The Dissertation Committee for Eric Thomas Wright
certifies that this is the approved version of the following dissertation:

Bayesian Learning Methods for Potential Energy Parameter Inference in Coarse-Grained Models of Atomistic Systems

Committee:

Robert Moser, Supervisor

Peter Rossky, Co-supervisor

J. Tinsley Oden

Leszek Demkowicz

Serge Prudhomme

Ron Elber

Bayesian Learning Methods for Potential Energy Parameter Inference in Coarse-Grained Models of Atomistic Systems

ERIC THOMAS WRIGHT, B.S., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

Acknowledgments

I owe much gratitude to Peter Rossky for the energy he has put into advising and guiding me through this dissertation research. His finely tuned, intuitive sense for chemical physics has been immeasurably helpful. I also gratefully acknowledge the many productive discussions with J. Tinsley Oden and Kathryn Farrell that have informed much of the work presented in this document. Special thanks are due to Serge Prudhomme, who advised me during the first year of my dissertation work, the remainder of my dissertation committee, who have provided their time and valuable feedback to me, and the support staff and administration of the Institute for Computational Engineering and Sciences. In the latter category, I particularly acknowledge Stephanie Rodriguez, the CSEM graduate coordinator, for her patience and dedication to helping students. Finally, I thank my wife, Mariana, and my parents for their understanding and support throughout my time in graduate school.

Bayesian Learning Methods for Potential Energy Parameter Inference in Coarse-Grained Models of Atomistic Systems

Publication No. _____

Eric Thomas Wright, Ph.D.
The University of Texas at Austin, 2015

Supervisors: Robert Moser
Peter Rossky

The present work addresses issues related to the derivation of reduced models of atomistic systems, their statistical calibration, and their relation to atomistic models of materials. The reduced model, known in the chemical physics community as a coarse-grained model, is calibrated within a Bayesian framework. Particular attention is given to developing likelihood functions, assigning priors on coarse-grained model parameters, and using data from molecular dynamics representations of atomistic systems to calibrate coarse-grained models such that certain physically relevant atomistic observables are accurately reproduced. The developed Bayesian framework is then applied in three case studies of increasing complexity and practical application. A freely jointed chain model is considered first for illustrative purposes. The next example entails the construction of a coarse-grained model for a liquid

heptane system, with the explicit design goal of accurately predicting a vapor-liquid transfer free energy. Finally, a coarse-grained model is developed for an alkylthiophene polymer that has been shown to have practical use in certain types of photovoltaic cells. The development therein employs Bayesian decision theory to select an optimal CG potential energy function. Subsequently, this model is subjected to validation tests in a prediction scenario that is relevant to the performance of a polyalkylthiophene-based solar cell.

Table of Contents

Chapter 1

Introduction

As computational capabilities have increased, so too has the demand to simulate physical phenomena over a tremendous range of length and time scales. The atomic and molecular scale[1] is a particularly important application area for computer modeling and simulation due to the difficulties of experimental and analytical methods at the molecular level. The challenge for computer codes here is a great one as well. Ab initio quantum chemistry computations promise accuracy in the prediction of electronic structure for one or a few molecules, but their complexity restricts their application to small systems.

Molecular mechanics methods significantly expand simulation range by using parameterized potential energy functions in a classical mechanics framework. In this case, the dynamics of a molecular system consisting of N atoms are simulated by approximately solving the Newton equations,

$$m_i \ddot{\boldsymbol{r}}_i = -\nabla_i U, \ i = 1, \ldots, N, \tag{1.1}$$

where m_i and $\boldsymbol{r}_i$ are the mass and position vector of the ith atom and U is the potential energy function. It is also necessary to supply initial conditions

[1] Roughly speaking, length scales from one angstrom up to several nanometers

for atom positions and velocities. A popular approximation scheme for solving Newton's equations in molecular dynamics (MD) is the velocity Verlet integrator which is defined by the steps,

$$\begin{aligned} \boldsymbol{r}_i(t+\Delta t) &= \boldsymbol{r}_i(t) + \Delta t\, \dot{\boldsymbol{r}}_i(t) + \frac{(\Delta t)^2}{2m_i}\boldsymbol{F}_i(t) \\ \dot{\boldsymbol{r}}_i(t+\Delta t) &= \dot{\boldsymbol{r}}_i(t) + \frac{\Delta t}{2m_i}\left(\boldsymbol{F}_i(t) + \boldsymbol{F}_i(t+\Delta t)\right). \end{aligned} \tag{1.2}$$

Here, $\boldsymbol{F}_i(t)$ is the force acting on atom i at time t as determined by the relevant gradient of the potential energy function. The Δt quantity indicates the size of the timestep used and is often on the order of one femtosecond. This time scale is set in reference to covalent bond vibrational modes which are generally the highest frequency motions encountered in MD simulations.

We note here that the exact equations of motion conserve total energy, with the velocity Verlet scheme doing so in an approximate sense. It is often the case, however, that one is interested in the behavior of an atomistic system when thermodynamic properties other than total energy are constrained. Simulations at constant temperature or pressure are useful examples, as these constraints more closely resemble the empirical scenarios frequently studied. For a description of such scenarios, we appeal to the results of equilibrium statistical mechanics. Here, a molecular system is viewed stochastically, with the relative likelihood of atomic configurations and momenta given by probability distributions. A canonical example is found in the case of a temperature constraint on a system with a fixed volume and number of particles. In thermal equilibrium at temperature T, the probability density for the configurational

states is given by the Boltzmann formula,

$$\varrho(\boldsymbol{r}) = \frac{\exp\left\{-U(\boldsymbol{r})/k_b T\right\}}{\int \exp\left\{-U(\boldsymbol{r})/k_b T\right\}\, d\boldsymbol{r}}, \tag{1.3}$$

where $\boldsymbol{r}$ represents a configuration of position vectors for all particles in the system and k_b is Boltzmann's constant. The denominator in eq. 1.3 is generally referred to as a partition function. Observable, thermodynamic quantities are now expressed as averages over the statistical mechanical probability distribution,

$$\langle g \rangle = \int g(\boldsymbol{r})\varrho(\boldsymbol{r})\, d\boldsymbol{r}, \tag{1.4}$$

where g is a general property depending on configuration. It is possible to augment the Newton equations of motion to include constraints in such a way as to generate the associated statistical mechanical equilibrium distributions in the limit of long time[2]. Time-stepping schemes derived from the modified equations of motion are then understood, assuming sufficient decoupling from initial conditions, to provide samples that are approximately distributed according to Boltzmann-type probability rules. Thus, the quantity in eq. 1.4 is practically estimated via the Monte Carlo approximation,

$$\langle g \rangle \approx \frac{1}{N_s} \sum_{j=1}^{N_s} g\left(\boldsymbol{r}^{(j)}\right), \tag{1.5}$$

where the $\boldsymbol{r}^{(j)}$ are MD samples of configuration, ideally independent of each other, and N_s is the number of samples comprising the mean value calculation.

[2]We note here that the system must satisfy the ergodic hypothesis in order for this statement to be true. That is, long time averages must be equivalent to "spatial" averages over the relevant phase space.

From the preceding discussion, it is clear that the potential energy function, U, is of fundamental importance since it determines molecular dynamics, and equivalently when in equilibrium situations, the probability density for configurational states. In an atomistic model, the potential energy is a function of atom configuration as well as a set of free parameters that must be calibrated from experimental data, ab initio simulation, or some other ostensibly more accurate model. In order to simulate systems of millions or billions of atoms, the atomistic model can be reduced to a yet simpler model by grouping multiple atoms into single sites. This procedure is known as "coarse-graining" and the resulting reduced model is known as a "coarse-grained model." In this text, we restrict our study to coarse-grained (CG) models that are derived from a source atomistic model, termed the "all-atom model." The task we consider in depth is that of determining a potential energy function for a CG model when given a well-defined procedure for mapping all-atom configurations into the coarse representation. The identification of such a mapping provides a basis on which we may compare predictions of the CG model with those of the all-atom model.

In principle, if one is provided the all-atom potential energy, U_{AA}, and a mapping $\mathcal{M}$ that produces CG configurations from all-atom configurations, then one has a statistical mechanical expression (e.g. in the canonical ensemble) for the corresponding CG potential energy,

$$\exp\left\{-U_{CG}(\boldsymbol{R})/k_bT\right\} \propto \frac{1}{Z_{AA}} \int_{\Gamma_{AA}} \exp\left\{-U_{AA}(\boldsymbol{r})/k_bT\right\} \delta\left(\boldsymbol{R} - \mathcal{M}(\boldsymbol{r})\right)\, d\boldsymbol{r}, \tag{1.6}$$

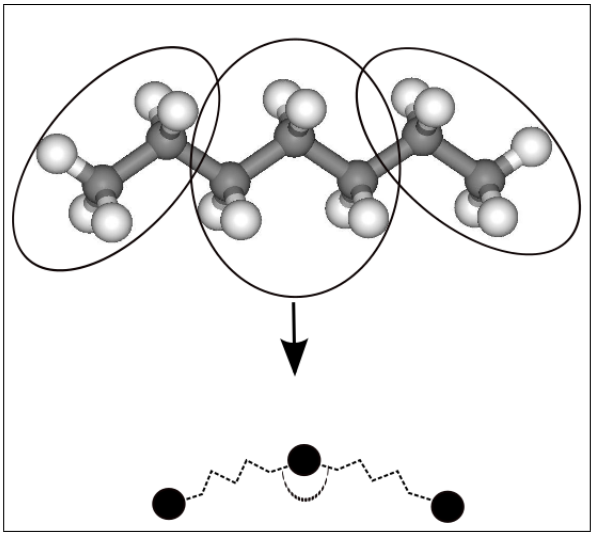

Figure 1.1: An example coarse-grained model of a heptane molecule. The top image shows the original all-atom system with CG groups identified by circles while the bottom image shows the resulting CG model

where $\boldsymbol{r}$ and $\boldsymbol{R}$ denote all-atom and CG configurations, respectively, Z_{AA} is the all-atom configurational partition function, and Γ_{AA} denotes the all-atom configurational phase space. The δ function notation indicates that the integration is over the portion of the all-atom configuration space in which $\mathcal{M}(\boldsymbol{r}) = \boldsymbol{R}$. Therefore, eq. 1.6 merely states that the CG canonical ensemble probability of state $\boldsymbol{R}$ should be proportional to the sum over the all-atom probabilities of states consistent with CG configuration. Denoting the right side of eq. 1.6 by $P_{AA}(\boldsymbol{R})$, we then have,

$$U_{CG}(\boldsymbol{R}) = -k_b T \log P_{AA}(\boldsymbol{R}) + \text{const.} \tag{1.7}$$

With U_{CG} in hand, molecular dynamics simulations can be undertaken using the reduced model. The coarse-grained potential defined in eq. 1.7 has the

property that its negative gradients with respect to CG particle positions are related to the mean all-atom forces acting between CG particles, where the mean is taken over the all-atom configurations consistent with R [72]. For this reason, U_{CG} is termed a *potential of mean force* (PMF). We return to this very important result in more detail later in the chapter. While the PMF is a convenient interpretation for the coarse-grained potential, eq. 1.7 is unfortunately a many bodied term that cannot be reasonably calculated for most systems of practical interest. Thus, many coarse-graining approaches seek to approximate this PMF with simpler, more easily computable terms. Several of these approaches are reviewed in the following section.

1.1 A Survey of Relevant Literature

1.1.1 All-Atom Models

We begin with a selection of popular all-atom models utilizing molecular mechanical potentials. Research related to the validation and calibration of atomistic "empirical energy functions" goes back several decades. Prominent projects in this area include CHARMM (Chemistry at HARvard Macromolecular Mechanics) [15], AMBER (Assisted Model Building and Energy Refinement) [79, 108], and OPLS (Optimized Potentials for Liquid Simulations) [45, 46], all of which consist of a potential energy functional form and a tabulated set of parameters for evaluating the potential energy in different molecular environments.

The energy formulas and parameter tables differ between CHARMM,

AMBER, and OPLS; however, the general character of the potential is common to all three. In all cases, the energy function is split into a sum of "bonded" and "non-bonded" parts. Interactions between atoms connected through a series of one, two, or three covalent bonds are contained in the bonded terms of the potential. The non-bonded portion contains interactions between atoms not connected through any series of covalent bonds and interactions between atoms connected by three or more covalent bonds. Non-bonded interactions are expressed as a sum of van der Waals dispersion and Coulomb-type electrostatic functional forms.

The potential energy parameters of CHARMM, AMBER, and OPLS are determined from a combination of experimental data and ab initio simulation. All three packages contain tables that give several values for each potential energy parameter depending on the molecular environment. An example of this environmental dependence is found in the CHARMM carbon-carbon equilibrium bond distance parameter. As this distance depends on the type of molecule the bond is in, CHARMM tabulates several values for it. For general alkanes (single bond between carbons), the distance is $1.53\,\mathring{A}$, while it is $1.34\,\mathring{A}$ for general alkenes (double bond between carbons) [106]. The carbon-carbon equilibrium distance is tabulated for many other specialized environments as well.

The described atomistic models are derived and calibrated in such a way as to be useful for a wide variety of molecular systems with different chemical compositions and thermodynamic states. In some cases, a model

specific to a particular set of molecules or chemical environment is desired. We encounter some particular examples of specialized atomistic models in the final chapters of this dissertation, but, in general, we utilize the 2005 revision of the OPLS potentials [6].

1.1.2 Coarse-Grained Models

Coarse-graining procedures have a relatively long history. For example, the Flory-Huggins lattice fluid model, which predicts thermodynamic properties of polymer solutions using a parameterized interaction between monomer and solvent units on a lattice, dates back to the early 1940s [27, 32]. Other relatively early examples of coarse-graining include the so-called "extended atoms" in the original versions of CHARMM and OPLS. Extended atom models simplify hydrocarbon representation by grouping carbon atoms and their attached hydrogen atoms into single sites [15, 45]. These models were originally published in the 1980s and the modern incarnations of CHARMM, AMBER, and OPLS still include extended atom models[3].

The modern notion of a general, parameterized CG model designed for use in a computer simulation is exemplified in a 1990 publication by Smit et al. [102]. In the aforementioned work, the authors present an off-lattice CG model of a water-oil interface that is commonly cited as the first of its kind for a lipid system. The Smit model is a prototype for much of the current work in CG modeling. In the remainder of the subsection, we review a selection of

[3]Although modern versions refer to these as "united atom models."

more contemporary CG models and calibration methods.

A general atomistic model calibration approach based on reproduction of mean forces from a higher resolution model is published in a 1994 paper by Ercolessi and Adams [24] and extended by Izvekov et al. in a 2004 publication [35]. These calibration methods are generally referred to as "force matching" methods. Force matching algorithms are developed for CG model calibration in a pair of papers by Izvekov and Voth from 2005 [36, 37]. In these works, the authors postulate parameterized CG potential energy functions and then calibrate them by choosing parameters that minimize the squared difference between all-atom force data and CG predicted forces at CG sites under comparable thermodynamic conditions. The authors and their collaborators refer to their calibration scheme as the "multiscale coarse-graining" (MS-CG) method. A wealth of literature details the extension and application of MS-CG to the simulation of various macro-molecular systems [33, 34, 38, 57, 96, 110]. Of particular interest is a 2008 publication in which Noid et al. show that the MS-CG method is part of a general theoretical framework for deriving CG models that are "physically consistent" with their source all-atom models [72]. Mullinax and Noid further derive an "extended ensemble" formalism for the MS-CG method in a publication from 2009 [69]. The extended ensemble framework enables CG model calibration that is, in principle, transferrable between systems with different atomistic topologies.

Another prominent CG model calibration method is the so-called "Boltzmann inversion" algorithm published by Reith et al. in 2003 [84]. The Boltz-

mann inversion procedure computes effective interaction potentials between CG sites using an iterative algorithm driven by the discrepancy between all-atom and CG pairwise potentials of mean force. The pairwise potential of mean force (PMF) is defined in statistical mechanics as the potential energy consistent with the average force along the lines connecting the centers of two fixed particles, or sites, where the average is taken over the ensemble of configurational states for the remaining $n-2$ particles. The PMF is closely related to another important quantity called the pair correlation function (PCF). In the statistical mechanical theory of materials, a PCF is a measure of the probability of observing a particle at a coordinate q relative to a given reference particle. One can show that the PMF, w, is related to the PCF, g, via

$$w(q) = -k_b T \log g(q), \tag{1.8}$$

where the variable q represents a generalized coordinate, such as a distance or an angle, that characterizes the pair interaction. The Boltzmann inversion algorithm begins with an initial guess for the CG interaction potential, $v_0(q)$, and then updates the potential at the ith step according to,

$$v_{i+1}(q) = v_i(q) - k_b T \log \left(\frac{g_i(q)}{g(q)} \right), \tag{1.9}$$

where g is the PCF predicted by simulating the all-atom system and g_i is the PCF that results from simulating the CG system with interaction potential v_i. The potential update is thus given by the difference between the all-atom and CG PMFs for the targeted interaction. Iteration terminates when the difference, $v_{i+1} - v_i$, is smaller than some specified tolerance. It

follows that, at convergence, the targeted interaction in the CG system approximately matches the all-atom PMF, and therefore, also the PCF predicted by the all-atom model. PCFs approximately determine some of the thermodynamic properties of a system[4], thus it can be desirable to reproduce them in CG models. There are a variety of applications in the literature that utilize Boltzmann inversion for CG model calibration [8, 44, 47, 88, 100]. Chapter 13 of McQuarrie's statistical mechanics text provides a readable background on pair correlation functions and potentials of mean force in the context of liquid systems [66].

Other notable methods for systematically determining effective CG interaction potentials include the reverse Monte-Carlo (RMC) method and the conditional reversible work (CRW) method. RMC methods are developed in a series of papers starting in 1988 [60, 61, 65] and applied to CG model calibration in a 2003 paper by Lyubartsev et al. [59]. The RMC method in the latter publication iteratively adjusts an effective interaction potential between CG sites until the PCF associated with the targeted interaction, which is predicted by a Monte-Carlo simulation of the CG system, converges to that predicted by the all-atom model. In contrast, the more recently developed CRW method [13, 14] non-iteratively determines a CG interaction potential between a pair of CG sites by computing the reversible work, in the all-atom system, associated with introducing interactions between the atomic constituents of the

[4]The pair correlation function exactly determines many system thermodynamic properties in the case of a pairwise additive potential energy function.

CG sites. CRW is essentially a PMF framework, as the reversible work done by introducing new interactions is related to the difference in PMF between a system in which the interactions between the CG sites of interest are excluded and one in which these interactions are included. The authors of CRW argue that the CG potentials produced by their method are transferable between different chemical and thermodynamic environments because their potentials are based on the free energies of effective pair interactions in the all-atom setting. However, we note that the authors assume a pairwise, additive CG potential, thus limiting the utility of the method.

Another general CG method of contemporary interest is M. S. Shell's relative entropy minimization method. Shell's 2008 publication introduces his method in the context of a target system and model system [93], where the model system is some reduced representation of the target system. The author notes the importance of reproducing the statistical ensemble of the target. Shell argues that the model optimally represents the target when the relative entropy, expressed as,

$$S_{rel} = \sum_i p_T(i) \log \frac{p_T(i)}{p_M(\mathcal{M}(i))} + S_{map}, \tag{1.10}$$

is minimized. In (1.10), $p(i)$ is the probability of configuration i in an ensemble, and T and M denote the target and model ensembles, $\mathcal{M}$ maps target structures into the model representation, and S_{map} is the so-called mapping entropy which is not dependent on the model probability. In the case that the target and model ensembles are canonical, the configuration probabilities have

simple forms related to Boltzmann factors. The author exploits the canonical structure to derive an optimality condition for a model potential energy, U_M, that is parameterized by a collection of adjustable values, $\{\lambda_i\}_{i=1}^k$. Specifically, the relative entropy expression in (1.10) is minimized when U_M satisfies, for all i,

$$\left\langle \frac{\partial U_M}{\partial \lambda_i} \right\rangle_M = \left\langle \frac{\partial U_M}{\partial \lambda_i} \right\rangle_T, \tag{1.11}$$

where $\langle \cdot \rangle_M$ denotes a canonical average in the model ensemble and $\langle \cdot \rangle_T$ is an average in the target ensemble. Shell then proposes a Newton-Raphson type iteration, coupled with MD simulation, to solve for the model parameter values that satisfy (1.11). In a later paper [17], Chaimovich and Shell suggest that the relative entropy framework provides a mechanism for reducing errors due to coarse-graining in observable quantities. If X is some observable that can be predicted from knowledge of the configuration in the all-atom and CG systems, then the authors suggest adding the term $aX(R)$, where a is a tunable parameter and R represents CG configuration, to the CG potential formulation. Then, due to (1.11), there is the optimality condition for X,

$$\langle X \rangle_{CG} = \langle X \rangle_{AA}. \tag{1.12}$$

Chaimovich and Shell note that more terms can be added to the CG potential energy to recover higher moments of the observable, X. In a 2012 publication, Carmichael and Shell extend the relative entropy minimization framework with a trajectory re-weighting scheme that accelerates the MD sampling procedure used to evaluate relative entropy gradients [16]. In the same paper, they apply the trajectory re-weighting algorithm to a set of CG models of a peptide,

where the CG potential energy function includes bonds described by harmonic springs and angle, torsion, and non-bonded terms that are represented by cubic splines. Thus, the calibration procedure determines the relative entropy-based optimal spring stiffnesses and lengths as well as the values of the other interactions at spline knots.

The methods reviewed so far are systematic attempts to calibrate general CG models of atomistic systems. Also of interest, however, are CG methods that employ a combination of expert intuition regarding the system of interest and systematic technique to produce CG models that reproduce specific, observable features of an atomistic model. A highly cited example of such is a CG model of a phospholipid and water system developed by Shelley et al. in a 2001 publication [94]. The authors first construct a CG model of water with the requirements that the model carry momentum in a way that is consistent with hydrodynamics, have the correct density, and have a liquid phase over the correct temperature range. They point out that their choice of model to meet the requirements is *not* unique, but nevertheless, will suit their purposes. To this end, they select a Lennard-Jones form to characterize the inter-molecular interactions between CG water particles and then calibrate the two Lennard-Jones parameters in a heuristic way that gives good agreement with experimental values of density and boiling temperature for water. In constructing the CG model of the phospholipid, the authors note that non-bonded interactions between the hydrophilic components of the molecule are crucial to obtaining the desired behavior. Thus, they take a systematic approach

to determining this set of interactions, using a Boltzmann inversion type of iteration scheme to derive the potentials for CG sites that are expected to be hydrophilic. More heuristic methods are used to model and calibrate the alkane chain parts of the phospholipid molecule. The authors and their collaborators derive similar CG models of lipid systems in subsequent publications [95, 97].

Another example of a set of application specific coarse-graining methods is found in the MARTINI project, which has attained relative prominence as a framework for constructing CG models of biomolecular systems. The original version of MARTINI, authored by Marrink et al. and published in a series of papers beginning in 2004 [62, 63], consists of a CG lipid model with specific CG potential energy formulas chosen by the authors and a tabulated set of parameters characterizing these potentials. The MARTINI model parameters are determined such that the resulting CG systems reproduce certain atomistic thermodynamic quantities, particularly oil/water partitioning free energies. The authors report using a trial and error procedure to select parameters such that the room temperature experimental densities, the mutual solubility of oil and water, and relative diffusion rates are reproduced. More recently, MARTINI has been extended to proteins [67] and carbohydrates [58]. As in the original MARTINI model, these models are calibrated primarily from atomistic partitioning free energies.

1.1.3 Model Calibration, Validation, and Uncertainty Quantification

A substantial part of the present work involves the application of standard results from the field of verification, validation, and uncertainty quantification (VV/UQ), particularly in the context of computer simulations in mechanics and physics. Much of the philosophy behind verification and validation is reviewed in a 2004 publication authored by Babuška and Oden [3]. Many of the technical terms common to the VV field are also defined in the aforementioned 2004 paper. The authors further describe model VV and illustrate its application with a few case study problems from solid mechanics and heat transfer in a 2005 publication [4]. The intertwined concepts of model validation and uncertainty quantification are reviewed in a pair of documents by Oden et al. in 2010 [75, 76]. In particular, these documents describe the quantification of uncertainty in a Bayesian framework. Bayesian ideas about probability and scientific inference are vital to the development of this dissertation document. We specifically cite the work of Edwin T. Jaynes as a strong philosophical influence on our application of Bayesian calibration and uncertainty quantification techniques [40, 42]. Further description of VV/UQ methods and philosophy can be found in [1, 74, 85].

1.2 Fundamental Theorems of Coarse-Graining

In this section, we review the consistency results, originally presented by Noid and Shell, that motivate the multiscale coarse-graining method/force matching and relative entropy minimization methods. The basic results therein are fundamental because they identify the theoretically optimal CG potential energy function, given a mapping from all-atom to CG configuration. In a mathematical sense, they give conditions for the almost everywhere equality of the statistical mechanical probability density functions associated with a CG representation and its source all-atom model.

We begin by defining the push-forward PDF of the all-atom statistical mechanical probability by the CG mapping function, $\mathcal{M}$,

$$\mathcal{M}_* \varrho_{AA}(\boldsymbol{R}) = \frac{1}{Z_{AA}} \int \exp\left\{-U_{AA}(\boldsymbol{r})/k_b T\right\} \delta\left(\boldsymbol{R} - \mathcal{M}(\boldsymbol{r})\right) \, d\boldsymbol{r}. \tag{1.13}$$

Intuitively, this quantity is the PDF on the configurational states of the CG representation that is implied by the all-atom potential energy function. Noid has noted that if a CG potential energy can be found such that $\varrho_{CG} = \mathcal{M}_* \varrho_{AA}$, then this potential is optimal, in a rigorous sense, because it results in a CG probability measure that reproduces all-atom measurements in the CG configuration space [72]. The basic theorem of Noid is the identification of this optimal potential with a potential of mean force arising from the CG mapping:

Theorem 1.2.1 (Noid)**.** *Let $\mathcal{M}$ be a center-of-mass mapping of the all-atom configurations into a CG representation such that each atom belongs to one,*

and only one, CG bead. Furthermore, let I label an arbitrary CG bead and also represent an index set over the atoms assigned to the bead. If ϱ_{CG} is the statistical mechanical PDF implied by the CG potential U_{CG}, then $\varrho_{CG} = \mathcal{M}_* \varrho_{AA}$ almost everywhere if and only if the CG potential energy satisfies,

$$-\nabla_I U_{CG}(\boldsymbol{R}) = \frac{\int \left(\sum_{i \in I} -\nabla_i U_{AA}(\boldsymbol{r})\right) \exp\left\{-U_{AA}(\boldsymbol{r})/k_b T\right\} \delta\left(\boldsymbol{R} - \mathcal{M}(\boldsymbol{r})\right) \, d\boldsymbol{r}}{\int \exp\left\{-U_{AA}(\boldsymbol{r})/k_b T\right\} \delta\left(\boldsymbol{R} - \mathcal{M}(\boldsymbol{r})\right) \, d\boldsymbol{r}}, \tag{1.14}$$

for all I and a.e. $\boldsymbol{R}$.

Proof. A version of this theorem, with more general assumptions on the CG mapping, is proved in detail in ref [72]. The crux of the proof is an integration-by-parts that transforms derivatives with respect to $\boldsymbol{R}$ into derivatives with respect to $\boldsymbol{r}$. □

This result shows that the optimal CG forces follow from an averaging procedure involving the net all-atom forces acting on each bead. For a given $\boldsymbol{R}$, this averaging happens over the portion of the all-atom configuration space that conserves the given CG configuration.

We now consider Shell's relative entropy minimization method and elucidate its connection to Noid's PMF. Shell's algorithm seeks to minimize the expression,

$$S_{rel} = \int \varrho_{AA}(\boldsymbol{r}) \log\left(\frac{\varrho_{AA}(\boldsymbol{r})}{\varrho_{CG}(\mathcal{M}(\boldsymbol{r}))}\right) \, d\boldsymbol{r} + \langle S_{map} \rangle_{AA}, \tag{1.15}$$

with respect to the CG statistical mechanical probability density, ϱ_{CG}. The quantity S_{map} is termed the "mapping entropy" and is a function of the all-

atom potential and the CG mapping, but not the CG potential. Its all-atom average is given by,

$$\langle S_{map} \rangle_{AA} = \int \varrho_{AA}(\boldsymbol{r}) \log \left\{ \int \mathbb{1}(\boldsymbol{r}') \, \delta \left(\mathcal{M}(\boldsymbol{r}') - \mathcal{M}(\boldsymbol{r}) \right) \, d\boldsymbol{r}' \right\} \, d\boldsymbol{r}, \tag{1.16}$$

where $\mathbb{1}$ simply denotes the indicator function on the whole all-atom configuration space. Thus, the quantity in the curly braces is related to the degeneracy of the CG map.

Theorem 1.2.2. *If a CG potential exists such that $S_{rel} = 0$, then $\varrho_{CG} = \mathcal{M}_* \varrho_{AA}$ almost everywhere.*

Proof. We first show that $\varrho_{CG} \circ \mathcal{M}$ can be renormalized by the argument of the mapping entropy to obtain a PDF on the all-atom configuration space. This is shown by a simple change of variables,

$$\begin{aligned} &\int \frac{\varrho_{CG}(\mathcal{M}(\boldsymbol{r}))}{\exp(S_{map}(\mathcal{M}(\boldsymbol{r})))} \, d\boldsymbol{r} \\ &= \int \frac{\varrho_{CG}(\boldsymbol{R})}{\exp(S_{map}(\boldsymbol{R}))} \left\{ \int \mathbb{1}(\boldsymbol{r}) \, \delta \left(\boldsymbol{R} - \mathcal{M}(\boldsymbol{r}) \right) \, d\boldsymbol{r} \right\} \, d\boldsymbol{R} \\ &= 1, \end{aligned} \tag{1.17}$$

where the quantity in curly braces is the differential volume element associated with the variable change. Thus, the mapping entropy factor cancels out and the original integrand is normalized. Let $\varrho_{CG}^{\mathcal{M}}(\boldsymbol{r}) = \varrho_{CG}(\mathcal{M}(\boldsymbol{r})) / \exp(S_{map}(\mathcal{M}(\boldsymbol{r})))$. It follows that,

$$\begin{aligned} S_{rel} &= \int \varrho_{AA}(\boldsymbol{r}) \log \left(\frac{\varrho_{AA}(\boldsymbol{r})}{\varrho_{CG}^{\mathcal{M}}(\boldsymbol{r})} \right) \, d\boldsymbol{r} \\ &\equiv D_{KL} \left(\varrho_{AA} \,\|\, \varrho_{CG}^{\mathcal{M}} \right), \end{aligned} \tag{1.18}$$

where D_{KL} denotes the Kullback-Leibler divergence [52]. Due to the Gibbs inequality[5], it follows that $S_{rel} = 0$ if and only if $\varrho_{AA} = \varrho_{CG}^{\mathcal{M}}$ in an almost everywhere sense. If the latter is true, then we can integrate over the all-atom space associated with an arbitrary CG configuration to obtain,

$$\begin{aligned} \mathcal{M}_* \varrho_{AA}(\boldsymbol{R}) &= \int \varrho_{CG}^{\mathcal{M}}(\boldsymbol{r})\, \delta\left(\boldsymbol{R} - \mathcal{M}(\boldsymbol{r})\right)\, d\boldsymbol{r} \\ &= \varrho_{CG}(\boldsymbol{R}). \end{aligned} \tag{1.19}$$

□

Corollary 1.2.3 (Relationship between Noid and Shell Optimal Potentials)**.** *Let $\mathcal{M}$ satisfy the same conditions as in thm. 1.2.1. If a CG potential exists such that $S_{rel} = 0$, then this potential is, up to an additive constant, the Noid PMF.*

It is interesting to note that the converse of corollary 1.2.3 is not generally true. That is, the Noid PMF as the CG potential does not necessarily imply $S_{rel} = 0$. The basic reason for this is that S_{rel} is equivalent to a Kullback-Leibler divergence in the all-atom configuration space between the all-atom PDF and a renormalized version of CG PDF that assigns the same probability density to every all-atom configuration belonging to the same pullback set, $\mathcal{M}^{-1}(\boldsymbol{R})$. This is shown in the proof of theorem 1.2.2. In order for this Kullback-Leibler divergence to be zero, the all-atom probability density would

[5]The Gibbs inequality is the result that $D_{KL}(\varrho_1 \| \varrho_2) \geq 0$, with equality if and only if the two PDFs are equal on sets with measure greater than zero. In order for the D_{KL} to be well-defined, it is also necessary to assume that if $\varrho_2 = 0$ on some set, then ϱ_1 vanishes on this set as well.

also have to make a constant assignment on these pullback sets. Thus, the Shell condition of $S_{rel} = 0$ is also a restriction on the all-atom probability density. It is important to observe, then, that Shell's globally optimal condition on the CG potential is not necessarily equivalent to that of Noid.

The preceding development begs the question: If the Noid PMF is the rigorously optimal CG potential, what is the use of any CG potential optimization technique that does not reference some kind of convergence to this PMF? The answer therein is related to practicality of actually computing the Noid PMF. Here, it is of vital importance to understand that the PMF is, in general, an M body potential when the CG system consists of M beads. The algorithms used by Noid and Shell are designed to converge in a subspace of the function space containing the true PMF. This approximation space often consists of additive potentials that are functions of a single, one dimensional generalized coordinate, due to practical computing limits. As the true PMF will not generally be in the span of the basis functions defining this approximation space, force-matching and relative entropy minimization in practice do not have control over the residual between their subspace optimal potentials and the rigorous PMF; that is, these methods do not determine the PMF to arbitrary precision. The implication of this statement is that these methods will not necessarily result in a CG potential that is able to reproduce a chosen observable of the all-atom model to the degree of precision set by the modeler. Thus, there is space within the field of CG potential optimization for methods that are practical in reference to the eventual prediction scenarios intended by

the designers and end-users of a particular CG model. The work presented in this document seeks to exist in that figurative space.

1.3 Document Overview

We propose a Bayesian framework for calibrating atomistic coarse-grained models that is driven by both observable information and principles of physical consistency. More specifically, the likelihood function of the Bayesian framework quantifies CG information about a specific set of observables relative to data gathered from an all-atom model. We intend these observables to be chosen according to expert intuition regarding the intended prediction scenarios of the CG model. The Bayesian prior information in our framework is determined from loose principles governing the physical consistency of the CG model with respect to the all-atom model.

The next chapter of this document describes theoretical aspects of applying Bayesian statistics to the coarse-graining problem. Of primary concern here is the assignment of likelihood functions and prior probability. The following chapter presents a case study involving CG calibration of a short chain molecule model. The final two chapters consider increasingly complicated scenarios where the Bayesian theory is applied, and augmented where necessary, to produce CG models that are able to predict quantities of practical interest in the realm of chemical physics. Chapter four considers a liquid heptane system wherein the eventual quantity of interest is a free energy of vapor-to-liquid trasfer. Finally, chapter five addresses a thiophene polymer that has

been shown to have practical use in the design of certain kinds of photovoltaic cells. Here, we construct a CG model that is designed to accurately predict structural quantities believed to be related to the efficiency of power generation in devices utilizing certain types of thiophene polymer.

Chapter 2

A Bayesian Theory of Coarse-Grained Model Calibration

In this chapter, we develop a Bayesian framework for calibrating a coarse-grained model of an atomistic system relative to a prescribed observable. The theory is developed in a general form with implementation details left to a later chapter. The major goal here is to describe the model calibration problem in an abstract setting and then to detail the ingredients of a Bayesian formulation, the prior and the likelihood function, as they apply to the CG model calibration procedure .

2.1 Model Calibration Concepts

We define an abstract mathematical model according to,

$$\mathcal{A}(\theta, \mathcal{S}, u(\theta, \mathcal{S})) = 0. \tag{2.1}$$

$\mathcal{A}(\cdot)$ is the collection of operators, constraints and conditions that define the mathematical form of the model. The set of parameters needed to characterize the model are contained in θ. These are the "degrees of freedom" of the model. We generally assume that the parameters can take on values in a subset of a real vector space, $\theta \in \Theta \subseteq \mathbb{R}^k$, for a model with k parameters. The scenario, $\mathcal{S}$,

is a description of the environment in which the model is expected to operate; it specifies the domains, initial/boundary conditions, and data source terms. The state variable, $u(\theta, \mathcal{S})$, is the solution of (2.1) for given parameters and scenario. As an example, consider a Lennard-Jones type atomistic system of identical particles in a microcanonical ensemble. The model formula, $\mathcal{A}$, consists of differential operators derived from the application of Newton's Second Law to each particle in the system. The model parameters are the Lennard-Jones constants, ϵ and σ. These constants enter the model equation through the Lennard-Jones potential energy form,

$$V(r; \epsilon, \sigma) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right].$$

The scenario, $\mathcal{S}$, contains the spatial and temporal domains, the number of particles, system total energy, and the initial positions and velocities of all the particles. The solution state variable, $u(\theta, \mathcal{S})$, gives the position and velocity vectors of each particle in the system as a function of time.

In the common case where the model parameter values are unknown, we must determine them through a calibration procedure. This entails finding parameter values such that predictions made by the model optimally "match" a set of observable data that is gathered independently of the model. We assume in this document that the observable data can be represented by a sequence of real numbers, $\{D_i\}_{i=1}^{n} \subset \mathbb{R}$. The D_i values correspond to a series of separate "observations" of the system of interest. Observable data could come from an empirical procedure, or in our case, from a different model that we trust to make accurate predictions of the observable. Generalizations can

be made to multi-dimensional observable data, but we examine the case of one-dimensional data for mathematical clarity. In order to compare the model with data, we must also assume that we can calculate observable values from the model through some functional of the solution state variable, $d[u(\theta, \mathcal{S})]$.

Procedures for choosing a "good" set of model parameters relative to a set of observable data abound. A well-known example since at least the time of Gauss and Laplace [2] is least squares minimization,

$$\{\hat{\theta}_{LS}\} \subseteq \left\{ \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \|D_i - d[u(\theta, \mathcal{S}_i)]\|^2 \right\}, \tag{2.2}$$

where we assume that the system of interest has a set of independent variables that we control in the observation setting and in the model through the scenario specification. Thus, i is an index for the different values of these controlled variables. We use set-inclusion notation to indicate that the least squares problem may not have a unique solution, or any solution. Determining $\hat{\theta}$ through (2.2) is potentially a challenging problem, as each evaluation of $d[u]$ at a specific scenario and parameter set requires a solution of (2.1). Despite the potential practical difficulty, the least squares solution gives, under some conditions, an optimal match to the data when the additive differences between observation and model, $D_i - d[u(\theta, \mathcal{S}_i)]$, are represented by uncorrelated random variables with equal variances and zero mean. The Gauss-Markov theorem contains the conditions and proof of this optimality when the model has a linear dependence on the parameters [48].

Least squares minimization is related to the more general method of maximum likelihood estimation (MLE). In a MLE method, one derives or

proposes a probability density, $\varrho\,(D\,|\,\theta)$, of obtaining a set of observed data from the model with a given set of parameters, θ. Here, the symbol D refers to a conjunction of n data samples. When this probability is regarded as a function of the parameter vector with fixed observed data, it is called a "likelihood function." Optimal parameter vectors are then those that maximize the likelihood function, given the observed data:

$$\{\hat{\theta}_{MLE}\} \subseteq \left\{ \underset{\theta \in \Theta}{\operatorname{argmax}}\ \varrho\,(D\,|\,\theta) \right\}. \tag{2.3}$$

There is much literature on MLE methods, but the modern understanding of the subject essentially begins with the publications of Sir R.A. Fisher. In his seminal 1922 paper [26], Fisher defines the term likelihood:

> The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of all observation should be that observed.

Given Fisher's definition, the likelihood function is determined for the simple probabilistic models of coin tossing, dice rolling, and urn sampling. In these cases, the likelihood involves a binomial or multinomial probability distribution regarded as a function of the long-running frequencies of certain event occurrences. When the model under consideration is deterministic, additional information is necessary to define a likelihood function. A common approach is to postulate that the discrepancy between the model and data is given by a

random variable with some generic properties, also known as a "noise model" [18]. When the discrepancy is assumed to be additive and distributed according to a gaussian with zero mean, the likelihood function takes the form,

$$\varrho_{like}\left(D_i \,|\, \theta\right) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{D_i - d[u(\theta, \mathcal{S}_i)]}{\sigma_i}\right)^2\right]. \tag{2.4}$$

If the D_i are regarded as independent samples and σ_i is the associated uncertainty, then the MLE from gaussian noise coincides with the least squares minimizer.

We note, however, that the likelihood formalism contains, not just information on optimal parameter values in reference to data, but also a representation of uncertainty in the data model. This observation motivates the construction of intervals in the parameter space which give a set of bounds containing the "true" parameter value to a some preset level of certainty. When quantification of uncertainty is important, parameter estimation in a maximum likelihood setting entails (1) the selection of a "sufficient statistic," which is a function of the sampled data and is maximally informative in relation to the unknown parameter, (2) deriving the sampling distribution associated with the chosen statistic, and (3) computing the MLE and possibly constructing confidence intervals in accordance with the sampling PDF which indicate parameter uncertainty at a chosen confidence level. For a more detailed development of sufficient statistics, we refer to Fisher's 1922 work; the concept of confidence interval estimation was developed in a seminal work by J. Neyman [70].

Maximum likelihood estimation procedures are data driven in the sense

that the optimal parameters and confidence intervals therein are determined by the form of the noise model and the observed data sets. If there is information that is relevant to the calibration of the model parameters, but not contained in the data, MLE does not take this cogent information into account. It is often advantageous to broaden the scope of the calibration and include so-called prior information that supplements the probability model of the data by itself. Bayesian methods, bringing a view of probability theory as a logical system of inference, are built to address this issue and are now introduced for our purposes.

The informative quantity in the Bayesian case is a probability distribution on the parameter space that is conditioned on two logical propositions: (1) a set of data regarding the observables characterizing the calibration and (2) so-called "prior information" which is information relevant to the model parameters that is not contained in the data. The prior information could take the form of a simple logical constraint on the parameters. For instance, it may be known from a previous data set, from a theory, or from a calculation that only a portion of the full parameter space produces a model that is physically meaningful in the given scenario. The PDF conditioned on the data and on this prior information is generally called a "posterior probability"; the reason for this terminology is explained shortly.

Before quantitatively defining the posterior, we note that the definition of probability in the Bayesian case is different from the traditional "frequentist" definition. Here we interpret probability as a degree of belief in the

truth of a proposition, as opposed to being a limiting frequency of an infinite sequence of random variable realizations. Probability has meaning in the Bayesian framework not because the proposition under evaluation is random, but because one may not be in possession of enough knowledge to deductively determine the truth of the proposition. It is in this sense that one may ask about the probability that θ takes on a certain value.

The posterior distribution is given by Bayes' Formula,

$$\varrho_{post}\left(\theta \mid DI\right) = \frac{\varrho_{like}\left(D \mid \theta I\right) \varrho_{prior}(\theta \mid I)}{\varrho_{evidence}\left(D \mid I\right)}, \tag{2.5}$$

where the notation, $\varrho(\cdot \mid DI)$, refers to a probability density conditioned on the logical conjunction of D, an observed data set, and I, a set of prior information. The quantity denoted ϱ_{prior} contains quantitative knowledge about the parameters due to the prior information, $\varrho_{evidence}$ is a normalizing factor called the "evidence", and ϱ_{like} is the familiar Fisher sort of likelihood function. The posterior distribution represents an update of the prior information upon consideration of the data, thus its name. The evidence factor is also known as a "marginal likelihood" because it is constrained to be:

$$\varrho_{evidence}\left(D \mid I\right) = \int \varrho_{like}\left(D \mid \theta I\right) \varrho_{prior}(\theta \mid I)\, d\theta. \tag{2.6}$$

It is argued by Edwin Jaynes that (2.5) is a consequence of a system of axioms that constitute a logical theory of probability [42]. For the purposes of the present document, we simply note that Bayes' formula can be derived from a simple set of postulates and we use it without trepidation.

We find from experience that the all-atom MD simulations we perform contain a great deal of information about coarse-grained model parameters prior to undertaking CG system simulation. We heretofore operate in a Bayesian framework where we may take advantage of abundant all-atom information and compute representations of parameter uncertainty through posterior parameter distributions. The remainder of the chapter is devoted to specifying likelihood functions and prior probability distributions relevant to CG model calibration.

2.2 Prior Probability for CG parameters

In this section, we develop some formalism for finding prior probability distributions of free parameters found in coarse-grained potential energy functions. The philosophical and mathematical underpinnings for these results are stated thoroughly by Edwin Jaynes in his posthumously published book, *Probability Theory: The Logic of Science* [42]. Following Jaynes, our main device is the *principle of maximum entropy*. Here, entropy refers to the inherent uncertainty in a given probability distribution. To motivate this statement, we consider a finite sample space $\{x_1, \ldots, x_n\}$ that has associated probabilities $\{p_1, \ldots, p_n\}$. We now seek a function, $H(p_1, \ldots, p_n)$, that characterizes uncertainty in the distribution. Following Shannon [92], the following properties are proposed for H:

I. $H(p_1, \ldots, p_n)$ is a non-negative real number,

II. $H(p_1, \ldots, p_n)$ is a continuous function of the p_i,

III. $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) > H\left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$ when $n > m$,

IV. $H(q_1, \ldots, q_m, p_2, \ldots, p_n) = H(p_1, \ldots, p_n) + p_1 H\left(\frac{q_1}{p_1}, \ldots, \frac{q_m}{p_1}\right)$ for $p_1 = \sum_{i=1}^{m} q_i$.

The third property indicates that, for uniform probabilities, H increases monotonically with the number of outcomes. The fourth requirement states that if one considers a subdivision of possibilities, then the entropy over the extended sample space is the entropy of the original, undivided space plus the entropy over the subdivision weighted by its probability. Shannon goes on to show that,

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i, \tag{2.7}$$

is the unique function, to within a multiplicative constant, satisfying the four properties. We note the resemblance of this expression to the Gibbs entropy common to statistical mechanics.

The continuum analog to the Shannon entropy is found from a limiting procedure wherein the spacing between the discrete samples points is described by a continuous density,

$$\eta(x_i) \equiv \lim_{n \to \infty} \frac{1}{n(x_{i+1} - x_i)}. \tag{2.8}$$

Accordingly, the discrete probabilities, for large n, approach a probability density function, ϱ,

$$\begin{aligned} p_i \approx & \varrho(x_i)(x_{i+1} - x_i) \\ \approx & \frac{\varrho(x_i)}{n\eta(x_i)}. \end{aligned} \tag{2.9}$$

The continuous expression for the entropy thus becomes[1],

$$\mathcal{H}[\varrho] = -\int_\Omega \varrho(x) \log \frac{\varrho(x)}{\eta(x)}\, dx, \tag{2.10}$$

where Ω denotes the sample space. The limiting density, η, is known as the *invariant measure*. Since it transforms in the same manner as ϱ, it is clear that the presence of η ensures that the continuous entropy is invariant to coordinate changes. The particular form of η depends on the underlying sample space, as is evident from the described limiting procedure. For instance, if the sample space is defined on radial values in a three dimensional spherical coordinate system, then the invariant measure is the associated density, $4\pi r^2$. Thus, we obtain an apparatus for finding a maximally uncertain probability distribution that encapsulates the given prior information: maximize the quantity in eq. 2.10 while maintaining the constraints associated with the prior information.

A remaining question is what form the prior constraints should take. A useful result is known in the case of a mean value constraint on a maximum entropy distribution. We refer to the following conclusion, originally attributed

[1]Placing the expression in eq. 2.9 into the Shannon entropy formula results in a "$\log n$" term that diverges as $n \to \infty$. To remedy this, we define the continuous entropy, $\mathcal{H}$, according to the limit, $\mathcal{H} \equiv \lim_{n\to\infty}(H - \log n)$

to Gibbs in the context of equilibrium statistical mechanics [41]: The PDF of maximum entropy, ϱ, subject to the constraint $\int_\Omega f(x)\varrho(x)\,dx = F$ for some function f, is given by:

$$\varrho(x) = \frac{\eta(x)\exp\{\lambda f(x)\}}{Z(\lambda)}, \tag{2.11}$$

where $Z(\lambda) = \int_\Omega \eta(x)\exp\{\lambda f(x)\}\,dx$ and λ is a Lagrange multiplier found from,

$$\frac{\partial \log Z}{\partial \lambda} = F. \tag{2.12}$$

In analogy to the maximum entropy formulation of classical statistical mechanics, we hereafter use mean value constraints in concert with entropy maximization to translate prior information into probabilistic representation.

Considering the problem of parameterizing a CG model from a corresponding all-atom model, we argue that the fully specified all-atom model provides a source of prior information. Suppose, for example, that we observe from an all-atom model (via limited MD runs, or perhaps from inspection of equilibrium configurations) that the distance between two sites, site A and site B, on a particular molecule tends to be near 1.5 $\mathring{A}$. From here, it seems reasonable to hypothesize that a CG model having sites A and B as degrees of spatial freedom could have a "bond" between A and B and this bond could have an equilibrium distance on the order of 1.5 $\mathring{A}$. The logic underlying this assertion is that if our CG model is to be statistically consistent with the all-atom model, then it seems unlikely that an A-B bond length that differs by orders of magnitude from 1.5 $\mathring{A}$ would produce a trustworthy representation of the

all-atom model. This sort of heuristic argument certainly does not constitute a rigorous calibration; however, it does serve as a valid *initial guess* regarding the properties of the coarse-grained A-B bond. In short, this argument is just an application of common sense when we have some limited understanding of the all-atom model.

In order to properly compute a posterior distribution for a set of CG parameters, we require not only an initial guess, but a probability distribution built on this initial guess. In particular, we would like a prior probability distribution that encodes the information that the CG A-B bond is probably of order 1.5 $\mathring{A}$. Simultaneously, we want this prior to be as uncertain as possible as to the actual numerical value of this bond distance. Applying the techniques of maximum entropy to prior probability assignment allow us to satisfy both of these demands. The basic idea is to consider the class of all probability distributions on $[0, \infty)$ having a mean of 1.5 $\mathring{A}$ and then to choose the distribution from this class with maximal entropy. That is, we maximize the functional,

$$H[\varrho] = -\int_0^\infty \varrho(R_{eq}) \log \frac{\varrho(R_{eq})}{4\pi R_{eq}^2}\, dR_{eq}, \tag{2.13}$$

over all ϱ such that $\langle R_{eq} \rangle_\varrho = 1.5\,\mathring{A}$. Since the integral in eq. 2.13 is over a distance in three dimensional space, the invariant measure includes a factor of $4\pi R_{eq}^2$. The distribution of maximal entropy then has the form,

$$\varrho(R_{eq}) = \frac{4\pi R_{eq}^2}{Z(\lambda)} e^{\lambda R_{eq}}, \tag{2.14}$$

in which,

$$Z(\lambda) = \int_0^\infty \exp(\lambda R_{eq})\, 4\pi R_{eq}^2\, dR_{eq}, \tag{2.15}$$

and λ is determined by solving the equation,

$$\langle R_{eq} \rangle_\varrho = \int_0^\infty R_{eq} \left(\frac{\exp(\lambda R_{eq})}{Z(\lambda)} \right) 4\pi R_{eq}^2\, dR_{eq} = R_{eq}^*, \tag{2.16}$$

where R_{eq}^* is a general constraint value on the mean (e.g. $R_{eq}^* = 1.5\,\mathring{A}$). Conveniently, a closed form solution for $\varrho(R_{eq})$ can be found in this case by carrying out the prescribed integrations,

$$\varrho(R_{eq}) = \frac{R_{eq}^2}{2} \left(\frac{3}{R_{eq}^*} \right)^3 \exp\left(-\frac{3R_{eq}}{R_{eq}^*} \right). \tag{2.17}$$

Thus, we find that $\varrho(R_{eq})$ is the PDF for a gamma distribution[2] with shape parameter $k = 3$ and scale parameter $\theta = R_{eq}^*/3$.

We may also be able to "guess" the spring constant of the CG A-B bond, in the case where we assign this CG interaction the form $K_r(R - R_{eq})^2$. According to the equipartition theorem of classical statistical mechanics, quadratic terms in the Hamiltonian contribute $\frac{1}{2}k_BT$ to the total average energy. So, we can loosely say that,

$$K_r \approx \frac{k_BT}{2\langle (R - R_{eq})^2 \rangle}. \tag{2.18}$$

The quantity $\langle (R - R_{eq})^2 \rangle$ is the second moment of the bond distance about its equilibrium value. In order to gain information about the order of magnitude

[2]The PDF of a gamma distribution can be characterized in terms of parameters k and θ with the functional form, $\varrho(x) = \frac{1}{\Gamma(k)\theta^k}\, x^{k-1} \exp\left(-\frac{x}{\theta}\right)$, where $\Gamma(\cdot)$ denotes the gamma function.

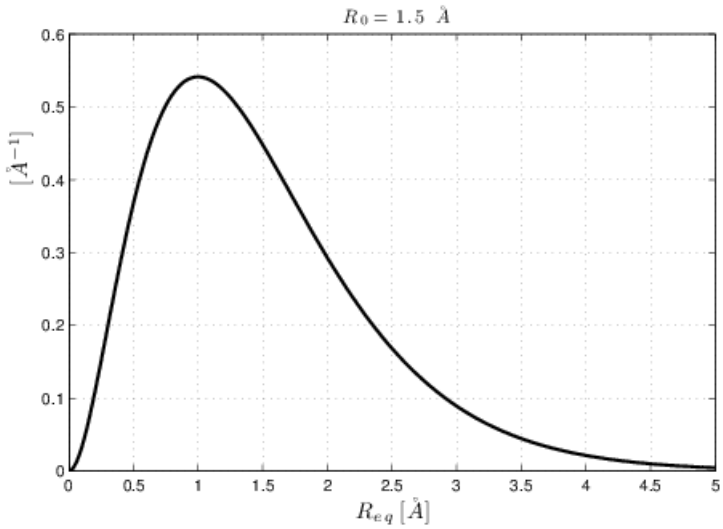

Figure 2.1: Maximum entropy prior PDF for bond equilibrium distance with prior information, $R^*_{eq} = 1.5\,\mathring{A}$

of K_r, we thus estimate the variance, σ^2_{AB}, of the A-B distance from an all-atom MD simulation. Applying the maximum entropy framework as in the case of the equilibrium distance yields the prior,

$$\varrho(K_r) = \frac{1}{K^*_r} \exp\left(-\frac{K_r}{K^*_r}\right), \tag{2.19}$$

where $K^*_r = k_B T / 2\sigma^2_{AB}$. This procedure may be repeated to find a maximum entropy prior for any force constant which multiplies a quadratic term in the CG Hamiltonian.

In analogy with the A-B site distance, we may have information from the all-atom model concerning the angle formed between sites A, B, and C. Suppose, for instance that the cosine of this angle tends to be near some number $C^* \in [-1, 1]$. Again utilizing the maximum entropy framework, we

find a prior for the A-B-C equilibrium angle, θ_0,

$$\varrho(\theta_0) = \frac{\sin(\theta_0)}{Z_\theta(\lambda_\theta)} \exp\left(\lambda_\theta \cos(\theta_0)\right), \tag{2.20}$$

with,

$$Z_\theta(\lambda_\theta) = \int_0^\pi \exp\left(\lambda_\theta \cos(\theta_0)\right) \sin(\theta_0)\, d\theta_0, \tag{2.21}$$

and λ_θ determined by,

$$\frac{1}{Z_\theta} \frac{dZ_\theta}{d\lambda_\theta} = C^*. \tag{2.22}$$

There is no closed form in this case, but we do find that

$$Z_\theta = \frac{2 \sinh \lambda_\theta}{\lambda_\theta}, \tag{2.23}$$

and λ_θ is a solution of the transcendental equation,

$$\coth \lambda_\theta - \frac{1}{\lambda_\theta} = C^*. \tag{2.24}$$

The results described in (2.14) - (2.20) are specific cases of the general result that the maximum entropy distribution for a variable x with k mean value constraints $\langle f_i(x) \rangle = F_i$, $i = 1, \ldots, k$ is given by,

$$\varrho(x) = \frac{\eta(x)}{Z(\lambda_1, \ldots, \lambda_k)} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x) \right), \tag{2.25}$$

where Z and $\{\lambda_i\}_{i=1}^k$ are found by satisfying normalization and mean value constraints.

In order to find prior probabilities for a Lennard-Jones type nonbonded interaction, we borrow the iterative Boltzmann inversion procedure for finding an initial guess for the pairwise PMF. That is, if the prescribed CG nonbonded

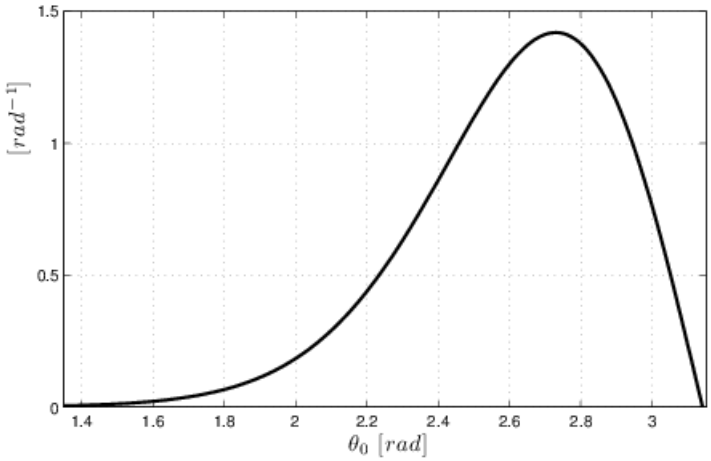

Figure 2.2: Maximum entropy distribution for an angle parameter with constraint, $\langle \cos\theta_0 \rangle = -0.8$

interaction between particles of type B has a Lennard-Jones functional form, we can compute a radial distribution function, $g_{BB}(r)$ from all-atom MD and then generate the trial PMF,

$$U^*_{BB}(r) = -k_b T \log g_{BB}(r). \tag{2.26}$$

If this trial PMF has a well defined minimum value at $r = r^*_{BB}$, then we may regard r^*_{BB} and $U^*_{BB}(r^*_{BB})$ as prior information regarding the Lennard-Jones parameters σ and ϵ (with appropriate proportionality constants applied that depend on the exponents employed in the Lennard-Jones form). Prior probability distributions for these parameters can then be obtained from applying the maximum entropy formalism, regarding the trial estimates of Lennard-Jones σ and ϵ as mean value constraints on the distributions.

The theme of this section is the use of the maximum entropy framework

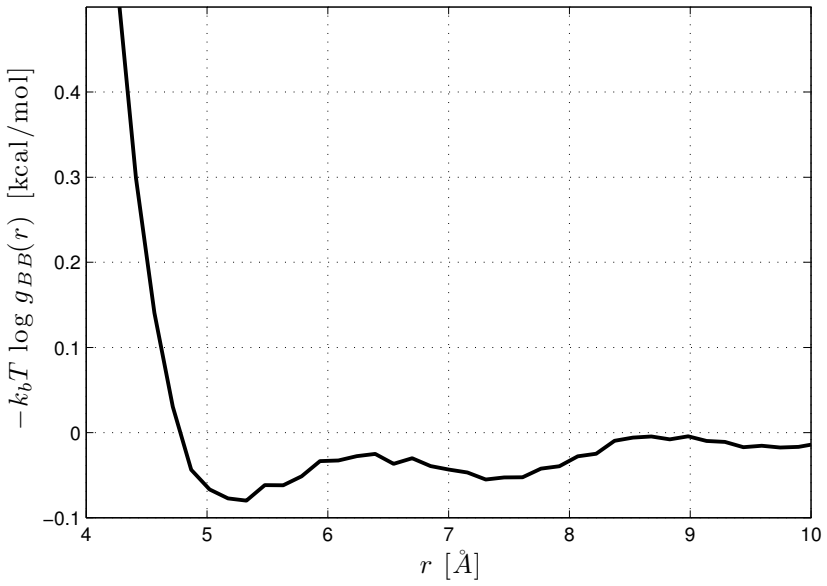

Figure 2.3: An example of a trial PMF derived from a radial distribution function. The RDF is estimated from all-atom MD simulation at $T = 300\,K$. In this case, $r^*_{BB} = 5.3\,\AA$ and $U^*_{BB}(r^*_{BB}) = -0.08$ kcal/mol.

to find prior probability distributions from simple mean value constraints. The constraints themselves are regarded as the cogent prior information to be encoded into the prior probabilities. We have shown several examples of finding maximum entropy priors for parameters one commonly finds in molecular mechanical type potential energy functions. We expect that similar techniques can be applied to prior probabilities for different types of potential energy parameters due to the generality of the maximum entropy formalism.

2.3 Calibration Observables and Likelihood Functions

To complete the picture for full Bayesian calibration of a CG model, we need to specify the data against which the model will be calibrated, as well as the mechanism for comparison of CG and all-atom predictions regarding the data. The former we will refer to as *calibration observables*, while the latter is accomplished by finding a suitable likelihood function for the calibration observables.

The nomenclature of *calibration observable* is chosen in order to delineate those observables determining the model parameters in a calibration setting from other quantities that may be used to test the *validity* of the calibrated CG model. We adopt the view here that the CG model is designed with the goal of computing a set of quantities of interest (QOIs). More specifically, a CG model is valid (or, more properly, *not invalid*) when it is able to accurately predict these QOIs, where the accuracy is relative to all-atom predictions. Therefore, we anticipate that choosing calibration observables that are correlated with, but are not the same as, the QOIs is vital to the problem of CG model validation. Of course, one may always increase the scope of CG model validation tests beyond the initial set of QOIs and in doing so may invalidate the CG model. The process of choosing calibration observables is thus an iterative one that requires feedback from the predictive capacity of the CG model with respect to QOIs. We also imagine that some cases of invalidation necessitate a refinement of the CG model itself (i.e. changing the mapping, potential energy function form, etc.). Although an inevitable direction for CG

research, this latter case is beyond the immediate scope of the present work.

We now address the issue of comparing all-atom and CG calibration observables through a likelihood function. Presently, we treat the case of just one calibration observable taking values on the real line. We further assume that the calibration observable can be expressed as the ensemble average of a function of statistical mechanical microstate in both the all-atom and CG models. The central question that must be answered is the following: What is the probability that the CG model could have generated a set of calibration observable values sampled from the all-atom model? In order to quantify the statistical properties of the calibration observable from the CG model, let f_{CG} be the function characterizing the observable and assume $\{\omega_{i,CG}\}_{i=1}^{n}$ are sampled microstates. An estimate of $\langle f_{CG} \rangle$ in the CG model is given by the sample mean,

$$S_{f,CG}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} f_{CG}\left(\omega_{i,CG}\right). \tag{2.27}$$

If n is "large enough," then the random variable realized by $S_{f,CG}^{(n)}$ has an approximately gaussian distribution about the true mean, $\mu_{f,CG}$, regardless of the exact form of the f_{CG} distribution function. This statement is a straightforward application of the central limit theorem and yields the probability distribution,

$$\varrho(S_{f,CG}^{(n)} \,|\, \mu_{f,CG}, \sigma_{f,CG}^2) = \sqrt{\frac{n}{2\pi\sigma_{f,CG}^2}} \exp\left(-\frac{n}{2} \frac{\left(S_{f,CG}^{(n)} - \mu_{f,CG}\right)^2}{\sigma_{f,CG}^2} \right), \tag{2.28}$$

where $\sigma_{f,CG}^2$ is the true variance of the f_{CG} random variable. If a sample mean is instead computed from *all-atom* microstate samples, $\omega_{i,AA}$, through

the relation,

$$S_{f,AA}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} f_{AA}\left(\omega_{i,AA}\right), \tag{2.29}$$

then $p(S_{f,AA}^{(n)} \,|\, \mu_{f,CG}, \sigma_{f,CG}^2)$ measures the probability that the all-atom sample mean, $S_{f,AA}^{(n)}$, is consistent with sample mean predictions made by the CG model. Thus, if m independent sample means are taken from the all-atom model, we obtain the likelihood function,

$$\varrho_{like}(D \,|\, \Theta, \mu_{f,CG}, \sigma_{f,CG}^2) = \left(\frac{n}{2\pi\sigma_{f,CG}^2}\right)^{m/2} \exp\left(-\frac{n}{2}\frac{\sum_{j=1}^{m}\left(S_{f,AA,j}^{(n)} - \mu_{f,CG}\right)^2}{\sigma_{f,CG}^2}\right), \tag{2.30}$$

where D represents the conjunction of the m all-atom sample means and Θ is the CG potential energy parameter set upon which the statistical properties of the CG model predictions are implicitly conditioned.

One issue that must be addressed is that the likelihood function in (2.30) is conditioned on $\mu_{f,CG}$ and $\sigma_{f,CG}^2$. Generally, one does not know these values exactly and must estimate them from CG model samples for a particular Θ. If these estimates have significant uncertainty, then this uncertainty should be encoded into the likelihood function. In short, one can include these "hyperparameters" in the Bayesian calibration (necessarily providing prior probabilities on $\mu_{f,CG}$ and $\sigma_{f,CG}^2$) and then integrate the posterior over $\mu_{f,CG}$ and $\sigma_{f,CG}^2$ to obtain the desired marginal posterior. In the examples we consider in the following chapter, $\mu_{f,CG}$ and $\sigma_{f,CG}^2$ can be estimated to very high precision (less than 1% error), so uncertainty associated with their estimation is not included in the calibration. Nevertheless, considerations of

hyperparameter uncertainty should generally not be ignored and are addressed in chapters four and five.

Chapter 3

Calibration of a Coarse-Grained Model of a Chain Molecule

3.1 The Freely Jointed Chain

The goal of this chapter is to implement the coarse-graining theory in a simple, illustrative model. To that end, we produce a CG model of a chain molecule that closely resembles a freely jointed chain (FJC). The close analogy with a freely jointed chain is desirable as a case study due to the latter's simple physical and statistical properties and because of its ubiquity as an idealized model for a general polymer chain. We subsequently describe the FJC as well as our slightly modified version.

Geometrically, a freely jointed chain is represented by a sequence of points in euclidean space in which neighboring points are constrained to be a set distance apart.

Definition 3.1.1 (**The Freely Jointed Chain**)**.** A freely jointed chain with N bonds of length l is the set, $\left\{\{r_i\}_{i=0}^{N} \subset \mathbb{R}^3 \colon \|r_i - r_{i+1}\| = l,\ i \in \{0, \ldots, N-1\}\right\}$. A *realization* of the FJC is any element of this set.

A realization of a chain can be generated by an N step random walk in $\mathbb{R}^3$, where r_0 is chosen randomly and each subsequent step must have length l.

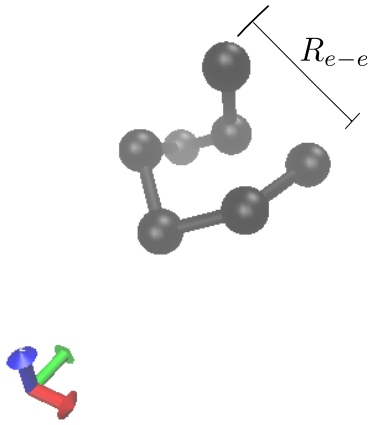

Figure 3.1: Freely jointed chain realization with $N = 6$, $l = 1.0\,\mathring{A}$

The "freely jointed" descriptor in the name comes about because there are no constraints imposed on angles formed by triples of neighboring points.

We study the freely jointed chain as a statistical mechanical ensemble by postulating that all realizations generated by fixed N and l are equally likely to occur. In this context, a property of interest for the chain is the end-to-end distance, $R_{e-e} = \|r_N - r_0\|$. The end-to-end distance thus labels macrostates of the FJC ensemble and there is a probability density associated with each end-to-end distance value. Conveniently, an exact formula for this probability distribution exists and is well known [9, 43]. We find from the book of Boyd and Phillips,

$$\rho_{N,l}(R_{e-e}) = \frac{2}{\pi}\int_0^\infty \left(\frac{\sin(ls)}{ls}\right)^N (sR_{e-e})\sin(sR_{e-e})\,ds, \tag{3.1}$$

where $\rho_{N,l}$ is the probability density. Intuition suggests that the probability density is low near $R_{e-e} = 0$ and $R_{e-e} = Nl$. This is because there are "few" chain realizations that are near full extension or curled back completely on themselves. Indeed, when $N > 2$, this is the case. Figure 3.2 displays the

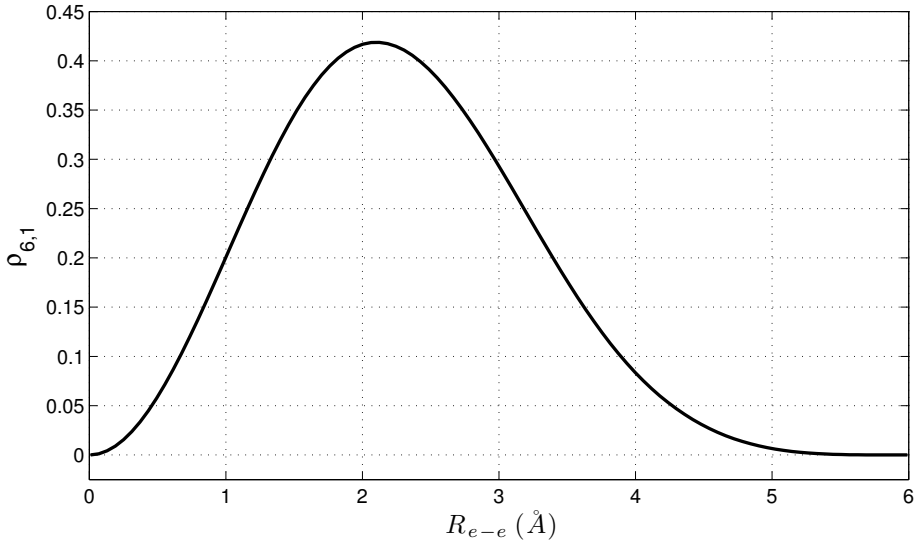

Figure 3.2: Freely jointed chain end-to-end PDF with $N = 6$, $l = 1.0\,\mathring{A}$

output of eq. 3.1 when $N = 6$; the density values are calculated by composite trapezoid rule numerical integration.

Subsequent sections of this chapter develop a coarse-grained model of a "flexible" freely jointed chain (FFJC). The end-to-end distance fills the role of the observable against which we calibrate this CG model. Consequently, a key part of selecting CG model parameters is the degree to which end-to-end distances predicted by a resultant CG model reproduce that of the flexible freely jointed chain.

3.2 Computational Models

3.2.1 The All-Atom Model

We implement an "all-atom" model of the FFJC with $N = 6$ and $l = 1.0\,\mathring{A}$ in the LAMMPS molecular dynamics simulator [82]. This model consists of seven point particles, each of nominal mass 1 amu, connected by six harmonic bonds. The bonds have the associated energy functions,

$$U(r_i, r_{i+1}) = k\left(\|r_i - r_{i+1}\| - r_{eq}\right)^2, \tag{3.2}$$

for $i \in \{0, \ldots, N-1\}$. In order to loosely enforce freely jointed chain constraints, we set $r_{eq} = 1.0\,\mathring{A}$ and $k = 500.0\,\text{kcal/mol} \cdot \mathring{A}^2$. The nearest neighbor distance constraint is, therefore, enforced in a weak sense in our model, instead of strongly through rigid bonds. This is the origin of the "flexible" modifier in our model nomenclature. We carry out molecular dynamics on the chain using a velocity Verlet time integrator with step size $\Delta t = 1.0$ fs. The system is maintained at temperature $T = 300$ K using a Langevin thermostat, which modifies the Newton equations of motion with a drag term proportional to particle velocity and a stochastic diffusion term. The coefficients of drag and diffusion in the Langevin equation are related to each other, and the imposed temperature, through the well-known fluctuation-dissipation theorem [66]. For implemention details of the Langevin thermostat, we refer to works of Schneider and Dunweg [22, 90]. We set the drag coefficient in our simulations to $\gamma = 100.0$ fs. It happens that the statistics of the FFJC found from MD simulation qualitatively match those of the exact FJC model, as can be

seen in figure 3.3.

One concern associated with using molecular dynamics to charac-

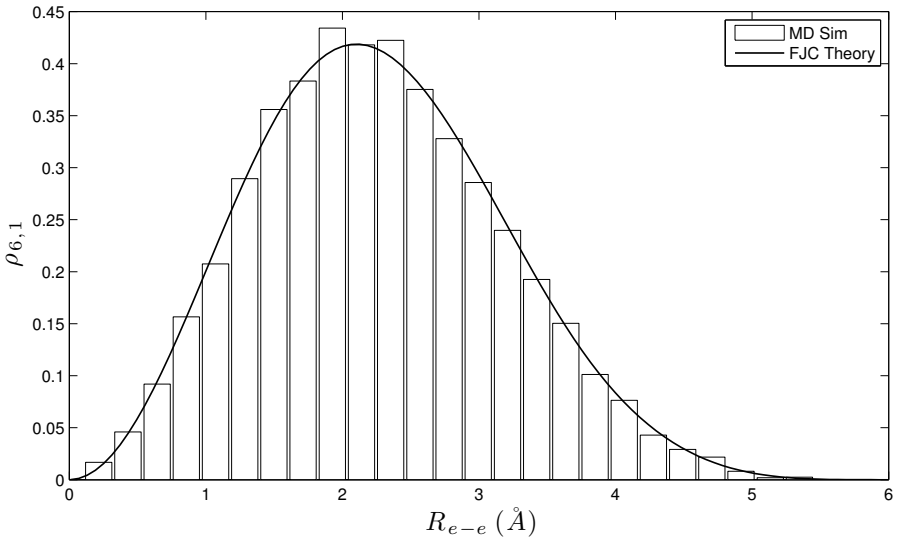

Figure 3.3: Comparison of the end-to-end distance distribution from the flexible freely jointed chain implemented in MD with that of the exact chain.

terize statistical properties of a system is correlation between samples taken from the MD code. Unbiased statistical estimators generally assume independence of samples, so it serves us to study the correlation times of the FFJC dynamics. We quantify this length by calculating the autocorrelation of the end-to-end vector time-series, $r_{06}^{(j)} = r_6^{(j)} - r_0^{(j)}$, where j is an index over the discrete timesteps in a simulation. The calculation determines a normalized covariance between the original time-series and a time-shifted copy,

$$a_k = \frac{\sum_{j=1}^{N_t-k}(r_{06}^{(j)} - \bar{r}_{06})^T(r_{06}^{(j+k)} - \bar{r}_{06})}{\sum_{j=1}^{N_t}(r_{06}^{(j)} - \bar{r}_{06})^T(r_{06}^{(j)} - \bar{r}_{06})}, \tag{3.3}$$

where N_t is the number of time-steps in the simulation, k is the number of time-steps to shift, known as the "lag," and $\bar{r}_{06}$ is the average end-to-end vector over the time span of the simulation. Figure 3.4 displays the autocorrelation, a_k, as a function of the time-step lag for the FFJC MD simulation. We draw

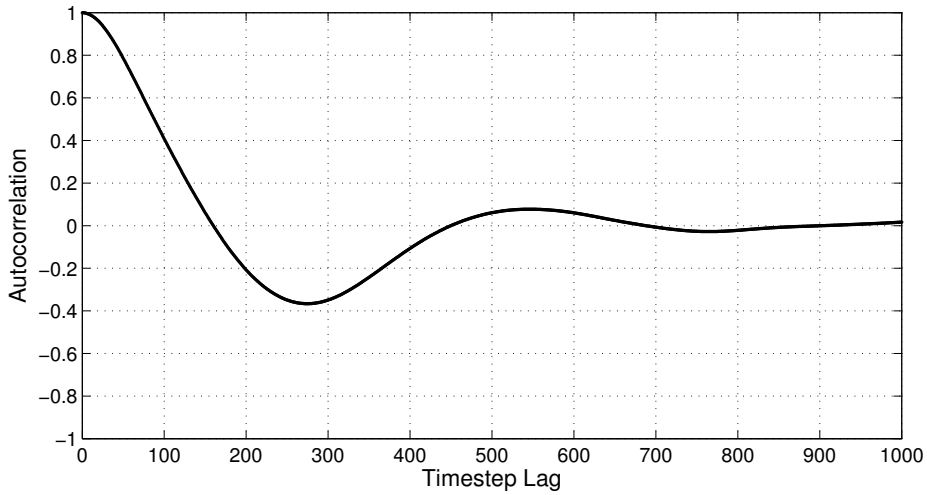

Figure 3.4: Autocorrelation in a time-series of the end-to-end vector in the freely jointed chain, with $N_t = 25{,}000$.

two conclusions from this experiment. First, any meaningful MD simulation should have $N_t \gg 100$. Second, it is prudent to "burn" time-steps between samples, that is, to allow the chain to de-correlate from its near past. We adopt the convention of saving a sample every 100 time-steps. Although a truly independent time-series of samples cannot be guaranteed, staggered samples over a long run of the simulation can be procured to adequately sample the configuration space.

3.2.2 The Coarse-Grained Model

We *choose* a coarse-grained model for the all-atom system that consists of three CG groups, or beads. Bead 1 contains the particles labeled by r_0 and r_1 in the all-atom model, bead two contains r_2, r_3, and r_4, while bead three has r_5 and r_6. We define coordinates, R_1, R_2, and R_3 for the three beads and set them to be the center of mass point of their respective beads.

More precisely, let m_k be the mass of the kth atom, where $k \in \{0, \ldots, 6\}$. For each CG bead, define the index set, $\mathfrak{I}_j = \{k \,|\, \text{atom } k \in \text{CG bead } j\}$ for $j \in \{1, 2, 3\}$. The CG coordinates can now be expressed in terms of the all-atom coordinates,

$$R_j = \sum_{k \in \mathfrak{I}_j} \frac{m_k}{\mathcal{M}_j} r_k, \tag{3.4}$$

where $\mathcal{M}_j = \sum_{k \in \mathfrak{I}_j} m_k$ is the total mass of bead j. In the language of chapter two, we have defined the mapping, M, from all-atom to CG phase space. This mapping can be represented by a matrix with entries defined by,

$$M_{jk} = \begin{cases} \frac{m_k}{\mathcal{M}_j}, & \text{if } k \in \mathfrak{I}_j \\ 0, & \text{otherwise.} \end{cases} \tag{3.5}$$

In the present case, $m_k = 1$. Thus, $\mathcal{M}_j$ is simply the number of atoms in CG bead j.

A potential energy for the CG model is chosen to consist of identical harmonic bonds between R_1-R_2 and between R_2-R_3,

$$U^{(CG)}(R_1, R_2, R_3) = K\left(\|R_1 - R_2\| - R_{eq}\right)^2 + K\left(\|R_2 - R_3\| - R_{eq}\right)^2. \tag{3.6}$$

Figure 3.5 shows a cartoon of the chosen CG model. When the parameters,

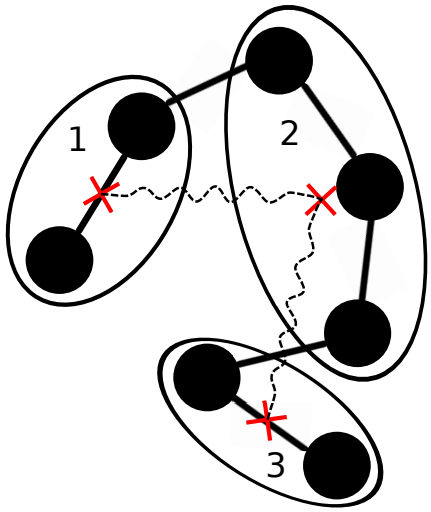

Figure 3.5: Coarse-Graining scheme for the flexible freely jointed chain. Red x's represent center-of-mass points for each CG bead. Dashed lines represent harmonic bonds between CG COM sites.

K and R_{eq}, are provided, the CG potential energy governs the interaction between CG beads, and can thus be used in an MD simulation of the CG system. The task at hand is to evaluate different choices for the CG model parameters. Subsequent sections utilize a Bayesian framework to compute a probability that the CG model, with a particular pair of K and R_{eq}, reproduces the end-to-end distance properties of the all-atom FFJC.

3.3 Bayesian Calibration

The first step in the calibration process is to gather prior information from all-atom MD runs. From a 750 ps trajectory, we find $R^*_{eq} = 1.265\,\mathring{A}$ and $K^* = 1.508$ kcal/mol $\cdot\, \mathring{A}^2$. The "soft" CG bond suggested by the prior information is not surprising given the freely jointed nature of the molecule

chain. The pdf of the prior is then given, according to (2.17) and (2.19), by,

$$p(R_{eq}, K) = \frac{R_{eq}^2}{2K^*} \left(\frac{3}{R_{eq}^*} \right)^3 \exp\left(-\frac{3R_{eq}}{R_{eq}^*} - \frac{K}{K^*} \right), \tag{3.7}$$

With the prior specified, the next issue is the likelihood function. We

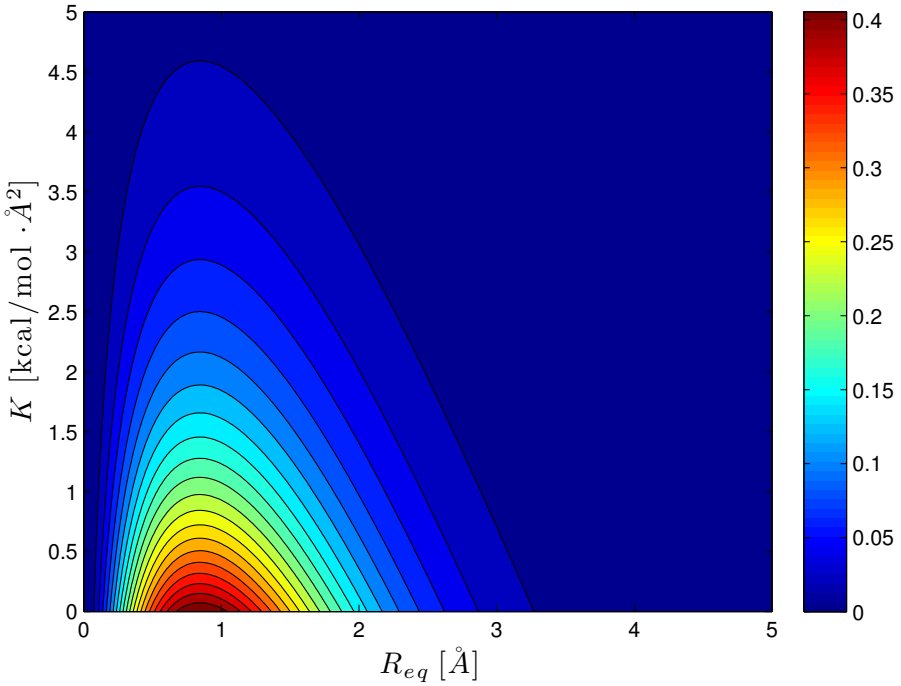

Figure 3.6: Prior PDF as given by (3.7) with $R_{eq}^* = 1.265\,\mathring{A}$ and $K^* = 1.508$ kcal/mol $\cdot\, \mathring{A}^2$

use the average end-to-end distance of the chain as our calibration observable for this particular example. The likelihood function is, therefore, given by the form in eq. 2.30. Calibration data consists of sample end-to-end distance means gathered from all-atom MD. We use $n = 120$ MD samples per mean and $m = 42$ independent mean estimates. Evaluation of the likelihood

function at given pair (R_{eq}, K) requires a MD simulation of the CG model in order to estimate $\mu_{e-e,CG}$ and $\sigma^2_{e-e,CG}$, the mean and variance of the CG end-to-end distance. The modularity of the LAMMPS MD code is used to great advantage here, as we are easily able to extract run statistics and alter the CG force field during an MD simulation using a linked library interface. In order to enhance our sampling of the CG end-to-end distance, we run 30 non-interacting replicas of the CG chain simultaneously. We observe that a likelihood function evaluation requires approximately 1 s of CPU time on a single core of a commodity workstation.

As there is no closed form expression for the posterior distribution in this case[1], we use a Markov chain Monte Carlo (MCMC) algorithm to generate independent samples from the posterior. Metropolis-Hastings type MCMC methods [87] are particularly suited to Bayesian posterior sampling, since they only require as input a function proportional to the actual posterior. In the present case, that function is the product of the prior pdf value and the likelihood value at a given (R_{eq}, K). We use the QUESO statistical analysis code to carry out "mutlilevel" MCMC sampling [83] of the posterior distribution. The modularity of LAMMPS allows us to compile the MD code directly into the sampler, which facilitates efficient communication between the Markov chain evolution code and the likelihood function evaluator. Since the MCMC sampler requires a likelihood function evaluation for each trial move of a Markov chain, the likelihood function is the rate limiting step in the posterior calcu-

[1]This is almost always the case for a calibration problem such as this.

lation. In our experience running on commodity, single CPU workstations, the time required to produce one posterior sample can be as much as half an hour. It should be noted, however, that this time includes the usual "burn in" period for the Markov chains.

An estimate of the posterior from MCMC samples is shown in fig. 3.7.

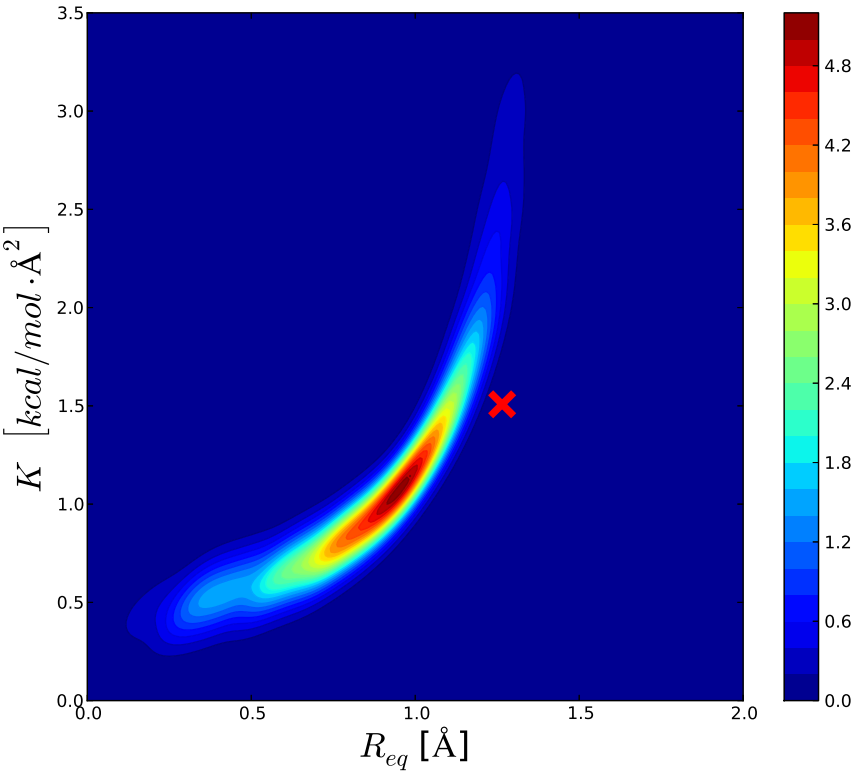

Figure 3.7: Kernel density estimate of Bayesian posterior from 15,000 MCMC samples. The Red 'X' indicates the position of the prior information, $R^*_{eq} = 1.265\,\mathring{A}$ and $K^* = 1.508$ kcal/mol $\cdot\,\mathring{A}^2$

From the figure, it is evident that considerations of end-to-end distance data

have both shifted and more sharply focused the parameter estimates in comparison to the prior information. A "sanity check" of the calibration results can be done by comparing end-to-end distance distributions from all-atom runs with those of CG models having high posterior probability density. The

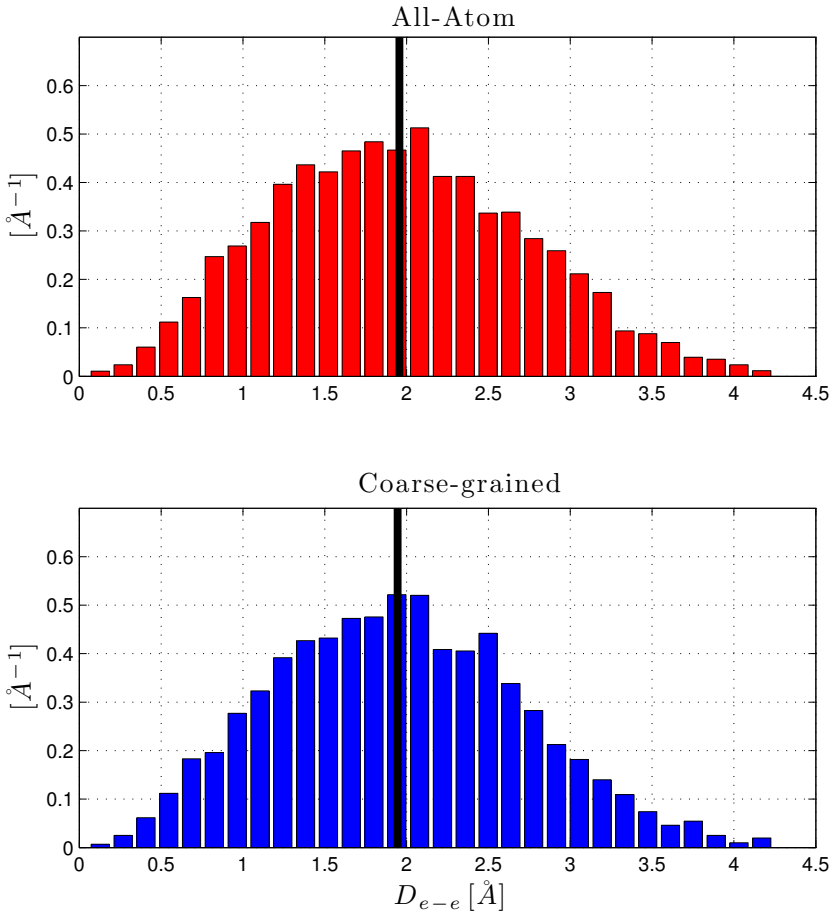

Figure 3.8: End-to-end distance histograms: All-atom and CG model with $R_{eq} = 0.97\,\mathring{A}$, $K = 1.1\,\text{kcal/mol} \cdot \mathring{A}^2$. The black lines indicate mean values.

results of one such check are displayed in fig. 3.8. In this case, the CG end-to-end distance average attains a 0.5% relative error with respect to the all-atom results. We note that in this example, the CG model is able to predict not only the mean of the end-to-end distance, but also the shape of the distribution[2].

3.4 Conclusions

We have shown in this section a full Bayesian calibration for a coarse-grained model of a flexible, freely jointed chain. In particular, we used "common sense" values from short all-atom MD simulations to find prior probabilities for CG harmonic bond constants and we chose average molecular end-to-end distance as the calibration observable characterizing the likelihood function. The posterior distribution resulting from MCMC sampling features a single peak in the CG bond parameter space, with a clear covariance in effect. A check of end-to-end distribution predictions for CG parameter values at the mode of the posterior probability density shows close agreement with all-atom results. We emphatically note, however, that the posterior gives us much more than just a set of reasonable bond constants to use in a CG model - the posterior also quantitatively represents our uncertainty in the calibration results. The utility of this uncertainty information is explored in subsequent chapters.

[2]The all-atom end-to-end distribution has a shorter "tail" than the analytical freely jointed chain result would suggest because the end-to-end distance is calculated from $\|(r_0 + r_1)/2 - (r_5 + r_6)/2\|$ to facilitate comparison with the CG model results.

Chapter 4

Liquid Heptane Case Study

We develop a Bayesian learning method for inferring the free parameters of coarse-grained (CG) potential energy functions in reference to training data from higher resolution atomistic models. Particular attention is given to informing a CG model toward an *a priori* chosen quantity of interest (QOI) and to validating the CG model in a way that takes into account quantified uncertainty in QOI prediction. We apply the method to a simple, but substantive CG model of liquid heptane (C_7H_{16}) in which the QOI is the Gibbs vapor-liquid transfer free energy. For this example, we employ a Markov chain Monte Carlo sampler alongside molecular dynamics simulations to estimate the joint posterior probability distribution of CG parameters that are associated with common molecular mechanical potential energy functions. Sequential Bayesian updates are used to improve the parameter inferences with respect to the transfer free energy. We then evaluate the accuracy and precision of the free energy predictions using samples from the parameter posterior.

4.1 Introduction

Coarse-grained (CG) models derived from source atomistic models have become a popular tool in the biochemical and materials science communities for extending the length and time scales reachable by molecular dynamics (MD) simulation. In the present work, we assume that a CG model is a reduced representation of a more detailed, "all-atom" model which lists all, or most, of the atoms comprising a substance along with a description of the connectivity of these atoms through chemical bonds. The CG representations we are concerned with are produced by a map which associates *a priori* chosen groups of atoms with single positions in three dimensional space. For instance, a CG mapping could associate groups of atoms with their respective centers-of-mass. The central task of CG modeling is then finding a potential energy function governing CG site interactions such that important features of the all-atom physics are retained. Once this is accomplished, simulations with the CG model are undertaken in scenarios heretofore inaccessible to the all-atom model due to its expense at large system size, long simulation time, or both.

The task of finding CG potential energy functions given a reduced representation has generally been approached as a regression analysis involving parameterized energy functions and optimization of a particular objective function in reference to training data from all-atom MD trajectories. Some popular examples include Boltzmann inversion [84], force matching [36, 72, 73], and relative entropy minimization [16, 93]. The defining differences between these methods lie in the choice of objective function. In the case of Boltzmann in-

version, the objective function involves a difference between CG and all-atom potentials of mean force that arise from relevant pair correlation functions. For force-matching, an ensemble average of the squared difference between CG and all-atom forces drives the objective function. Relative entropy methods minimize an objective function involving the relative entropy, or Kullback-Leibler divergence, between CG and all-atom statistical mechanical probability distributions. For a more thorough discussion of these and other CG methods, we refer to the excellent perspective article of W. Noid [71].

The goal quantity in the aforementioned CG methodologies is the optimal vector of potential energy parameters, as determined by the choice of objective function, that is necessary to fully specify the CG Hamiltonian. In contrast, we develop a Bayesian method for potential energy inference that computes probability distributions for the model parameters. There are a number of advantages to this approach: (1) Uncertainties in CG model predictions of physical observables and quantities of interest (QOIs) are naturally quantified due to the probabilistic representation of information inherent in Bayesian methodology, (2) The CG model can "learn" in the sense that previously inferred parameter distributions can be improved or adapted when new types of all-atom training data are supplied, (3) Competing CG models can be compared to each other on a rigorous, quantitative basis using methods of Bayesian model plausibility and selection. Uncertainty analysis, as introduced in (1), relates to issues of *model validation*; that is, determining the confidence one has in the ability of the model to predict QOIs to within pre-

set tolerances based on the accuracy with which the model predicts specific observables [75, 77]. We demonstrate CG model validation in tandem with Bayesian learning in this work. For an exploration of Bayesian model selection as it applies to CG modeling, we refer to the work of Farrell and Oden [25]. We also note the work of Koutsourelakis and Bilionis for their application of Bayesian inference to multiscale modeling of general dynamical systems [50].

In the following section, we provide a brief review of Bayesian methodology and derive prior probabilities and likelihood functions relevant to CG modeling. The basis of our prior probability representations is the principle of maximum entropy, which gives a maximally spread out probability distribution that satisfies constraints of our choosing. Likelihood functions follow from applying the central limit theorem to molecular dynamics estimation of observables deemed important to the eventual prediction of a specific QOI. We then devote a section to developing a CG model of liquid heptane that is designed to predict, to within a preset tolerance, vapor-liquid transfer free energy at standard temperature and pressure. Within this development, principles of model validation and Bayesian learning via posterior updating are explored.

4.2 A Bayesian Theory for Coarse-Grained Parameter Learning

The most basic idea underlying Bayesian methodology is that probability is a quantitative degree of belief in the truth of a given proposition.

This contrasts with the "frequentist" definition of probability as a limiting frequency of occurrence for an event in a large number of repeated trials. In the Bayesian setting, probabilistic statements are updated upon consideration of data through Bayes' formula,

$$\overbrace{\varrho(\boldsymbol{\theta}\,|\,\boldsymbol{d})}^{\text{posterior}} = \frac{\overbrace{\varrho(\boldsymbol{d}\,|\,\boldsymbol{\theta})}^{\text{likelihood}}\overbrace{\varrho(\boldsymbol{\theta})}^{\text{prior}}}{\int d\boldsymbol{\theta}\,\varrho(\boldsymbol{d}\,|\,\boldsymbol{\theta})\varrho(\boldsymbol{\theta})}, \tag{4.1}$$

where $\boldsymbol{\theta}$ is the proposition of interest, in our case a vector of model parameters, and $\boldsymbol{d}$ represents observed data. In eq. 5.1, we assume $\boldsymbol{\theta}$ and $\boldsymbol{d}$ take values in a continuum, so the $\varrho(\cdot)$ indicate probability density functions (PDFs). Statements of the form $\varrho(\boldsymbol{x}\,|\,\boldsymbol{y})$ indicate the usual shorthand for conditional probability wherein the degree of belief in $\boldsymbol{x}$ is conditioned on the truth of the fixed statement, $\boldsymbol{y}$. We take the term "Bayes update" to mean the procedure by which the posterior probability, conditioned on the data, is produced from a given prior probability, which is independent of the data. The likelihood function is the connective device in this context, giving the probability of the data as a function of the parameter vector. This updating procedure can be iterated when new data is available by using the previous update's posterior as the current prior. For a thorough review of Bayesian probability theory, we refer to the book of E. T. Jaynes [42]. The likelihood function concept has a long history in frequentist statistics; we refer to the seminal work of R. A. Fisher on this topic [26].

To specify the prior probability in our framework, we state two requirements: (1) The prior should be be *maximally uncertain*, as measured by the

Shannon entropy functional [92], over the range of the CG model parameters and (2) initial guesses for parameter values should be encoded into the prior through mean value constraints. These requirements certainly do not constitute the only way to define a prior probability; however, the procedure they imply is in close correspondence with information theoretic formulations of statistical mechanics [40]. Hence, we find the mean value constrained maximum entropy solution to be a reasonable and useful choice. It is now a straightforward problem in constrained optimization to find the PDF maximizing the Shannon entropy,

$$H[\varrho] = -\int d\theta_i \, \varrho(\theta_i) \log \frac{\varrho(\theta_i)}{\eta(\theta_i)}, \tag{4.2}$$

such that $\int d\theta_i \, \varrho(\theta_i) \, f(\theta_i) = \mu_i$. Here, $\eta(\theta_i)$ is the so-called invariant measure which contains the relevant differential volume factor for θ_i, f is a known function of the parameter for which prior information is obtainable, and μ_i is the prescribed mean value. The differential volume contains the Jacobian for integration in three dimensional space in the case of parameters that correspond to locations in space, such as equilibrium positions for generalized coordinates. Solutions of the optimization problem take the form [40, 42],

$$\varrho(\theta_i) = \frac{\eta(\theta_i)}{Z(\mu_i)} \exp\left\{\lambda(\mu_i) \, f(\theta_i)\right\}, \tag{4.3}$$

where λ is a Lagrange multiplier, and Z is a normalization constant. We note that the joint prior for the parameter vector is simply a product of the component priors. That is, the parameters are initially assumed independent in the absence of prior information specifying parameter covariance.

The final issue we encounter in defining the prior is the origin of the prescribed mean value constraints. Here, we find that short, all-atom MD trajectories that are *a posteriori* mapped into the CG representation provide a wealth of prior information, or initial guesses, for CG potential energy parameters. Particularly, we calculate statistical properties, such as means, modes and variances, from observed distributions for generalized coordinates associated with model parameters in the CG representation. We emphasize that the goal of these calculations is to provide *order-of-magnitude* information on the parameters. The maximum entropy requirement for prior probability implies that the resulting distributions are necessarily broad, and thus, highly accurate estimates of the parameters are not desired at this initial stage.

In defining the likelihood function, we restrict our attention to data that can be expressed as an ensemble average from MD simulation. This includes a wide variety of observables that are of practical interest to molecular modelers, thus the following development is quite general. To begin, we note that the distribution of data sampled from a CG simulation is asymptotically gaussian due to the central limit theorem for convergence of sample means,

$$\varrho(d \,|\, n, \boldsymbol{\theta}, \mu, \sigma^2) \approx \sqrt{\frac{n}{2\pi\, \sigma^2(\boldsymbol{\theta})}} \exp\left\{ -\frac{n}{2} \frac{(d - \mu(\boldsymbol{\theta}))^2}{\sigma^2(\boldsymbol{\theta})} \right\}. \tag{4.4}$$

Here, n is the number of independent samples comprising the sample mean, d, while μ and σ^2 are the true mean and variance for the observable. The former is the usual statistical mechanical ensemble average which corresponds to a thermodynamic quantity; the latter arises from the inherent uncertainty in the statistical mechanical ensemble. Since the CG Hamiltonian governs the

physics producing these values, there is an explicit dependence on the CG model parameters. In practice, the μ and σ^2 from eq. 4.4 are not known exactly, but estimated from a finite number of samples taken from CG MD simulation. We incorporate the uncertainty from this estimation by including the hyper-parameters, μ and σ^2, in the parameter inference and integrating over their full ranges. Using Bayes' formula, the parameter posterior now takes the form,

$$\varrho(\boldsymbol{\theta} \,|\, n, m, d) \propto \int_0^\infty d\sigma^2 \int_{-\infty}^\infty d\mu \, \varrho(d \,|\, n, \boldsymbol{\theta}, \mu, \sigma^2) \, \varrho(\boldsymbol{\theta}, \mu, \sigma^2 \,|\, m), \tag{4.5}$$

where we now condition on the number m of CG MD samples used to estimate the hyper-parameters. Following the usual rules of conditional probability, the joint prior in eq. 4.5 factors into three products,

$$\varrho(\boldsymbol{\theta}, \mu, \sigma^2 \,|\, m) = \varrho(\mu \,|\, m, \boldsymbol{\theta}, \sigma^2) \, \varrho(\sigma^2 \,|\, m, \boldsymbol{\theta}) \, \varrho(\boldsymbol{\theta}). \tag{4.6}$$

Assuming the MD estimates $\widehat{\mu}_m$ and $\widehat{\sigma}^2_m$ have been computed, we assign the priors,

$$\varrho\left(\mu \,|\, m, \boldsymbol{\theta}, \sigma^2\right) = \sqrt{\frac{m}{2\pi\sigma^2}} \exp\left\{ -\frac{m}{2} \, \frac{\left(\mu - \widehat{\mu}_m(\boldsymbol{\theta})\right)^2}{\sigma^2} \right\}, \tag{4.7}$$

and,

$$\varrho\left(\sigma^2 \,|\, m, \boldsymbol{\theta}\right) = \frac{1}{\widehat{\sigma}^2_m(\boldsymbol{\theta})} \exp\left\{ -\frac{\sigma^2}{\widehat{\sigma}^2_m(\boldsymbol{\theta})} \right\}. \tag{4.8}$$

The PDFs in eqs. 4.7 and 4.8 are maximum entropy densities with respect to given constraints. In the case of eq. 4.7, the prior mean is constrained to the estimated observable mean and the variance is constrained to σ^2/m, in line with the central limit theorem for convergence to a mean. The distribution

of maximum entropy with mean and variance constrained in this manner is a Gaussian. Eq. 4.8 is an example of eq. 4.3 where a prior is found from a mean value constraint. With the priors specified, the integrals in eq. 4.5 can be evaluated analytically,

$$\overbrace{\varrho(\boldsymbol{\theta} \mid n, m, d)}^{\text{posterior}} \propto \overbrace{\frac{1}{2\,\widehat{\gamma}_{mn}(\boldsymbol{\theta})}\,\exp\left\{-\frac{|d-\widehat{\mu}_m(\boldsymbol{\theta})\,|}{\widehat{\gamma}_{mn}(\boldsymbol{\theta})}\right\}}^{\text{reduced likelihood}} \times \overbrace{\varrho(\boldsymbol{\theta})}^{\text{prior}}, \tag{4.9}$$

where $\widehat{\gamma}_{mn}(\boldsymbol{\theta}) = \widehat{\sigma}_m(\boldsymbol{\theta})\sqrt{\frac{1}{2}\,\left(\frac{1}{m}+\frac{1}{n}\right)}$. The calculations leading to this result are detailed in the appendix for the interested reader.

Eq. 4.9 gives the final symbolic expression for the posterior. Here, the reduced likelihood has the form of a Laplace distribution [49]; the consequence of accounting for hyper-parameter uncertainty is that the sampling distribution for the data has broader tails than the original Gaussian. It is assumed that the data, d, in the likelihood expression is supplied by averaging n independent samples drawn from all-atom MD simulation. Thus, the likelihood is maximized for CG parameters that produce an observable estimate that is identical to that of the all-atom model. In the case of multiple observables, we assume that data for each observable can be gathered independently of the others so that the full likelihood expression is simply a product of single observable likelihoods. We note also that no uncertainty is assumed in the all-atom data. Such a quantity could be included in the likelihood by adding a term, σ^2_{AA}, to the variance in eq. 4.4; however, this term is of limited consequence in this context because we restrict our development to observables than can be calculated to high accuracy in the all-atom model and because we already

consider a wide variance range by integrating over the hyper-parameters.

It is clear from eq. 4.9 that an exact, closed form expression for the posterior is not generally available, as $\widehat{\mu}$ and $\widehat{\gamma}_{mn}$ depend on MD simulations with varying parameter vectors. We therefore utilize Metropolis-Hastings (MH) Markov chain Monte Carlo (MCMC) techniques to obtain numerous independent samples from the exact posterior [104]. Once acquired, these samples are used as the seeds for future posterior updates via Bayes' Rule and also to calculate distributions of QOIs for the purposes of prediction and validation. The latter is accomplished through the integral,

$$\varrho(q \,|\, \boldsymbol{d}) = \int d\boldsymbol{\theta}\, \varrho(q \,|\, \boldsymbol{\theta})\, \varrho(\boldsymbol{\theta} \,|\, \boldsymbol{d}), \tag{4.10}$$

where q represents a scalar QOI and $\varrho(q \,|\, \boldsymbol{\theta})$ is the PDF for QOI values computed by a CG model with given parameter vector. The posterior QOI is thus expressed as a posterior-weighted average of predictions from all possible parameter vectors. In practice, we approximate eq. 4.10 from posterior samples $\left\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_p}\right\}$ via the Monte Carlo expression,

$$\varrho(q \,|\, \boldsymbol{d}) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} \kappa\left(q - \widehat{q}\,(\boldsymbol{\theta}_i)\right), \tag{4.11}$$

where κ is a kernel PDF with mean zero and $\widehat{q}$ is an estimate of the QOI computed by the CG model. The kernel function is often chosen to be gaussian with a variance parameter that depends on the input samples. We refer to the book of Silverman for a development of such "kernel density estimation" methods [101].

The theory is now in place for computing PDFs of CG potential energy parameters given all-atom data for chosen observables, and subsequently, for making probabilistic predictions of QOIs. In the next section, we apply the Bayesian theory to a CG model of heptane, developing a maximum entropy prior from simple constraints and a likelihood function from a small collection of observables. Following eqs. 4.10 and 4.11, PDFs for a free energy are computed, providing a goal-oriented basis on which to evaluate and improve the CG model.

4.3 Application to a Liquid Heptane Model

The reference atomistic model for our CG demonstration is the TraPPE model for alkanes [64]. TraPPE consists of a united atom[1] representation and molecular mechanical potentials parameterized to reproduce free energies of transfer and other quantities related to phase equilibria. We pick a simple CG designation of atomistic heptane into three beads, as shown in fig. 5.3a. The two end beads comprise $CH_3 - CH_2$ atom groups, while the middle bead contains the group, $CH_2 - CH_2 - CH_2$. To describe the configuration of these three beads from an atomistic configuration, we map the atom groups to their respective centers-of-mass. Accordingly, the CG model has two end particles assigned to one particle type, 'E,' and a middle particle assigned its own distinct type, 'M.' Fig. 5.3b displays this representation with effective chemical

[1]The TraPPE model does not include hydrogen atoms; their cumulative effect is contained in "united atoms" CH_3 and CH_2.

bonds shown between adjacent beads. We propose a simple, molecular me-

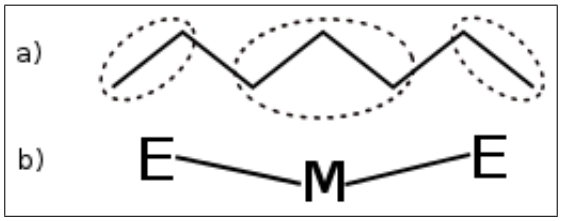

Figure 4.1: (a) Atomistic heptane in skeletal representation with CG bead designations and (b) corresponding CG heptane with particle types E and M.

chanical potential energy function for the CG model: linear springs govern the E-M bonds and the E-M-E angle, while a Lennard Jones 9-6 form with a cutoff at 14.0 $\mathring{A}$ defines the non-bonded interactions. Thus, the potential is the sum of three terms:

$$\begin{aligned} U_{CG}(\boldsymbol{x};\boldsymbol{\theta}) = & \sum_{i \in \text{bond}} k_r (r_i(\boldsymbol{x}) - r_0)^2 \\ & + \sum_{i \in \text{angle}} k_\varphi (\varphi_i(\boldsymbol{x}) - \varphi_0)^2 \\ & + \sum_{i,j \in \text{pair}} 4\,\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}(\boldsymbol{x})} \right)^9 - \left(\frac{\sigma_{ij}}{r_{ij}(\boldsymbol{x})} \right)^6 \right] \mathbb{I}_{\left\{ r_{ij} < 14 \mathring{A} \right\}}, \end{aligned} \tag{4.12}$$

where $\boldsymbol{x}$ denotes the vector of 3-D positions constituting an "atomic" configuration in the CG system and r_i, φ_i, and r_{ij} are generalized coordinates indicating, respectively, bond length, angle between connected bonds, and pair distance between sites not connected by bonds. In this case, the parameter vector $\boldsymbol{\theta}$ consists of eight quantities: r_0, k_r, φ_0, k_φ, σ_{EE}, ϵ_{EE}, σ_{MM}, and ϵ_{MM}. We set the E-M Lennard-Jones parameters via the common mixing rules,

$$\begin{aligned} \sigma_{EM} &= (\sigma_{EE} + \sigma_{MM})/2 \\ \epsilon_{EM} &= \sqrt{\epsilon_{EE}\,\epsilon_{MM}}. \end{aligned} \tag{4.13}$$

The physical scenario for parameter inference is given by 128 heptane molecules in a cubic volume with periodic boundary conditions and held at a temperature of 300 K and 1 atm of pressure. Tail corrections are applied to the energy and pressure to account for the Lennard-Jones cutoff [99]. In both all-atom and CG simulations, the temperature is controlled by a Langevin thermostat [22, 90] while the pressure is maintained with a Parrinello-Rahman type barostat [78]. We use the LAMMPS MD code with a time step of 1.0 fs for all physics simulations [82]. The all-atom system is initialized on a lattice and then equilibrated for one nanosecond, enough time for the mean energy and domain volume to reach stable values. Prior information is then gathered from a 100 ps MD trajectory that is mapped into the CG representation. From this trajectory, we extract initial guesses for the bond and angle parameters by estimating the mean and variance of the corresponding generalized coordinate distributions. The mean values are associated with equilibrium positions, while the variances are related to the spring constants. Lennard-Jones parameters are informed by computing radial distribution functions, along with their corresponding potentials of mean force, for E-E and M-M pairs. Maximum entropy priors are then calculated using the initial guesses as mean value constraints, as per eq. 4.3. The invariant measures use the form $4\pi r^2$ for distance parameters r_0, σ_{EE}, and σ_{MM} and the form $\sin\varphi_0$ for angle parameter φ_0. In all other cases, the invariant measure is assumed constant. The resulting PDFs are shown in fig. 4.2. We refer to section 2.2 for a more detailed account of prior PDF construction. We note the wide spread in the

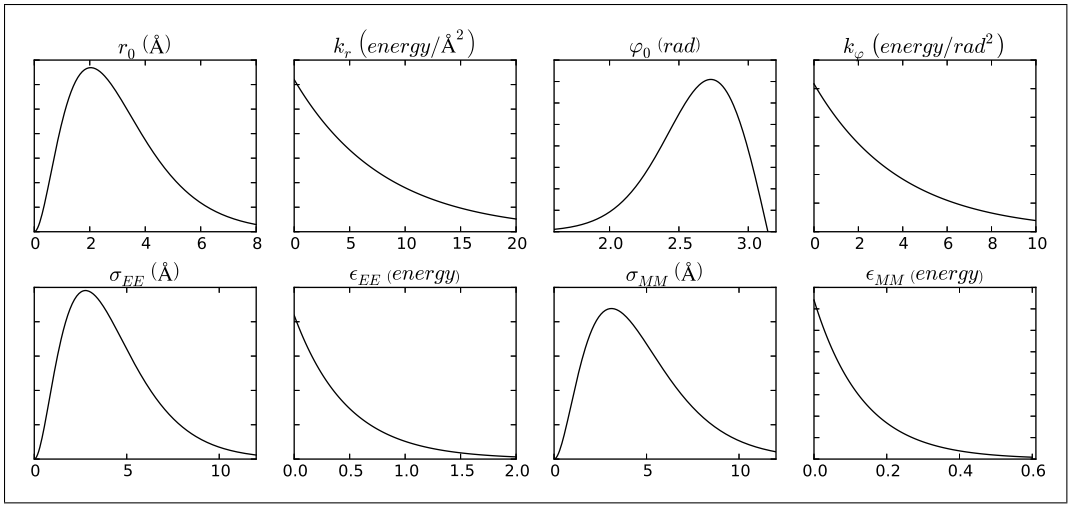

Figure 4.2: 1-D prior PDFs for CG heptane parameters. Energy units are kcal/mol.

prior PDFs, indicating relatively large uncertainty for parameter values in the initial representation of knowledge.

To improve on the prior knowledge, we undertake a Bayes' update using all-atom training data for two observables: mass density and molecular interaction energy. The latter is computed by summing all non-bonded energy between a single heptane molecule and the remaining 127 molecules, plus a correction energy arising from the finite Lennard-Jones cutoff. The rationale behind selecting these observables is that the CG model should be informed about configurational entropy and interaction energy when predicting vapor-liquid transfer free energy; the chosen observables are easily computed in simulation and provide relevant information about these two components of the free energy. Average values for these observables are computed from $n = 50$ samples gathered over a nanosecond of all-atom simulation. Although

the observables attain stability after much shorter time spans, we run for a nanosecond to aid in sample independence. The means computed from the samples then serve as data for likelihood functions having the form given in eq. 4.9. Given the data, posterior samples are produced using the QUESO[2] multilevel MCMC code [83]. As each Markov chain transition requires an estimation of CG hyper-parameters, the LAMMPS MD code is compiled into the QUESO likelihood function to allow MD simulation in the course of Markov chain construction. The CG MD, for each sampled parameter vector, is started from a reference state and further equilibrated for 200 ps. Hyper-parameters are then estimated from $m = 60$ samples gathered over 60 ps. To speed posterior sampling, Markov chains are run in parallel on 56 CPU cores. Figure

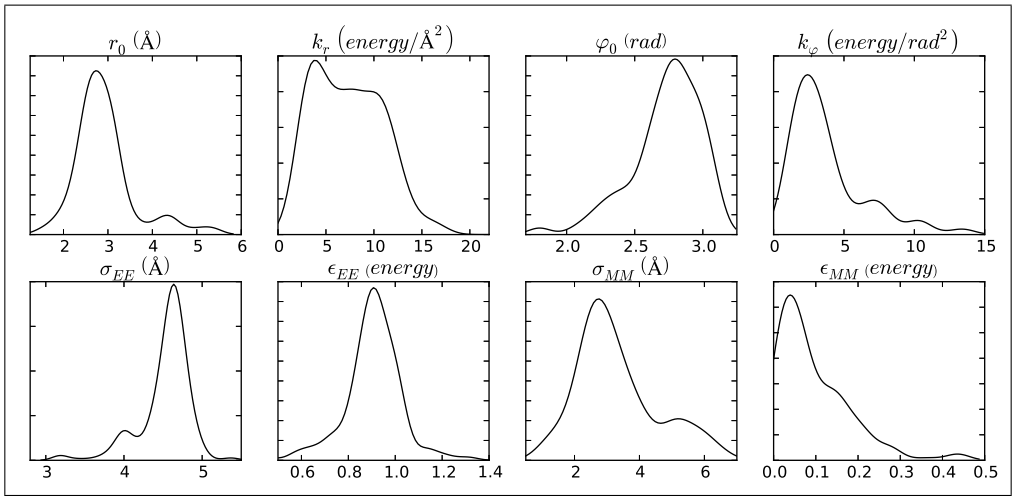

Figure 4.3: 1-D posterior PDFs for CG heptane parameters from kernel density estimates. Energy units are kcal/mol.

[2]Quantification of Uncertainty for Estimation, Simulation and Optimization

4.3 displays kernel density estimates of the 1-D posteriors derived from 2800 independent samples. We note, however, that the 1-D views show a limited subset of the information available from the full, 8-D joint posterior. For instance, the joint posterior samples show a strong, negative correlation between σ_{EE} and σ_{MM} which is expected since that the system density is a targeted observable.

We now examine the distribution of Gibbs transfer free energy that results from the estimated parameter posterior. Transfer free energies are calculated using a variant of the free energy perturbation method. Here, the "perturbation" is the introduction of a new heptane molecule into the simulation box. The calculated free energy is thus a difference involving two states: an initial state where the new molecule is intermolecularly uncoupled from the original system and a final state with full coupling. To mitigate phase space overlap issues, 14 intermediate states are defined. We use a "soft-core" Lennard Jones potential in the intermediate states to improve MD sampling and to remove the singularity arising from the repulsive term in the usual Lennard Jones form [105]. Free energy differences between states are then computed using the Bennett acceptance ratio (BAR) method [7] as implemented in the pyMBAR python code [98]. We find that four nanoseconds of MD simulation per state is sufficient for our calculations, with an estimated uncertainty of approximately 0.02 kcal/mol for each full transfer free energy computation. Figure 4.4 displays an estimate of the QOI PDF based on free energy calculations at 300 different parameter samples from the joint poste-

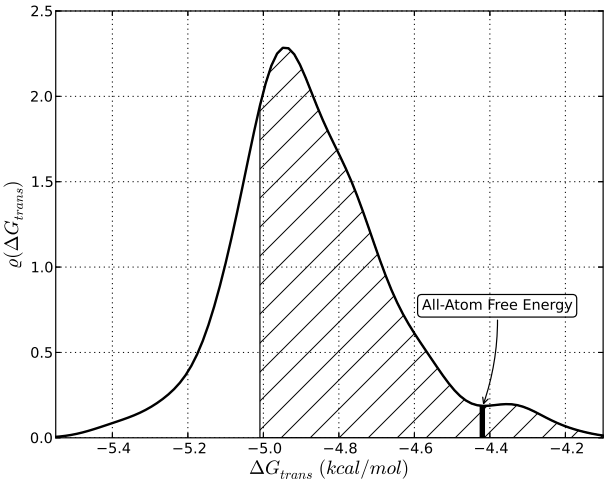

Figure 4.4: Kernel density estimate of CG transfer free energy PDF from 300 free energy calculations. The black bar marks the all-atom free energy, while the hatched area shows the portion of the PDF within k_bT of the all-atom value.

rior. Using more samples does not appreciably change the features of this PDF. To evaluate the predictive capacity of the CG model, we compute the proportion of QOI probability density within k_bT of the all-atom free energy. The latter is calculated from all-atom simulation and found to be -4.42 +/− 0.01 kcal/mol. We note that this is within 0.25 kcal/mol of both the configurational bias Monte Carlo simulation result of Martin et al. [64] and the experimental measurements of Eikens [23]. We then find that approximately 77% of the free energy PDF is within k_bT of the all-atom value, as illustrated in fig. 4.4.

The next consideration is whether the former parameter inference can be improved in reference to the free energy. It is observed that the posterior

distribution for CG heptane radius of gyration is relatively wide, and furthermore, posterior free energy and radius of gyration are strongly correlated, as show in figure 4.5. This suggests that the CG model could be better informed

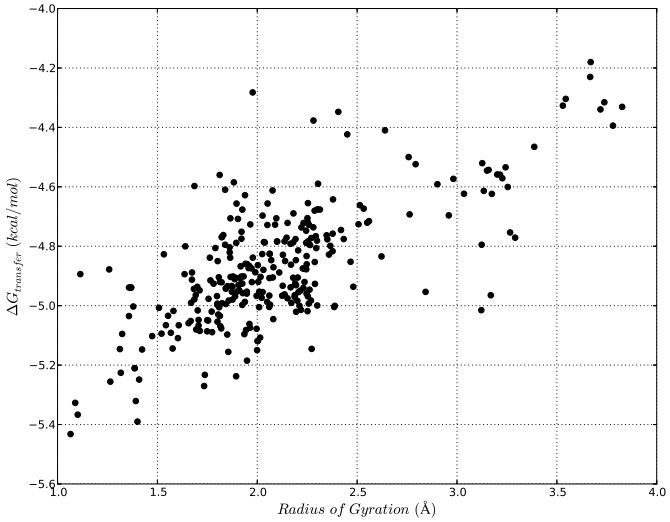

Figure 4.5: Radius of gyration value vs. transfer free energy for 300 posterior samples.

in relation to configurational entropy if the radius of gyration were included in the training data. Starting from the posterior samples previously determined, a Bayes' update using all-atom radius of gyration data results in new posteriors estimated from 16,000 samples, as shown in figure 4.6, and an improved free energy prediction. Figure 4.7 displays the updated free energy PDF; we now find about 99% of the probability density within k_bT of the all-atom result. Clearly, the update results in a quantitatively better inference for the CG potential energy parameters and we can say, with high confidence, that the

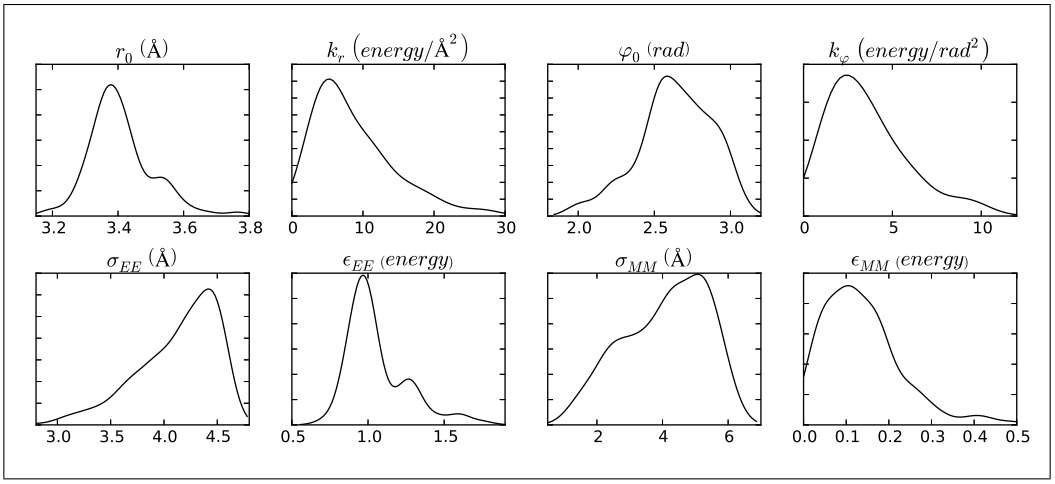

Figure 4.6: Improved 1-D posteriors with radius of gyration training. Energy units are kcal/mol.

proposed CG model predicts the transfer free energy to within chemical accuracy when trained with all-atom density, molecular interaction energy, and radius of gyration.

4.4 Discussion

Uncertainty is an essential feature of CG model construction. In reducing the number of atomistic degrees-of-freedom, we introduce a degeneracy of coarse models that reproduce subsets of features from the full resolution model. In this work, we directly address that degeneracy in relation to CG potential energy parameterization. Thus, we avoid computing an optimal vector of parameters relative to a particular objective function and seek, instead, to answer the questions, (1) How well determined by the data are the model parameters? (2) What consequence does parameter uncertainty have on the calculation of

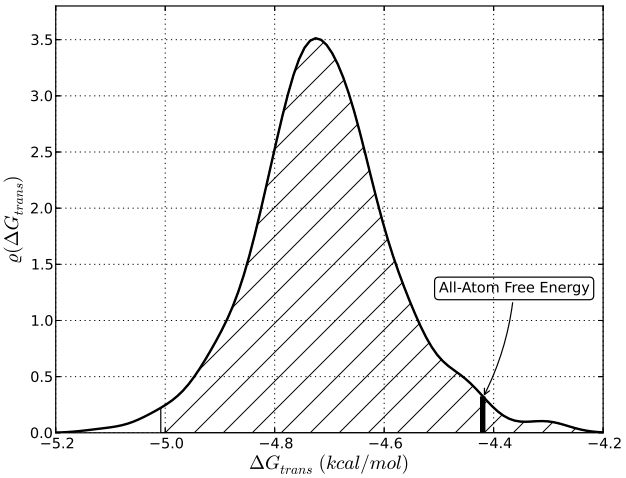

Figure 4.7: Improved estimate of CG transfer free energy PDF from 300 free energy calculations. The black bar marks the all-atom free energy, while the hatched area shows the portion of the PDF within k_bT of the all-atom value.

quantities that the model is designed to predict? (3) Can we improve parameter inferences in reference to a QOI when new or different information becomes available? The tools of Bayesian statistics give us a robust platform for answering these questions, as we show in this work for a CG model of heptane. In particular, we illustrate the value of representing parameter uncertainty to the issue of CG model validation. Here, we go beyond validating particular parameter values and asses the *information* that determines these values. We find in this case that a set of three easily computed and intuitively grounded observables suffices to train a simple CG model toward chemically accurate and precise prediction of a free energy. We do not believe this conclusion to be obvious given the coarseness of the model, its simple molecular mechanical

potential energy ansatz, and the complex nature of calculating free energy integrals.

The rigorous consideration of uncertainty, although providing a great deal more information to a CG model than parameter optimization only, does this at significant computational expense. Even with a 56 core cluster, gathering the posterior samples used in this work requires around three weeks of continual calculation. This is because the MCMC incurs MD simulation at each chain transition in order to evaluate the likelihood function. We argue, however, that this sluggish performance can be greatly improved and that the benefits of quantifying parameter uncertainty, particularly to issues of validation, are worth extra computation in many cases. To address the former, we note that the sampling algorithm used here is of a "brute force" type. Order-of-magnitude improvements in speed are likely to be found by constructing surrogate likelihood response surfaces that require sparser MD simulation to evaluate. Furthermore, a careful choice of physical scenario for parameter inference can result in significantly faster MD simulations in the course of likelihood evaluation. As to the utility of extra computation in the name of uncertainty quantification, we observe that the increasing emphasis on molecular dynamics simulations to make predictions in biophysical scenarios and material design necessitates the construction of multiscale models. The variety of these applications means that it is also increasingly important to validate the accuracy and precision of model predictions for a diverse set of QOIs in a multitude of physical scenarios, many of which the model may not

be explicitly trained to calculate. In line with this last point, we would also like to be able to improve the model, if possible, for customized applications in a goal-oriented way. The cost of working in a Bayesian framework is more than justified, in our opinion, when a quantitative assessment of uncertainty is vital to the predictive power of a model.

Chapter 5

Polythiophene Case Study

We develop a coarse-grained (CG) model of 3-polyethylthiophene (P3ET) utilizing common molecular mechanical potential energy functions. The free parameters therein are calibrated via a Bayesian learning approach that references training data from all-atom molecular dynamics (MD) simulations. In such an approach, we use Markov chain Monte Carlo (MCMC) sampling to estimate posterior probability distributions for CG parameters that are conditioned on important all-atom observable values and then apply Bayesian decision theory to select a "best" parameter set relative to the posterior expectation of a chosen loss function. The CG model determined by this procedure is finally subjected to validation tests involving the aggregation properties of P3ET oligomers. These tests are particularly pertinent to models of certain organic photovoltaic (OPV) materials in which aggregates of thiophene polymer function as electron donating material.

5.1 Introduction

In parallel with recent advances in organic photovoltaic (OPV) materials research is an increasing interest in the atomic scale behavior of thiophene

polymers in aggregated states. Certain thiophene polymers, in mixture with other constituents, have been shown to be effective electron donors upon absorption of visible light. In particular, so called bulk heterojunction (BHJ) cells containing phase separated blends of poly-alkyl-thiophene and fullerene make up some of the most promising OPVs for commercial solar cell development [10, 11, 109]. The amount of electrical power deliverable by such devices

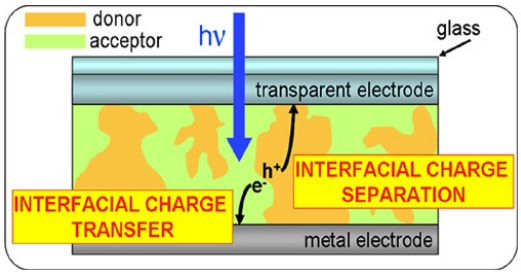

Figure 5.1: Basic representation of a bulk heterojunction solar cell

is governed by several processes at the atomic and device scale. Among the most important are the efficiencies of light absorption, charge separation, and charge transport through the BHJ [5, 39, 89]. Many aspects of these processes are difficult to understand through purely empirical means, and hence, the construction of accurate models is imperative to the continued improvement of device fabrication techniques and eventual solar cell performance [19, 29, 103]. Of particular note for the present work are efforts to build molecular mechanical, atomistic models of poly-alkyl-thiophenes using information from quantum chemical calculations [21, 68]. These models enable the study of polymer conformation in physical scenarios involving the aggregation of multiple chains, as one finds in electron donating regions of BHJs. Here, the space of likely

conformations is important to understand, as conformation is widely believed to influence the optical and electronic properties of polythiophenes.

As the typical BHJ device thickness is on the order of 50-100 nm [5], atomic scale modeling is a logistic and computational challenge. Coarse-graining, the process of reducing certain collections of atoms to single sites, is thus one method of choice for reducing model complexity in molecular dynamics simulation of BHJ constituents. The principle challenge in producing a coarse-grained (CG) model is generally in finding a suitable potential energy function that describes the physical interactions between CG sites in a way that is "compatible" with a source atomistic model [71]. In this work, we construct a CG mapping and potential energy for poly-ethyl-thiophene, a representative member of the class of poly-alkyl-thiophenes of interest in OPV device research. Our purpose with this model is to enable efficient study of poly-akyl-thiophene in aggregate states. Specifically, we are interested in reproducing all-atom conformational properties such as the inter-monomer dihedral angle distribution, single chain radius-of-gyration and end-end distance, and minimal contact distances between monomers on separate chains. These properties give indications of polymer geometry, packing, and, in the case of the dihedal angle distribution, the degree of planarity within the polymer chains. Increased planarity in consecutive sections of polythiophene is associated with a greater degree of pi electron delocalization, and subsequently, more favorable optical properties for the purposes of photocurrent generation; thus, it is assumed that the inter-monomer dihedral angle is a very important

observable to preserve in a CG model.

In our previous work, we introduced a Bayesian learning method for inferring CG potential energy parameters from all-atom training data. The data was assumed to come from MD estimates for a chosen set of ensemble averages; these averages correspond with observables deemed important for the eventual application of the CG model. A major benefit here is that the resulting CG potential is ostensibly optimized with respect to the chosen observables. Additionally, the posterior probabilities generated by the Bayesian inversion provide a measure of the uncertainty in the parameter estimation and also, importantly, a basis from which to improve the CG parameters with regard to other observables as needed. We utilize, and build upon, this Bayesian framework to calibrate a CG model of poly-ethyl-thiophene with respect to a small set of conformational observables. The two major additions to the framework that we undertake here are (1) the derivation of a likelihood function in the case that the all-atom data is an estimated probability distribution for an observable, in contrast to the case where the data is an estimated mean of an observable and (2) the use of Bayesian decision theory to determine the "best" parameter set relative to the posterior parameter expectation of a chosen loss function. The former is relevant to CG modeling of polythiophene because we would like to reproduce the all-atom inter-monomer dihedral angle distribution and the latter is necessary for the validation and reporting of a single, specified CG model.

5.2 Bayesian Theory for CG Model Calibration

A Bayesian view of CG model calibration begins with Bayes' formula for the posterior probability density of a model parameter set conditioned on data,

$$\overbrace{\varrho(\boldsymbol{\theta}\,|\,\boldsymbol{d})}^{\text{posterior}} = \frac{\overbrace{\varrho(\boldsymbol{d}\,|\,\boldsymbol{\theta})}^{\text{likelihood}}\,\overbrace{\varrho(\boldsymbol{\theta})}^{\text{prior}}}{\int d\boldsymbol{\theta}\,\varrho(\boldsymbol{d}\,|\,\boldsymbol{\theta})\varrho(\boldsymbol{\theta})}, \tag{5.1}$$

where $\boldsymbol{\theta}$ refers to a vector of CG potential energy parameters and $\mathbf{d}$ refers to an observed all-atom data set. The calibration of the model parameters first necessitates a specification of a prior probability for the parameters as well as a probability density for data sampled from a model with a given parameter vector. When the latter quantity is regarded as a function of parameter vector with given data, it is known as a likelihood function. The prior probability carries information about admissible model parameters without regard to the observed data set. We refer to our earlier work for a detailed development of prior probabilities and likelihood functions in the case of data that corresponds to a MD estimation of an ensemble average. The major themes developed therein are (1) use of maximum entropy distributions for prior probabilities that encode minimal parameter information from all-atom simulation through mean value constraints and (2) likelihood functions for all-atom MD data arising from the central limit theorem for convergence of sample means.

Here, we augment the likelihood function development with a treatment that takes into account more general features of the data for a chosen observable than just an estimated mean. This likelihood function is, in fact,

maximized by minimizing an estimated Kullback-Leibler divergence between the all-atom and CG distributions of the observable. To show this result, we define the observable, φ, and assume that values of this observable can be calculated from any configuration of beads in a given, fixed CG representation. There are thus two different ways to define a distribution on the possible values of the observable: (1) through a given CG potential energy function, U_{CG} and (2) through the all-atom potential energy along with a mapping, $\mathcal{M}$, that takes all-atom configurations into the CG representation. Both ways determine ensemble dependent probabilities for the observable values, but they do so from different statistical mechanical bases. As such, it is reasonable to compare the two distributions with the intuitive goal of selecting a CG parameter set such that the observable distribution arising from the CG potential best matches that resulting from the all-atom potential. We now label these distributions according to their probability density functions (PDFs), $\varrho_{CG}(\varphi \,|\, \boldsymbol{\theta}, T)$ and $\varrho_{AA}(\varphi \,|\, \mathcal{M}, T)$, where we have conditioned on the CG model parameters and the CG mapping as well as the ensemble temperature, T. The "closeness" of these two distributions can be measured using the Kullback-Leibler divergence which we express as a difference between two expectations with respect to ϱ_{AA},

$$D_{KL}(\boldsymbol{\theta}) = \langle \log \varrho_{CG}(\cdot \,|\, \boldsymbol{\theta}) \rangle_{AA} \, - \, \langle \log \varrho_{AA} \rangle_{AA} \, . \tag{5.2}$$

As both quantities on the right side of eq. 5.2, which are actually entropies, involve computation of average values, they are amenable to estimation by way of MD simulation. Using the machinery of our previous work, we find a

sampling PDF for estimates of the cross-entropy term, $\langle \log \varrho_{CG}(\cdot \,|\, \boldsymbol{\theta}) \rangle_{AA}$, that are computed from n all-atom MD samples,

$$\varrho(d \,|\, \boldsymbol{\theta}, n) = \frac{1}{2} \sqrt{\frac{n}{s_n^2(\boldsymbol{\theta})}} \exp\left\{ -\sqrt{\frac{n}{s_n^2(\boldsymbol{\theta})}} \left| d - \widehat{H}_n(\boldsymbol{\theta}) \right| \right\}. \tag{5.3}$$

This result follows from the central limit theorem for convergence of sample means and an integration over unknown hyperparameters. In eq. 5.3, $\widehat{H}_n$ is the MD estimate of the cross-entropy,

$$\widehat{H}_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log\, \varrho_{CG}\left(\varphi_{AA,i} \,|\, \boldsymbol{\theta}\right), \tag{5.4}$$

where $\{\varphi_{AA,i}\}_{i=1}^n$ are all-atom MD samples of φ. The quantity s_n^2/n, where s_n^2 is the variance of the samples, $\{\log\, \varrho_{CG}\left(\varphi_{AA,i}\right)\}_{i=1}^n$, is a measure of the uncertainty inherent in the cross-entropy calculation. If the data, d, in eq. 5.3 is defined to be an MD estimate of the all-atom entropy,

$$d = -\sum_{i=1}^{n} \log \varrho_{AA}\left(\varphi_{AA,i} \,|\, \mathcal{M}\right), \tag{5.5}$$

then it follows from eq. 5.2 and the Gibbs inequality that the misfit term in the resulting likelihood function is an MD estimate of D_{KL},

$$\left| d - \widehat{H}_n(\boldsymbol{\theta}) \right| = \widehat{D}_{KL,n}(\boldsymbol{\theta}). \tag{5.6}$$

It is important to note in the above development that, in practice, one does not know the exact densities for ϱ_{AA} and ϱ_{CG}; these densities are themselves estimated from MD simulation. Thus we actually use the estimates, $\widehat{\varrho}_{AA,n}$ and $\widehat{\varrho}_{CG,m}$, where n and m denote numbers of all-atom and CG

MD samples used to calculate the density estimates. In particular, we utilize the algorithm of Wang et al. to estimate D_{KL} via k-nearest neighbor approximations of the probability densities [107]. A notable consequence of estimating the probability densities is higher uncertainty in the sampling distribution. We include this uncertainty in an approximate sense through the s^2 parameter in the likelihood. Instead of s_n^2, we calculate s_{nm}^2 which is the variance of the samples, $\{\log \widehat{\varrho}_{CG,m}(\varphi_{AA,i})\}_{i=1}^n$. In this paper, we consider a one-dimensional, bounded observable wherein probability densities can be accurately approximated for D_{KL} estimation. Here, the preceding treatment of density uncertainty is adequate for our purposes; however, in more complex cases, the potentially large amount of uncertainty contributed to the likelihood function by density estimation may require a more careful treatment.

Parameter calibration via Bayes' formula necessitates finding the posterior parameter distribution from a specified prior and likelihood function. However, the posterior distribution does not on its own specify which parameter value is the "best" for the model at hand. The latter goal is often accomplished by posing a mapping from pairs of parameter vectors into the real numbers called a "loss function" and finding the parameter values that minimize the expectation of the loss function under the posterior probability. Denoting the loss function by $L(\boldsymbol{\theta}, \boldsymbol{\lambda})$, this procedure is specified by the expression,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \int d\boldsymbol{\theta}\, L(\boldsymbol{\theta}, \boldsymbol{\lambda}) \varrho(\boldsymbol{\theta} \,|\, \boldsymbol{d}). \tag{5.7}$$

The intuition underlying the loss function is that it measures the cost associated with using a sub-optimal parameter vector, thus quantifying the notion of a "best" set of parameters. For the remaining calculations in this paper, we propose the simple loss function,

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \|\boldsymbol{\theta} - \boldsymbol{\lambda}\|, \tag{5.8}$$

which is the ℓ^2, or euclidean, distance between $\boldsymbol{\theta}$ and the optimal candidate, $\boldsymbol{\lambda}$. As noted in previous work, a closed form expression for the posterior PDF is not generally available, and hence, we resort to gathering posterior samples, $\{\boldsymbol{\theta}_i\}_{i=1}^{N_p}$, via a Markov chain Monte Carlo (MCMC) algorithm. The optimal parameter vector is thus estimated by,

$$\widehat{\boldsymbol{\theta}} \approx \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N_p} \|\boldsymbol{\theta}_i - \boldsymbol{\lambda}\| \right\}. \tag{5.9}$$

The parameter vector associated with eq. 5.9 is, in fact, the geometric median of the N_p posterior samples and can be found from iterative techniques such as Weiszfeld's algorithm [51].

5.3 All-Atom and Coarse-Grained Models of Thiophene Polymers

The alkyl-thiophene variety we consider here consists of a five sided thiophene ring with an alkane side-chain originating from the 3 position in the carbon ring. Figure 5.2 shows a skeletal representation of 3-3'-polyhexylthiophene, also known as P3HT. Atomistic models of polythiophene are a topic of contem-

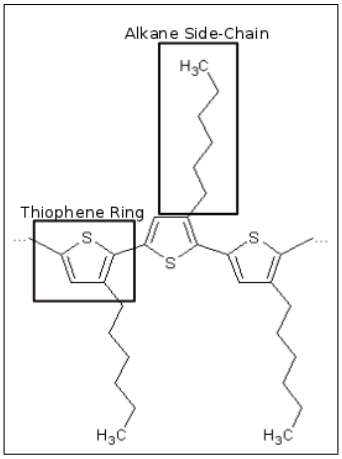

Figure 5.2: Atomic constituents and bond topology of 3-3'-polyhexylthiophene

porary research as stock atomistic potentials such as CHARMM and OPLS fail to describe relevant physics related to the distribution of electrical charge on the thiophene rings and the potential energy associated with twisting around the inter-monomer dihedral angle. Thus, it is necessary to consider atomistic potentials developed from first principle quantum chemical calculations that have been carried out specifically for alkylthiophenes. For the part of the potential governing the inter-monomer dihedral, as well as the deformations involving two and three bonded atoms, we use the model of DuBay et al. [21] due to their relatively comprehensive study of the polythiophene torsional landscape in the presence of alkane side-chains of varying length. Of note in their findings is a sizable energetic penalty for planar conformations due to steric hindrance from alkane side chains. For coulombic point charges we take values from Moreno et al. [68]. The remainder of the potential, which comprises side-chain terms, is taken from stock alkane parameterizations in

OPLS-2005 [6]. Since the DuBay model itself is posed as a modification to the OPLS parameterization, this choice for side-chain parameters is a reasonable one and is, in fact, the route they take as well. We have implemented the described atomistic potential in the LAMMPS MD simulation program [82] and, hence, this potential forms the basis from which we calibrate our CG potential energy functions.

We now take up the description of a suitable CG representation of 3-3' polyethylthiophene. The choice of ethyl side chains is due to the computational ease of modeling ethyl groups and, also, to the observation from the study of DuBay et al. that ethyl chains are of sufficient volume to generate the steric hindrance that so influences the conformational space of alkylthiophenes. We choose a three site per monomer CG scheme that is as coarse as possible while still resolving the ethyl side chain and the inter-monomer dihedral angle. Our coarse representation is shown in figure 5.3, with the three CG bead types indicated by letters 'S', 'R', and 'C'. To compare all-atom MD trajectories with

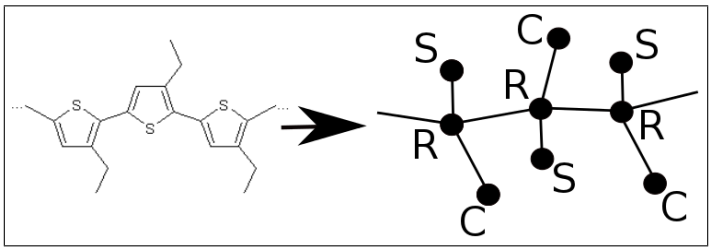

Figure 5.3: Left: P3ET atomistic model. Right: coarse representation with three sites per monomer. 'S' stands for sulfur bead, 'R' for ring and 'C' for side-chain.

those of our CG model, we define a mapping that sends the sulfur atom to the

'S' bead position, the center-of-mass (COM) of the remaining ring atoms to the 'R' position, and the COM of the ethyl side-chain to the 'C' position. We further pose a potential energy functional form consisting of harmonic spring forms for bond and angle terms, a four term OPLS style cosine series for the S-R-R-S inter-monomer dihedral angle, and a 12-6 Lennard-Jones form for interactions involving sites not connected by bonds and those separated by three or more bonds. Thus, the free parameters of the CG potential consist of stiffnesses and equilibrium positions for the harmonic spring terms, four Fourier coefficients for the dihedral, and a Lennard-Jones σ and ϵ for each of the three bead types. Mixing rules for the non-bonded terms are geometric in ϵ and arithmetic in σ: $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$, $\sigma_{ij} = (\sigma_i + \sigma_j)/2$.

5.4 CG Potential Energy Calibration

To find values for the free parameters of the CG potential, we first select a physical scenario for the calibration. We anticipate the need for numerous, sequential MD simulations in this scenario, thus we choose a setting of minimal complexity that can still deliver discriminating information. A system consisting of two interacting P3ET trimers at 300 K makes up this scenario. Prior information for the CG parameters is then gathered from all-atom MD trajectories for the dual trimer that are mapped into the described CG representation. For location parameters, that is, equilibrium bond and angle positions and Lennard-Jones σ parameters, we estimate the average values of the generalized coordinates associated with these parameters from the

MD trajectory. Spring constants for the bond and angle terms are estimated from an energy equipartition relation involving the variances of the associated generalized coordinates,

$$k_Q \approx \frac{k_b T}{2 s_Q^2}, \tag{5.10}$$

where Q denotes the generalized coordinate of interest and s_Q^2 is the sample variance for Q over the MD trajectory. The prior information values for the bonds and angles are given in tables 5.1 and 5.2. Prior information for

Bond Parameters	r_0 $(\mathring{A})$	k_r $(kcal/mol \cdot \mathring{A}^2)$
S-R	1.825	88.6
R-R	4.094	22.52
R-C	2.902	37.0

Table 5.1: Prior information for CG bonds: $\frac{1}{2}k_r(r - r_0)^2$

Angle Parameters	α_0 (deg)	k_α $(kcal/mol \cdot rad^2)$
S-R-C	141.7	14.0
S-R-R	69.0	19.44
R-R-R	131.2	7.34
C-R-R	79.6	11.24

Table 5.2: Prior information for CG angles: $\frac{1}{2}k_\alpha(\alpha - \alpha_0)^2$

Lennard-Jones ϵ parameters is found by observing average interaction energies between associated beads belonging to distinct trimers; table 5.3 lists the values for this prior information. We take a different approach for the four dihedral parameters associated with the S-R-R-S dihedral,

$$U_{SRRS}(\phi) = \sum_{k=1}^{4} \frac{V_k}{2} \left(1 + (-1)^{k-1} \cos k\phi\right). \tag{5.11}$$

Lennard-Jones Parameters	σ ($\mathring{A}$)	ϵ ($kcal/mol$)
S	3.64	0.25
R	5.99	0.21
C	6.47	0.11

Table 5.3: Prior information for CG non-bonded parameters: 12-6 Lennard Jones

Here, we take prior information directly from the quantum chemical calculations of DuBay et al. In particular, we note an energy barrier of 10 kcal/mol separating the lowest energy torsion state from the cis-planar state[1]. In the all-atom model, a non-bonded steric hindrance between the hydrogen connected to the 4 position on the thiophene ring and an adjacent side-chain accounts for the majority of this barrier. As the CG model lacks a representation of this hydrogen, we seek to build this cis-planar barrier into the CG dihedral potential instead of the non-bonded part of the potential. Setting restrictions on the V_1 parameter gives control over the barrier, since it is peaked only in the cis-planar state. Thus, we insist that the prior for V_1 have a mean constrained to 10 kcal/mol, in line with the calculation of DuBay et al., and assign zero probability to negative values. This is the extent of the dihedral prior information we use.

As in the previous work, we quantify the prior information with mean constrained, maximum entropy distributions on the parameters we include in the Bayesian inversion. Presently, this includes the dihedral coefficient, V_1,

[1]In this case, the cis-planar state is achieved when the dihedral angle is zero. For a dimer in this state, both alkyl side-chains are on the same side of the thiophene rings.

the C-R-R angle parameters, and the Lennard-Jones parameters. The basic functional form for these PDFs is that of a gamma distribution. Uniform priors between -20 and 20 kcal/mol are used for V_2, V_3, and V_4, reflecting prior ignorance over the dihedral energy scale. We do not expect the bond parameters, nor the S-R-C, S-R-R, and R-R-R angle parameters, to influence the final results in an interesting way, so we simply set them to their nominal values in the prior information and exclude them from the inversion procedure. This is equivalent to posing strong priors that approach delta functions centered on the nominal values. The C-R-R angle parameters are included in the inversion because they influence the hindrance effect between the side-chain bead and beads on adjacent monomers, an effect that impacts the torsional conformations accessible to the polymer as a whole. With all of these issues taken into account, the parameter space for the calibration is defined by the 12 dimensional domain implied by the prior PDFs.

It remains to specify the observables providing data to the parameter calibration. Three observables are used in total. We choose two that give information on the aggregation properties of P3ET: the average distance between the trimer COMs and the average interaction energy between them. The third observable, anticipated in the previous section's likelihood function development, is the entropy of the inter-monomer dihedral angle distribution. We assume logical independence of the data sets associated with these three observables so that the likelihood function factors into a simple product of one observable likelihoods. The likelihoods associated with the first two observ-

ables are derived in earlier work and have the form of Laplace distributions. The third follows from eq. 5.3 and is given by,

$$\varrho(d\,|\,\boldsymbol{\theta}, n, m) = \frac{1}{2}\sqrt{\frac{n}{s^2_{nm}(\boldsymbol{\theta})}}\,\exp\left\{-\sqrt{\frac{n}{s^2_{nm}(\boldsymbol{\theta})}}\,\widehat{D}_{KL,nm}(\boldsymbol{\theta})\right\}, \tag{5.12}$$

where the presence of m indicates the density estimation originating from CG MD samples.

The parameter posterior is sampled via a MCMC method implemented in the QUESO code [83]. All-atom data for the first two observables consists of 75 samples each of average trimer separation and interaction energy taken from 1 ns trajectories. For the dihedral angle PDF estimate, 2400 dihedral angle samples are taken from a 10 ns all-atom trajectory. Evaluation of the likelihood function further requires MD estimates from the CG model. In practice, this results in MD simulation undertaken at every Markov chain transition. Average trimer separation, interaction energy, dihedral angle PDF, as well as the uncertainties associated with these observables, are estimated from 250 ps of CG MD simulation to complete the evaluation of likelihood. One dimensional posterior estimates for each parameter are shown in figure 5.4. We note here that the full 12 dimensional posterior includes covariances among the parameters that are not possible to show using 1-D representations, but these covariances are nonetheless present in the MCMC samples. The Bayes' estimate of the CG potential energy parameter values is then calculated by estimating the geometric median of the posterior samples. We have insured in this median calculation that enough MCMC samples have been gathered to give a stable estimate that does not change appreciably with additional

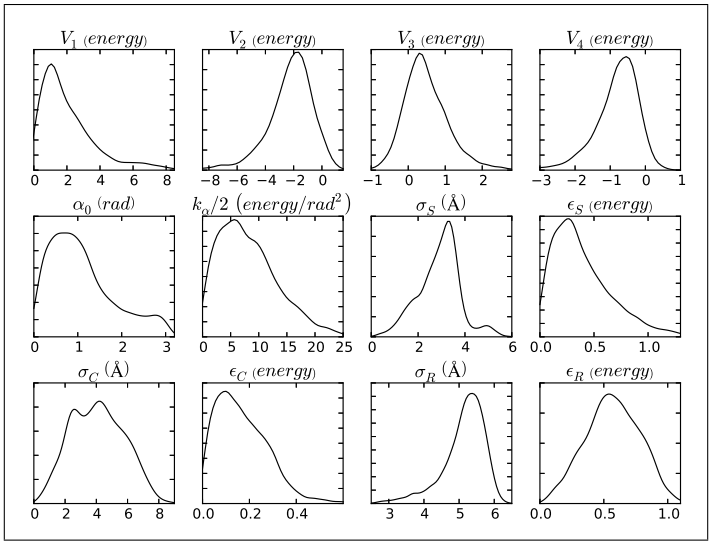

Figure 5.4: 1-D parameter posteriors estimated from 46,500 samples gathered from MCMC. Energy units are kcal/mol.

posterior samples. Table 5.4 displays these values for the parameters included in the calibration; omitted bond and angle parameters remain identical to the prior estimates. With the CG potential energy specified, MD simulation can be carried out to predict values of important observables in the reduced representation. Figure 5.5 shows a comparison of the cosine of the inter-monomer dihedral angle PDFs from all-atom and CG Bayes MD for the dual trimer system comprising the calibration scenario. The Kullback-Leibler divergence between all-atom and CG dihedral cosine PDFs is small, as expected, estimated from trapezoid rule quadrature to be about 0.02. In the following section, we consider validation tests of the Bayes' estimated CG potential that involve a more complicated physical scenario.

Dihedrals	V_1	V_2	V_3	V_4 $(kcal/mol)$
S-R-R-S	1.92	-2.18	0.421	-0.775

Angles	α_0 (deg)	k_α $(kcal/mol \cdot rad^2)$
C-R-R	62.92	14.6

LJ 12-6	σ $(\mathring{A})$	ϵ $(kcal/mol)$
S	2.87	0.41
R	5.13	0.52
C	4.30	0.19

Table 5.4: Bayes' posterior estimate of CG potential energy parameters

5.5 CG Model Validation Tests

We now subject the Bayes' estimated CG potential energy to validation tests that push the model to make predictions outside of the calibration realm. If this CG model is to be realistically used to accelerate MD investigations of polythiophene aggregation properties, there must be some such test of predictive capacity. Thus, we consider a system of four interacting decamer chains in thermal equilibrium at 300 K. In this validation scenario, we pose four quantities of interest: radius-of-gyration and end-end distance of single polymer chains, minimum distances between thiophene rings belonging to different polymer chains, and the inter-monomer dihedral angle distribution. The minimal ring distance for the ith thiophene ring is calculated according to,

$$d_{min,i} = \min \left\{ \|r_i - r_j\|_2 \colon c(j) \neq c(i) \right\}, \tag{5.13}$$

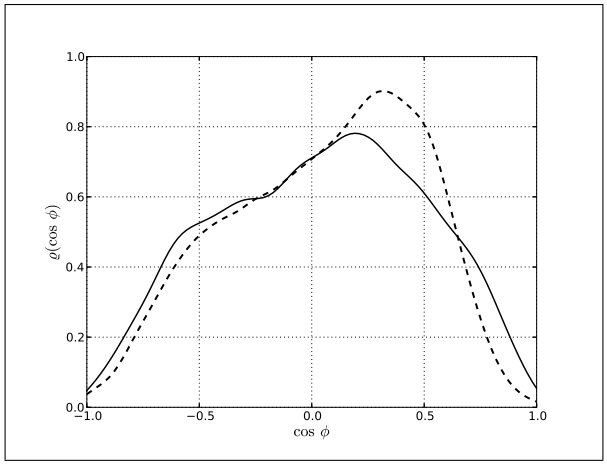

Figure 5.5: Kernel density estimated inter-monomer dihedral angle cosine PDFs for the dual trimer calibration scenario. The Solid line is from all-atom MD, the dashed line is from CG MD with the Bayes estimated potential. $D_{KL}(\varrho_{AA}\|\varrho_{CG}) \approx 0.02$.

where r_i and r_j are the position vectors of 'R' beads in the CG representation and $c(\cdot)$ is an integer representing the chain that contains a given thiophene ring.

In order to make quantitative comparisons, we run the validation scenario with the all-atom model and with the CG Bayes model, calculating the QOIs from 10 ns MD trajectories in each model. We simulate both systems using replica exchange dynamics to bolster sampling of the phase spaces; replica temperatures are set according to a geometric spacing rule-of-thumb [105]. A summary of the all-atom, CG comparisons is displayed in table 5.5. Each entry in the table is an estimate of a mean from MD simulation. The all-atom estimates are calculated by first mapping the MD trajectory to the CG representation. Comparisons of radius-of-gyration and end-end distance

QOI	All-atom ($\mathring{A}$)	Coarse-grained ($\mathring{A}$)
$\widehat{R}_g$	7.15 ± 0.1	6.72 ± 0.1
$\widehat{R}_{ee}$	15.1 ± 0.64	14.1 ± 0.64
$\widehat{d}_{min}$	5.21	5.58

Table 5.5: All-atom and CG QOI calculations from MD for average radius-of-gyration, end-end distance, and minimal thiophene ring distance in the validation scenario. Reported uncertainties are at the 95% confidence level. Uncertainties below 0.1 $\mathring{A}$ are not reported.

indicate relative agreement between all-atom and CG ensembles for conformational properties at the polymer chain scale, with hairpin-like shapes being the most common conformational feature in both cases. In the case of $\widehat{d}_{min}$, which is an average over all rings and all MD samples in a trajectory, we note that in both all-atom and CG cases, 80% of the samples making up the estimates are between 4 and 6 $\mathring{A}$. This observation motivates the statement that the polymer packing behavior in the CG setting is similar to that in the all-atom model, albeit with a slightly larger ring-ring separation on average. It is also interesting to examine the angle between normal vectors for thiophene ring pairs that satisfy the d_{min} condition in eq 5.13. In the all-atom case, it is straightforward to define a vector normal to the plane of a thiophene ring; in the CG representation, we estimate the plane normal for ring i from the $\overrightarrow{R_iS_i}$ vector and the $\overrightarrow{R_iR_{i+1}}$ vector, except at the right edge where $\overrightarrow{R_iR_{i-1}}$ is used. The contour plot in figure 5.6 shows probability density estimates for $\cos\beta$, the cosine of the angle between plane normals corresponding to thiophene rings satisfying the d_{min} condition, versus the minimal ring pair distance. These

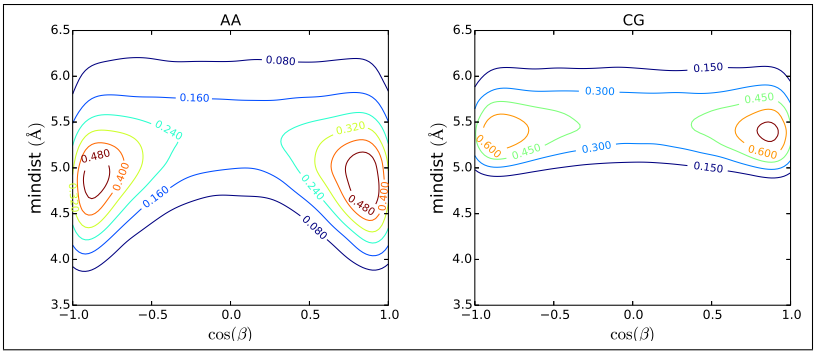

Figure 5.6: Kernel density estimates of the $\cos\beta$ vs. d_{min} joint PDF for all-atom and CG cases.

plots show that in both all-atom and CG cases, the minimal ring pairs tend to stack on top of each other in relative planar or anti-planar ($\cos\beta \approx \pm 1$) configurations.

Figure 5.7 shows a comparison of the S-R-R-S dihedral angle PDF from all-atom and CG MD. We observe that the CG potential tends to concentrate the dihedral probability density closer to the peak of the PDF, in the cis-distored state, while the all-atom distribution is more spread out, but also peaked in the cis-distorted state. As a consequence, the CG potential underestimates the probabilities of cis-planar and trans-planar conformations; the approximate probability of the inter-monomer dihedral being within 30 degrees of either planar state is 4% in the all-atom model and 1.5% in the CG model. Despite these issues, general agreement between the two distributions is still intact, with an estimated Kullback-Leibler divergence of 0.06. Another, more intuitive way to quantify the "degree of disagreement" between the two distributions is with the total variation distance, d_{TV}. This metric computes

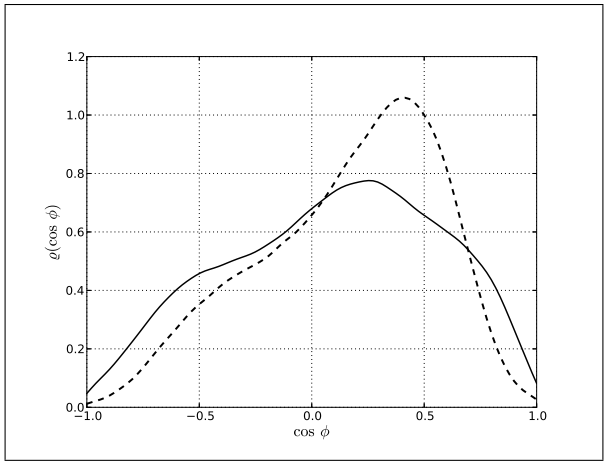

Figure 5.7: Kernel density estimated inter-monomer dihedral angle cosine PDFs for the four decamer validation scenario. The Solid line is from all-atom MD, the dashed line is from CG MD with the Bayes estimated potential. $D_{KL}(\varrho_{AA}\|\varrho_{CG}) \approx 0.06$

the largest possible probability difference that the two distributions can assign the same event and is, in this case, computed according to,

$$d_{TV}\left(\varrho_{AA}, \varrho_{CG}\right) = \frac{1}{2}\int_{-1}^{1} d(\cos\phi)\,\left|\varrho_{AA}\left(\cos\phi\,|\,\mathcal{M}\right) - \varrho_{CG}\left(\cos\phi\,|\,\widehat{\boldsymbol{\theta}}\right)\right|, \quad (5.14)$$

where we have indicated conditioning on the CG mapping, $\mathcal{M}$, and the optimal CG potential energy parameters, $\widehat{\boldsymbol{\theta}}$. For the PDFs in fig. 5.7, $d_{TV} \approx 0.126$. Thus, the greatest difference in probability assignment that the distributions can make is about 12.6%. These observations give us quantitative motivation to state that the equilibrium conformations sampled by the CG model are similar to those sampled in the all-atom model, with some minor discrepancies where noted. We finally point out that the CG MD simulation of the validation scenario required fewer CPU hours, by a factor of 36, than the equivalent

all-atom MD on our parallel processing workstations, thus demonstrating an obvious benefit of using the reduced model.

5.6 Discussion

We begin the section by noting that coarse-grained representations and potentials for varieties of polythiophene have been previously posed and calibrated in the works of Lee [54] and Huang [30, 91]. Both of these groups build models for 3-polyhexylthiophene (P3HT) that coarse-grain thiophene rings into single sites, thus their models lack the resolution to precisely define an inter-monomer dihedral angle. The Huang group is able to approximately define the inter-monomer dihedral using side-chain sites on adjacent monomer units in the dihedral definition. However, the atomistic torsional energy profile that is used to parameterize their CG model indicates potential energy minima at planar conformations, thus conflicting sharply with the source of atomistic information for our CG model. This planar minima conclusion seems to be based on density functional theory (DFT) calculations undertaken by Darling and Sternberg in 2009 [19]. DuBay et al. address this 2009 work as well, concluding that the planar minima found in that case are, in fact, local minima in the torsional energy landscape and that the global minima involve distorted dihedral angles. Due to the more current information in the atomistic model of DuBay et al., on which our CG model is based, we regard our model as having a more accurate representation than the Huang model for the inter-monomer dihedral angle potential. The Huang model is also rela-

tively complex in comparison, using anharmonic spring potentials for bonds and angles which require the specification of up to four parameters for each interaction type. Additionally, the non-bonded potentials therein are gleamed from Boltzmann inversion and are hence specified by tabulated energies at equally spaced intervals of 0.1 $\mathring{A}$. Our model is able to describe the major features of the polythiophene conformational space with an order of magnitude fewer number of parameters; furthermore, our CG potential energy consists entirely of functional forms already implemented in most major MD codes, making it easy to physically interpret, report, and share with other people studying thiophene polymers.

It is instructive, at this point, to note some key differences between computing optimal CG potential energy parameters via Bayes' estimation, as we have done in this work, and through other established algorithms. In particular, the popular methods of Boltzmann inversion, multi-scale coarse graining/force matching, and relative entropy minimization present some interesting contrasts. First and foremost, the goal all-atom quantity to reproduce using these methods is the potential of mean force (PMF) that arises from Boltzmann-weighted averaging of the net all-atom forces on CG beads. The average therein takes place over the portion of the all-atom phase space that conserves CG bead configuration. This PMF, if it can be found, is the best possible CG potential energy function, from the standpoint of statistical

mechanics, because it satisfies the relation,

$$\frac{1}{Z_{PMF}} \exp\left\{-\frac{V_{PMF}(\boldsymbol{R})}{k_b T}\right\} = \frac{1}{Z_{AA}} \int d\boldsymbol{r} \, \exp\left\{-\frac{U_{AA}(\boldsymbol{r})}{k_b T}\right\} \, \delta\left(\mathcal{M}(\boldsymbol{r}) - \boldsymbol{R}\right), \tag{5.15}$$

where $\boldsymbol{R}$ and $\boldsymbol{r}$ represent CG and all-atom configurations, the Z quantities are relevant partition functions, and $\delta(\cdot)$ selects the all-atom configurations conserving CG configuration $\boldsymbol{R}$. This relation implies that coarse-grained simulations undertaken with V_{PMF} as the potential energy function are guaranteed to reproduce any configurational observable from the all-atom model, modulo the all-atom-to-CG mapping. W. Noid has proved this identification of the PMF with the "consistent" CG potential energy. Unfortunately, V_{PMF} is generally an unknown function of all the CG coordinates. In practice, force-matching, relative entropy minimization, and Boltzmann inversion construct iterative approximations to the PMF in vector spaces that generally assume additive interactions of functions of single generalized coordinates. Hence, the exact, many-body PMF is not generally in the asymptotic span of the of the function spaces used. Consequently, differences between general ensemble averages in the all-atom and resulting CG models are not necessarily bounded by any constants related to the convergence criteria of the potential energy optimization algorithm. Put another way, finding the potential energy in a space of additive, pairwise functions that minimizes a residual involving mean forces, relative entropy, or select pair correlation functions does not guarantee that the resulting CG potential energy will reproduce the all-atom value for a general observable. A rigorous determination for general observables can-

not be made unless the potential energy approximation spaces are enlarged to contain the true PMF, but this is, as yet, not a practical possibility.

The Bayes estimation scheme introduced here and in previous work is decidedly different in philosophy and motivation than the PMF methods mentioned. We make no attempt, nor any claim, to produce a method that converges to the exact PMF arising from a CG mapping. Rather, the Bayes estimation procedure addresses potential shortcomings of PMF methods by targeting observables that are physically relevant, in a practical sense, to designers and end-users of a particular CG model. In this setting, the method seeks to quantify the degree to which CG potential energy parameters are determined by constraining the values of a small collection of ensemble averages. The resulting posterior probability distributions allow us not only to quantify uncertainty in parameter estimates, but also to iteratively build more comprehensive posterior representations of parameter knowledge given new varieties of observable information. We've shown in this paper for a polythiophene model, and in previous work for a heptane model, that the Bayes estimation procedure can produce CG models with predictive capacity in reference to QOIs that are informed to some degree by the calibration observables. It is important to note that the success of these examples depended on the ability of the method to find sets of CG parameter vectors such that the calibration observable values could all be accurately reproduced in the CG setting. If the CG potential energy space is not rich enough to reproduce these features, then it must be expanded in some way to include them. Here, PMF methods may

be able to provide valuable prior information on the nature of the potential energy function space enrichment that is necessary. Due to the substantial difficulties associated with CG modeling of complex chemical systems in biophysical and materials science, it is likely that the most useful CG models will be posed and calibrated using a hybrid approach that references the consistency principles embodied in Noid's PMF theorem while still taking into account the practical importance of optimizing potential energy parameters in reference to observables that are relevant to the original design goals of the CG model. We have pursued the latter, with loose consistency imposed by maximum entropy priors defined from mean value constraints. Many subsequent improvements are likely as the different perspectives on the coarse-graining problem are brought to bear.

Appendix

The objective of this appendix is to show that, given certain maximum entropy priors, a gaussian sampling distribution admits a Laplace distribution upon integration over its hyper-parameters. We begin with the gaussian form likelihood that results from applying the central limit theorem to a sample mean computed from n independent samples,

$$\varrho(d \,|\, n, \boldsymbol{\theta}, \mu, \sigma^2) = \sqrt{\frac{n}{2\pi\,\sigma^2}} \exp\left\{ -\frac{n}{2} \frac{(d-\mu)^2}{\sigma^2} \right\}, \tag{A.1}$$

where μ and σ are the true mean and variance for the sampling PDF. In the CG calibration setting, these values are implicitly dependent on the CG potential energy parameters, $\boldsymbol{\theta}$. Consequently, the true mean and variance are unknown and must be estimated from MD samples gathered from the CG model. We eliminate these values, now referred to as hyper-parameters, from the likelihood expression by applying appropriate prior probabilities and integrating over the parameter domains. We now define the sampling PDF according to,

$$\varrho(d \,|\, n, m, \boldsymbol{\theta}) \equiv \int_0^\infty d\sigma^2 \int_{-\infty}^\infty d\mu\, \varrho(d \,|\, n, \boldsymbol{\theta}, \mu, \sigma^2)\, \varrho(\mu, \sigma^2 \,|\, m, \boldsymbol{\theta}), \tag{A.2}$$

where we now condition on the number m of CG MD samples used to estimate the hyper-parameters. The prior on the right side of eq. A.2 is then

factored according to, $\varrho(\mu, \sigma^2 \,|\, m, \boldsymbol{\theta}) = \varrho(\mu \,|\, m, \boldsymbol{\theta}, \sigma^2)\, \varrho(\sigma^2 \,|\, m, \boldsymbol{\theta})$, where the single parameter priors are assigned from maximum entropy requirements,

$$\begin{aligned} \varrho\left(\mu \,|\, m, \boldsymbol{\theta}, \sigma^2\right) &= \sqrt{\frac{m}{2\pi\sigma^2}} \exp\left\{ -\frac{m}{2} \frac{\left(\mu - \widehat{\mu}_m(\boldsymbol{\theta})\right)^2}{\sigma^2} \right\}, \\ \varrho\left(\sigma^2 \,|\, m, \boldsymbol{\theta}\right) &= \frac{1}{\widehat{\sigma}_m^2(\boldsymbol{\theta})} \exp\left\{ -\frac{\sigma^2}{\widehat{\sigma}_m^2(\boldsymbol{\theta})} \right\}. \end{aligned} \tag{A.3}$$

The quantities, $\widehat{\mu}_m$ and $\widehat{\sigma}_m^2$, are the MD estimates computed from CG simulation. Expanding terms and completing the square in the first integral yields,

$$\begin{aligned} &\frac{\sqrt{nm}}{2\pi\sigma^2} \int_{-\infty}^{\infty} d\mu \, \exp\left\{ -\frac{n}{2} \frac{(d-\mu)^2}{\sigma^2} - \frac{m}{2} \frac{(\mu - \widehat{\mu}_m)^2}{\sigma^2} \right\} \\ =&\frac{\sqrt{nm}}{2\pi\sigma^2} \exp\left\{ -\frac{nm\,(d-\widehat{\mu}_m)^2}{2(n+m)\sigma^2} \right\} \int_{-\infty}^{\infty} d\mu \, \exp\left\{ -\frac{n+m}{2\sigma^2} \left(\mu - \frac{nd + m\widehat{\mu}_m}{n+m} \right)^2 \right\} \\ =&\sqrt{\frac{nm}{2(n+m)\pi\sigma^2}} \exp\left\{ -\frac{nm\,(d-\widehat{\mu}_m)^2}{2(n+m)\sigma^2} \right\}. \end{aligned} \tag{A.4}$$

This last line of (A.4) indicates that the integral over μ produces a new gaussian PDF with mean $\widehat{\mu}_m$ and variance $(n^{-1} + m^{-1})\sigma^2$.

We now show, by way of a duality argument, that the final integral over the unknown variance implies a random variable with a Laplace distribution. That is, we show that the characteristic function, or the Fourier transform, of the integral is equivalent to that of a Laplace PDF. To simplify the expressions, we let $\kappa^2 = n^{-1} + m^{-1}$ and $x = d$. Furthermore, we denote the gaussian PDF of mean zero and variance σ^2 by $G(x; \sigma^2)$. We begin the calculation by noting that the following iterated integral has a clearly finite value,

$$\int_0^{\infty} d\sigma^2 \int_{-\infty}^{\infty} dx \, \left| e^{itx}\, G(x - \widehat{\mu}_m; \kappa^2\sigma^2)\, \varrho\left(\sigma^2 \,|\, m, \boldsymbol{\theta}\right) \right| = 1 < \infty. \tag{A.5}$$

It follows that the conditions of the Fubini theorem for iterated integrals are satisfied, and hence we can compute the characteristic function by reversing the integration order,

$$\begin{aligned}\varphi(t) &\equiv \int_{-\infty}^{\infty} dx\, e^{itx} \int_0^{\infty} d\sigma^2\, G(x-\widehat{\mu}_m;\kappa^2\sigma^2)\, \varrho\left(\sigma^2 \,|\, m,\boldsymbol{\theta}\right) \\ &= \int_0^{\infty} d\sigma^2\, \varrho\left(\sigma^2 \,|\, m,\boldsymbol{\theta}\right) \int_{-\infty}^{\infty} dx\, e^{itx}\, G(x-\widehat{\mu}_m;\kappa^2\sigma^2) \\ &= \int_0^{\infty} d\sigma^2\, \varrho\left(\sigma^2 \,|\, m,\boldsymbol{\theta}\right)\, \exp\left\{ i\,\widehat{\mu}_m t - \frac{1}{2}\kappa^2\sigma^2 t^2 \right\},\end{aligned} \tag{A.6}$$

where, in the last line, we have substituted the known characteristic function for a gaussian PDF. The final result is obtained by carrying out the integration,

$$\begin{aligned}\varphi(t) =& \frac{e^{i\,\widehat{\mu}_m t}}{\widehat{\sigma}_m^2} \int_0^{\infty} d\sigma^2\, \exp\left\{ -\left(\frac{1}{\widehat{\sigma}_m^2} + \frac{1}{2}\kappa^2 t^2 \right) \sigma^2 \right\} \\ =& \frac{e^{i\,\widehat{\mu}_m t}}{1 + \frac{1}{2}\kappa^2\,\widehat{\sigma}_m^2\, t^2}.\end{aligned} \tag{A.7}$$

The computed characteristic function is the Fourier dual of a Laplace PDF with mean $\widehat{\mu}_m$ and variance $\kappa^2\,\widehat{\sigma}_m^2$. Thus, we find,

$$\varrho(d \,|\, n, m, \boldsymbol{\theta}) = \frac{1}{2\,\gamma_{mn}(\boldsymbol{\theta})}\, \exp\left\{ -\frac{|d - \widehat{\mu}_m(\boldsymbol{\theta})\,|}{\gamma_{mn}(\boldsymbol{\theta})} \right\}, \tag{A.8}$$

with $\gamma_{mn}(\boldsymbol{\theta}) = \widehat{\sigma}_m(\boldsymbol{\theta})\sqrt{\frac{1}{2}\,\left(\frac{1}{m} + \frac{1}{n}\right)}$. The identification of the hyper-parameter integral with a Laplace PDF is thus established.

Bibliography

[1] M.L. Adams et al. *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verication, Validation, and Uncertainty Quantification*. National Academic Press, Washington, D.C., 2012.

[2] J. Aldrich. Doing least squares: perspectives from Gauss and Yule. *International Statistical Review*, 66(1):61–81, 1998.

[3] I. Babuška and J. T. Oden. Verification and validation in computational engineering and science: basic concepts. *Computer Methods in Applied Mechanics and Engineering*, 193(36-38):4057–4066, September 2004.

[4] I. Babuška and J. T. Oden. The reliability of computer predictions: can they be trusted? *International Journal of Numerical Analysis and Modeling*, 1(1):1–18, 2005.

[5] N. Banerji, S. Cowan, M. Leclerc, E. Vauthey, and A. J. Heeger. Exciton formation, relaxation, and decay in pcdtbt. *Journal of the American Chemical Society*, 132(49):17459–70, December 2010.

[6] J. L. Banks et al. Integrated modeling program, applied chemical theory (IMPACT). *The Journal of Computational Chemistry*, 26(16):1752–1780, 2005.

[7] C. H. Bennett. Efficient estimation of free energy differences from monte carlo data. *The Journal of Computational Physics*, 22:245–268, 1976.

[8] M. R. Betancourt and S. J. Omovie. Pairwise energies for polypeptide coarse-grained models derived from atomic force fields. *The Journal of Chemical Physics*, 130(19):195103, May 2009.

[9] R. H. Boyd and P. J. Phillips. *The Science of Polymer Molecules*. Cambridge University Press, 1996.

[10] C. J. Brabec, S. Gowrisanker, J. J. M. Halls, D. Laird, S. Jia, and S. P. Williams. Polymer-fullerene bulk-heterojunction solar cells. *Advanced materials*, 22(34):3839–56, September 2010.

[11] M. A. Brady, G. M. Su, and M. L. Chabinyc. Recent progress in the morphology of bulk heterojunction photovoltaics. *Soft Matter*, 7(23):11065, 2011.

[12] M. Breza, V. Lukes, and I. Vrabel. On the dependence of optical properties on conformational changes in oligothiophenes. I. Electron absorption spectra. *Journal of Molecular Structure. Theochem*, 572:151–160, 2001.

[13] E. Brini, V. Marcon, and N. F. A. van der Vegt. Conditional reversible work method for molecular coarse graining applications. *Physical Chemistry Chemical Physics*, 13(22):10468–74, June 2011.

[14] E. Brini and N. F. A. van der Vegt. Chemically transferable coarse-grained potentials from conditional reversible work calculations. *The Journal of Chemical Physics*, 137(15):154113, October 2012.

[15] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *The Journal of Computational Chemistry*, 4(2):187–217, 1983.

[16] S. P. Carmichael and M. S. Shell. A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly. *The Journal of Physical Chemistry B*, 116(29):8383–93, July 2012.

[17] A. Chaimovich and M. S. Shell. Coarse-graining errors and numerical optimization using a relative entropy framework. *The Journal of Chemical Physics*, 134(9):094112, March 2011.

[18] S. H. Cheung, T. A. Oliver, E. E. Prudencio, S. Prudhomme, and R. D. Moser. Bayesian uncertainty analysis with applications to turbulence modeling. *Reliability Engineering & System Safety*, 96(9):1137–1149, September 2011.

[19] S. B. Darling and M. Sternberg. Importance of side chains and backbone length in defect modeling of poly(3-alkylthiophenes). *The Journal of Physical Chemistry B*, 113(18):6215–8, May 2009.

[20] S. Difley and T. Van Voorhis. Exciton/charge-transfer electronic couplings in organic semiconductors. *Journal of Chemical Theory and Computation*, 7(3):594–601, March 2011.

[21] K. H. DuBay, M. L. Hall, T. F. Hughes, C. Wu, D. R. Reichman, and R. A. Friesner. Accurate force field development for modeling conjugated polymers. *Journal of Chemical Theory and Computation*, 8(11):4556–4569, November 2012.

[22] B. Dünweg and W. Paul. Brownian dynamics simulations without Gaussian random numbers. *International Journal of Modern Physics C*, 2(03):817–827, 1991.

[23] D. I. Eikens. *Applicability of Theoretical and Semi-empirical Models for Predicting Infinite Dilution Activity Coefficients*. PhD thesis, University of Minnesota, 1993.

[24] F. Ercolessi and J. Adams. Interatomic potentials from first-principles calculations: the force-matching method. *Europhysical Letters*, 26:583, 1994.

[25] K. Farrell and J. T. Oden. Calibration and validation of coarse-grained models of atomic systems: application to semiconductor manufacturing. *The Journal of Computational Mechanics*, 54(1):3–19, May 2014.

[26] R.A. Fisher. On the mathematical foundations of theoretical statistics.

Philosophical Transactions of the Royal Society of London, A(222):309–368, 1922.

[27] P. J. Flory. Thermodynamics of high polymer solutions. *The Journal of Chemical Physics*, 10(1):51, 1942.

[28] M. G. Guenza. Theoretical models for bridging timescales in polymer dynamics. *Journal of Physics: Condensed Matter*, 20(3):033101, 2008.

[29] D. Hu, J. Yu, K. Wong, B. Bagchi, P. J. Rossky, and P. F. Barbara. Collapse of stiff conjugated polymers with chemical defects into ordered, cylindrical conformations. *Nature*, 405(6790):1030–3, June 2000.

[30] D. M. Huang, R. Faller, K. Do, and A. J. Moule. Coarse-grained computer simulations of polymer / fullerene bulk heterojunctions for organic photovoltaic applications. *Journal of Chemical Theory and Computation*, 6:526–537, 2010.

[31] D. M. Huang, A. J. Moule, and R. Faller. Characterization of polymer-fullerene mixtures for organic photovoltaics by systematically coarse-grained molecular simulations. *Fluid Phase Equilibria*, 302(1-2):21–25, March 2011.

[32] M. L. Huggins. Solutions of long chain compounds. *The Journal of Chemical Physics*, 9(5):440, 1941.

[33] S. Izvekov. Towards an understanding of many-particle effects in hydrophobic association in methane solutions. *The Journal of Chemical Physics*, 134(3):034104, January 2011.

[34] S. Izvekov, P. W. Chung, and B. M. Rice. The multiscale coarse-graining method: assessing its accuracy and introducing density dependent coarse-grain potentials. *The Journal of Chemical Physics*, 133(6):064109, 2010.

[35] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth. Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: a new method for force-matching. *The Journal of chemical physics*, 120(23):10896–913, June 2004.

[36] S. Izvekov and G. A. Voth. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry. B*, 109(7):2469–73, February 2005.

[37] S. Izvekov and G. A. Voth. Multiscale coarse graining of liquid-state systems. *The Journal of Chemical Physics*, 123(13):134105, October 2005.

[38] S. Izvekov and G. A. Voth. Modeling real dynamics in the coarse-grained representation of condensed phase systems. *The Journal of Chemical Physics*, 125(15):151101, October 2006.

[39] A. E. Jailaubekov, A. P. Willard, J. R. Tritsch, W. L. Chan, N. Sai, R. Gearba, L. G. Kaake, K. J. Williams, K. Leung, P. J. Rossky, and X. Zhu. Hot charge-transfer excitons set the time limit for charge separation at donor/acceptor interfaces in organic photovoltaics. *Nature Materials*, 12(1):66–73, January 2013.

[40] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

[41] E. T. Jaynes. Gibbs vs. Boltzmann entropies. *American Journal of Physics*, 33(5):391–398, 1965.

[42] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[43] R. L. Jernigan and P. J. Flory. Distribution functions for chain molecules. *The Journal of Chemical Physics*, 50(10):4185, 1969.

[44] M. Jochum, D. Andrienko, K. Kremer, and C. Peter. Structure-based coarse-graining in liquid slabs. *The Journal of Chemical Physics*, 137(6):064102, August 2012.

[45] W. L. Jorgensen, J. D. Madura, and C. J. Swenson. Optimized intermolecular potential functions for liquid hydrocarbons. *Journal of the American Chemical Society*, 106(22):6638–6646, 1984.

[46] W. L. Jorgensen and D. S. Maxwell. Development and testing of the OPLS all-atom force field on conformational energetics and prop-

erties of organic liquids. *Journal of the American Chemical Society*, 118(15):11225–11236, 1996.

[47] H. A. Karimi Varzaneh and H. J. Qian. IBIsCO: A molecular dynamics simulation package for coarse-grained simulation. *Journal of Computational Chemistry*, 32(7):1475–87, 2011.

[48] T. Kariya and H. Kurata. *Generalized Least Squares*. Wiley, 2004.

[49] S. Kotz, T. Kozubowski, and K. Podgorski. *The Laplace Distribution and Generalizations*. Birkhäuser, 2001.

[50] P. S. Koutsourelakis and E. Bilionis. Scalable bayesian reduced-order models for simulating high-dimensional multiscale dynamical systems. *Multiscale Modeling and Simulation*, 9(1):449–485, 2011.

[51] H. W. Kuhn. A note on fermat's problem. *Mathematical Programming*, 4:98–107, 1973.

[52] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[53] C. K. Lee and C. W. Pao. Solubility of [6,6]-phenyl-c 61-butyric acid methyl ester and optimal blending ratio of bulk heterojunction polymer solar cells. *The Journal of Physical Chemistry C*, 116:12455–12461, 2012.

[54] C. K. Lee, C. W. Pao, and C. W. Chu. Multiscale molecular simulations of the nanoscale morphologies of p3ht:pcbm blends for bulk heterojunction organic photovoltaic cells. *Energy & Environmental Science*, 4(10):4124, 2011.

[55] J. Lee, K. Vandewal, S. R. Yost, M. E. Bahlke, L. Goris, M. A. Baldo, J. V. Manca, and T. Van Voorhis. Charge transfer state versus hot exciton dissociation in polymer-fullerene blended solar cells. *Journal of the American Chemical Society*, 132(34):11878–80, September 2010.

[56] N. Leonenko, L. Pronzato, and V. Savani. A class of rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182, October 2008.

[57] P. Liu, Q. Shi, H. Daumé, and G. A. Voth. A bayesian statistics approach to multiscale coarse graining. *The Journal of Chemical Physics*, 129(21):214114, 2008.

[58] C. A Lopez, A. J. Rzepiela, Alex H. de Vries, L. Dijkhuizen, P. H. Huenenberger, and S. J. Marrink. MARTINI coarse-grained force field : Extension to carbohydrates. *The Journal of Chemical Theory and Computation*, 5(12):3195–3210, 2009.

[59] A. P. Lyubartsev, M. Karttunen, I. Vattulainen, and A. Laaksonen. On coarse-graining by the inverse monte carlo method: Dissipative particle dynamics simulations made to a precise tool in soft matter modeling. *Soft Materials*, 1(1):121–137, January 2003.

[60] A. P. Lyubartsev and A. Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse monte carlo approach. *Physical Review E*, 52(4):3730–3737, 1995.

[61] A. P. Lyubartsev and A. Laaksonen. Effective potentials for ion-dna interactions. *The Journal of Chemical Physics*, 111(24):11207, 1999.

[62] S. J. Marrink, A. H. de Vries, and A. E. Mark. Coarse grained model for semiquantitative lipid simulations. *The Journal of Physical Chemistry B*, 108(2):750–760, January 2004.

[63] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. The MARTINI force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–24, July 2007.

[64] M. G. Martin and J. I. Siepmann. Predicting multicomponent phase equilibria and free energies of transfer for alkanes by molecular simulation. *The Journal of the American Chemical Society*, 7863(10):8921–8924, 1997.

[65] R. L. Mcgreevy and L. Pusztai. Reverse monte carlo simulation: A new technique for the determination of disordered structures. *Molecular Simulation*, 1(6):359–367, 1988.

[66] D. A. McQuarrie. *Statistical Mechanics*. University Science Books, 2000.

[67] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink. The MARTINI coarse-grained force field: Extension to proteins. *The Journal of Chemical Theory and Computation*, 4(5):819–834, May 2008.

[68] M. Moreno, M. Casalegno, G. Raos, S. V. Meille, and R. Po. Molecular modeling of crystalline alkylthiophene oligomers and polymers. *The Journal of Physical Chemistry B*, 114(4):1591–602, February 2010.

[69] J. W. Mullinax and W. G. Noid. Extended ensemble approach for deriving transferable coarse-grained potentials. *The Journal of Chemical Physics*, 131(10):104110, 2009.

[70] J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937.

[71] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics*, 139(9):090901, September 2013.

[72] W. G. Noid, J. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of Chemical Physics*, 128(24):244114, June 2008.

[73] W. G. Noid, P. Liu, Y. Wang, J. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth. The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *The Journal of Chemical Physics*, 128(24):244115, June 2008.

[74] W. L. Oberkampf, T. G. Trucano, and C. Hirsch. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*, 57(5):345, 2004.

[75] J. T. Oden, R. Moser, and O. Ghattas. Computer predictions with quantified uncertainty, part I. *SIAM News*, 43(9), 2010.

[76] J. T. Oden, R. Moser, and O. Ghattas. Computer predictions with quantified uncertainty, part II. *SIAM News*, 43(10):2008–2011, 2010.

[77] J. T. Oden, E. E. Prudencio, and P. T. Bauman. Virtual model validation of complex multiscale systems: Applications to nonlinear elastostatics. *Computer Methods in Applied Mechanics and Engineering*, 266:162–184, November 2013.

[78] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182, 1981.

[79] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode

analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, September 1995.

[80] M. Pickholz and G. Giupponi. Coarse grained simulations of local anesthetics encapsulated into a liposome. *The Journal of Physical Chemistry B*, 114(20):7009–15, 2010.

[81] M. S. Pinsker. *Information and information stability of random variables and processes*. Holden Day, Inc., San Francisco, CA, 1964.

[82] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *The Journal of Computational Physics*, 117:1–19, 1995.

[83] E. E. Prudencio and S. H. Cheung. Parallel adaptive multilevel sampling algorithms for the bayesian analysis of mathematical models. *International Journal for Uncertainty Quantification*, 2(3):215–237, 2012.

[84] D. Reith, M. Pütz, and F. Müller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *The Journal of Computational Chemistry*, 24(13):1624–1636, 2003.

[85] P. Roach. *Verification and Validation in Computational Science and Engineering*. Hermosa Press, Albuquerque, N.M., 1998.

[86] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832, 1956.

[87] S. M. Ross. *Introduction to Probability Models*. Academic Press, 9th edition, 2007.

[88] A. Samiotakis, D. Homouz, and M. S. Cheung. Multiscale investigation of chemical interference in proteins. *The Journal of Chemical Physics*, 132(17):175101, May 2010.

[89] M.C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A.J. Heeger, and C.J. Brabec. Design rules for donors in bulk-heterojunction solar cells: Towards 10% energy-conversion efficiency. *Advanced Materials*, 18(6):789–794, March 2006.

[90] T. Schneider and E. Stoll. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B*, 17:1302–1322, Feb 1978.

[91] K. N. Schwarz, T. W. Kee, and D. M. Huang. Coarse-grained simulations of the solution-phase self-assembly of poly(3-hexylthiophene) nanostructures. *Nanoscale*, 5(5):2017–27, March 2013.

[92] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[93] M. S. Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of Chemical Physics*, 129(14):144108, October 2008.

[94] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein. A coarse grain model for phospholipid simulations. *The Journal of Physical Chemistry B*, 105(19):4464–4470, 2001.

[95] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, P. B. Moore, and M. L. Klein. Simulations of phospholipids using a coarse grain model. *The Journal of Physical Chemistry B*, 105(40):9785–9792, October 2001.

[96] Q. Shi, S. Izvekov, and G. A. Voth. Mixed atomistic and coarse-grained molecular dynamics: simulation of a membrane-bound ion channel. *The Journal of Physical Chemistry B*, 110(31):15045–8, August 2006.

[97] W. Shinoda, R. DeVane, and M. L. Klein. Zwitterionic lipid assemblies: molecular dynamics studies of monolayers, bilayers, and vesicles using a new coarse grain force field. *The Journal of Physical Chemistry B*, 114(20):6836–49, May 2010.

[98] M. R. Shirts and J. D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, October 2008.

[99] M. R. Shirts, D. L. Mobley, J. D. Chodera, and V. S. Pande. Accurate and efficient corrections for missing dispersion interactions in molecular simulations. *The Journal of Physical Chemistry B*, 111(45):13052–63, November 2007.

[100] J. R. Silbermann, S. H. L. Klapp, M. Schoen, N. Chennamsetty, H. Bock, and K. E. Gubbins. Mesoscale modeling of complex binary fluid mixtures: towards an atomistic foundation of effective potentials. *The Journal of Chemical Physics*, 124(7):74105, February 2006.

[101] B. Silverman. *Density Estimation for Statistics and Data Analysis*. CRC Press, 1986.

[102] B. Smit, P. Hilbers, and K. Esselink. Computer simulations of a water/oil interface in the presence of micelles. *Nature*, 348:624, 1990.

[103] B. G. Sumpter and V. Meunier. Can computational approaches aid in untangling the inherent complexity of practical organic photovoltaic systems? *Journal of Polymer Science Part B: Polymer Physics*, 50(15):1071–1089, August 2012.

[104] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

[105] D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team. *GROMACS User Manual version 4.6.3*, 2013.

[106] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, and E. Darian. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *The Journal of Computational Chemistry*, 31(4):671–690, 2010.

[107] Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k -nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

[108] P. K. Weiner and P. A. Kollman. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *The Journal of Computational Chemistry*, 2(3):287–303, 1981.

[109] X. Yang, J. Loos, S. C. Veenstra, W. J. H. Verhees, M. M. Wienk, J. M Kroon, M. A. J. Michels, and R. Janssen. Nanoscale morphology of high-performance polymer solar cells. *Nano Letters*, 5(4):579–83, April 2005.

[110] J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth. Coarse-grained peptide modeling using a systematic multiscale approach. *Biophysical journal*, 92(12):4289–303, June 2007.