THE UNIVERSITY OF CHICAGO

LEVERAGING FREE ENERGY IN MOLECULAR DYNAMICS: APPLICATIONS AND NEW APPROACHES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE PRITZKER SCHOOL FOR MOLECULAR ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
SELAMI EMRE SEVGEN

CHICAGO, ILLINOIS
AUGUST 2019

For my mother

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

Thanks to Ana Beiriger, my partner in life. Without you this would have been a much darker journey.

Thanks to Esin Pere Sirel and Haluk Kilimci, my grandparents who supported me through my undergraduate studies. Your kindness paved the way from Istanbul to Chicago.

Thanks to Gene, Antonia and Alexandra Beiriger. I could not have asked for a more supportive family. You celebrated my every victory and supported me through every failure.

Thanks to my ever supportive Allenites, in no particular order: Joe Frumkin, Amanda Hayward, Alex Aavang, Allison Story Gulick, Eric DeBellis, Tara McGovern, Joe Boersma, Serena Brodsky, Abigail McEwen, Bobby Till, Erica Yuen, Alex Kulyk, Alon Stotter, Lindsay Sonenthal, Debbie Shi, Jason Blumstein, Kathleen VonDerHaar, Ben VonDerHaar, Tyler Hedlund, Jocelyn Sullivan-Hedlund, Dana Sievert, Daniel Hinze, Scott Durand, and Brock Russell. You all contributed to this more than you realize.

Thanks to my friends and colleagues in the department who supported me, in no particular order: Moshe Dolejsi, Alec Bowen, Zack Jarin, Arin Greenwood, Ryan McAvoy, Elyse Watkins, Yu Kambe, Mike Webb, Nick Jackson, Josh Moller, Federico Giberti, Cody Bezik, Gabriela Basel, Viviana Palacio Betancur, and a special thanks to Hythem Sidky for being an inspiring collaborator and a font of wisdom. My learning would have been incomplete without you to draw from.

Thanks to Ashley Guo for her unwavering support through the entire process.

Thanks to my advisors, Prof. Juan de Pablo and Prof. Jeffrey Hubbell for your continuous guidance and support, and to Prof. Giulia Galli for your advice and support through our collaboration. Your expertise and optimism were essential in exploring disparate avenues in research.

Finally, thanks to my mother, Neslihan Dincel. It would be foolish to attempt to list all of your contributions. I dedicate this work to you.

ABSTRACT

Free energy is the driving force behind countless processes ranging from the biological to the industrial. Large differences in free energy drive processes forward, while large barriers impede transitions. Accurate determination of these differences and barriers allow researchers to calculate key properties. We begin with such an application. Using free energy methods in molecular dynamics, we characterize a block copolymer that forms micelles via crystallization-driven self-assembly. Through a range of free energy calculations where we determine the relative stability of micelles as a function of size, we calculate the equilibrium size, shape, and stability of these micelles. We then turn our attention to the methodology that powers this kind of analysis: free energy calculations in molecular dynamics. Given that free energy is often the quantity of interest in a system studied via molecular dynamics, the length of time these methods take to estimate the free energy strongly influences the computational cost of the studies. We present two methods that leverage self-regularizing neural networks to very rapidly estimate underlying free energy during a molecular dynamics simulation. The first method builds upon an already successful method, Adaptive Biasing Force (ABF), by better handling the error inherent in its estimates and by providing exploratory bias in unvisited regions. The second method further builds on the first by incorporating the frequency of visits in phase space, in addition to the forces, to the final estimate of the free energy for an even faster, more robust estimate. Finally, we seek to expand the reach of these methods by introducing an easy-to-use, powerful and scalable framework for applying these methods to first principles molecular dynamics, and a hierarchical transfer methodology to rapidly converge such calculations.

CHAPTER 1

INTRODUCTION

In Chapter 2, a mechanism for micelle formation is considered in which nanocrystalline domains are formed at the core of the micelles from short ethylene sulfide oligomers, leading to exceptionally stable, uniform micellar structures. The structure and thermodynamic properties of the resulting micelles are examined through a combination of experiments, theory, and simulations. We find that in oligo(ethylene sulfide)-*b*-poly(ethylene glycol) amphiphiles, as few as three ethylene sulfide monomers are sufficient to form a highly crystalline core, surrounded by a water-soluble ethylene glycol corona of arbitrary size. Sulfur-sulfur interactions induce formation of rhombohedral lattice crystalline regions, which exhibit well-defined intramolecular and intermolecular order. An atomistic model is used to determine the free energy of the micelles, and the critical micelle concentration (CMC) is found to be extremely small, on the order of 10^{-8} mol/L. The size distribution of these micelles is nearly monodisperse. The crystalline core also includes amorphous regions that could serve as hosts for other molecules. Taken together, these properties serve to underscore that controlled crystallization provides a useful and under-exploited mechanism for assembly of ultra-stable micelles in applications including drug delivery and immunology.

In chapter 3, a machine learning-assisted method is presented for molecular simulation of systems with rugged free energy landscapes. The method is general and can be combined with other advanced sampling techniques. In the specific implementation proposed here, it is illustrated in the context of an Adaptive Biasing Force (ABF) approach where, rather than relying on discrete force estimates, one can use a self-regularizing artificial neural network to generate continuous, estimated generalized forces. By doing so, the proposed approach addresses several shortcomings common to adaptive biasing force and other algorithms. Specifically, the neural network enables (1) smooth estimates of generalized forces in sparsely-sampled regions, (2) force estimates in previously unexplored regions, and (3)

continuous force estimates with which to bias the simulation, as opposed to biases generated at specific points of a discrete grid. The usefulness of the method is illustrated with three different examples, chosen to highlight the wide range of applicability of the underlying concepts. In all three cases, the new method is found to enhance considerably the underlying traditional Adaptive Biasing Force approach. The method is also found to provide improvements over previous implementations of neural network-assisted algorithms.

In chapter 4, another machine learning-based adaptive free energy method is described for use in systems with complex, potentially high-dimensional free energy landscapes. This method, Force-Biasing Using Neural Networks (FUNN), builds upon the one described in the previous chapter. The method's main strength lies in being able to learn both from state visit frequencies and from the generalized force estimates of the system. To this end, a self-integrating artificial neural network is described, which generates an estimate of the free energy directly from its derivatives. This combined approach is found to be robust, faster, and more accurate than relying only on frequency-based or generalized force-based estimation. Combined with overfill protection and support for sparse data storage and training, the method can safely ignore high energy regions and effectively scale to a large number of collective variables.

In chapter 5, we present a seamless coupling of a suite of codes designed to perform advanced sampling simulations, with a first principles molecular dynamics (MD) engine. As an illustrative example, we discuss results for the free energy and potential surfaces of the alanine dipeptide obtained using both local and hybrid density functionals (DFT), and we compare them with those of a widely used classical force field, Amber99sb. In our calculations, the efficiency of first principles MD using hybrid functionals is augmented by hierarchical sampling, where hybrid free energy calculations are initiated using estimates obtained with local functionals. We find that the free energy surfaces obtained from classical and first principles calculations differ. Compared to DFT results, the classical force

field overestimates the internal energy contribution of high free energy states, and it underestimates the entropic contribution along the entire free energy profile. Using the string method, we illustrate how these differences lead to different transition pathways connecting the metastable minima of the alanine dipeptide. In larger peptides, those differences would lead to qualitatively different results for the equilibrium structure and conformation of the molecules.

CHAPTER 2

NANOCRYSTALLINE OLIGO(ETHYLENE SULFIDE)-*B*-POLY(ETHYLENE GLYCOL) MICELLES: STRUCTURE AND STABILITY

2.1 Introduction

Amphiphiles consist of at least two differently interacting domains, for example a hydrophobic and a hydrophilic region. Sodium dodecyl sulfate (SDS), a widely studied anionic surfactant, is representative of such a system. Amphiphilic molecules can self-assemble in solution, thereby providing a basis for their use as detergents[9, 17], drug delivery vehicles[95, 80], immunomodulators[114], and molecular reporters[79]. An extensive body of theoretical, experimental and, more recently, simulation work, has sought to arrive at a better understanding of micelle formation.

Block copolymers have also received considerable attention in the context of micelle formation. A majority of past work, however, has focused on amphiphilic systems consisting of an amorphous core, stabilized by hydrophobic or electrostatic interactions. These "polymerosome" systems are of interest owing to their ability to self-assemble into multi-domain structures, or to exhibit additional functionality through judicious modification of the underlying polymer chemistry [82, 66, 91]. Unfortunately, amorphous-core micelles are generally susceptible to changes in concentration, leading to wide size distributions that render them unsuitable for numerous applications. It is generally appreciated that block copolymers can adopt ordered morphologies through a variety of mechanisms, some of which involve crystallization. In that case, the micelles exhibit semi-crystalline cores, which impart remarkable consistency in size and superior tunability of critical properties such as dissociation rate, thermodynamic stability, and critical micelle concentration (CMC)[38]. It has also been shown that, in the context of polymers, crystallization driven self-assembly can lead to formation

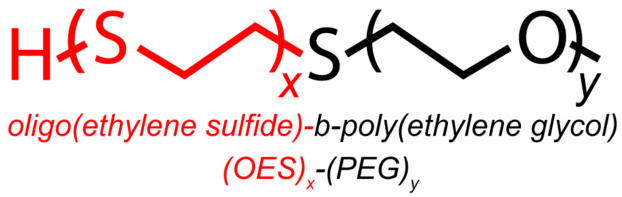


Figure 2.1: The two blocks that make up the block copolymer of interest

of monodisperse and hierarchical structures [33, 37].

In this work, a combined experimental and computational strategy is adopted to describe and understand the origin of self-assembly into semi-crystalline micelles. Specifically, we consider diblock copolymers of oligo(ethylene sulfide)-b-poly(ethylene glycol), which have been recently shown to form micelles, fibrils and hydrogels that exhibit crystalline signatures[10]. The small ethylene sulfide block has been hypothesized to induce crystal formation, yet remain water soluble at low repeat numbers (n=3-9). A longer PEG block (n=44-112) can be manipulated to tune aggregation number, dissociation rate, size and shape (Fig. 2.1). As such, these materials provide an easy-to-synthesize, well-controlled system for precision assembly of water stable aggregates. The resulting structures are capable of hosting drugs in their core, thereby lending themselves for applications in controlled drug delivery and immunology.

The two overarching aims of this investigation are, first, to provide unambiguous evidence of the crystallinity of OES-*b*-PEG micelles, and its role in aggregation. Note that in prior work crystallinity was only hypothesized to be responsible for assembly[10]. Second, we seek to understand and characterize the aggregation behavior, in order to determine the characteristic sizes that are accessible in the context of ultra-small, ultra-stable micelles. Such information is critical for potential applications in immunology, where penetration into specialized lymph node channels having a cutoff of approximately 8nm is necessary[94]. Third, we calculate the CMC via statistical mechanical methods, as the CMC of previously synthesized OES-*b*-PEG micelles could not be determined by standard experimental techniques.

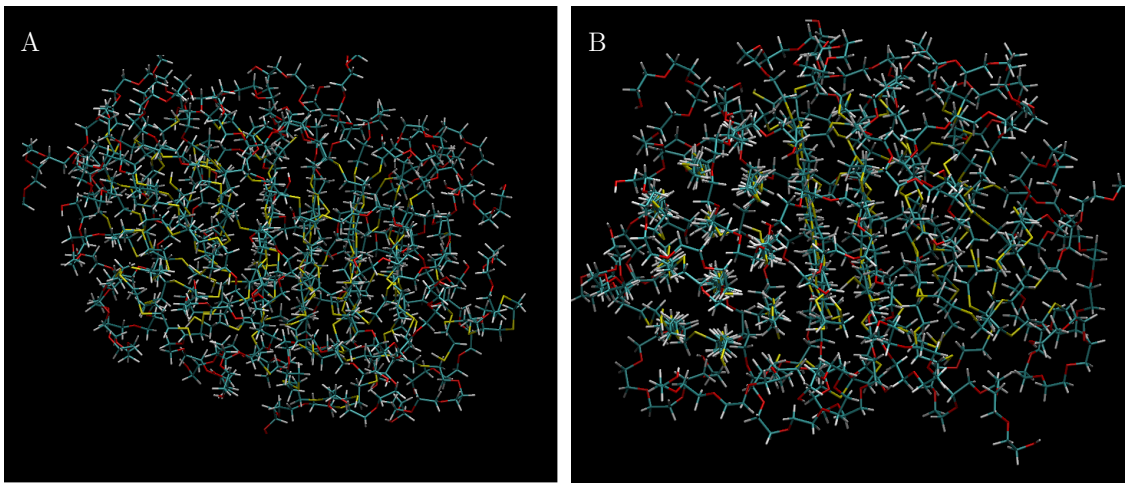
In what follows, we begin by showing that the molecular model of OES-*b*-PEG introduced here is able to capture spontaneous self-assembly in water. We then characterize the crystallinity of aggregate cores, and examine how it relates to the crystalline structure of pure ethylene sulfide in the solid phase. We provide experimental evidence from grazing incidence wide-angle X-ray scattering (GIWAXS) and from differential scanning calorimetry that is consistent with theoretical predictions. By conducting simulations of the potential of mean force required to extract individual molecules from a micelles, we examine the effects of both PEG and OES block size on micelle stability. We also characterize the aggregation behavior, and provide an estimate for the CMC as well as the average aggregate size and its distribution as a function of concentration. We conclude with an analysis of the crystallinity of the corresponding micelle cores, and how it affects micelle characteristics.

2.2 Results and Discussion

Aggregation of Copolymer

After the derivation of dihedral parameters and the validation of the forcefield (see Methods section), several different block-copolymers were simulated in water starting from initial random configurations, thereby allowing us to follow the process of spontaneous self-assembly (Fig. 2.2). All polymers considered here were found to self-assemble for all the conditions examined (Fig. 2.2C); the OES blocks spontaneously formed semi-crystalline cores. Semi-crystalline cores are consistent with previous observations, where some degree of crystallinity was hypothesized to exist within the cores, yet the micelles were experimentally shown to be capable of carrying small-molecule drugs - a feature that is consistent with the presence of amorphous domains [10].

Brubaker et al.[10] inferred the formation of crystals in the micelle cores on the basis of differential scanning calorimetry (DSC) measurements for pure SH-OES_3-OH. To further establish the link between pure SH-OES_3-OH and OES-*b*-PEG crystallinity, we carried out



C

	PEG$_3$	**PEG$_5$**	**PEG$_{10}$**	**PEG$_{44}$**
OES$_3$	Self-Assembly	Self-Assembly	Self-Assembly	Self-Assembly
OES$_5$	Not Tested	Self-Assembly	Self-Assembly	Not Tested

Figure 2.2: Spontaneous self-assembly of OES-*b*-PEG in water. PEG repeats are shown in red, OES repeats are in yellow. Water molecules are hidden for clarity. **A)** OES$_3$-*b*-PEG$_3$ after 40ns, starting from a random configuration. **B)** OES$_5$-*b*-PEG$_5$ after 40ns, starting from a random configuration. C) Summary of conditions tested for spontaneous self-assembly.

all-atom simulations of both SH-OES$_3$-OH and SH-OES$_5$-OH. Both systems crystallized, forming domains that closely resemble the structures that form in the micelle cores (Fig. S2). We then performed a series of melting point simulations of SH-OES$_3$-OH to compare to experimental DSC data. Simulations of crystalline SH-OES$_3$-OH were held at different temperatures in 10K increments. Experimentally, SH-OES$_3$-OH exhibits a first order transition in the range between 100-138C from a crystalline solid to a liquid. In our simulations, we observe melting at a slightly higher temperature, 147C, but not at 137C (Fig. 2.3). We attribute this difference to the different cooling and heating rates that are accessible in experiments and simulations.

Crystalline Structure via GIWAXS

OES$_3$-*b*-PEG$_{44}$ was synthesized and spuncoat to verify the predicted crystalline structure. GIWAXS is used as a powerful tool for experimental verification of the observed structures

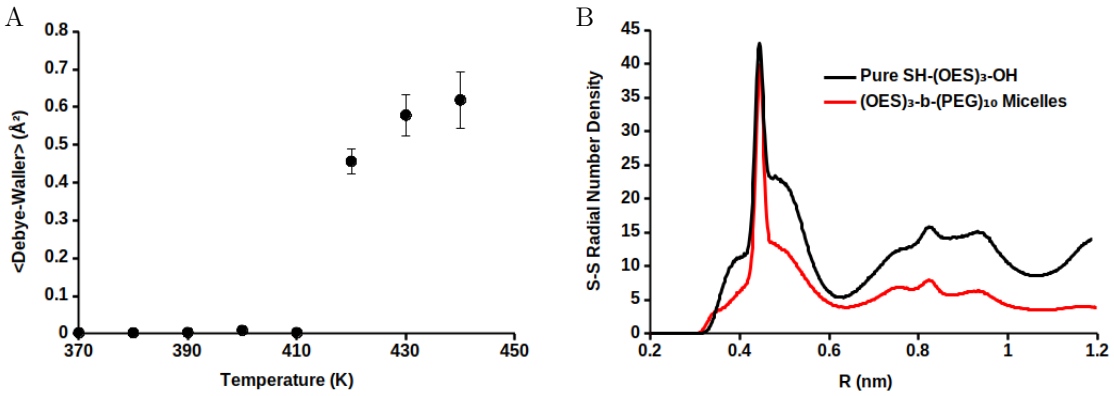


Figure 2.3: **A)** Average Debye-Waller Factor of SH-OES_3-OH calculated at different temperatures from separate simulations. A melting transition is observed between 410-420K (137-147C), which is consistent with previous experimental findings. **B)** Comparison of S-S radial number density distributions of pure, crystalline SH-OES_3-OH and OES_3-*b*-PEG_10 micelles. Both show the same hexagonal ordering (Fig. 2.4, S2). The decay in micelle distribution is due to the finite size of the micelle core.

in simulation. The dominant coherent scattering came from sulfur sulfur events due to sulfurs cross section, which is more than double that of either oxygen or carbon. The sulfur - sulfur separation along the backbone of the polymer results in the first primary peak at $Q = 1.34\text{Å}^{-1}$ (Qc). Coinciding with this peak is also the PEG oxygen oxygen peak, which appears at a similar spacing [77]. The PEG alone could not be responsible for such a large peak, as is verified by the low intensity of the other PEG peaks which appear as a doublet at $Q = 1.82\text{Å}^{-1}$, 1.88Å^{-1}. The sulfur sulfur separation between adjacent crystallized chains appears as the second primary peak $Q = 1.63\text{Å}^{-1}$ (Qcc). The additional peaks present at $Q = 2.6\text{Å}^{-1}$, 3.0Å^{-1} correspond to the underlying gold substrate. Both PEG peaks and sulfur-sulfur peaks are fully consistent with the expected ordering and distances from simulations of pure OES_3, as well as the rhombohedral structure of sulfur atoms observed in the micelle cores (Fig. 2.4). The Qc peak is sharper than that of Qcc, further validating the simulated structures, which exhibit a looser hexagonal packing when viewed from above. Simulations of micelles predict separations of $3.82 \pm 0.042\text{Å}$ and $4.47 \pm 0.016\text{Å}$ for Qcc and Qc, respectively. Overall, structures observed in simulations are consistent with GIWAXS

results, serving to validate the model and calculations of micelle cores presented in this work.

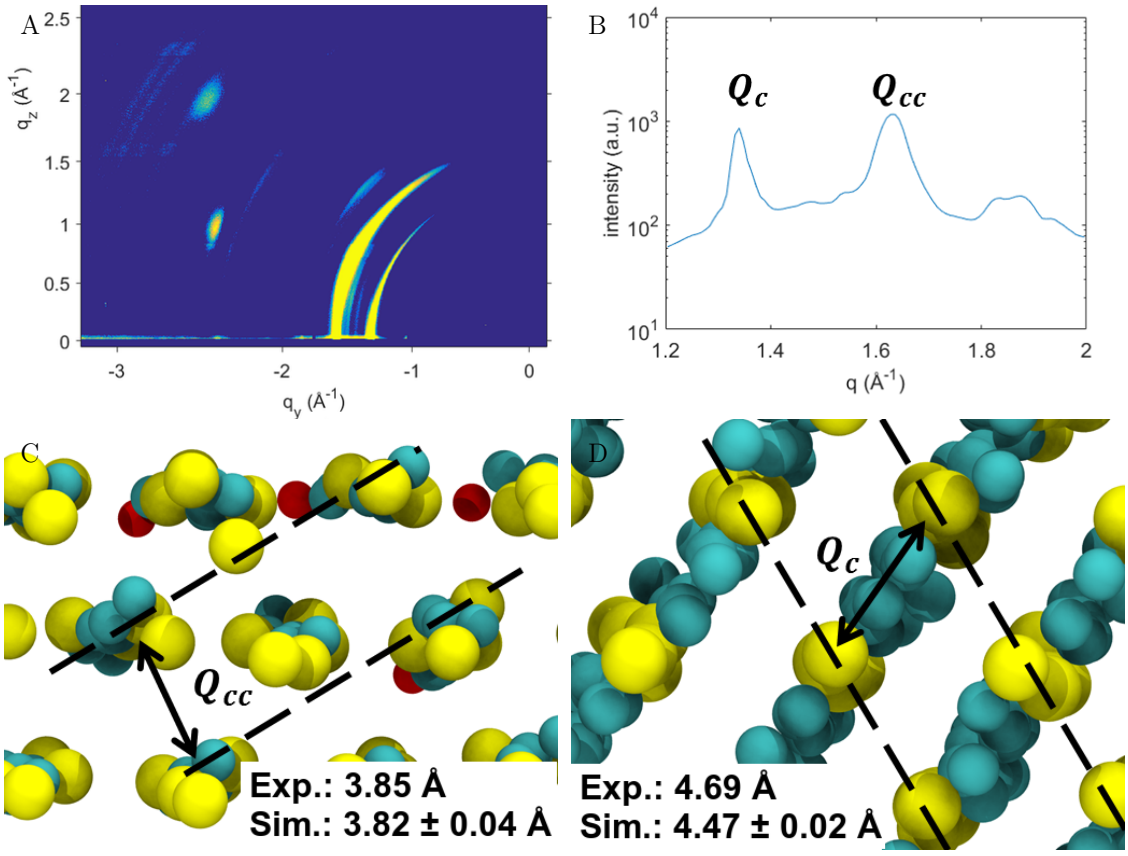


Figure 2.4: **A)** GIWAXS of OES_3-*b*-PEG_{44} thin films. **B)** Integrated intensity and assigned peaks. **C, D)** Top down (each node is a molecule, view is down the backbone) and side projections (each angled rod is a polymer backbone from the side) of the crystalline OES_3 from a simulation configuration. GIWAXS results are consistent in spacing and arrangement with simulation predictions.

Potential of Mean Force (PMF)

PMF simulations probe how a system's energy changes as it progresses on a defined 'reaction coordinate'. First, to quantify the strength of the sulfur-sulfur interactions that lead to micelle formation, we determined the free energy of interaction (or potential of mean force, PMF) between two OES_3 blocks in water, with and without a PEG_{10} tail. The interactions are pronounced, on the order of 6 kT even in the absence of an ordered structure, and the PEG tail does not significantly influence their magnitude (Fig. 2.5).

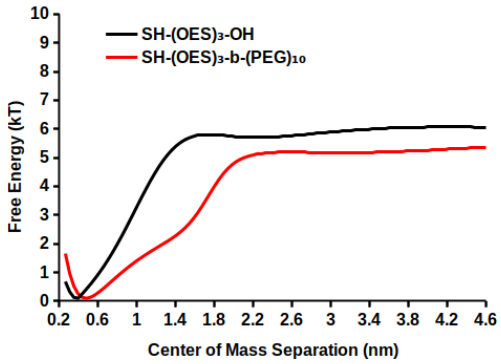


Figure 2.5: Potential of Mean Force of two OES_3 blocks in water, with and without a PEG_{10} tail. The PEG_{10} tail does not significantly influence the PMF, which is mainly driven by the sulfur-sulfur interactions.

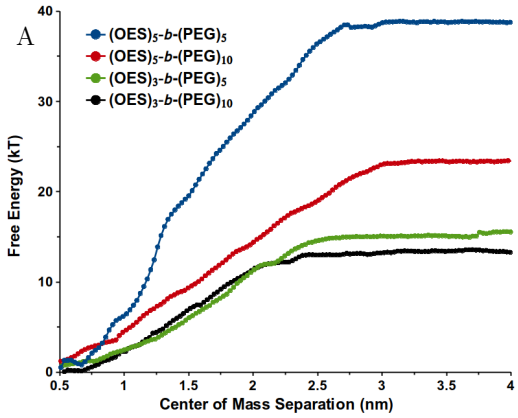
In order to estimate how OES and PEG block sizes influenced the stability of the micelle, we performed umbrella sampling on micelles formed from four different block copolymer configurations. OES block size significantly affects the association strength of individual molecules to the micelle. In contrast, the PEG block size does not significantly influence the association strength (Fig. 2.6). The increased dependence of micelle crystallinity on OES block size is due to the dipolar interactions between OES blocks, which are stronger than those between PEG blocks. Compared to PEG, OES exhibits a much higher crystallinity and melting point (210 °C vs 66 °C) [105, 97].

One can obtain an estimate for the CMC, given the association strength, via:

$$\Delta G_{micelle} = \frac{-RT}{N} ln[micelle] + RTln(CMC) \tag{2.1}$$

This equation provides a useful way to gauge how CMC depends on association strength and therefore on block copolymer composition. However, a more in-depth analysis is required for an accurate estimation of the CMC, necessitating the calculation of the size distribution of micelles for a given block-copolymer configuration. With the goal of designing ultra-small (<8nm diameter) and ultra-stable micelles, we turned to $(OES)_3$-*b*-$(PEG)_{10}$ as a potential

A



B

(kT)	$\mathbf{OES_3}$	$\mathbf{OES_5}$
$\mathbf{PEG_5}$	13.37 ± 0.38	30.8 ± 6.3
$\mathbf{PEG_{10}}$	16.98 ± 2.38	33.83 ± 10.08

Figure 2.6: **A)** A representative graph showing one umbrella sampling run from each different configuration for Aggregation Number N=24 micelles. **B)** Free Energy of removing a single molecule from an N=24 micelle for different block copolymer configurations. Each run was repeated three times with different micelles.

candidate, based on preliminary experimental and simulation data. Given the remarkably low expected CMC, simulating a true equilibrium system would require an unfeasibly large system. Thus, we adopt the alternate approach described in the Methods section. In order to estimate how micelle free energy of formation changes with aggregation number (N), a series of umbrella sampling runs were performed at varying N. We obtain 1.4E-8 mol/L as the CMC, consistent with previous observations that OES-*b*-PEG micelles have CMCs below the experimental limit of detection (Fig. 2.7).

The final interpolation scheme used to estimate the free energy barrier to removing a molecule at a given aggregation number N was simply linear, as the results were insensitive to the choice of interpolation scheme. The results display classical CMC-like behavior. Below 1.4E-8 mol/L, all dissolved polymer exist as individual molecules. In contrast, starting at 1.4E-8 mol/L the molecule concentration is no longer one-to-one with total polymer concentration, as most polymers start forming micelles. Mole fractions as a function of aggregation number clearly show this behavior: for concentrations below the CMC, all polymers exist

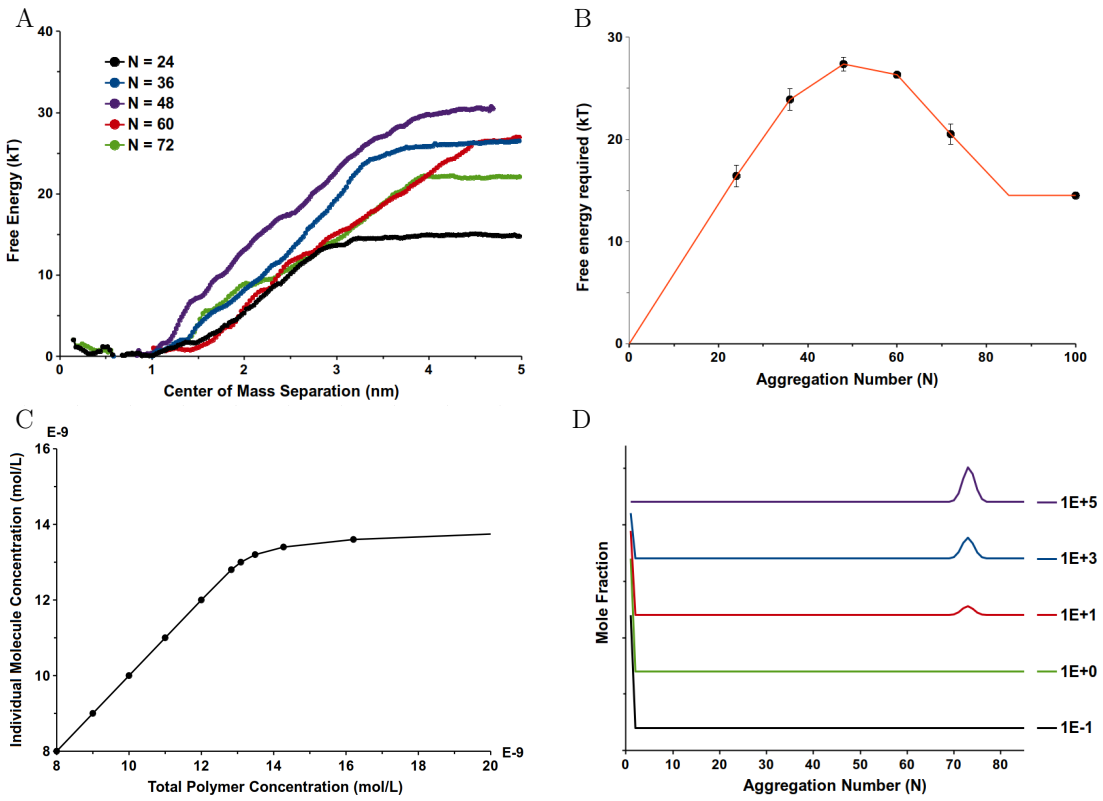


Figure 2.7: Calculation of the free energy of formation of OES_3-*b*-PEG_{10} micelles for different aggregation numbers N, from a micelle of N-1 and an individual molecule, and emergent CMC-like behavior. **A)** Representative umbrella sampling runs a molecule is extracted from a OES_3-*b*-PEG_{10} micelle at various aggregation numbers. Each run was repeated four times (not shown here). **B)** Free energy for extraction of a molecule from a OES_3-*b*-PEG_{10} micelle at various aggregation numbers. **C)** Free molecules in solution as a function of total concentration of polymer. The CMC is approximately 1.4E-8 mol/L. **D)** Mole fractions of different aggregate sizes, given at various total concentrations as multiples of the CMC.

as isolated entities. Above the CMC, a peak at N=74 starts forming. It is remarkable how insensitive the average aggregation size is to concentration the peak center at N=74 barely shifts over five orders of magnitude. As expected from Cryo-EM images of other OES-*b*-PEG micelles, the size distribution is extremely narrow [10]. These favorable characteristics are absent from a chemically similar, but amorphous analogue of OES-*b*-PEG, poly(propylene sulfide)-*b*-poly(ethylene glycol) (PPS-*b*-PEG). PPS-*b*-PEG exhibits a wealth of morphologies, but cannot form micelles that are as stable or as small as those formed by OES-*b*-PEG [112]. Note, however, that despite these favorable characteristics that emerge from crys-

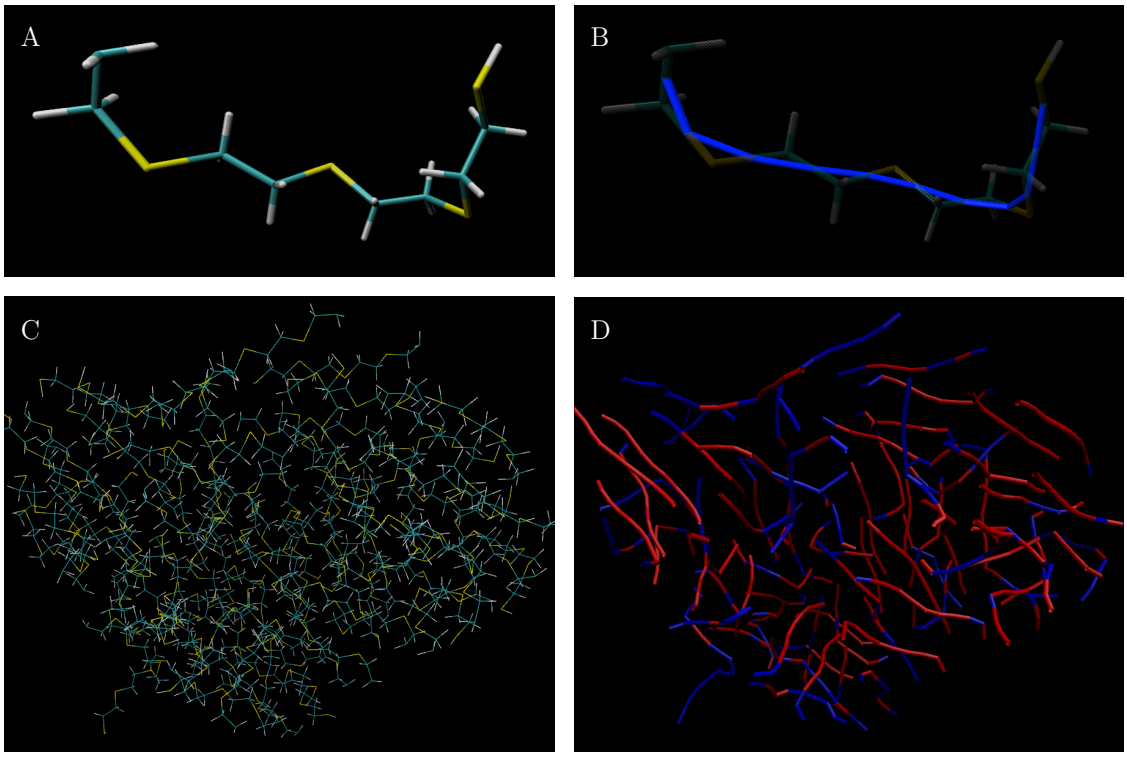


Figure 2.8: Illustration of the method used to map the polymer backbone to rods for analysis of micelle cores. **A)** A unimer with only $(OES)_3$ visible. **B)** Backbone rods in blue, superimposed on the unimer. **C)** A micelle core from a simulation snapshot with PEGs and waters removed. **D)** Rods mapped onto the micelle core, color coded for amorphous (blue) and crystalline (red) regions.

tallinity, the cores of the micelles also exhibit significant amorphous domains. Existence of these amorphous regions endows these micelles with drug and dye loading capabilities.

Crystallinity and Asphericity

In order to investigate micelle core compositions, we utilized two algorithms. First, polymer backbones were mapped onto rods with well-defined lengths and orientations (Fig. 2.8). Then, a clustering algorithm that successfully detects different crystal grains in micelle cores and their orientations was developed and utilized (Fig. 2.9).

Twenty micelles were simulated at N=74 for analysis of crystallinity, and its effect on micelle shape and structure. First, crystalline and rodlike content, average ordering, and size of each micelle were analyzed (Table 2.3).

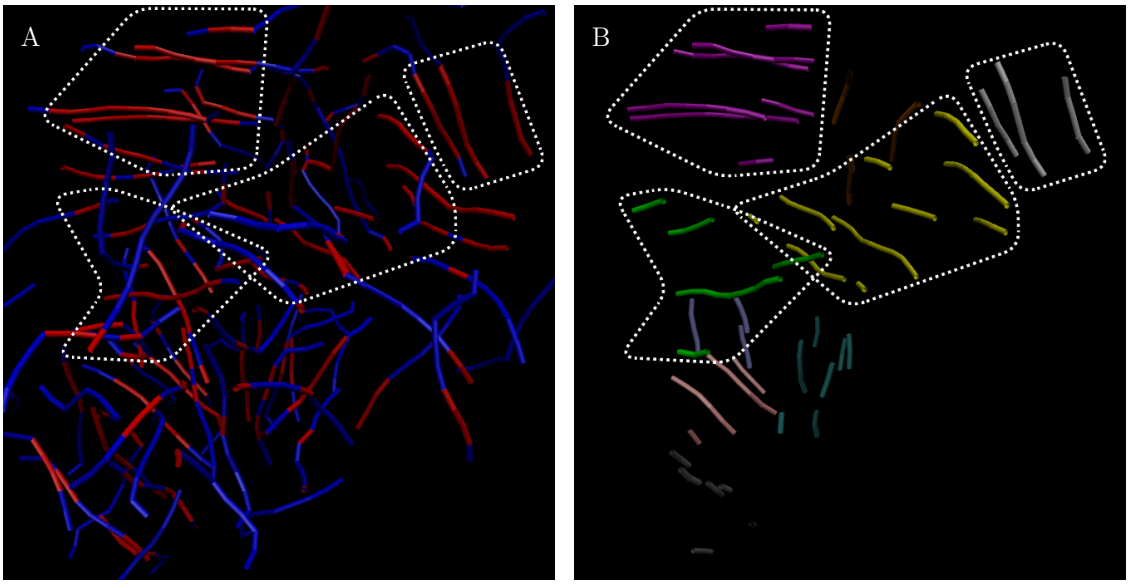


Figure 2.9: Illustration of results from the clustering algorithm. **A)** Core of a micelle after coarse-graining, with amorphous regions in blue and crystalline regions in red. **B)** Clustering algorithm results, with each crystal grain shown in a different color.

There is considerable variability in the crystalline content and relative order (S) from micelle to micelle. However, the rod-like content is remarkably uniform. In our observations of the system, most often it is only the free end and the OES-to-PEG linkage segments that break the rod-like character. Indeed, the average rod-like character is 83.9%, which is close to $8/10 = 80\%$ - the theoretical percentage if each segment was rod-like everywhere except at the end and the OES-*b*-PEG linkages. This indicates the strong preference of OES to align in a straight line, regardless of its contribution to a well-formed crystal grain.

Our analysis of asphericity in OES_3-*b*-PEG_{10} reveals highly spherical micelles within the desired size constraints (Table 2.3). Given OES-*b*-PEGs tendency to form fibrils at other configurations, we tried to establish a link between asphericity and various properties of the micelle. Furthermore, asphericity could greatly hinder these micelles' ability to enter the 8nm channels they are designed to navigate in biomedical applications. It is therefore critical to understand which properties affect asphericity in order to control it.

We hypothesized that shape, size and distribution of crystal grains within the core would

significantly influence the shape of the micelle overall. Surprisingly, despite a strong relation between micelle core shape and overall micelle shape, we found no correlation between crystalline percentage, crystal grain sizes, size distributions, or overall order of the core with micelle asphericity. Thus, we further analyzed each micelle by binning the angle of each crystalline rod with every other crystalline rod in a given micelle. This yields a distribution of angles that deviates from the expected entropic curve of $sin(a)/2$ based on the crystal grain orientation in the system. We fit the resulting histogram to a sine expansion of the form (Fig. 2.10):

$$k_1 sin(x) + k_3 sin(3x) + k_5 sin(5x) + k_7 sin(7x) \tag{2.2}$$

We found that the coefficients of this expansion, especially k_1 and k_3, hold information about relative alignment of structures, and thus the asphericity (Fig. 2.10). We further define the parallel factor as all the elements that contribute towards 0° and 180° configurations:

$$ParallelFactor = k_3 + k_5 + k_7 \tag{2.3}$$

Parallel factor is the best predictor of the asphericity of micelles and micelle cores in our system (Fig. 2.10).

Our current hypothesis is that configurations that favor adjacent, perpendicular crystal grains should favor more aspherical structures, and thus potentially fibrillar configurations for different ratios of OES to PEG. Given that OES_5 terminated polymers qualitatively favor such structures over OES_3 terminated ones and that, experimentally, increasing OES block size favors fibrils over spherical micelles [10], a potential mechanism is that increasing OES block size leads to larger, more adjacent and perpendicular crystal grains, which give rise to increasingly aspherical structures. A more thorough investigation of formation of fibrillar micelles, and the kinetic and thermodynamic properties of that process, and the effects of

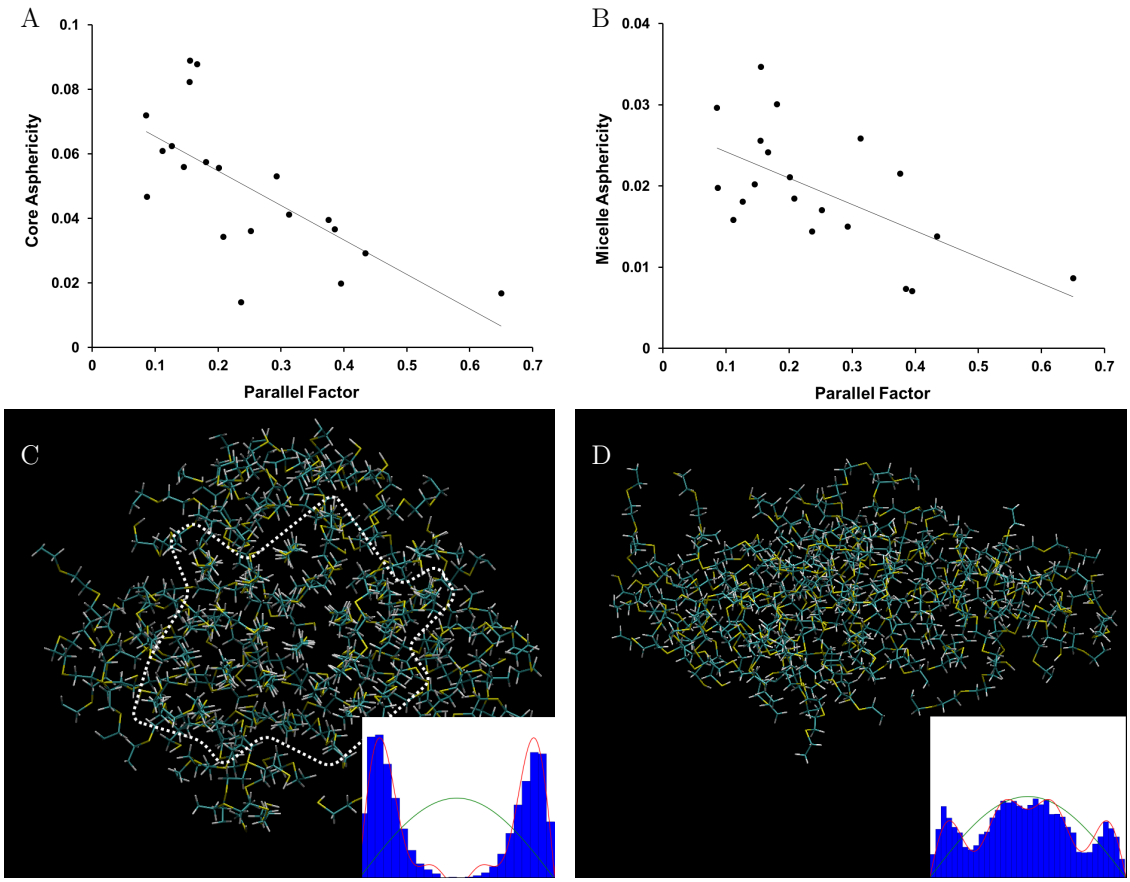


Figure 2.10: Analysis of 20 micelle simulations: Parallel factor correlation with core and micelle asphericity, and two representative examples. **A)** Parallel factor correlates with core asphericity. $R^2 = 0.47$ **B)** Parallel factor correlates with micelle asphericity. $R^2 = 0.38$ **C)** Simulation 3, which has a large cluster. Large clusters lead to spherical micelles, and have rod angle distributions with peaks corresponding to parallel alignment (inset). **D)** Simulation 7, which has several small clusters. High number of grains tend to align perpendicular to each other, and form pancake micelles. Rod angle distributions have peaks both in perpendicular and parallel arrangement (inset). The green curve represents the expected entropic distribution of $sin(a)/2$, the red curve is the fitted sine expansion.

block sizes on spherical versus fibrillar structures is currently underway.

2.3 Conclusions

We performed all-atom simulations of crystallization-driven self-assembly of OES-*b*-PEG copolymer micelles in water. These aggregates form ordered, semicrystalline cores with regions of rhombohedral character. This ordering mirrors that of pure SH-OES_3-OH and SH-OES_5-OH deduced from solid phase simulations, and is consistent with GIWAXS measurements of spuncoat films of OES_3-*b*-PEG_{44}. As analysis of the effects of OES and PEG block sizes on stability reveals an easily tunable behavior for both the CMC and mean escape times from these micelles. Further examination of OES_3-*b*-PEG_{10} micelles is indicative of remarkable stability, extremely small sizes, and a narrow size distribution. The calculated CMC of OES_3-*b*-PEG_{10} is just 1.4E-8 mol/L, and the average size is 6.776 $\pm$ 0.214 nm in diameter along the widest dimension. These characteristics should be contrasted with those of OES-*b*-PEGs amorphous analogues, such as PPS-*b*-PEG. In spite of their large degree of crystallinity, the remaining amorphous regions are sufficiently large to enable drug or dye loading for applications [10]. These ultra-small, ultra-stable micelles have the potential to enable novel applications in drug delivery and immunology due to their unprecedented size, coupled with their stability and ability to solubilize drugs and fluorescent reporters. Moving forward, there are multiple design avenues that one could pursue for alternative applications, including cross-linking [82] or forming multi-block, multi-domain configurations [66].

2.4 Methods

Derivation of OPLS-AA Parameters for the S-C-C-S and S-C-C-O Dihedral Potentials

Atomistic simulations have been used in the past to calculate the free energy of association and mean escape time from sodium dodecyl sulfate (SDS) micelles, finding good agreement with experiment [117]. We also adopt an all-atom representation of the OES-*b*-PEG molecules in this work. We begin by constructing a suitable force field for our simulations. Parameters for S-C-C-S and S-C-C-O dihedral potentials, in particular, are necessary in order to accurately model the behavior of OES-*b*-PEG. Ryckaert-Bellemans dihedral parameters for S-C-C-S and S-C-C-O were derived from density functional theory (DFT) calculations using Gaussian. [96] Energy as a function of dihedral angle was scanned using the b3lyp functional and 6-311+g(d) basis set in two-degree increments for CH_3-S-CH_2-CH_2-S-CH_3 and CH_3-S-CH_2-CH_2-O-CH_3. Then, the same scan was performed in GROMACS 4.6.7 [48] with available OPLS-AA parameters, with the dihedral potential disabled. Finally, the energy difference (Gaussian - GROMACS) was fitted to Ryckaert-Bellemans parameters using least squares.

Table 2.1: Newly Derived Ryckaert-Bellemans Dihedral Potentials for S-C-C-S and S-C-C-O

	C_0	C_1	C_2	C_3	C_4	C_5
S-C-C-S	4.066	-9.200	-10.568	8.783	6.188	-1.386
S-C-C-O	-2.915	-15.110	-2.437	4.574	1.877	3.814

The new force field parameters were used in conjunction with the OPLS force field, and were validated by reproducing the density and enthalpy of vaporization of B-Mercaptoethanol and 1,2-ethanedithiol at several conditions. As shown in Table 2.2, the resulting parameters provide a reasonable description of these two liquids.

Crystallization and Melting Point Simulations of OH-$(OES)_3$-SH, OH-$(OES)_5$-SH

Table 2.2: Comparison of Predictions from Simulations with Experimental Results

Parameter	Temp (K)	Pressure(atm)	Experiment	Simulation	%Error
1,2-ethanedithiol					
Density	298.15	1	1123 g/L	1152 ± 0.23 g/L	2.64
Enthalpy of Vap.	298.15	1	44.7 kJ/mol	45.94 ± 0.013 kJ/mol	2.77
	418.2	1	37.93 kJ/mol	38.97 ± 0.014 kJ/mol	2.75
β-Mercaptoethanol					
Density	298.15	1	1115 g/L	1147 ± 0.13 g/L	2.86
Enthalpy of Vap.	298.15	1		63.1134 kJ/mol	
	430	1		51.7159 kJ/mol	
	293 440	1	54.1 kJ/mol		

In order to assess the ability of the force field to reproduce crystallinity in the systems considered here, one hundred molecules of OH-$(OES)_3$-SH or OH-$(OES)_5$-SH were annealed from 600K to 300K at 10K/ns and then held at 300K for 400ns using GROMACS 4.6.7. In both systems, crystalline order was reached spontaneously. Several simulations were then started from the crystalline configuration, and annealed at 10K/ns to a range of final, physiologically relevant temperatures. Each simulation was held at its target temperature for 400ns. Simulations were performed in an NPT ensemble using a modified velocity rescaling thermostat that preserves the correct underlying statistical-mechanical distribution[12] and the Parrinello-Rahman barostat[85].

Micelle Formation in Water

For all configurations, copolymer molecules were randomly distributed in TIP4P water[1] and were simulated in an NPT ensemble.

PMF Simulations

For calculations of the PMF of OES_3, two molecules were solvated and equilibrated in TIP4P water [1]. A reaction coordinate for PMF simulations was defined as the difference between the center of mass of the OES_3 blocks of each molecule. The PMF was then calculated using the Artificial Neural Network sampling method[102] using GROMACS-

2016.4 and SSAGES[99].

For calculations of the PMF in micelles, micelles of a given aggregation number were pre-assembled, solvated, and equilibrated in TIP4P water [1]. A reaction coordinate for PMF simulations was defined as the difference between the center of mass of a randomly selected copolymer molecule and the center of mass of the remaining molecules in the micelle. The selected copolymer molecule was pulled away from the center of a position-restrained micelle by 5nm over 3.5ns to generate initial configurations. Umbrella sampling windows were formed from configurations closest to 0.05nm apart from each other for the first 2.5 nanometers, then 0.1nm apart for the rest of the reaction coordinate. Window size selection was determined by ensuring that sufficient overlap exists between adjacent histograms. A 5nm maximum was deemed sufficient as the PMF flattened out. Pulling experiments were carried out using umbrella sampling with built-in GROMACS tools [110]; a Weighted Histogram Analysis Method (WHAM) was used to combine umbrellas into a potential of mean force[60].

Calculation of Critical Micelle Concentration and Average Aggregation Number

The free energy of association/dissociation of a molecule with a micelle was calculated for different aggregation numbers, namely N = 24, 36, 48, 60, 72 via umbrella sampling according to:

$$micelle_{N+1} \rightleftharpoons micelle_N + unimer \tag{2.4}$$

The free energy was determined from:

$$\Delta G_N = -RTln(K_N) = -RTln\left(\frac{[micelle_N][unimer]}{[micelle_{N+1}]}\right) \tag{2.5}$$

We used linear interpolation to obtain values for all N in the range N=1-100. The $N \rightarrow large$

value was deduced from an infinite fibril simulation (See below). The results were not found to change appreciably for different types of interpolation schemes. We then solved (2.5) simultaneously for all N at several different concentrations of the polymers.

Infinite Fibril Simulation

To simulate a long fibril, 120 molecules of OES_3-*b*-PEG_{10} were first solvated and equilibrated from a semi-random configuration that was designed to favor an infinite, fibril-like shape, connected to itself through the periodic boundary conditions. The fibril was simulated for an additional 20ns using a semi-isotropic pressure coupling in the NPT ensemble, thereby allowing it to adjust its thickness. Finally, umbrella sampling was performed as described above to obtain the PMF of dissociation of individual molecules from the fibril.

Simulations for Studies of Asphericity and Crystallinity

Micelles of aggregation number N=74 were formed from OES3-*b*-PEG_{10} molecules in a random configuration in water, where polymer molecules were initially biased to reside in proximity. They were then allowed to aggregate over at least 200ns, and each system was simulated for an additional 100ns as needed, until measured metrics were observed to be converged over the last 100ns of simulation time.

Calculation of Asphericity

Asphericity was calculated as:

$$b = \frac{(\lambda_x - \lambda_y)^2 + (\lambda_x - \lambda_z)^2 + (\lambda_y - \lambda_z)^2}{2(\lambda_x^2 + \lambda_y^2 + \lambda_z^2)} \tag{2.6}$$

where λ_x^2, λ_y^2, and λ_z^2 correspond to the principal moments of the gyration tensor.

Mapping of Polymers to Rod Segments for Analysis

For the analysis of packing and crystallinity, we mapped the polymer backbones onto

rod segments that represent the molecules orientation. These rods were constructed by connecting the midpoints of two adjacent C-C or C-S bonds along the backbone through the molecule (Fig. 2.8). This yields a straight rod for a molecule that does not bend along its backbone. For molecules that do bend, the resulting rod segments facilitate analysis by virtue of their well-defined local orientation. This scheme maps 10 rod segments onto the $(OES)_3$-S- end of the OES_3-*b*-PEG_{10} block copolymer.

Definition of Crystalline Regions and Construction of Crystal Grains

For each rod segment, a neighbor list was constructed that consisted of every other rod within a radius X, which was chosen to be slightly larger than the hexagonal packing lattice parameter determined from simulations and GIWAXS measurements. Then, each rod that was aligned to within $cos(a) > 0.96$ with at least two other neighbors from different molecules was labeled as crystalline (Fig. 2.8).

Furthermore, each segment was defined as rodlike if it was aligned to within $cos(a) > 0.96$ of the prior or following rod segment in the same molecule. The cutoff value was selected by testing a range of values and picking the least error-prone cutoff via visual inspection for several cases. Then, for each aggregate, regions of aligned rods were grouped into crystal grains for further analysis using a clustering algorithm. Specifically, a seed pool from all crystalline rods in the system is created; the algorithm then picks a seed at random and initiates a new crystal grain. A recursive subroutine is launched that first removes the rod from the seed pool, adds it to the crystal grain, and launches the same subroutine for each neighbor of the rod. Thus, the algorithm crawls over individual crystal grains until all rods are removed from the seed pool and added to the grain. The process is repeated starting from another random seed until there are no more seeds left in the pool.

Definition of the S Order Parameter The order parameter S was defined as follows:

$$S = \left\langle \frac{3cos^2\theta - 1}{2} \right\rangle \tag{2.7}$$

It was used to quantify overall order within the micelle cores. Here, θ corresponds to the angle between any given two rods included in the analysis, and S ranges from -0.5 to 1, with 0 corresponding to random, uncorrelated directions, 1 to perfect angular alignment, and -0.5 to antiparallel alignment.

Synthesis of $(OES)_3$-$(PEG)_{44}$

Mono-methoxy-PEG-thioacetate (mPEG-TA) was transferred to a Schlenk tube under argon and dissolved in tetrahydrofuran (THF). The solution was stirred at room temperature under nitrogen for 30 min following addition of sodium methoxide (0.5 M solution in methanol, 1.1 equiv). After sodium methoxide activation, various equivalents of ethylene sulfide monomer were added. The ring-opening oligomerization reaction was terminated with excess glacial acetic acid. The block copolymer product was obtained with 80-90% yield after washing, filtration, precipitation in diethyl ether, and vacuum drying. ^{1}H NMR spectroscopy was performed in $CDCl_3$ on the Bruker AVANCE (500 MHz) platform with Topspin software: δ = 3.83 - 3.44 (s, broad, OC-H_2-CH_2), 3.37(s, OCH_3), 2.87 (m, CH_2SH) 2.85-2.76 (m, SCH−2CH−2), 2.74 (td,CH_2CH_2SH).

X-Ray Characterization

Silicon wafers were obtained from WRS Materials and cleaned in hot piranha (**Explosion Danger**) prior to use. To prepare thick films of micelles, a one weight percent solution of block copolymer in THF was spun-coat onto a freshly cleaned wafer at 1k RPM. The film thickness was determined to be nominally 100 nanometers by single angle ellipsometry using a J.A. Woolam Alpha-SE ellipsometry with a Cauchy model. Within one day of preparation,

Grazing Incidence Wide Angle X-ray Scattering (GIWAXS) was performed at the 8-ID-E line of the Advanced Photon Source, Argonne National Laboratory (Ref Jiang 2012). The X-ray energy was 10.82 KeV ($\lambda = 115pm$), and the scattering was captured with a Pilatus detector 228 mm from the sample. Samples were exposed for 30 seconds, as longer exposures were found to result in significant degradation, and shorter exposures resulted in a signal-to-noise ratio below 10. Data were analyzed using GIXSGUI, which allowed the correction for flat field and air scattering [53]. A standard Lorentz correction for randomly oriented 3D objects was also applied.

2.5 Appendix

Table 2.3: Crystallinity, Size and Ordering of N=74 OES_3-*b*-PEG_{10} Micelles

Simulation	% Crystalline	% Rodlike	S average	Micelle Diameter (nm)
1	0.342	81.6	0.218	6.95
2	0.289	82.6	0.179	6.54
3	0.406	82.4	0.252	6.53
4	0.326	83.4	0.242	6.59
5	0.330	80.7	0.218	6.99
6	0.419	84.9	0.314	6.51
7	0.374	79.6	0.256	6.62
8	0.412	84.8	0.251	6.78
9	0.331	82.7	0.250	7.03
10	0.408	83.2	0.289	6.70
11	0.331	81.2	0.230	6.63
12	0.238	82.6	0.180	7.24
13	0.345	81.1	0.204	7.11
14	0.396	84.3	0.249	6.87
15	0.269	80.1	0.133	7.00
16	0.415	85.0	0.303	6.82
17	0.345	80.8	0.188	6.30
18	0.396	83.9	0.291	6.80
19	0.350	83.8	0.273	7.03
20	0.395	84.1	0.247	6.91
Average	0.356	83.9	0.238	6.80
St. Dev.	0.051	0.17	0.046	0.24

Table 2.4: Normalized Principal Moments of Inertia, Asphericity of N=74 OES_3-*b*-PEG_{10} Micelles

Simulation	I_1/I_{tot}	I_2/I_{tot}	I_3/I_{tot}	Asphericity
1	0.51	0.54	0.67	1.97E-02
2	0.47	0.60	0.65	2.42E-02
3	0.48	0.55	0.68	3.01E-02
4	0.51	0.55	0.66	1.85E-02
5	0.51	0.56	0.66	1.70E-02
6	0.50	0.55	0.67	2.15E-02
7	0.41	0.59	0.69	5.98E-02
8	0.50	0.53	0.69	2.96E-02
9	0.54	0.56	0.63	7.35E-03
10	0.47	0.59	0.59	1.58E-02
11	0.50	0.59	0.63	1.38E-02
12	0.47	0.60	0.65	2.59E-02
13	0.52	0.58	0.63	8.66E-03
14	0.49	0.60	0.64	1.81E-02
15	0.49	0.57	0.66	2.02E-02
16	0.49	0.61	0.62	1.50E-02
17	0.50	0.59	0.64	1.44E-02
18	0.53	0.58	0.62	7.04E-03
19	0.50	0.55	0.66	2.11E-02
20	0.47	0.59	0.54	7.35E-03
Average	0.49	0.57	0.65	2.11E-02
St. Dev.	0.03	0.02	0.03	1.17E-02

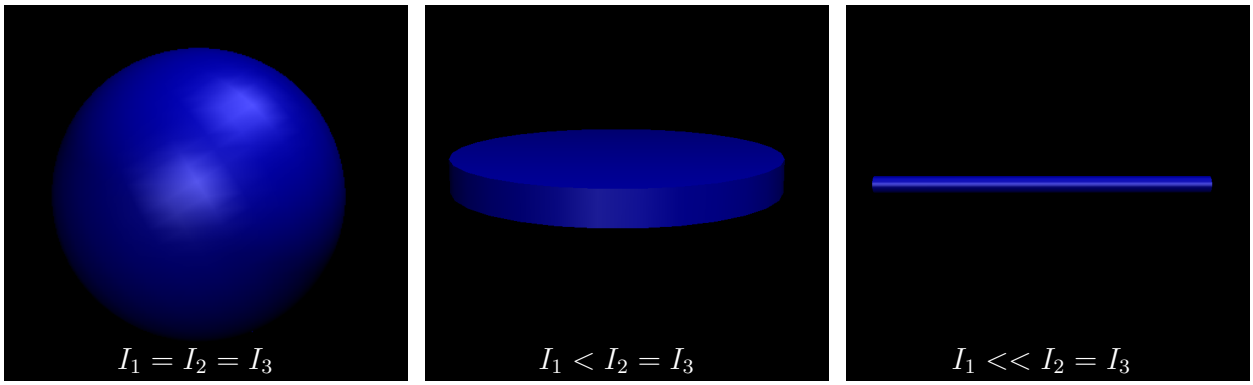


Figure 2.11: Illustration of shapes with different magnitudes of principal moments of inertia. Most micelles are closest to a sphere, with a few adopting shapes closer to a pancake.

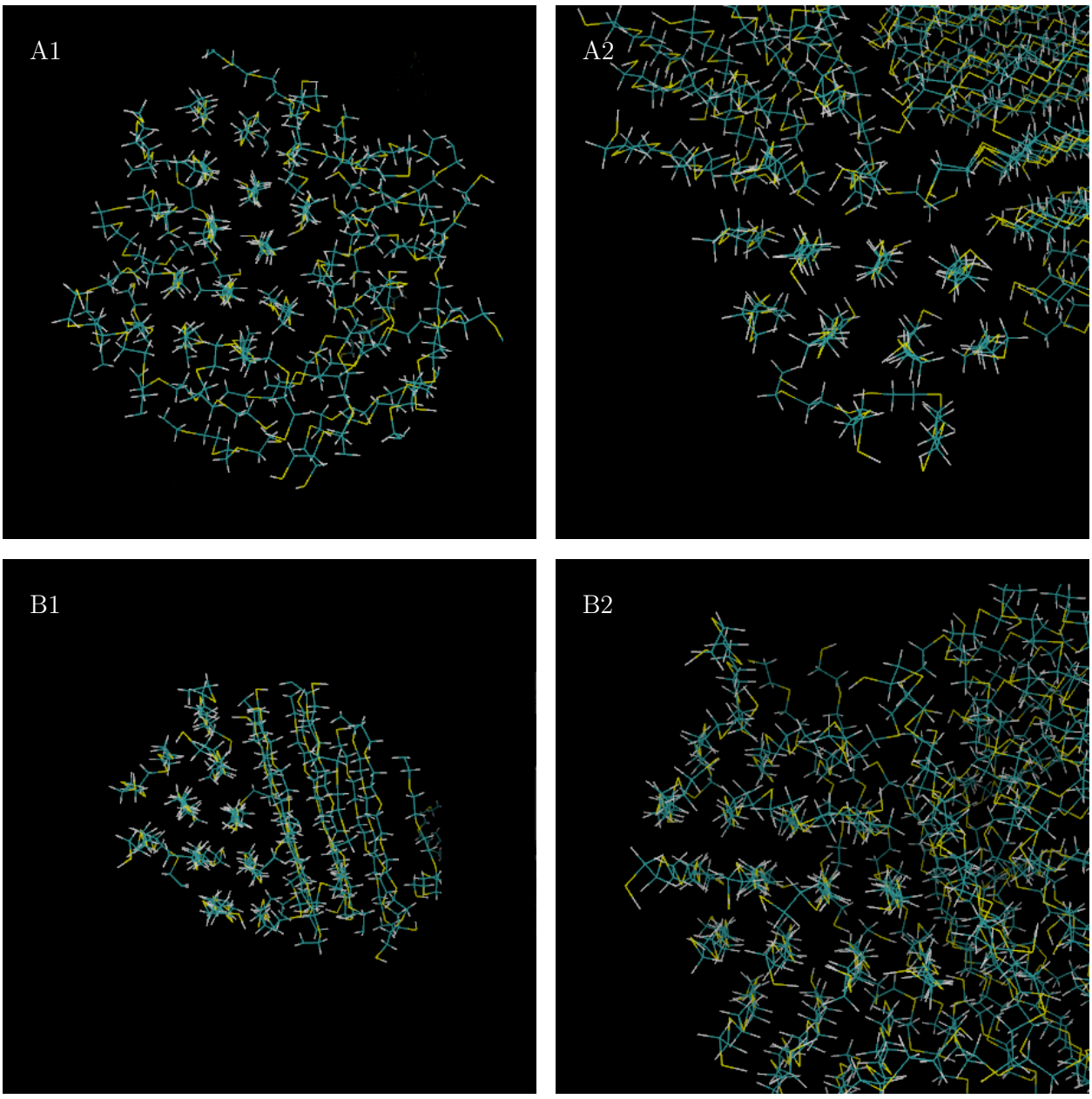


Figure 2.12: **A)** Comparison of a spontaneously formed micelle core of OES_3-*b*-PEG_5 and crystalline SH-OES_3-OH. Water molecules and PEG chains are removed from the first image for clarity.
B) Comparison of a spontaneously formed micelle core of OES_5-*b*-PEG_5 and crystalline SH-OES_5-OH. Water molecules and PEG chains are removed from the first image for clarity.

CHAPTER 3

FORCE-BIASING USING NEURAL NETWORKS

3.1 Introduction

Complex fluids, materials, or macromolecules are generally characterized by rugged free energy landscapes. Simulating such systems therefore requires that enhanced sampling methods be used to overcome the barriers that separate local free energy minima and explore configuration space efficiently. Enhanced sampling methods can also provide a direct measure of the free energy landscape, thereby leading to a deeper understanding of systems of interest. A good example of an algorithm that is commonly used to improve sampling is provided by umbrella sampling [107], where it is possible to force a system to visit a distinct state by applying a harmonic restraint. Multiple restrained simulations may then be stitched together (e.g. with a weighted histogram approach, or WHAM [59]) to generate an estimate of the underlying free energy landscape. In systems characterized by a broad or high-dimensional collective variable space, however, a naïve implementation of umbrella sampling can quickly become intractable. More recent methods, such as metadynamics [61] and its variants [89, 6, 24, 104], represent the free energy landscape as a sum of Gaussians, which are generated on the fly as a simulation visits different regions of collective variable space. In such approaches, it is important that judicious choices be made for the Gaussian heights, widths, and deposition rates, lest the convergence of a calculation be hampered [104]. Other techniques that circumvent the use of Gaussians include Basis Function Sampling (BFS) [115], Green's Function Sampling (GFS) [116], and Variationally Enhanced Sampling (VES) [109], all of which express free energy landscapes in terms of a set of orthogonal basis functions. Unfortunately, such methods are susceptible to errors near sharp features and near boundaries of the free energy landscape [75]. Note that in all of the algorithms mentioned above, an estimate of the free energy is constructed in terms of the frequency with which

distinct states are visited.

Adaptive Biasing Force (ABF) methods adopt a fundamentally different approach in that they seek to estimate the free energy landscape from its derivatives, which are given in the form of generalized forces [20]. The local nature of these mean forces can improve considerably the performance of ABF relative to that of algorithms that rely on histograms of visits to different states. Various improvements of ABF have been proposed, including an extended ABF method [63, 30] and a Metadynamics-ABF hybrid technique that is augmented with Gaussian process regression (GPR) [78]. The use of GPR for free energy reconstruction in [78] is representative of a trend to rely on machine learning techniques to enhance sampling in molecular simulations; other recent examples include utilization of artificial neural networks in the NN2B method [34] and in representation of high-dimensional free energy surfaces [98]. In this work, we build on the advantages of ABF and introduce a neural network to perform force-biasing. More specifically, the mean forces involved in the proposed method, which is referred to as "Force-bias Using Neural Networks" (FUNN) sampling, are estimated using an artificial neural network.

3.2 Method Description

The details of the ABF method have been described in the literature [21, 31, 18]; here we merely provide a brief summary on which to base our subsequent discussion. One starts with an expression for the mean force given in [21]:

$$\frac{dA}{d\xi} = -\left\langle \frac{d}{dt}(\mathbf{w}\cdot\mathbf{p}) \middle| \xi \right\rangle \tag{3.1}$$

where A is the free energy, ξ is a collective variable, and $\mathbf{p}$ is the vector of atomic momenta. $\mathbf{w}$ is an arbitrary vector field that satisfies

$$\mathbf{w} \cdot \nabla \xi = 1. \tag{3.2}$$

Here $\nabla \xi$ is the gradient of the collective variable ξ with respect to the 3N atomic coordinates. We choose $w_i = \dfrac{\partial x_i}{\partial \xi}$. This framework can be extended to multiple collective variables, and Equation 3.2 recast as $\mathbf{J}_\xi \mathbf{W} = \mathbf{I}$, where entry $[J_\xi]_{ij} = \dfrac{\partial \xi_i}{\partial x_j}$ and where J_ξ has $3N$ columns corresponding to the atomic coordinates and N_ξ rows corresponding to the number of collective variables. In order to calculate $\mathbf{W}$ while satisfying $\mathbf{J}_\xi \mathbf{W} = \mathbf{I}$, we choose $\mathbf{W}$ to be the right pseudoinverse of $\mathbf{J}_\xi$, or

$$\mathbf{W} = {\mathbf{J}_\xi}^\top (\mathbf{J}_\xi {\mathbf{J}_\xi}^\top)^{-1} \tag{3.3}$$

Typically $\mathbf{W}$ is selected in such a way as to avoid performing tedious second derivatives of collective variables with respect to Cartesian coordinates, which arise in the original ABF formulation. While second derivatives of simple collective variables may be straightforward, second order derivatives of complex collective variables may become much more difficult [18]. Note that this choice of $\mathbf{W}$ differs from that prescribed in [21], which is $\mathbf{W} = \mathbf{M}_\xi \mathbf{J}_\xi \mathbf{M}^{-1}$, where $\mathbf{M}_\xi^{-1} = \mathbf{J}_\xi \mathbf{M}^{-1} \mathbf{J}_\xi^\top$ and $\mathbf{M}$ consists of the atomic masses along the diagonal. This is equivalent to the expression in (3.3) weighted by mass. Equation (3.3) remains valid in cases with ghost atoms or virtual sites, while the mass-weighted pseudoinverse is not. In our experience, our choice of $\mathbf{W}$ works well in a variety of contexts and has been implemented in the SSAGES enhanced sampling package (along with the mass-weighted alternative) [100].

In conventional ABF, the mean force $\dfrac{dA}{d\xi}$ is estimated as the running average of $\dfrac{d}{dt}(\mathbf{W}^\top \mathbf{p})$ on a grid $\mathbf{F}(k)$, where the time derivative is calculated using a finite difference scheme; k denotes the closest bin in grid $\mathbf{F}$ corresponding to collective variable ξ. Any previously applied external bias is subtracted before updating the mean force estimates on the grid. The next applied external bias is calculated from the estimated mean force as $-\dfrac{\mathbf{F}(k)\nabla\xi}{n(k)}$,

where $n(k)$ denotes the number of times that bin k has been visited.

As mentioned above, ABF cannot bias the system on the basis of states or configurations that have not been observed yet. Furthermore, a system can easily diffuse away from a newly encountered free energy barrier before successfully climbing that barrier. Also note that ABF relies on a discrete grid to store local estimates of the mean force. Such a grid is used for both biasing the system and for estimating the underlying free energy landscape. A choice must therefore be made between precision and speed: a low resolution ABF grid may converge more rapidly, but the mean force in a particular bin is then smeared across a broader region of collective variable space, leading to errors. Estimates of the mean force are necessarily poor in the early stages of ABF due to the low number of samples in any particular bin; in some cases, fluctuations of the estimated mean forces may lead to instabilities and incorrect biasing. In practice, force estimates are artificially suppressed when a particular bin has been visited fewer than N_0 times; for example, Darve et al. recommend that the mean force estimate in the kth bin be multiplied by $\min(1, \frac{n(k)}{N_0})$, where N_0 is typically chosen to be 100-200 samples [21], which we adopt in our conventional ABF implementation [100].

To circumvent these issues, we propose to generate estimates of the mean force that are stored on a grid, as in ABF, but the grid is instead used to train an artificial neural network that produces a continuous estimate of the mean force across collective variable space. This estimate extends to regions that may not have been explored yet, thereby introducing a "scouting" biasing force that can accelerate sampling. Following in the tradition of mountain metaphors, a traveler in search of a mountain to climb does not turn around upon reaching the piedmont to diffusively wander in the flat plain that preceded the hills. She notices the slope changing beneath her feet, and anticipates that some mountains might lie somewhere ahead, even if she does not yet see them with her own eyes. In this approach, biasing forces are calculated from the continuous estimate of the mean force obtained via the neural network, allowing for the biasing force to be adaptive beyond the resolution of the underlying

grid. The use of a regularized neural network prevents the instabilities associated with noisy mean force estimates in sparsely sampled bins that are encountered during the early stages of ABF, without artificially tempering force estimates over the first N_0 samples.

In practice, an initial sweep must be performed before training the neural network, where $\frac{d}{dt}(\mathbf{W}^\top \mathbf{p})$ in the kth bin is summed into $\mathbf{F}$ while tracking the number of visits to the bin $n(k)$. While this may be performed without an external bias, it is advantageous to carry out the initial sweep with conventional ABF to avoid training the neural network with mean force estimates from a narrow strip of collective variable space; we liken the use of ABF in the initial sweep to giving the estimated mean forces a "push" before the bias supplied through the neural network takes over. The mean force estimate in the kth bin is simply $\frac{\mathbf{F}(k)}{n(k)}$. The entire grid of estimated mean forces and their k locations in collective variable space serve as our training set. Although there are countless possible neural network topologies, we choose to use a fully connected network.

The activation ϕ_i^m of any particular layer m is given by

$$\phi_i^m = f\left(\sum_{j=1}^{M} w_{ji}^m \phi_j^{m-1} + b_i^m\right) \tag{3.4}$$

where w_{ji} and b are a layer weight and bias, respectively, and i and j are indices over the number of neurons in the current and previous layers. For the activation function f, we choose the common sigmoidal activation function for all layers except the output layer, for which we use a linear activation. Backpropagation is used to train the neural network, the output of which is an estimate of the mean force which we denote $\mathbf{Q}$. Regularization improves generalizability and prevents overfitting of the neural network; if we denote the ith mean force estimate from the grid as P_i and the ith mean force estimate from the neural network output as Q_i, our loss function is

$$E = \beta \sum_i (P_i - Q_i)^2 + \alpha \sum_j {w_j}^2 \tag{3.5}$$

where the first term is the squared error between network predictions of the mean force and the target values stored in the histogram summed over all data points, and the second term penalizes network weights. Specifically, we use Bayesian regularization, which was used in the work by Sidky and Whitmer [103], and is crucial for the viability of this sampling method. Importantly, it is advantageous in requiring minimal input parameters from the end user and eliminating the need for using a subset of the sampling data as a validation set. The latter point is particularly helpful in difficult regions of collective variable space where utilizing all collected estimated mean forces is preferable, rather than using a subset for validation. In this framework, we assume Gaussian priors on the network weights and Gaussian errors on the estimates of the generalized force. To minimize loss, one must find weights that maximize the conditional probability of the weights, given the observed generalized force data; a generalizable model is obtained through this self-regularization of network weights. The loss function is minimized by solving for optimal values of α and β, which may be expressed in terms of a value γ, representing the effective number of well-determined network parameters [69, 68]. A Levenberg-Marquardt optimization routine is used to iteratively update α, β, and γ. This proceeds until either γ stabilizes, the Marquardt trust region radius exceeds 10^{10}, or a certain number of maximum training iterations has been reached [65, 70, 81]. We refer readers to [69, 68, 103] for further details on Bayesian regularization. While the exact training time depends on network topology and grid size, we find that it does not contribute significantly to the total runtime of most research systems (for example, each round of training for a 60×60 grid on a two-layer network topology of 16-12, such as the one used in the solvated alanine dipeptide example below, takes under a minute).

For each subsequent sweep, the system is biased using $-\mathbf{Q}(\xi^*)\nabla\xi^*$, where ξ^* is the instantaneous value of the collective variable, again collecting entries in $\mathbf{F}(k)$ and $n(k)$.

Herein lies a first advantage of FUNN: the trained network returns an estimate of the mean force at any point in collective variable space $\mathbf{Q}(\xi^*)$, regardless of whether ξ^* has already been visited by the simulation or whether ξ^* lies somewhere between grid points. Here $\mathbf{Q}$ supplies a hypothetical mean force estimate, even for regions that have yet to be explored, and the applied bias is calculated from the estimated mean force at ξ^*, not the nearest kth grid point. Even if $\mathbf{Q}$ supplies a poor mean force estimate in a previously unexplored region, this estimated mean force performs the critical job of driving preliminary biasing in this region. Future estimates in this region are then corrected, as further samples are accrued in the histogram corresponding to this region and the neural network is retrained.

The applied external biasing forces are easily accounted for before logging the current estimate of the true generalized force by adding the current mean force from the neural network to the instantaneous force:

$$\mathbf{F}(k) = \mathbf{F}_{\text{old}}(k) + \frac{d}{dt}(\mathbf{W}^\top \mathbf{p}) + \mathbf{Q}(\xi) \tag{3.6}$$

At the end of each sweep, $\mathbf{Q}$ is updated by training the network on the current generalized mean force estimates *from the grid*, using $\dfrac{\mathbf{F}(k)}{n(k)}$. This cycle of collecting mean force estimates on the grid and re-training the neural network to generate mean force estimates across collective variable space is iterated until convergence of the mean force landscape, which can then be integrated to recover the underlying free energy landscape.

3.3 Examples

To illustrate the application and performance of FUNN sampling, three examples of increasing complexity are considered: (1) a particle on a 2D surface consisting of a sum of 50 randomly deposited Gaussian functions, emulating a rough free energy landscape; (2) the isomerization of a simple alanine dipeptide in explicit water; and (3) the sampling of a

polymer chain along its first three Rouse modes.

3.3.1 Langevin Particle in 2D Potential

The first example consists of a single Langevin particle in a 2-dimensional square box with side length of 4 LJ units. All simulations were performed using SSAGES [100] coupled to LAMMPS (11 Aug 2017) [92]. A potential field of 50 randomly generated Gaussians was used to generate the surface features, shown in panel (c) in Figure 3.1.

Figure 3.1 shows the free energy landscape from both ABF and FUNN simulations; one can appreciate that the convergence of FUNN is several times faster than that of ABF. Free energies obtained from FUNN are shown in the left column (a-b) and compared to results from ABF at the equivalent simulation time in the right column (d-e). Free energy landscapes are calculated at 500 time-units (top row) and 1000 time-units (middle row). After only 500 time-units, FUNN has visited almost all of phase space, whereas ABF has not left its starting basin. After 1000 time-units, FUNN has essentially converged, while ABF has barely left its starting minimum, leaving most of phase space unexplored. The relative error of the simulated free energy landscape predicted by FUNN decays rapidly in the initial states of the simulation, whereas for traditional ABF it decays more slowly.

3.3.2 Alanine Dipeptide in Water

The second example, an alanine dipeptide molecule in explicit water, was simulated using SSAGES, but in this case it was coupled to GROMACS 5.1.3 [2], using the AMBER99SB force field [49]. The box size was 3 nm×3 nm×3 nm with 880 TIP3P water molecules, and a timestep of 2 fs was used. Temperature and pressure were controlled using GROMACS' stochastic velocity rescaling thermostat [11] at 298.15 K and Parrinello-Rahman barostat [84] at 1 bar.

The results are shown in Figure 3.2. As before, the exploration of collective variable

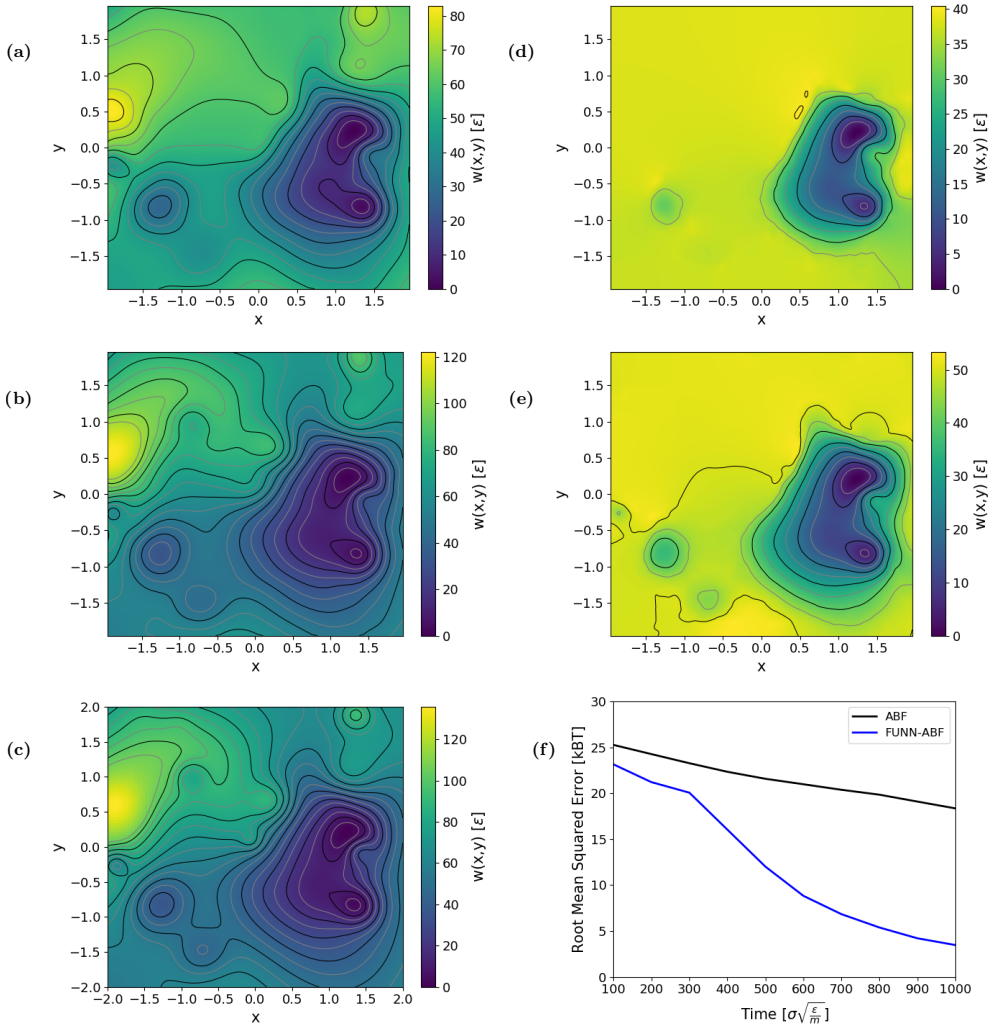


Figure 3.1: Comparison of FUNN to ABF on a 60×60 grid with a topology of 16-12 (2 hidden layers, containing 16 and 12 neurons) and a sweep interval of 5 LJ units on a surface of 50 randomly deposited Gaussian functions. (a) Results obtained from FUNN at 500 LJ timesteps, and (b) at 1000 LJ timesteps. (c) Exact surface for the 50 Gaussian landscape. (d) Results obtained from ABF at 500 LJ timesteps, and (e) at 1000 LJ timesteps. (f) Comparison of root-mean-square error over CV space with respect to the exact surface as a function of simulation time for both methods.

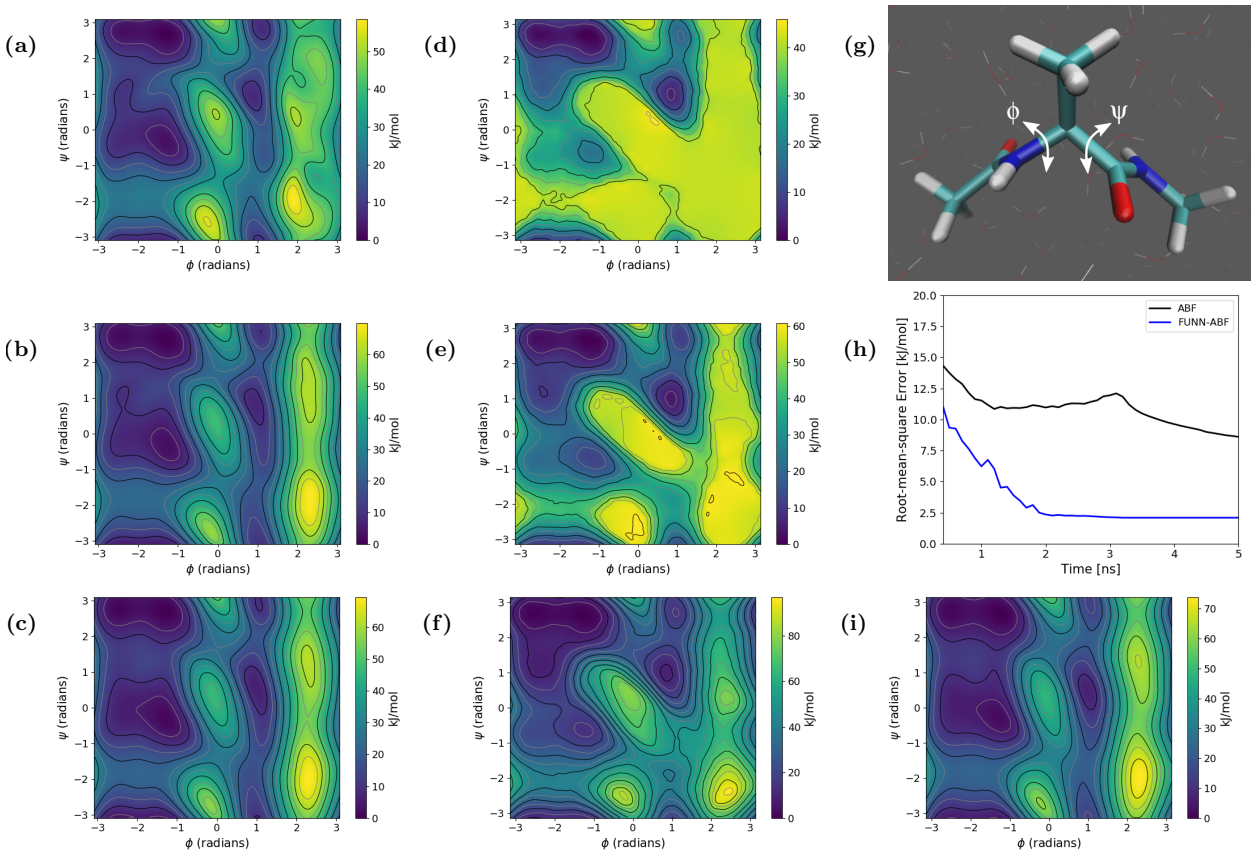


Figure 3.2: Comparison of FUNN to ABF on solvated alanine dipeptide, at 1, 2 and 5 ns from top to bottom. (a-c) Result obtained from FUNN on a 60×60 grid using a topology of 16-12 and sweep interval of 5 ps. (d-f) Result obtained from ABF on a 60×60 grid. (g) Snapshot from the simulated alanine dipeptide in explicit water. (h) Comparison of root-mean-square error for FUNN and ABF over the CV space as a function of simulation time for both methods. Error is calculated with respect to a long ABF simulation shown in panel (i). (i) Reference surface obtained from a long-time ABF simulation (120 ns).

space is faster using FUNN than using a traditional ABF method. As little as one ns into the simulation, FUNN has virtually visited all available states, and a rough but reasonable estimate of the free energy across collective variable space is already available. In contrast, most high-energy regions remain unexplored in ABF. After 5 ns, FUNN provides a converged estimate of the free energy surface, whereas ABF has not converged (as evidenced by the incorrect height of the $(\phi, \psi) = (0, 0)$ peak and the features around $(\phi, \psi) = (1, -1)$). An additional 90 ns of simulation time are necessary for ABF to converge. The relative error decays rapidly in the first 2 ns of simulation time for FUNN, whereas more than 100 ns are required from traditional ABF simulations to reach a comparable accuracy.

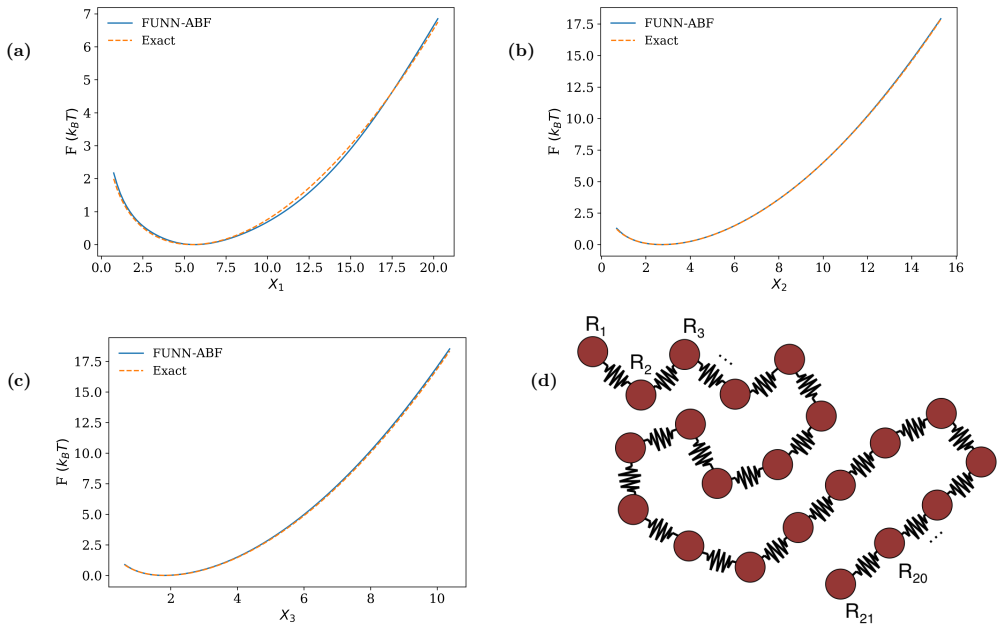


Figure 3.3: (a-c) Comparison of FUNN to analytic solution of the first three Rouse modes of a 21-bead Gaussian chain. (d) Schematic of the simulated polymer chain, consisting of 21 beads connected by harmonic bonds.

3.3.3 Rouse Modes of Coarse-Grained Polymer Chain

As a final example, we sampled the first three Rouse modes of a coarse-grained polymer chain. The polymer is represented as 21 Lennard-Jones particles with mass and parameters

set to 1.0 (see 3.3d). Langevin dynamics were used in this case, with a temperature of 0.66, a damping constant of 0.5, and a time step of 0.005. The results from sampling along these three collective variables was then compared to the analytic result in Figure 3.3.

One sees that there is excellent agreement between the exact result for the free energy and that obtained from FUNN across all three Rouse modes, serving to demonstrate that FUNN recovers the pertinent free energy profiles efficiently and accurately.

3.4 Conclusions

In summary, we have introduced a new method for performing molecular simulations that relies on learning the generalized mean forces acting on a system with an artificial neural network. FUNN builds upon the framework for conventional ABF while addressing its shortcomings. Specifically, the use of a neural network enables mean force approximations and adaptive biasing even in previously un-sampled regions of collective variable space. Furthermore, the choice of Bayesian regularization allows for minimal input from the user, eliminates the need to select a subset of the generated data as a validation set, and prevents poor mean force estimates in sparsely-sampled histogram bins. A continuous estimate for the mean force is obtained from the neural network output, allowing for force estimates and adaptive biasing with a finer resolution than that used for the underlying grid. These improvements are reflected in FUNN's superior convergence rate, as demonstrated in three examples that are representative of glassy materials, molecular conformational transitions in aqueous solvents, and polymeric materials. We note that a sparse option is also available, where the artificial neural network is trained while ignoring empty histogram bins; this is particularly useful for systems in higher dimensions. Finally, the network parameters can easily be used to query the network at arbitrary resolution after sampling is complete, allowing one to integrate the generalized forces using a more dense grid with ease.

CHAPTER 4

COMBINED FORCE FREQUENCY

4.1 Introduction

Self-assembly, conformational changes, binding events and similar processes of interest are often limited by free energy barriers. Therefore, simulating these processes requires methods beyond brute-force simulations. Enhanced sampling methods can be used to overcome these free energy barriers while simultaneously constructing an estimate of the barrier heights, leading to a deeper understanding of the kinetics and dynamics of the system. Perhaps one of the most popular of these enhanced sampling methods is umbrella sampling[107], which has been a cornerstone of free energy calculations. In umbrella sampling, multiple copies of the simulation are restrained along one or more collective variables (CVs) of interest to collect biased statistics. These statistics are unbiased after the simulations end using methods such as the weighed histogram analysis method[59]. More recently, adaptive methods have been employed frequently, such as Metadynamics,[61] where the estimate of the free energy is continuously refined and used to bias the system towards rarely visited, higher energy configurations. Metadynamics constructs the inverse of the underlying free energy surface as a function of the CVs by depositing Gaussians at a preset interval, and the system eventually behaves nearly diffusively in CV space. Though the original Metadynamics algorithm has been successful in a large number of systems, it still suffers from several drawbacks, such as non-convergent behavior and occasional challenges in picking method parameters[104]. Some of its drawbacks have been addressed via improvements to the original algorithm[89, 6, 24, 104]. Other adaptive free energy methods have also been proposed, such as Basis Function Sampling[115], Green's Function Sampling[116] and Variationally Enhanced Sampling[109], with their own set of advantages and disadvantages. Finally, neural networks have been used to generate a free energy estimate directly from a frequency histogram in ANN method[103].

Adaptive Biasing Force (ABF)[20] stands out in contrast to these methods, in particular due to how the estimate of the free energy is constructed. ABF uses an estimate of the generalized force (the derivative of the free energy with respect to collective variables) directly, rather than estimate the free energy surface from frequency of state visits. In ABF, the generalized force is estimated directly from momentum changes in the system. Though ABF is a fast and robust method, it comes with its own disadvantages such as the lack of an explicit expression for the free energy estimate, the need to select an appropriate damping hyperparameter, and the need for manual integration after the simulation. Some of these shortcomings have been successfully addressed, and significant speed-up over conventional ABF has been achieved through the use of machine learning techniques[78, 34]. In particular, FUNN[40] uses an ANN to estimate the generalized force over CV space, which yields significant improvements over the traditional ABF algorithm. Though FUNN yields noticeable speed-up over ABF and overcomes some the original algorithm's shortcomings, it still lacks an explicit expression for the free energy estimate at any given time i.e. manual integration is still necessary to obtain the free energy estimate. Such an explicit expression would be necessary for techniques such as CV-reweighing and replica exchange[51].

In this paper, we present a new method that combines the frequency-based approach to free energy estimation with generalized force-based estimation using machine learning techniques. In meta-eABF[32] as well as this work, learning from both the frequency of CV space state visits and the generalized force estimates results in faster and more accurate estimation of the underlying free energy surface compared to only one or the other as in aforementioned methods. However, in contrast to meta-eABF, we accomplish the unification without the introduction of a fictitious particle as in eABF [64], and thus avoid hyperparameter selection, oversmoothing and other associated problems that such an approach brings. In this neural network-based approach, we first extend the expression for the objective function to allow learning a function from its derivatives (here, the free energy from generalized force

estimates), which allows the construction of a free energy estimator function directly from the estimates of the derivative of the free energy, thereby solving the issue of lacking an explicit expression for the free energy from generalized force-based methods. Thus, no manual integration is necessary as the network directly provides an estimate for the free energy, even though the free energy estimate is generated using from the generalized force estimate. Then, we combine this network with another that learns from the unbiased histogram of state visits in CV space as described in ANN sampling[103]. The outputs of these networks are blended using the relative complexity of the networks as a proxy for quality of fit to produce a final estimate. In addition, we implement overfill protection[25] and sparse storage of histograms to allow scaling of the method to higher dimensions in collective variables without running into memory limitations, and to easily handle nonphysical regions of phase space often encountered using CVs such as coordination number.

Overall, this method combines the speed of generalized-force based methods such as ABF[21] or FUNN[40] with the advantages of frequency-based methods such as Metadynamics[61] or ANN method[103]. Notably, improvements over aforementioned force-based methods include removing the need to dampen early-time estimates via a hyperparameter, removing the need to manually integrate to obtain free energy, and having an explicit expression for free energy at all times, enabling replica exchange or reweighing. We also include support for overfill protection[25] to automatically avoid high-energy regions, and sparse training and data structures for higher-dimensional systems, all the while being faster than state-of-the-art force-based methods. As with FUNN[40] and ANN method[103], the network training time remains negligible compared to simulation time for all but the smallest systems.

Combined Force Frequency (CFF) method is implemented in Software Suite for Advanced General Ensemble Simulations,[100] which enables its use with a wide range of simulation engines.

4.2 Method Description

Artificial neural networks are powerful, non-linear function estimators that find use in a wide variety of fields. Generally, training an ANN requires large amounts of data, powerful hardware and time, and subsequent validation of the model. However, smaller, self-regularizing networks which train much faster have been successfully used for regression tasks, and have been successfully used in a similar context[103]. Here, similarly to ANN method and FUNN, we use a Bayesian self-regularizing network[69], which eschews the requirement for a validation set, avoids overfitting, and handles the noise inherent to the ABF estimate remarkably well [40].

In Combined Force Frequency method, an overall free energy estimate is constructed by combining frequency of state visits in CV space with an estimate generated from the generalized force. Both visit frequency and the generalized force are first collected as discretized estimates using $N-$dimensional grids, where N is the number of CVs. These grids are updated every timestep as visits and local estimates of the generalized force are collected. The grids are then each fed into separate neural networks to produce estimates of the underlying free energy. Thus, we obtain two independent estimates, one from CV space state visit frequencies, and the other from the local estimates of generalized force. The network outputs are then combined into a final estimate using a weighed average based on effective number of network parameters.

Training is performed every X timesteps of a simulation, where X is the sweep length determined by the user, and generates continuous functions with well-defined derivatives with respect to the collective variables to bias the system.

The frequency-based portion of the estimate is constructed as in ANN method[103]. For the force-based estimation, we require an explicit expression for the free energy estimate from the generalized force so that an unbiased histogram of state visits in CV space may be generated. We first start with the estimate of the generalized force as described by Darve et

al.[21].

$$\frac{dA}{d\xi} = -\left\langle \frac{d}{dt}(\mathbf{w}\cdot\mathbf{p})\middle|\xi\right\rangle \tag{4.1}$$

We wish to train the Bayesian self-regularizing ANN in a way that produces an estimate of the free energy from its derivatives with respect to the collective variables, which is the generalized force given in (4.1). To achieve this, we derive back-propagation for an extended objective function which includes terms for the ANN output with respect to its inputs. Including these terms allows training the network to learn a function (the free energy) from a training set of the target function's derivatives (the generalized force). The objective function used in ANN sampling and FUNN is

$$E = \beta \sum_i (P_i - Q_i)^2 + \alpha \sum_j {w_j}^2 \tag{4.2}$$

where the first term, scaled by β, is the squared error on network predictions to their targets (P_i are the targets, and Q_i are the network outputs, where i indexes a grid point). The second term, scaled by α, is a penalty on network weights to prevent overfitting (w_j are the network weights in no particular order). The ratio of α/β is γ, which controls the network complexity and is adjusted dynamically during training as described in Sidky et al.[103]. We extend (4.2) to include terms for the error of the output derivative with respect to the neural network inputs:

$$E = \beta \sum_i ((P_i - Q_i)^2 + (F_i - \dot{Q}_i)^2) + \alpha \sum_j {w_j}^2 \tag{4.3}$$

where P_i and F_i are the ith estimate of the target function's value and its derivative, respectively, and Q_i and $\dot{Q}_i$ are the neural network output and its derivative with respect to the neural network inputs.

Then, we derive the back-propagation algorithm that results from this objective function.

The derivative of the error with respect to the $N-1^{th}$ can be expressed as a function of the Nth layer:

$$\frac{\partial E}{\partial \phi_{N-1}} = \frac{\partial E}{\partial \phi_N}\frac{\partial \phi_N}{\partial \phi_{N-1}} + \frac{\partial E}{\partial \dot{\phi}_{N-1}}\frac{\partial \dot{\phi}_N}{\partial \phi_{N-1}} \tag{4.4}$$

where ϕ_N refers to the output of the Nth layer of the network, and a dot above a layer represents the derivative of the layer output with respect to the neural network inputs. The derivative of the error with respect to the $N-1^{th}$ layer's derivative with respect to the neural network inputs can similarly be related to the Nth layer:

$$\frac{\partial E}{\partial \dot{\phi}_{N-1}} = \frac{\partial E}{\partial \dot{\phi}_N}\frac{\partial \dot{\phi}_N}{\partial \dot{\phi}_{N-1}} \tag{4.5}$$

These expressions are propagated backwards to obtain derivatives of the error with respect to each layer. Derivatives of the layer outputs with respect to the neural network inputs ($\dot{\phi}$) are calculated during the forward pass. For the initial layer, we have:

$$\frac{\partial \phi_0^x}{\partial A^y} = \delta_{xy} \tag{4.6}$$

where ϕ_0^x refers to the x^{th} input node and A^y refers to the y^{th} entry of an input vector. Then, for linear layers we have:

$$\dot{\phi}_N = \sum w\dot{\phi}_{N-1} \tag{4.7}$$

$$\frac{\partial \phi_N}{\partial \dot{\phi}_{N-1}} = w \tag{4.8}$$

$$\frac{\partial \dot{\phi}_N}{\partial \dot{\phi}_{N-1}} = 0 \tag{4.9}$$

where w refers to weights for a given layer. Finally, for nonlinear layers:

$$\dot{\phi}_N = \sum w\sigma' \dot{\phi}_{N-1} \tag{4.10}$$

$$\frac{\partial \phi_N}{\partial \dot{\phi}_{N-1}} = w\sigma'' \dot{\phi}_{N-1} \tag{4.11}$$

$$\frac{\partial \dot{\phi}_N}{\partial \dot{\phi}_{N-1}} = w\sigma' \tag{4.12}$$

where σ is the activation function and σ' and σ'' are its first and second derivatives. Notably, the second derivative of the activation function is required, which prohibits the use of some activation functions, such as ReLU. In this work, *tanh* was used as the activation function.

Overall, a neural network with the objective function (4.3) and backpropagation derived above is trained on the generalized force data provided by (4.1), while a second neural network is trained using (4.2) on an unbiased histogram as described in Sidky et al.[103]. We take a weighed average of the two estimates, where the ratio of the γ parameters of each network, which are proxies for the complexity of the networks, is used as the weights. Weighing by γ allows a more complex network to provide a heavier bias, for instance in very early in the calculation when forces are too inconsistent to provide a good estimate, the frequency-based estimate provides the majority of the bias. The weights quickly equalize and the mixing approaches 1:1 in most cases (Fig. 4.3a). Furthermore, it is possible to use either network output or their combined output as the final estimates of the free energy. The combined output is used throughout this paper, but the estimates are often comparable (Fig. 4.3b).

We note that it is possible to train a single network on both the frequency and the force data. However, due to the discrepancy in early time estimates generated from forces and

frequencies, this approach does not work as well as training two separate networks (Fig. 4.2).

In addition, we implement the 'overfill protection' framework as generally described for histogram-based methods by Dickinson et al.[24]. To implement overfill protection, the unbiased histogram of state visits in CV space, which is used to construct the training set for the frequency-based network, is modified such that areas with low enough visit frequency that would correspond to free energies higher than the desired cutoff are artificially filled up to the cutoff. Essentially, system is not biased towards configurations that are higher in energy than the cutoff, allowing rapid exploration of areas of interest in phase space while ignoring very high energy regions.

The method also supports a sparse mode, where histograms are stored using sparse data structures, and the ANNs are only trained on non-zero entries. Combining overfill protection with sparse mode, this method can scale well to higher dimensions in collective variables if one limits the exploration of the free energy surface to an effectively lower-dimensional region consisting of lower free energy regions.

4.3 Examples

We illustrate the various features of the CFF method on a simple test system, alanine dipeptide in explicit water (Fig. 4.1a). We first show how generalized force and frequency networks contribute to convergence and overall estimate (Fig. 4.2). We then show that CFF method is faster than state-of-the-art force based methods while possessing the advantages of frequency-based methods (Fig. 4.3). Finally, we showcase overfill protection with sparse training in an intuitive example showing the results for several overfill cutoff settings (Fig. 4.5). Finally, we test the method on a model of polymer diffusion through a narrow pore for a range of pore sizes as a system representative of hidden energy barriers (Fig. 4.1b).

Throughout the paper, we use a short-hand to refer to method details when we report results. The format is 'network layers_sweep length_grid dimensions', where network layers

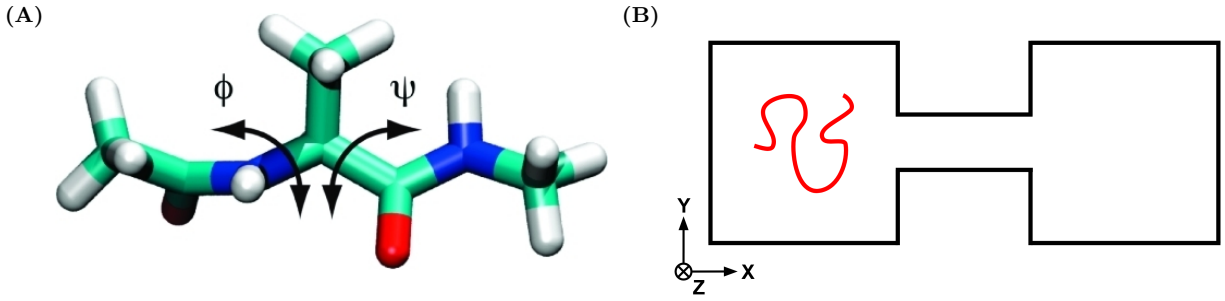


Figure 4.1: A) The dihedral angles that are the two canonical collective variables that describe alanine dipeptide. B) Schematic for polymer diffusion through a pore.

is the number of nodes in each hidden layer of the networks used to train on both estimates, sweep length is number of timesteps in between training the networks, and grid dimensions refers to the discretization used to store the state visits in CV space and the generalized force estimates. For example, 12-8_5000_30 $\times$ 30 refers to networks with two hidden layers, with 12 and 8 neurons in the first and second hidden layer respectively, which are trained every 5000 timesteps of a simulation on data stored on a 30 $\times$ 30 grid. The dimensionality of the grid corresponds to the number of CVs defined for the system.

4.3.1 Alanine Dipeptide

We use alanine dipeptide in explicit water as a test system to illustrate the features of the method. All alanine dipeptide simulations were performed using GROMACS 2016.5[2] linked to SSAGES[100] where the method is implemented, using the AMBER99SB force field[49]. The box size was 3 nm $\times$ 3 nm $\times$ 3 nm with 880 TIP3P[54] water molecules, with a timestep of 2 fs. Temperature and pressure were controlled using GROMACS' stochastic velocity rescaling thermostat[11] at 298.15 K and Parrinello-Rahman barostat[84] at 1 bar.

In the CFF method, we have access to an estimate of the free energy generated purely from an unbiased histogram of state visits in CV space, and another one from the generalized force. The combined estimate is a weighed average of the two, where the weights are the relative neural network complexities as described by the γ parameter. The individual estimates

from forces and frequencies evolve over time as the simulation progresses, which together make up the combined estimate (Fig. 4.2). Depending on the network configurations, random initialization, and the system, the histogram-based estimate may have a higher weight at early time when the generalized force-estimates are very noisy, but the ratio of the mixing quickly goes to 0.5 (equal weights) for most configurations (Fig. 4.3a).

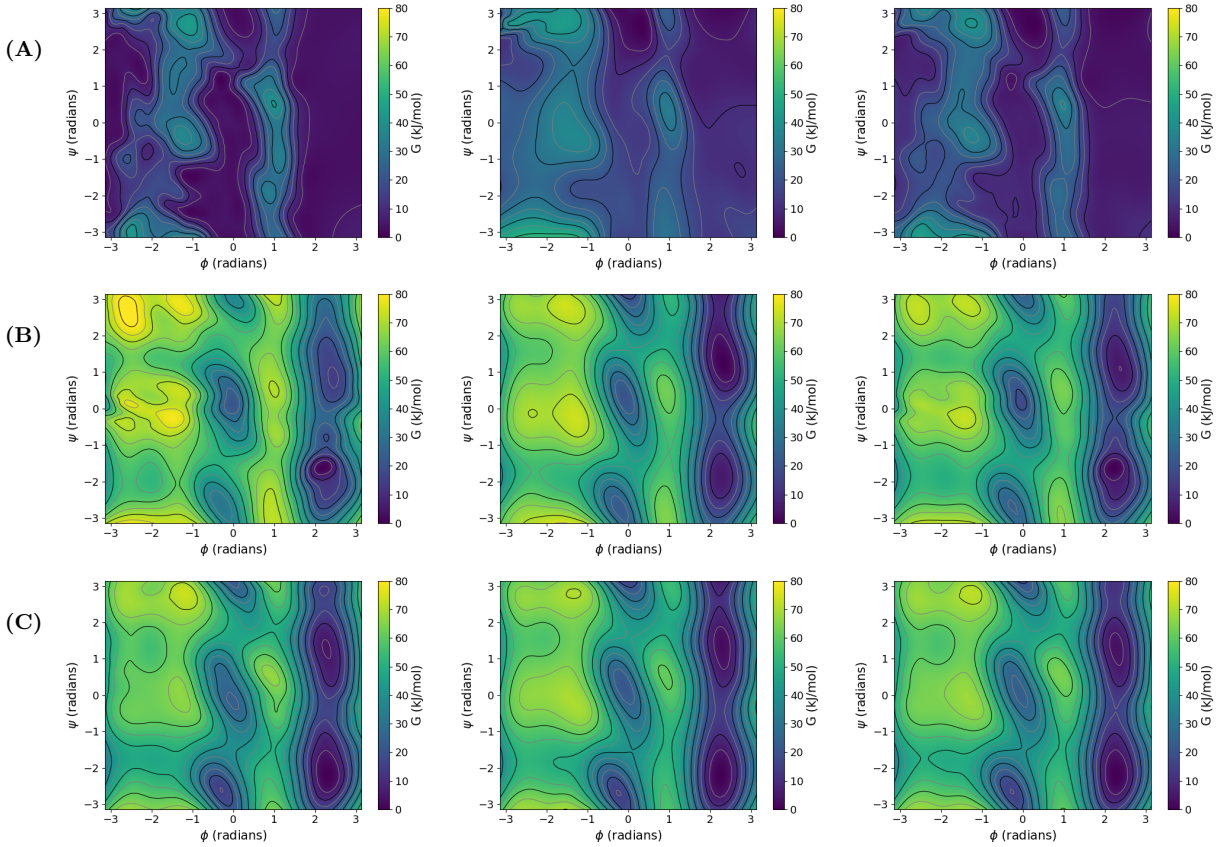


Figure 4.2: Contributions from histogram and generalized force estimates to the overall free energy surface of alanine dipeptide for 12-8_5000_30 × 30. State visit-based (left column), generalized force-based (middle column) and combined estimate of the free energy (right column) at A) 0.1 ns. B) 0.5 ns. and C) 1 ns.

Generally, unbiased state visit-based estimate captures finer features of the underlying surface, whereas the generalized force-based estimate very quickly maps out the larger features and higher barriers. The combined estimate during a CFF-accelerated simulation is often more accurate than the individual estimates, though the difference is often small and

the user is free to pick any estimate (Fig. 4.3b). In any case, all estimates will benefit from the acceleration provided by the jointly generated bias in CFF. We illustrate this speed-up over using only the state visit-based or the generalized force-based biasing in Fig. 4.4.

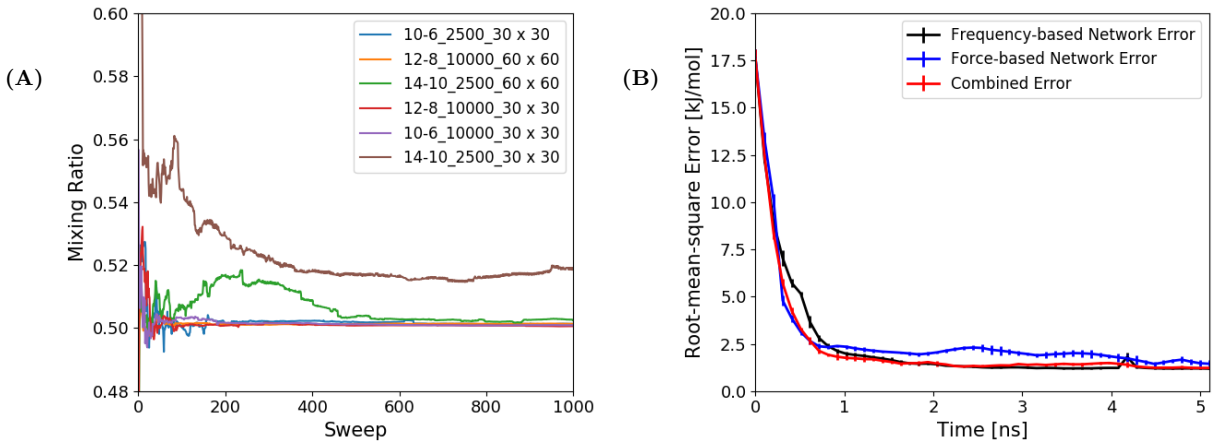


Figure 4.3: Mixing details and contributions to the overall performance. A) Mixing ratio for a sweep of network configurations on alanine dipeptide. B) Error of CFF method free energy estimates generated from the unbiased state visit frequency, generalized force, and their mixture for 12-8_5000_30 $\times$ 30.

Finally, we provide an example of overfill protection. Overfill protection is designed to allow scaling to high number of CVs, and to safely ignore nonphysical, high energy regions of phase space. When overfill protection is enabled, the bias is turned off above a user-selected cutoff, and unbiased histogram of state visits in CV space are modified to cap the maximum free energy difference in the system. The end result is that the system will only explore and generate an estimate for regions of free energy below the cutoff from the lowest free energy state discovered. As an example, we run the same alanine dipeptide system at a range of overfill settings (Fig. 4.5). As expected, with a cap of 20 kJ/mol, the system does not leave the initial basin during the short simulation. At 40 kJ/mol, the system explores both minima, and the favorable paths that connect the them. Comparable settings could be used in higher dimensions with sparse functionality to limit the system to lower energy channels and efficiently explore complex surfaces. At 60 kJ/mol, some of the lower barriers are crossed

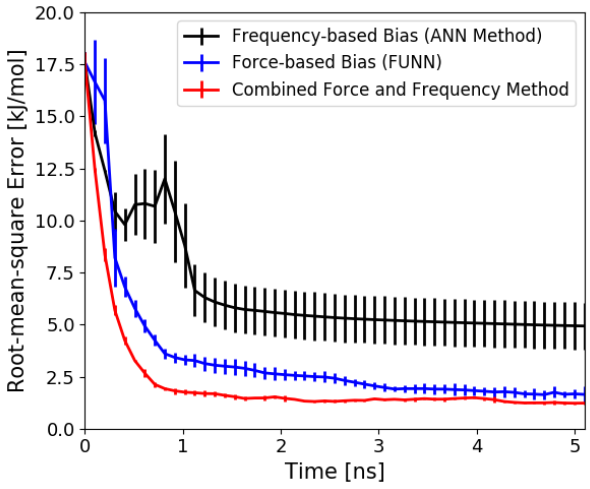


Figure 4.4: Convergence rate of CFF method compared to biasing based purely on frequency or based purely on generalized force for a network of 12-8_5000_30 × 30 for alanine dipeptide in water.

fully, but the highest energy regions are still not accessed. Notably, with the exception of the very high energy regions, the surface is well explored in just 0.5ns with overfill enabled.

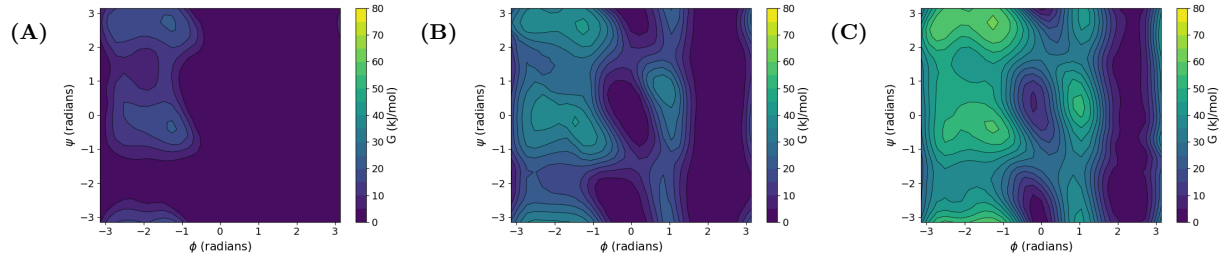


Figure 4.5: Overfill protection enabled simulations of alanine dipeptide for 12-8_5000_30 × 30. Free energy results at 0.5 ns with overfill set to A) 20 kJ/mol, B) 40 kJ/mol and C) 60 kJ/mol.

4.3.2 Polymer Diffusion Through a Pore

Finally, we test CFF method on a system with hidden energy barriers. We use LAMMPS to simulate a Kremer-Grest[57] polymer in 3 dimensions with a single 50 bead chain in a simulation box consisting of two large regions of $60 \times 60 \times 35\sigma$ connected by a 30σ long pore of variable cross-sectional area. All monomers are connected to a Langevin heat bath at $T = 1.5k_B/\epsilon$. For the FENE potential, we use k = $30\epsilon/\sigma^2$ and $R_0 = 1.5\sigma$.

Without any biasing, the probability of the polymer extending, aligning with the pore in the Y and Z dimensions, finding the correct orientation and then diffusing through the pore is extremely low. To accelerate these transition events, we apply the CFF method to some, but not all of these degrees of freedom. The end-to-end distance and center of mass in the X dimension were selected as the two collective variables, leaving pore location in Y-Z space and alignment with the pore as two hidden collective variables. Overfill protection is enabled and set to 100ϵ for these simulations which aids in simulation stability when encountering walls by limiting the maximum bias. Despite the hidden barriers, it is possible to obtain the free energies for the transition through the pore for a range of pore sizes in $5.0 * 10^6$ LJ timesteps (Fig. 4.6). We integrate out end-to-end distance for convenience, but the full 2D surfaces are available in the supplemental material (Fig. 4.7).

As expected, narrowing the pore makes the process increasingly unfavorable. Smaller pores not only have a smaller proportion of the available volume, but they also limit the number of configurations available during the crossing process. The end-to-end distance collective variable was selected to aid in unfolding the polymer, as we hypothesized that rod-like configurations would be much more likely to successfully diffuse through the pore. However, end-to-end distance did not fully capture the configurational limitations, due to a range of once-folded configurations with continuous end-to-end distances. However, the free energy calculation converged successfully despite the suboptimal choice of collective variable, even for the smallest pore size.

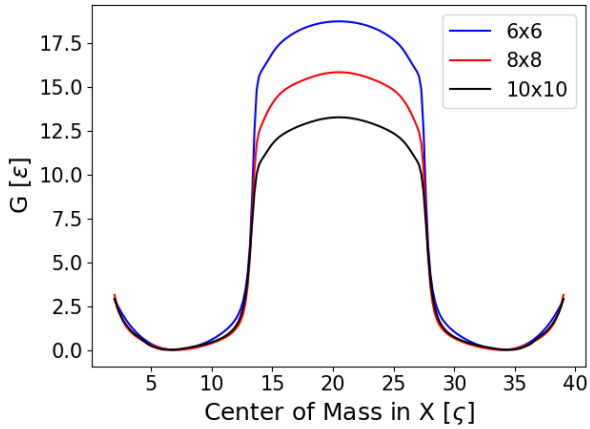


Figure 4.6: Free energy surfaces for a 50-bead Kremer-Grest polymer diffusing through a pore of 10×10 , 8×8 and 6×6 σ at $5.0 * 10^6$ LJ timesteps.

4.4 Conclusions

In summary, we introduced the Combined Force Frequency (CFF) method, a neural network-based method that learns the free energy surface from both frequencies of visits in CV-space and the generalized forces, unlike most free energy methods which use one or the other exclusively. CFF is faster than state-of-the-art force-based methods while still possessing advantages of frequency-based methods such as an explicit expression for free energy. Furthermore, Bayesian self-regularization and the ability to adjust on-the-fly the ratio of force-based or frequency-based estimation reduces user-specified hyperparameters, and renders the method minimally sensitive to various network configurations. In addition, CFF method supports overfill protection along with support for sparse storage of data, allowing CFF method to scale better to higher number of collective variables for some systems, and to automatically avoid unphysical, high-energy regions within its phase-space. Overall, CFF method is a feature rich, easy to use, fast and powerful method for free energy calculations.

4.5 Appendix

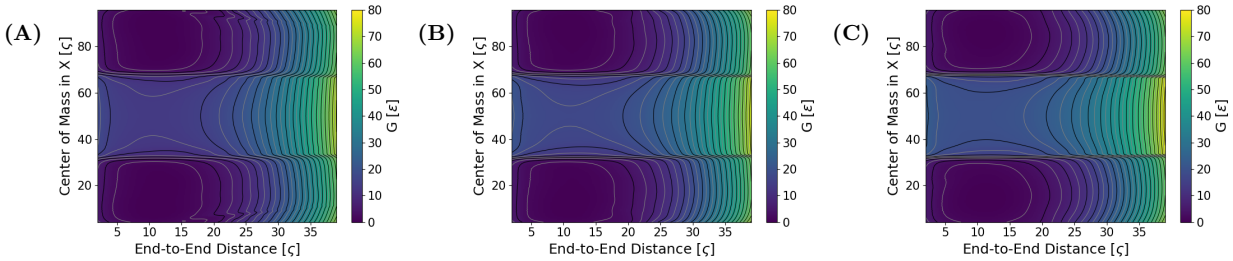


Figure 4.7: Free energy surfaces for a 50-bead Kremer-Grest polymer diffusing through a pore of A) 10×10, B) 8x8 and C) 6x6 at $5.0 * 10^6$ LJ timesteps.

CHAPTER 5

HIERARCHICAL COUPLING OF FIRST PRINCIPLES MOLECULAR DYNAMICS WITH ADVANCED SAMPLING METHODS

5.1 Introduction

Molecular dynamics (MD) simulations are routinely used to study a wide spectrum of physical phenomena, ranging from protein folding [16, 76] to glass formation [93, 43], from self-assembly [87] and nucleation[15, 36, 35], to chemical problems such as reactions in solution and at interfaces [88]. In many of these studies relevant processes are often rare events', where the characteristic timescale associated with a given transition is not accessible within a reasonable amount of simulation time, due to the presence of large free energy barriers that separate local minima along a rough free energy landscape.

In recent years, a wide range of advanced sampling methods has been developed to overcome the challenges associated with rugged free energy landscapes. These methods are generally aimed at accelerating the exploration of free energy profiles along given reaction coordinates, at finding transition pathways, and at computing dynamical properties such as transition rates [23, 22, 106, 62, 7]. For the most part, these techniques have been applied in conjunction with classical molecular dynamics, where the quantum-mechanical interactions between atoms are described in terms of pre-determined functions of the coordinates called force fields. These functions are often parametrized by relying on a combination of results from quantum mechanical calculations and experimental data [55, 50, 5, 71, 14]. Force fields enable simulations of large systems and, when coupled to sophisticated sampling methods, can be used to estimate the structure and thermodynamic properties of relatively complex fluids and materials. Force fields, however, can be inaccurate[29] and suffer from several drawbacks. Chemical reactions, for example, cannot be directly simulated by using the most

common force fields[55, 50]. Furthermore, force fields are fitted to experimental data under specific situations or states, and their transferability to conditions different from those used to parameterize the model is not insured [46, 67].

The use of first-principles molecular dynamics (FPMD) overcomes the transferability issues present in classical interatomic potentials, and allows for the description of bond breaking and formation, and, in principle, chemical reactions. In FPMD, interatomic forces are computed on-the-fly[13, 74], leading to more computationally demanding calculations than classical MD simulations. However the combination of accelerated techniques for electronic structure calculations developed in the last decades, and the computational power of modern architectures suggest that it may now be possible to use FPMD, in conjunction with sampling techniques, for calculations of free energy surfaces and the pathways along which rare transitions take place.

Only a few examples of FPMD and advanced sampling simulations have been so far reported in the literature[52, 4, 42], and several interesting results have begun to emerge from such efforts. Examples include FPMD combined with blue moon sampling, metadynamics and umbrella sampling to address problems such as chemical reactions, phase transitions and ionic or molecular dissociation[47, 58, 73, 90, 39, 27]. However, the level of theory used in these calculations has been limited to Density Functional Theory within the Generalized Gradient Approximation (GGA) or the Local Density Approximation (LDA), as more elaborate exchange-correlation functionals were deemed too computationally expensive.

In this work, we describe the hierarchical transfer of free energy estimates to enable free energy calculations using FPMD at hybrid DFT level of theory, through the coupling between Qbox[41], a C++/MPI scalable parallel code for first-principles molecular dynamics simulations, and SSAGES[100], an open-source C++11 based package for sampling simulations. The coupling includes calculations at the DFT-GGA (PBE[86]) and hybrid DFT (PBE0[3]) level of theory. SSAGES offers a selection of flat histogram, string, and flux methods that

are directly applicable to FPMD simulations. The client-server mode functionality of Qbox allows SSAGES to launch, control and combine several Qbox instances in multiple walkers fashion, thereby significantly accelerating the convergence of the sampling process. As a proof-of-concept, we present results for the free energy calculations of the alanine dipeptide (ADP) in the gas phase, and we compare them to those obtained using a classical force field. Taking full advantage of the flexibility that SSAGES and Qbox offer, we are able to significantly decrease the time required to generate a free energy surface (FES) using the hybrid PBE0 functional, by initializing hybrid simulations from the PBE free energy calculations and by using multiple walkers. We find that the classical and quantum-level FES are qualitatively similar, but exhibit important quantitative differences. In particular, the classical force field yields higher barriers than FPMD, and leads to a different minimum free energy pathway connecting the main minima on the FES; these results are consistent with those of Ref. [19], where a lower level of quantum mechanical theory was used to describe the alanine dipeptide. By analyzing the entropic and internal energy contributions to the free energy, we identify the origins of the quantitative differences between the classical and quantum free energy surfaces. Compared to the first principles calculations, Amber99sb overestimates the internal energy contribution, in particular at the transition states in high free energy regions, and it underestimates the entropy of the system.

In addition, we employ the finite temperature string method (FTS) [26] to estimate the transition pathway between the different minima of the ADP molecule. SSAGES supports several variants of the string method, along with flux methods, which are helpful for studies of reaction pathways characterized by high dimensionality[111, 83, 45, 44]. We find that both the predicted transition pathway as well as the free energy barriers along the pathway are noticeably different when computed classically and quantum mechanically.

The rest of the paper is organized as follows: In section 2 we first present our results for the free energy surface and path calculations at the first principles and classical levels. We

then discuss the structural and energetic differences of ADP obtained with different levels of theory. Section 3 concludes the paper.

5.2 Methods

5.2.1 Details of molecular dynamics simulations.

Classical MD calculations were performed using SSAGES coupled to Gromacs 5.1.3, using the Amber99sb force field. A box size of 40x40x40 Å was used. The simulations were carried out in the NVT ensemble, using GROMACS and the stochastic velocity rescaling thermostat [11] coupled at 0.1 times constant to 300K. A timestep of 2 fs was used for all classical MD calculations.

First principles calculations were performed using SSAGES coupled to Qbox 1.63.5. The box size was set at 15.875x15.875x15.875 Å. A plane-wave cutoff of 60 Ry with electronic structure tolerance of 1e-4 a.u. was used. A Bussi-Donadio-Parrinello thermostat was used to control the temperature, coupled to 300K[11]. A timestep of 0.24 fs was used for all classical MD calculations.

5.2.2 Advanced sampling methods

The adaptive biasing force method was used with 42x42 bins for the ϕ and ψ variables, with the initial linear damping set to 200 samples for all calculations. Up to 16 walkers contributed to a joint histogram in first principles runs. In order to smooth possible noise, we convolved the histogram with a small Gaussian filter.

A finite temperature string method was used with 24 nodes, initialized uniformly along the line (ϕ=-3.0,ψ=3.0) - (ϕ=3.0,ψ=-3.0). Each block consisted of 1000 timesteps, and the string constant was set to 0.1. The smoothing constant was set to 0.1. The free energy along the string was calculated from the adaptive biasing force results for analysis.

5.2.3 Simulation details

Classical MD runs were performed using a single walker and ran for a total of 100 ns for ABF and 10 ns per string node for string method simulations. The calculations took less than a day on a common workstation.

All FPMD free energy calculations were performed on 16 Intel E5-2680v4 processors, each processor hosting one walker. PBE level calculations of the FES were performed using 16 walkers, for a total trajectory time of 1.5ns. The wall time for the calculation was 2 weeks. PBE0 level calculations of the FES were initialized from the PBE results to accelerate convergence and ran for a total of 100 ps across 16 walkers. The wall time for the calculation was 2.5 weeks. The initial contribution from PBE was removed entirely from the final PBE0 result.

The FPMD string method was performed for 50 ps per string node. FPMD string method calculations were performed on 12 Intel E5-2680v4 processors, each processor hosting 2 nodes.

In all cases, calculations were performed in vacuum.

5.3 Results

5.3.1 Advanced sampling of alanine dipeptide using first principles molecular dynamics

The free energy surface of the ADP has been studied extensively with classical force fields and, more recently, using metadynamics with tight-binding Hamiltonians[7, 101, 72, 19]. The collective variables (CV) used to describe the system are the two dihedral angles, ϕ and ψ, shown in Fig. 5.1. Using these two CVs, it is possible to identify three different minima in a Ramachandran plot, describing the secondary structure of the peptide. The first minimum, in which the peptide is almost planar, is labeled β and is located at (ϕ=-2.5,ψ=2.5) radians. The second and third minima, both stabilized by an intra-molecular hydrogen bond, are

denoted as C_{7eq} and C_{7ax} and are approximately located at (ϕ=-1.2,ψ=1.2) and (ϕ=1,ψ=-1.2) radians, respectively.

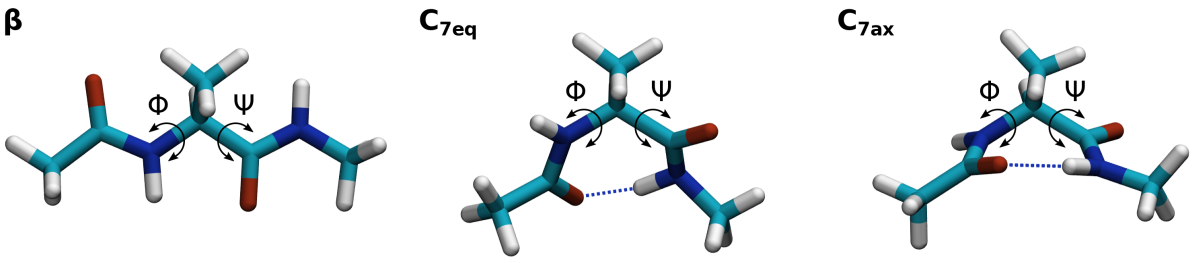


Figure 5.1: Representation of the three metastable minima, β, C_{7eq} and C_{7ax} of the alanine dipeptide together with the two angles (ϕ,θ) used to bias and analyze our calculations.

In Fig.5.2 we show a comparison between the FES of the ADP obtained with a classical force field, Amber99sb, and with first principles molecular dynamics using density functional theory with two different functionals, PBE and PBE0. To calculate the free energy surface (FES) as a function of the two CVs, we use the Adaptive Biasing Force (ABF) method as implemented in SSAGES [100, 22]. ABF was chosen over other methods due to how it generates the estimate of the free energy. ABF estimates the derivatives of the free energy directly from changes in momentum each timestep. We found it to be particularly advantageous in simulations using small timesteps, such as in FPMD, compared to methods that rely on frequency of visits, as ABF can refine the estimate locally even when diffusion through the phase space is slow, whereas other methods might be diffusion limited. In order to accelerate sampling, up to 16 individual walkers were used for tight binding MD calculations (see Methods for details).

There are clear differences between the FES calculated with DFT-PBE and the one calculated with the classical force field, Amber99sb (Fig. 5.2A). The difference is especially noticeable near the maximum located at (ϕ=2.8,ψ=2.8) in the Ramachandran plot, which is less pronounced in FPMD. In addition, the Amber99sb force field predicts a much larger barrier that spans the entirety of ψ at $\phi = 2$, likely restricting conformational transitions across that dihedral angle. The smaller barrier observed in DFT calculations is consistent

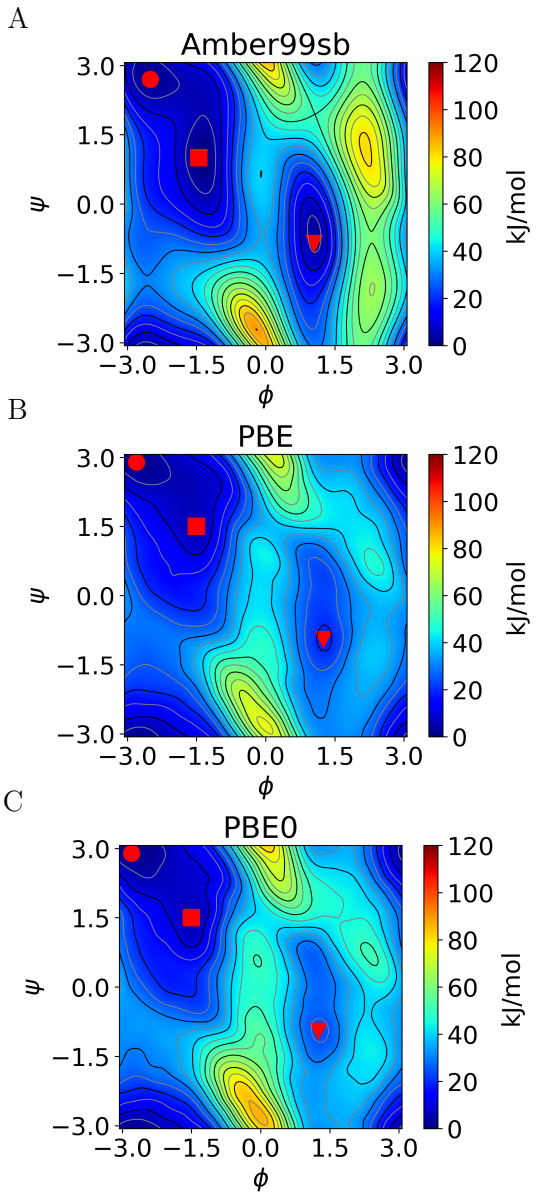


Figure 5.2: Comparison of the Free Eenrgy Surface (FES) obtained from classical and first principles molecular dynamics using the Adaptive Biasing Force method. A) Classical FES from Amber99sb force field. B) First principles result obtained at the PBE level of theory. C) First principles result obtained at the PBE0 level of theory. While there are small quantitative differences between the PBE and PBE0 calculations, the Amber99sb results differs from both. In particular, Amber99sb predicts a higher barrier in the $\phi = 2$ region. The position of the β, C_{7eq} and C_{7ax} minima are defined here as •, ■, and ▼, respectively (see Fig. 5.1.

with earlier FPMD simulations of the alanine dipeptide [19].

In order to investigate the dependence of the FPMD results on the functional chosen, the FES was recalculated using an hybrid functional (PBE0). Since these calculations are significantly more demanding than those carried out with PBE, we estimated the initial PBE0 FES from the converged ABF histogram obtained from PBE. This strategy allows the simulations to converge considerably faster. The PBE contribution to the FES was removed at the end of the hybrid simulation, yielding the pure PBE0 result. The morphology of the PBE0 FES is similar to that of PBE, but the PBE0 functional predicts slightly higher barriers to transitions between $\beta \longrightarrow C_{7ax}$ and $C_{7eq} \longrightarrow C_{7ax}$, as well as a higher relative free energy for the C_{7ax} minimum than PBE(see Fig. 5.2B and 5.2C).

While obtaining a converged FES is relatively straightforward for a small number of collective variables, it may quickly become prohibitive for systems requiring many collective variables when using flat histogram methods such as ABF. The string method and its variants are instead particularly useful for identifying transition pathways and calculating the free energy along them, as these methods focus only on the states along the transition path. This feature allows string methods to only sample states relevant to the transition, thereby enabling calculations with increasing numbers of CVs.

To illustrate how the string method may be useful in the context of FPMD simulations, we performed finite temperature string method simulations using DFT-PBE for the alanine dipeptide[111]. The computed FES suggests that there are two possible transitions from β to C_{7ax}: one involves the formation of the metastable state C_{7eq}, and then a transition to C_{7ax}, without breaking the intramolecular hydrogen bond. A second transition involves the direct formation of an intramolecular hydrogen bond to the C_{7ax} state, starting from the β state. We thus initialized the string to include both transitions, from (ϕ=0,ψ=0) to (ϕ=3.1,ψ=-3.1). The converged string was found to be consistent with the FES obtained by ABF, being perpendicular to the contours of the FES, as illustrated in Fig 5.3A. At a

qualitative level, the PBE string path resembles the classical result, but the transition at $\phi = 0$ deviates noticeably from the classical result, as shown in Fig. 5.3b.

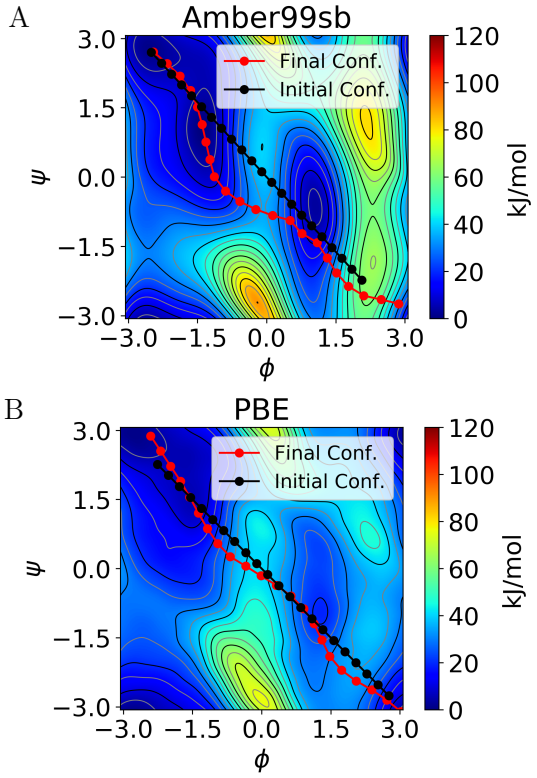


Figure 5.3: Finite temperature string method results overlayed on the free energy surface. In black and red are reported the initial and final configuration, respectively. A) Results from classical molecular dynamics using Amber99sb force field B) Results from FPMD using the PBE functional. The pathways are qualitatively similar, yet there are differences in their positions.

It is instructive to plot the free energy along the reaction coordinate, i.e., the string in this case, to illustrate the differences in barrier heights between PBE and Amber99sb. Figure 5.4 shows that Amber99sb calculations predict a higher barrier height for the $C_{7ax} \longrightarrow \beta$ compared to PBE. The evaluation of the FES along the string in the classical and PBE simulation reveals that the lowest transition pathway connecting $\beta \longrightarrow C_{7ax}$ is different in the two cases: using Amber99sb, the transition occurs following the path $\beta \longrightarrow C_{7eq} \longrightarrow C_{7ax}$, whereas in FPMD simulations the secondary path $\beta \longrightarrow C_{7ax}$ is favored, as the barrier for the transition $C_{7eq} \longrightarrow C_{7ax}$ is higher. We emphasize that the Amber99sb and

PBE calculations predict the same structure in the metastable minima and at the transition states, but different transition pathways. In a larger system with more dihedrals, e.g. a longer polypeptide, the difference in the folding-unfolding path predicted classically and quantum mechanically could be even more significant.

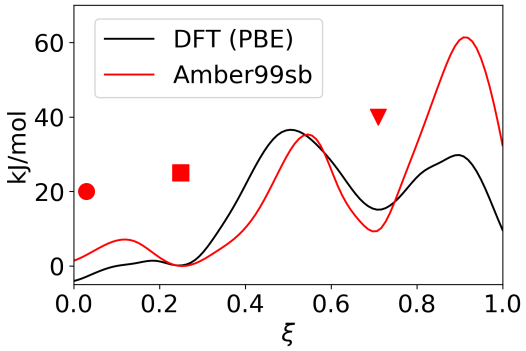


Figure 5.4: Free energy along transition paths calculated using finite temperature string for Amber99sb (red) and DFT PBE (black). The two levels of theory predict two different transition pathways: while Amber99sb predicts the minimum free energy pathway to be $\beta \longrightarrow C_{7eq} \longrightarrow C_{7ax}$, the PBE case predicts it to be $\beta \longrightarrow C_{7ax}$ due to different barrier heights. The β, C_{7eq} and C_{7ax} minima are denoted as ●, ■, and ▼, respectively (see Fig. 5.1).

5.3.2 Analysis of Differences Between Classical and first principles Results

To understand the origin of the differences between DFT and the Amber force field described above, we performed additional structural and energetic analyses.

We first quantified differences in the geometrical structure predicted by Amber99sb, PBE and PBE0, by computing the average Root Mean Square Displacement (RMSD) between the accessible structures at a given point (ϕ,ψ) in CVs space:

$$RMSD(\phi,\psi) = \frac{1}{N(\phi,\psi)} \sum_{i}^{N(\phi,\psi)} \sum_{j\neq i}^{N(\phi,\psi)} \sum_{k}^{Nat} \sqrt{(\mathbf{r}_k^i - \mathbf{r}_k^j)^2} \tag{5.1}$$

where N is the total number of configurations at a given (ϕ, ψ) point, and Nat is the total number of atoms composing the backbone of the peptide. The RMSD parameter is small if the N configurations are all very similar, whereas higher values indicate higher diversity. We have not included hydrogen atoms in the parameter definition, as they do not provide any relevant information on the structure. The results for Amber99sb, PBE and PBE0 are reported in Fig 5.5.

The three levels of theory yield similar qualitative features. The RMSD surfaces are characterized by two regions of low RMSD, one roughly corresponding to the C_{7eq} and a second one located approximately at (ϕ=2,ψ=-2). The positions where the RMSD reaches a maximum correspond approximately to the high free energy states.

The most important contribution to the general form of the plot, and especially to the low RMSD 'channel' going from the C_{7eq} to (ϕ=3,ψ=-3), is the intra-molecular hydrogen bond that forms in the peptide (Fig. 5.6). Given the hydrogen bond formation, the position of the basin corresponding to C_{7eq} is not a surprise, as the intra-molecular hydrogen bond limits the flexibility of the peptide. To our surprise, however, the second basin does not exactly correspond to the C_{7ax} minimum, and its position is closer to the barrier along the $\beta \longrightarrow C_{7ax}$. Despite quantitative differences, however, the distribution of geometries that the three levels of theory predict are similar, and we conclude that they do not represent the main origin for the difference in the FES reported in Fig 5.2.

Minor differences can also be noticed in the distance between the hydrogen and the oxygen involved in the intramolecular hydrogen bond. As can be noted in Fig. 5.6, the transition between the $C_{7eq} \longrightarrow C_{7ax}$ does not involve the breaking of the intra-molecular hydrogen bond. The barrier between the two metastable states is most likely due to steric hindrance. The hydrogen bond distance is also correlated with the RMSD illustrated in Fig. 5.5: both contour plots show the same channel which follows the $C_{7eq} \longrightarrow C_{7ax}$ transition, illustrating that the intra-molecular hydrogen bond tends to decrease the molecular mobility.

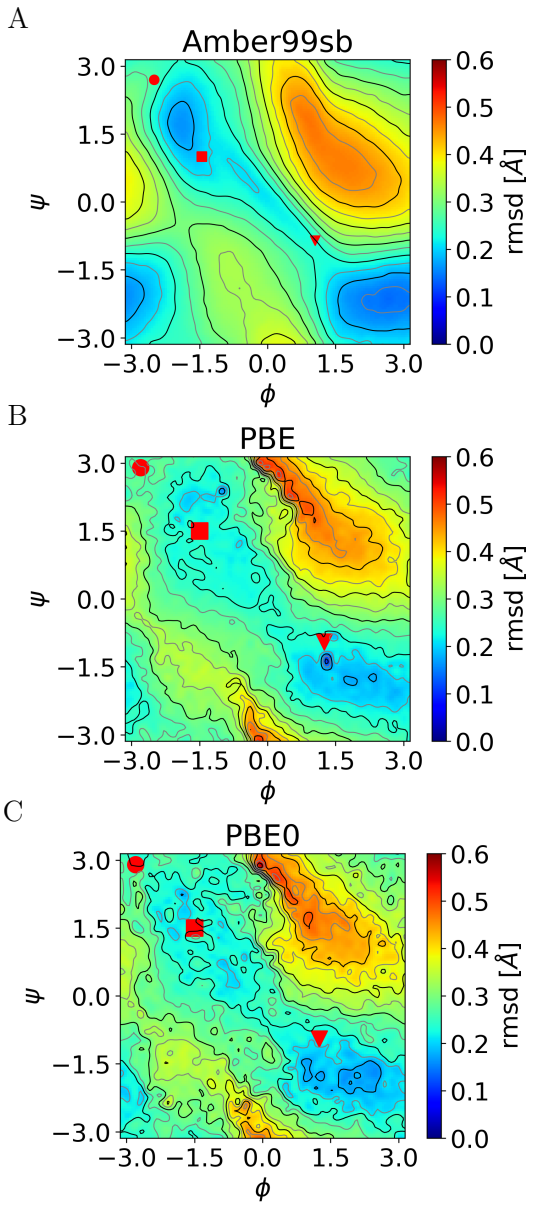


Figure 5.5: The average diversity of configurations adopted in phase space during a molecular dynamics simulations, as obtained with classical force field and with first principle molecular dynamics. A) Average local root mean square displacement (see Eq. 5.1) from Amber99sb. B) Average local root mean square displacement (see Eq. 5.1) from first principles within the PBE functional. The two surfaces are quantitatively different. The β, C_{7eq} and C_{7ax} minima are denoted as •, ■, and ▼, respectively (see Fig. 5.1.

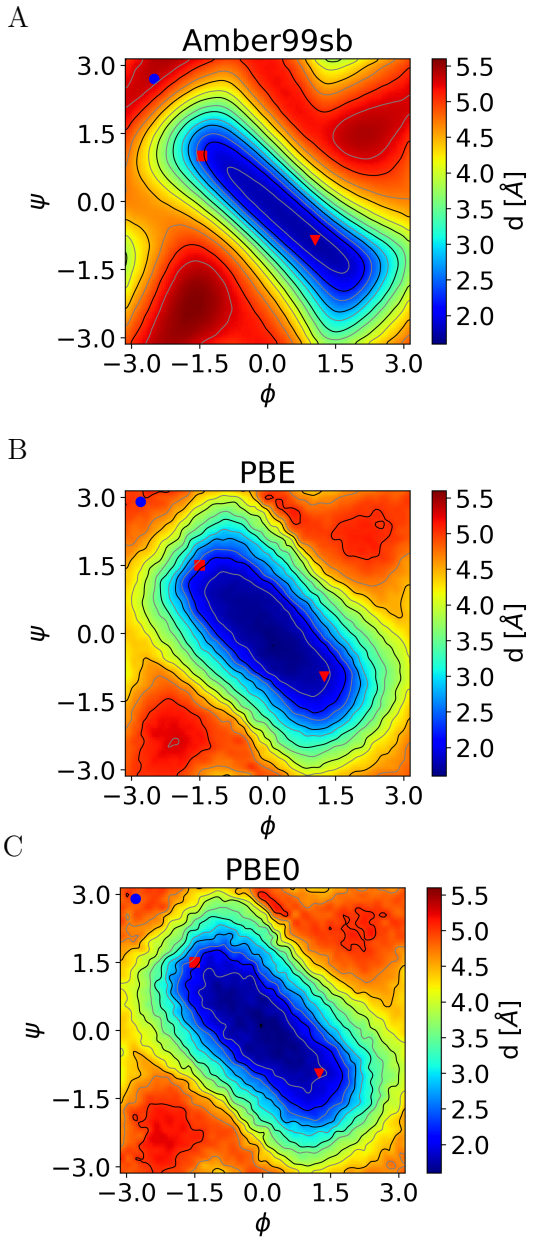


Figure 5.6: Average distance between the oxygen and the hydrogen stabilizing the C_{7ax} and C_{7eq} structures. The minimum distance (roughly 1.8 Å, located at(ϕ=0.0,ϕ=0.0)) is very close to the saddle point for the transition $C_{7eq} \longrightarrow C_{7ax}$. The Amber99sb force field predicts a slightly different average distance than the PBE and PBE0 simulations. The β, C_{7eq} and C_{7ax} minima are denoted as •, ■, and ▼, respectively (See Fig. 5.1).

Next we investigated whether the FES differences originate from different internal energy and/or entropic contributions in the classical and FPMD simulations. We partition the FES as:

$$\Delta A(\phi, \psi) = \Delta U(\phi, \psi) - T\Delta S(\phi, \psi) \tag{5.2}$$

The internal energy contribution (U) to the FES was calculated as the local running-average of the internal energy computed during the MD simulations on a grid. An estimate of the entropic contributions, $T\Delta S$, was obtained from the difference $(\Delta A - \Delta U)$. Fig. 5.7 shows the potential energy surfaces, which are qualitatively similar, and resemble the free energy surfaces obtained using ABF calculations. The classical force field predicts a higher internal energy than the PBE and PBE0 functionals in the region corresponding to the $\phi = 2$ dihedral, which is reflected in the FES. The differences observed between PBE and PBE0 FES are mirrored here, as the barriers predicted by the PBE0 functional are higher than those predicted at the PBE level of theory. This effect is also visible for the C_{7ax} minimum, which is less stable when using PBE0. It is important to emphasize that the DFT and classical PES differ the most in the low probability states. This difference is also maintained when DFT and Amber99sb are compared with a higher level of theory such as CCSD(T) [56]. Notably, there is a consensus of all he methods, regarding the free energy difference between the different metastable minima. However, there are minor local minima predicted by MP2/cc-pVTZ//MP2/6-31G** calculations [113], for which our simulations with PBE and PBE0, along with B3LYP and CCSDT/CBS-aVDZ give consistent energies [56]. Interestingly, the PBE0 functional captures most of the features observed in the MP2/cc-pVTZ//MP2/6-31G** calculation of the alanine dipeptide PES [113]. A comparison between energies at all the identified minima are reported in the SI, in Table SI-1 and Fig. SI-1.

The entropic contributions to the free energy also exhibit significant differences. The classical case predicts a low entropy, in contrast to PBE and PBE0. The entropy maxima

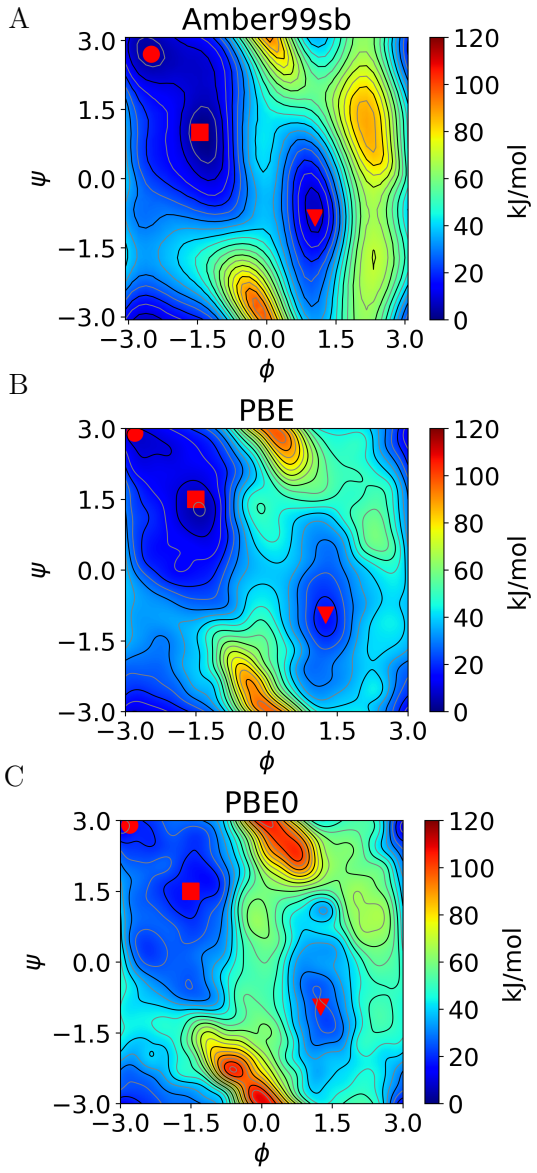


Figure 5.7: Comparison of potential energy surfaces. A) Potential energy surface from the classical force field Amber99sb B) Results from first principles calculations using the PBE functional. C) Results from first principles calculations using the PBE0 functionals. The classical force field predicts a higher barrier in the region identified by $\phi = 2$ than DFT calculations. The PBE0 functional predicts a higher barrier than PBE, as well as a less stable C_{7ax} minimum. The β, C_{7eq} and C_{7ax} minima are denoted as •, ■, and ▼, respectively (See Fig. 5.1).

for all the three cases are located in proximity of the FES maxima. Two of the minima correspond to the C_{7eq} and C_{7ax}, while the third one located at (0.5,-2.5) does not correspond to a minimum in the FES or to a well-defined structure. Even though they are on an entirely different scale, the morphologies of the surfaces are similar in the classical, PBE and PBE0 cases, with the minima and maxima in the same positions (Fig. SI-2). Most likely, the classical force field predicts such a low entropic contribution because the entropy is not explicitly included in the fitting of the force field. We note that the entropic contributions may be important to obtain a correct description of large peptides, in particular for folding and unfolding processes.

We end our discussion of the results by emphasizing that our calculations do not imply or prove that the FES computed from FPMD simulations are superior to those of a classical force field for the description of peptide molecules. The classical force field has been parametrized by using extensive amounts of experimental data for peptides, both isolated and in aqueous solutions. Instead, we wish to stress two important points. First, our premise is that a general aim of condensed matter theory is to one day be in a position to predict the properties of macroscopic fluids and complex materials from first principles. Free energies are essential for such pursuits, and our calculations serve to establish that one can now predict free energies and free energy pathways by relying on advanced sampling methods that, until now, have been almost exclusively used in the context of classical force fields. Second, we note that when Ramachandran plots are used to parameterize classical force fields by relying on quantum mechanical calculations, they are expressed in terms of the internal energy. For such endeavors, a more appropriate quantity is the free energy. Our work illustrates that for a representative dipeptide the free energy surface differs considerably from the internal energy surface, particularly in the vicinity of high energy barriers, where entropic contributions to the free energy can differ appreciably.

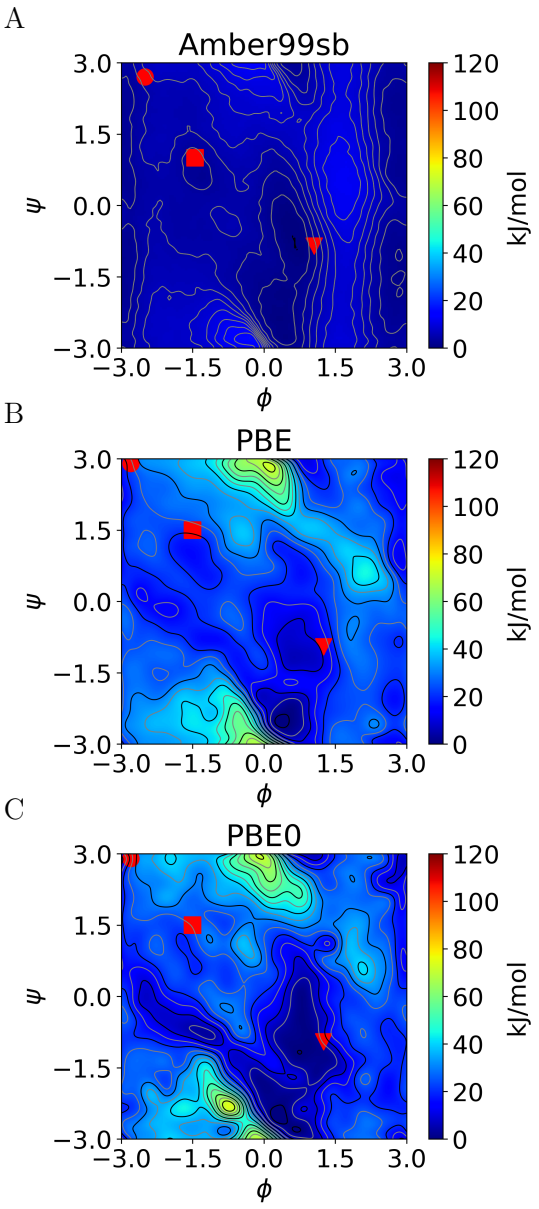


Figure 5.8: Comparison of the entropic term $T\Delta S$ as obtained from classical and FPMD. A) Entropy surface from the classical force field Amber99sb. B) Entropy surface from first principles calculations using the PBE functional. C) Entropy surface from first principles calculations using the PBE0 functional. Refer to Fig. SI-2 in the Supplementary Information for a comparison of the purely entropic term.

5.4 Conclusions

Our first aim in this work was to present a methodology to calculate free energy surfaces using first-principles molecular dynamics with demanding functionals, leveraging the coupling between Qbox, a DFT code capable of doing FPMD simulations with local and hybrid functionals, with SSAGES, a general engine-agnostic code to perform enhanced sampling calculations. Some of these free energy methods have gained popularity through their availability in public domain software packages, e.g. Plumed or Colvars[108, 8, 28]. However, the coupling presented here increases the number of enhanced sampling methods usable with FPMD, and it facilitates their use thanks to the architecture of the coupling. We emphasize that the two codes are not compiled together, but work in a client-server mode. In this way, the two codes have minimal dependencies, and it is straightforward to execute multiple walkers calculations. Thanks to the scalability of Qbox, its fast evaluation of hybrid functionals, and the possibility or restarting SSAGES calculations from a previous results, one can resort to a hierarchical sampling approach. In this paradigm, it is possible to restart hybrid functionals calculations from previous semi-local functional ones, greatly diminishing the computational burden needed to converge the simulations.

To illustrate the efficacy of our coupling, we calculated the FES for alanine dipeptide in vacuum at the PBE and PBE0 levels of theory. To the best of our knowledge, this is the first FES calculated using a hybrid functional. We also calculated the minimum free energy path using finite temperature string method at the PBE level of theory. A key strategy in obtaining a converged free energy surface at such high level of theory is the hierarchical transfer of free energy estimates. Leveraging the implementation of ABF in SSAGES, it was possible to easily initialize the ABF histogram for PBE0 from the converged PBE result, making PBE0 nearly diffusive from the start. We note that the same strategy could be applied to investigate how dispersion correction influences the result. In particular, it would be possible to perform the same calculation with the OPTB88 functional by restarting from

the PBE result, greatly diminishing the computational resources required.

Transferring the estimate from PBE, obtained in 1.5ns total simulation time, to PBE0 reduces total required simulation time to just 100ps - a 15x improvement. A 1.5ns calculation at PBE0 level of theory would have taken more than 30 weeks instead of 2.5 weeks used to obtain the result with the transfer. Initial PBE contribution was completely removed from the final result.

The results are consistent with similar calculations performed at a lower level of theory[19], but show significant deviations from classical results determined using a classical force field (Amber99sb). In particular, the classical calculation overestimates the internal energy contribution in an entire region of the CVs space ($\psi = 2$), predicting a free energy surface with a higher barrier than that of the PBE/PBE0 calculations. This higher barrier causes a change in the transition pathway between the $\beta \longrightarrow C_{7ax}$ minima which passes through C_{7eq} in the classical case, but does not in the FPMD case. Furthermore, the difference manifests as a significantly higher barrier across the transition in the classical case. In addition, Amber99sb underestimates the entropic contribution to the free energy surface, when compared to the semi-local and hybrid case.

The result for the ADP case is of importance because force fields are generally fitted to properly represent equilibrium properties, such as the dihedral angles of the backbone of a protein. However, they are not always designed to include information about transition states and high free energy states. In order to establish whether the classical or the first principles calculations are better representatives of reality, calculations of solvated systems will be required and these are work in progress.

Both codes are open-source and are readily available at https://github.com/MICCoM/SSAGES-public and at http://qboxcode.org/. The coupling used in this work is implemented and can be setup off-the-shelf, allowing the application of many flat histogram, string and flux methods to first principles molecular dynamics. We believe that the combination of

these techniques will provide a new way of solving exciting problems, from chemical reactions to structural characterization of solid-liquid interfaces. Force field construction with these methods is particularly appealing, as one can match the classical and first principles free energy surfaces, rather than solely fitting to the potential energy surface.

CHAPTER 6

CONCLUSION

In this work, we have shown an application of enhanced sampling on molecular dynamics to obtain free energies, and using these free energies to predict material properties. Specifically, calculating the free energy of dissociation of a single molecule from a micelle in a self-assembling system allows the estimation of the relative stability of the micelle. Performing such calculations for a range of sizes using a statistical mechanical model of micelle formation provides further information, allowing predictions of equilibrium micelle size, micelle size distribution, and stability. Further studies of the micelle core provide answers into the mechanisms of micelle formation.

This study is one of countless possible ways of leveraging free energy to infer properties of interest, including but not limited to kinetics, population densities across phase space, phase diagrams, and protein conformational changes. However, as the systems and processes studied get more complex, the computational demands to calculate relevant free energies also increase dramatically. Therefore, faster free energy methods are one of the most efficient ways of expanding the feasibility limit for simulating larger systems. To this end, two new free energy methods are developed and made available to the community. These methods take advantage of neural networks to better estimate free energy during a simulation and significantly accelerate the exploration of phase space.

Finally, we present a framework and hierarchical approach for bringing new free energy methods to the first principles molecular dynamics community, where such approaches remain underutilized. In this framework, SSAGES is coupled to QBox to inject easy-to-use and powerful enhanced sampling methods into a fast and scalable first principles molecular dynamics package. The coupling enables multi-level parallelization through SSAGES on top of the inherent layer of QBox for the best possible performance on a range of hardware configurations and simulation setups. In addition, free energy calculations are further ac-

celerated through the hierarchical approach, where a cheap estimate is constructed using a lower level of theory and then refined at a higher level of theory for higher accuracy. Such hierarchical sampling is easy to perform in SSAGES, and requires minimal post-processing to obtain a free energy from high level theory simulations for a much lower computational cost.

References

[1] Jose LF Abascal and Carlos Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *The Journal of Chemical Physics*, 123(23):234505, 2005.

[2] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, September 2015.

[3] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The pbe0 model. *The Journal of chemical physics*, 110(13):6158–6170, 1999.

[4] Wanda Andreoni and Alessandro Curioni. New advances in chemistry and materials science with cpmd and parallel computing. *Parallel Computing*, 26(7):819–842, 2000.

[5] Volodymyr Babin, Claude Leforestier, and Francesco Paesani. Development of a first principles water potential with flexible monomers: Dimer potential energy surface, vrt spectrum, and second virial coefficient. *Journal of chemical theory and computation*, 9(12):5395–5403, 2013.

[6] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100(2):1–4, 2008.

[7] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*, 100(2):020603, 2008.

[8] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A Broglia, and Michele Parrinello. Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10):1961–1972, 2009.

[9] Clement Bordier. Phase separation of integral membrane proteins in triton x-114 solution. *Journal of Biological Chemistry*, 256(4):1604–1607, 1981.

[10] Carrie E. Brubaker, Diana Velluto, Davide Demurtas, Edward A. Phelps, and Jeffrey A. Hubbell. Crystalline oligo(ethylene sulfide) domains define highly stable supramolecular block copolymer assemblies. *ACS Nano*, 9(7):6872–6881, 2015.

[11] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126, 2007.

[12] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007.

[13] Richard Car and Mark Parrinello. Unified approach for molecular dynamics and density-functional theory. *Physical review letters*, 55(22):2471, 1985.

[14] Chen Chen, Christopher Arntsen, and Gregory A Voth. Development of reactive force fields using ab initio molecular dynamics simulation minimally biased to experimental data. *The Journal of chemical physics*, 147(16):161719, 2017.

[15] Bingqing Cheng and Michele Ceriotti. Bridging the gap between atomistic and macroscopic models of homogeneous nucleation. *The Journal of chemical physics*, 146(3):034106, 2017.

[16] John D Chodera and Frank No. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology*, 25:135 – 144, 2014.

[17] Don W Cleveland, STUART G Fischer, MARC W Kirschner, and ULRICH K Laemmli. Peptide mapping by limited proteolysis in sodium dodecyl sulfate and analysis by gel electrophoresis. *Journal of Biological Chemistry*, 252(3):1102–1106, 1977.

[18] Jeffrey Comer, James C. Gumbart, Jérôme Hénin, Tony Lelievre, Andrew Pohorille, and Christophe Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B*, 119(3):1129–1151, 2015.

[19] Jerome Cuny, Kseniia Korchagina, Chemseddine Menakbi, and Tzonka Mineva. Metadynamics combined with auxiliary density functional and density functional tight-binding methods: alanine dipeptide as a case study. *Journal of Molecular Modeling*, 23(3):72, Feb 2017.

[20] E. Darve and A. Pohorille. Calculating free energies using average force. *J. Chem. Phys.*, 115(20):9169–9183, 2001.

[21] Eric Darve, David Rodríguez-Gómez, and Andrew Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.*, 128(14):1–13, 2008.

[22] Eric Darve, David Rodríguez-Gómez, and Andrew Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *The Journal of chemical physics*, 128(14):144120, 2008.

[23] Christoph Dellago, Peter Bolhuis, and Phillip L Geissler. Transition path sampling. *Advances in chemical physics*, 123(1), 2002.

[24] Bradley M. Dickson. Approaching a parameter-free metadynamics. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 84(3):1–4, 2011.

[25] Bradley M Dickson. Overfill protection and hyperdynamics in adaptively biased simulations. *Journal of chemical theory and computation*, 13(12):5925–5932, 2017.

[26] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Finite temperature string method for the study of rare events. *The Journal of Physical Chemistry B*, 109(14):6688–6693, 2005. PMID: 16851751.

[27] Bernd Ensing, Marco De Vivo, Zhiwei Liu, Preston Moore, and Michael L Klein. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Accounts of chemical research*, 39(2):73–81, 2006.

[28] Giacomo Fiorin, Michael L Klein, and Jerome Hénin. Using collective variables to drive molecular dynamics simulations. *Molecular Physics*, 111(22-23):3345–3362, 2013.

[29] Daan Frenkel and Berend Smit. Understanding molecular simulations: from algorithms to applications. Technical report, Academic Press, 2002.

[30] Haohao Fu, Xueguang Shao, Christophe Chipot, and Wensheng Cai. Extended adaptive biasing force algorithm. an on-the-fly implementation for accurate free-energy calculations. *Journal of Chemical Theory and Computation*, 12(8):3506–3513, 2016.

[31] Haohao Fu, Xueguang Shao, Christophe Chipot, and Wensheng Cai. Extended Adaptive Biasing Force Algorithm. An On-the-Fly Implementation for Accurate Free-Energy Calculations. *J. Chem. Theory Comput.*, 12(8):3506–3513, 2016.

[32] Haohao Fu, Hong Zhang, Haochuan Chen, Xueguang Shao, Christophe Chipot, and Wensheng Cai. Zooming across the free-energy landscape: Shaving barriers, and flooding valleys. *The journal of physical chemistry letters*, 9(16):4738–4745, 2018.

[33] T. Gadt, N. S. Ieong, G. Cambridge, M. A. Winnik, and I. Manners. Complex and hierarchical micelle architectures from diblock copolymers using living, crystallization-driven polymerizations. *Nature Materials*, 8(2):144–50, 2009.

[34] Raimondas Galvelis and Yuji Sugita. Neural Network and Nearest Neighbor Algorithms for Enhancing Sampling of Molecular Dynamics. *J. Chem. Theory Comput.*, 13(6):2489–2500, 2017.

[35] Federico Giberti, Matteo Salvalaglio, Marco Mazzotti, and Michele Parrinello. 3bry nucleation: from dimers to needle-like clusters. *Crystal Growth & Design*, 2017.

[36] Federico Giberti, Matteo Salvalaglio, and Michele Parrinello. Metadynamics studies of crystal nucleation. *IUCrJ*, 2(2):256–266, 2015.

[37] J. B. Gilroy, T. Gadt, G. R. Whittell, L. Chabanne, J. M. Mitchels, R. M. Richardson, M. A. Winnik, and I. Manners. Monodisperse cylindrical micelles by crystallization-driven living self-assembly. *Nature Chemistry*, 2(7):566–70, 2010.

[38] Lidija Glavas, Peter Olsén, Karin Odelius, and Ann-Christine Albertsson. Achieving micelle control through core crystallinity. *Biomacromolecules*, 14(11):4150–4156, 2013.

[39] Ciro A Guido, Fabio Pietrucci, Gregoire A Gallet, and Wanda Andreoni. The fate of a zwitterion in water from ab initio molecular dynamics: Monoethanolamine (mea)-co2. *Journal of chemical theory and computation*, 9(1):28–32, 2012.

[40] Ashley Guo, Emre Sevgen, Hythem Sidky, Jonathan Whitmer, and Juan De Pablo. Adaptive enhanced sampling with funn: Force-biasing using neural networks. *Bulletin of the American Physical Society*, 2018.

[41] Francois Gygi. Architecture of qbox: A scalable first-principles molecular dynamics code. *IBM Journal of Research and Development*, 52(1.2):137–144, 2008.

[42] Jürgen Hafner. Ab-initio simulations of materials using vasp: Density-functional theory and beyond. *Journal of computational chemistry*, 29(13):2044–2078, 2008.

[43] Julian Helfferich, Ivan Lyubimov, Daniel Reid, and Juan J de Pablo. Inherent structure energy is a good indicator of molecular mobility in glasses. *Soft matter*, 12(27):5898–5904, 2016.

[44] Graeme Henkelman and Hannes Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of chemical physics*, 113(22):9978–9985, 2000.

[45] Graeme Henkelman, Blas P Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of chemical physics*, 113(22):9901–9904, 2000.

[46] Joo Henriques and Marie Skep. Molecular dynamics simulations of intrinsically disordered proteins: On the accuracy of the tip4p-d water model and the representativeness of protein disorder models. *Journal of Chemical Theory and Computation*, 12(7):3407–3415, 2016. PMID: 27243806.

[47] Jeffrey A Herron, Yoshitada Morikawa, and Manos Mavrikakis. Ab initio molecular dynamics of solvation effects on reactivity at electrified interfaces. *Proceedings of the National Academy of Sciences*, page 201604590, 2016.

[48] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.

[49] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., and Bioinf.*, 65(3):712–725, 2006.

[50] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

[51] Koji Hukushima and Koji Nemoto. Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

[52] Jürg Hutter, Marcella Iannuzzi, Florian Schiffmann, and Joost VandeVondele. Cp2k: atomistic simulations of condensed matter systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(1):15–25, 2014.

[53] Zhang Jiang. Gixsgui: a matlab toolbox for grazing-incidence x-ray scattering data visualization and reduction, and indexing of buried three-dimensional periodic nanostructured films. *Journal of Applied Crystallography*, 48(3):917–926, 2015.

[54] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[55] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc*, 118(45):11225–11236, 1996.

[56] Young Kee Kang and Hae Sook Park. Assessment of ccsd (t), mp2, dft-d, cbs-qb3, and g4 (mp2) methods for conformational study of alanine and proline dipeptides. *Chemical Physics Letters*, 600:112–117, 2014.

[57] Kurt Kremer and Gary S Grest. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *The Journal of Chemical Physics*, 92(8):5057–5086, 1990.

[58] Yoshiyuki Kubota, Toshiharu Ohnuma, and Tomáš Bučko. Carbon dioxide capture in 2-aminoethanol aqueous solution from ab initio molecular dynamics simulations. *The Journal of Chemical Physics*, 146(9):094303, 2017.

[59] Shankar Kumar, Djamal Bouzida, Robert H. Swendsen, Peter A. Kollman, and John M. Rosenberg. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[60] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[61] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[62] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[63] Adrien Lesage, Tony Lelièvre, Gabriel Stoltz, and Jérôme Hénin. Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method. *The Journal of Physical Chemistry B*, 121(15):3676–3685, 2017.

[64] Adrien Lesage, Tony Lelievre, Gabriel Stoltz, and Jérôme Hénin. Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method. *The Journal of Physical Chemistry B*, 121(15):3676–3685, 2016.

[65] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164168, Jan 1944.

[66] Zhibo Li, Ellina Kesselman, Yeshayahu Talmon, Marc A Hillmyer, and Timothy P Lodge. Multicompartment micelles from abc miktoarm stars in water. *Science*, 306(5693):98–101, 2004.

[67] Alexander P. Lyubartsev and Alexander L. Rabinovich. Force field development for lipid membrane simulations. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1858(10):2483 – 2497, 2016.

[68] David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[69] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[70] Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[71] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H De Vries. The martini force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry B*, 111(27):7812–7824, 2007.

[72] Simone Marsili, Alessandro Barducci, Riccardo Chelli, Piero Procacci, and Vincenzo Schettino. Self-healing umbrella sampling: a non-equilibrium approach for quantitative free energy calculations. *The Journal of Physical Chemistry B*, 110(29):14011–14013, 2006.

[73] Roman Martoňák, Davide Donadio, Artem R Oganov, and Michele Parrinello. Crystal structure transformations in sio2 from classical and ab initio metadynamics. *Nature materials*, 5(8):623–626, 2006.

[74] Dominik Marx and Jürg Hutter. *Ab initio molecular dynamics: basic theory and advanced methods*. Cambridge University Press, 2009.

[75] Michael Mcgovern and Juan De Pablo. A boundary correction algorithm for metadynamics in multiple dimensions A boundary correction algorithm for metadynamics in multiple dimensions. 084102(2013), 2013.

[76] Yinglong Miao, Ferran Feixas, Changsun Eun, and J. Andrew McCammon. Accelerated molecular dynamics simulations of protein folding. *Journal of Computational Chemistry*, 36(20):1536–1549, 2015.

[77] M Mihailov, B Bogdanov, and G Davarska. X-ray diffraction analysis of high molecular poly (ethylene oxide) films moulded at different temperatures. *Acta polymerica*, 36(9):481–483, 1985.

[78] Letif Mones, Noam Bernstein, and Gábor Csányi. Exploration, sampling, and reconstruction of free energy surfaces with gaussian process regression. *Journal of Chemical Theory and Computation*, 12(10):5100–5110, 2016.

[79] Willem JM Mulder, Gustav J Strijkers, Geralda AF Van Tilborg, David P Cormode, Zahi A Fayad, and Klaas Nicolay. Nanoparticulate assemblies of amphiphiles and diagnostically active materials for multimodality imaging. *Accounts of Chemical Research*, 42(7):904–914, 2009.

[80] Gavin T Noble, Jared F Stefanick, Jonathan D Ashley, Tanyel Kiziltepe, and Basar Bilgicer. Ligand-targeted liposome design: challenges and fundamental considerations. *Trends in Biotechnology*, 32(1):32–45, 2014.

[81] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[82] Rachel K O'Reilly, Craig J Hawker, and Karen L Wooley. Cross-linked block copolymer micelles: functional nanostructures of great potential and versatility. *Chemical Society Reviews*, 35(11):1068–1083, 2006.

[83] Albert C Pan, Deniz Sezer, and Benoît Roux. Finding transition pathways using the string method with swarms of trajectories. *The journal of physical chemistry B*, 112(11):3432–3440, 2008.

[84] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

[85] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[86] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[87] Juan R Perilla, Boon Chong Goh, C Keith Cassidy, Bo Liu, Rafael C Bernardi, Till Rudack, Hang Yu, Zhe Wu, and Klaus Schulten. Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology*, 31:64 – 74, 2015.

[88] Tuan Anh Pham, Yuan Ping, and Giulia Galli. Modelling heterogeneous interfaces for solar water splitting. *Nature materials*, 2017.

[89] Stefano Piana and Alessandro Laio. A Bias-Exchange Approach to Protein Folding. *J. Phys. Chem. B*, 111(17):4553–4559, 2007.

[90] Fabio Pietrucci and Antonino Marco Saitta. Formamide reaction network in gas phase and solution via a unified theoretical approach: Toward a reconciliation of different prebiotic scenarios. *Proceedings of the National Academy of Sciences*, 112(49):15030–15035, 2015.

[91] Anaïs Pitto-Barry, Nigel Kirby, Andrew P Dove, and Rachel K O'Reilly. Expanding the scope of the crystallization-driven self-assembly of polylactide-containing polymers. *Polymer Chemistry*, 5(4):1427–1436, 2014.

[92] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comp. Phys.*, 117:1–19, 1995.

[93] Daniel R Reid, Ivan Lyubimov, MD Ediger, and Juan J De Pablo. Age and structure of a model vapour-deposited glass. *Nature communications*, 7, 2016.

[94] Ramon Roozendaal, Thorsten R Mempel, Lisa A Pitcher, Santiago F Gonzalez, Admar Verschoor, Reina E Mebius, Ulrich H von Andrian, and Michael C Carroll. Conduits mediate transport of low-molecular-weight antigen to lymph node follicles. *Immunity*, 30(2):264–276, 2009.

[95] Annette Rösler, Guido WM Vandermeulen, and Harm-Anton Klok. Advanced drug delivery devices via self-assembly of amphiphilic block copolymers. *Advanced Drug Delivery Reviews*, 53(1):95–108, 2001.

[96] Jean-Paul Ryckaert and André Bellemans. Molecular dynamics of liquid alkanes. *Faraday Discussions of the Chemical Society*, 66:95–106, 1978.

[97] Yuji Sasanuma, Hajime Ohta, Ikuko Touma, Hiroki Matoba, Yugo Hayashi, and Akira Kaito. Conformational characteristics of poly (ethylene sulfide) and poly (ethylene oxide): Solvent dependence of attractive and repulsive gauche effects. *Macromolecules*, 35(9):3748–3761, 2002.

[98] Elia Schneider, Luke Dai, Robert Q. Topper, Christof Drechsel-Grau, and Mark E. Tuckerman. Stochastic neural network approach for learning high-dimensional free energy surfaces. *Phys. Rev. Lett.*, 119:150601, Oct 2017.

[99] Hythem Sidky, Yamil J Colón, Julian Helfferich, Benjamin J Sikora, Cody Bezik, Weiwei Chu, Federico Giberti, Ashley Z Guo, Xikai Jiang, Joshua Lequieu, et al. Ssages: Software suite for advanced general ensemble simulations. *The Journal of chemical physics*, 148(4):044104, 2018.

[100] Hythem Sidky, Yamil J Colón, Julian Helfferich, Benjamin J Sikora, Cody Bezik, Weiwei Chu, Federico Giberti, Ashley Z Guo, Xikai Jiang, Joshua Lequieu, et al. Ssages: Software suite for advanced general ensemble simulations. *The Journal of chemical physics*, 148(4):044104, 2018.

[101] Hythem Sidky and Jonathan K Whitmer. Learning free energy landscapes using artificial neural networks. *arXiv:1712.02840*, 2017.

[102] Hythem Sidky and Jonathan K Whitmer. Learning free energy landscapes using artificial neural networks. *The Journal of chemical physics*, 148(10):104111, 2018.

[103] Hythem Sidky and Jonathan K Whitmer. Learning free energy landscapes using artificial neural networks. *The Journal of chemical physics*, 148(10):104111, 2018.

[104] Sadanand Singh, Chi cheng Chiu, and Juan J. de Pablo. Flux Tempered Metadynamics. *J. Stat. Phys.*, 145(4):932–945, 2011.

[105] Y Takahashi, H Tadokoro, and Y Chatani. Structure of polyethylene sulfide. *Journal of Macromolecular Science, Part B: Physics*, 2(2):361–367, 1968.

[106] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[107] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Chem.*, 23(2):187–199, 1977.

[108] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Bussi Giovanni. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.

[109] O. Valsson and M. Parrinello. A variational approach to enhanced sampling and free energy calculations. *Phys. Rev. Lett.*, 113(1), 2014.

[110] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, 2005.

[111] Eric Vanden-Eijnden and Maddalena Venturoli. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *The Journal of chemical physics*, 130(19):05B605, 2009.

[112] Diana Velluto, Davide Demurtas, and Jeffrey A Hubbell. Peg-b-pps diblock copolymer aggregates for hydrophobic drug solubilization and release: cyclosporin a as an example. *Molecular Pharmaceutics*, 5(4):632–642, 2008.

[113] Zhi-Xiang Wang and Yong Duan. Solvation effects on alanine dipeptide: A mp2/cc-pvtz//mp2/6-31g** study of (phi, psi) energy maps and conformers in the gas phase, ether, and water. *Journal of Computational Chemistry*, 25(14):1699–1716, 2004.

[114] Yi Wen and Joel H Collier. Supramolecular peptide vaccines: tuning adaptive immunity. *Current Opinion in Immunology*, 35:73–79, 2015.

[115] Jonathan K. Whitmer, Chi Cheng Chiu, Abhijeet A. Joshi, and Juan J. De Pablo. Basis function sampling: A new paradigm for material property computation. *Phys. Rev. Lett.*, 113(19):1–5, 2014.

[116] Jonathan K. Whitmer, Aaron M. Fluitt, Lucas Antony, Jian Qin, Michael McGovern, and Juan J. De Pablo. Sculpting bespoke mountains: Determining free energies with basis expansions. *J. Chem. Phys.*, 143(4), 2015.

[117] Fang Yuan, Shihu Wang, and Ronald G Larson. Potentials of mean force and escape times of surfactants from micelles and hydrophobic surfaces using molecular dynamics simulations. *Langmuir*, 31(4):1336–1343, 2015.