 Explicit treatment of solvent molecules

In explicit solvent MD simulations, solvent molecules are represented with a water model. Ref [20] lists 46 distinct water models. In this dissertation, TIP3P[21] is used in all explicit solvent

simulations. TIP3P contains 3 particles that represent one oxygen atom and two hydrogens for a water molecule. Although, there are some models that have a higher number of particles per water molecule, however, they make the non-bonded computations more costly.

1.4.2 Implicit treatment of solvent environment

Explicit MD simulation of biomolecules uses the most detailed information of solvent atoms, which makes it costly. An alternative way is to remove solvent molecules, and only include the solvent degrees of freedom, and instead estimate the net effect of solvent environment on the solute atoms.[22] To do that, free energy cost of solvating the solute should be calculated, which has three components, i.e., electrostatic, non-polar, and cost of cavity formation, as given in equation 1-4

$$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{non-polar} \qquad 1\text{-}4$$

Figure 1-2 shows the schematic diagram of decomposing solvation free energy into its electrostatic and non-polar components in a thermodynamic cycle. In this diagram, the solute molecule is transferred from vacuum (white area) to solvent environment (gray area) in two different thermodynamic paths. Since free energy change is a path-independent thermodynamic property, the change in free energy from direct insertion (the first path) and step-by-step insertion (the second path) should be equivalent to each other. In step-by-step insertion path, first the atomic charges in the solute are turned off ($-\Delta G_{elec}^{\varepsilon=1}$), then the uncharged solute is inserted from vacuum to solvent environment, resulting in non-polar solute-solvent and solvent-solvent interactions ($\Delta G_{non-polar}$). Finally, the charges in the solute are turned on ($\Delta G_{elec}^{\varepsilon=80}$).

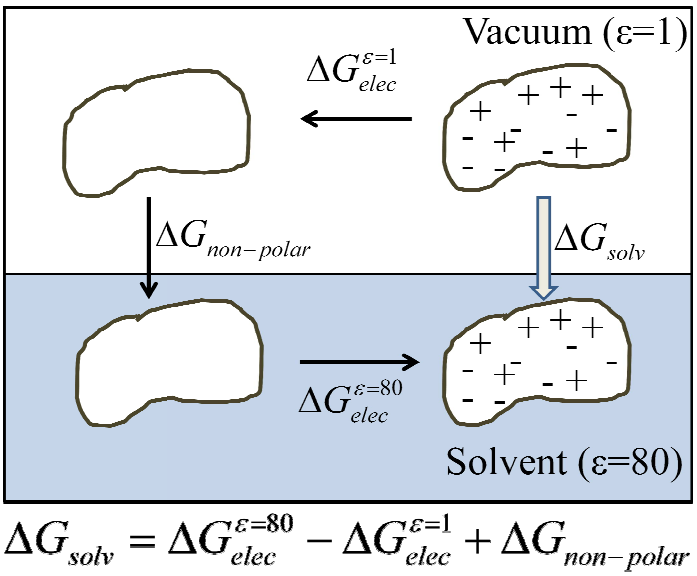

Figure 1-2 Schematic diagram of decomposing solvation free energy in a thermodynamic cycle.

The electrostatic component of solvation free energy is calculated by Generalized Born (GB) theory, using the formulation proposed by Still et al.[23] as follows

$$\Delta G_{elec} = -k\left(\frac{1}{\varepsilon_{solute}} - \frac{1}{\varepsilon_{solvent}}\right)\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i\alpha_j \exp(-r_{ij}^2 / 4\alpha_i\alpha_j)}} \qquad 1\text{-}5$$

where ε_{solute} , $\varepsilon_{solvent}$ are the dielectric constants of the solute and solvent environments, respectively. α_i represents the Born radius of atom *i*, which is average distance of atom *i* to the solvent-exposed surface of the solute. The more buried atom *i* is, the larger Born radius it has. The main computational cost in the GB term is spent on Born radii calculation. Several methods have been proposed that estimate Born radii in different ways[24-27] that differ in computational complexity, and accuracy. In one category of such methods, a numerical integration has to be performed over atomic molecular volume to obtain Born radii.[25, 26] The latest GBMV (GB Molecular Volume) formulation[26] is shown to have very good accuracy compared to Poisson-

Boltzmann (PB) theory in computing the self-polarization energies for solute atoms.[26] Born radii in GBMV are calculated as follows

$$\alpha_i = \frac{1}{(1-\sqrt{2}/2)I_4 + I_7}$$

$$I_4 = \frac{1}{R_i} - \frac{1}{4\pi}\int_{solute, r>R_i} \frac{1}{r^4} dV \qquad \text{1-6}$$

$$I_7 = \left(\frac{1}{4R_i^4} - \frac{1}{4\pi}\int_{solute, r>R_i} \frac{1}{r^7} dV \right)^{1/4}$$

where r is distance of grid points to atom i, and I_4 and I_7 are integration terms of r to atom i to the 4^{th} and 7^{th} power, respectively.

The non-polar component of solvation free energy in equation 1-4, accounts for cost of cavity formation in solvent, as well as van der Waals (vdW) interactions between solute and solvent. In most GB implementations, this term is approximated by solvent-accessible surface area (SASA)

$$\Delta G_{non-polar} = \sum_i \gamma_i \, . \, SASA_i \qquad \text{1-7}$$

where the scaling factor γ represents the surface tension coefficient.

1.4.3 Implicit membrane model

In an implicit membrane environment, the dielectric constant is not homogenously uniform. A schematic diagram of variation of the dielectric constant across membrane is shown in Figure 1-3.

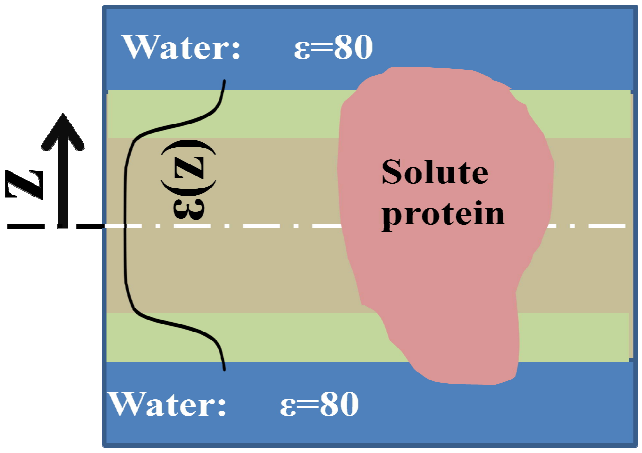

Figure 1-3 Schematic diagram of variation of dielectric constant across membrane bilayer.

Therefore, a special treatment is needed to account for effects of heterogeneous dielectric environment on atomic Born radii and solvation free energy. Heterogeneous Dielectric Generalized Born (HDGB)[28, 29] is an extension of GBMV for implicit membrane environments, which models dielectric constant across bilayer normal as a function of distance from bilayer center, Z. Further, Born radii calculation is also computed as a function of dielectric constant, and the HDGB energy formulation is given according to equation 1-8

$$\Delta G_{elec,HDGB} = -k\left(\frac{1}{\varepsilon_{solute}} - \frac{1}{\varepsilon_{solvent}(\varepsilon_i,\varepsilon_j)}\right)\times \sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i(\varepsilon_i)\alpha_j(\varepsilon_j)\exp(-r_{ij}^2/4\alpha_i(\varepsilon_i)\alpha_j(\varepsilon_j))}} \quad 1\text{-}8$$

and the equation for Born radius is modified as follows

$$\alpha_i(\varepsilon_i,\varepsilon_p) = \frac{1}{C_0 A_4 + C_1\left(\frac{3\varepsilon_w}{3\varepsilon_w + 2\varepsilon_p}\right)A_7} + D + \frac{E}{\varepsilon_w + 1} \quad 1\text{-}9$$

The non-polar contribution is also modified by scaling the non-polar term by a continuous profile S(z). The shape of ε(z) and S(z) profiles as adopted from Sayadi and Feig[30], as shown in Figure 1-4

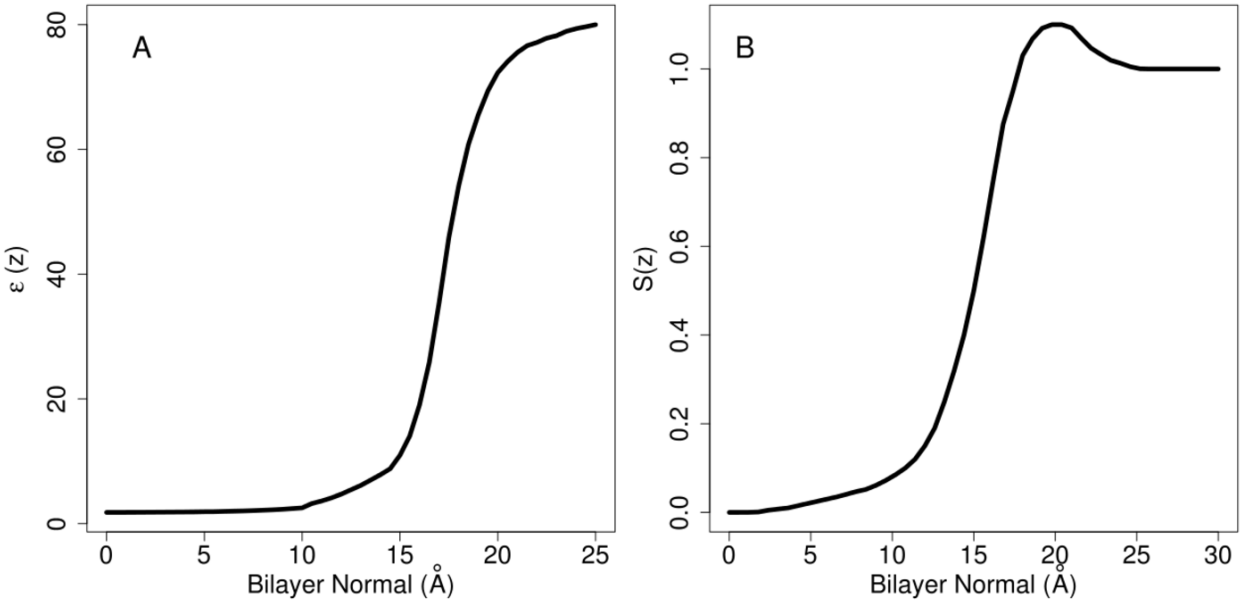

Figure 1-4 The dielectric profile (A), and non-polar scaling profile (B) for HDGB model

1.5 Free Energy Calculation

In order to estimate free energy landscape of a thermodynamic reaction along a reaction coordinate using molecular simulation, one needs to sufficiently sample configurational space to reach equilibrium conditions between two thermodynamic states. In molecular simulation of liquid environment, Gibbs free energy can be stated as follows

$$\Delta G = \Delta U + p\Delta V - T\Delta S = \Delta H - T\Delta S \qquad 1\text{-}10$$

where ΔU, ΔV, and ΔS are the change in internal energy, volume and entropy of the system, respectively, and p and T are pressure and temperature. Entropy of a thermodynamic system is directly related to the total number of configurations accessible to it.[31] As the system size gets

larger, the number of accessible configurations becomes larger; as a result, sampling all configurations with traditional MD simulations to estimate free energy becomes infeasible. Therefore, enhanced sampling methods for free energy calculations have been developed, which can be classified into three general categories; modified potential, modified sampling, and modified dynamics.[3]

In this dissertation, umbrella sampling[32] is used for free energy calculation, which belongs to the first category. In umbrella sampling, the potential function is modified to bias sampling along a specific thermodynamic direction, also known as the reaction coordinate (ξ). The choice of reaction coordinate is critical for this method. The reaction coordinate should drive the system from one desired state to another. The reaction coordinate is divided into equal bins, and the system is simulated with additional potential function for each bin

$$U_{umb}(\xi) = \frac{1}{2} K_{umb} \left(\xi - \xi_i\right)^2 \qquad 1\text{-}11$$

where K_{umb} is the force constant. Multiple biased systems are run with different equilibrium value of reaction coordinate for the i^{th} bin (ξ_i). Free energy along the reaction coordinate can be recovered from the ensemble of biased sampling using weighted histogram analysis method (WHAM)[33]. Energy of the unbiased system can be written as

$$E_{unbiased}(R^N) = E_{biased}(R^N) - U_{umb}(\xi) \qquad 1\text{-}12$$

Given energy of a particular configuration, the probability of finding the system having reaction coordinate ξ is

$$p(\xi) = \frac{\int \delta(\xi - \xi_0) \exp(-\beta E_{unbiased}(R^N)) dR^N}{\int \exp(-\beta E_{unbiased}(R^N)) dR^N} \qquad 1\text{-}13$$

where $\beta = 1/k_B T$. substituting for unbiased energy, the probability distribution can be recovered from biased sampling as follows

$$p(\xi) = C \exp\left(\beta U_{umb}(\xi)\right) p_{biased}(\xi) \qquad 1\text{-}14$$

where constant C is added due to integration. Then, potential of mean force (PMF) can be obtained from the probability distribution of reaction coordinate given above

$$\begin{aligned} W(\xi) = -k_B T \ln p(\xi) &= C - U_{umb}(\xi) - k_B T \ln p_{biased}(\xi) \\ &= C - U_{umb}(\xi) + W_{biased}(\xi) \end{aligned} \qquad 1\text{-}15$$

1.6 Dissertation Scope

The theoretical backgrounds described in previous sections provide a valuable tool to study biological systems. In this dissertation, MD simulations have been used for two main tasks, protein structure refinement and characterizing peptide-membrane interactions within membrane environment. In the next section, a brief description of tasks is provided.

1.6.1 Protein Structure Refinement

Refinement of protein structures using computational methods has remained a challenging task. Given a low resolution 3-dimensional model of a protein from homology modeling, the goal is to further refine the structure towards a high-resolution native-like model.[34] Successful refinement of protein targets using computational techniques can have a huge impact on future biological and pharmaceutical research. Critical Assessment of techniques in protein Structure Prediction (CASP) is a biennial world-wide competition which provides a benchmark for researchers to effectively test their method on new protein targets. The refinement category of CASP, called CASPR, selects the best homology models from tertiary structure prediction

groups as input for refinement. Over the past rounds of CASP, refinement has remained a difficult task, with most participating groups showing little to no improvement.[35, 36]

The challenge in protein structure refinement is two folds; sampling conformational space of a given protein model, and model scoring and selection from a bag of sampled conformations.[37] Divers approaches involving a combination of physics-based[37-40] and knowledge-based[41-43] approaches have been employed to address these challenges using Monte Carlo (MC) or MD simulations. For model selection, a number of scoring functions have also been proposed[44, 45], such as DFIRE[45, 46], Seder[47], and RW-plus[48]. However, structure selection still remains a difficult task since scoring functions are not accurate enough for discriminating models at close resolutions. Therefore, efforts in improving protein models would make the model qualities worse on average, even with long MD simulations of up to 100 μs.[34]

We established a robust protocol, which for the first time showed positive improvement on average over CASP targets.[37, 38] Given an initial model, we used harmonic restraints on selected C_α atoms, and ran explicit solvent MD simulations using CHARMM36 force-field. An ensemble of conformations is obtained from MD simulations, which then for structure selection, we scored the models, and selected a subset based on an optimized criterion. Finally, the average structure from this subset was shown to have consistent improvement from the initial model. The details of our protocol is given in chapter 2, the results of our method applied to CASP10 for blind prediction is provided in chapter 3, and results of CASP11 in chapter 4.

1.6.2 Understanding peptide-membrane interactions

Membrane proteins play critical roles in cellular processes and signaling pathways that are crucial for cell survival. The rapidly increasing number of experimental structures of membrane

proteins has shed light on their biological importance. However, experimental techniques for understanding behavior of peptides and proteins have limitations, especially when dealing with bilayer membranes. As a result, computational approaches provide valuable tools.

In the second part of this dissertation, we characterized the peptide-membrane interactions. Previously, experimental data for free energy of insertion of amino acids into membrane bilayer was characterized.[49] Then, MacCallum et al. obtained the insertion profiles of amino acid side-chain analogs into membrane bilayer through molecular dynamics simulations.[2] The free energy profiles of insertion of amino acid side-chain analogs provided a valuable source for understanding polarities and behavior of side-chain analogs in bilayer environment, as well as a benchmark for comparison and parameterization of computational tools for membrane, such as implicit membrane models. While, the mentioned studies gave very useful biological insights on the interactions of peptides with membrane, they did not consider the interactions of amino acids with each other within membrane environment. Therefore, in the first attempt to address amino acid interactions within membrane, de Jong et al.[1] considered different pairs of amino acid side-chain analogs in three different environments, water, n-octanol, and decane. These environments have different polarities, and decane to some extent represents the hydrophobic region of membrane bilayers. However, the interactions of amino-acid pairs in a real bilayer have not been addressed so far.

Knowing the importance of such interactions for understanding proteins structure and function, we tried to characterize the interactions of amino acid side-chain analog pairs within membrane environment. Considering all possible pairs of amino-acids could be very expensive, therefore, we only selected four amino acids, i.e. Phe, Val, Ser and Asn. These four amino acids resemble a wide range of amino-acid sizes and polarities. A pair of each side-chain analog is placed in

bilayer at different distances from bilayer normal (Z), and their association is studied using umbrella sampling molecular dynamics by pulling them apart at a fixed Z. Due to the polarity of some of the side-chain analog pairs placed in bilayer interior, water defect and membrane deformation was observed for some cases. While these phenomena are described in later chapters in detail, however, convergence issues are raised if neighboring windows are not sampling the same flat/deformed bilayer states.

As a result of bilayer deformation in neighboring umbrellas, it is necessary to address the effect of amino acid side-chain analog pairs on bilayer deformation. Therefore, in order to study this physical process, we developed a new biasing potential that can effectively deform bilayer. The number of water molecules in a cylinder along the bilayer normal axis is computed in a continuous, rather than discrete fashion, using a smooth switching function. Then, number density of water molecules is computed, and used as a new reaction coordinate in umbrella sampling. Full description of this biasing potential and its applications in studying density driven processes are provided in chapter 5.

Chapter 2

Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles

Vahid Mirjalili, Michael Feig

Adapted from

Journal of Chemical Theory and Computation,
V2, p. 1294-1303, 2012

2.1 Abstract

A molecular dynamics (MD) simulation based protocol for structure refinement of template-based model predictions is described. The protocol involves the application of restraints, ensemble averaging of selected subsets, interpolation between initial and refined structures, and assessment of refinement success. It is found that sub-microsecond MD-based sampling when combined with ensemble averaging can produce moderate but consistent refinement for most systems in the CASP targets considered here.

2.2 Introduction

Much progress has been made towards predicting the tertiary structure of proteins from their amino-acid sequence.[50-52] By far the most success has been found with template-based modeling (TBM) methods[53-55] where information from known experimental structures is utilized. Traditionally, TBM would use a single homologous protein for which a structure is available, but the best methods combine structural information from multiple templates in a variety of different algorithms.[50, 56-60] Using such methods, structures for most soluble proteins can be obtained today with high accuracy as long as sufficiently close structural templates can be found in the Protein Data Bank.[61] Nevertheless, the resulting models for non-trivial cases often retain structural errors with respect to experimental structures that limit the use of such models in further studies. For example, TBM-derived structures are often problematic as drug design targets[62, 63] or as starting structures for detailed mechanistic studies via molecular dynamics simulations and other computational methods.[64]

Structure refinement methods aim at the further improvement of TBM-based models towards experimental accuracy.[35, 36, 65] Because TBM-based models already utilize knowledge from

related structures, most refinement algorithms that have been proposed rely on physics-based techniques, in particular molecular dynamics (MD) simulations.[65-68] Although successful examples of MD-based refinement have been reported in the past,[40, 51, 60, 66-72] consistent success appears to be hindered by a combination of insufficient sampling,[60, 73, 74] force field inaccuracies,[67, 75] and an inability to reliably identify refined structures that may be generated during the course of an MD simulation.[60, 70, 75-78] To address these issues, statistical potentials[41, 68, 79, 80] and optimized force fields[67, 81, 82] have been used as well as effective sampling techniques such as replica-exchange[40, 41, 66, 71] and self-guided Langevin dynamics[83] simulations. In some studies it was possible to generate improved structures by as much as 0.5 Å in root-mean-square deviation (RMSD) in one out of five models,[40, 41] but reliable identification of a single refined structure remained difficult. Recently, Fan et al.[71] have shown that by mimicking the electrostatic effects with chaperone Hamiltonian replica-exchange MD simulation can generate refined structures for 10 out of 15 targets with improvements of more than 1 Å RMSD for the secondary structure elements, but again reliable selection of refined structures without knowledge of the native state remained challenging. However, on average models selected based on a statistical potential function, Distance-scaled Finite Ideal gas REference (DFIRE),[44, 84] could be improved by 0.25 Å from the initial models.[71]

A common observation is that unrestrained MD simulations of template-based models almost invariably end up drifting away from the native structure.[66, 70] Refinement is more likely to occur when structures are restrained,[66, 70] but the drawback of using restraints is that the degree to which structures can be refined is limited. The most extensive test of MD-based refinement published so far involved simulations up to 100 μs for CASP8 (Critical Assessment

of techniques for protein Structure Prediction) and CASP9 refinement targets.[70] In that work from the Shaw group, the final structures were not improved on average but refinement could be achieved by using a cluster-based selection method to reach 1% in terms of GDT-TS (Global Distance Test-Total Score)[85] for conformations extracted from simulations exceeding 10 μs in length. Better structures with sometimes much more significantly improved GDT-TS scores were generated in these simulations but could not be identified reliably.[70]

Finally, Zhang et al.[81] used a fragment-guided MD technique, in which different fragments of target proteins were restrained to their homologous templates. Using this technique, improvements in GDT-HA (GDT-High Accuracy) scores were possible for targets with initial GDT-HA scores of greater than 50. However, for CASP8 and CASP9 targets average improvement was limited to only 0.6% in terms of GDT-HA and the improvement in RMSD was insignificant.

Here, we are presenting a structure refinement protocol that combines MD-based sampling in explicit solvent using the latest CHARMM (Chemistry at HARvard Molecular Mechanics) force field[9], a scoring protocol that identifies the most native-like structures, and ensemble averaging to mimic the conditions under which experimental structures are obtained. Using this protocol, we are able to consistently refine CASP8 and CASP9 targets with relatively modest computational resources.

In the following, the computational methods are described before results are presented and discussed.

2.3 Methods

We have performed all-atom molecular dynamics (MD) simulations for 26 refinement targets from CASP8 and CASP9. The targets used here as test sets are listed in Table 2-1. The initial structures were provided by the CASP organizers and represent predicted models of high accuracy for the respective targets that were submitted during CASP. Along with the initial coordinates, the CASP organizers also provided information for many targets about regions that refinement should focus on. This information was used here to apply restraints on the remaining parts of the structure considered to be accurate. For targets where a refinement residue range was not provided during CASP we determined a residue restraint list during the respective CASP rounds when knowledge of the experimental structures was not yet available under the assumption that the core secondary structure elements are likely to be more correct than other parts of the structure. The resulting list of restraints for each target is given in Table 2-1. For 16 targets the restraint regions were selected based on CASP suggestions, and for the remaining 10 targets restraints were based on core secondary structure elements.

For each initial structure, missing hydrogens were built using the HBUILD module in CHARMM.[86, 87] The protein structures were then solvated in a cubic box of water with a minimum distance of 10 Å between any protein atom and the edge of the box. The systems were neutralized by adding Na^+ or Cl^- as counterions to balance the overall charge. All of the systems were equilibrated by minimization followed by heating through short simulations over 1 ps at 50 K, 100 K, 150 K, 200 K, 250 K, and 298 K. Subsequent production simulations were carried out at 298 K and 1 bar pressure in the NTP (constant number of particles, temperature, and pressure) ensemble over different simulation lengths up to 200 ns.

Table 2-1 CASP8 and CASP9 refinement targets used here as test cases with the total number of residues and Cα-RMSD of the initial models from the respective native structures. Restraint regions denote residues for which harmonic restraints were applied to maintain structures near their initial structures. The targets were sorted according to increasing RMSD values. The regions suggested by CASP are shown in bold.

Target	# of res.	RMSD (Å)	GDT-HA	Restraint regions
TR592	105	1.26	72.9	**17-29;36-46;**58-67**;76-121**
TR453	87	1.47	71.3	**5-34;45-91**
TR432	130	1.65	77.5	**1-84;93-130**
TR462a	75	1.76	57.7	1-5;10-16;21-30;35-42;50-53;57-60;64-75
TR594	140	1.82	67.0	**1-71;82-101;114-140**
TR614	121	1.87	71.5	11-33;53-64;75-109
TR435	137	1.89	67.9	15-19;26-27;38-66;75-87;92-94;98-103;113-133;137-151
TR530	80	1.99	69.1	**36-44;56-74;80-115**
TR488	95	2.11	75.0	**1-11;17-95**
TR469	63	2.18	63.5	3-7;11-28;33-50;54-65
TR462b	68	2.42	48.9	76-83;88-91;97-106;114-124;127-129;133-136;140-143
TR389	135	2.64	63.3	10-15;22-34;49-55;68-73;81-82;100-109;116-126
TR464	69	2.73	59.8	**18-37;44-56;61-86**
TR569	79	3.01	52.2	**1-25;44-49;62-79**
TR454	192	3.24	42.3	5-24;29-34;40-44;50-71;77-107;113-138;147-167;176-196
TR567	142	3.44	58.3	**4-21;28-47;55-59;67-74;90-101;109-145**
TR574	102	3.58	40.0	28-35;49-57;71-73;79-81;85-91;97-106
TR557	125	4.06	46.8	**1-11;21-40;49-52;73-100;107-125**
TR429a	79	4.31	54.8	22-37;44-57;68-80;89-93;98-100
TR517	159	4.64	53.6	**1-62;89-159**
TR606	123	4.85	52.6	**56-144**
TR429b	76	4.98	30.3	101-104;108-111;115-122;128-154;162-176
TR624	69	5.19	35.9	**5-11;16-20;34-51;57-73**
TR568	97	6.15	35.8	**62-77;91-94;107-108;124-158**
TR622	122	6.47	51.9	**1-96**
TR576	138	6.85	45.3	**25-56;66-119**

The CHARMM36 force field[9] was used in combination with the TIP3 water model[88]. The CHARMM36 force field was recently introduced as an improved version of the previous CHARMM22/CMAP force field[89, 90]. The main differences are improved sampling of

backbone propensities in better agreement with experimental data, in particular NMR J-coupling data, and improved side chain torsions, also to improve agreement with experimental data.[9] In all simulations, periodic boundaries were applied and particle-mesh Ewald summation was used to calculate electrostatic interactions using a grid spacing of 1 Å. Direct-space electrostatic and Lennard-Jones interactions were truncated using a switching function between 8.5 Å and 10 Å. All simulations used holonomic constraints on bonds involving hydrogens so that a 2 fs integration time step could be used. Simulations were carried out with and without restraints according to Table 2-1. Restraints were applied through a harmonic force on C_α atoms with a force constant of 1 $\mathrm{kcal/mol/Å^2}$.

Because part of our refinement protocol involves averaging over structural ensembles, a second set of simulations was carried out to allow side chains in the averaged structures to relax while maintaining the backbone geometries. This was accomplished by resolvation of the refined structures followed by minimization over 5000 steps and two short MD simulations at 10 K and 100 K, each for 40 ps. During these minimization and MD simulations, all C_α atoms were restrained with a force constant of 100 $\mathrm{kcal/mol/Å^2}$. The quality of the structures before and after the final refinement simulations was assessed using the MolProbity structure validation web service[91].

All of the systems were initially setup using CHARMM[86, 87] and the MMTSB (Multiscale Modeling Tools for Structural Biology) Tool Set[92]. Production simulations were carried out using NAMD[93]. Analysis was carried out using a combination of CHARMM, the MMTSB Tool Set, and custom scripts and programs.

2.4 Results

Molecular dynamics simulations were carried out for the CASP8 and CASP9 refinement targets starting from the template-based models provided during the respective CASP rounds for the CASPR refinement competition. Simulations were run with and without restraints and over different lengths of 24 ns, 200 ns, or eight times 3 ns to compare the effect of different amounts of sampling. The conformations sampled for each target during these simulations were then subjected to different selection and averaging protocols with the goal to obtain refined structures. Each protocol and the corresponding results are described in more detail in the following.

2.4.1 Final and Best Structures

The most straightforward MD-based refinement protocol would consist of simply considering the final structure at the end of a given MD run. Tables 2-2 and 2-3 show the change in RMSD and GDT-HA, respectively, relative to the native structures for the final structures under different conditions. We show here changes in both RMSD and GDT-HA[94] values because they emphasize different aspects. GDT-HA represents the fraction of residues in the model that are within a short RMSD cutoff from a reference structure. Improvements in GDT-HA characterize to what extent the fraction of high-quality parts of a given structure is increased while ignoring parts of a structure that are of poor quality. RMSD changes capture the entire structure including bad parts of the structure. Often, GDT-HA and RMSD are highly correlated but in some cases, we find refinement in one measure but not in the other and vice versa. The first observation from the results in Tables 2-2 and 2-3 is that without restraints most of the structures move away from the native structure, some significantly, despite the relatively short simulation length of 24 ns. However, for the few cases where the final structure is refined, the improvement can also be quite significant, by about 1 Å for two targets. The occasional success

but overall failure with unrestrained MD simulations is consistent with similar findings by other groups.[70] When restraints are applied during simulations of the same length, the number of refined targets increases from 5 to 9 (out of total of 26 cases considered here) but while the restraints prevent large deviations away from the native they also limit to what extent structures can be improved.

Table 2-2 Changes in RMSD (Å) from the experimental structure relative to the RMSDs of the initial models during MD simulations with and without restraints over different simulation lengths. For all cases, the ΔRMSD for the final conformation and the overall lowest RMSD are given. Improved cases with negative ΔRMSD values are highlighted in bold.

Target	NO RESTRAINTS		WITH RESTRAINTS					
	24 ns		24 ns		8 × 3 ns		200 ns	
	Final	*Best*	*Final*	*Best*	*Final*	*Best*	*Final*	*Best*
TR592	0.42	***-0.05***	***-0.07***	***-0.18***	***-0.12***	***-0.16***	***-0.12***	***-0.20***
TR453	0.34	0.08	0.16	***-0.09***	0.17	***-0.10***	0.44	***-0.09***
TR432	1.39	***-0.12***	***-0.13***	***-0.30***	***-0.26***	***-0.34***	***-0.18***	***-0.31***
TR462a	0.53	0.04	0.44	0.04	***-0.04***	***-0.26***	0.42	***-0.07***
TR594	1.37	0.13	0.75	0.00	0.05	***-0.12***	0.34	0.00
TR614	1.11	0.03	0.40	***-0.13***	0.25	***-0.11***	0.08	***-0.13***
TR435	0.13	***-0.31***	0.30	0.03	0.07	***-0.08***	0.78	0.03
TR530	***-0.27***	***-0.64***	0.18	***-0.27***	***-0.22***	***-0.40***	0.26	***-0.35***
TR488	1.29	***-0.08***	0.00	***-0.25***	***-0.16***	***-0.23***	***-0.13***	***-0.26***
TR469	0.72	***-0.14***	***-0.02***	***-0.19***	***-0.09***	***-0.20***	0.15	***-0.19***
TR462b	0.23	***-0.16***	0.10	***-0.11***	***-0.02***	***-0.14***	0.17	***-0.11***
TR389	0.81	0.01	***-0.27***	***-0.62***	***-0.11***	***-0.51***	0.31	***-0.62***
TR464	0.89	***-0.14***	***-0.02***	***-0.16***	0.03	***-0.15***	***-0.12***	***-0.23***
TR569	0.46	***-0.03***	***-0.24***	***-0.50***	***-0.26***	***-0.47***	***-0.28***	***-0.69***
TR454	0.89	***-0.31***	0.06	***-0.15***	***-0.10***	***-0.19***	***-0.12***	***-0.20***
TR567	***-1.00***	***-1.46***	0.02	***-0.18***	***-0.03***	***-0.11***	***-0.06***	***-0.20***
TR574	1.82	0.15	1.07	0.07	0.09	***-0.50***	1.16	***-0.40***
TR557	***-0.01***	***-0.75***	***-0.58***	***-0.67***	***-0.35***	***-0.57***	***-0.61***	***-0.84***
TR429a	2.32	***-1.19***	0.20	***-0.20***	***-0.08***	***-0.21***	***-0.03***	***-0.26***
TR517	3.05	0.03	0.50	***-0.12***	0.15	***-0.17***	0.46	***-0.12***
TR606	1.76	0.01	1.63	***-0.28***	0.58	***-0.93***	***-0.80***	***-1.51***
TR429b	***-0.35***	***-0.59***	***-0.02***	***-0.17***	***-0.04***	***-0.25***	0.01	***-0.23***
TR624	***-0.90***	***-1.83***	***-0.21***	***-0.68***	***-0.03***	***-0.37***	***-0.63***	***-0.89***
TR568	0.59	0.13	0.07	***-0.31***	0.32	***-0.10***	0.29	***-0.43***
TR622	0.20	0.03	0.23	***-0.05***	0.06	***-0.72***	1.63	***-0.32***
TR576	1.01	0.51	0.74	0.49	0.70	0.37	0.28	0.00
Avg.	0.72	**-0.26**	0.20	**-0.19**	0.02	**-0.27**	0.14	**-0.33**
#better	**5**	**15**	**9**	**21**	**15**	**25**	**11**	**23**

Table 2-3 Changes in GDT-HA from the experimental structure relative to the GDT-HA values of the initial models during MD simulations as in Table 2-2. Improved cases with positive ΔGDT-HA values are highlighted in bold.

	NO RESTRAINTS		WITH RESTRAINTS					
	24 ns		24 ns		8 × 3 ns		200 ns	
	Final	*Best*	*Final*	*Best*	*Final*	*Best*	*Final*	*Best*
TR592	-10.5	***6.2***	***4.5***	***8.3***	***4.1***	***6.4***	***5.7***	***9.1***
TR453	-5.5	***4.0***	-4.0	***5.8***	***1.7***	***4.0***	***0.9***	***5.8***
TR432	-22.7	***1.4***	-2.5	***5.0***	***3.9***	***4.8***	***2.5***	***6.4***
TR462a	-6.0	***1.3***	-1.3	***5.7***	-0.3	***7.3***	-1.0	***7.3***
TR594	-23.8	-4.5	-0.7	***3.8***	***0.9***	***2.1***	-0.7	***4.1***
TR614	-16.2	-1.1	0.0	***4.6***	***0.4***	***2.8***	-1.4	***6.3***
TR435	-6.2	***0.2***	-4.9	***0.2***	-1.6	***1.3***	-3.5	***1.8***
TR530	-1.6	***2.8***	-0.9	***3.1***	***1.3***	***4.4***	-2.8	***3.1***
TR488	-1.3	***6.6***	***2.4***	***6.6***	***5.0***	***5.8***	***5.3***	***7.1***
TR469	-16.7	-4.8	-2.4	***2.0***	-1.6	***2.8***	-5.6	***3.2***
TR462b	***1.5***	***8.5***	-1.1	***2.2***	***0.7***	***4.0***	-1.5	***2.6***
TR389	-18.9	-7.3	-6.9	-1.9	-5.8	-0.6	-7.3	-1.9
TR464	-4.7	***3.3***	0.0	***5.4***	-0.4	***3.6***	***1.5***	***6.2***
TR569	-7.0	***3.2***	0.0	***6.0***	***1.3***	***5.7***	-0.6	***7.6***
TR454	-11.6	***1.3***	-1.3	***2.3***	***1.3***	***3.0***	***0.3***	***4.0***
TR567	-3.0	***1.6***	-0.4	***4.8***	***2.5***	***4.2***	***2.8***	***5.3***
TR574	-7.9	-2.5	***1.7***	***4.2***	***3.2***	***3.7***	***0.7***	***6.4***
TR557	***1.0***	***7.2***	***2.2***	***7.4***	***3.8***	***6.6***	***5.2***	***9.0***
TR429a	***0.4***	***10.5***	***2.8***	***12.1***	***5.6***	***11.7***	***8.9***	***14.5***
TR517	-1.6	***2.4***	-1.9	***3.0***	***2.0***	***3.6***	-2.4	***3.0***
TR606	-7.9	-2.2	***1.8***	***4.7***	-0.4	***3.5***	-1.4	***5.7***
TR429b	***2.6***	***7.9***	-0.7	***2.6***	0.0	***4.6***	-1.0	***4.3***
TR624	***6.9***	***11.2***	***4.4***	***6.2***	***0.4***	***4.0***	***2.2***	***6.2***
TR568	-8.0	***3.9***	***1.0***	***3.9***	***0.3***	***3.4***	***0.3***	***4.6***
TR622	-14.1	0.0	***2.7***	***6.0***	***4.8***	***6.0***	***2.5***	***7.4***
TR576	-11.6	-4.2	-1.1	***2.0***	-0.7	***0.5***	-1.3	***2.9***
Avg.	-7.5	**2.2**	-0.3	**4.5**	**1.2**	**4.2**	**0.3**	**5.5**
#better	**5**	**18**	**9**	**25**	**18**	**25**	**13**	**25**

Extending the sampling to 200 ns further increases the number of structures that were refined at the end to 11 (according to RMSD) or 13 (according to GDT-HA). However, even better results were found when the average final structures from many short simulations (8 × 3 ns) were

considered with now more than half of the structures being refined. The use of multiple short simulations is expected to improve sampling over a single long simulation[41, 95] and our results suggest that increased sampling does lead to improved success with refinement. This is in agreement with previous findings.[70] It is interesting to note that when selecting the average final structure from the 8 × 3 ns simulations, we already find an average improvement in GDT-HA score by 1.2, comparable to the results reported by the Shaw group after much longer simulations.

As shown in Figures 2-6 and 2-7 (supplementary material), the RMSD and GDT-HA scores fluctuate significantly during the simulations and while the final structures are often not improved, there are improved structures at other times during the simulation for many targets. Tables 2-2 and 2-3 also show the improvement in RMSD and GDT-HA for the best structures (in terms of RMSD or GDT-HA) that were sampled during the simulations. Without restraints, only about half of the targets are refined at some point during the trajectory, but with restraints refined structures are found for almost all of the targets, in particular during the longer 200 ns simulation and during the multiple short simulations. The average maximum improvement in terms of GDT-HA is again similar to the values for the simulations from the Shaw group after about 10 μs. This finding raises the possibility that such long simulations may not be necessary to achieve refinement and that other methodological factors may be more critical.

Table 2-4 Changes in RMSD (Å) and GDT-HA upon selecting structures with the lowest DFIRE score and correlation coefficients of RMSD or GDT-HA vs. iRMSD or DFIRE. Correlation coefficients larger than 0.30 (RMSD) or less than -0.30 (GDT-HA) are highlighted in bold.

Target	200 ns				8 × 3 ns			
	Δ RMSD	Δ GDT-HA	Correlation RMSD/GDT-HA		Δ RMSD	Δ GDT-HA	Correlation RMSD/GDT-HA	
			vs iRMSD	vs DFIRE			vs iRMSD	vs DFIRE
TR592	***-0.06***	***0.5***	0.10/0.02	0.04/-0.10	***-0.06***	***0.2***	0.16/-0.09	**0.35**/-0.27
TR453	0.16	***0.3***	**0.95/-0.33**	0.19/-0.17	0.30	-1.4	**0.89/-0.43**	**0.35/-0.30**
TR432	***-0.12***	***2.1***	-0.03/0.02	0.06/-0.11	***-0.04***	***0.4***	-0.25/0.24	-0.01/-0.17
TR462a	0.21	***4.7***	**0.51/-0.43**	0.25/**-0.51**	0.31	-0.7	-0.11/-0.21	-0.16/-0.14
TR594	0.18	***2.0***	**0.61**/0.07	**0.30**/-0.25	0.07	-1.4	**0.50**/-0.06	0.17/-0.15
TR614	0.38	-4.2	0.05/-0.02	0.22/**-0.32**	0.29	***0.7***	-0.03/-0.01	0.06/**-0.33**
TR435	0.20	-2.7	**0.95**/0.15	**0.57**/0.03	0.08	-3.1	**0.71/-0.37**	**0.34**/-0.21
TR530	0.96	-3.4	**0.93/-0.55**	-0.02/-0.03	0.03	***0.3***	0.16/-0.14	0.15/-0.27
TR488	***-0.13***	***2.9***	-0.20/0.23	0.01/-0.14	***-0.10***	***0.3***	-0.24/0.24	0.06/-0.11
TR469	0.09	-3.2	**0.46**/-0.27	0.11/-0.22	***-0.04***	-0.8	-0.10/-0.12	0.22/-0.26
TR462b	0.30	-3.3	**0.57/-0.43**	-0.15/0.04	0.02	-1.5	**0.43/-0.48**	0.24/-0.23
TR389	0.30	-7.1	**0.71**/-0.22	0.27/-0.15	***-0.51***	-5.8	0.08/**-0.49**	**0.62**/-0.28
TR464	***-0.13***	***0.4***	**-0.37**/0.18	0.12/-0.03	0.04	-2.2	-0.14/0.00	-0.06/0.07
TR569	***-0.37***	***3.8***	**-0.45**/0.21	0.01/-0.13	***-0.03***	0.0	**-0.70**/0.08	0.18/0.00
TR454	***-0.09***	***0.1***	**0.37**/-0.16	**0.35/-0.37**	***-0.19***	***0.8***	0.13/-0.07	0.09/-0.16
TR567	***-0.05***	***0.7***	-0.10/-0.11	-0.07/-0.08	***-0.02***	***0.7***	0.05/-0.20	0.05/0.03
TR574	1.08	***2.0***	**0.64**/-0.03	0.15/-0.25	***-0.09***	-2.0	**0.32**/0.22	**0.48**/-0.18
TR557	***-0.56***	***6.0***	**-0.30/0.36**	-0.16/0.00	***-0.03***	***1.6***	**-0.66/0.34**	-0.20/0.06
TR429a	***-0.14***	***9.7***	0.20/-0.09	**0.32/-0.32**	0.06	***6.5***	**0.36**/-0.04	0.04/-0.23
TR517	0.03	-1.3	**0.48**/0.00	**0.43**/0.01	0.01	***1.1***	**0.51**/-0.12	**0.45**/-0.15
TR606	***-0.96***	***0.4***	-0.04/0.08	**0.80**/-0.14	***-0.26***	-2.2	**0.43**/-0.02	**0.55**/-0.01
TR429b	***-0.09***	***1.0***	0.19/-0.06	**0.41**/-0.27	***-0.10***	***0.3***	**0.71/-0.50**	**0.48/-0.44**
TR624	***-0.44***	***1.8***	**-0.46**/-0.05	0.06/-0.05	0.32	-2.2	0.16/-0.03	-0.12/0.05
TR568	0.03	***2.6***	-0.14/0.15	0.02/-0.20	0.14	***1.6***	**0.36**/-0.20	0.29/-0.26
TR622	0.04	***3.9***	**0.82**/-0.23	**0.77**/-0.28	0.27	***0.8***	-0.23/-0.06	0.09/0.05
TR576	0.29	-2.5	-0.22/0.04	**0.40**/-0.10	0.84	-4.9	**0.47**/0.03	0.07/-0.08
Avg.	0.04	**0.7**	0.24/-0.06	0.21/-0.16	0.05	-0.5	0.15/-0.10	0.18/-0.15

2.4.2 Lowest-scoring Structures

Since refined structures were generated during most of the simulations, the next question we investigated was whether application of a scoring function to an ensemble of structures extracted from the MD runs would allow us to identify the most native-like, and therefore refined

structures. Table 2-4 shows the change in RMSD and GDT-HA with respect to the experimental structures when selecting the conformation with the lowest DFIRE score. We chose DFIRE as one of the best-performing scoring functions that has been widely applied in structure prediction applications.[44, 84] The results indicate that selecting structures based on the lowest DFIRE score has similar performance or is even slightly worse than simply taking the final structures. This is not entirely surprising when considering the correlation coefficients between RMSD or GDT-HA and the DFIRE score. Although the correlation coefficients largely have the correct sign (positive for RMSD, negative for GDT-HA), their small magnitude – with a few exceptions – suggests that it would be difficult to reliably select a single structure. We also considered other scoring functions (data not shown) and found similar results.

2.4.3 Ensemble-averaged Structures

Next, we considered that experimental structures are the product of conformational averaging rather than representing single snapshots. Consequently, we obtained average structures from the MD-generated structure ensembles. Figure 2-1 shows the effect of averaging different percentages of the MD-generated structures that were sorted either according to their DFIRE score or based on their distance from the initial structure (iRMSD). We find that averaging generally outperforms selecting a single structure, while averaging over the 10% of structures with the lowest DFIRE scores results in a maximum improvement in GDT-HA by 2.6, which is about half of what could be achieved theoretically if the best conformation could be selected from each trajectory. However, when considering RMSD, an even smaller ensemble of only the 1% best-scoring structures results in a maximum improvement by 0.04 Å. Interestingly, selecting structures according to low iRMSD values, i.e. averaging over structures that have moved the least from the initial structure, also results in refinement. The rationale for that finding is that

when structures start to deviate significantly from the initial template-based model, they are much more likely to move away from the native structure than towards it.

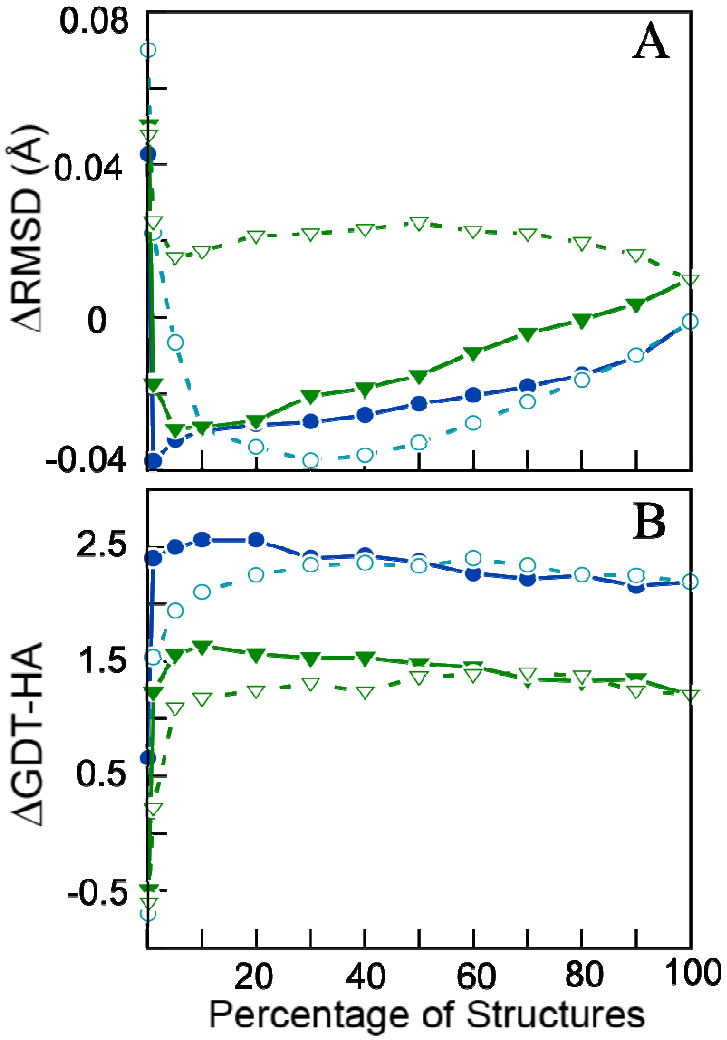

Figure 2-1 Change in RMSD with respect to native structure (A) and in GDT-HA (B) upon averaging different subsets of structures sorted by either DFIRE scores or iRMSD. Results from the 200 ns MD runs are shown in blue (circles) and from 8x3 ns sampling in green (triangles). Open symbols denote iRMSD-based selection; closed symbols refer to DFIRE-based selection.

The observation that both DFIRE and iRMSD appear to be suitable metrics to identify ensembles of structures that when averaged provide structures that are likely closer to the native state, prompted us to consider a combination of both scores for selecting a subset of structures to be

averaged. Since the range of these two scores is different, we first normalized the values by subtracting the mean and dividing by their respective standard deviations for a given set of structures. We then chose values in an open arc segment as illustrated in Fig. 2-2. Given the identity line through the origin (dashed line in Fig. 2-2), structures were chosen within a given angle $\theta/2$, around the line to the origin and at a minimum radial distance ρ from the center of the distribution.

To find optimal values of (ρ, θ), we varied ρ from 0.2 to 1.9 with increments of 0.1, and changed the angle θ from 30 to 200 degrees at increments of 10. For each target, we extracted the structures that lie in the aforementioned region, and then calculated the average structure. Figure 2-3 shows the average improvements in RMSD and GDT-HA as functions of ρ and θ. As optimal values that maximize both RMSD and GDT-HA we chose ρ=1.2 and θ=120°. Using these values, the RMSD is improved by 0.07 Å and GDT-HA scores by 2.6. The improvements in RMSD and GDT-HA for individual targets using this criterion are given in Table 2-5. We find that GDT-HA is not further improved over simply selecting the 10% of the structures with the lowest DFIRE score but the improvement in RMSD appears to be more significant.

A drawback of structure averaging is that further refinement is necessary afterwards to generate stereochemically good models. As an alternative protocol, we also selected the ensemble structure closest to the subset averages. The data given in Table 2-8 shows that on average there is no improvement in RMSD and there is only a small improvement in GDT-HA for structures taken from the 200 ns simulation. This suggests that averaging rather than selecting a single structure is a key to the success of the refinement protocol described here.

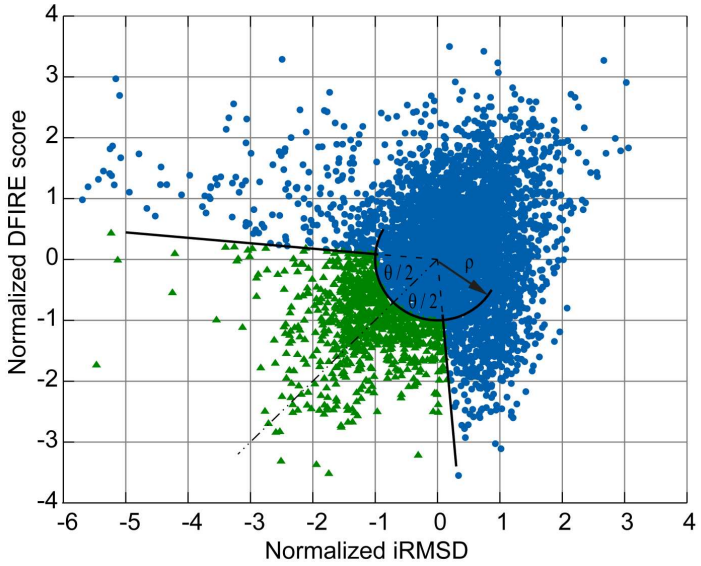

Figure 2-2 Subset selection based on combination of DFIRE and iRMSD scores (normalized by their respective standard deviations). Selected structures (green triangles) are outside the circle with radius (ρ) and within the segment with angle (θ).

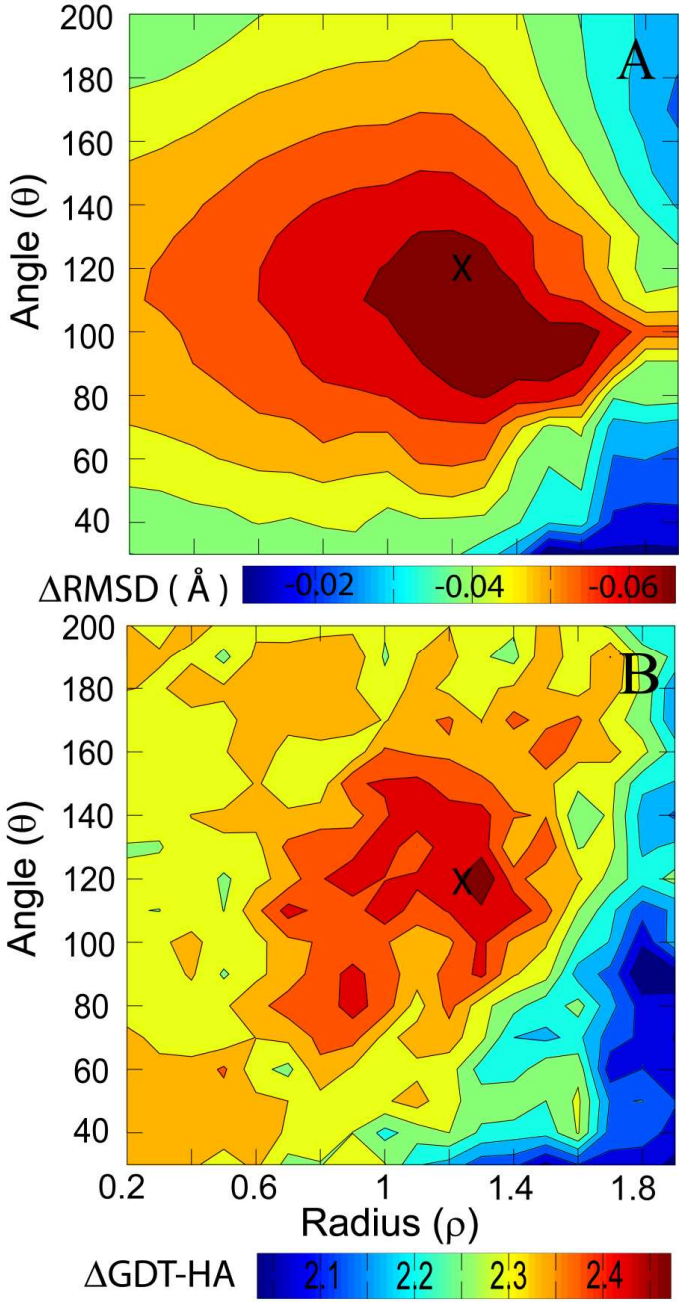

Figure 2-3 Change in RMSD with respect to native structure (A) and GDT-HA (B) as a function of radius (ρ), and angle (θ). Parameters considered to be optimal and used subsequently for subset averaging are indicated by 'X'.

Table 2-5 Change in RMSD (Å) and GDT-HA upon averaging over selected subsets (see text) with and without additional structure interpolation. Averages were calculated for all targets and for those where the correlation coefficient of iRMSD vs. DFIRE is less than 0.4 (indicated by*)

Target	200 ns					8 × 3 ns				
	Corr. iRMSD vs. DFIRE	Subset Average		Structure Interpolation		Corr. iRMSD vs. DFIRE	Subset Average		Structure Interpolation	
		Δ RMSD	Δ GDT-HA	Δ RMSD	Δ GDT-HA		Δ RMSD	Δ GDT-HA	Δ RMSD	Δ GDT-HA
TR592	0.01	***-0.14***	***6.2***	***-0.13***	***4.3***	0.14	***-0.12***	***3.1***	***-0.11***	***1.9***
TR453	0.14	0.09	***2.9***	0.04	***2.9***	0.20	0.03	***2.3***	0.00	***2.9***
TR432	-0.03	***-0.19***	***4.4***	***-0.19***	***3.7***	0.18	***-0.14***	***3.5***	***-0.12***	***3.7***
TR462a	**0.49***	0.20	***4.0***	0.13	***3.0***	**0.68***	0.08	***0.7***	0.03	***0.3***
TR594	-0.05	0.15	***2.0***	0.09	***1.3***	0.13	0.01	***0.7***	***-0.01***	***0.7***
TR614	**0.45***	0.24	***3.9***	0.08	***4.2***	**0.54***	0.33	-0.4	0.22	***0.7***
TR435	**0.59***	0.23	-1.8	0.14	-0.9	0.36	***-0.01***	-0.9	***-0.02***	-0.2
TR530	-0.07	***-0.16***	***0.9***	***-0.16***	***0.6***	0.11	***-0.17***	***2.2***	***-0.15***	***1.6***
TR488	0.04	***-0.12***	***5.0***	***-0.11***	***4.5***	-0.06	***-0.13***	***4.2***	***-0.12***	***4.5***
TR469	-0.16	***-0.02***	-0.8	***-0.03***	***0.0***	0.06	***-0.06***	-2.4	***-0.06***	-0.8
TR462b	-0.28	0.07	***0.7***	0.00	***2.6***	0.37	***-0.03***	***2.2***	***-0.06***	***3.3***
TR389	0.17	***-0.43***	-2.6	***-0.48***	-1.5	0.27	***-0.14***	-2.2	***-0.16***	-0.8
TR464	0.09	***-0.01***	***1.1***	***-0.01***	***0.7***	0.14	0.03	***0.4***	0.02	0.0
TR569	0.06	***-0.29***	***1.0***	***-0.27***	***1.0***	-0.19	***-0.07***	***0.3***	***-0.06***	***1.0***
TR454	0.23	***-0.09***	***1.7***	***-0.09***	***1.8***	0.14	***-0.08***	***1.3***	***-0.07***	***2.1***
TR567	0.23	***-0.06***	***3.3***	***-0.06***	***2.5***	-0.04	***-0.02***	***3.3***	***-0.02***	***2.6***
TR574	-0.01	0.24	***3.9***	0.10	***2.7***	0.12	***-0.04***	***1.5***	***-0.06***	***1.2***
TR557	0.12	***-0.56***	***4.2***	***-0.49***	***3.6***	0.33	***-0.18***	***3.8***	***-0.15***	***3.0***
TR429a	0.18	***-0.09***	***9.3***	***-0.10***	***8.5***	0.10	***-0.08***	***6.1***	***-0.08***	***6.9***
TR517	0.28	0.22	***1.3***	0.12	***1.4***	**0.50***	0.03	***2.4***	0.02	***2.0***
TR606	-0.19	***-1.04***	***2.6***	***-1.00***	***3.3***	0.32	***-0.01***	***0.2***	***-0.03***	***0.4***
TR429b	0.28	***-0.12***	***0.3***	***-0.13***	0.0	**0.49***	***-0.15***	***1.7***	***-0.13***	***1.3***
TR624	-0.09	***-0.33***	***4.0***	***-0.29***	***3.6***	-0.07	0.00	***0.4***	***-0.01***	0.0
TR568	-0.09	0.02	***3.1***	***-0.02***	***2.8***	**0.48***	0.18	***1.0***	0.14	***1.0***
TR622	**0.84***	0.14	***5.8***	0.07	***5.4***	0.26	0.17	***4.3***	0.12	***3.9***
TR576	0.31	0.31	***0.4***	0.21	***0.4***	**0.52***	0.68	-1.3	0.52	-0.9
Avg.	--	**-0.07**	**2.6**	**-0.10**	**2.4**	--	0.00	**1.5**	**-0.01**	**1.6**
Avg.*	--	**-0.12**	**2.5**	**-0.14**	**2.3**	--	**-0.05**	**1.7**	**-0.06**	**1.9**
#better	--	**15**	**23**	**16**	**23**	--	**16**	**21**	**18**	**20**

2.4.4 Structure Interpolation

As a result of subset averaging described above we can generate refined structures for a majority of cases (15-16 out of 26 in terms of RMSD and 21-23 in terms of GDT-HA, see Table 2-5). The idea we followed next was that whether it would be possible to refine structures further by extrapolating the 3N-dimensional vector between the initial model and the refined structures. More specifically, we consider the vector difference between the C_α coordinates in the initial model, $\vec{R}_{C_\alpha}^{(init)}$, and the ones obtained from the ensemble-averaged structures $\vec{R}_{C_\alpha}^{(avg)}$, most of which are refined relative to the initial model. Note, that the average structure is already superimposed to the initial model as a result of how the ensemble average was generated. We then tested whether a new set of coordinates obtained according to Eq. 2-1 would increase the degree of refinement:

$$\vec{R}_{C_\alpha}^{(new)} = (1-\alpha)\vec{R}_{C_\alpha}^{(init)} + \alpha\,\vec{R}_{C_\alpha}^{(avg)} \qquad 2\text{-}1$$

where α is a scaling factor. Here, α=0 corresponds to the initial model, and α=1 corresponds to the ensemble-averaged structure. Values of α between 0 and 1 would correspond to interpolation between the initial and refined structures, values beyond 1 would be extrapolation beyond the refined structures. Figure 2-4 shows the effect of applying Eq. 1 on the overall change in GDT-HA and RMSD. We find the optimum value of α to be α=0.6 for maximizing improvements in RMSD, and α=1 for GDT-HA. This result was surprising as we expected that values of α>1 may improve structures further. However, closer inspection of which targets are most affected by the structure interpolation approach suggests that scaling coordinates according to Eq. 2-1 has a stronger effect on the RMSD of targets where the RMSD increased during the refinement stage (see Fig. 2-5), i.e. structures that were made worse during the refinement. On the other hand,

there was less of an impact on the structures that could be refined. Hence, the overall effect is an average improvement. It is unclear to what extent this is a general finding but as a result of applying the structure interpolation method (with α=0.8) we find further improvement in terms of RMSD. However, GDT-HA becomes slightly worse when the structure interpolation method is applied.

The restraints applied during the MD simulations were either given by the CASP organizers or determined by us (see Table 2-1). An interesting question is whether the origin of the restraint list had an impact on the refinement success. The changes in RMSD and GDT-HA after refinement for the targets with CASP-suggested restraints were -1.4 Å and 2.6, respectively, but somewhat less, -0.04 Å and 2.0, respectively, for the targets where we selected the restraints. Hence, refinement is most successful if sampling can be targeted to the regions known to be deviating most from the native.

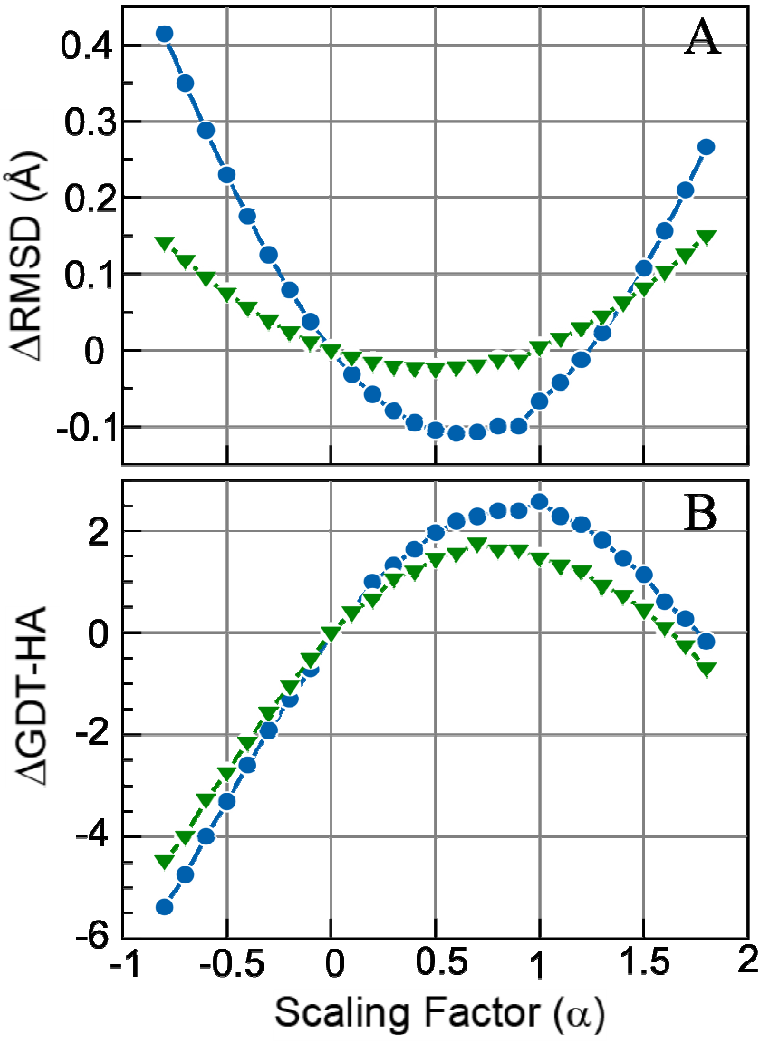

Figure 2-4 Change in RMSD with respect to native structure (A) and GDT-HA (B) upon structure interpolation between the initial (α=0.0) and the subset-averaged structures (at α=1.0). Results from 200 ns MD runs are shown in blue (circles) and from 8x3 ns sampling in green (triangles).

2.4.5 Quality Assessment

Finally, we considered whether it is possible to predict in which cases refinement is successful and when structures become worse as a result of refinement. Motivated by a previous analysis using a correlation-based metric,[96, 97] we considered the correlation between the two scores iRMSD and DFIRE, both of which are available without knowledge of the native structure. The rationale for using this score is that because iRMSD is often correlated with RMSD (see Table 2-

4), a correlation between DFIRE and iRMSD is indicative of a correlation between DFIRE and RMSD. Figure 2-5 shows the change in RMSD after refinement as a function of this correlation coefficient. It can be seen that all of the significantly refined structures have a correlation coefficient between -0.4 and 0.4 while higher correlation coefficients larger than 0.4 correlate with a lack of refinement. Significant correlation between DFIRE and RMSD (and by proxy with iRMSD) most likely occurs when structures move by a significant extent. It appears from this analysis that in those cases the motion is likely to be away from the native structure rather than towards it. Using a DFIRE/iRMSD correlation coefficient of <0.4 as a criterion that refinement has been successful, we identify four cases, TR435, TR462A, TR614, and TR622, that are outside this range and for which refinement was therefore assumed not to be successful. If we use the initial model (ΔRMSD=0) for these targets instead of the 'refined' structures, the average change in RMSD from the native improves further, to -0.12 (without structure interpolation) and to -0.14 (with structure interpolation). The effect on GDT-HA is less clear, because the improvement is actually slightly decreased for the 200 ns set but it improves for the 8 × 3 ns sampling set.

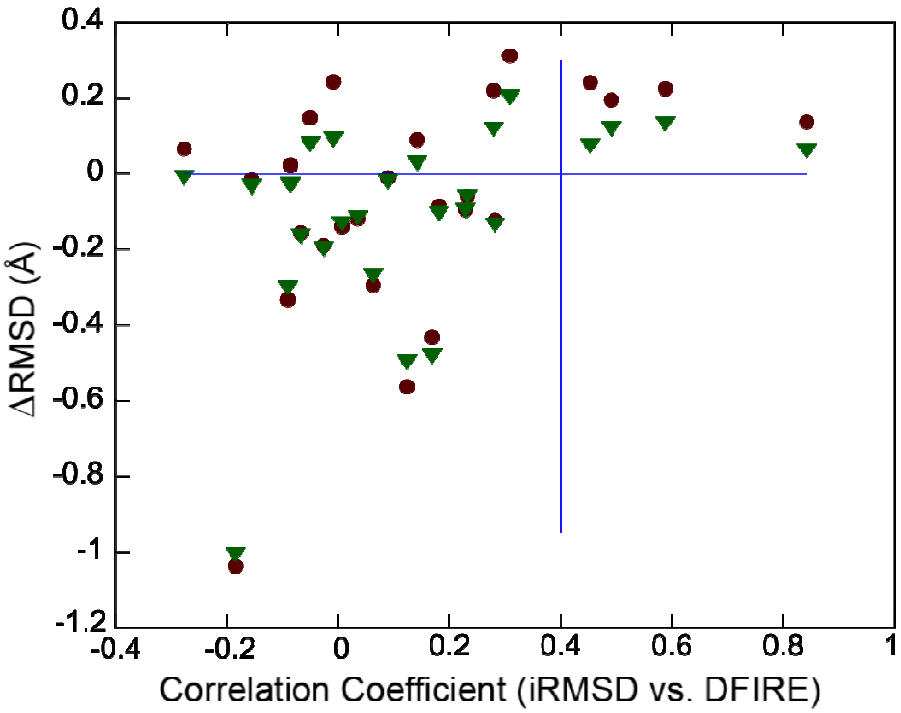

Figure 2-5 Change in RMSD with respect to native structure as a function of correlation between iRMSD and DFIRE scores with (green triangles) and without (red squares) structure interpolation.

2.4.6 Final Refinement of Averaged Structures

So far, the structural analysis has focused on the C_α coordinates. As a result of the averaging and structure interpolation procedures, the generated structures are of poor quality in terms of bond geometries, clashes, etc. which is readily apparent when submitting those models to structural analysis tools (see Table 2-6). In order to generate overall high quality structures, we performed additional short MD simulations where the C_α atoms were constrained to maintain the overall improvement in structure but where other atoms were allowed to relax. The quality of the final models was improved dramatically (see Table 2-7) to result in high-quality refined structures. After the final step, the average change in RMSD was still -0.08 Å, and the change in GDT-HA was 2.3. For comparison with other studies, we also calculated the average improvement in GDT-TS for the final structures to be 1.6.

Table 2-6 Quality measures of averaged structures before (*Avg*) and after (*MD*) refinement via restrained MD simulations.

Target	Clash score		% poor rotamers		% Ramach. outliers		C_β dev.		% bad bonds		% bad angles		MolProbity score	
	Avg	*MD*	*Avg*	*MD*	*Avg*	*MD*	*Avg*	*MD*	*Avg*	*MD*	*Avg*	*MD*	*Avg*	*MD*
TR592	147.8	3.0	5.6	4.2	8.8	0.0	64	2	82.7	0.0	41.4	0.0	3.9	2.0
TR453	435.8	2.2	11.4	5.7	8.4	0.0	77	3	92.9	0.0	75.0	0.0	4.6	1.9
TR432	295.0	0.9	5.5	4.6	7.2	0.8	112	2	85.8	0.0	76.4	0.0	4.2	1.3
TR462a	445.7	0.0	14.3	0.0	15.3	4.2	63	3	94.6	0.0	93.2	0.0	4.7	1.0
TR594	253.2	3.1	9.7	5.3	14.2	6.0	116	5	91.2	0.0	74.3	0.7	4.4	2.3
TR614	722.7	8.2	45.7	10.6	31.0	10.6	119	15	100	0.0	100	6.1	5.6	2.9
TR435	382.9	1.4	13.0	4.6	7.6	0.8	114	3	88.7	0.0	75.2	0.8	4.6	1.8
TR530	219.0	3.9	9.1	3.0	10.4	2.6	63	1	80.8	0.0	65.4	0.0	4.2	2.0
TR488	314.1	0.7	9.0	7.5	4.4	3.3	72	3	89.3	0.0	61.3	0.0	4.3	1.8
TR469	366.8	1.1	2.3	4.6	3.4	1.7	49	0	78.7	0.0	75.4	0.0	3.9	1.5
TR462b	686.6	1.8	20.4	6.1	19.7	4.6	58	2	98.5	0.0	92.7	1.5	5.2	2.1
TR389	537.0	4.2	17.4	10.1	14.0	5.4	124	11	96.2	0.0	92.4	3.8	5.0	2.6
TR464	126.5	0.0	0.0	2.0	1.5	0.0	44	0	79.4	0.0	50.0	0.0	2.7	0.7
TR569	320.3	2.7	8.6	1.7	6.7	1.3	57	3	85.7	0.0	75.3	1.3	4.3	1.6
TR454	243.4	1.0	4.4	1.5	5.9	0.5	149	0	86.7	0.0	56.9	0.5	4.0	1.0
TR567	168.4	2.2	6.7	1.0	4.5	0.8	101	3	75.0	0.0	46.3	0.7	3.9	1.3
TR574	515.8	3.9	28.2	6.4	23.2	5.1	94	6	97.0	0.0	90.1	4.0	5.2	2.5
TR557	336.6	2.6	11.1	2.0	9.0	2.5	111	5	91.9	0.0	83.9	0.8	4.5	1.8
TR429a	656.0	1.6	26.5	4.4	34.2	7.9	76	6	10	0.0	98.7	3.9	5.3	2.1
TR517	363.7	2.7	11.5	5.3	8.9	3.8	148	8	96.9	0.0	87.4	0.6	4.5	2.1
TR606	590.1	6.3	38.4	6.1	30.6	5.0	118	15	95.9	0.0	100	4.9	5.4	2.6
TR429b	551.9	3.2	14.3	6.4	28.4	8.1	71	8	92.1	0.0	96.1	4.0	5.0	2.4
TR624	330.1	3.6	15.5	5.2	9.1	0.0	58	2	95.6	0.0	88.2	0.0	4.6	2.1
TR568	455.5	2.6	11.0	2.4	17.2	4.3	87	4	97.9	0.0	96.8	1.1	4.8	1.8
TR622	472.8	5.5	37.4	7.7	29.3	3.5	110	7	97.5	0.0	96.6	0.9	5.3	2.5
TR576	702.3	5.5	38.2	11.8	24.8	6.8	127	13	98.5	0.0	99.3	1.5	5.4	2.8
Avg:	409.2	**2.8**	16.0	**5.0**	14.5	**3.4**	92	**5**	91.1	**0.0**	80.3	**1.4**	4.6	**1.9**

Table 2-7 Summary of the average improvements in RMSD (Å) and GDT-HA for all the attempted methods for structure selection out of 8×3 ns and 200 ns simulation sets; Best in trajectory is given as a reference for the maximum possible improvement.

Method:	Δ RMSD (Å)		Δ GDT-HA	
	8 × 3 ns	200 ns	8 × 3 ns	200 ns
Best in trajectory	**-0.27**	**-0.33**	**4.2**	**5.5**
Final Structure	0.02	0.14	**1.2**	**0.3**
Lowest DFIRE	0.05	0.04	-0.5	**0.7**
Average over 10% lowest DFIRE	**-0.03**	**-0.04**	**1.6**	**2.6**
Average over 1% lowest iRMSD	0.01	**-0.04**	**1.4**	**2.4**
Subset average from combined DFIRE/iRMSD scores	0.00	**-0.07**	**1.5**	**2.6**
Closest structure to subset average	0.07	0.01	-0.6	0.6
Subset average and structure interpolation	**-0.01**	**-0.10**	**1.6**	**2.4**
Subset average/interpolation with correlation-based filtering	**-0.06**	**-0.14**	**1.9**	**2.3**

2.5 Discussion and Conclusion

We are presenting here a new protocol for structure refinement that is based on MD simulations, but adds a new scoring and averaging protocol. A summary of the performance with different structure selection methods is presented in Table 2-7. Overall, the refinement results reported here are moderate, but what we consider most important is that we are able to consistently refine the large majority of structures rather than making a significant fraction worse as in earlier attempts at structure refinement. The overall refinement results are better than those reported recently by the Shaw group despite the much shorter simulations used here which may be due to a number of different reasons. The force field that was used here is a recently updated version of the CHARMM force field that appears to outperform most other available force fields in other tests.[9] Furthermore, the use of ensemble averages instead of single structures appears to lead to significant improvements that may compensate for the much more limited sampling compared to

the work by Shaw et al. With respect to the sampling, we find that nearly equivalent refinement can be achieved with multiple short simulations rather than a single long simulation. This is consistent with previous findings,[41, 95] but is a point that merits further investigation since it is generally much easier to run many short simulations than one very long simulation on commonly available computer platforms. We also attempted here to employ an extrapolation scheme to further refine structures –which was not successful so far – and an assessment criterion to determine whether structure refinement is successful –which does appear to have merit.

Another question is whether the refinement success is biased by how the starting structures were generated. The targets considered here were selected by the CASP organizers from the best predictions during the CASP competition. While this limits the methods by which the models were generated to a few top groups, an effort was made to avoid selecting models from only one participating group. Hence, the models used as starting structures here represent some degree of diversity in terms of how they were created. Since we see consistent refinement across most of the targets we assume that refinement success is independent of the exact way the structures were initially prepared. Furthermore, similar results for sampling from 200 ns simulations vs. 8 x 3 ns simulations suggests that just a few nanoseconds were enough to equilibrate the structures sufficiently.

Finally, it would be interesting to see whether repeated application of the protocol presented here can be used in an iterative protocol to achieve more significant refinement. These are areas that we will focus on in more detail in future studies.

2.6 Acknowledgment

We would like to thank Nan Liu for the initial setup and generation of some of the simulation data presented here. Funding from NIH GM084953 and NSF CBET 0941055 is acknowledged. Computer resources were used at XSEDE facilities (TG-MCB090003) and at the High-Performance Computing Center at Michigan State University.

2.7 Supporting Information

Table 2-8 Change in RMSD (Å) and GDT-HA for the structure closest to the subset average relative to their respective values from the initial models.

Target	200 ns		8 × 3 ns	
	Δ RMSD	Δ GDT-HA	Δ RMSD	Δ GDT-HA
TR592	***-0.08***	***3.1***	***-0.03***	-0.2
TR453	0.15	-0.3	0.07	-0.3
TR432	***-0.20***	***0.4***	***-0.11***	***0.4***
TR462a	0.20	***2.7***	0.11	***0.7***
TR594	0.24	-0.2	0.04	-0.9
TR614	0.44	-1.1	0.19	-2.5
TR435	0.26	-3.8	0.08	-3.3
TR530	***-0.11***	-1.6	***-0.09***	-2.2
TR488	***-0.14***	***2.1***	***-0.07***	***0.5***
TR469	0.03	-2.8	0.00	-3.2
TR462b	0.11	***0.4***	0.08	***0.4***
TR389	***-0.21***	-5.2	0.03	-4.9
TR464	0.03	***0.7***	0.04	-0.4
TR569	***-0.30***	-1.0	***-0.03***	***0.3***
TR454	***-0.08***	***1.4***	***-0.05***	-0.1
TR567	***-0.05***	***2.6***	0.01	***2.8***
TR574	0.79	***0.7***	0.14	-2.7
TR557	***-0.50***	***1.4***	**-0.16**	***2.2***
TR429a	***-0.11***	***8.9***	0.03	***2.4***
TR517	0.31	-1.0	0.07	***0.8***
TR606	***-0.48***	***1.2***	0.02	-2.9
TR429b	***-0.13***	-0.3	***-0.06***	**0.3**
TR624	***-0.42***	***4.0***	0.06	0.0
TR568	***-0.06***	***1.6***	0.22	***0.5***
TR622	0.21	***2.5***	0.39	-0.6
TR576	0.40	-0.7	0.72	-3.3
Avg.	0.01	**0.6**	0.07	-0.6

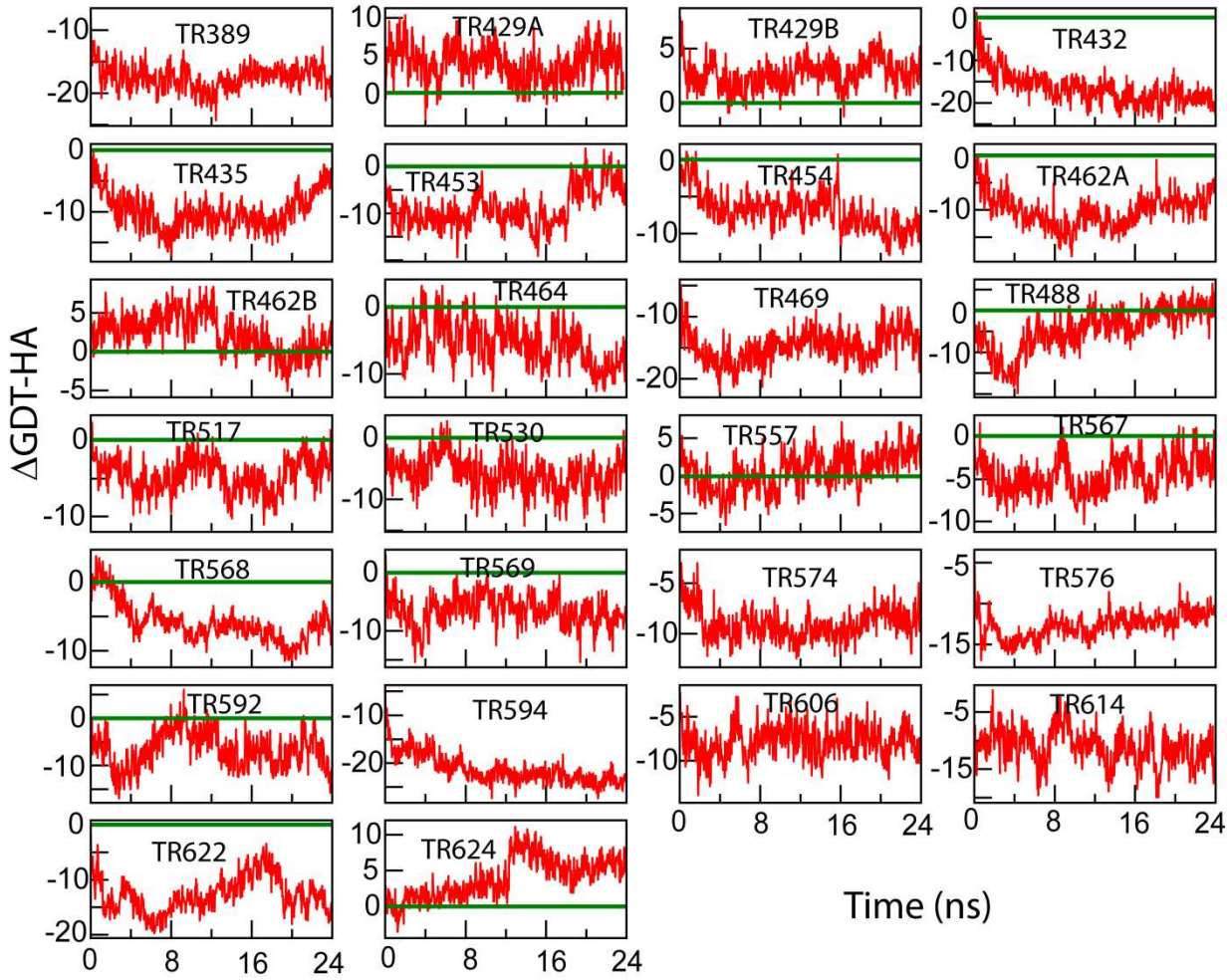

Figure 2-6 Change in GDT-HA of all CASP8 and CASP9 targets after refinement without imposing restraints using C36ff.

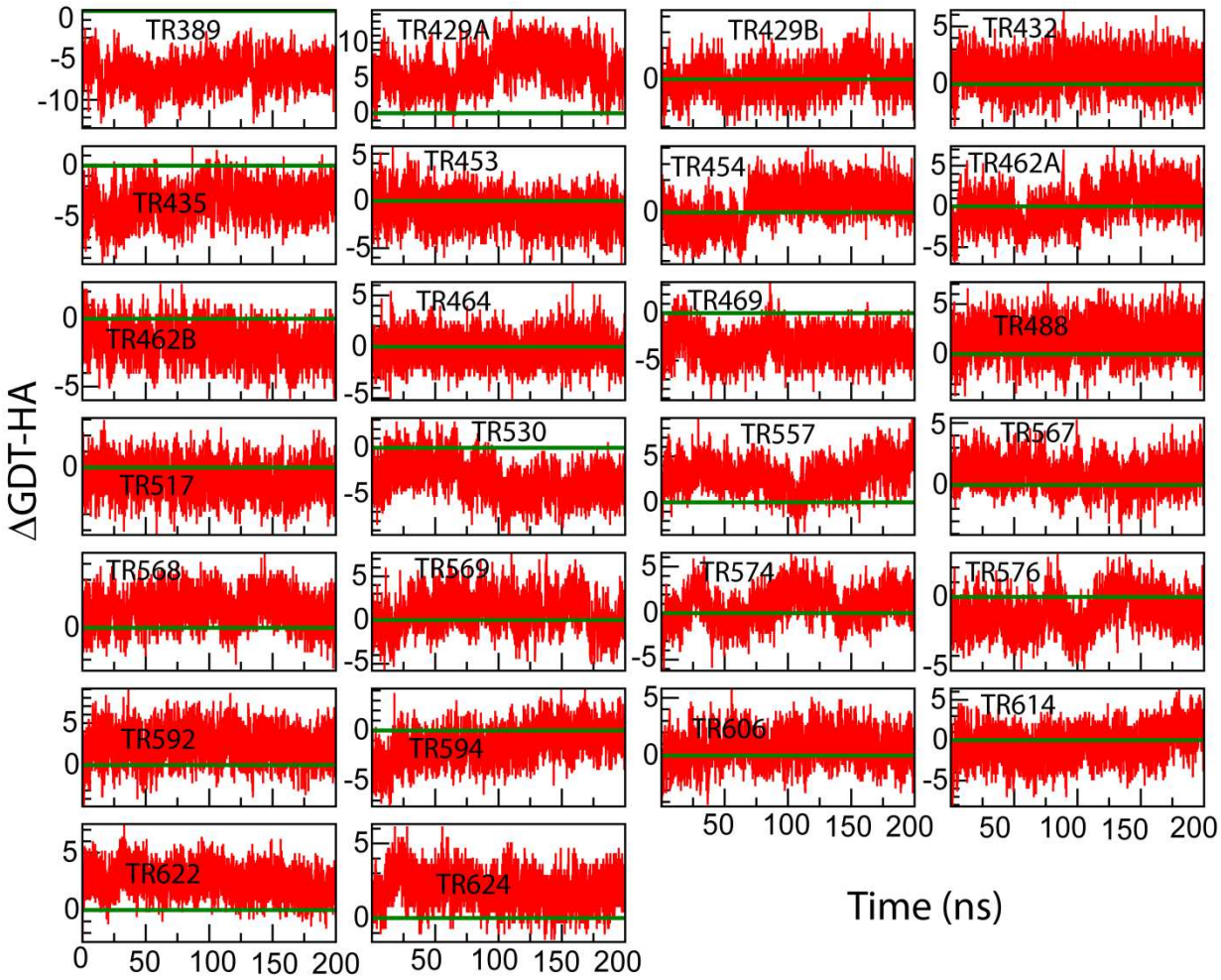

Figure 2-7 Change in GDT-HA vs. time for all targets with 200 ns simulation with imposed restraints.

Chapter 3

Physics Based Protein Structure Refinement via Data Mining through Multiple Molecular Dynamics Trajectories and Structure Averaging

Vahid Mirjalili, Keenan Noyes, Michael Feig

Adapted from

Proteins, Structure, Function and Bioinformatics,
V82, p196-207, 2014

3.1 Abstract

We used molecular dynamics (MD) simulations for structure refinement of CASP10 targets. Refinement was achieved by selecting structures from the MD-based ensembles followed by structural averaging. The overall performance of this method in CASP10 is described and specific aspects are analyzed in detail to provide insight into key components. In particular, the use of different restraint sets, sampling from multiple short simulations vs. a single long simulation, the success of a quality assessment criterion, the application of scoring vs. averaging, and the impact of a final refinement step are discussed in detail.

3.2 Introduction

Two decades of CASP (Critical Assessment of Techniques for Protein Structure Prediction) have documented significant progress with predicting the structure of proteins from their amino acid sequences.[55, 98-102] This can be attributed to the development of new techniques but an increasing number of structures in the Protein Data Bank (PDB)[103] are at least an equally important factor.[104-107] The most reliable method for protein structure prediction is template based modeling.[102, 105, 108, 109] The resulting models are often overall correct, but deviate from experimental structures in detail with typical root mean square deviations (RMSD) of 2-6 Å due to intrinsic errors when constructing models based on template structures.[35, 36] Therefore, recent attention has shifted towards the refinement of template-based models to improve their accuracy and generate models that are suitable for biological and pharmaceutical studies.[57]

A variety of methods for the refinement of template-based models have been proposed, with the majority involving some combination of sampling and scoring with an emphasis on physics-based methods, such as molecular dynamics[65, 70, 110, 111]. At the same time, knowledge

based methods have also been proposed[77, 82, 112, 113]. The challenges with typical structure refinement protocols are two-fold: 1) Sampling has to progress at least in part towards the native structure; and 2) improved structures generated by the sampling method have to be reliably selected. In terms of sampling, different strategies have been explored. The application of restraints on some regions of the protein judged to be of higher quality than other regions often leads to improved sampling of refined structures.[70] Other strategies have involved enhanced sampling methods such as replica exchange MD simulation[114] and self-guided Langevin dynamics[83] as well as implicit and explicit solvent simulations.[111, 115] A key issue is the quality of the force field which ultimately determines whether refined structures are likely to be generated. In the past, force fields have been optimized specifically for refinement[67, 116], but improvements in general biomolecular force fields[117, 118] are expected to also impact the ability to carry out successful structure refinement.

While sampling methods are often able to generate refined models, these are typically not found at the end of a given sampling run but instead at intermediate time points. The challenge is then to find those refined structures from the ensemble of structures generated at the sampling stage. The force fields used for sampling, while physically accurate, are often too noisy to reliably identify single structures or small subsets of structures that are most native-like. Instead, a number of statistical potential functions have been used for scoring decoy structures, such as DFIRE,[119, 120] GOAP,[121] DOPE,[122] and OPUS-PSP[122]. All of these scoring functions have shown promise in selecting native-like structures from an ensemble, but struggle with consistently selecting refined structures.[37, 114, 123]

Despite considerable efforts, effective structure refinement protocols have remained elusive. During the last round of CASP, CASP9, there were only a few groups that were able to

outperform a naive prediction of simply resubmitting the initial model given by the organizers to be refined[35]. Furthermore, refinement progress was very modest and predictions from the most successful groups lacked consistency as some targets were refined significantly, while others were made worse. Further efforts since CASP9 include very long MD simulation by the D. E. Shaw group[70]. In that work, it was clearly shown that without restraints the initial models are likely to drift away from the native structure making refinement largely impossible. When restraints were applied, the sampling of refined structures became possible but the reliable selection of refined structures remained a significant obstacle. Overall, structures selected based on cluster size and/or energetic criteria were improved on average 1% in terms of GDT-TS. A similar level of performance was reported by Zhang et al.,[110] in which they combined knowledge-based information with physics-based MD simulations and applied a fragment-guided method with distance restraints used on global and local structural templates from the PDB. Gront et al.[124] recently provided a comprehensive review of refinement methods ranging from physics based to knowledge based methods and concluded that refinement is more challenging when starting structures are already within 2-3 Å from the native structure. In that paper it was also noted that knowledge-based methods may have an advantage because they are parameterized based on experimental structures which are the target of refinement protocols vs. physics-based methods that aim at capturing the protein dynamics at the global minima of the energy landscape.

The distinction between (simulation-generated) protein dynamics and experimentally-obtained structures may become increasingly important as refinement methods aim to reproduce experimental structures at high accuracy. One particular issue is that experimental structures reflect ensemble- and time-averaged conformations rather than instantaneous snapshots.

Following this idea, we have recently devised a structure refinement protocol that obtains refined structures from ensemble averages over selected subsets instead of single snapshots[37]. When this protocol was applied to ensembles from extensive MD-based sampling with the recently updated CHARMM36 force field in combination with explicit water, significant and consistent refinement became possible when tested on CASP8 and CASP9 targets. Here, we describe the blind application of such a refinement protocol during CASP10.

In the following we will first describe the methodology before presenting and discussing results obtained during CASP10 and from subsequent post-analysis.

3.3 Methods

The initial models from CASP10 were preprocessed by adding missing hydrogens using the HBUILD module in CHARMM.[86] Protonation states of His residues (if present), were determined by visual inspection. The pKa values of other titratable residues (Glu, Asp, Lys, Arg) were determined using the PROPKA web server[125, 126] followed by visual inspection. All proteins were subsequently solvated in a cubic box of water with at least 9 Å cutoff to the edge of the box. The systems were neutralized by adding Na^+ or Cl^- to balance the net charge of the systems.

The solvated systems were then subjected to molecular dynamics (MD) simulations with periodic boundary conditions. The non-bonded interactions were cut off using the switching method between 8.5 to 10 Å, along with particle-mesh Ewald (PME) summation using a grid spacing of 1 Å for long range electrostatic interactions. The simulations were performed under NPT condition using Langevin dynamics at a temperature of 298 K with a Langevin piston to maintain constant pressure at 1 bar. A time step of 2 fs was used with the SHAKE algorithm to

fix bonds involving hydrogen atom. The CHARMM36 force field[118] was used to model the proteins in conjunction with the TIP3 water model[88].

All of the simulations used some form of restraints. Two types of restraints were used for almost all of the targets; type 1 consisted of weak restraints (with a force constant of 0.05 kcal/mol/Å^2) applied to all C_α atoms; type 2 involved strong restraints (with a force-constant of 1 kcal/mol/Å^2) applied to C_α atoms of only the regions that were assumed to be reliable in the starting model. For targets, where CASP organizers indicated which regions to refine, we followed their suggestions. In other cases, we assumed that secondary structure elements are likely to be more reliable and applied restraints to those while leaving loops flexible. Table 3-1 shows the regions which were selected for the strong restraints. In some cases, a combination of weak and strong restraints was used by applying strong restraints on selected residues but weak restraints on the rest. Due to the presence of zinc ions in TR754, the first set was modeled with weak restraints on all C_αs except for the region around the zinc fingers.

The heating and equilibration protocol involved 10 stages: First, simulations were carried out at 50 K using C_α restraints according to Table 3-1 with a force constant of 2 kcal/mol/Å^2 and a force constant of 0.5 kcal/mol/Å^2 for all other C_α atoms. The temperatures and force constants were subsequently increased/decreased in 10 ps steps to (100 K, 2/0.5 kcal/mol/Å^2), (200 K, 2/0.5 kcal/mol/Å^2) (200 K, 1.5/0.2 kcal/mol/Å^2), (200 K, 1/0.1 kcal/mol/Å^2), (200 K, 1/0.05 kcal/mol/Å^2), (250 K, 1/0.05 kcal/mol/Å^2), (298 K, 1/.0.05 kcal/mol/Å^2), (298 K, 1/0.01 kcal/mol/Å^2), and (298 K, 1/0 kcal/mol/Å^2). The structure at the end of the final stage was used as the starting point for all of the production runs.

Production simulations consisted of 10 or 20 replicate MD simulations, each 20 ns long and starting from the same starting structure using either strong or weak restraints (see Table 3-1). We ran multiple short simulations instead of a single long simulation to maximize sampling given limited availability of computer resources.[95] During post-analysis we also carried out single long simulations (200 ns) using the restraint types listed as set 1 (Table 3-1).

Ensembles of structures were generated from the simulations, containing 500 snapshots for each of the MD trajectories. Structures in each replica ensemble were analyzed in terms of the RMSD from the initial model (iRMSD) and their DFIRE scores.

Following our previously established protocol[37] (see Fig. 3-1), we began by using the correlation coefficient between iRMSD vs. DFIRE as a quality assessment score. Replicas with correlation coefficients greater than 0.4 were discarded from subsequent analyses. From the remaining replicas a subset of structures with combined minimal iRMSD and DFIRE scores[37] were then selected. Briefly, the selection criteria is based on normalized iRMSD and DFIRE scores to be within an angle $\theta/2$ around the identity line and outside a circle of radius ρ from the center of the distribution, corresponding to the lower left corner of the scatter plot of iRMSD vs. DFIRE.[37] The criterion used in CASP experiment was, however, slightly different than what was used for testing the protocol on CASP8 and CASP9 targets because of additional optimization.[37] Here, we used $\rho=1$, and $\theta=100^{\circ}$. An average structure was then calculated from the selected subset of structures followed by a structure interpolation. This was accomplished by taking the point on the vector between the corresponding C_α atoms in the average and the initial model, with its distance to the initial model to be 0.55 of the vector length. The coordinates of all other atoms were copied from the initial model.

Table 3-1 Type of restraints applied on C_α atoms for the two simulation sets; strong (1 kCal/mol/Å^2), weak (0.05 kCal/mol/Å^2) and a combination of both. The strong force constants are only applied to the selected regions.

Target	Set 1: 20×20 ns	Set 2: 10×20 ns	Strongly Restrained Regions
TR644	Combined	Strong	53:56, 61:66, 71:75, 84:87, 115:119, 129:132, 140:142, 151:153
TR655	Strong	Weak	21:50, 65:90, 94:141, 164:180
TR661	Weak	-	-
TR662	Weak	Strong	5:16, 38:50, 66:79
TR663	Combined	Strong	79:140, 182:204
TR671	Strong	Weak	38:54, 77:80, 85:89, 96, 108:125
TR674	Weak	Strong	284:288, 300:305, 310:312, 318:320, 333:335
TR679	Strong	Weak	1:24, 46:145, 157:186, 198:223
TR681	Strong	Weak	21:40, 51:57, 65:87, 102:118, 128:144, 153:157, 171:172, 200:224
TR688	Combined	Strong	46:54, 67:76, 89:98, 113:122, 137:145, 160:167, 182:190
TR689	Strong	Weak	14:21, 33:39, 48:59, 64:72, 81:89, 116:118, 143:147, 156:160, 165:169, 181:190, 197:207, 211:218, 226:234
TR696	Weak	Strong	18:22, 27:35, 41:43, 50:51, 58:60, 69:73, 93:96, 101:105
TR698	Strong	Weak	1:16, 36:89, 101:119
TR699	Weak	Strong	8:11, 37:45, 53:55, 86:94, 103:135, 161, 205:206, 219:234
TR704	Weak	Strong	25:32, 40:42, 50:55, 61:64, 81:83, 100:102, 113:120, 128:132, 141:149, 161:166, 188:189, 193:200, 204:209, 217:226, 236:237, 242:246
TR705	Weak	Strong	40:42, 65:67, 82:85, 90:91, 110:114, 119:126
TR708	Weak	Strong	24:27, 45:60, 66:70, 99:101, 113:119, 125:129, 136:152, 172:183
TR710	Weak	Strong	27:50, 67:83, 100:117, 135:152, 168:185, 201:220
TR712	Strong	Weak	38:79, 90:115, 130:140, 156:223
TR720	Weak	Strong	27:29, 53:71, 80:86, 91:103, 108:114, 127:139, 144:147, 154:157, 162:176
TR723	Strong	Weak	39:73, 99:112
TR724	Weak	Strong	135:136, 152:157, 198:202, 210:216, 232:238
TR738	Strong	Weak	1:38, 88:90, 103:249
TR747	Weak	Strong	24:26, 46:49, 55:59, 68:71, 80:83, 92:94, 103:109, 114:121
TR750	Weak	Strong	1:6, 28:29, 48:57, 64:66, 78:93, 98:100, 121:137, 168:182
TR752	Strong	Combined	1:40, 51:99, 111:124, 129:156
TR754	Weak	Strong	25, 33, 63:76 (weak restraints are not applied to the zinc fingers)

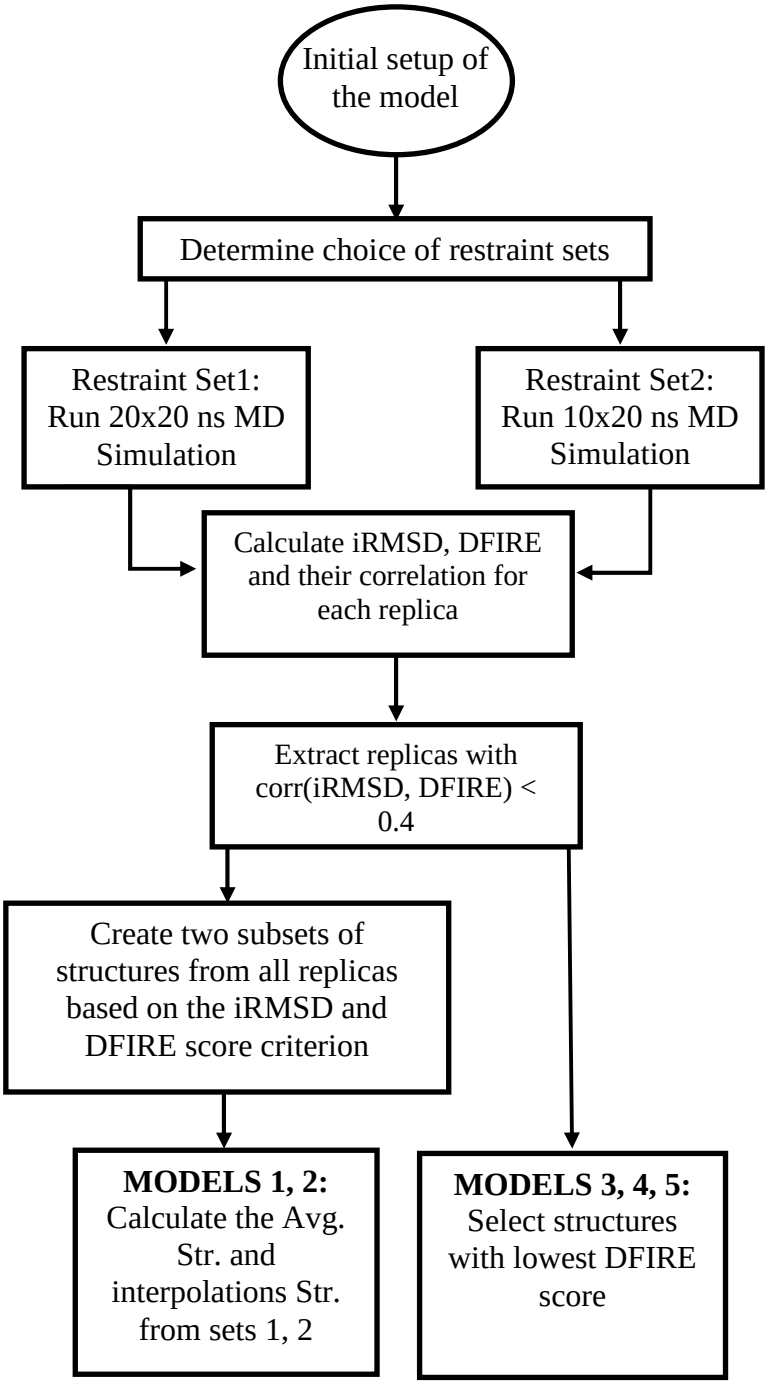

Figure 3-1 Flowchart of the refinement protocol, from simulation to model selection

The resulting structure was then solvated again, neutralized by adding appropriate charges, and subjected to 5,000 energy minimization steps followed by 40 ps of MD simulation at 100 K with restraints on all C_αs and a force constant of 100 kcal/mol/Å^2. The purpose of the final MD simulations was to relax structural artifacts due to the averaging procedure and generate structures that are of high stereochemical quality.

The application of the above protocol to simulations with restraint sets 1 and 2 resulted in models 1 and 2 submitted to CASP. Models 3-5 were selected from the trajectory snapshots with low DFIRE score but outside the region of the scatter plot used for averaging with the idea that some of these structures may be refined more extensively compared to models 1 and 2.

All of the molecular dynamics simulations were carried out with the NAMD molecular dynamics package in conjunction with the MMTSB tool set[92] which was also used for analysis along with custom scripts. The protein structures were visualized via the PyMol molecular visualization software.[127, 128]

3.4 Results

The MD-based refinement protocol described in the methods section was applied to 27 CASP10 refinement targets. The protocol was not applied to one target, TR722, which was modeled unsuccessfully using an entirely different procedure.

Table 3-2 Refinement results showing the best observed structures in the trajectories, the first submitted model, the best of five submitted models, and best model among models 3-5.

Target	Best in 30x20ns trajectories		First submitted model		Best of five models		Best of models 3-5	
	ΔRMSD (Å)	ΔGDT-HA	ΔRMSD (Å)	ΔGDT-HA	ΔRMSD (Å)	ΔGDT-HA	ΔRMSD (Å)	ΔGDT-HA
TR644	**-0.94**	**11.0**	**-0.03**	**2.8**	**-0.55**	**5.3**	0.04	-1.4
TR655	**-0.26**	**2.4**	0.04	**0.3**	0.00	**0.3**	0.20	-0.9
TR661	**-0.25**	**6.1**	**-0.03**	**1.9**	**-0.03**	**1.9**	0.17	-2.2
TR662	**-0.54**	**13.0**	**-0.20**	**5.3**	**-0.25**	**6.7**	**-0.25**	**6.7**
TR663	**-0.41**	**4.6**	**-0.12**	**2.6**	**-0.15**	**3.6**	**-0.15**	**3.3**
TR671	**-0.62**	**5.4**	**-0.01**	**0.6**	**-0.25**	**2.8**	0.09	**2.8**
TR674	**-0.78**	**7.0**	0.00	**4.9**	**-0.06**	**4.9**	**-0.06**	-3.4
TR679	**-0.55**	**4.8**	0.01	**0.6**	**-0.03**	**3.3**	0.12	**1.0**
TR681	**-0.13**	**5.2**	**-0.04**	**1.1**	**-0.15**	**5.4**	**-0.15**	**5.4**
TR688	**-0.14**	**6.9**	0.01	**1.5**	**-0.02**	**2.2**	0.02	-0.1
TR689	**-0.25**	**2.3**	**-0.10**	**3.5**	**-0.13**	**4.9**	**-0.12**	**2.3**
TR696	**-0.81**	**11.0**	**-0.13**	**3.5**	**-0.33**	**4.8**	**-0.33**	**4.8**
TR698	**-0.32**	**3.6**	**-0.02**	-0.4	**-0.02**	**-0.4**	0.09	-0.6
TR699	**-0.33**	**4.1**	**-0.09**	**4.6**	**-0.09**	**4.6**	**-0.07**	**3.7**
TR704	**-0.57**	**7.8**	**-0.17**	**3.9**	**-0.23**	**5.6**	**-0.23**	**5.6**
TR705	**-0.51**	**10.7**	**-0.14**	**6.0**	**-0.24**	**7.3**	**-0.24**	**7.3**
TR708	**-0.84**	**1.3**	0.09	**2.7**	0.09	**2.9**	0.10	-2.4
TR710	**-0.20**	**11.1**	**-0.04**	**4.3**	**-0.06**	**4.3**	**-0.04**	**2.1**
TR712	**-0.54**	**3.1**	**-0.08**	**3.4**	**-0.14**	**5.0**	**-0.14**	**5.0**
TR720	**-1.85**	**5.1**	0.02	**2.7**	**-0.99**	**3.2**	**-0.99**	**1.1**
TR723	**-0.71**	**11.8**	**-0.13**	**6.5**	**-0.39**	**9.7**	**-0.39**	**9.7**
TR724	**-1.49**	**8.5**	**-0.01**	**2.6**	**-0.48**	**3.7**	--	--
TR738	**-0.37**	**10.6**	**-0.20**	**6.0**	**-0.30**	**9.5**	**-0.09**	**3.5**
TR747	**-0.44**	**13.1**	**-0.10**	**0.8**	**-0.10**	**0.8**	0.10	-0.6
TR750	**-0.43**	**11.8**	**-0.16**	**4.8**	**-0.16**	**4.8**	**-0.04**	**2.5**
TR752	**-0.30**	**3.1**	**-0.12**	**1.4**	**-0.12**	**1.4**	**-0.05**	-0.7
TR754	**-0.35**	**2.6**	0.09	-6.3	0.09	-6.3	0.09	-7.4
Avg.	**-0.55**	**7.0**	**-0.06**	**2.6**	**-0.19**	**3.8**	**-0.09**	**1.8**

3.4.1 Overall CASP10 Performance

Five models were submitted for each of the targets. The first and second models resulted from ensemble averaging. The other models were selected based on favorable DFIRE scores (see methods section). Table 3-2 shows the changes upon refinement, ΔRMSD and ΔGDT-HA, with respect to the initial models provided by CASP for the first submitted model and the best of all five models, respectively. The average change in RMSD for the first models is -0.06 Å, and the

average change in GDT-HA is 2.6. More importantly, 20 out of 27 targets improved in RMSD, and 25 targets improved in terms of GDT-HA. This performance is similar to what we found previously when testing the protocol on CASP8 and CASP9 targets.[37] When selecting the best out of five structures, the average improvement in terms of RMSD is -0.19 Å, and in terms of GDT-HA is 3.8. Looking at all five models, 24 targets are improved with respect to RMSD, and all targets except for TR754 are improved in GDT-HA. The overall best refinement case has an RMSD value that is improved by almost 1 Å (TR720) and GDT-HA improvements by nearly 10 units (TR723 and TR738). These results suggest that with this refinement protocol it is possible to consistently generate significantly refined structures from the initial template-based models. The only target where the predicted structure was significantly worse than the starting structure was TR754, where the presence of zinc ions presumably complicated the scoring with DFIRE.

Furthermore, Table 3-2 lists the best observed structures in terms of RMSD and GDT-HA throughout all 30×20 ns trajectories. Because the best cases were not necessarily picked out for submission, this information provides a theoretical limit of how much refinement could have been achieved with a perfect scoring function. Significant refinement of 1.85 Å in TR720 is observed, as well as several cases with improvements in GDT-HA higher than 10%. On average, improvement in RMSD is 0.55 and 7.0 for GDT-HA. On the other hand looking at the best of five models, we see that the RMSD and GDT-HA are improved by 34% and 69% of the maximum possible improvements, respectively. To our knowledge, no single-structure selection protocol can achieve such a result. Interestingly, there are a few cases where the refined structures are actually better (in terms of GDT-HA) than the best single structure from the trajectories (TR689, TR699, TR708, TR712). This indicates that the averaging procedure used here leads to additional refinement over just selecting the best structure from a given ensemble.

Figure 3-2 shows four of the best modeled targets, TR662, TR674, TR723, and TR738. While most of the secondary structure elements are fixed, some of the loop regions were refined to conformations intermediate between the initial model and the experimental reference. This suggests that refinement is proceeding towards the right direction but it is clear that further progress is needed to fully reach experimental accuracy.

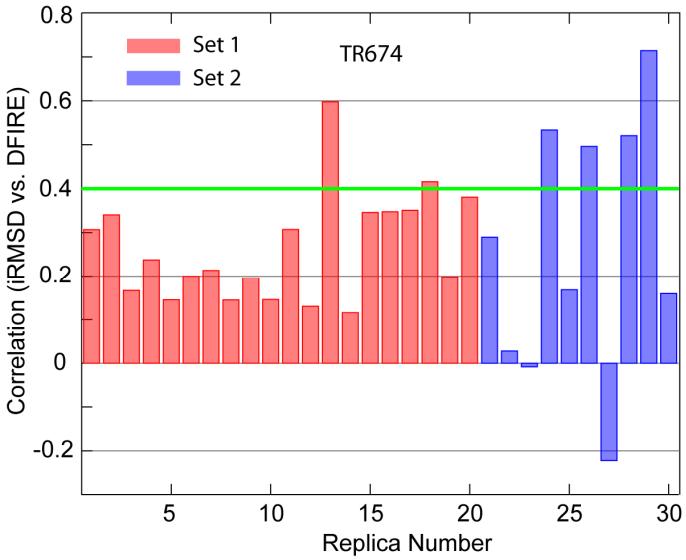

Figure 3-2 Correlation of iRMSD vs. DFIRe scores of individual replicas for TR674; Set 1 (replicas 1:20) is shown with red boxes and set 2 (replicas 21:30) with blue. Replicas with corr>0.4 are discarded for model selection.

3.4.2 Model Selection based on Lowest DFIRE and Highest iRMSD

For models 3-5, we selected structures with the lowest DFIRE score and higher iRMSD values. The rationale was that ideally the lowest DFIRE scores would identify the most native structures while higher iRMSD values would allow for more significant refinement but also risks larger deviations away from the native. This is in contrast to the more conservative criterion used for the subset ensemble selection based on small iRMSD values. Table 3-2 shows the best structures among the submitted models 3-5. It can be seen that while there are indeed some cases with

significantly refined structures (TR720, TR723) and overall average improvement in both RMSD and GDT-HA, there are also many cases where no refinement was achieved. Although the best of three models were analyzed here, the results remain inferior to the single model 1 obtained from ensemble averaging.

3.4.3 Quality Assessment using Correlation between iRMSD and DFIRE

One aspect of our refinement protocol is to estimate whether a given set of samples likely includes significantly refined structures. As discussed in more detail in our previous paper,[37] we identified the correlation coefficient between DFIRE and iRMSD as a suitable metric. Correlation coefficients above 0.4 appeared to be correlated with poor refinement performance;[37] we applied this criterion here to discard trajectories where this condition was satisfied from further analysis. To further assess the validity of this assumption, we compare in Table 3-3 the fractions of improved structures in terms of RMSD and GDT-HA for replicas where the correlation is less than 0.4, with those where the correlation is greater than or equal 0.4. While the results vary greatly for individual targets, there is on average a modest enrichment in terms of both RMSD and GDT-HA, both by about 6%, when discarding samples from replicas where the correlation coefficient is above 0.4. This suggests that the quality assessment procedure used here adds value and it could be used in the future to guide the generation of additional trajectories for cases where refinement appears to be difficult as suggested by many replicas with correlation coefficients above the 0.4 threshold.

Table 3-3 Fraction of improved trajectory frames in replicas classified by correlation between iRMSD and DFIRE score; Fraction of improved frames in trajectories with correlation <0.4 that are larger by 10% than fractions with correlation≥0.4 are highlighted.

Target	Fraction (%) of traj. frames improved in RMSD		# replicas with Corr≥0.4	Fraction (%) of traj. frames improved in GDT-HA	
	Corr < 0.4	Corr ≥ 0.4		Corr < 0.4	Corr ≥ 0.4
TR644	46.5	N.A.	0	14.1	N.A.
TR655	4.4	3.7	16	2.3	0.5
TR661	12.6	N.A.	0	8.5	N.A.
TR662	**70.4**	8.4	6	**76.1**	33.8
TR663	28.0	22.7	15	49.0	41.0
TR671	**18.7**	5.4	12	**13.2**	3.0
TR674	**23.1**	5.2	6	9.3	1.3
TR679	15.3	35.9	3	16.7	30.7
TR681	0.4	0.0	2	2.2	0.1
TR688	1.7	0.5	3	19.8	17.3
TR689	**33.4**	10.5	7	0.7	0.3
TR696	**55.1**	33.8	1	**33.6**	15.6
TR698	60.3	59.7	4	13.6	13.2
TR699	29.2	26.8	1	1.6	0.2
TR704	53.2	NA	0	37.9	N.A.
TR705	**33.7**	3.0	4	**55.3**	20.5
TR708	3.4	9.8	6	0.2	0.1
TR710	26.3	NA	0	79.4	N.A.
TR712	28.4	20.0	1	5.5	0.6
TR720	27.8	39.2	8	15.6	32.1
TR723	57.4	68.7	3	55.7	47.0
TR724	37.6	99.6	1	22.9	60.8
TR738	85.7	99.9	2	78.8	99.8
TR747	**38.2**	8.0	5	**20.8**	5.7
TR750	**54.3**	22.4	6	**76.6**	36.2
TR752	**35.7**	13.9	6	6.1	0.7
TR754	1.0	0.5	22	0.1	0.1
Avg.	32.7	26.0		26.5	20.0

3.4.4 Restraint Choice

We used different restraints out of the following three choices: 1) strong restraints on selected C_αs; 2) weak restraints on all C_αs; 3) a combination of strong restraints on selected regions and weak restraints on the rest. The first choice is most appropriate for cases where there is specific information about which regions require refinement. In the case of CASP, this information was provided for some targets. However, for other targets – and more general applications of

structure refinement methods – such information may not be available. Therefore, we evaluated how the choice of restraints affected the results. In order to compare results in a consistent fashion, we used only the first 10 replicas of each set. Some targets use strong, partial restraints for the first set with 20 replicas while for other targets weak, complete restraints were used for the first set (see Table 3-1). Therefore, the total number of replicas that were used for each restraint type does not match among different targets. Furthermore, not all targets were run with strong, partial and weak, complete restraints. Those targets were excluded from the comparison (see Table 3-4).

Table 3-4 Refinement results of different restraints for the best observed structure in terms of RMSD (Å) and GDT-HA; comparing strong (1 kCal/mol/Å^2) restraint on selected residues vs. weak (0.05 kCal/mol/Å^2) restraint on all C_αs, and a combination of both. Cases associated with * indicate targets that had suggestions from CASP on which regions need refinement.

Target	Strong Restraint		Weak on all C_α		Strong + weak	
	ΔRMSD (Å)	ΔGDT-HA	ΔRMSD (Å)	ΔGDT-HA	ΔRMSD (Å)	ΔGDT-HA
TR644	**-0.94**	**5.1**	--	--	**-0.49**	**4.4**
TR655 *	**-0.26**	**1.3**	**-0.14**	-0.3		
TR661	--	--	**-0.22**	**3.9**		
TR662	**-0.22**	**5.0**	**-0.54**	**10.7**		
TR663 *	**-0.22**	**3.6**	--	--	**-0.34**	**3.6**
TR671	**-0.21**	**1.7**	**-0.62**	**4.3**		
TR674	**-0.78**	**3.8**	**-0.40**	**4.7**		
TR679 *	**-0.55**	**2.9**	**-0.23**	**2.4**		
TR681	**-0.02**	**1.6**	**-0.13**	**3.5**		
TR688	**-0.09**	**4.1**	--	--	**-0.13**	**5.0**
TR689	**-0.15**	-0.2	**-0.24**	**0.9**		
TR696	**-0.81**	**6.0**	**-0.50**	**7.8**		
TR698 *	**-0.32**	**2.3**	**-0.16**	-1.5		
TR699	**-0.32**	**1.2**	**-0.33**	**1.7**		
TR704	**-0.37**	**4.5**	**-0.57**	**6.6**		
TR705	**-0.48**	**8.3**	**-0.45**	**5.7**		
TR708	**-0.84**	**0.0**	**-0.43**	-0.1		
TR710	**-0.16**	**6.7**	**-0.19**	**8.8**		
TR712 *	**-0.48**	**2.3**	**-0.23**	-1.9		
TR720	**-1.85**	**2.8**	**-0.21**	**2.8**		
TR723	**-0.71**	**9.7**	**-0.35**	**7.8**		
TR724	**-1.49**	**7.4**	**-0.40**	**3.1**		
TR738 *	**-0.34**	**8.5**	**-0.37**	**6.3**		
TR747	**-0.19**	**3.3**	**-0.44**	**9.2**		
TR750	**-0.26**	**4.5**	**-0.42**	**10.2**		
TR752 *	**-0.30**	**2.5**	--	--	**-0.24**	**1.0**
TR754	**-0.21**	-0.7	**-0.35**	**2.57**		
Avg. of common rows:	**-0.50**	**3.8**	**-0.35**	**4.3**	Not enough data	Not enough data
Avg. * (CASP sugg.)	**-0.35**	**3.3**	**-0.23**	**1**		
Avg. (no sugg.)	**-0.53**	**3.9**	**-0.38**	**5.2**		

Table 3-4 shows the best structures in terms of RMSD (Å) and GDT-HA from all the 10 replicas for a given restraint type. Average values were calculated only for the 22 targets, which have

both strong and weak restraint sets. The analysis suggests that in terms of best structures that were generated, strong, partial restraints may be roughly equivalent to using weak, complete restraints. Interestingly, the RMSD seems to be improved more with strong, partial restraints while GDT-HA scores appear to be improved more with weak, complete restraints. It is instructive to further separate the analysis into targets where the CASP organizers suggested regions to be refined vs. targets where no such information was given. We find that the degree of refinement was actually greater for the targets where no information was given, indicating that the additional information given during CASP10 was not essential for successful refinement. However, we also note that in the cases where information was available about which regions to refine, the application of partial restraints clearly outperformed weak restraints on all residues. On the other hand, targets where no information was given resulted in significantly better GDT-HA scores with weak, overall restraints than with partial restraints based on secondary structures. This suggests that an optimal strategy may be to use partial restraints if information is available which regions require refinement while applying weak restraints for all residues otherwise. While Table 3-4 focuses on the best structures that are generated, Table 3-5 shows the result of refinement when the entire protocol is applied. The overall trends match those of Table 3-4.

Table 3-5 Refinement results of different restraints using the established structure generation protocol; comparing strong (1 kCal/mol/Å^2) on selected residues vs. weak (0.05 kCal/mol/Å^2) restraint on all C_α, and a combination of both

Target	Strong restraint		Weak on all C_α		Strong + weak	
	ΔRMSD (Å)	ΔGDT-HA	ΔRMSD (Å)	ΔGDT-HA	ΔRMSD (Å)	ΔGDT-HA
TR644	**-0.54**	**3.0**	--	--	**-0.36**	**3.2**
TR655 *	**-0.05**	0.0	**-0.01**	-1.7		
TR661	--	--	**-0.03**	**2.3**		
TR662	**-0.03**	**1.3**	**-0.20**	**4.7**		
TR663 *	0.54	**3.1**	--	--	**-0.11**	**2.8**
TR671	0.02	**1.1**	**-0.05**	**1.4**		
TR674	**-0.11**	**3.4**	0.00	**5.3**		
TR679 *	0.03	**0.1**	**-0.05**	**3.0**		
TR681	**-0.06**	-0.1	**-0.04**	**1.3**		
TR688	**-0.02**	**2.3**	--	--	0.01	**1.4**
TR689	**-0.11**	**3.6**	**-0.14**	**4.4**		
TR696	**-0.24**	**2.8**	**-0.17**	**4.0**		
TR698 *	**-0.01**	-0.4	0.03	-1.3		
TR699	**-0.14**	**4.0**	**-0.04**	**4.2**		
TR704	**-0.10**	**2.3**	**-0.18**	**3.9**		
TR705	**-0.15**	**5.2**	**-0.14**	**4.4**		
TR708	0.11	**2.8**	0.08	**2.2**		
TR710	**-0.05**	**2.6**	**-0.05**	**4.4**		
TR712 *	**-0.08**	**3.5**	**-0.07**	**4.3**		
TR720	**-0.54**	**1.1**	0.01	**2.4**		
TR723	**-0.22**	**4.6**	**-0.13**	**6.3**		
TR724	**-0.51**	**3.7**	**-0.02**	**2.8**		
TR738 *	**-0.20**	**6.1**	**-0.23**	**7.5**		
TR747	**-0.08**	**0.3**	**-0.11**	**0.8**		
TR750	**-0.11**	**2.9**	**-0.16**	**4.0**		
TR752 *	**-0.13**	**1.7**	--	--	**-0.10**	**1.0**
TR754	0.16	-7.4	0.07	-5.5		
Avg. of common rows:	**-0.11**	**2.0**	**-0.07**	**2.9**	***Not enough data***	***Not enough data***
Avg. * (CASP sugg.)	0.01	**2.0**	**-0.07**	**2.4**		
Avg. (no sugg.)	**-0.14**	**2.1**	**-0.08**	**3.5**		

3.4.5 Simulation Time: Single MD vs. Multiple Short MDs

Finally, we compared the sampling efficiency of two sets of simulations in order to assess the benefits of using multiple short simulations vs. a single long MD simulation. During CASP10 we ran multiple short simulations because of time and resource constraints. A single long simulation was run after completion of CASP for over 200 ns for each target continued from the first replica in set 1 using the same restraints as for the short simulations. We then compared the output of the 10×20 ns simulations with the results from the single 200 ns simulations. Note that the restraints can be either of strong, partial type, or weak, complete type. Figure 3-3 shows the cumulative minimum ΔRMSD and cumulative maximum ΔGDT-HA averaged over all 27 targets for the single 200 ns simulations and 10×20 ns simulations. The results of the 10 replicas in multiple short simulations are combined at each time slot, so at each time t, the cumulative minimum ΔRMSD and maximum ΔGDT-HA values are calculated from the t/10 portion of all of the 10 trajectories. There is an expanding gap between the single and multiple trajectories where the multiple short simulations outperform the long simulation both in terms of RMSD and GDT-HA.

Furthermore, in Table 3-6 we compare the refinement performance by using structures either from a single 200 ns simulation or from multiple 10×20 ns simulations. We tested two selection protocols, using the lowest DFIRE score and subset selection and averaging followed by structure interpolation as described above. Selecting structures with the lowest DFIRE score performs poorly in both cases. However, applying our protocol improves the average RMSD with a similar level of accuracy (-0.08 Å), while the average improvement in GDT-HA is actually slightly higher in the case of the single 200 ns simulations (2.9 vs. 2.5 for 10×20 ns simulations). Given the increased sampling of refined structures with multiple short simulations, this is somewhat surprising and warrants further investigation. Assuming that the differences are

statistically significant, it may be that much longer simulations generate a broader sampling that when averaged result in a structure that is closer to the experimentally averaged structures.

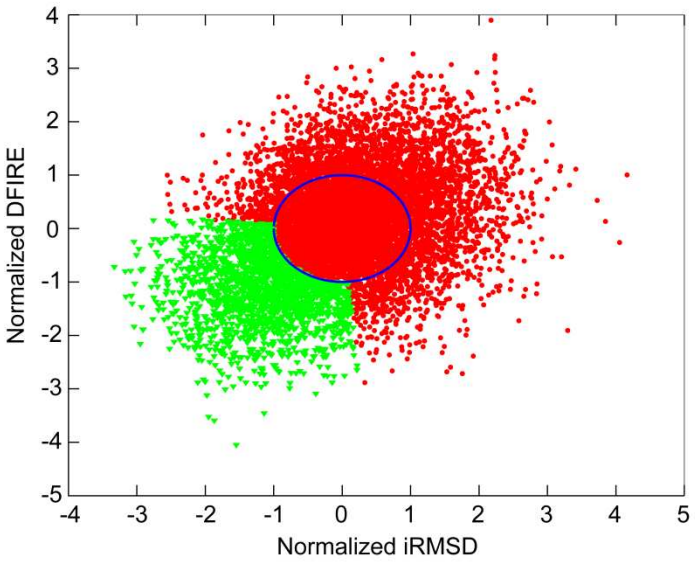

Figure 3-3 Subset selection based on normalized iRMSD and DFIRE scores from the replicas in set 1 with corr(iRMSD,DFIRE)<0.4 for TR674. The subset shown in the lower left corner of the scatter-plot with green triangles are selected to calculate the average structure and model selection.

Table 3-6 Summary of comparing refinement results between 10×20 ns simulations and single 200 ns simulations having the same restraint conditions. The results are averaged over 27 CASP10 targets.

		200 ns	10×20 ns
Best ΔRMSD in trajectory		**-0.29**	**-0.34**
Best ΔGDT-HA in trajectory		**5.5**	**6.0**
Lowest DFIRE score	ΔRMSD	0.06	0.03
	ΔGDT-HA	0.0	-0.1
Subset Avg. + Str. Interp.	ΔRMSD	**-0.08**	**-0.08**
	ΔGDT-HA	**2.9**	**2.5**

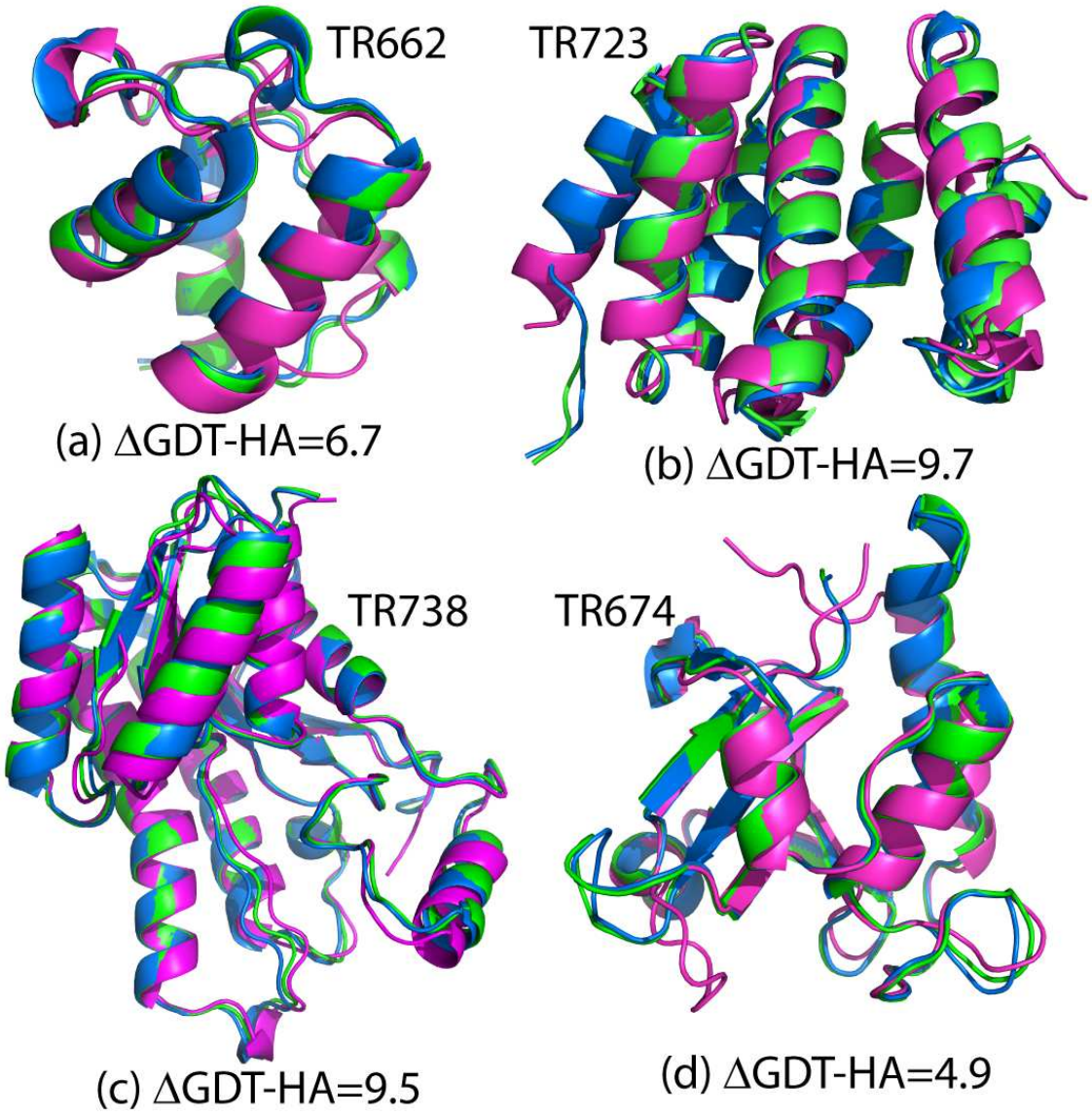

Figure 3-4 Initial model (blue), refined (green) and native (magenta) for (a) TR662, (b) TR723, (c) TR738 and (d) TR674

3.4.6 Final Stage of Refinement

As mentioned in the methods section, structure averaging and interpolation cause some unphysical conformation with bad bonds, angles, dihedrals and steric clashes. Therefore, an extra stage of refinement is required to generate stereochemically acceptable structures. Table 3-7 shows the MolProbity measures for individual targets before (avg) and after final refinement stage (MD). This final refinement had only a small effect on RMSD from native and GDT-HA,

as before this stage, the average change in RMSD was -0.08 Å before the final stage and -0.06 Å after the final stage while the average GDT-HA did not change during the final refinement stage.

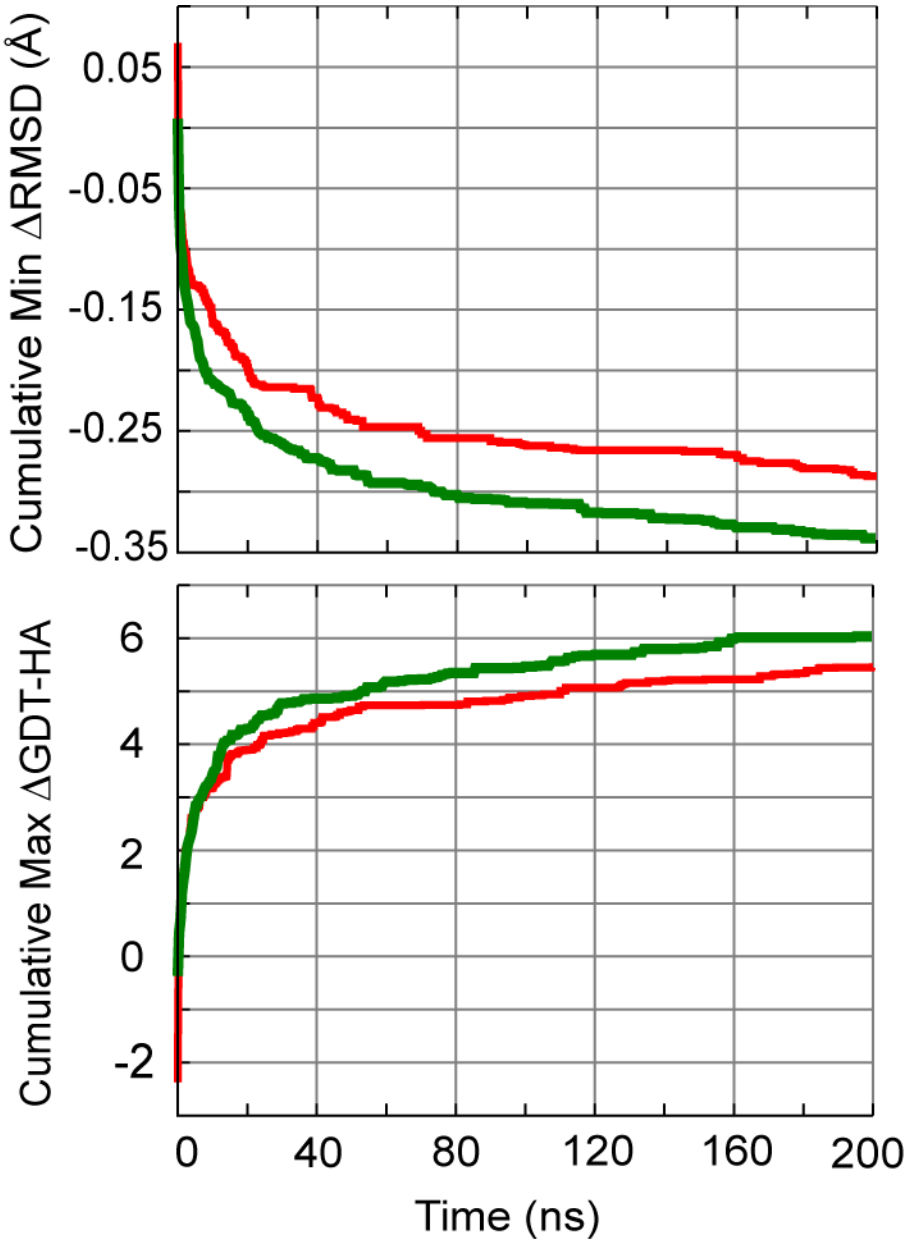

Figure 3-5 Sampling efficiency toward native, showing the cumulative minimum ΔRMSD (top) and cumulative maximum ΔGDT-HA (bottom), comparing the single 200 ns MD simulations (red) vs. 10×20 ns MD simulations (green) averaged over 27 CASP10 targets.

Table 3-7 MolProbity results of the structure obtained from the averaging and structure interpolation

Target	MolProbity	
	Avg.	Final
TR644	4.16	**1.42**
TR655	4.83	2.33
TR661	4.00	**1.03**
TR662	4.48	**1.49**
TR663	4.80	2.51
TR671	4.69	2.52
TR674	4.34	**1.99**
TR679	3.54	**1.74**
TR681	4.45	**1.74**
TR688	4.09	**1.64**
TR689	4.26	2.00
TR696	4.82	2.14
TR698	3.88	**1.47**
TR699	4.34	2.10
TR704	4.72	**1.32**
TR705	5.07	2.36
TR708	4.11	**1.44**
TR710	4.06	**1.14**
TR712	3.79	**1.82**
TR720	4.60	**1.28**
TR723	4.44	**1.80**
TR724	4.87	**1.90**
TR738	3.88	**0.88**
TR747	3.91	**1.12**
TR750	4.12	**1.61**
TR752	3.02	**1.05**
TR754	5.15	2.59
Avg.	4.31	**1.72**

3.5 Conclusion

We applied a recently established molecular dynamics-based structure refinement protocol to CASP10 targets. Overall, we were able to reliably refine most of the targets both in terms of RMSD and GDT-HA relative to the experimental structures. The key components of our protocol are the use of restraints during MD simulations, the selection of trajectories based on a

quality assessment score, and the generation of refined structures following structure subset selection and averaging.

We compared the results of using strong restraints on selected residues vs. weak restraints on all C_αs, and concluded that using strong restraints on selected C_αs leads to improved RMSD values, while weak restraints can improve GDT-HA measures better.

Another question in MD based refinement is the time scale of the simulation, and in this study we compared the sampling in multiple short MD simulations vs. one single long simulation, and we observed that multiple short MD simulations may have a higher sampling efficiency.

Although our protocol outperformed other refinement methods, overall, the improvements in RMSD and GDT-HA measures in refining protein structures are still relatively minor and it is clear that further progress is needed. One possibility is to take advantage of the consistent and reliable refinement obtained here and extrapolate along the initial direction. Another direction is the further improvement of structure selection methods since for many targets significantly more refined structures were generated than what we submitted as predictions.

Chapter 4

Protein Structure Refinement on CASP11 Targets

4.1 Introduction

Previously, we developed a robust protocol for MD-based protein structure refinement.[37] As our protocol was tested on CASP10 targets, 23 out of 27 cases we had improvements in ΔGDT-HA by taking advantage of subset selection algorithm and structure averaging.[38] Our work has remained state of the art solution to protein structure refinement to date. However, the improvements were still minor, reaching only up to 3% in ΔGDT-HA on average, and maximum improvement per target was limited to 5.5%. In order to make MD-based structure refinement a practical approach for structure determination procedure, improvements in ΔGDT-HA of about 20% is required. Therefore, we seek the limitations and bottlenecks in our approach.

Computational methods for structure refinement of proteins rely on performance of two aspects: conformational sampling and structure selection.[37] Therefore, improvements in both categories are necessary. Our analysis showed that restraints impose a strong limit on conformational sampling. We realized that using weaker restraints on all C_α atoms for targets in CASP10 gave better or nearly the same performance in refinement compared to strong restraint on selected C_α atoms.[38] Therefore, we modified our protocol and used weaker restraints in MD simulations. Applying weak restraints will greatly enhance the sampling efficiency, since the protein backbone has more freedom to respond to its environment. Yet, we also investigated the structure selection category, by comparing the performance of our protocol using different scoring functions. Many scoring functions were considered, such as ITScore[129], RW+[48], DFIRE[46], GOAP[130], OPUS-PSP[131], DOPE[132], and Seder[47]. The final protocol was refined and optimized based on its performance on CASP8, CASP9, and CASP10 targets.

4.2 Methods

The refinement category of CASP11 provided 37 protein targets, ranging from 62 to 288 residues. Table 4.1 shows detailed information of targets, such as number of residues, initial GDT-HA, and suggestions from CASP on which regions need further refinement.

Table 4-1 List of refinement targets in CASP11

Target	Number of residues	PDB	Initial GDT-HA	CASP Suggestions on regions to refine
TR217	224	4WED	62.8	
TR228	84			
TR274	194	4QB7	29.0	
TR280	96	4QDY	59.9	
TR283	168	4CVH	41.2	
TR759	62	4Q28	44.3	
TR760	201	4PQX	57.3	
TR762	257	4Q5T	70.6	
TR765	76	4PWU	57.9	
TR768	143	4OJU	64.0	
TR769	97	2MQ8	56.2	
TR772	198	4QHZ	52.4	
TR774	155	4QB7	37.8	
TR776	219	4Q9A	62.8	
TR780	95	4QDY	54.0	
TR782	110	4QRL	64.8	
TR783	243	4CVH	57.5	
TR786	217	4QVU	47.9	
TR792	80			
TR795	136			
TR803	134			
TR810	243			Residues 137-149
TR811	251			
TR816	68			
TR817	265	4WED	65.8	
TR821	255	4R7S	48.3	
TR822	117			
TR823	288			
TR827	193			
TR828	84			
TR829	67			N-terminal residues 2-9
TR833	108	4R03	61.3	
TR837	121			
TR848	138	4R4G	58.0	
TR854	70			
TR856	159			
TR857	96	2MQC	32.8	

As table 4-1 shows, the experimental structures for 22 targets are released on Protein Data Bank (PDB). The majority of targets did not have any suggestions from CASP on which regions need to be refined further. In the following sections, we describe our protocol applied to these protein models.

4.2.1 Conformational sampling

Two rounds of MD sampling were performed for each target. In all the simulations, CHARMM36[9] force field was used, along with TIP3[21] water model. In the first round, the initial models underwent 5,000 energy minimization steps, and heated to temperatures of 10, 50, 100, 200, and 298 K within 42 ps MD simulations. Then, multiple replica MD simulations in explicit solvent starting from the given initial structures were run for 30 ns per target per replica. Having 40 replicas per target, total simulation time in the first round summed up to 1.2 μs for each target. The restraint force constant was chosen to be 0.05 kcal/mol/Å^2 on all C_α atoms, unless suggestions from CASP exist for that protein, in which case, the restraint on suggested regions were relaxed (no restraint) and the force constant for the remaining regions was 0.1 kcal/mol/ Å^2. An ensemble containing 30,000 structures was generated from all 40 replicas. Then, according to our optimized/tuned subset selection algorithm and structure averaging described in the next section, a final refined structure is obtained. This model constitutes the first submitted model to CASP, and the starting structure for the second round of refinement.

In the second round of refinement, the refined and initial structures are geometrically aligned, and regions that have moved more than a certain threshold are identified. For this purpose, a moving window of size 3 residues is applied and the RMSD between the two structures subject to that window is computed. Consecutive regions that have RMSD more than threshold of 3 Å are identified. The selected regions are considered as potential regional targets for improvement

in the second round. Table 4-2 shows the number of regions, and the range of selected regions for each target. A set of 3 independent MD simulations of 20 ns were run for each region, while the residues outside the considered region were restrained with force constant of 1 kcal/mol/Å. For each region an ensemble of 1500 conformations was generated. The same subset selection algorithm, which was used in the first round, is applied to each ensemble, and an average structure for each region is obtained.

4.2.2 Subset selection and average structure

The model selection algorithm is based on our previous work, in which a subset of structures was selected based on two scoring functions. Here, we modified the criterion for subset selection as follows. We used iRMSD (RMSD from the initial model) and RW+ as two scoring functions. These scores were standardized by subtracting their mean and dividing by their standard deviation. Then, the subset of points, which satisfy the following condition in polar coordinates, are selected

$$S = \left\{ p_i \mid r_i > 1 \;\&\; 200^\circ < \theta_i < 270^\circ \right\}$$

where r_i and θi are the polar coordinates of point p_i in the space formed by iRMSD and RW+ scores. The selected structures are geometrically superimposed to the initial model, and an average structure is computed. The average structure is further refined to remove clashes and unphysical bonds and angles, by a MD simulation of length 200 ps, and restraints applied to all C_α atoms with force constant 10 kcal/mol/Å^2.

Table 4-2 Selected regions for the second round of sampling for each target

Target	Number of regions	Regions
TR217	10	277-287; 303-313; 316-326; 326-334; 347-357; 367-376; 404-414; 435-445; 447-455; 481-488;
TR228	4	236-245; 245-255; 269-277; 284-290;
TR274	14	187-197; 198-206; 206-212; 213-223; 243-253; 253-259; 264-267; 267-277; 277-287; 293-303; 309-312; 314-323; 337-347; 357-367;
TR280	6	135-145; 157-163; 176-185; 190-196; 201-206; 214-217;
TR283	11	247-254; 254-262; 273-283; 286-296; 298-308; 317-327; 331-341; 349-357; 359-363; 369-377; 396-404;
TR759	5	46-51; 52-62; 63-73; 73-79; 89-99;
TR760	12	41-50; 50-58; 65-74; 81-91; 94-101; 109-118; 130-140; 149-157; 165-174; 187-196; 205-208; 217-224;
TR762	12	24-29; 38-46; 49-54; 60-68; 117-127; 162-169; 183-187; 209-218; 224-229; 239-247; 254-259; 266-274;
TR765	5	37-47; 54-64; 70-74; 76-82; 95-101;
TR768	10	24-29; 36-44; 46-56; 63-73; 82-90; 99-104; 104-111; 118-128; 137-147; 155-160;
TR769	4	6-16; 32-39; 41-48; 81-90;
TR772	14	69-77; 78-88; 99-106; 106-110; 121-131; 137-147; 153-161; 161-165; 166-174; 174-180; 193-203; 220-230; 232-240; 243-252;
TR774	12	31-34; 35-45; 49-56; 60-69; 69-75; 75-85; 86-95; 95-105; 121-131; 136-144; 144-153; 156-166;
TR776	13	38-43; 60-66; 66-73; 73-77; 85-90; 107-117; 128-134; 135-144; 160-164; 175-180; 180-190; 228-235; 236-246;
TR780	7	40-48; 50-57; 57-67; 67-75; 81-91; 91-101; 109-119;
TR782	6	38-41; 43-48; 51-54; 60-68; 85-95; 100-107;
TR783	18	1-4; 13-23; 26-32; 42-48; 52-58; 63-73; 76-85; 91-101; 125-135; 142-145; 150-158; 158-168; 182-189; 189-198; 199-209; 212-221; 224-232; 232-237;
TR786	15	37-44; 45-55; 67-77; 91-100; 106-114; 114-122; 125-133; 133-142; 144-150; 155-165; 170-177; 177-185; 189-199; 223-233; 242-247;
TR792	4	6-14; 21-30; 46-56; 57-66;
TR795	9	19-28; 36-46; 61-66; 68-77; 87-92; 95-98; 102-107; 113-122; 128-133;
TR803	9	1-5; 30-38; 44-50; 50-60; 61-69; 70-80; 87-96; 101-111; 114-124;
TR810	13	137-140; 140-149; 152-159; 224-229; 233-241; 266-273; 277-285; 285-293; 293-299; 305-310; 320-326; 348-353; 364-373;
TR811	13	10-18; 21-31; 53-60; 60-63; 69-79; 83-88; 130-140; 155-163; 170-180; 196-205; 210-218; 230-237; 244-249;
TR816	3	25-35; 47-50; 56-62;
TR817	10	44-54; 54-62; 67-75; 122-131; 134-144; 148-158; 186-196; 200-210; 245-255; 280-288;
TR821	10	35-44; 45-55; 69-76; 82-90; 115-123; 128-136; 137-147; 156-166; 180-189; 217-225;
TR822	9	2-7; 8-17; 21-26; 31-41; 59-68; 68-73; 74-84; 99-106; 107-112;
TR823	14	5-8; 17-26; 30-35; 45-52; 73-82; 113-123; 129-136; 138-148; 148-155; 159-169; 190-200; 201-207; 243-251; 278-282;
TR827	13	31-40; 46-54; 54-64; 67-76; 84-92; 94-101; 113-118; 126-134; 137-142; 162-172; 176-181; 184-190; 191-201;
TR828	6	137-145; 145-153; 161-169; 177-185; 185-194; 205-211;
TR829	5	2-6; 23-32; 32-39; 42-52; 52-61;
TR833	4	46-53; 64-68; 92-100; 114-124;
TR837	7	8-17; 17-25; 37-44; 61-71; 75-83; 89-99; 110-115;
TR848	4	48-53; 96-106; 139-148; 155-164;
TR854	3	37-47; 64-69; 77-87;
TR856	8	1-8; 15-24; 43-53; 61-71; 83-92; 92-100; 118-126; 138-145;
TR857	0	

4.2.3 Model submission in CASP11

For all the targets, we used the final refined model obtained after structure averaging from the first round of simulations as model 1. The second model is taken from the consensus structure obtained via the average of both average structures from the first and second rounds. The last 3 models are selected from the average structures of individual regions in the second round of sampling that have the highest RMSD from the initial model.

4.3 Results and Discussion

The experimental structures for 22 targets are released in the PDB data bank. We have assessed the quality of our submitted models by calculating the change GDT-HA of submitted models from those of the initial models. Larger ΔGDT-HA means higher improvement in structure quality compared to the experimental structure is achieved. Table 4-3 shows ΔGDT-HA for all 5 models for each target. The average improvement in GDT-HA for the first model is 4.31% and for the best of five models is 4.71%. Comparing the average improvement of the first model to that of CASP10 suggests that our refinement protocol has improved from what we used in CASP10 with only 2.8% improvement. On the other hand, the best of 5 models does not show much difference to the first model. This indicates the shortcoming of our second round of refinement, in which the targets had restraint with force constant 1 kcal/mol/$Å^2$ on selected regions.

Significant improvement in 3 targets, i.e. TR759, TR765, and TR821, is observed with more than 10% change in GDT-HA. By carefully examining these targets, we can see that the improvement is made in the all parts of the protein including loops and secondary structure elements. Therefore, applying strong restraints on secondary structure elements would have limited refinement as was the case in our previous protocol. The initial, refined and native models for

TR759 are shown in Fig. 4-1. This level of refinement has not been observed in the past. While our results show significant improvement, it is still not clear how our results for targets TR759, TR765, and TR821 are compared against other groups. It could be that other groups were also able to achieve such high improvements. Therefore, a through comparison of the results in CASP11 would give insights on the effectiveness of refinement methods.

Table 4-3 GDT-HA results and MolProbity and ClashScore measures for quality assessment of submitted models

Target	ΔGDT-HA					
	Model 1	Model 2	Model 3	Model 4	Model 5	Best of 5
TR217	0.34	0.00	0.46	-0.58	0.11	0.46
TR228						
TR274	-0.50	-2.70	-1.50	-0.50	-2.50	-0.50
TR280	5.70	4.90	3.60	4.20	4.90	5.70
TR283	0.30	0.30	0.00	0.50	-0.20	0.50
TR759	**12.30**	**14.30**	**13.10**	**15.20**	**12.70**	**15.20**
TR760	0.40	-1.00	0.60	0.40	-0.10	0.60
TR762	-4.00	-5.60	-6.20	-5.00	-5.10	-4.00
TR765	**19.70**	**20.70**	**20.10**	**18.80**	**18.10**	**20.70**
TR768	6.10	6.50	4.70	4.50	4.40	6.50
TR769	1.80	0.80	1.00	1.30	1.00	1.80
TR772	1.40	0.90	0.50	0.80	-0.60	1.40
TR774	3.00	2.20	1.80	3.70	2.70	3.70
TR776	6.50	5.50	3.60	4.90	4.40	6.50
TR780	4.20	2.90	1.10	2.90	1.80	4.20
TR782	**8.60**	7.50	**8.40**	**9.30**	**8.60**	**9.30**
TR783	5.00	5.50	3.20	2.80	3.90	5.50
TR786	4.60	3.90	4.30	4.80	3.90	4.80
TR792						
TR795						
TR803						
TR810						
TR811						
TR816						
TR817	-0.94	-0.94	-1.03	-1.41	-1.32	-0.94
TR821	**12.16**	**12.75**	**12.26**	**11.96**	**12.26**	**12.75**
TR822						
TR823						
TR827						
TR828						
TR829						
TR833	2.30	0.90	2.80	3.00	-2.50	3.00
TR837						
TR848	1.99	0.91	1.45	1.09	-1.63	1.99
TR854						
TR856						
TR857	3.90	3.60	2.10	4.40	3.10	4.40
Avg.	4.31	--	--	--	--	4.71

Figure 4-1 Comparing the initial (green) and refined (cyan) models of TR759 to its experimentally observed crystal structure

4.4 Conclusion and Future Work

In CASP11 we modified our protocol by extending the amount of sampling by molecular dynamics simulation, using weaker restraints, and tuned the subset selection algorithm. As a result, we observed greater progress in structure refinement. The average improvement in GDT-HA for the first model was significantly higher than our previous result in CASP10. Yet, in three targets we have achieved improvements of more than 10% in GDT-HA. We believe that such significant improvements would be impossible using strong restraints. On the other hand, the second round of refinement with strong refinement on selected regions which led to models 2-5 did not show much difference to the first model.

Chapter 5

Density Biased Sampling: A Robust Computational Method for Studying Pore Formation in Membranes

Vahid Mirjalili, Michael Feig

Submitted to

Journal of Chemical Theory and Computation

5.1 Abstract

A new reaction coordinate to bias molecular dynamics simulation is described which allows enhanced sampling of density-driven processes, such as mixing and de-mixing two different molecular species. The methodology is validated by comparing the theoretical entropy of de-mixing two ideal gas species and then applied to induce deformation and pore formation in phospholipid membranes within an umbrella sampling framework. Comparison with previous biased simulations of membrane pore formation suggest overall quantitative agreement but the density-based biasing potential results in a different, more realistic transition pathway than in previous studies.

5.2 Introduction

Advanced computational methods have long attracted the attention of biophysicists to shed light on the behavior of biological systems. The computer simulation of proteins, membranes, and nucleic acids are a powerful technique for understanding the physical characteristics of these complex systems.[8] Despite advances in computer power, the time scales required for studying many physical phenomena are still beyond the possibilities for the majority of the scientific community. However, the use of enhanced sampling methods[133-137] can overcome such limitations. One example where enhanced sampling is needed is the pore formation and deformation of lipid membranes.[138-147] Pore formation is involved in a variety of biological processes, such as signal transduction and small molecule transports,[138-140, 147] but it is also highly-relevant in the context of toxins and antimicrobial peptides that induce membrane pores to cause cell leakage and ultimately kill cells.[39, 40, 148]

A common strategy for overcoming kinetic barriers is the use of umbrella sampling techniques[32], where a main challenge is the choice of a suitable reaction coordinate. Geometric properties such as distances, angles, or dihedrals between groups of atoms have been widely used, but some physical processes are not described well by such simple reaction coordinates. As a result, enhanced sampling simulations using such coordinates may be less effective for these systems. For example, density-driven processes may not be described well by traditional reaction coordinates. Membrane pore formation is one such process where the application of enhanced sampling methods has been challenging.[149] In one previous study, the pore radius was incorporated as a reaction coordinate in a molecular dynamics framework,[149, 150] and the free energy cost of pore formation was measured using the potential of mean constraint field (PMCF) approach[151]. Furthermore, Bennett et al.[152] investigated the mechanism of pore formation initially by long equilibrium MD simulations followed by umbrella sampling where a single phosphorous atom in one of the lipids was pulled to the center. However, both choices of the reaction coordinate could be problematic as they make assumptions about how the membrane structure deforms upon pore formation.

A natural reaction coordinate for studying membrane pore formation is the density of water molecules within the membrane in the area where pore formation takes place. Using the water density instead of a structural property of the membrane avoids biasing membrane structure unnecessarily but still provides enhanced sampling across the key kinetic barrier, *i.e.* water penetration into the membrane. Here, we are describing the development of a density-based reaction coordinate and its application in umbrella sampling simulations of membrane pore formation. The method introduced here biases the density of a group of atoms in a volume of interest, such as a cylinder. Therefore, our density biasing potential function can be used not just

for studying membrane pores but it is also applicable more generally for reaching a target density for a given molecular species relative to another species in any context. This methodology was implemented in the CHARMM biomolecular software package[86].

In the remainder of this paper, we will provide a detailed description of the density biasing potential, followed by validation of our method by comparing entropic components of de-mixing free energy of two ideal gases with theoretical estimates. Then, this method is applied to a pure DPPC membrane bilayer system to demonstrate its potential for estimating free energies of membrane pore formation.

5.3 Methods

5.3.1 Density Biasing Potential

In this section, we provide the mathematical basis of the density biasing potential function. Given the coordinates $\vec{q}_i$ for atom i, the total number of atoms in any arbitrary sub-volume of interest V can be calculated by integrating the product of a volume function $f(\vec{r})$ with the Dirac delta function: $\delta(\vec{r} - \vec{q}_i) f(\vec{r})$ for all atoms:

$$\Gamma_V = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_{i=1}^{N}\delta(\vec{r} - \vec{q}_i) f(\vec{r})\, dr^3 = \sum_{i=1}^{N} f(\vec{q}_i) \qquad (1)$$

where $f(\vec{r})$ returns one inside V while it switches smoothly to zero on the boundaries, and stays zero for all the points outside the volume (see below). In general, any differentiable volume function can be used to define $f(\vec{r})$, however, simpler functions are preferred since they are easier to implement in a molecular dynamics framework.

The volume of interest in our study is a cylinder with radius R_{cyl} and height Z_{cyl} with its axis aligned to the bilayer normal (Fig. 1A). Therefore, we use cylindrical coordinates and decompose the volume function into radial and axial components so that:

$$\Gamma_V = \sum_{i=1}^{N} f_{radial}(r_i) . f_{axial}(z_i) \tag{2}$$

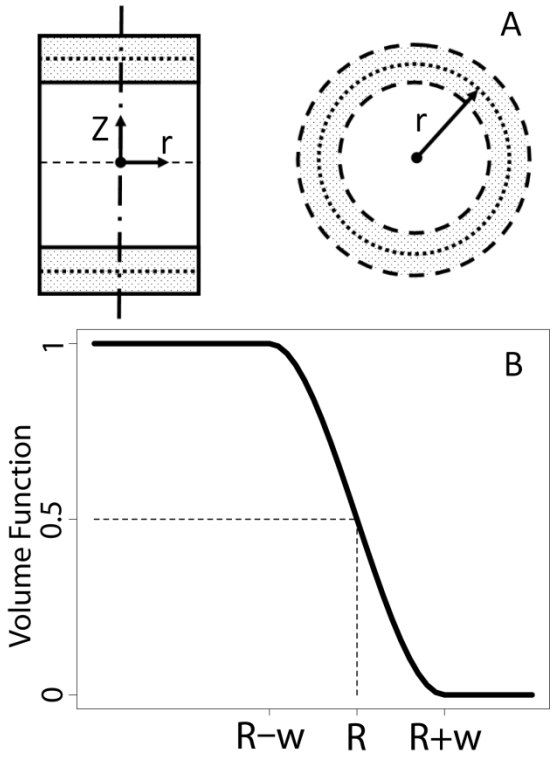

Figure 5-1 A: Schematic representation of the biasing cylinder aligned to the bilayer normal. The center of the switching region is indicated with dashed lines; B: Volume function used in axial and radial directions.

Choosing the switching function as a third degree polynomial used in CHARMM PBEQ[3] and GBSW[31, 153] modules results in the following differentiable volume function:

$$f_{radial}(r) = \begin{cases} 1 & r \le R_{cyl} - w \\ \frac{1}{2} - \frac{3}{4w}(r - R_{cyl}) + \frac{1}{4w^3}(r - R_{cyl})^3 & \left|r - R_{cyl}\right| < w \\ 0 & r > R_{cyl} + w \end{cases} \tag{3}$$

$$f_{axial}(z)=\begin{cases}1 & Z_{low}+h\le z\le Z_{up}-h\\ \frac{1}{2}-\frac{3}{4h}(z-Z_{up})+\frac{1}{4h^3}(z-Z_{up})^3 & \left|z-Z_{up}\right|<h\\ \frac{1}{2}+\frac{3}{4h}(z-Z_{low})-\frac{1}{4h^3}(z-Z_{low})^3 & \left|z-Z_{low}\right|<h\\ 0 & z>Z_{up}+h \text{ or } z<Z_{low}-h\end{cases} \quad (4)$$

where *w* and *h* are the switching distances for the radial and axial terms, respectively. Figure 1B shows the shape of radial component of volume function; the axial component has a similar shape.

The number density ρ_V is calculated by normalizing Γ_V by the cylinder volume. The potential energy is then calculated for a given value of target density ρ_t with the force constant k

$$U=\frac{k}{2}(\rho_V-\rho_t)^2 \quad (5)$$

The corresponding force components can be obtained from the gradient of the potential term

$$\vec{F}_i=\vec{\nabla}_i U=\frac{k}{V}(\rho_V-\rho_t)\,\vec{\nabla}_i\Gamma_V \quad (6)$$

$$\vec{\nabla}_i\Gamma_V=\left(f_{axial}(z_i)\frac{\partial f_{radial}(r_i)}{\partial x_i},\ f_{axial}(z_i)\frac{\partial f_{radial}(r_i)}{\partial y_i},\quad f_{radial}(r_i)\frac{\partial f_{axial}(z_i)}{\partial z_i}\right) \quad (7)$$

The details of the derivative components are provided in the Appendix.

5.3.2 Simulation Details

5.3.2.1 Method Validation

For validation of our computational method, we investigated the mixing entropy of two noble gas species. 200 helium atoms were placed in a box, 40 of which were tagged to make two distinguishable species with identical parameters. The box dimensions were 200×200×50 Å^3. A density biasing cylinder with a radius of 50 Å was placed in the center of the box with the

cylinder axis aligned with the z axis. The switching distance in the radial direction was set to 1 Å. The cylinder height was considered to be infinite; therefore the biasing potential did not vary along the z-axis. The number densities were normalized by the equilibrium number of particles in the cylinder volume. In order to fully separate the two molecular species, the reaction coordinate in the density biasing potential was constructed as the difference between the densities of the tagged and untagged species in the cylinder. In this case, an increase in the value of the reaction coordinate can be due to either increasing the number of tagged species or decreasing the number of untagged ones assuming that the total number of particles in the cylinder is constant on average over time.

For this system, the equilibrium value of the reaction coordinate is -6×10^{-5} Å^{-3} for the fully mixed state and 1×10^{-4} Å^{-3} for the fully separated state. Therefore, using umbrella sampling, the reaction coordinate was varied from -5.1×10^{-5} to 7.1×10^{-5} Å^{-3} in increments of 2.54×10^{-7} Å^{-3}. Each umbrella window was simulated for 20 ns with a force constant of 10^7 kcal/mol/Å^6 and a time step of 2 fs. The last 16 ns from each window were used to construct the PMF as a function of the reaction coordinate using WHAM analysis.

A theoretical estimate of the mixing entropy for two molecular gas species *A* and *B* is given by

$$\Delta S = nR(x_A \log(x_A) + x_B \log(x_B)) \tag{8}$$

where *x* is the mole fraction of each species, *n* is the total number of moles, and *R* is the universal gas constant. The total change in entropy is given by

$$\Delta S_{tot} = \overline{n_{tot,V_1}}\Delta S(x_{A,V_1}, x_{B,V_1}) + \overline{n_{tot,V_2}}\Delta S(x_{A,V_2}, x_{B,V_2}) \tag{9}$$

where $\overline{n_{tot,V_1}}$ and $\overline{n_{tot,V2}}$ are the average total number of atoms in volumes V_1 and V_2 at equilibrium, respectively. In order to compare the theoretical mixing entropy with our

computational approach, we evaluated the theoretical estimate as a function of the mole fraction of species *A* in volume V_1 in the process of going from a fully separated state (*i*) to a partially mixed state (*ii*) as shown in Fig. 2. The mole fraction is then converted to the reaction coordinate (ξ) used in the umbrella sampling simulations according to:

$$\xi = \frac{\overline{n_{tot,V_1}}(x_{A,V_1} - x_{B,V_1})}{V_1} \tag{10}$$

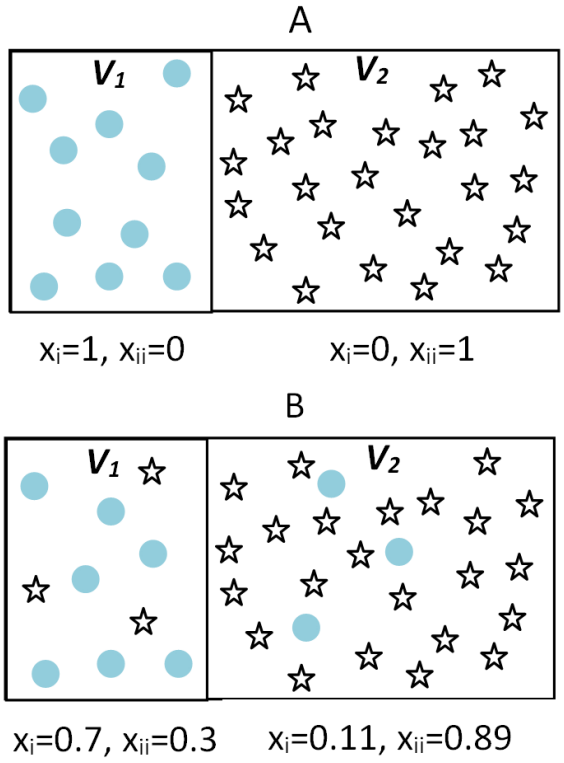

Figure 5-2 Schematic representation of the mixing process for a simple two-component noble gas mixture that is fully demixed (A) and partially mixed (B).

5.3.2.2 Membrane Simulations

A pure membrane bilayer was constructed by web-based CHARMM-GUI membrane builder[154], containing 288 dipalmitoyl phosphatidylcholine (DPPC) and 8376 water molecules placed in a periodic box of size 95.2×95.2×66.6 $Å^3$. The x-y dimensions were adjusted to match

the experimental value of 63 $Å^2$ for the area per lipid of DPPC in the fluid phase.[155, 156] The z dimension was chosen large enough to avoid boundary artifacts. The CHARMM36 force field[157] was used along with the TIP3 water model[21]. Lennard-Jones interactions were cut off at 9 Å (with a switching function beginning at 8 Å). Particle-Mesh Ewald summation[158] was used for long-range electrostatic interactions with a 9 Å cutoff for the direct sum. A time step of 2 fs was used in combination with the SHAKE algorithm.[159] The initially flat bilayer was heated in steps at 50K, 100K, 200K, 250K, and 323K, each for 100 ps with a Nosé-Hoover thermostat and barostat (target pressure of 1 bar) to maintain an NPT ensemble. The center of mass of the bilayer was restrained to the plane at z=0 with a force constant of 100 kcal/mol/$Å^2$. The final equilibrated system was then used to study membrane deformation and pore formation with our density biased sampling method.

5.3.2.3 One-sided deformation of a membrane bilayer

The density biasing approach was applied to the DPPC membrane bilayer system. A cylinder with a radius of 8 Å was aligned to the bilayer normal (z) axis. The cylinder spanned from z=-2.5 Å to z=+15 Å, and the radial and axial switching distances were set to 1 Å and 5 Å, respectively. Umbrella sampling simulations were performed with 10 windows, increasing the number density of water molecules per unit area in the cylinder from 1.1×10^{-3} to 2.17×10^{-2} $Å^{-3}$. A force constant of 9.2×10^5 kcal/mol/$Å^6$ was used. To prevent deformation in the lower leaflet, a plane potential with a force constant of 100 kcal/mol/$Å^2$ was applied to the phosphates of the lower leaflet if their distance to bilayer center was less than 8 Å. Each umbrella was simulated for 50 ns.

5.3.2.4 Pore formation in a membrane bilayer

In order to create a pore in a membrane bilayer, we expanded the cylinder from the previous case to cover both leaflets, *i.e.* from z=-18 Å to z=+18 Å. The radius of the cylinder was chosen as

r=6 Å, and the radial and axial switching distances were set to 2 Å and 8 Å, respectively. The parameters were adjusted based on initial trial simulations in order to achieve double-sided pore formation. 20 umbrella windows were used to vary the number density of water molecules in the cylinder from 6.7×10^{-3} $Å^{-3}$ to 2.25×10^{-2} $Å^{-3}$, using a force constant of 5.18×10^{6} kcal/mol/$Å^{6}$. Each umbrella was simulated for 50 ns. The total simulation time for pore formation was 1 μs.

5.3.2.5 Parameter Selection

While our method can be used for a diverse set of applications, the biasing potential parameters would have to be adjusted accordingly. We will provide guidance here how to choose the two key parameters, height and radius, for the case of a cylindrical biasing volume.

Generally, the cylinder height should encompass and extend beyond the region where the density is meant to be changed. For membrane simulations, a short cylinder height would be appropriate to induce one-sided deformation while longer cylinders are necessary to induce transmembrane pores. Furthermore, for one-sided deformations, the lower bound of the cylinder was fixed at z=-2.5 Å to let water molecules reach the bilayer center without forming complete pores. In the helium gas de-mixing simulations, the cylinder height was chosen bigger than the box size to avoid gradients along the z axis.

The cylinder radius should be chosen large enough so that the cylinder extends beyond the pore or deformation that is meant to be formed. Otherwise, the biasing potential may affect the shape of the deformation. On the other hand, a cylinder radius that is too large may not be effective in inducing pore formation because large membrane deformation could also satisfy a bias towards increased water densities within the cylinder. Because it was not entirely clear *a priori* which radius and cylinder height would be optimal, we conducted a series of test simulations with varying radii and cylinder heights until pore formation was accomplished successfully.

Finally, the force constants and window spacing were optimized by trial error. We found that the final values were similar as those predicted by the criterion given by Park and Im[97].

5.3.2.6 Implementation

The density biasing method using a cylinder-based volume function was implemented in the CHARMM biomolecular software package[86], version c40a1. Although not implemented so far, it would be easy to extend the method to other geometries such as a rectangular box with switching regions on each edge or a spherical geometry.

5.4 Results and Discussion

5.4.1 Mixing entropy of two-component gas

The free energy cost of separating two noble gas species was calculated using theoretical and computational methods. Since the two species have identical properties, there is no change in the mean of internal energy of the system upon separating the two species. Figure 3 compares theoretical estimates of $-T\Delta S$ according to Eqs. 9 and 10 with the change in free energy computed using the density biased sampling method. The reference point in this figure is the fully mixed state which has the highest entropy. This state corresponds to a mole fraction of $x_A=0.22$. If the theoretical estimate assumes a perfectly uniform particle distribution to obtain the number of particles in the cylinder (Fig 3 – theory A), the ΔG from the simulation underestimates the theory significantly. The agreement improves when the actual average number of particles in the umbrella windows that corresponds to the de-mixed states is used in the theoretical estimate (Fig 3 – theory B). The remaining small discrepancy is due to a non-

negligible virial term that results from a pressure difference inside and outside the cylinder during the umbrella simulateons in response to the application of the biasing potential. A correction by adding $-\Delta(PV)$, calculated from the average external pressures from simulations of the fully mixed and fully de-mixed states as reported by CHARMM, brings the theoretical and simulation estimates in near-perfect agreement. We note, that the simulated system is not an ideal gas because of weak attractive interactions and volume exclusion effects as a result of the Lennard-Jones interaction potential. This would lead to a small correction of the theoretical estimate that is expected to be smaller or on the same order as the uncertainties in the free energies obtained from the simulations. Therefore, the simple test case validates the density biasing potential introduced here.

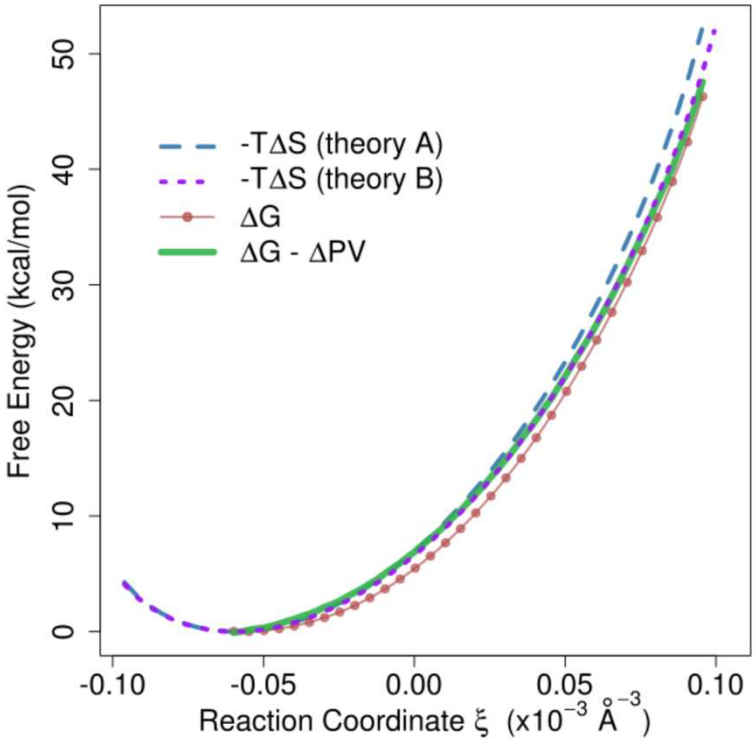

Figure 5-3 Free energy cost of mixing two noble gas species as a function of the biasing reaction coordinate ξ based on theory (mixing entropy) and simulation (free energy calculated from umbrella sampling simulations). Theory A is using uniform density to estimate total number of

particle in cylinder, whereas theory B uses the empirical average number of particles observed during the simulations.

5.4.2 Membrane Simulations

We will now demonstrate the application of the density biasing approach to simulations of membrane bilayers. As described in detail in the Methods section, the density biasing potential was applied to water molecules within a cylinder encompassing a section of a phospholipid bilayer. Figure 4 demonstrates how local membrane thickness, calculated as the average z coordinate of the phosphorous atoms in a cylinder of radius 8 Å, responds to the water density in the cylinder when varied in umbrella sampling simulations. The strong correlation reaffirms that water density within the bilayer is a suitable reaction coordinate for inducing membrane deformations. Figure 5 shows snapshots of the membrane bilayer after 50 ns molecular dynamics simulation with the density biasing potential set to increasing target values. The increasing degree of membrane deformation is readily apparent and we note that the deformation appears to proceed with a slight bending on both leaflets (Fig. 5C), presumably because this lowers the overall free energy for these intermediate states. However, a further increase in the water density results in a pronounced one-sided deformation with little apparent perturbation on the opposing leaflet. This is shown in Fig. 5F. Another feature of the deformation process is that it progresses from an initially wide and shallow deformation to a narrow and deep deformation, presumably due to a balance between the elastic properties of the membrane bilayer and the free energy costs of forming water defects within the membrane. The deformation of the bilayer is further quantified in Figure 6A, where the bilayer thickness at the deformation location is shown for each umbrella window. The first umbrella is simulated with equilibrium flat bilayer conditions, therefore, no deformation is observed. However, other umbrellas increase the density of water molecules, which induces a deformation in bilayer. From the umbrella sampling, a free energy

profile was obtained by weighted histogram analysis method (WHAM)[33]. Figure 6B shows the resulting potential of mean force (PMF) as a function of the water density in the cylinder. As would be expected, increasing the number of water molecules, and thereby deforming the bilayer, is highly unfavorable in terms of free energy with a cost exceeding 40 kcal/mol for a one-sided water defect that extends to the center of the membrane. As shown below, the cost of forming a pore is about half so that without any restraints on the lower leaflet (see methods section) the bilayer would not be expected to stably maintain a one-side deformation.

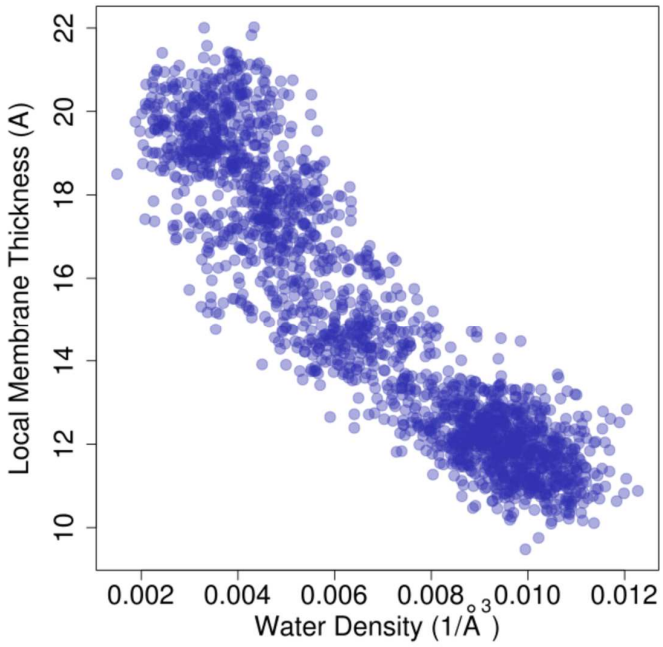

Figure 5-4 Local membrane bilayer thickness of the upper leaflet vs. water density per unit volume from biased sampling of one-sided membrane deformations.

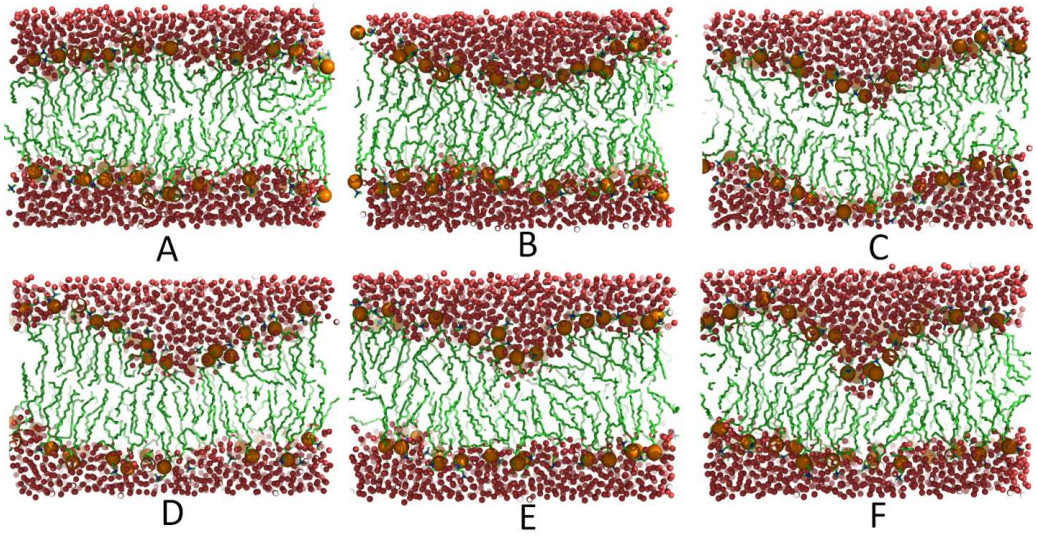

Figure 5-5 Snapshots illustrating the one-sided deformation process from a flat bilayer state to a fully deformed state at water densities of 0.0016 (A), 0.0073 (B), 0.0111 (C), 0.0143 (D), 0.0167 (E), and 0.0170 Å-3 (F); Red spheres represent water molecules, brown spheres represent phosphorous atoms of the lipids, and lipid tails are shown in green.

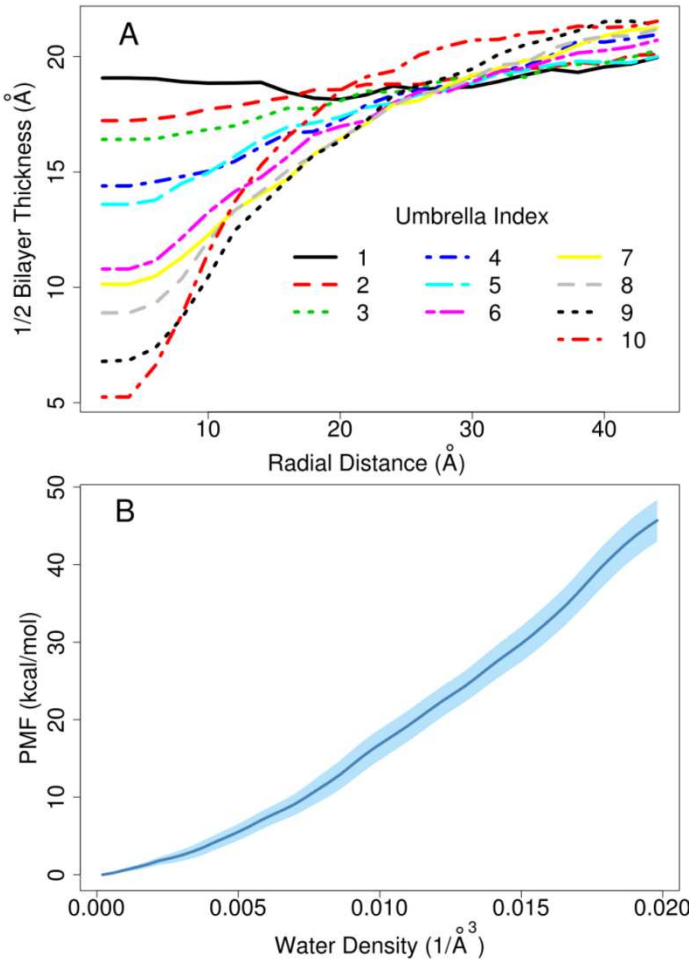

Figure 5-6 Average bilayer thickness in radial slabs for each umbrella window as a function of radial distance from the pore center. B: Free energy profile for one sided bilayer deformation as a function of water density in the cylinder. Standard error values obtained by calculating the PMF profiles over 10 2-ns subsets from the umbrella sampling simulation are shown as light blue shades.

Finally, we applied the density biasing method across the entire DPPC bilayer in order to induce pore formation. Snapshots of the bilayer after 50 ns molecular dynamics simulations with increasing water density biases are shown in Fig. 7. Similar to what has been described previously[152], pore formation starts by bending both leaflets inward. A water wire forms initially (Fig. 7D). The lipid head groups then rearrange and form the familiar hourglass shape of a stable pore once a critical pore radius is passed (Fig. 7E). A transition involving an initial water wire is consistent with results from the equilibrium simulations by Bennett et al.[152] The

average number density profiles of water molecules across the bilayer normal for a flat bilayer and a bilayer with a stable pore (with average water density of 0.0216 $Å^{-3}$) are compared in Fig. 8. By integrating over the difference between the two curves, it is found that 148 water molecules exist in the pore. This result is comparable with the 124 water molecules obtained by Leontiadou et al.[140], in which they applied mechanical stress (surface tension) to form a pore in a DPPC bilayer.

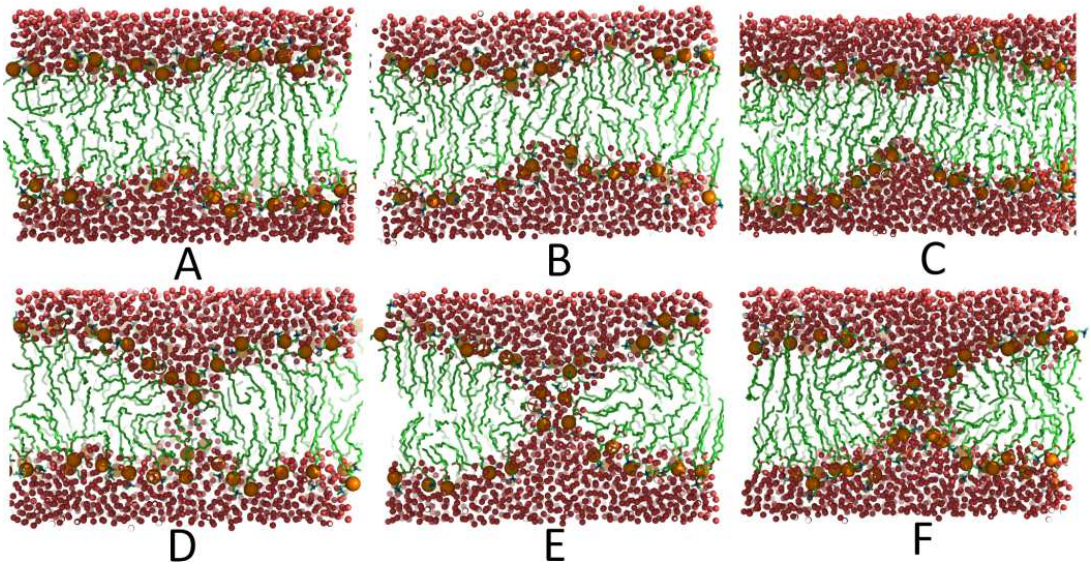

Figure 5-7 Snapshots illustrating the pore formation process from a flat bilayer state to a stable pore at water densities of 0.0067 (A), 0.0144 (B), 0.0159 (C), 0.0168 (D), 0.0196 (E), and 0.0222 Å-3 (F) with coloring as in Figure 5-5.

We computed the pore size by assuming perfect cylindrical shape between z=-8 to z=8 Å, and a uniform water density in that region. The average number of water molecules in the region was found to be 117.7 in the last umbrella. The resulting pore radius is found to be 8.8 Å. Similar analyses assuming perfect cylinder for water wire result in pore radius of 4.2 Å.

Figure 9 shows the PMF of pore formation as a function of water density in the aforementioned cylinder. Again, pore formation is energetically unfavorable as expected. A plateau free energy of 22.2 (+/- 0.4) kcal/mol is reached at a critical water density of 0.018 $Å^{-3}$ once a stable pore is

formed. This result is close to the value of 19.02 kcal/mol reported by Bennett et al. for DPPC.[152] The agreement is excellent, especially when considering differences in force fields. We further decomposed the free energy into enthalpic and entropic contributions. The change in enthalpy is estimated by computing the average potential energy of the system and we found that pore formation is enthalpically favorable by 46±1 kcal/mol. The simple de-mixing test case above suggests that there may be an additional ΔPV term but for a partially de-mixed system the contribution is estimated to be less than 1 kcal/mol and it is therefore neglected here. This implies an entropic cost (-TΔS) of pore formation of about 68 kcal/mol.

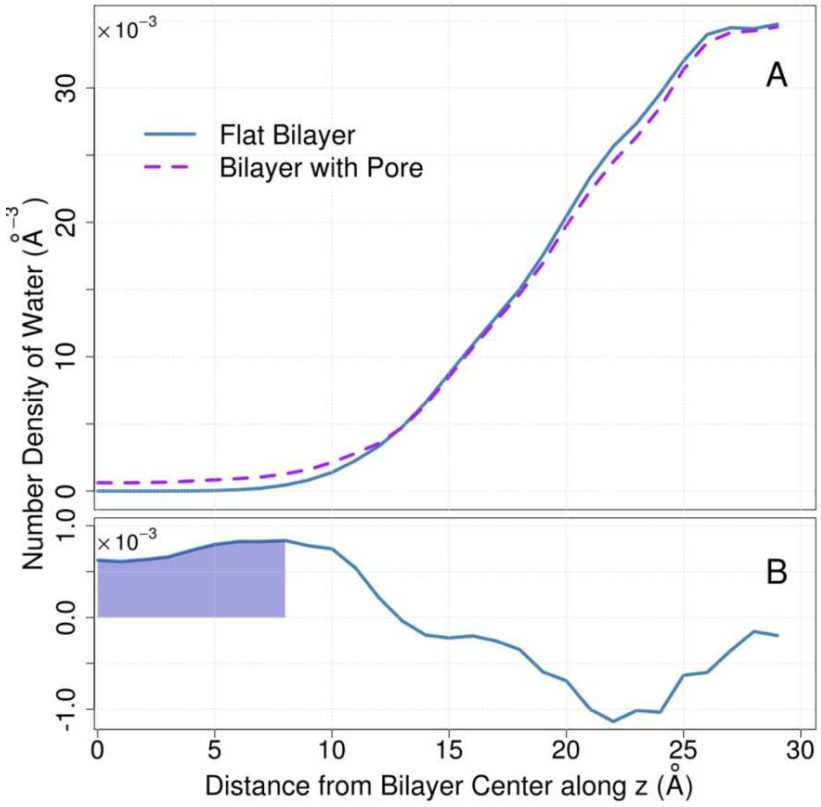

Figure 5-8 Number density of water molecules across bilayer normal compared between a flat bilayer and a bilayer with a stable pore (A), and their differences (B).

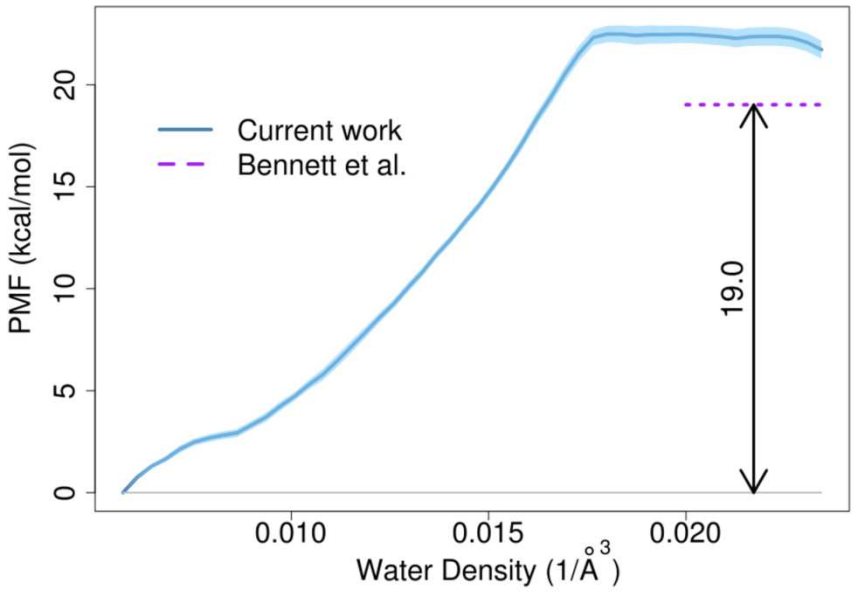

Figure 5-9 Free energy of pore formation as a function of water density in the cylinder from density-biased sampling with errors indicated as in Figure 5-6B. A previous result from Bennett et al. is shown for comparison.

As mentioned above, one motivation for inducing membrane pores via water density biasing rather than biasing the membrane structure directly was to avoid artifacts that could affect the pore formation pathway and thereby the energy profiles obtained from umbrella sampling. Figure 10 compares the water density (ξ), our reaction coordinate that imposes minimal bias on the membrane structure, with the average distance of the two closest phosphates from the bilayer center (λ). The latter relates to previous biased simulation studies where the distance of a single phosphate group from the bilayer center was used.[152] Poor correlation between the two reaction coordinates suggests that there could be mechanistic differences when either of the two reaction coordinates is used to induce pore formation. With the density biasing term, a typical transition path (indicated in red in Fig. 10) would delay a transition of phosphates to the bilayer center until a critical water density is reached at which point there is a sharp, cooperative transition that leads to formation of a full pore. On the other hand, we speculate that forming the pore by pulling down a phosphate group would follow a path indicated in green in Fig. 10 where

phosphates approach the center of the bilayer early and a sharp, cooperative transition could be absent. Figure 11 shows two intermediate conformations with extreme low λ values that may be intermediates on such a transition path. In these conformations, the membrane exhibits large deformations on one leaflet, and the water molecules are dragged into the center along with the lipid headgroups, as shown in Fig. 11. Since free energies are state functions, overall energies of pore formation are of course independent of the path taken. However, the free energy profile along the transition path and any mechanistic insight obtained from such simulations does depend on the path taken as a result of the biasing potential.

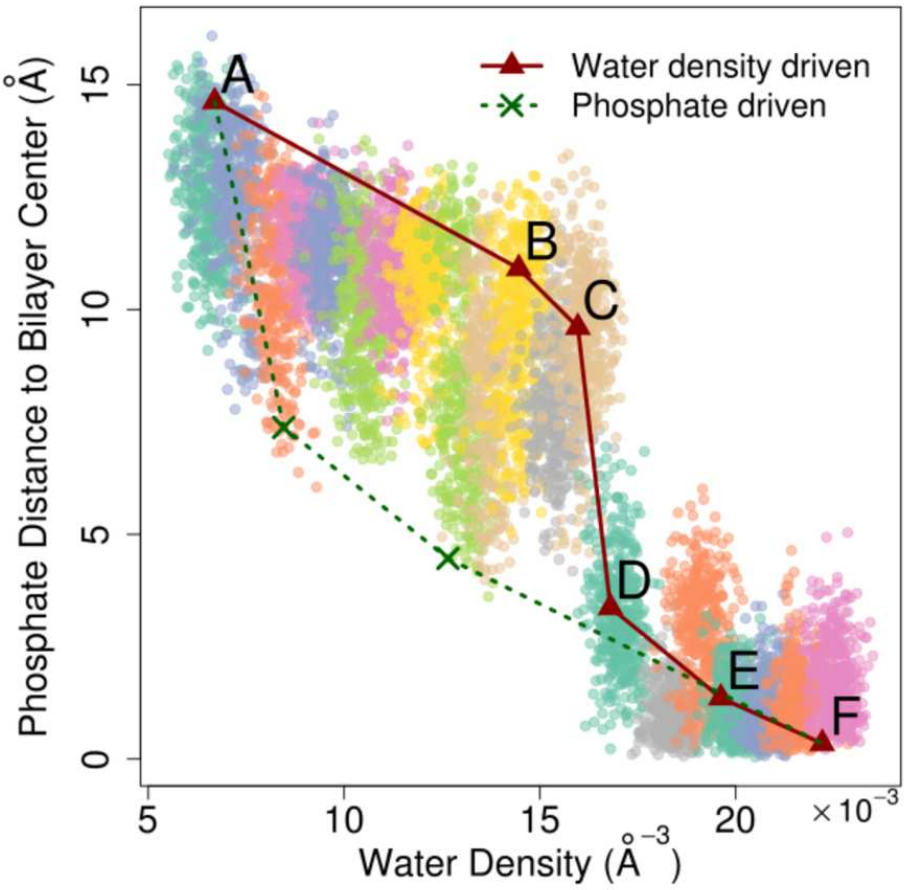

Figure 5-10 Average z coordinate of the two closest lipid phosphates from the bilayer center vs. water density within pore cylinder illustrating different mechanisms between density-driven and phosphate-driven pore formation bias. Sampling from each umbrella is shown in different colors.

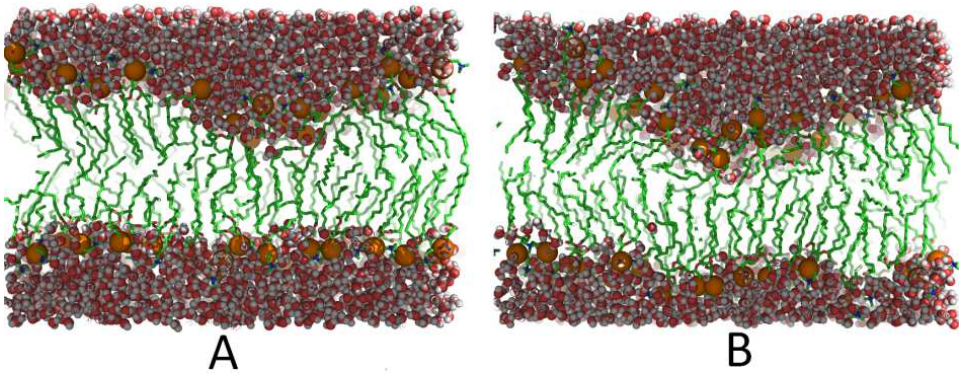

Figure 5-11 Intermediate bilayer states with low average distance of phosphates to the bilayer center.

The proposed method in this work applies a minimal bias to induce a pore in membrane. There is no assumption made about the shape of the pore or the density distribution inside the cylinder. However, the performance of this method is sensitive to the choice of cylinder parameters as described above. Therefore, we believe that this method is more universally applicable to membrane pore formation and deformations in response to interactions with other molecules, especially in cases where it is not clear *a priori* how exactly the membrane responds to such molecules.

The variation of the water density in our method is reminiscent of grand canonical ensemble methods[41, 132] that have been widely used to simulate the mixing process of model fluids[34, 160]. However, because de-mixing and bilayer pore formation processes maybe either thermodynamically unfavorable or kinetically hindered, enhanced sampling techniques such as umbrella sampling would still be required. Furthermore, a global variation of the chemical potential for water in a membrane-bilayer system may not necessarily lead to pore formation since water molecules could be added in the bulk region while a targeted change of a local chemical potential would eventually result in a method similar to ours but with the additional complications of the grand-canonical machinery.

Finally, while the method presented here focuses on overcoming the kinetic barriers in creating membrane deformations and pores, it may not fully address overcoming the slow relaxation times of lipid motions. Therefore, mechanistic studies of membrane pore formation would likely require longer simulations and/or a combination with other enhanced sampling techniques such as replica exchange sampling that can accelerate lipid motions to guarantee full convergence of deformed bilayer systems.

5.5 Conclusions

We have developed a new computational technique to bias the density of a group of molecular species, or the difference in densities of two molecular groups. The method was validated for the case of de-mixing two ideal gas species. Furthermore, we applied the new biasing term in the context of membrane pore formation. We believe that biasing the water density rather than structural properties of the membrane is less likely to introduce artifacts. Furthermore, the density biasing approach allows the study of one-sided deformations which has not been described with umbrella sampling techniques previously. The density biasing function is also more broadly applicable to any system involving the mixing or de-mixing of molecular species with respect to each other. Possible applications include lipid raft formation, co-solvent effects, and studies of concentration gradients in complex systems.

Chapter 6

Interactions of Amino Acid Analogs within Membrane Environments

Vahid Mirjalili, Michael Feig

Submitted to

Journal of Physical Chemistry B

6.1 Abstract

The interactions among four amino acid analog pairs (Asn, Ser, Phe, and Val) in the membrane environment are investigated using umbrella sampling molecular dynamics simulations. The physical characteristics of interactions among the amino acid pairs at the bound states and transition states were analyzed, and useful insights are gained by observing the differences in the relative population of the bound state conformations at different distances from the bilayer interface. It is shown that the distance from the bilayer interface dictates the interactions between the polar pairs and their conformations. Furthermore, the binding free energy obtained from all atom explicit simulations of each pair with respect to the bilayer normal distance is calculated. The results of this study can potentially be used for parameterization of other membrane models, as we have compared the results with three implicit membrane models.

6.2 Introduction

Membrane proteins are involved in a variety of cellular processes such as molecular transport and signaling pathways, and they are the target of many pharmaceutical studies. Membrane proteins are embedded in lipid bilayers that support and modulate their activity.[161-163] As with soluble proteins, the interactions among the amino acids and the environment are the primary determinants of membrane protein structure and function.[163] Yet, knowledge gaps remain about the fundamental nature of amino acid interactions within the membrane environment. Experimentally, such interactions are notoriously difficult to probe while computational studies have been hindered by the long time scales and complexity of bilayer systems. Therefore, many computational studies of

amino acids in bilayer environments rely on simplifications such as coarse-graining[164, 165] or implicit membrane[134, 166-168] models to facilitate the conformational sampling of membrane protein systems[169-172] at the expense of neglecting subtle details of amino acid lipid interactions.

The association and insertion of small peptides in aqueous and lipid environments has been the subject of several studies.[2, 170-185] In particular, amino acid insertion into membrane bilayers is fairly well understood. MacCallum et al. investigated the insertion of 17 amino acid side chains from the bulk water into membrane region and calculated the distribution of amino acid analogs with respect to the distance from the bilayer center.[2] The membrane insertion free energy profiles for each amino acid group (aliphatic, aromatic, and polar side chains) were compared, and it was found that the energetic minimum of aliphatic side chains is at the center of the bilayer, while the free energy minimum for aromatic side chains (Trp, Tyr, Phe) is located near the lipid carbonyl group. Polar residues (Asn, Gln, Ser, and Thr) have large positive free energies at the bilayer center that can be resolved in part by allowing water penetration into the lipid bilayer.[2] Membrane deformations are especially dramatic when charged amino acids are inserted as described most clearly for the case of arginine insertion.[2]

The energetics of amino acid interactions within the membrane is less well understood. Kim and Im[174] studied the interactions of transmembrane (TM) helices with lipid bilayers. They decomposed the PMF of helix tilt angles into entropic and helix-lipid interactions, and concluded that helix-lipid interactions provide a driving force for helix orientation under positive hydrophobic mismatch conditions.[174] In a recent study, Castillo et al.[177] studied the association of two WALP23 peptides in three lipid

membrane systems using the MARTINI coarse grained model[164]. In that study, the peptide-peptide, peptide-lipid, and lipid-lipid interactions upon peptide binding were analyzed and characterized in terms of their thermodynamic behavior. They reported that association of WALP23 peptides is favored by more than 20 kJ/mol, without any free energy barrier separating associated and dissociated states.[177] In a more systematic study by de Jong et al.[1], the dimerization of amino acid side-chain pairs was simulated using different force fields in water, *n*-octanol, and decane as mimics of lipid membranes. The general features of favoring association of polar compounds and disfavoring association of hydrophobic compounds in decane and octanol were reproduced but it remains unclear how their results transfer to actual lipid bilayer environments.

In order to better understand amino acid interactions in lipid bilayers, this work describes the association of pairs of four amino acid analogs, acetamide (Asn), methanol (Ser), toluene (Phe), and propane (Val) in DPPC bilayers from extensive biased-sampling all-atom computer simulations. The results provide association free energy profiles and detailed insight into the coupling between inserted amino acid pairs and membrane deformations. Furthermore, the energy profiles were compared with common implicit solvent models[24, 186, 187] to assess their ability to reproduce amino acid interactions within the membrane.

6.3 Materials and Methods

Pairs of four amino acid analogs were considered in this study: toluene-toluene, acetamide-acetamide, methanol-methanol, and propane-propane (Fig. 6-10 supplementary materials). Molecular dynamics umbrella sampling simulations were used to study the interactions among each pair at different positions along the bilayer normal: z=0, 4, 8, 12, 16, 20, 24 Å. The center of mass was restrained to the respective z values using a harmonic potential function with a force constant of 50 kcal/mol/Å^2. The center of mass distance between the analogs was then varied from 3 Å to 15 Å with increments of 0.5 Å. At each distance, umbrella sampling was carried out[32] using a force constant of 5 kcal/mol/Å^2 to maintain the respective distances. Initial systems were set up by placing the pair of molecules inside two spheres that were created inside the membrane at different values of z. Two sets of umbrella sampling simulations were carried out. In one set (called forward sampling), the amino acid analogs were initially equilibrated at a distance of 5 Å, and then after 2 ns simulation, the pair distance was decreased to 4.5, 4, 3.5 and 3 Å as well as increased to 5.5, 6, 6.5 … 15 Å in subsequent simulations. In the second set (called backward sampling), the pair was initially equilibrated for 2 ns at a distance of 15 Å and then pulled to increasingly shorter distances up to 3 Å.

6.3.1 Explicit Solvent Simulations

A membrane bilayer consisting of 288 DPPC molecules was constructed and enclosed in a periodic box with a fixed lateral size of 95.24 Å × 95.24 Å. The non-bonded interactions were calculated within a cutoff distance of 10 Å (switched to zero between 8.5 to 10 Å), and for long range electrostatic interactions particle-mesh Ewald (PME) with a grid spacing of 1 Å was used. The simulations were performed using the NAMD

molecular dynamics package[188], under NPAT conditions using Langevin dynamics with a temperature of 323 K, and a constant normal pressure of 1 bar. A time step of 2 fs was used in conjunction with SHAKE. The CHARMM36 force field[9] was used to model the lipids, the CHARMM General force field (CGenFF)[11] for the amino acid analogs, and the TIP3 water model[189] was used.

Initial configurations were minimized for 500 steps and then heated and equilibrated to temperatures of 20 K, 100 K, 250 K and 323 K for 2 ps, 2 ps, 2 ps, and 10 ps, respectively under the restraining potentials with respect to the pair distance and the z position of the pair. The overall center of mass of the lipids was also restrained to zero using a force constant of 100 kcal/mol/$Å^2$. Subsequent umbrella runs were started from the previous 2 ns production run, and equilibrated and heated to 100 K, 250 K, and 323 K with their corresponding umbrella potential.

The first 2 ns of each simulation was discarded, and the rest of the data was used for calculating PMFs. To assess convergence, the root mean squared deviations (RMSD) between the potentials of mean force (PMF) at a given distance were compared between the forward and backward sets. Simulations were initially carried out for 6 ns per umbrella and continued in both sets until an RMSD value of less than 0.2 kcal/mol was achieved (see Figure 6-11). For some umbrellas this required as much as 200 ns with explicit solvent and lipids (see Table 6-2). Generally, polar compounds required more sampling because of coupling with membrane deformations as discussed below. The total simulation time for acetamide and methanol pairs were 10.4 μs and 8.0 μs, respectively, whereas for toluene and propane the aggregate simulation times were 6.8 μs and 2.1 μs. Finally, weighted histogram analysis method (WHAM)[33] was used to generate a

composite unbiased PMF from the individual umbrellas along the entire range of pair distances.

6.3.2 Implicit Solvent Simulations

Three implicit solvent models were considered in this study, HDGB, GBSW, and IMM1, the implicit membrane extension of EEF1. The implicit solvent simulations were run using CHARMM[86] following the same umbrella sampling protocol as with the explicit lipids and solvent but with a shorter time of 1.5 ns per umbrella that was sufficient to satisfy the convergence criterion. All the initial systems underwent 50 steepest descent energy minimization steps followed by 500 adopted basis Newton Raphson method. Then the systems were heated to 100 K, 200 K, and 323 K for 500 MD steps. The production runs were performed for 1.5 ns in each direction. For HDGB simulations, the dielectric and non-polar profiles along the Z axis were adopted from Sayadi et al.[169] (also shown in table 6-3 supplementary materials). A scaling factor of 0.015 kcal/mol/Å^2 was used to obtain non-polar solvation free energies proportional to the solvent-accessible surface area (SASA).[26] For GBSW simulations, the implicit membrane thickness was set to 28 Å, and a switching length of 0.3 Å was used. In the case of IMM1 model, a membrane thickness of 28 Å was used. For IMM1, the amino acid analog parameters were directly adopted from their corresponding amino acids in the EEF1 model without further modifications. With the given parameters, all three implicit solvent models are meant to approximate the energetics of a DPPC bilayer.

6.3.3 Bilayer Deformation Simulation

For certain separation distances and certain values of z, membrane bilayer deformations were observed with acetamide and methanol pairs (see results section). In most cases,

forward and backward umbrellas exhibited the same behavior (deformed or undeformed membrane), but in a few cases bistable behavior was observed where forward and backward sampling did not converge to the same state and where the membrane was deformed in one case but not the other. In order to be able to generate a complete free energy profile we carried out additional umbrella biasing simulations at a fixed distance and z value but varying the degree of bilayer deformation.

To connect states with different degrees of membrane deformation we employed a recently introduced density-biasing approach[190]. In this method, an imaginary cylinder is placed along the bilayer normal axis. A volume function *V* is defined with two independent radial and axial components with a value of 1 inside the cylinder that is smoothly switched to zero to points outside the cylinder. The integral of the volume function over all water molecules gives the number of water molecules within the volume, which once normalized by the cylinder volume, is used as the reaction coordinate where low water density corresponds to an undeformed bilayer and high water density indicates deformation. In this case, a cylinder with radius 8 Å was used, spanning from z=-2.5 to z=15 Å, with the switching region set to 1 and 5 Å in radial and axial directions, respectively. Umbrella sampling was then used to vary the water density in the cylinder from 1.1e-3 $Å^{-3}$ to 17.1e-3 $Å^{-3}$ over eight umbrella windows with a force constant of 1.225e6 kcal/mol/$Å^{-6}$. Due to convergence issues, we increased the number of umbrellas to 16 for methanol at z=4 Å. An additional restraint was applied to the phosphates of the lower leaflet if their distance to bilayer center was less than 8 Å, which prevents deformation in the lower leaflet. Density biased molecular dynamics simulations were carried out for 48 ns for each umbrella. The water-density biasing simulations were

combined with the distance-based umbrella simulations to generate 2D PMFs as a function of the pair distance (ξ) and water density (ρ) using WHAM[33]. Final 1D PMF profiles as a function of the pair distance (ξ) were obtained by Boltzmann averaging according to Eq. 1

$$\Delta G'(\xi) = -k_B T \log\left\langle \exp\left(-\frac{\Delta G(\xi,\rho)}{k_B T} \right) \right\rangle \quad (1)$$

6.4 Results and Discussion

Results from extensive biased molecular dynamics simulations are presented that describe the pairwise interactions between acetamide, methanol, toluene, and propane pairs at different distances from the center of a lipid bilayer. Although the main focus of this study is on the amino acid interactions within lipid bilayers, we observed significant coupling with the lipid bilayer structure, which will be described first before continuing to amino acid association energetics and structural details.

6.4.1 Membrane Deformations

Because none of the compounds are charged, we initially assumed that membrane deformations would be modest and limited to cases where the analogs are near the membrane surface. However, we found significant membrane deformations even for deeply inserted acetamide and methanol pairs as shown in Figure 1. In the case of acetamide, deep deformations of the bilayer are observed consistently at z=4 Å and z=8 Å. When the acetamide pair is at the center (z=0), deformations are observed in some of the umbrellas and only at some pair distances suggesting a bi-stable scenario where

deformed and undeformed membrane states are similarly favorable but separated by a significant kinetic barrier. Methanol pairs also result in membrane deformations at z=4 Å and z=8 Å but not at z=0. At z=4 Å, the sampling is again bi-stable with all of the backward sampling umbrellas showing a deformed membrane while the membrane is deformed only at three pair distances in the forward sampling umbrellas. The non-polar compounds toluene and propane do not lead to water insertion when inserted deeply but when fixed at z=16 Å and z=20 Å the bilayer expands to accommodate the hydrophobic pairs. The membrane deformation largely disappears when the pairs are placed even further away from the center at z=24 Å.

In order to further understand the bi-stable membrane deformation states for acetamide and methanol, we carried out additional density-biasing umbrella sampling simulations along the deformation reaction coordinate for acetamide and methanol pairs at z=0 and z=4 Å and at short pairwise distances where the bi-stable behavior was observed. The results are shown in Figure 2. In both cases, two states are found, separated by a kinetic barrier. In the case of acetamide, deformed and undeformed membranes are similarly favorable; for methanol the deformed membrane appears to be slightly more favorable when the pair is placed at z=4 Å. Water density biasing simulations were also carried out for additional pair distances of 5 and 6 Å for acetamide in order to be able to connect the forward and backward umbrella sampling sets (see below).

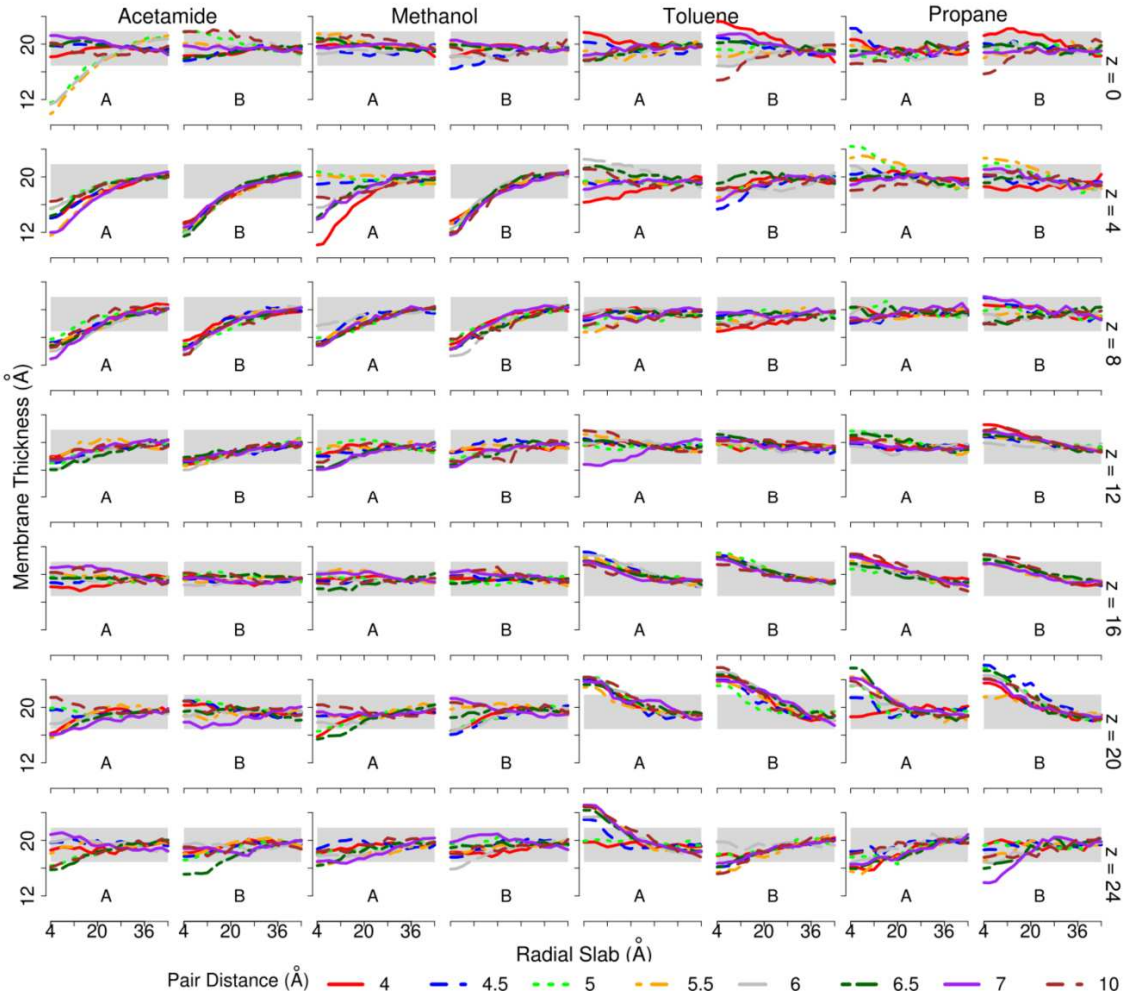

Figure 6-1 Local membrane thickness of the upper leaflet as a function of the radial distance from the center of the amino acid analog pairs at different distances from the center of the membrane. A: sampling in forward direction; B: sampling in backward direction. Local thickness is calculated as average z of phosphorous atoms in the upper leaflet which fall into radial slabs of width 4 Å.

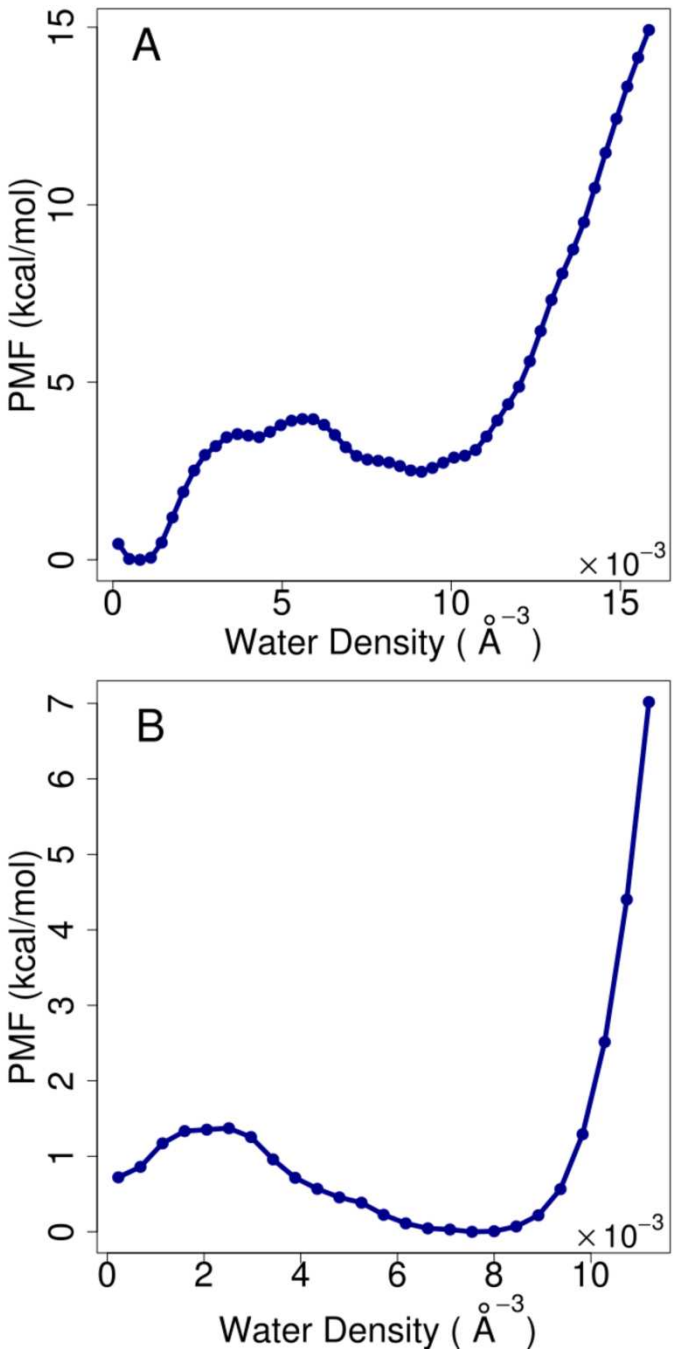

Figure 6-2 Potentials of mean force as a function of water density to reflect membrane deformation. A: acetamide pair at z=0 and d=5.5 Å; B: methanol pair at z=0 and d=4.5 Å.

6.4.2 Association Free Energy Profiles

A main goal of this study is to obtain free energy profiles for amino acid side chain analog association within lipid bilayer environments. Umbrella sampling along the pair

distance reaction coordinate was carried out at different membrane insertion depths and a comparison between forward and backward sampling umbrella runs was used to assess satisfactory convergence. As shown in Figure 6-11, convergence, defined as an RMSD of less than 0.2 kcal/mol between forward and backward runs, was achieved for almost all windows except for acetamide and methanol at certain short distances and deep membrane insertions. These cases correspond to the bi-stable membrane deformation scenario described above where both deformed and undeformed membranes are similarly favorable but transitions between the two states are not sampled in the pair distance umbrella simulations. The additional water-density biasing simulations described above provide access to that transition and a combination of the pair distance umbrella runs with the water-density umbrella runs was necessary to obtain a complete energetic picture. In order to do so, two-dimensional PMFs as a function of pairwise distance and water density were constructed from the combined sampling (see Fig. 6-12 and 6-13) and then integrated using Boltzmann averaging along the density reaction coordinate to obtain correct one-dimensional PMFs as a function of the pair distance. When compared to the naïve case where the pair distance umbrella runs are simply combined without considering that in fact disconnected states are sampled, the corrected PMFs differ by 0.25-0.5 kcal/mol (see Fig. 6-12 and 6-13). For other pairs, distances, and membrane insertions, such a correction was not necessary because forward and backward sampling umbrella appear to have reached convergence.

The complete association energy profiles as a function of pair distance and membrane insertion are presented in Figures 6-3A, 4A, 5A, and 6A. These profiles include the corrected PMF profiles for acetamide at z=0 Å and for methanol at z=4 Å. We note that

because the pairs were fixed at certain insertion depths the present simulations do not provide information about the relative free energies along the z direction. Instead, the PMFs are combined so that the contact pair has the same free energy at all values of z. Information about membrane insertion free energies is available from previous studies while adequate sampling of membrane insertion along with separation within the membrane would have greatly increased the need for additional sampling beyond what we can accomplish with the resources available to us. Overall, the free energy analysis confirms what would be expected qualitatively: both acetamide and methanol have a deep minimum when forming a contact pair inside the membrane but separating the pair becomes increasingly favorable towards the edge of the bilayer where the polar molecules can interact with water rather than with each other. At z=0 the acetamide pair is stabilized by as much as 2.5 kcal/mol while the methanol pair is stabilized by about 1.5 kcal/mol. Toluene and propane pairs on the other hand are slightly more favorable when separated in the membrane by about 0.25 kcal/mol while favoring weak association at the edge of the bilayer as would be expected for hydrophobic compounds. For all compounds there is a ‘desolvation’ peak immediately after separating the contact pair with an energetic penalty of 0.5 to 1 kcal/mol.

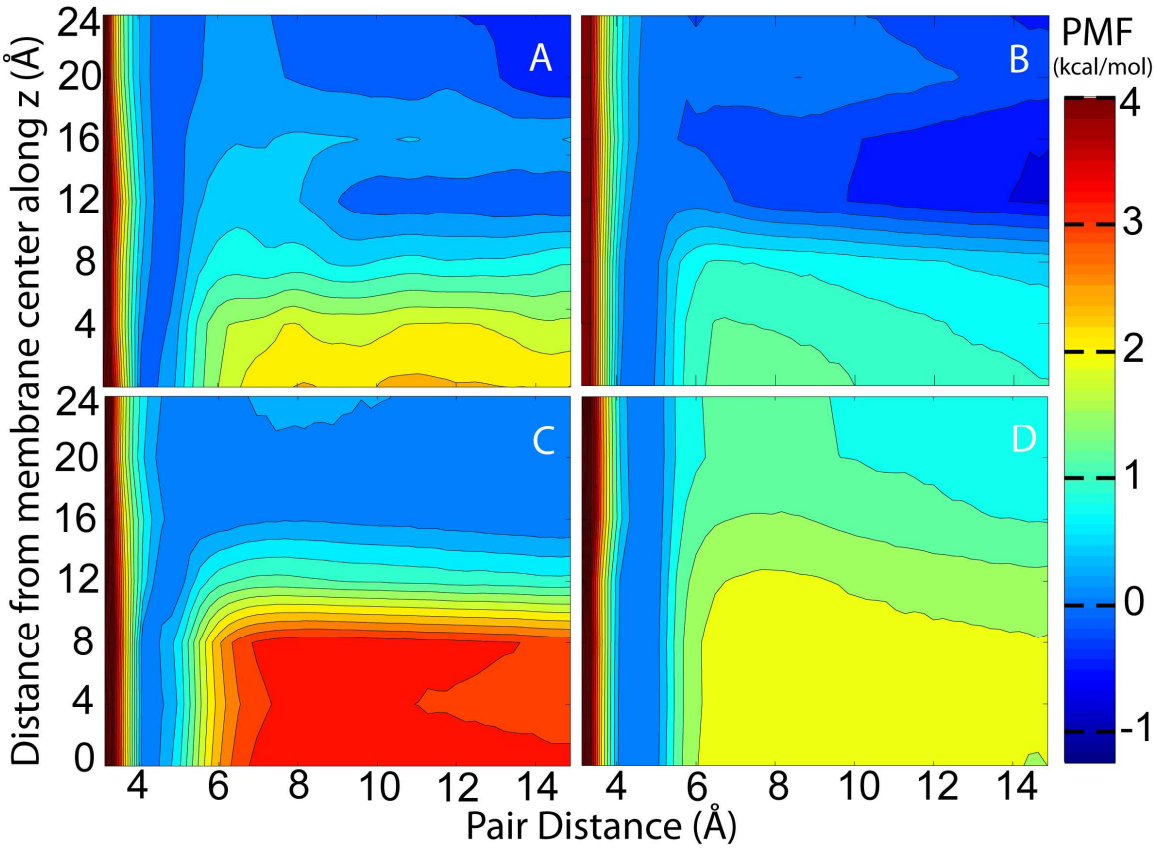

Figure 6-3 Potential of mean force for acetamide as a function of pair distance at different insertion depth into the lipid bilayer from simulations with A) explicit solvent and lipids B) HDGB implicit membrane C) GBSW implicit membrane and D) EEF1 implicit membrane models; For each insertion depth, the bound state was used as the reference with an energy of zero.

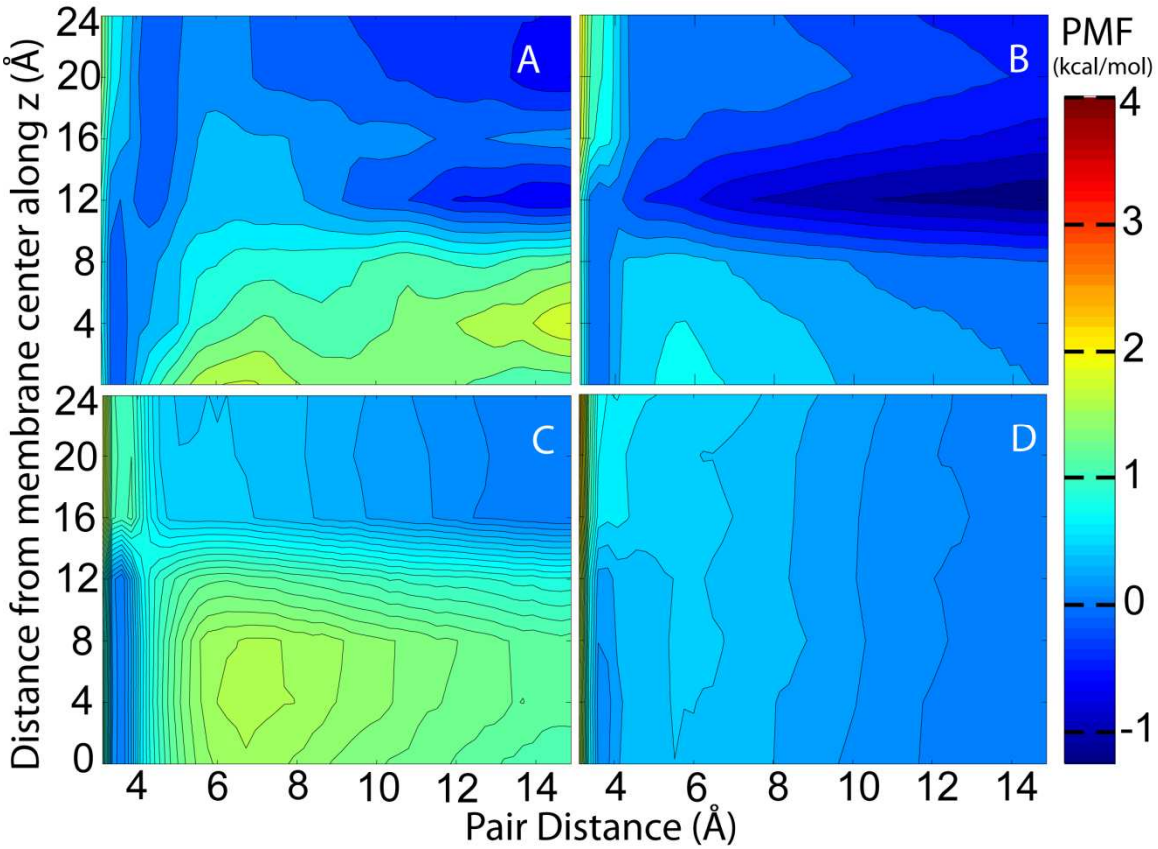

Figure 6-4 PMF of methanol as in Fig. 6-3.

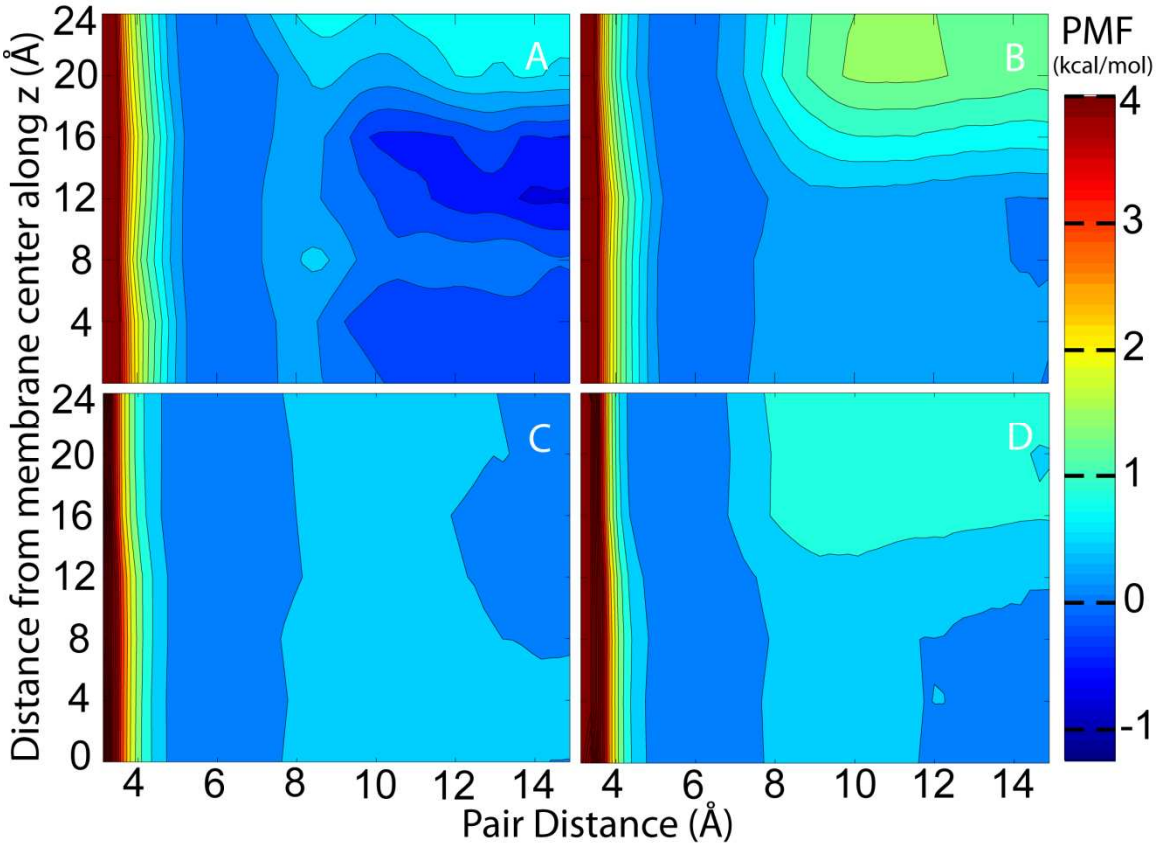

Figure 6-5 PMF of toluene as in Fig. 6-3.

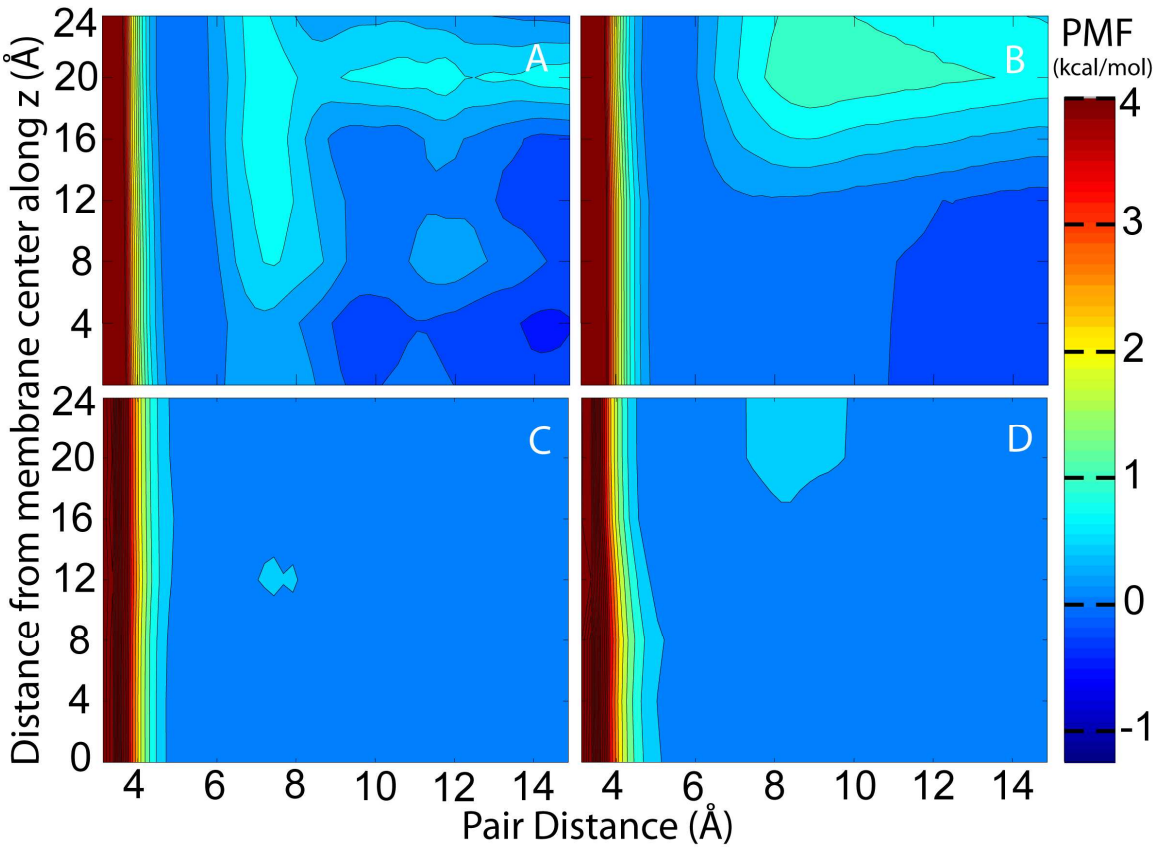

Figure 6-6 PMF of propane as in Fig. 6-3.

The pair binding free energies obtained from the PMF profiles as the difference between the free energy at the contact pair and at the greatest pair distance considered here can be compared to previous results for pair formation in different solvents by de Jong et al[1] with GROMOS[13] and OPLS[14] force fields. More specifically, we compare our results at z=0, 12, and 24 Å insertion depths to the values obtained in decane, octanol and water, respectively. Overall, the agreement is good especially if one considers differences in force fields, the oversimplification of using decane and octanol as mimics of lipid bilayer environments, and the missing contribution due to membrane deformations with the simple hydrophobic solvents. However, taking the data at face value, it appears that

the agreement with the OPLS results is better while GROMOS may be overestimating contact pair formation in decane except for propane.

Table 6-1 Binding free energies in kcal/mol obtained from explicit simulations (CHARMM) at different Z distances as the difference between the free energy for the contact pair and the average energy for distances greater than 10 Å. Standard errors are given in parentheses. Values obtained at z=0, 12 Å, and 24 Å are compared with values obtained previously in decane, octanol, and water by de Jong et al.[1].

		0	4	8	12	16	20	24
Acetamide	CHARMM	-2.54 (0.02)	-1.78 (0.04)	-0.65 (0.05)	-0.17 (0.03)	-0.15 (0.06)	0.03 (0.03)	0.18 (0.01)
	GROMOS[1]	-4.21			0.06			0.13
	OPLS[1]	NA			-0.31			0.04
Methanol	CHARMM	-1.47 (0.03)	-1.74 (0.06)	-0.46 (0.11)	0.18 (0.03)	0.13 (0.05)	0.22 (0.02)	0.24 (0.01)
	GROMOS	-2.72			-0.10			0.34
	OPLS	-1.37			-0.19			0.32
Toluene	CHARMM	0.07 (0.01)	0.19 (0.02)	-0.15 (0.03)	0.33 (0.05)	0.26 (0.02)	-0.68 (0.03)	-0.86 (0.08)
	GROMOS	-1.09			0.21			-0.29
	OPLS	-0.31			-0.03			-0.47
Propane	CHARMM	-0.24 (0.02)	-0.33 (0.03)	-0.53 (0.04)	-0.92 (0.09)	-0.89 (0.09)	-0.86 (0.05)	-0.32 (0.02)
	GROMOS	-0.04			-0.38			-0.06
	OPLS	-0.06			-0.11			-0.06

6.4.3 Contact Pair Formation of Polar Compounds

Closer inspection of the conformations of the polar side-chain pairs (acetamide and methanol) indicate a conformational bias at the bound state as a function of the presence or absence of water molecules around the pair, while conformational analysis of the hydrophobic compounds, on the other hand, did not reveal any noticeable difference along the bilayer normal. We refer to the bound state as the closest pair distance where the association profile is still favorable, while the longest pair distance is referred to as the free state. We observed that relative population of different conformations of acetamide and methanol pairs at the bound state are directly related to the number of

hydrogen bonds they form with water molecules. By clustering acetamide pair conformations at the local minimum of the free energy profiles (d=4 Å), we distinguished three different conformations that could form 0, 1 and 2 hydrogen bonds within the pair. Figure 7 shows the relative population of conformations that form two or one hydrogen bond, as a function of the distance from the bilayer center. At z=0, no hydrogen bond is formed with water molecules because the membrane is not deformed, and as a result the percentage of conformations forming two hydrogen bonds within the pair is 25%. This value decreases as the pair moves to z=4 Å, due to membrane deformation that allow the formation of hydrogen bonds with water.

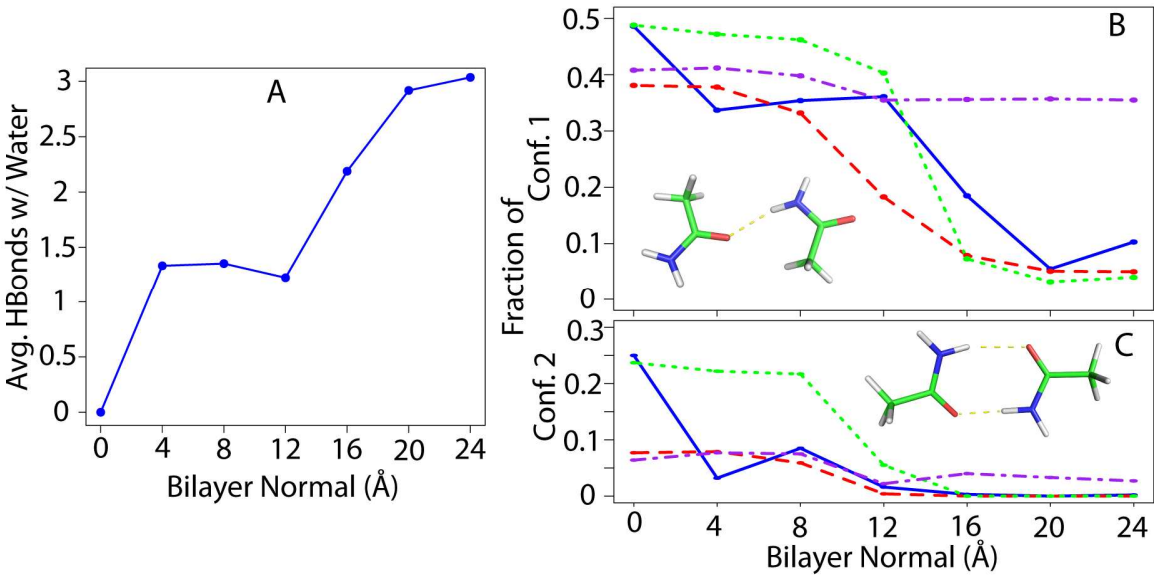

Figure 6-7 Conformational analysis of acetamide pair at the bound state; A) average hydrogen bonds formed between acetamide pair and water molecules as a function of bilayer normal distance, B) fraction of conformations that form one intra-pair hydrogen bond, C) fraction of conformations forming two intra-pair hydrogen bonds

Methanol shows a shift in the contact pair distance from 3.5 Å for z values below 10 Å to a distance of 4.5 Å for z values above 12 Å (see Fig. 4A). The corresponding conformations are shown in Figure 8. At deeper insertion depths the methyl groups are

exposed to the hydrophobic environment while self-interactions between the two hydroxyl groups are maximized, leading to a shorter center of mass distance. On the other hand, at shallower insertion depths, the hydroxyl groups is exposed to the environment while the methyl groups interact with each other so that they are shielded from the more polar environment.

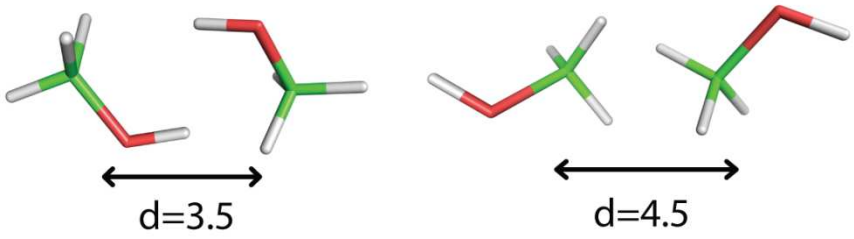

Figure 6-8 Conformational analysis of methanol pair at two possible bound states; A) average hydrogen bonds formed with water, B) fraction of conformation 1 at bound distance 3.5 Å (blue) and at 4.5 Å (red). Results of explicit simulations are shown in solid lines, HDGB in dashed lines, GBSW in dotted line, and IMM1 in dash-dotted line

6.4.4 Comparison with Implicit Membrane Models

The data presented here is especially useful for parameterizing simplified models of membrane environment. Implicit membrane models have been previously parameterized using amino acid side chain insertion free energies but so far little attention has been paid to how well implicit membrane models can capture interactions of solutes within the membrane. Figures 3-6 compare the association free energy PMFs for acetamide, methanol, toluene, and propane with HDGB, GBSW, and IMM1 to the explicit solvent results. Very qualitatively, the main trends are more or less reproduced, but, in detail, there are quite significant differences. For example, GBSW greatly overestimates the binding free energy of acetamide in the membrane while the acetamide contact pair is still more favorable than the separated pair at z=24 Å. In the case of methanol, both

HDGB and IMM1 do not find a significant favorable binding energy at z=0, only GBSW captures the explicit lipid trend correctly. HDGB and GBSW do capture the shift from favoring the hydroxyl-interacting close distance contact pairs at deep insertion to the methyl-interacting longer contact pair beyond 10-12 Å while IMM1 does not. For the non-polar compounds the differences are less dramatic but nevertheless significant when compared to the explicit lipid simulations. For example, GBSW shows little variation as a function of z while HDGB appears to overemphasize the attraction of hydrophobic pairs near the aqueous phase.

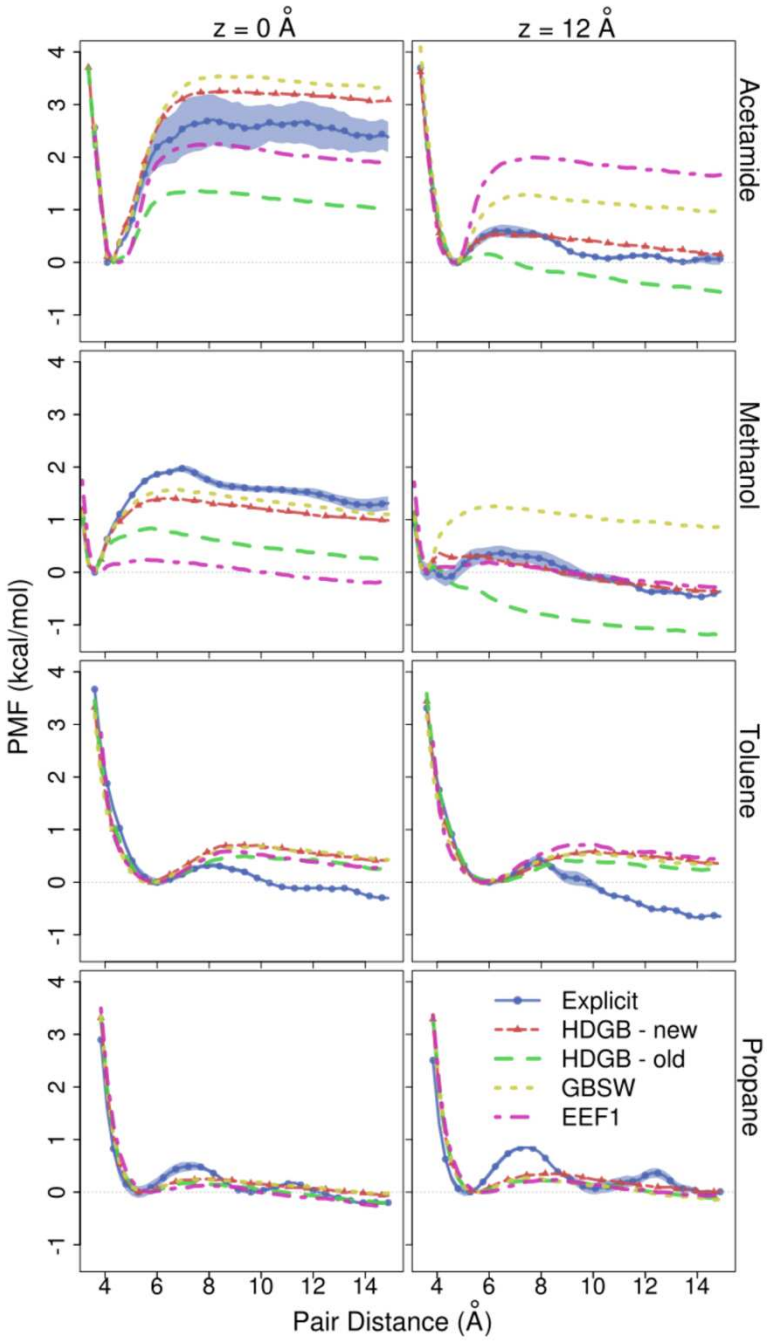

Figure 6-9 PMF profiles for acetamide, methanol, toluene, and propane at Z=0 and Z=12 as a function of the pair distance obtained from explicit, HDGB, GBSW, and EEF1 models

Based on the new data from this study we attempted to improve the parameterization of the HDGB model that was previously developed in our group. Specifically, we adjusted

the dielectric profile as well as the overall scaling factor γ for the non-polar contribution to improve agreement with the pair distance free energies within the membrane while maintaining good agreement with membrane insertion free energies of single amino acid side chain analogs. The overall scaling factor was set to 0.02 kcal/mol/Å^2. The optimized dielectric profile is given in Table 6-3 along with the (unmodified) non-polar profile. Figure 9 focuses on the distance profiles at z=0 and z=12 Å for the four analog pairs with the original and improved HDGB model. As can be seen, it is possible to significantly improve the agreement between the implicit membrane model and the explicit lipid results. At the same time, amino acid insertion profiles for 14 amino acid side-chain analogs are in similar agreement with results from explicit simulation[2] and experimental measurements[49] as for the previous HDGB parameterization (see Fig. S5). Nevertheless, with the modified parameters, the association free energy is now overestimated for acetamide at z=0 while dissociated toluene is still not favorable enough, especially for z=12 Å. The use of an implicit model that would allow membrane deformations such as the DHDGB model[134] that may improve the agreement with the explicit lipid results. Another possibility is the inclusion of implicit van der Waals interactions that are expected to become more important in the membrane environment as the role of electrostatics decreases due to the hydrophobic environment.

6.5 Conclusions

In this study, we are presenting a detailed energetic and structural analysis of amino acid side chain analog interactions within lipid bilayer environments which has received little attention in previous studies. Qualitatively, we confirm expected trends of polar compounds associating strongly inside lipid bilayers compared to hydrophobic compounds. Furthermore, we present detailed quantitative data about the energetics of pair formation at different membrane insertion depths that required a careful analysis of the coupling between amino acid pair interactions and membrane deformations.

The presented data is especially useful for the validation and parameterization of simplified membrane models. We show that established implicit membrane models have difficulties to reproduce the association energetics described here. However, it was possible to improve the HDGB model to better reproduce the new data from this study while maintaining good insertion free energy profiles. In future studies we will aim to further improve the implicit membrane model by considering membrane deformations and implicit van der Waals terms.

6.6 Acknowledgement

Computational resources for this work were provided by XSEDE facilities (TG-MCB090003) and High Performance Computing Center at Michigan State University (HPCC@MSU).

6.7 Supplementary Materials

Table 6-2 Simulation time in nanoseconds for explicit simulations of each amino-acid analog pair under different umbrella potential; the simulation time listed is used in forward and backward directions.

		3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12	12.5	13	13.5	14	14.5	15
Acetamide	0	20	20	30	20	20	200	120	20	20	20	20	20	20	10	10	10	20	20	30	20	20	200	120	20	20
	4	30	70	30	50	200	90	150	50	50	30	30	20	20	10	10	10	30	70	30	50	200	90	150	50	50
	8	20	20	20	20	20	20	20	20	20	20	20	20	20	10	20	10	20	20	20	20	20	20	20	20	20
	12	20	20	20	20	20	20	20	20	20	20	20	20	20	10	10	10	20	20	20	20	20	20	20	20	20
	16	20	20	30	30	30	30	30	30	30	30	30	30	30	40	14	10	20	20	30	30	30	30	30	30	30
	20	20	20	20	20	20	20	20	20	20	20	20	20	20	10	10	10	20	20	20	20	20	20	20	20	20
	24	20	20	20	20	20	20	20	20	20	20	20	20	20	10	10	10	20	20	20	20	20	20	20	20	20
Methanol	0	20	20	20	20	20	20	20	20	50	20	20	20	20	10	10	10	20	20	20	20	20	20	20	20	50
	4	20	20	20	200	30	30	20	20	20	20	20	20	20	10	10	10	20	20	20	200	30	30	20	20	20
	8	20	20	92	40	20	20	20	20	20	30	20	30	20	10	10	10	20	20	92	40	20	20	20	20	20
	12	20	20	20	20	20	20	20	20	20	20	20	20	20	10	10	10	20	20	20	20	20	20	20	20	20
	16	20	20	50	20	20	20	20	20	20	20	20	20	20	10	10	10	20	20	50	20	20	20	20	20	20
	20	20	20	20	20	20	20	20	20	20	20	20	20	20	10	10	10	20	20	20	20	20	20	20	20	20
	24	20	20	20	20	20	20	20	20	20	20	20	20	20	10	10	10	20	20	20	20	20	20	20	20	20
Toluene	0	12	12	12	12	12	12	12	12	12	12	12	12	12	10	10	10	12	12	12	12	12	12	12	12	12
	4	12	12	12	20	12	12	12	12	12	12	12	12	12	10	10	10	12	12	12	20	12	12	12	12	12
	8	12	12	12	12	12	20	12	12	12	12	12	12	12	10	10	10	12	12	12	12	12	20	12	12	12
	12	12	12	12	12	40	40	40	40	20	20	30	20	20	10	62	10	12	12	12	12	40	40	40	40	20
	16	70	70	70	70	40	40	40	40	40	40	40	40	40	40	10	30	70	70	70	70	40	40	40	40	40
	20	12	12	12	12	12	12	20	12	12	12	12	12	12	10	10	10	12	12	12	12	12	12	20	12	12
	24	12	12	12	12	12	12	12	30	30	12	12	12	12	10	10	10	12	12	12	12	12	12	12	30	30
Propane	0	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
	4	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
	8	6	6	6	6	6	20	6	6	6	6	6	20	6	6	20	6	6	6	6	6	6	20	6	6	6
	12	6	6	16	16	16	16	16	16	30	30	30	16	30	30	20	6	6	6	16	16	16	16	16	16	30
	16	6	30	30	30	30	30	30	30	40	100	40	60	30	20	20	6	6	30	30	30	30	30	30	30	40
	20	6	6	6	6	6	6	6	6	20	20	6	6	6	6	20	20	6	6	6	6	6	6	6	6	20
	24	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6

Table 6-3 Improved HDGB parameters, dielectric profile (ε) and non-polar profile (γ).

Z	ε(z) - old	ε(z) - new
0.0	1.80	1.15
0.5	1.80	1.15
1.0	1.81	1.16
1.5	1.82	1.17
2.0	1.83	1.18
2.5	1.84	1.19
3.0	1.85	1.20
3.5	1.86	1.21
4.0	1.87	1.22
4.5	1.89	1.24
5.0	1.91	1.26
5.5	1.93	1.28
6.0	1.97	1.32
6.5	2.00	1.35
7.0	2.04	1.49
7.5	2.09	1.64
8.0	2.15	1.80
8.5	2.22	1.87
9.0	2.31	1.96
9.5	2.41	2.06
10.0	2.53	2.18
10.5	3.23	2.28
11.0	3.63	2.38
11.5	4.13	2.58
12.0	4.73	2.89
12.5	5.43	3.42
13.0	6.13	4.00
13.5	6.98	5.08
14.0	7.84	6.04
14.5	8.80	7.50
15.0	10.96	10.06
15.5	14.05	13.75
16.0	19.04	19.04
16.5	25.85	25.85
17.0	35.38	35.38
17.5	45.88	45.88
18.0	54.11	54.11
18.5	60.79	60.79
19.0	65.52	65.52
19.5	69.42	69.42
20.0	72.31	72.31
20.5	74.07	74.07
21.0	75.53	75.53
21.5	76.63	76.63
22.0	77.14	77.14
22.5	77.83	77.83
23.0	78.22	78.22
23.5	78.92	78.92
24.0	79.35	79.35
24.5	79.66	79.66
25.0	80.00	80.00

z	γ(z)
0.0	0.0000
0.6	0.0001
1.2	0.0002
1.8	0.0010
2.4	0.0050
3.0	0.0075
3.6	0.0100
4.2	0.0150
4.8	0.0200
5.4	0.0250
6.0	0.0300
6.6	0.0350
7.2	0.0410
7.8	0.0470
8.4	0.0520
9.0	0.0610
9.6	0.0720
10.2	0.0850
10.8	0.1000
11.4	0.1200
12.0	0.1500
12.6	0.1900
13.2	0.2500
13.8	0.3200
14.4	0.4000
15.0	0.5000
15.6	0.6200
16.2	0.7500
16.8	0.8753
17.4	0.9500
18.0	1.0308
18.6	1.0680
19.2	1.0921
19.8	1.1000
20.4	1.1000
21.0	1.0927
21.6	1.0690
22.2	1.0468
22.8	1.0328
23.4	1.0197
24.0	1.0130
24.6	1.0052
25.2	1.0005
25.8	1
26.4	1
27.0	1
27.6	1
28.2	1
28.8	1
29.4	1
30.0	1

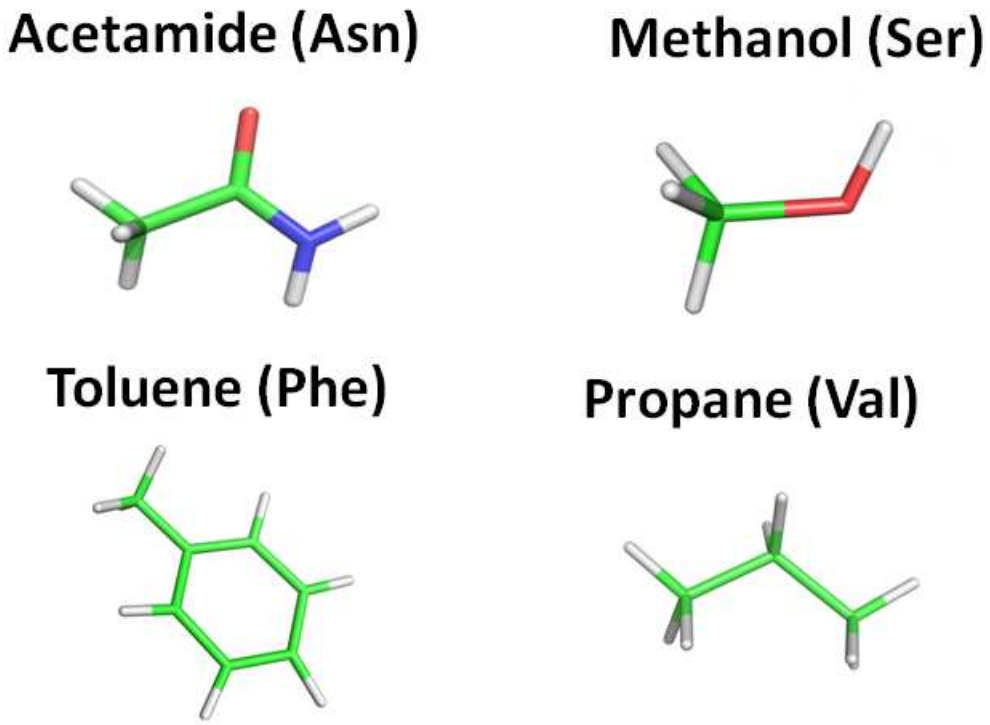

Figure 6-10 Amino acid analogs used in this study, acetamide (Asn), methanol (Ser), toluene (Phe) and propane (Val).

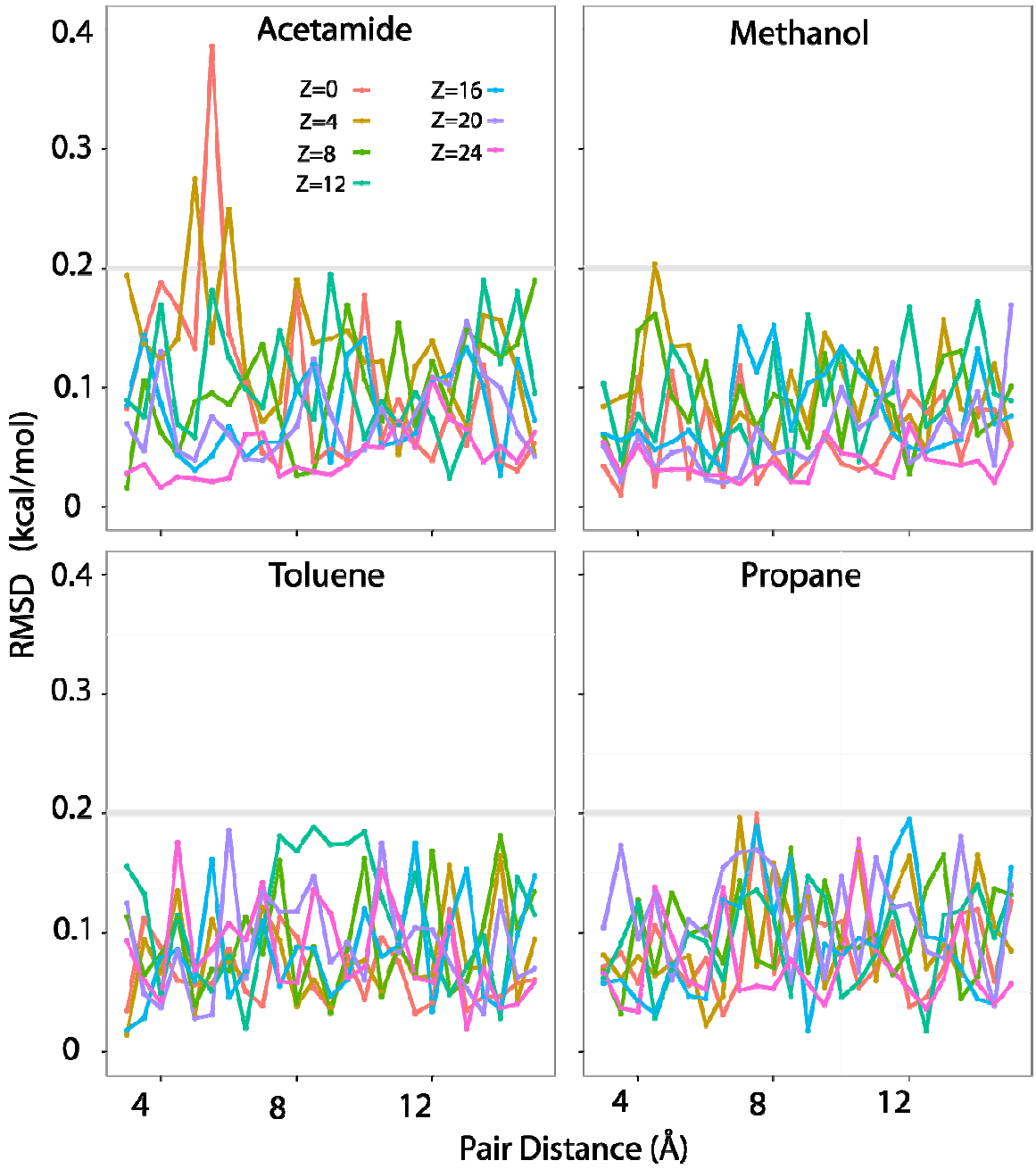

Figure 6-11 Root mean squared deviation between the PMF profiles of biased simulations in forward and backward pulling directions.

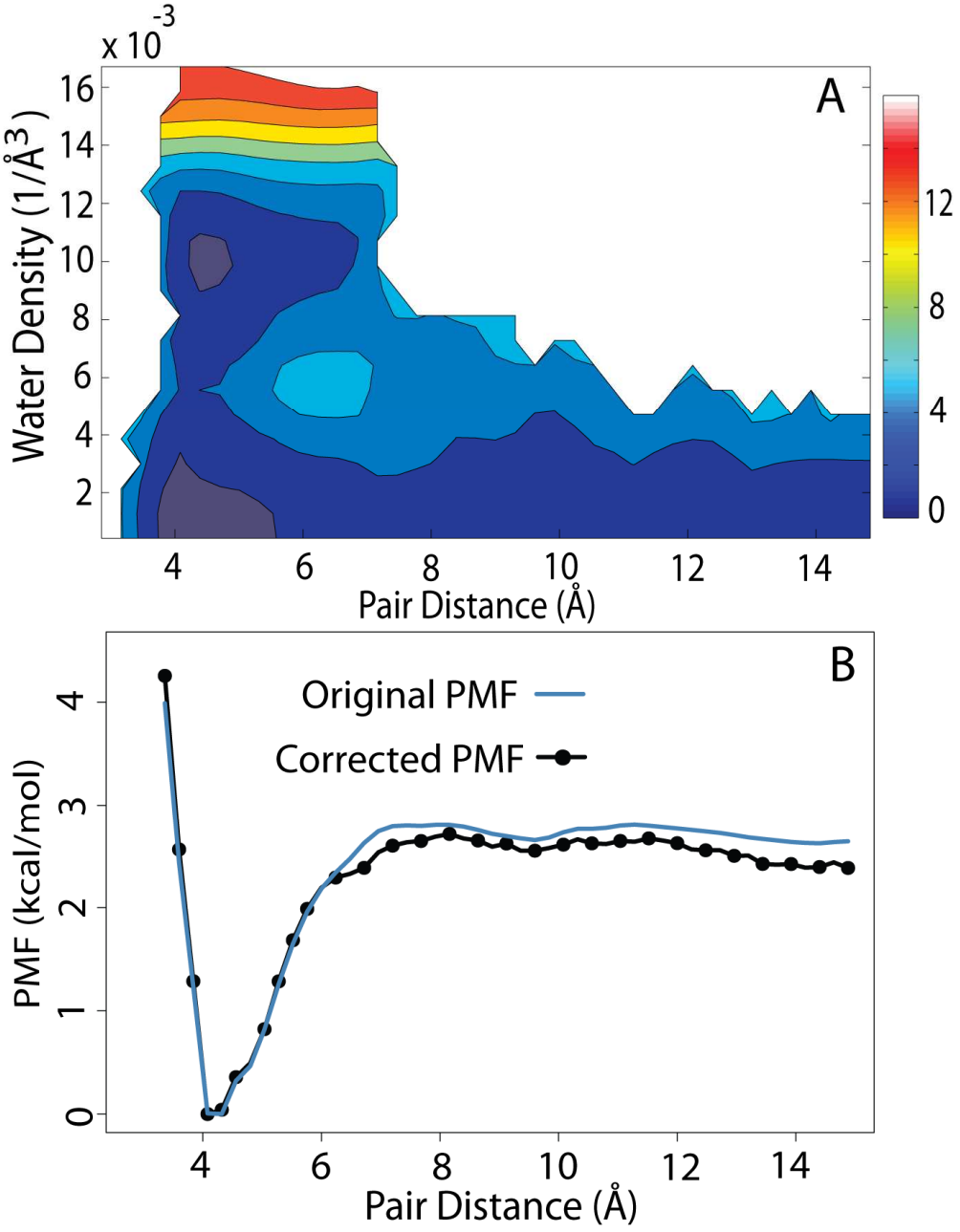

Figure 6-12 A: 2D PMF of acetamide pair association at z=0 and water density in biasing cylinder; B: Corrected 1D PMF as a function of pair distance after Boltzmann averaging along the water density reaction coordinate.

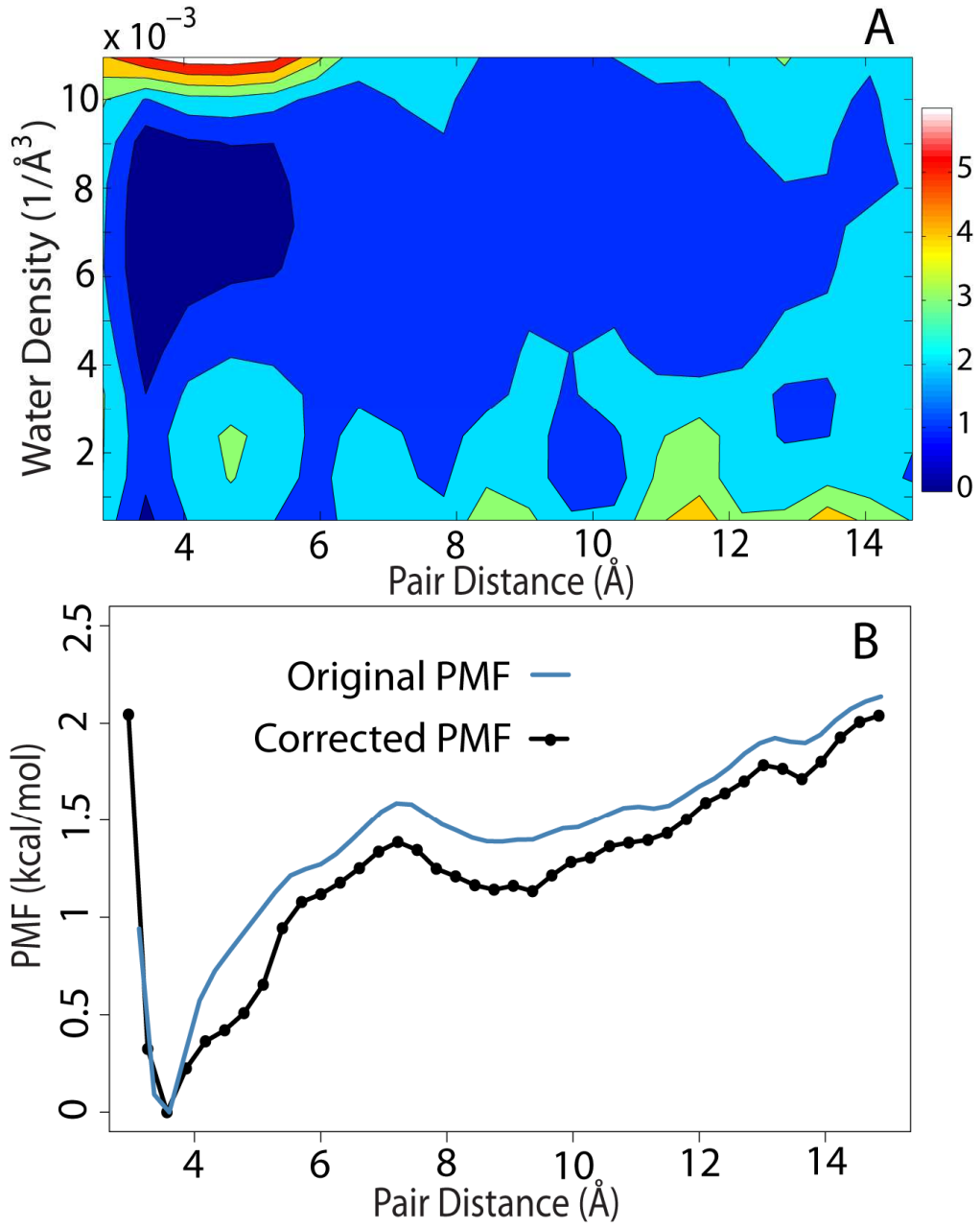

Figure 6-13 A: 2D PMF of methanol pair association at z=4 Å and water density in biasing cylinder; B: Corrected 1D PMF as a function of pair distance after Boltzmann averaging along the water density reaction coordinate.

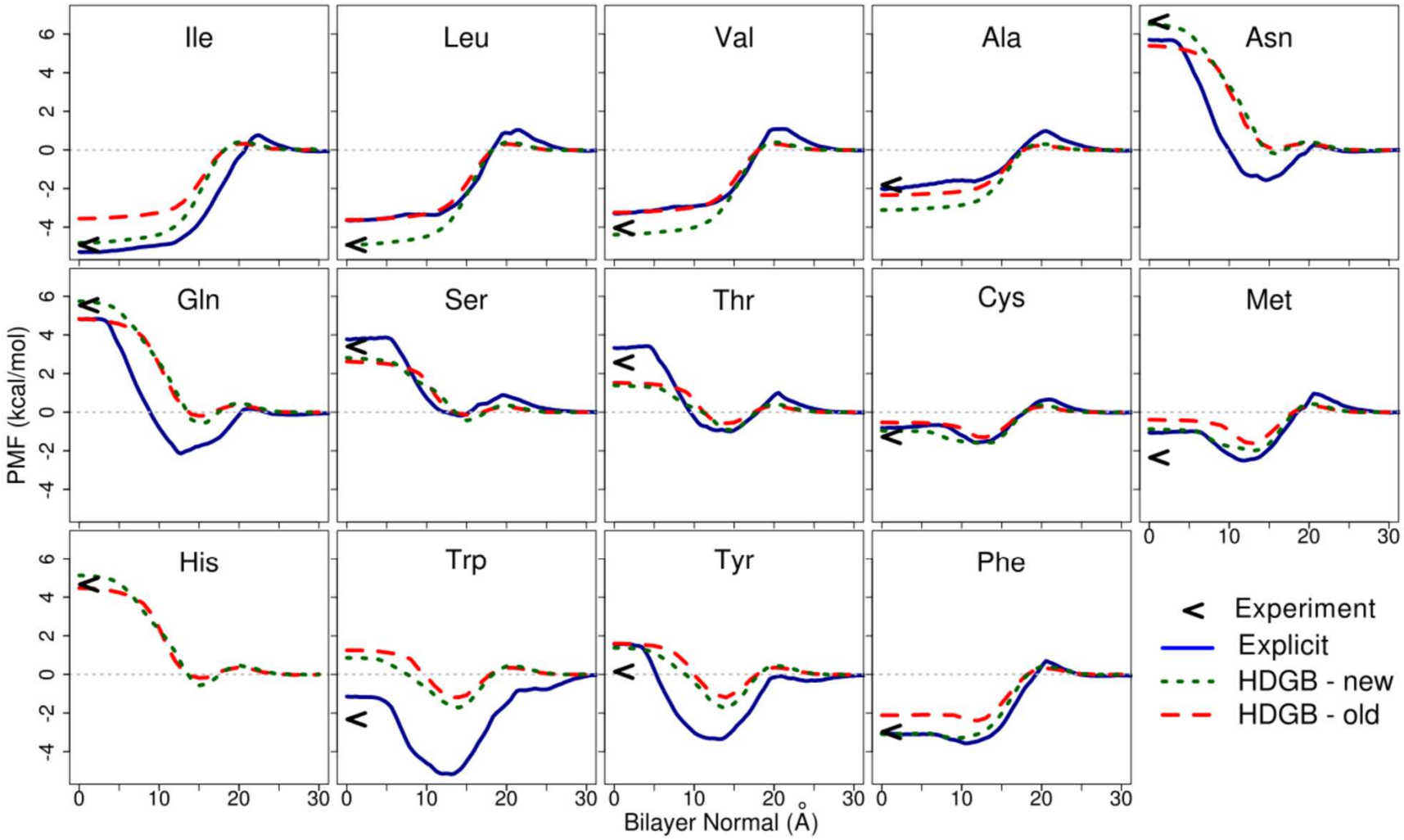

Figure 6-14 Free energy profiles of insertion of single amino acid side-chain analogs using HDGB simulations with old and improved parameters compared with results of explicit simulation and experimental measurements.

Chapter 7

Conclusion and Perspective

Molecular dynamics simulation is a powerful computational technique that gives useful insights to understand the physical characteristics and detailed dynamical information of biophysical systems, such as proteins, membranes, and nucleic acids. Molecular dynamics combined with enhanced sampling methods, such as umbrella sampling, can be used to estimate the free energy and other thermodynamic properties of such systems.

In this dissertation, we used molecular dynamics to investigate two goals in biophysical systems. The first goal was to improve/refine protein structures starting from a homology model and develop a robust MD-based protocol for structure refinement, and the second goal was to understand and characterize amino-acid interactions within membrane environments.

A robust protocol for structure refinement of proteins models was developed. This protocol was applied to CASP10 targets, and 23 out of 27 targets were successfully refined. To this date, our methodology has remained state of the art solution to protein structure refinement. The key winning factor in our method is optimal subset selection and structure averaging, which was introduced for the first time in protein structure refinement. With the aforementioned success of our method, we applied this protocol to CASP11 with minor modifications; Sampling by MD was extended to 1.2 μs, as well as changes in restraint and subset selection algorithm. Finally, the outcome of this method with such modification has resulted in further improvement. Indeed, breakthrough results are achieved, in which in 3 cases, we have obtained up to 20% improvements in GDT-HA. While the results of structure refinement in CASP11 will be addressed thoroughly in future, we need to understand the effect of our modifications to the protocol.

As the second aim of this dissertation, we characterized the association free energy of four amino acid side-chain analog pairs (acetamide (Asn), methanol (Ser), toluene (Phe) and propane (Val)) within membrane bilayer at different distances from bilayer center. Throughout this study, it was observed that acetamide and methanol can create two separate states. The bilayer could be flat or deformed with the same position of those analog pairs, while the relative free energy between flat and deformed state was unknown. Therefore, in order to measure the free energy difference between flat and bilayer states with polar compounds placed at specific bilayer normal, we developed a new computational tool to study free energy of bilayer deformation under umbrella sampling framework. This methodology uses density of water molecules in a cylinder as a reaction coordinate. With application of this methodology, the association free energy surface of acetamide and methanol were corrected through Boltzmann averaging of PMF profiles as functions of association pair distance and water density.

Polar compounds in this study, i.e. acetamide and methanol, showed favorable binding free energy at bilayer center, while this effect diminishes as the pair is moved toward water region. On the other hand, non-polar compounds, toluene and propane showed the opposite behavior. This result provides a useful benchmark for understanding peptide-membrane interactions, as well as a valuable tool for comparison and parameterization of other membrane models. In order to improve the performance of HDGB model, we have also re-parameterized HDGB and modified the dielectric profile by comparing the association free energies of these analog pairs against the obtained PMF profiles in explicit simulations. The new HDGB dielectric profile is made available through this study.

Prior to this study, the field of protein structure refinement had remained steady with very little progress. We stand as the pioneers of structural subset selection and structure averaging for

protein structure refinement. The computational approach developed for membrane deformation and pore formation provides a useful tool for study of membrane bilayer stability under different stress conditions. The protocols and tools developed in this dissertation are freely available to the greater scientific community.

REFERENCES

REFERENCES

[1] D.H. de Jong, X. Periole, S.J. Marrink, Dimerization of Amino Acid Side Chains: Lessons from the Comparison of Different Force Fields, Publisher, City, 2012.

[2] J.L. MacCallum, W.F.D. Bennett, D.P. Tieleman, Distribution of Amino Acids in a Lipid Bilayer from Computer Simulations, Publisher, City, 2008.

[3] S.A. Adcock, J.A. McCammon, Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins, Publisher, City, 2006.

[4] L. Fang, H.J. Cho, C. Chan, M. Feig, Binding site multiplicity with fatty acid ligands: Implications for the regulation of PKR kinase autophosphorylation with palmitate: Palmitate-Kinase Binding Multiplicity, Publisher, City, 2014.

[5] B. Wang, Alexander V. Predeus, Zachary F. Burton, M. Feig, Energetic and Structural Details of the Trigger-Loop Closing Transition in RNA Polymerase II, Publisher, City, 2013.

[6] M. Feig, Z.F. Burton, RNA polymerase II with open and closed trigger loops: active site dynamics and nucleic acid translocation, Publisher, City, 2010.

[7] E.I. Chudyk, M.A.L. Limb, C. Jones, J. Spencer, M.W. van der Kamp, A.J. Mulholland, QM/MM simulations as an assay for carbapenemase activity in class A β-lactamases, Publisher, City, 2014.

[8] M. Karplus, J.A. McCammon, Molecular dynamics simulations of biomolecules, Publisher, City, 2002.

[9] R.B. Best, X. Zhu, J. Shim, P.E.M. Lopes, J. Mittal, M. Feig, A.D. MacKerell, Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone , ψ and Side-Chain χ1 and χ2 Dihedral Angles, Publisher, City, 2012.

[10] R.W. Pastor, A.D. MacKerell, Development of the CHARMM Force Field for Lipids, Publisher, City, 2011.

[11] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, A.D. Mackerell, CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields, Publisher, City, 2010.

[12] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, Publisher, City, 1995.

[13] C. Oostenbrink, A. Villa, A.E. Mark, W.F. Van Gunsteren, A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6, Publisher, City, 2004.

[14] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids, Publisher, City, 1996.

[15] W.L. Jorgensen, J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, Publisher, City, 1988.

[16] S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, A.H. de Vries, The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations, Publisher, City, 2007.

[17] P. Kar, S.M. Gopal, Y.-M. Cheng, A. Predeus, M. Feig, PRIMO: A Transferable Coarse-Grained Force Field for Proteins, Publisher, City, 2013.

[18] S.M. Gopal, S. Mukherjee, Y.-M. Cheng, M. Feig, PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy, Publisher, City, 2010.

[19] A.D. Mackerell, M. Feig, C.L. Brooks, Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, Publisher, City, 2004.

[20] B. Guillot, A reappraisal of what we have learnt during three decades of computer simulations on water, Publisher, City, 2002.

[21] P. Mark, L. Nilsson, Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K, Publisher, City, 2001.

[22] J. Chen, C.L. Brooks, J. Khandogin, Recent advances in implicit solvent-based methods for biomolecular simulations, Publisher, City, 2008.

[23] W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, Publisher, City, 1990.

[24] W. Im, M.S. Lee, C.L. Brooks, Generalized born model with a simple smoothing function, Publisher, City, 2003.

[25] M.S. Lee, F.R. Salsbury, C.L. Brooks, Novel generalized Born methods, Publisher, City, 2002.

[26] M.S. Lee, M. Feig, F.R. Salsbury, C.L. Brooks, New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations, Publisher, City, 2003.

[27] A. Onufriev, D. Bashford, D.A. Case, Modification of the Generalized Born Model Suitable for Macromolecules, Publisher, City, 2000.

[28] M. Feig, W. Im, C.L. Brooks, Implicit solvation based on generalized Born theory in different dielectric environments, Publisher, City, 2004.

[29] S. Tanizaki, M. Feig, A generalized Born formalism for heterogeneous dielectric environments: Application to the implicit modeling of biological membranes, Publisher, City, 2005.

[30] M. Sayadi, M. Feig, Role of conformational sampling of Ser16 and Thr17-phosphorylated phospholamban in interactions with SERCA, Publisher, City, 2013.

[31] D.A. McQuarrie, Statistical mechanics, Harper & Row, New York, 1975.

[32] G.M. Torrie, J.P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling, Publisher, City, 1977.

[33] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, P.A. Kollman, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method, Publisher, City, 1992.

[34] A. Raval, S. Piana, M.P. Eastwood, R.O. Dror, D.E. Shaw, Refinement of protein structure homology models via long, all-atom molecular dynamics simulations, Publisher, City, 2012.

[35] J.L. MacCallum, A. Pérez, M.J. Schnieders, L. Hua, M.P. Jacobson, K.A. Dill, Assessment of protein structure refinement in CASP9, Publisher, City, 2011.

[36] J.L. MacCallum, L. Hua, M.J. Schnieders, V.S. Pande, M.P. Jacobson, K.A. Dill, Assessment of the protein-structure refinement category in CASP8, Publisher, City, 2009.

[37] V. Mirjalili, M. Feig, Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles, Publisher, City, 2012.

[38] V. Mirjalili, K. Noyes, M. Feig, Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging, Publisher, City, 2014.

[39] A. Jagielska, L. Wroblewska, J. Skolnick, Protein model refinement using an optimized physics-based all-atom force field, Publisher, City, 2008.

[40] M.S. Lin, T. Head-Gordon, Reliable Protein Structure Refinement Using a Physical Energy Function, Publisher, City, 2011.

[41] J. Zhu, H. Fan, X. Periole, B. Honig, A.E. Mark, Refining homology models by combining replica-exchange molecular dynamics and statistical potentials, Publisher, City, 2008.

[42] G. Chopra, N. Kalisman, M. Levitt, Consistent refinement of submitted models at CASP using a knowledge-based potential, Publisher, City, 2010.

[43] J.P.G.L.M. Rodrigues, M. Levitt, G. Chopra, KoBaMIN: a knowledge-based minimization web server for protein structure refinement, Publisher, City, 2012.

[44] Y. Yang, Y. Zhou, Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions, Publisher, City, 2008.

[45] C. Zhang, S. Liu, H. Zhou, Y. Zhou, The Dependence of All-Atom Statistical Potentials on Structural Training Database, Publisher, City, 2004.

[46] H. Zhou, Y. Zhou, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, Publisher, City, 2009.

[47] E. Faraggi, A. Kloczkowski, A global machine learning based scoring function for protein structure prediction: Global Knowledge-Based Function for Protein Structure, Publisher, City, 2014.

[48] J. Zhang, Y. Zhang, A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction, Publisher, City, 2010.

[49] A. Radzicka, R. Wolfenden, Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution, Publisher, City, 1988.

[50] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, Publisher, City, 2010.

[51] P. Bradley, K.M.S. Misura, D. Baker, Toward High-Resolution de Novo Structure Prediction for Small Proteins, Publisher, City, 2005.

[52] A. Šali, T.L. Blundell, Comparative Protein Modelling by Satisfaction of Spatial Restraints, Publisher, City, 1993.

[53] B.K. Lance, C.M. Deane, G.R. Wood, Exploring the potential of template-based modelling, Publisher, City, 2010.

[54] K. Joo, J. Lee, S. Lee, J.-H. Seo, S.J. Lee, J. Lee, High accuracy template based modeling by global optimization, Publisher, City, 2007.

[55] J. Moult, A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction, Publisher, City, 2005.

[56] Y. Zhang, I-TASSER server for protein 3D structure prediction, Publisher, City, 2008.

[57] Y. Zhang, Protein structure prediction: when is it useful?, Publisher, City, 2009.

[58] D. Fischer, 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor, Publisher, City, 2003.

[59] K. Ginalski, A. Elofsson, D. Fischer, L. Rychlewski, 3D-Jury: a simple approach to improve protein structure predictions, Publisher, City, 2003.

[60] K.M.S. Misura, D. Baker, Progress and challenges in high-resolution refinement of protein structure models, Publisher, City, 2005.

[61] H. Berman, K. Henrick, H. Nakamura, J.L. Markley, The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data, Publisher, City, 2007.

[62] T.Y. Liu, G.W. Tang, E. Capriotti, Comparative Modeling: The State of the Art and Protein Drug Target Structure Prediction, Publisher, City, 2011.

[63] M. Takeda-Shitaka, D. Takaya, C. Chiba, H. Tanaka, H. Umeyama, Protein structure prediction in structure based drug design, Publisher, City, 2004.

[64] A. Giorgetti, D. Raimondo, A.E. Miele, A. Tramontano, Evaluating the usefulness of protein structure models for molecular replacement, Publisher, City, 2005.

[65] M.R. Lee, J. Tsai, D. Baker, P.A. Kollman, Molecular dynamics in the endgame of protein structure prediction, Publisher, City, 2001.

[66] J. Chen, C.L. Brooks, Can molecular dynamics simulations provide high-resolution refinement of protein structure?, Publisher, City, 2007.

[67] A. Jagielska, L. Wroblewska, J. Skolnick, Protein model refinement using an optimized physics-based all-atom force field, Publisher, City, 2008.

[68] M.S. Lee, M.A. Olson, Assessment of detection and refinement strategies for de novo protein structures using force field and statistical potentials, Publisher, City, 2007.

[69] H. Fan, A.E. Mark, Refinement of homology-based protein structures by molecular dynamics simulation techniques, Publisher, City, 2004.

[70] A. Raval, S. Piana, M.P. Eastwood, R.O. Dror, D.E. Shaw, Refinement of protein structure homology models via long, all-atom molecular dynamics simulations, Publisher, City, 2012.

[71] H. Fan, X. Periole, A.E. Mark, Mimicking the action of folding chaperones by Hamiltonian replica-exchange molecular dynamics simulations: Application in the refinement of de novo models, Publisher, City, 2012.

[72] R. Ishitani, T. Terada, K. Shimizu, Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations, Publisher, City, 2008.

[73] A.W. Stumpff-Kane, K. Maksimiak, M.S. Lee, M. Feig, Sampling of near-native protein conformations during protein structure refinement using a coarse-grained model, normal modes, and molecular dynamics simulations, Publisher, City, 2008.

[74] D.E. Kim, B. Blum, P. Bradley, D. Baker, Sampling Bottlenecks in De novo Protein Structure Prediction, Publisher, City, 2009.

[75] C.M. Summa, M. Levitt, Near-native structure refinement using in vacuo energy minimization, Publisher, City, 2007.

[76] D.W. Li, R. Bruschweiler, Dynamic and Thermodynamic Signatures of Native and Non-Native Protein States with Application to the Improvement of Protein Structures, Publisher, City, 2012.

[77] G. Chopra, C.M. Summa, M. Levitt, Solvent dramatically affects protein structure refinement, Publisher, City, 2008.

[78] G. Chopra, N. Kalisman, M. Levitt, Consistent refinement of submitted models at CASP using a knowledge-based potential, Publisher, City, 2010.

[79] H. Lu, J. Skolnick, Application of statistical potentials to protein structure refinement from low resolution Ab initio models, Publisher, City, 2003.

[80] C. Zhang, S. Liu, Y.Q. Zhou, Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential, Publisher, City, 2004.

[81] J. Zhang, Y. Liang, Y. Zhang, Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling, Publisher, City, 2011.

[82] J. Zhu, L. Xie, B. Honig, Structural refinement of protein segments containing secondary structure elements: Local sampling, knowledge-based potentials, and clustering, Publisher, City, 2006.

[83] M.A. Olson, S. Chaudhury, M.S. Lee, Comparison Between Self-Guided Langevin Dynamics and Molecular Dynamics Simulations for Structure Refinement of Protein Loop Conformations, Publisher, City, 2011.

[84] Y. Yang, Y. Zhou, Specific interactions for ab initio folding of protein terminal regions with secondary structures, Publisher, City, 2008.

[85] A. Zemla, C. Venclovas, J. Moult, K. Fidelis, Processing and analysis of CASP3 protein structure predictions, Publisher, City, 1999.

[86] B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu,

W. Yang, D.M. York, M. Karplus, CHARMM: The biomolecular simulation program, Publisher, City, 2009.

[87] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, Publisher, City, 1983.

[88] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER, Publisher, City, 1983.

[89] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus, All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins, Publisher, City, 1998.

[90] A.D. MacKerell, M. Feig, C.L. Brooks, Improved treatment of the protein backbone in empirical force fields, Publisher, City, 2004.

[91] V.B. Chen, W.B. Arendall, III, J.J. Headd, D.A. Keedy, R.M. Immormino, G.J. Kapral, L.W. Murray, J.S. Richardson, D.C. Richardson, MolProbity: all-atom structure validation for macromolecular crystallography, Publisher, City, 2010.

[92] M. Feig, J. Karanicolas, C.L. Brooks, MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology, Publisher, City, 2004.

[93] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, K. Schulten, Scalable molecular dynamics with NAMD, Publisher, City, 2005.

[94] A. Zemla, LGA: a method for finding 3D similarities in protein structures, Publisher, City, 2003.

[95] L.S.D. Caves, J.D. Evanseck, M. Karplus, Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin, Publisher, City, 1998.

[96] M.I. Zavodszky, A.W. Stumpff-Kane, D.J. Lee, M. Feig, Scoring confidence index: statistical evaluation of ligand binding mode predictions, Publisher, City, 2009.

[97] A.W. Stumpff-Kane, M. Feig, A correlation-based method for the enhancement of scoring functions on funnel-shaped energy landscapes, Publisher, City, 2006.

[98] V. Mariani, F. Kiefer, T. Schmidt, J. Haas, T. Schwede, Assessment of template based protein structure predictions in CASP9, Publisher, City, 2011.

[99] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, Publisher, City, 2010.

[100] A. Kryshtafovych, C. Venclovas, K. Fidelis, J. Moult, Progress over the first decade of CASP experiments, Publisher, City, 2005.

[101] J. Moult, K. Fidelis, B. Rost, T. Hubbard, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP) - Round 6, Publisher, City, 2005.

[102] Y. Zhang, J. Skolnick, The protein structure prediction problem could be solved using the current PDB library, Publisher, City, 2005.

[103] K. Henrick, Z.K. Feng, W.F. Bluhm, D. Dimitropoulos, J.F. Doreleijers, S. Dutta, J.L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, C.L. Lawson, J.L. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, E.L. Ulrich, W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin, H.M. Berman, Remediation of the protein data bank archive, Publisher, City, 2008.

[104] A. Bazzoli, A.G.B. Tettamanzi, Y. Zhang, Computational Protein Design and Large-Scale Assessment by I-TASSER Structure Assembly Simulations, Publisher, City, 2011.

[105] A. Fiser, A. Sali, MODELLER: Generation and refinement of homology-based protein structure models, in: C.W. Carter, R.M. Sweet (Eds.) Macromolecular Crystallography, Pt D, Elsevier Academic Press Inc, San Diego, 2003, pp. 461-+.

[106] D. Xu, Y. Zhang, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field, Publisher, City, 2012.

[107] Y. Zhang, J. Skolnick, Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins, Publisher, City, 2004.

[108] J. Ko, H. Park, C. Seok, GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions, Publisher, City, 2012.

[109] X.T. Qu, R. Swanson, R. Day, J. Tsai, A Guide to Template Based Structure Prediction, Publisher, City, 2009.

[110] J. Zhang, Y. Liang, Y. Zhang, Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling, Publisher, City, 2011.

[111] J.H. Chen, C.L. Brooks, Can molecular dynamics simulations provide high-resolution refinement of protein structure?, Publisher, City, 2007.

[112] D. Bhattacharya, J.L. Cheng, 3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization, Publisher, City, 2013.

[113] J. Rodrigues, M. Levitt, G. Chopra, KoBaMIN: a knowledge-based minimization web server for protein structure refinement, Publisher, City, 2012.

[114] J. Zhu, H. Fan, X. Periole, B. Honig, A.E. Mark, Refining homology models by combining replica-exchange molecular dynamics and statistical potentials, Publisher, City, 2008.

[115] M.A. Olson, M.S. Lee, Structure refinement of protein model decoys requires accurate side-chain placement, Publisher, City, 2013.

[116] L. Wroblewska, A. Jagielska, J. Skolnick, Development of a physics-based force field for the scoring and refinement of protein models, Publisher, City, 2008.

[117] J.W. Ponder, D.A. Case, Force fields for protein simulations, Publisher, City, 2003.

[118] R.B. Best, X. Zhu, J. Shim, P.E.M. Lopes, J. Mittal, M. Feig, A.D. MacKerell, Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles, Publisher, City, 2012.

[119] Y.D. Yang, Y.Q. Zhou, Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions, Publisher, City, 2008.

[120] Y.D. Yang, Y.Q. Zhou, Specific interactions for ab initio folding of protein terminal regions with secondary structures, Publisher, City, 2008.

[121] H.Y. Zhou, J. Skolnick, GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction, Publisher, City, 2011.

[122] M.Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures, Publisher, City, 2006.

[123] H. Fan, X. Periole, A.E. Mark, Mimicking the action of folding chaperones by Hamiltonian replica-exchange molecular dynamics simulations: Application in the refinement of de novo models, Publisher, City, 2012.

[124] D. Gront, S. Kmiecik, M. Blaszczyk, D. Ekonomiuk, A. Kolinski, Optimization of protein models, Publisher, City, 2012.

[125] M.H.M. Olsson, C.R. Søndergaard, M. Rostkowski, J.H. Jensen, PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions, Publisher, City, 2011.

[126] C.R. Søndergaard, M.H.M. Olsson, M. Rostkowski, J.H. Jensen, Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values, Publisher, City, 2011.

[127] W.L. DeLano, PyMOL molecular viewer: Updates and refinements, Publisher, City, 2009.

[128] W.L. DeLano, J.W. Lam, PyMOL: A communications tool for computational models, Publisher, City, 2005.

[129] S.-Y. Huang, X. Zou, An iterative knowledge-based scoring function for protein-protein recognition, Publisher, City, 2008.

[130] H. Zhou, J. Skolnick, GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction, Publisher, City, 2011.

[131] M. Lu, A.D. Dousis, J. Ma, OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing, Publisher, City, 2008.

[132] M.-y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures, Publisher, City, 2006.

[133] S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, A.H. de Vries, The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations, Publisher, City, 2007.

[134] A. Panahi, M. Feig, Dynamic Heterogeneous Dielectric Generalized Born (DHDGB): An Implicit Membrane Model with a Dynamically Varying Bilayer Thickness, Publisher, City, 2013.

[135] M. Feig, Implicit Membrane Models for Membrane Protein Simulation, in: A. Kukol (Ed.) Molecular Modeling of Proteins, Humana Press, 2008, pp. 181-196.

[136] Y.Z. Ohkubo, Taras V. Pogorelov, Mark J. Arcario, Geoff A. Christensen, E. Tajkhorshid, Accelerating Membrane Insertion of Peripheral Proteins with a Novel Membrane Mimetic Model, Publisher, City, 2012.

[137] G. Brannigan, P.F. Philips, F.L.H. Brown, Flexible lipid bilayers in implicit solvent, Publisher, City, 2005.

[138] R.A. Böckmann, B.L. de Groot, S. Kakorin, E. Neumann, H. Grubmüller, Kinetics, Statistics, and Energetics of Lipid Membrane Electroporation Studied by Molecular Dynamics Simulations, Publisher, City, 2008.

[139] A.A. Gurtovenko, J. Anwar, I. Vattulainen, Defect-Mediated Trafficking across Cell Membranes: Insights from in Silico Modeling, Publisher, City, 2010.

[140] H. Leontiadou, A.E. Mark, S.J. Marrink, Molecular Dynamics Simulations of Hydrophilic Pores in Lipid Bilayers, Publisher, City, 2004.

[141] J.D. Litster, Stability of lipid bilayers and red blood cell membranes, Publisher, City, 1975.

[142] J.N. Sachs, P.S. Crozier, T.B. Woolf, Atomistic simulations of biologically realistic transmembrane potential gradients, Publisher, City, 2004.

[143] D. Tieleman, The molecular basis of electroporation, Publisher, City, 2004.

[144] D.P. Tieleman, H. Leontiadou, A.E. Mark, S.-J. Marrink, Simulation of Pore Formation in Lipid Bilayers by Mechanical Stress and Electric Fields, Publisher, City, 2003.

[145] J.C. Weaver, Y.A. Chizmadzhev, Theory of electroporation: A review, Publisher, City, 1996.

[146] L. Delemotte, M. Tarek, Molecular Dynamics Simulations of Lipid Membrane Electroporation, Publisher, City, 2012.

[147] Z. Levine, P.T. Vernier, Life Cycle of an Electropore: Field-Dependent and Field-Independent Steps in Pore Creation and Annihilation, Publisher, City, 2010.

[148] K.A. Dill, J.L. MacCallum, The Protein-Folding Problem, 50 Years On, Publisher, City, 2012.

[149] T.V. Tolpekina, W.K. den Otter, W.J. Briels, Nucleation free energy of pore formation in an amphiphilic bilayer studied by molecular dynamics simulations, Publisher, City, 2004.

[150] J. Wohlert, W.K. den Otter, O. Edholm, W.J. Briels, Free energy of a trans-membrane pore calculated from atomistic molecular dynamics simulations, Publisher, City, 2006.

[151] E. Guàrdia, R. Rey, J.A. Padró, Potential of mean force by constrained molecular dynamics: A sodium chloride ion-pair in water, Publisher, City, 1991.

[152] W.F.D. Bennett, N. Sapay, D.P. Tieleman, Atomistic Simulations of Pore Formation and Closure in Lipid Bilayers, Publisher, City, 2014.

[153] D. Bhattacharya, J. Cheng, 3Drefine: Consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization, Publisher, City, 2013.

[154] T. Nugent, D. Cozzetto, D.T. Jones, Evaluation of predictions in the CASP10 model refinement category: Assessment of Model Refinement Predictions, Publisher, City, 2014.

[155] N. Kučerka, S. Tristram-Nagle, J.F. Nagle, Closer Look at Structure of Fully Hydrated Fluid Phase DPPC Bilayers, Publisher, City, 2006.

[156] J.F. Nagle, S. Tristram-Nagle, Structure of lipid bilayers, Publisher, City, 2000.

[157] J.B. Klauda, R.M. Venable, J.A. Freites, J.W. O'Connor, D.J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A.D. MacKerell, R.W. Pastor, Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types, Publisher, City, 2010.

[158] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems, Publisher, City, 1993.

[159] J.-P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, Publisher, City, 1977.

[160] A.B. Larsen, J.R. Wagner, A. Jain, N. Vaidehi, Protein Structure Refinement of CASP Target Proteins Using GNEIMO Torsional Dynamics Method, Publisher, City, 2014.

[161] R.S. Ostrom, P.A. Insel, The evolving role of lipid rafts and caveolae in G protein-coupled receptor signaling: implications for molecular pharmacology, Publisher, City, 2004.

[162] W.F.D. Bennett, D.P. Tieleman, Computer simulations of lipid membrane domains, Publisher, City, 2013.

[163] O.S. Andersen, R.E. Koeppe, Bilayer Thickness and Membrane Protein Function: An Energetic Perspective, Publisher, City, 2007.

[164] S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, A.H. de Vries, The MARTINI force field: Coarse grained model for biomolecular simulations, Publisher, City, 2007.

[165] P. Kar, S.M. Gopal, Y.-M. Cheng, A. Predeus, M. Feig, PRIMO: A Transferable Coarse-Grained Force Field for Proteins, Publisher, City, 2013.

[166] W. Im, M. Feig, C.L. Brooks, An Implicit Membrane Generalized Born Theory for the Study of Structure, Stability, and Interactions of Membrane Proteins, Publisher, City, 2003.

[167] T. Lazaridis, Effective energy function for proteins in lipid membranes, Publisher, City, 2003.

[168] S. Tanizaki, M. Feig, A generalized Born formalism for heterogeneous dielectric environments: Application to the implicit modeling of biological membranes, Publisher, City, 2005.

[169] M. Sayadi, S. Tanizaki, M. Feig, Effect of Membrane Thickness on Conformational Sampling of Phospholamban from Computer Simulations, Publisher, City, 2010.

[170] A. Panahi, M. Feig, Conformational Sampling of Influenza Fusion Peptide in Membrane Bilayers as a Function of Termini and Protonation States, Publisher, City, 2009.

[171] N.R. Latorraca, K.M. Callenberg, J.P. Boyle, M. Grabe, Continuum Approaches to Understanding Ion and Peptide Interactions with the Membrane, Publisher, City, 2014.

[172] G. Brannigan, F.L.H. Brown, Contributions of Gaussian Curvature and Nonconstant Lipid Volume to Protein Deformation of Lipid Bilayers, Publisher, City, 2007.

[173] S. Esteban-Martín, J. Salgado, The Dynamic Orientation of Membrane-Bound Peptides: Bridging Simulations and Experiments, Publisher, City, 2007.

[174] T. Kim, W. Im, Revisiting Hydrophobic Mismatch with Free Energy Simulation Studies of Transmembrane Helix Tilt and Rotation, Publisher, City, 2010.

[175] J. Lee, W. Im, Transmembrane Helix Tilting: Insights from Calculating the Potential of Mean Force, Publisher, City, 2008.

[176] S.H. Park, S.J. Opella, Tilt Angle of a Trans-membrane Helix is Determined by Hydrophobic Mismatch, Publisher, City, 2005.

[177] N. Castillo, L. Monticelli, J. Barnoud, D.P. Tieleman, Free energy of WALP23 dimer association in DMPC, DPPC, and DOPC bilayers, Publisher, City, 2013.

[178] P. Lagüe, M.J. Zuckermann, B. Roux, Lipid-Mediated Interactions between Intrinsic Membrane Proteins: Dependence on Protein Size and Lipid Composition, Publisher, City, 2001.

[179] F.J.M. de Meyer, M. Venturoli, B. Smit, Molecular simulations of lipid-mediated protein-protein interactions, Publisher, City, 2008.

[180] A. Benjamini, B. Smit, Robust Driving Forces for Transmembrane Helix Packing, Publisher, City, 2012.

[181] W. Im, C.L. Brooks, Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations, Publisher, City, 2005.

[182] S. Choe, K.A. Hecht, M. Grabe, A Continuum Method for Determining Membrane Protein Insertion Energies and the Problem of Charged Residues, Publisher, City, 2008.

[183] H. Aranda-Espinoza, A. Berman, N. Dan, P. Pincus, S. Safran, Interaction between inclusions embedded in membranes, Publisher, City, 1996.

[184] S. Mondal, G. Khelashvili, H. Weinstein, Not Just an Oil Slick: How the Energetics of Protein-Membrane Interactions Impacts the Function and Organization of Transmembrane Proteins, Publisher, City, 2014.

[185] D.P. Tieleman, J.L. MacCallum, W.L. Ash, C. Kandt, Z. Xu, L. Monticelli, Membrane protein simulations with a united-atom lipid and all-atom protein model: lipid–protein interactions, side chain transfer free energies and model proteins, Publisher, City, 2006.

[186] J.L. Knight, C.L. Brooks, Surveying implicit solvent models for estimating small molecule absolute hydration free energies, Publisher, City, 2011.

[187] T. Lazaridis, M. Karplus, Effective energy function for proteins in solution, Publisher, City, 1999.

[188] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, K. Schulten, Scalable molecular dynamics with NAMD, Publisher, City, 2005.

[189] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, Publisher, City, 1983.

[190] V. Mirjalili, M. Feig, Density-Biased Sampling: A Robust Computational Method for Studying Pore Formation in Membranes, Publisher, City.