Calibration and Validation of Coarse-Grained Models of Atomic Systems: Application to Semiconductor Manufacturing

Kathryn Farrell and J. Tinsley Oden

Institute for Computational Engineering and Sciences
The University of Texas at Austin
oden@ices.utexas.edu, kfarrell@ices.utexas.edu

March 18, 2014

Abstract

Coarse-grained models of atomic systems, created by aggregating groups of atoms into molecules to reduce the number of degrees of freedom, have been used for decades in important scientific and technological applications. In recent years, interest in developing a more rigorous theory for coarse graining and in assessing the predictivity of coarse-grained models has arisen. In this work, Bayesian methods for the calibration and validation of coarse-grained models of atomistic systems in thermodynamic equilibrium are developed. For specificity, only configurational models of systems in canonical ensembles are considered. Among major challenges in validating coarse-grained models are 1) the development of validation processes that lead to information essential in establishing confidence in the model's ability predict key quantities of interest and 2), above all, the determination of the coarse-grained model itself; that is, the characterization of the molecular architecture, the choice of interaction potentials and thus parameters, which best fit available data. The all-atom model is treated as the "ground truth," and it provides the basis with respect to which properties of the coarse-grained model are compared. This base all-atom model is characterized by an appropriate statistical mechanics framework in this work by canonical ensembles involving only configurational energies. The all-atom model thus supplies data for Bayesian calibration and validation methods for the molecular model. To address the first challenge, we develop priors based on the maximum entropy principle and likelihood functions based on Gaussian approximations of the uncertainties in the parameter-to-observation error. To address challenge 2), we introduce the notion of model plausibilities as a means for model selection. This methodology provides a powerful approach toward constructing coarse-grained models which are most plausible for given all-atom data. We demonstrate the theory and methods through applications to representative atomic structures and we discuss extensions to the validation process for molecular models of polymer structures encountered in certain semiconductor nanomanufacturing processes. The powerful method of model plausibility as a means for selecting interaction potentials for coarse-grained models is discussed in connection with a coarse-grained hexane molecule. Discussions of how all-atom information is used to construct priors are contained in an appendix.

Keywords: Bayesian statistics, coarse graining, canonical ensemble, calibration, validation, uncertainty quantification, model selection

1 Introduction

Interest in the study of physical events at atomistic scales has dramatically expanded in recent years due to developments in high-performance computing, computational modeling and simulation, and experimental science, all pushed forward by important advances in biology, medicine, drug design, nanomanufacturing, and material science. The universally accepted approach for modeling virtually all atomic systems is to employ atomistic molecular dynamics (MD) simulations, implemented using any of several hardened MD codes, generally available to the scientific community to compute estimates of ensemble averages of key quantities of interest. However, the enormous size and complexity of atomistic MD models needed to capture events at scales prevalent in most simulations of scientific or technological importance far exceeds the capacity of today's largest supercomputers or even those envisioned decades into the future. Thus, methods for reducing the number of degrees of freedom of atomistic models to sizes manageable using MD by aggregating atoms into equivalent molecular models is viewed as a necessary approach to MD-studies of long time- and length-scale processes. These lumped or aggregated, lower-dimensional models that function on coarser spatial and temporal scales are called *coarse-grained* (CG) models and the process that produces them is called *coarse graining.*

A relatively large literature exists on various methods of deriving CG models of all-atom (AA) systems. An exhaustive review of literature on CG models of biomolecular systems that contains almost 600 references [61]. The common aim of the design of CG models is, of course, to preserve in some sense key properties of the underlying AA systems. According to [87], early CG methods appeared in the 1940s in [27] and [33], with more modern approaches designed for computer simulations exemplified in the 1990 publication [83]. In more recent times, a variety of methods have been proposed for calibrating CG models, including force-matching methods [37–40], extending the 1994 work [24]. These classes of methods are referred to by Izvekov, Voth, *et al.* as “multiscale coarse-graining methods,” and have been applied to several types of molecular systems (see, *e.g.* [35, 36, 40, 53, 81, 88]). Along these lines, the works [60, 62, 63] develop a formalism for deriving CG models that are “physically consistent” with underlying AA models when CG parameters are calibrated via force matching conditions. In another class of CG methods, referred to as iterative Boltzmann inversion and developed in [77], CG parameters are chosen to match specific probability distributions of the AA systems. The Reverse Monte Carlo method (RMC) and the Conditional Reverse Work (CRW) method described in [9, 10, 56, 57, 59] and applied to CG model calibration [55] in 2003 also represent proposed approaches for deriving CG models.

A general approach to developing and calibrating consistent CG models is the method of minimum relative entropy proposed in [80] and extended in [12, 13]. There it is argued that the CG model optimally represents the AA model when their relative entropy, measured by the Kullback-Leibler divergence between their configurational probability density functions for canonical ensembles, is a minimum. Chaimovich and Shell also describe a scheme for accelerating MD sampling used in evaluating gradients of the relative entropy.

In this paper, we develop a general Bayesian framework for constructing CG approximations of atomistic systems and for calibrating and validating CG models based on data drawn from underlying AA models. Several Bayesian-based methods for model validation have been proposed in recent years. The works [3, 7, 43, 44, 47, 48] must be mentioned. Bayesian methods for calibration and validation of multiscale models and additional references of Bayesian approaches are given in [68]. In considering coarse-grained models, however, the processes are further exacerbated by the uncertainty in the forms of the models themselves, the interaction potentials not being rigorously defined for components of the CG model. An approach toward resolving this problem using the concept of model plausibilities is discussed in Section 7. Examples of CG model calibration and validation processes for representative models of polymeric materials are given in Section 8. Applications of the calibration and validation procedures to CG models of polymeric structures encountered in semiconductor nanomanufacturing are also discussed in Section 8 along with an example of a CG model of hexane, demonstrating the use of Bayesian model plausibilities for model selection. Brief concluding comments are given in Section 9. The use of AA data to construct priors is discussed in an appendix.

Our general goal is not only to lay down principles for constructing meaningful CG models that preserve key properties of the AA models on which they are based, but also to develop meaningful calibration and validation processes for CG models. For simplicity, we focus on configurational atomistic models of systems in thermodynamic equilibrium.

2 Framework for Coarse-Graining

We consider an atomic system consisting of n atoms occupying a fixed volume $V \subset \mathbb{R}^3$ at temperature T, in thermodynamic equilibrium. The most probable states of this all-atom (AA) system are characterized by the Boltzmann distribution for canonical ensembles,

$$\rho_{AA}(\mathbf{r}^n) = \exp\{-\beta(u(\mathbf{r}^n) - a)\}, \tag{1}$$

where $\mathbf{r}^n$ is the set of particle coordinates ($\mathbf{r}^n = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$), $\beta = 1/k_BT$, k_B being Boltzmann's constant and T the absolute temperature, $u(\mathbf{r}^n)$ is the potential energy of the AA system, and a is the Helmholtz free energy ($a = -\beta^{-1} \ln Z_{AA}$, Z_{AA} being the partition function). In general, the goal of constructing such an AA model is to compute certain thermodynamic properties of the system, or, more specifically, to predict certain *quantities of interest (or "QoIs") Q; e.g.*

$$Q = \langle q \rangle_{AA} = \int_{\Gamma_{AA}} \rho_{AA}(\omega) q(\omega) \, d\omega, \tag{2}$$

where $q(\omega)$ ($\equiv q(\mathbf{r}^n)$) is a given phase-function (examples are given later), $\langle q \rangle_{AA}$ is its ensemble average in the AA system, and Γ_{AA} is the phase space of the AA model. Here, and occasionally hereafter, if no confusion is likely, we use the compact notation, $\omega = \mathbf{r}^n$.

Owing to the enormous size and complexity of the AA model, we wish to replace it by a reduced-order coarse-grained (CG) model that will hopefully deliver results that are both consistent with and close to those of the AA model. The CG model consists of $N(<< n)$ molecules occupying volume V, with temperature T. The coordinate positions are denoted $\mathbf{R}^N = (\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N)$ and it is convenient to regard the coordinates of each molecule "bead" $\mathbf{R}_I$ in the CG model as the image of a surjective map G of the AA coordinates $\mathbf{r}^n$ to the CG coordinates $\mathbf{R}^N$, as depicted in Figure 1: $\mathbf{R}_I = G(\mathbf{r}^n)$, $I = 1, 2, \ldots, N$; see, for example, Noid *et al* [62, 63], Shell [80], and the references therein.

The CG probability distribution is

$$\rho_{CG}\left(\boldsymbol{\theta}; \mathbf{R}^N\right) = \exp\left\{-\beta\left(U\left(\boldsymbol{\theta}; \mathbf{R}^N\right) - A\left(\boldsymbol{\theta}\right)\right)\right\}, \tag{3}$$

where $\boldsymbol{\theta}$ is a vector of parameters embedded in the characterization of the interaction potential energy $U(\cdot, \cdot)$ and free energy $A(\cdot)$ of the CG model. Hereafter, we shall usually write for compactness in notation, $\mathbf{R}^N = G(\mathbf{r}^n) = G(\omega)$. The CG approximation of the QoI (2) for given $\boldsymbol{\theta}$ is then

$$Q_{CG}(\boldsymbol{\theta}) = \langle q(\boldsymbol{\theta}) \rangle_{CG} = \int_{\Gamma_{CG}} \rho_{CG}\left(\boldsymbol{\theta}; G(\omega)\right) q\left(G(\omega)\right) \, dG(\omega), \tag{4}$$

where Γ_{CG} is the phase space of the CG model and $G(\omega) = G\left(\mathbf{r}^n\right) = \mathbf{R}^N$.

The parameters $\boldsymbol{\theta}$ are generally random variables distributed with respect to some probability measure μ_θ so $Q_{CG}(\boldsymbol{\theta})$ is also a random variable. One measure of its value is the expectation with respect to a distribution π,

$$\mathbb{E}_\pi\left[Q_{CG}\right] = \int Q_{CG}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{5}$$

The measure distribution π may be taken to be a prior or posterior distribution generalized in the validation process. We could also, of course, compute variances and other moments to further characterize the estimated QoI.

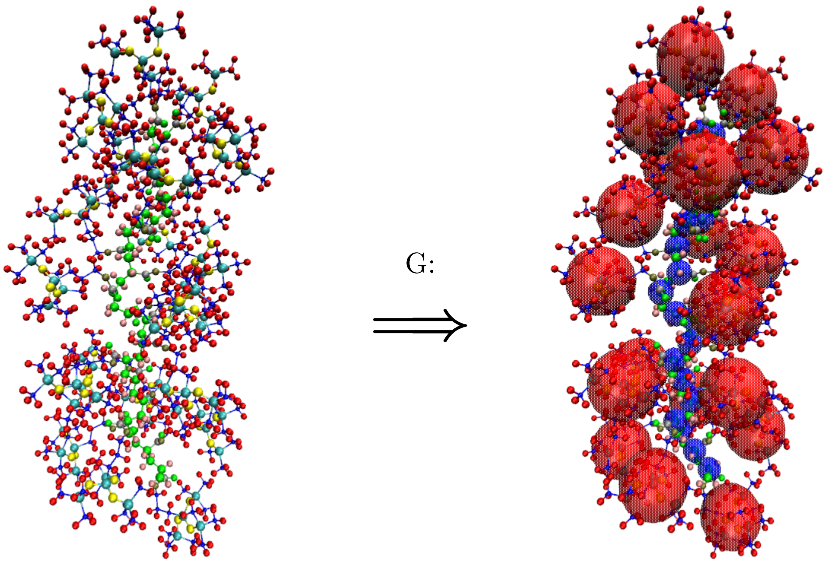


Figure 1: The map G from an all-atom (AA) model to a coarse-grained (CG) model.

The CG model of the AA system is completely characterized by (3) and, for the chosen QoI, by (4) and (5). It involves the identification of the N beads (equivalently, the map G), the form of the potential $U\left(\boldsymbol{\theta};\mathbf{R}^N\right)$, and the parameters $\boldsymbol{\theta}$. For any fixed choice of N and G, we wish to determine $U(\cdot,\cdot)$ and $\boldsymbol{\theta}$ so that the CG model is not only thermodynamically consistent with the AA model but that it also delivers sufficiently accurate approximations of the QoIs (4). We regard the first of these goals (consistency with the AA model) as primarily a matter of *model calibration*, the observational data being supplied by the AA model, while the second (estimating the accuracy of the QoI approximation Q_{CG}) is one of *model validation* (see, *e.g.* [1–3, 66–68, 78]), with data again being furnished by the AA model. Generally we will assume that the number N of beads and that the AA to CG map G is fixed, but if the resulting model is found to be invalid for given QoIs, then obviously, these features of the model must be changed.

3 Bayesian Framework for Statistical Calibration and Validation

3.1 Bayesian Inference as a Framework for Model Validation

While Bayesian approaches to problems of statistical inference have been the subject of debate over two-and-a-half centuries, their rise in popularity in broad areas of science and engineering in recent years has been, in part, due to the general framework they provide for conceptualizing foundational issues in the applications of the scientific method in the presence of uncertainties. These include uncertainties in observational data and in model parameters, while exploiting, when possible, prior information known about the system under study.

3.2 Bayes' Rule

Given observational data $\mathbf{D} = \{D_i\}_{i=1}^{L}$ gleaned from measurements or, as in the present case, furnished by the AA model, and given that additional prior information may be known about the parameters in the form of a true postulate I (and here following Jaynes' insistence that we recognize all relevant prior information at our disposal [41]), and also given prior information on the parameters in the form of a probability density function $\pi(\boldsymbol{\theta}|I)$, and, finally, given a likelihood function $\pi_{like}(\mathbf{D}|\boldsymbol{\theta}, I)$, then Bayes' rule can be written in the form,

$$\pi_{post}(\boldsymbol{\theta}|\mathbf{D}, I) = \frac{\pi_{like}(\mathbf{D}|\boldsymbol{\theta}, I)\pi(\boldsymbol{\theta}|I)}{\pi_{evidence}(\mathbf{D}|I)}, \tag{6}$$

where $\pi_{post}(\boldsymbol{\theta}|\mathbf{D}, I)$ is the posterior probability density function and the evidence $\pi_{evidence}(\mathbf{D}|I)$ is the marginal likelihood:

$$\pi_{evidence}(\mathbf{D}|I) = \int \pi_{like}(\mathbf{D}|\boldsymbol{\theta}, I)\pi(\boldsymbol{\theta}|I)\, d\boldsymbol{\theta}. \tag{7}$$

In what follows, we do not always list prior knowledge of information I for notational simplicity.

The posterior defines the Bayesian update of the parameters made possible by the knowledge of the prior and the data, and the likelihood is a measure of how well the model maps parameters into data.

3.3 The Prediction Pyramid

The processes leading to the prediction of a QoI have been said to be analogous to traversing a hypothetical pyramid, the prediction pyramid, through a sequence of experiments and model predictions (see [1, 3, 66–68]). Beginning with unit calibration tests at the base of the pyramid performed in calibration scenarios S_c with calibration observational data $\mathbf{D}_c$, one progresses up the pyramid to subsystem level validation scenarios S_v with validation observational data $\mathbf{D}_v$, and then moves to the full prediction scenario S_p at the peak of the pyramid where the QoI resides. By "scenarios" we mean the solution domain and the boundary and initial conditions in which the model is implemented, the model itself being a mathematical abstraction of a theory (an inductive hypothesis), generally independent of the parameters. The implication is that as one moves up the pyramid, the amount of observational data decreases (which may not always be the case). The QoI is not an observable, as explained in [68].

Determining the proper calibration and validation scenarios begins in the prediction scenario. This process is illustrated in Figure 6 for the application discussed in Section 8. Within the prediction scenario, a cube of polymeric material, are complex polymer chains which become the validation scenarios, shown in Figure 6c. These chains can be broken down even further as in Figure 6d to create the calibration scenarios.

The purpose of the calibration tests, by definition, is to calibrate the parameters for unit components that make up the model by fitting model predictions with calibration data $\mathbf{D}_c$ in the generally simpler calibration scenarios S_c. Since there are uncertainties in both the data and the parameters, this becomes a statistical calibration process in which probability density functions (pdfs) are sought in the form of posteriors. Unfortunately, at this stage in the analysis of the CG model, this process is generally impossible without additional information because *the model itself is unknown: we do not know which of the many possible choices of interaction potentials should be used for the given molecular structure of the model for the S_c scenarios.*

The validation experiments, on the other hand, are in theory designed to 1) challenge the model by testing the validity of the assumptions on which the full-system model is based, and 2) bring into play the influence of the choices of the QoI on the parameters as well. In this setting, model validation is interpreted as follows: *the process of determining the confidence one has in a model's ability to predict QoIs based on the accuracy with which the model can predict specific observables in validation scenarios.*

4 Bayesian Model Calibration

4.1 Construction of Priors via the Principle of Maximum Entropy

In the absence of any prior information on parameters, but for the specification of a finite set of possible values, it is customary to employ uniform priors, representing "complete ignorance" [41], in which parameter values fall in intervals possibly suggested by experimental or virtual data (in this case, AA data). In [42], Jeffreys argues that complete ignorance of a continuous variable θ known to be positive is best represented by assigning a uniform prior to its logarithm, leading to priors of the form $\pi(\theta|I) \propto \theta^{-1}$, $0 \leq \theta \leq \infty$, a prior that cannot be normalized but is nevertheless used successfully (*e.g.* [41], p. 182).

A more satisfactory approach to construct priors when some features of the parameter distribution are known, such as the mean $\langle\theta_i\rangle$, is to employ the principle of maximum entropy. The information entropy (or Shannon entropy [79]) of a probability density p is defined by

$$H(p) = -\int p(\theta) \log p(\theta)\, d\theta \tag{8}$$

for the continuous case and

$$H(p) = -\sum_{i=1}^{m} p(\theta_i) \log p(\theta_i) \tag{9}$$

for the discrete case, where m is the number of samples of the parameter pdf. It is argued in [41] (see also [17]) that the information entropy is the only reasonable measure of the amount of uncertainty in p ("reasonable" meaning it satisfies four basic conditions laid down in [79]), and that the correct prior distribution maximizes $H(p)$.

Let us suppose that the mean of a parameter θ can be inferred by information from the AA model, which is in fact generally possible, as will be seen later. We seek a pdf π such that the entropy $H(\pi)$ is maximized subject to the constraints that $\sum_i \pi_i = 1$ and that the distribution π delivers the known mean. A straightforward calculation yields the prior distribution,

$$\pi_i = \frac{1}{\langle\theta\rangle} \exp(-\theta_i / \langle\theta\rangle). \tag{10}$$

If, in addition, the variance, σ_θ^2, is known from the AA data, then a similar calculation reveals that the maximum entropy prior is a Gaussian distribution

$$\pi_i = \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(\frac{-(\theta_i - \langle\theta\rangle)^2.}{2\sigma_\theta^2}\right). \tag{11}$$

This approach can be generalized when more information can be inferred from the AA data.

4.2 Calibration Likelihoods

According to [26], "The likelihood that any parameter should have any assigned value is proportional to the probability that if this were so, the totality of all observations should be that observed." Likelihood functions can be constructed by assigning a probability distribution p to the error representing the difference between the observational data $\mathbf{D}$ and the parameter-to-observation map provided by the model. To define the observational data, we select L independent and identically distributed samples of the internal energy $u(\omega)$ at sites $\{\omega_i\}_{i=1}^{L}$ in Γ_{AA} and set the calibration data vector $\mathbf{D}_c = \mathbf{D}$, with $D_i = \beta u(\omega_i)$, $i = 1, 2, \ldots, L$. The parameter-to-observation map in this case is the vector

$$\mathbf{d}(\boldsymbol{\theta}) = \beta \left\{ U(\boldsymbol{\theta}; G(\omega_1)), U(\boldsymbol{\theta}; G(\omega_2)), \ldots, U(\boldsymbol{\theta}; G(\omega_L)) \right\}^T. \tag{12}$$

If p_ε is the probability density function that describes the error $\boldsymbol{\varepsilon} = \mathbf{D}_c - \mathbf{d}(\boldsymbol{\theta})$, due to observational noise and model inadequacy, then the likelihood is

$$\pi_{like}(\mathbf{D}_c|\boldsymbol{\theta}) = p_\varepsilon(\mathbf{D}_c - \mathbf{d}(\boldsymbol{\theta})). \tag{13}$$

A common assumption is that p_ε is a normal distribution $\sim \mathcal{N}(0, \mathbf{\Gamma}_{noise}^{-1})$, in which case

$$\pi_{like}(\mathbf{D}_c|\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}(\mathbf{D}_c - \mathbf{d}(\boldsymbol{\theta}))^T \mathbf{\Gamma}_{noise}^{-1}(\mathbf{D}_c - \mathbf{d}(\boldsymbol{\theta}))\right\}, \tag{14}$$

where $\mathbf{\Gamma}_{noise}$ is the covariance matrix representing the noise plus model error. A common approximation is $\mathbf{\Gamma}_{noise} = \sigma^2\mathbf{I}$, where the variance σ^2 is added to the list of parameters to be calibrated using available data.

4.3 Calibration Scenarios

With the likelihood and priors identified, we calibrate the CG model in one or more calibration scenarios S_c with data $\mathbf{D}_c$ furnished by the AA model and compute the calibration posteriors

$$\pi_{post}^c(\boldsymbol{\theta}) = \pi_{post}^c(\boldsymbol{\theta}|\mathbf{D}_c) = \frac{\pi_{like}(\mathbf{D}_c|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{D})}, \tag{15}$$

where $\pi(\mathbf{D})$ is the evidence. To evaluate such posteriors, we use QUESO, an efficient, parallel code that employs several sampling algorithms, particularly MCMC [75].

5 Bayesian Model Validation

The validation scenarios S_v correspond to more complex system level models which employ calibration posteriors as priors and, ideally, are designed to reflect the choice of the QoI on the accuracy with which the CG model predicts validation observables $\mathbf{D}_v$. The validation observables are also generated through samples drawn from the AA model, now in a way that approximates the QoI and now in an appropriate validation scenario. In addition, one must avoid overfitting by designing validation tests to produce an information gain, in the sense of Shannon, as described in [79] over the calibration process.

We now identify the parameter-to-observation misfit as

$$\boldsymbol{\varepsilon}_v(\boldsymbol{\theta}) = \mathbf{D}_v - \mathbf{d}_v(\boldsymbol{\theta}), \tag{16}$$

where, for example,

$$\begin{aligned} \mathbf{D}_v &= \{q(\omega_i)\rho_{AA}(\omega_i)\}_{i=1}^L, & \omega \in \Gamma_{v,AA} \quad (17)\\ \mathbf{d}_v(\boldsymbol{\theta}) &= \{q(G(\omega_i))\rho_{CG}(\boldsymbol{\theta}; G(\omega_i))\}_{i=1}^L, & G(\omega) \in \Gamma_{v,CG}. \quad (18) \end{aligned}$$

Here, $q(\omega)$ is the phase function defining the QoI and $\Gamma_{v,CG}$ is the subspace of the phase space associated with the validation scenario. A possible likelihood function is

$$\pi^{v}_{like}(\mathbf{D}_v|\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\varepsilon}_v(\boldsymbol{\theta})^T\boldsymbol{\Gamma}_v^{-1}\boldsymbol{\varepsilon}_v(\boldsymbol{\theta})\right\}, \tag{19}$$

with $\boldsymbol{\Gamma}_v$ an appropriate covariance matrix. The validation update for the parameters is

$$\pi^{v}_{post}(\boldsymbol{\theta}) = \pi^{v}_{post}(\boldsymbol{\theta}|\mathbf{D}_v) = \frac{\pi^{v}_{like}(\mathbf{D}_v|\boldsymbol{\theta})\pi^{c}_{post}(\boldsymbol{\theta}|\mathbf{D}_c)}{\pi(\mathbf{D}_v)}, \tag{20}$$

where, as previously mentioned, the validation prior is the calibration posterior.

6 Model Validity

Let us now consider a sequence of validation experiments with respective scenarios $S_{v1}, S_{v2}, \ldots$ and phase spaces $\Gamma_{v1,CG} \subset \Gamma_{v2,CG} \subset \ldots \subset \Gamma_{CG}$. With parameters characterized by posteriors of the form (20), we solve for the best approximations of the QoI that can be provided by each scenario, which are always limited by the fact that only the low-dimensional phase spaces $\Gamma_{v,CG}$ are available to us. Since the parameters are random variables, the QoI can be computed for each randomly chosen parameter vector. This yields a QoI for scenario S_{vk} given by the pdf

$$\pi(Q_{CG,k}|\boldsymbol{\theta}) = \int_{\Gamma_{CG,k}} q(G(\omega))\rho_{CG}(\boldsymbol{\theta}; G(\omega))\, dG(\omega), \tag{21}$$

where it is understood that the results are conditioned on all data. The expected value of the quantity of interest is then

$$\mathbb{E}_{\pi^{vk}_{post}}\left[\pi(Q_{CG,k}|\boldsymbol{\theta})\right] = \int_{\Theta} \pi^{vk}_{post}(\boldsymbol{\theta}|\mathbf{D}_{vk})\pi(Q_{CG,k}|\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{22}$$

However, for S_{vk} and $\Gamma_{CG,k}$, the "observational value" of this quantity is known:

$$Q_{AA,k} = \int_{\Gamma_{AA,k}} q(\omega)\rho_{AA}(\omega)\, d\omega. \tag{23}$$

So the accuracy with which the model can predict the QoI at this stage is determined by the error measure,

$$\gamma_k = \left|Q_{AA,k} - \mathbb{E}_{\pi^{vk}_{post}}\left[\pi(Q_{CG,k}|\boldsymbol{\theta})\right]\right|. \tag{24}$$

We remark that if the QoI approximation $Q_{AA,k}$ is itself a probability distribution as opposed to a real number, we use as an error measure the Kullback-Leibler divergence, $\gamma_k = D_{KL}(\pi(Q_{AA,k})\|\pi(Q_{CG,k}|\boldsymbol{\theta}))$, where

$$D_{KL}(\pi(Q_{AA,k})\|\pi(Q_{CG,k}|\boldsymbol{\theta})) = \int \pi(Q_{AA,k}) \log\frac{\pi(Q_{AA,k})}{\pi(Q_{CG,k}|\boldsymbol{\theta})}\, dQ. \tag{25}$$

We must now make a (subjective?) decision of whether or not the accuracy of the model, as indicated by the magnitude of γ_k, is sufficient to declare the model "valid" (or, more accurately, "not invalid"). If a tolerance γ_{tol} is established for this purpose, we merely check if $\gamma_k \leq \gamma_{tol}$. If not, a continuation up the validation pyramid to $S_{v,k+1}$ may be undertaken, and we compute a new error γ_{k+1}, with generally $\gamma_{k+1} < \gamma_k$. If $\gamma_{k+1} \leq \gamma_{tol}$, the model is deemed valid. If no scenario explored can meet the preset tolerance, then the CG model must be declared invalid and either abandoned or improved by increasing the number N of molecular beads and/or their configuration. An improvement in predictions manifested in the decrease in γ_k as the volume of the phase space is enlarged, and with $\gamma_{k+l} < \gamma_{tol}$ for some l gives one confidence that the model can predict the QoI for the full prediction scenario S_p.

6.1 Solution of the Forward Problem

In all of the numerical calculations we describe in sections to follow, we follow standard practice and invoke the ergodic hypothesis, evaluating ensemble averages such as (21) and (23) using MD models. Calculations are implemented using the Sandia code Largescale Atomistic/Molecular Massively Parallel Simulator (LAMMPS), which delivers approximations of canonical ensembles based on Nosé-Hoover thermostats [74].

7 Model Plausibility

A major challenge in constructing a CG model from an AA system arises from the many choices that must be made in aggregating atom groups. One must define (*i.e.* choose) the mapping from the atomistic coordinates into the CG coordinates, $G(\omega)$. In defining this map, one must choose the number N of CG sites and, most importantly, the interactions between the CG sites.

A general approach to model selection is embodied as the notion of model plausibility, which provides not only a basis for choosing interaction potentials but also their parameters. The idea is to view the sets of possible combinations of CG sites and interaction parameters as different models of molecular structures, each with different parameters, so that one has a set $\mathcal{M}$ of model classes, $\mathcal{M} = \{(M_1, \boldsymbol{\theta}_1), (M_2, \boldsymbol{\theta}_2), \ldots, (M_K, \boldsymbol{\theta}_K)\}$. For the set $\mathcal{M}$ of CG model classes, a Bayesian rule can be written for each model pair, $(M_j, \boldsymbol{\theta}_j)$, in the set:

$$\pi(\boldsymbol{\theta}_j|\mathbf{D}, M_j) = \frac{\pi(\mathbf{D}|\boldsymbol{\theta}_j, M_j)\pi(\boldsymbol{\theta}_j|M_j)}{\pi(\mathbf{D}|M_j)}, \quad j = 1, 2, \ldots, K \tag{26}$$

where $\pi(\mathbf{D}|M_j)$ is the *evidence* of model M_j:

$$\pi(\mathbf{D}|M_j) = \int \pi(\mathbf{D}|\boldsymbol{\theta}_j, M_j)\pi(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j. \tag{27}$$

Here, $\mathbf{D}$ is the AA calibration data, $\mathbf{D}_c$. By appealing to what some call a "higher form" of Bayes theorem (see, *e.g.*, [41]), one can define the *posterior plausibility* of model M_j in the set $\mathcal{M}$ for given data $\mathbf{D}$:

$$\rho(M_j|\mathbf{D}, \mathcal{M}) = \frac{\pi(\mathbf{D}|M_j)\pi(M_j|\mathcal{M})}{\pi(\mathbf{D}|\mathcal{M})}. \tag{28}$$

Here, $\pi(M_j|\mathcal{M})$ is the prior plausibility that model M_j is true among those in $\mathcal{M}$, and $\pi(\mathbf{D}|\mathcal{M})$ is the marginal probability over the model classes. Relations such as these are discussed in [8, 68, 69].

The plausibility (28) provides an immediate means to determine which model in the set best fits the data. Indeed, since

$$\sum_{j=1}^{K} \rho(M_j|\mathbf{D}, \mathcal{M}) = 1, \tag{29}$$

the model(s) closest to unity are deemed the most plausible. In particular, if

$$\rho(M_j|\mathbf{D}, \mathcal{M}) > \rho(M_k|\mathbf{D}, \mathcal{M}), \tag{30}$$

then M_j is more plausible than M_k for the data $\mathbf{D}$.

8 A Model Application: Analysis of Polymer Components in Nanomanufacturing

For specificity, we now describe calibration and validation processes for models of polymer materials encountered in certain nanomanufacturing processes. Advancements in technology have lead to the miniaturization of modern electronics and their components, such as semiconductors. Techniques such as optical projection lithography can print features smaller than 100 nanometers onto these components. However, as the size of these features continues to decrease, the cost of manufacturing using traditional techniques increases. Step and Flash Imprint Lithography (SFIL) offers an affordable alternative to patterning the nanoscale features onto semiconductors. The behavior of the polymeric etch barrier critical in the fabrication process is targeted as a model problem class for demonstrating all atom, coarse-grained, and macroscale modeling issues explored in this work. Figure 6b depicts one realization of an approximately 10 nm cube of polymeric material consisting of chains of molecules making up the etch barrier used in the Step and Flash Imprint Lithography (SFIL) process for manufacturing semiconductors.

Before the SFIL process can begin, a transfer layer, consisting of an organic polymer, is spin-coated onto a silicon substrate wafer. An etch barrier solution is deposited onto the transfer layer, as shown schematically in Figure 2. It is important that this solution has low viscosity and is *photocurable*, *i.e.* its polymerization is initiated by light. The polymerization process is detailed later. A translucent template, which contains the pattern to be imprinted, is lowered, trapping the etch barrier in its imprint geometry. Polymerization starts when the template is illuminated with ultraviolet light. After polymerization is complete, the template is removed, leaving the desired pattern in the etch barrier. In Step 5, the base layer of the etch barrier is removed with a halogen plasma etch, exposing the transfer layer beneath. The desired features are then etched into the transfer layer with an anisotropic oxygen reactive ion etch (RIE). Finally, the imprint is etched into the substrate and what remains of the etch and transfer barriers is washed away, leaving the desired pattern in the silicon substrate [4, 6, 21, 22].

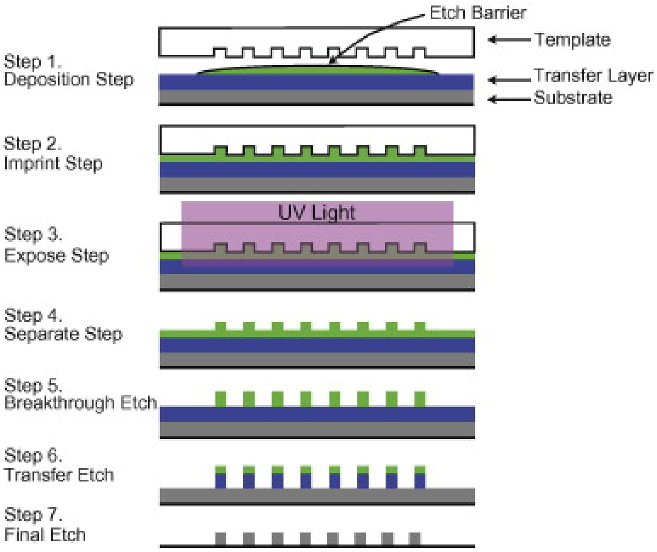


Figure 2: Illustration of the steps involved in the SFIL procedure [6]

.

The etch barrier is an organosilicon solution which contains four components. The first is a silylated monomer, a monoacrylate, which we call "monomer 1" or "M1." This monomer ensures that the etch barrier is not washed away during O_2 reactive-ion-etch exposure.The second component is a t-Butyl acrylate, which lowers the viscosity of the etch barrier solution. We call this organic monomer "monomer 2" or "M2." The "crosslinker" monomer ("XL") is an ethylene glycol diacrylate, which provides thermal stability and improves the cohesive strength of the etch barrier solution. Finally, we have the photoinitiator, which dissociates when exposed to ultraviolet light to initiate polymerization. We use "I" as shorthand to refer to the initiator molecule. The chemical structure of these molecules is given in Figure 3 and Figure 4. Further details can be found in [4, 21].

Note that monomers 1 and 2 and the crosslinker have a common component,

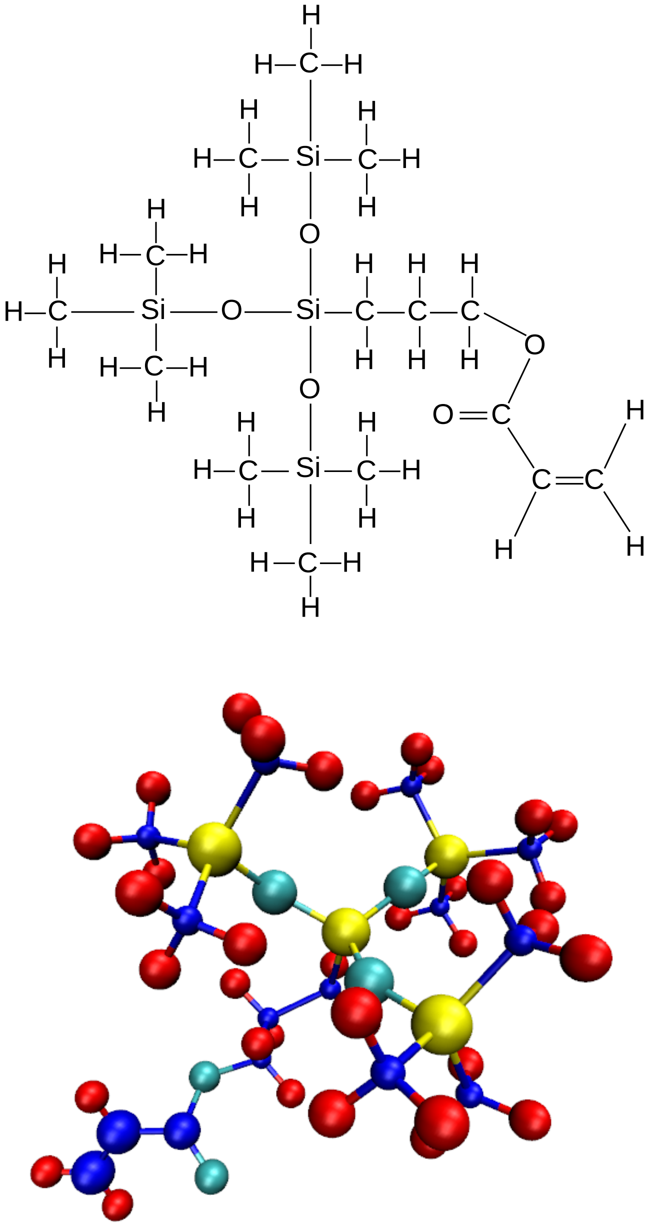
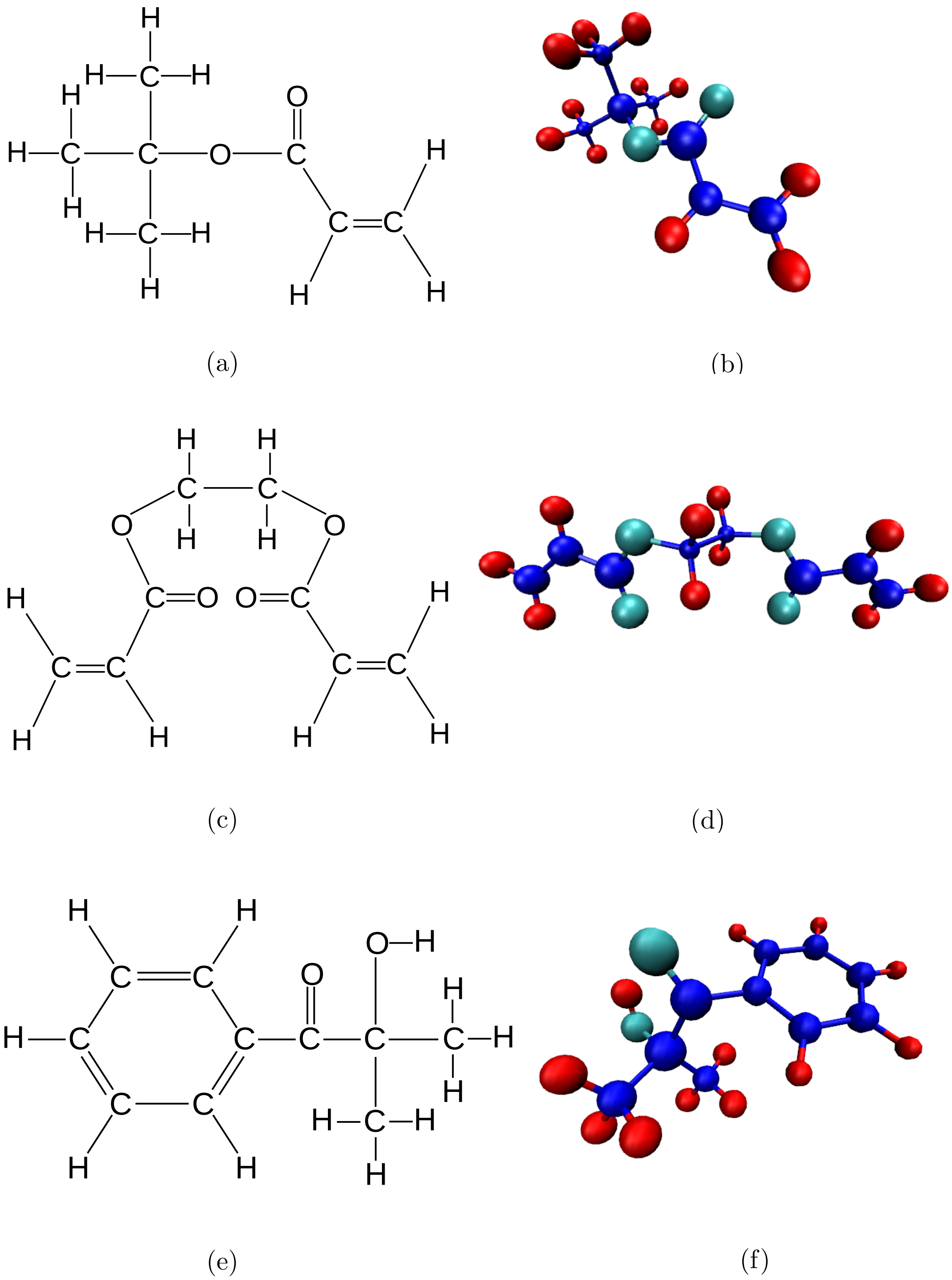


Figure 3: (top) Chemical structure of M1; (bottom) Three-dimensional rendering of M1.



(a) (b) (c) (d) (e) (f)

Figure 4: Chemical structures of (a) M2, (c) XL, and (e) the initiator; Three-dimensional images of (b) M2, (d) XL, and (f) the initiator.

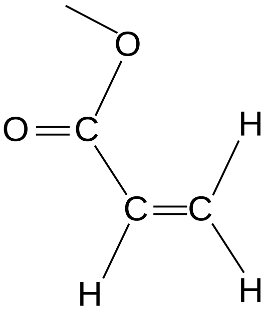


Figure 5: Component common to the molecules M1, M2, and XL.

shown in Figure 5. In fact, the crosslinker has two of these components. The fact that this structure unit repeats throughout the molecules implies that one of the CG beads can be defined to be this component. This bead is called the *linker bead*, or “L,” not to be confused with the crosslinker. The rest of each molecule goes into the *residue beads*. The residue bead for M1 is denoted “R1,” for M2 the residue bead is labeled “R2,” and the middle section of the crosslinker is expressed as “X.” Since the photoinitiator does not contain this component, it will be broken up into two beads (which become the radicals in the polymerization process), denoted “I1” and “I2.”

Each realization of the polymer cube is the result of a Kinetic Monte Carlo calculation in which charge-neutral conformations of the molecular structure are generated, where the appropriate mass fractions of the monomer constituents are maintained. In this scenario, we take the mass fractions as follows: M1 ~ 0.4445, M2 ~ 0.3716, XL ~ 0.1541, and I ~ 0.0298, as described in [4]. As an example of a goal of the simulation, we take as the QoI the total energy of the cube.

8.1 Calibration, Validation, and Prediction Scenarios

An idealized model of typical etch-barrier flanges is depicted in Figure 6a. As noted earlier, a simple example of the target quantity of interest is the total potential energy per unit volume of the material, computed over the representative cube shown in Figure 6b, which is the result of one realization of the polymerization process. Thus, in this case $q(\omega) = u(\omega)$ and Γ_{AA} consists of phase-space realizations confined to the representative cube.

Within the prediction scenario S_p (the cube), polymer chains of increasing complexity and size can be identified. These represent possible validation scenarios. Examples are given in Figure 8 and are depicted in the overall process in Figure 6c. For computational efficiency, the calibration phase can be broken up into three small calibration scenarios. The first calibration scenario is a chain of three M1s, the second is a chain of three M2s, and in the third, a crosslinker connects two small chains of M2s, as shown in Figure 7.

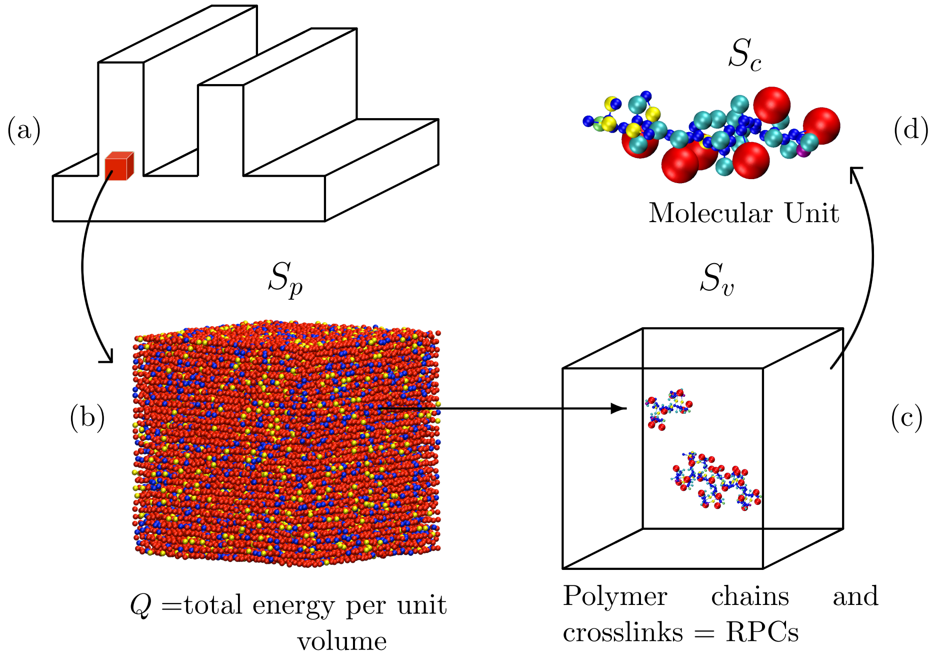

(a)
S_c
(d)
Molecular Unit
S_p
S_v
(b)
(c)
Q =total energy per unit volume
Polymer chains and crosslinks = RPCs

Figure 6: Illustration of the prediction, validation, and calibration scenarios for predictive models of polymer materials in an etch-barrier flange: a) the flange; b) one realization of a cube, the total energy being the QoI; c) validation RPCs and d) molecular units used in the calibration scenario

8.2 The Calibration Scenario

Before calibration can begin, any prior information that can be derived from the AA system needs to be included in the prior pdf of the parameters, $\pi(\boldsymbol{\theta}|I)$, while allowing as much uncertainty as possible, as discussed in Section 4.1. Details on how information is derived from the AA system is discussed in the appendix.

Once prior pdfs for each parameter have been defined, calibration can begin. Generally speaking, the parameters present in the first scenario will be calibrated using data from the all-atom (AA) system, $\mathbf{D}_{c1}$ and priors according to either (10) or (11) are assigned. In the second scenario, a new set of data from the AA system, $\mathbf{D}_{c2}$, will be used. For the parameters present in this scenario that also appeared in the previous scenario, the posteriors from S_{c1} will be used as priors in S_{c2}. The new parameters will be assigned maximum entropy priors. That is,

$$\pi_2\left(\theta_i\right) = \left\{ \begin{array}{ll} \pi_{post}^{c1}\left(\theta_i|\mathbf{D}_{c1}\right) & if\ \theta_i \in \boldsymbol{\theta}_{c1} \\ \pi(\theta_i) & otherwise \end{array} \right. , \tag{31}$$

where $\boldsymbol{\theta}_{c1}$ is the vector of parameters present in S_{c1}. Similarly, the prior parameter

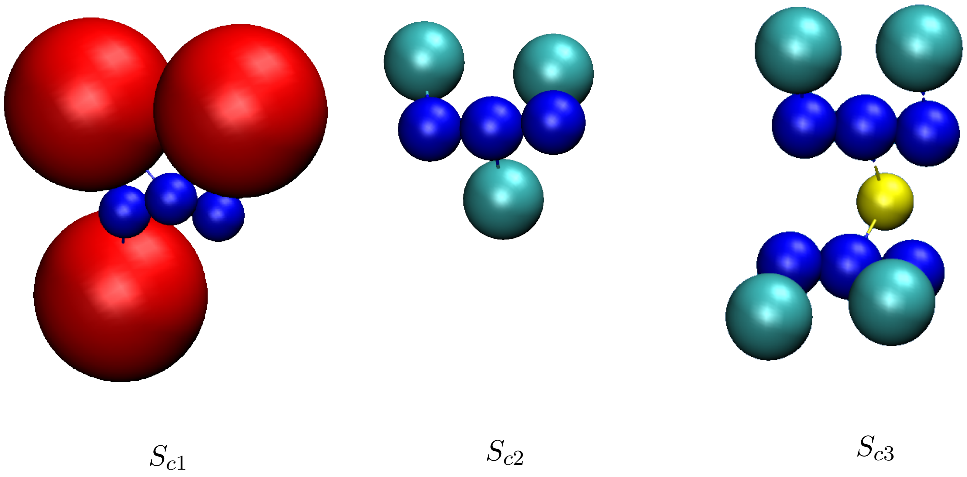


Figure 7: Series of calibration scenarios to be used in the SFIL model problem

distributions in the third calibration scenario are

$$\pi_3\left(\theta_i\right) = \begin{cases} \pi_{post}^{c1}\left(\theta_i|\mathbf{D}_{c1}\right) & if\ \theta_i \in \boldsymbol{\theta}_{c1}/\boldsymbol{\theta}_{c2} \\ \pi_{post}^{c2}\left(\theta_i|\mathbf{D}_{c2}\right) & if\ \theta_i \in \boldsymbol{\theta}_{c2} \\ \pi(\theta_i) & otherwise \end{cases} . \tag{32}$$

That is, any parameter in S_{c3} that appears in S_{c1} but not in S_{c2} will be assigned a prior equal to the posterior of that parameter in S_{c1}. Any parameter in S_{c3} that appears in S_{c2}, regardless of whether or not it appears in S_{c1}, will be assigned a prior equal to the posterior of that parameter in S_{c2}. Any new parameter is assigned a maximum entropy prior.

Consider as an example the bond parameters in each of these scenarios. In S_{c1}, there exist R1−L and L−L bonds, in S_{c2}, the bonds R2−L and L−L are present, and S_{c3} contains R2−L, L−L, and X−L bonds. Maximum entropy priors are used in S_{c1}. However, since L−L exists in S_{c2}, the posteriors from S_{c1} now become the priors for S_{c2} and the parameters are updated. Since R2−L is not present in S_{c1}, the parameters for this bond will be given maximum entropy priors. When moving on to S_{c3}, the posteriors for the R2−L and L−L bonds from S_{c2} will be used as priors. Note that the posteriors for L−L in S_{c2} contain more (updated) information than the posteriors from S_{c1}, which is why $\pi_{post}^{c2}\left(\boldsymbol{\theta}_{L-L}|\mathbf{D}_{c2}\right)$ is used over $\pi_{post}^{c1}\left(\boldsymbol{\theta}_{L-L}|\mathbf{D}_{c1}\right)$. As the X−L bond does not appear in either S_{c1} or S_{c2}, its parameters are given maximum entropy priors.

8.3 Model Choices

Let it be assumed that the number of coarse-grained particles is given. Then the models differ only in the representation of the potential energy function, U. Here, it is also assumed that the functional form of the potential energy function assumes the OPLS functional form, as described in [25, 45, 46],

$$\begin{aligned} U(\omega) &= \sum_{bonds} k_r(r(\omega)-r_0)^2 + \sum_{angles} k_\theta(\theta(\omega)-\theta_0)^2 \\ &\quad + \sum_{dihedrals}\sum_{n=1}^{4} \frac{V_n}{2}\left[1+(-1)^{n-1}\cos(n\varphi(\omega))\right] \\ &\quad + \sum_{non-bonded} 4\epsilon \left[\left(\frac{\sigma}{r(\omega)}\right)^{12} - \left(\frac{\sigma}{r(\omega)}\right)^{6}\right] f, \end{aligned} \tag{33}$$

where $k_r, k_\theta, r_0, \theta_0, \epsilon, \sigma$ are model parameters and f is a weighting function, equal to 0.5 for Lennard-Jones interactions between atoms that are separated by exactly three bonds, and 1.0 otherwise. By including or excluding different interaction types (bonds, angles, dihedrals, non-bonded), different models are created. A complete tabulation of all possible models is given in Table 1.

Most of these models, however, can be eliminated by considering prior information on the physics of the system. For example, there has to be some physical

Model	**Bonds**	**Angles**	**Dihedrals**	**Non-Bonded**	**# of Parameters**
M_1	rigid			✓	12
M_2	✓				18
M_3	✓			✓	30
M_4	rigid	✓			32
M_5	rigid	✓		✓	44
M_6	✓	✓			50
M_7	✓	✓		✓	62
M_8	rigid		✓		96
M_9	rigid		✓	✓	108
M_{10}	✓		✓		114
M_{11}	✓		✓	✓	126
M_{12}	rigid	✓	✓		128
M_{13}	rigid	✓	✓	✓	140
M_{14}	✓	✓	✓		146
M_{15}	✓	✓	✓	✓	158

Table 1: Tabulation of the interactions that are included in each of the available CG models. A check-mark in the column of an interaction implies the use of that interaction in the model. When bonds are not included, they are treated as rigid. When angles, dihedrals, and non-bonded terms are not included in the model, they have no contribution to the total energy of the system.

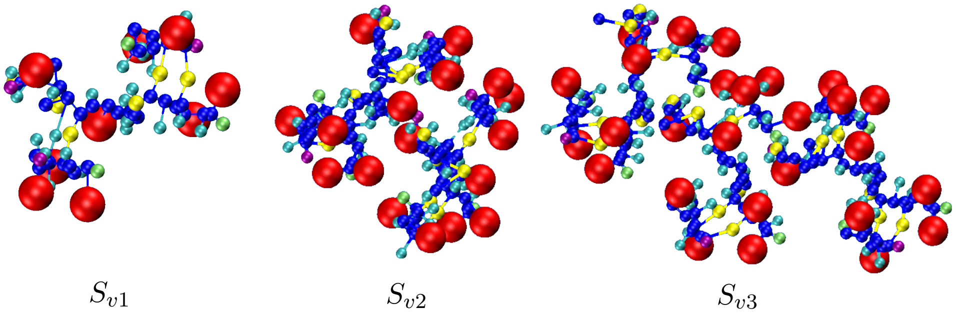


Figure 8: RPCs (Representative Polymer Chains) in a sequence of validation scenarios of increasing size

force that keeps two bonded molecules together. Therefore, bonds must be included in the model, with the covalent bonds represented by harmonic spring potentials or by the use of rigid bars, signifying no relative motion. Furthermore, the notion of the size of beads is given by the Lennard-Jones radius σ. Thus this interaction should also be included in the model. It should be noted that in the context of a cube of SFIL material, the contribution to the potential energy of the dihedral interactions will be quite small. In addition, the structural contribution of the dihedral potentials will be dominated by the packing constraints of being in a cube of material. These observations coupled with the fact that there are 96 dihedral parameters in the full system implies the calibrating dihedral parameters may not be worth the computational cost. With these restrictions, the possible models are given in Table 2. If necessary, this table can be expanded by considering different variations of the LJ interaction. The OPLS functional form traditionally uses a LJ 12-6 potential, but perhaps a softer potential, such as LJ 9-6 would be more appropriate.

The conclusion of the validation process now follows the recipe discussed earlier. With the plausibilities of the models in Table 2 computed, the parameter are updated and the most plausible models are determined. These are used to compute the calibration posteriors for the molecular units, which become priors for the RPCs in the validation scenarios. These are used to calculate the approximate QoIs (the

Model	**Bonds**	**Angles**	**Dihedrals**	**LJ 12-6**	**# of Parameters**
M_1	rigid			✓	12
M_2	✓			✓	30
M_3	rigid	✓		✓	44
M_4	✓	✓		✓	62

Table 2: Table of possible models to represent the CG system of SFIL. Each model will be calibrated and its plausibility calculated.

validation observables) via (21) and their expected value via (22). The approximate AA QoI (23) is computed using LAMMPS and the error measure γ_k is computed using (24). Again, if $\gamma_k \leq \gamma_{tol}$, the model is deemed "valid" (or "not invalid"). What remains is to spell out more details in the critical plausibilty calculations. We discuss this in the next subsection for a lower degree of freedom example, hexane.

8.4 Example: Hexane

As a proof-of-concept example, consider a molecule of hexane, C_6H_{14}, shown in Figure 9. The CG system consists of three beads, each containing two carbon atoms and their attached hydrogen atoms.

Since the CG model has three beads, possible model choices are created by including or excluding the harmonic bond, the harmonic angle, and the Lennard-Jones 12-6 potential, given in (33). As discussed in Section 8.3, the Lennard-Jones potential must be included to give the CG system a notion of bead size. Bonds must also be included, whether as harmonic springs or rigid bonds. This leaves only three possible models. Therefore, the Lennard-Jones 9-6 potential is also introduced, producing six possible models, shown in Table 3.

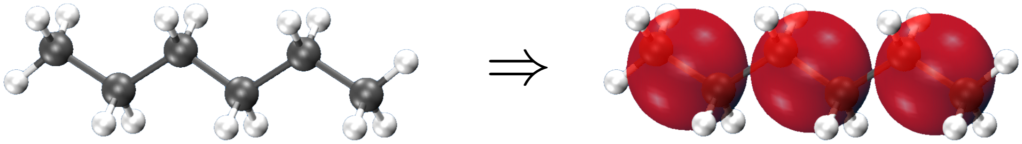


Figure 9: Schematic of the mapping from AA hexane (left) to the CG representation of hexane (right). Each CG bead contains two carbon atoms and their accompanying hydrogen atoms.

Model	Bonds	Angles	LJ 9-6	LJ 12-6	# of Parameters
M_1	✓			✓	4
M_2	rigid	✓		✓	4
M_3	✓	✓		✓	6
M_4	✓		✓		4
M_5	rigid	✓	✓		4
M_6	✓	✓	✓		6

Table 3: Table of possible representations of the CG potential energy for hexane. Each model will be calibrated and its plausibilities will be calculated and compared to quantitatively determine which model best represents the AA data.

We consider as a calibration scenario a single hexane molecule in a constant volume and temperature ensemble. No molecules are being added to or taken away from the system, and the number of CG beads describing the molecule does not change. Thus we are considering a canonical ensemble. For calibration data, we take $D_i = u(\omega_i)$, $i = 1, \ldots, L$. That is, each data point is the potential energy of the system at a configuration of the molecule. Prior information on the CG parameters can be deduced from the AA system using techniques discussed in the appendix. Prior distributions, which are used during the calibration of each model, are shown in Figure 10.

To complete the Bayesian inversion framework, a Gaussian likelihood of the form (14) is used. Since the models differ only in the representation of the potential energy function, the inversion process differs between each model choice through the parameter-to-observation map $\mathbf{d}(\boldsymbol{\theta})$. For model M_j, we define $d_i(\boldsymbol{\theta}_j) = U(\boldsymbol{\theta}_j; G(\omega_i))$. That is, each data point is the CG potential energy evaluated using parameters $\boldsymbol{\theta}_j$, in a configuration of the molecule when mapped onto the CG system.

For each of the six model choices, the parameters are updated using (26), and the plausibility is calculated using (28). For hexane, the normalized plausibilities are

$$\rho_1 \approx 0.5, \quad \rho_2 = 0, \quad \rho_3 \approx 0, \quad \rho_4 \approx 0.5, \quad \rho_5 = 0, \quad \rho_6 \approx 0 \tag{34}$$

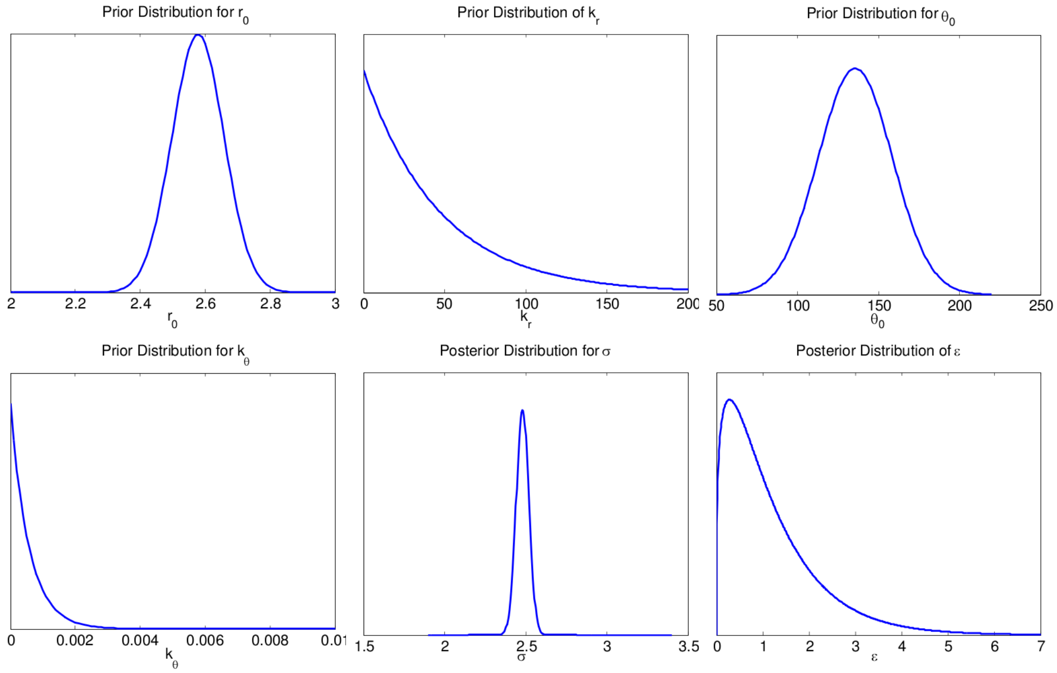


Figure 10: Prior distributions of all parameters that may be considered for any choice of representation of the potential energy for CG hexane.

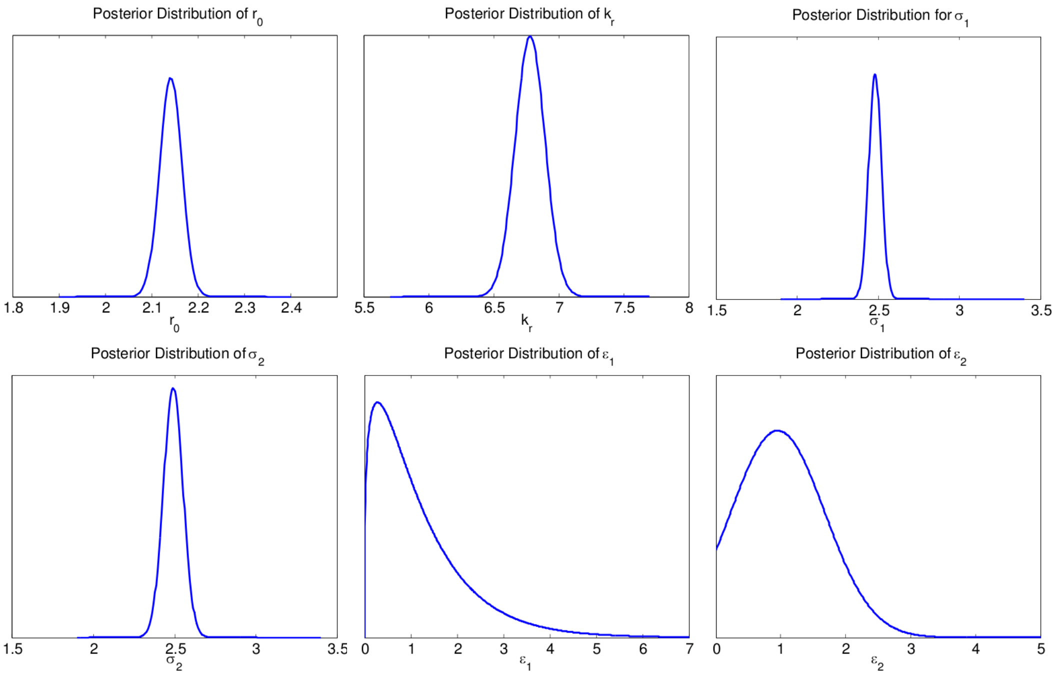


Figure 11: The three-bead CG system of hexane yields a choice of six parametric model classes. For each, the parameters are calibrated according to the aforementioned Bayesian calibration. The evidence resulting from Bayesian inversion is used to calculate the plausibility of each model. Posteriors produced by one of the most plausible models are shown here

From this it is easy to see that Model 1 and Model 4 are equally plausible models. The posteriors for Model 1 are shown in Figure 11. In both models covalent bonds are represented by harmonic springs and angular interactions are neglected. The difference between the two models lies in the expression of the Lennard-Jones interaction, however, the hexane model is insensitive to this difference.

It is important to note that in this case, the most plausible model for approximating the potential energy in the prediction scenario is *not* the model that has the most parameters. Should either of these two models be validated, the computation cost of predicting the QoI in a much larger system will be lower than simply choosing the model with the most parameters, as intuition may suggest we do.

As a validation scenario, consider now octadecane, a longer polyethylene chain with 18 carbon atoms and 38 hydrogen atoms. In accordance with (20), the calibration posterior now becomes the validation prior in the Bayesian update. As was the case in the calibration step, each validation data point is the potential energy, now of octadecane, at a configuration sampled from the canonical distribution and a Gaussian likelihood is used.

Parameters are updated for Model 1 using (20). The updated posteriors are then used to run the CG system in the canonical ensemble. The kernel density estimate of the pdfs of the AA observable, $\pi(Q_{AA,v})$, as well as $\pi(Q_{CG,v}|\boldsymbol{\theta}_1)$ for Model 1 are shown in Figure 12. Although it is clear that $\pi(Q_{AA,v})$ is not accurately predicted, this can be verified quantitatively using techniques discussed in Section 6. If we take as the QoI the ensemble average of the potential energy, $Q_{AA,v} = 62.4610$ and $\mathbb{E}_{\pi^v_{post}}[\pi(Q_{CG,v}|\boldsymbol{\theta}_1)] = 6.0905$, yielding an error of 90.2%. It is reasonable to take $\gamma_{v,1} = 0.15(Q_{AA,v})$ (a tolerance of 15% error), rendering Model 1 invalid. If instead the QoI is considered to be the distribution of potential energy, the D_{KL} is computed. Determining a proper tolerance for this pseudo-measure is less straightforward, but we relate it to the variance in the AA distribution by assigning $\gamma_{v,2} = 0.12\sigma^2_{AA}$. Then

$$D_{kl}(\pi(Q_{AA,v})\|\pi(Q_{CG,v}|\boldsymbol{\theta}_1)) > \gamma_{v,2}, \tag{35}$$

again rendering Model 1 invalid. A similar validation experiment can be done for the remaining models, revealing that none of them is valid.

To remedy this situation, the class of models may be redefined. It should be noted that within a single CG bead, the minimum energy configuration of the contained atoms yields a non-zero potential energy. This contribution is lost in (33). Therefore, a new parameter, the internal bead energy, A is introduced into (33). The hexane model class may be redefined to be $\tilde{\mathcal{M}} = \{(\tilde{M}_1, \tilde{\boldsymbol{\theta}}_1), \ldots, (\tilde{M}_6, \tilde{\boldsymbol{\theta}}_6)\}$, with the parametric models characterized according to Table 4. The plausibility calculation can be recomputed, yielding

$$\tilde{\rho}_1 \approx 0, \quad \tilde{\rho}_2 = 0, \quad \tilde{\rho}_3 \approx 0.5, \quad \tilde{\rho}_4 \approx 0, \quad \tilde{\rho}_5 = 0, \quad \tilde{\rho}_6 \approx 0.5. \tag{36}$$

The validation experiment, previously described, yields updated parameters, which are then used to run the CG system in the canonical ensemble. The kernel density estimate of $\pi(Q_{CG,v}|\tilde{\boldsymbol{\theta}}_3)$ is shown in Figure 12. Although it appears bi-modal, it is completely contained in the range of $\pi(Q_{AA,v})$. Furthermore,

$$\left|Q_{AA,v} - \mathbb{E}_{\pi^v_{post}}\left[\pi(Q_{CG,v}|\tilde{\boldsymbol{\theta}}_3)\right]\right| < \gamma_{v,1} \tag{37}$$

Model	Bonds	Angles	LJ 9-6	LJ 12-6	A	# of Parameters
$\tilde{M}_1$	✓			✓	✓	5
$\tilde{M}_2$	rigid	✓		✓	✓	5
$\tilde{M}_3$	✓	✓		✓	✓	7
$\tilde{M}_4$	✓		✓		✓	5
$\tilde{M}_5$	rigid	✓	✓		✓	5
$\tilde{M}_6$	✓	✓	✓		✓	7

Table 4: Table of possible representations of the CG potential energy for hexane. Each model will be calibrated and its plausibilities will be calculated and compared to quantitatively determine which model best represents the AA data.

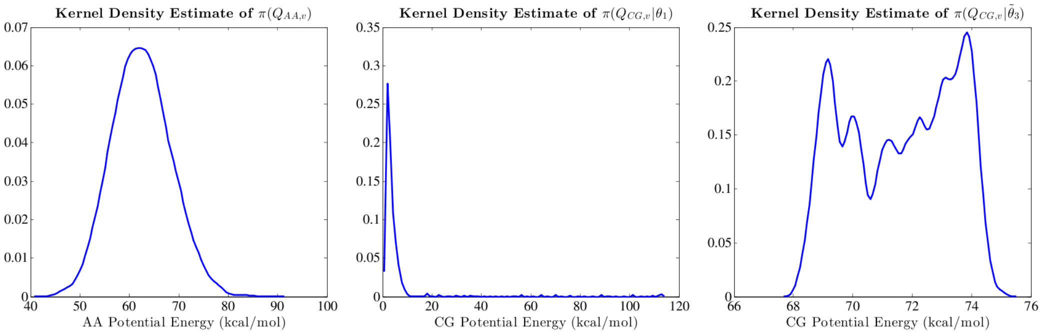


Figure 12: Kernel density estimates of the probability distribution functions of the potential energy of C_8H_{18} calculated in the all-atom system (left), the coarse-grained system as defined by model M_1 (middle), and the coarse-grained system as defined by model $\tilde{M}_3$ (right).

and

$$D_{kl}(\pi(Q_{AA,v})\|\pi(Q_{CG,v}|\tilde{\boldsymbol{\theta}}_3)) < \gamma_{v,2}, \tag{38}$$

where $\gamma_{v,1}$ and $\gamma_{v,2}$ are defined as before. Therefore, $\tilde{M}_3$ is considered valid (not invalid).

It is noted here that if we consider as a model class the most plausible models from $\mathcal{M}$ and those from $\tilde{\mathcal{M}}$, *e.g.* $\hat{\mathcal{M}} = \{(M_1, \boldsymbol{\theta}_1), (M_4, \boldsymbol{\theta}_4), (\tilde{M}_3, \tilde{\boldsymbol{\theta}}_3), (\tilde{M}_6, \tilde{\boldsymbol{\theta}}_6)\}$, the plausibilities may be compared. Then

$$\rho_1 \approx 0, \qquad \rho_4 \approx 0, \qquad \tilde{\rho}_3 \approx 0.5, \qquad \tilde{\rho}_6 \approx 0.5, \tag{39}$$

which agrees with intuition following the numerical results above.

9 Concluding Comments

A broad theoretical and computational framework for the selection and validation of coarse-grained models of atomistic systems is laid down in this work which uses Bayesian inference and information theoretics to deal with model and parameter uncertainties. The problem of selecting the CG model itself, generally ignored in CG construction, is addressed using the notion of model plausibilities, an idea that has been used in low-dimensional statistics for many years. We demonstrate the use of Bayesian model plausibilities in calculations of parameters and most plausible models for a hexane molecule coarse-grained with three beads. We also lay the groundwork for future studies of complex polymer structures encountered in Step and Flash Imprint Lithography for semiconductor nanomanufacturing. These tools are also useful, and possibly essential, for developing methods for validation of a large class of multiscale models.

Acknowledgements

This work was partially supported by the Mathematical Multifaceted Integrated Capability Centers (MMICCs) effort within the Applied Mathematics activity of the U.S. Department of Energy's Advanced Scientific Computing Research program under Award Number DE-5C0009286. We like to thank the many fruitful and enlightening discussions with Peter Rossky and Eric wright, from which we've benefited greatly, and our collaboration with Serge Prudhomme.

References

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

[2] I. Babuška and J. T. Oden. Verification and validation in computational engineering and science: Part I, basic concepts. *Computer Methods in Applied Mechanics and Engineering*, 193(1):4047–4068, 2004.

[3] I. Babuška, R. Tempone, and F. Nobile. A systematic approach to model validation based on Bayesian updates and prediction-related rejection criteria. *Computer Methods in Applied Mechanics and Engineering*, 197:2517–2539, 2008.

[4] T. Bailey, S. Johnson, S. Srreenivasan, J. Ekerdt, C. Willson, and D. Resnick. Step and flash imprint lithography: An efficient nanoscale printing technology. *Journal of Photopolymer Science and Technology*, 15(3):481–486, 2002.

[5] P. T. Bauman. *Adaptive multiscale modeling of polymeric materials using goal-oriented error estimation, Arlequin coupling, and goals algorithms.* PhD thesis, University of Texas at Austin, 2008.

[6] P. T. Bauman, J. T. Oden, and S. Prudhomme. Adaptive multiscale modeling of polymeric materials with Arlequin coupling and Goals algorithms. *Computational Methods in Applied Mechanics and Engineering*, 198:799–818, 2009.

[7] M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. M. Cafeo, C. Cavendish, L. C.-H., and J. Tu. A framework for validation of computer models. *Technometrics*, 49(2):138–154, May 2007.

[8] J. L. Beck and K.-V. Yuan. Model selection using response measurements: Bayesian probabilistic approach. *Journal of Engineering Mechanics*, 130(2):192–203, 2004.

[9] E. Brini, V. Marcon, and N. F. A. van der Vegt. Conditional reversible work method for molecular coarse graining applications. 13(22):10468–10474, 2011.

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

[11] 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. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[12] 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):8383–8393, 2012.

[13] 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, 2011.

[14] M. Clyde, H. Desimone, and G. Parmigiani. Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91(435):1197–1208, 1996.

[15] M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

[16] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[17] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, 2nd edition, 2006.

[18] R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13, 1946.

[19] B. S. Daan Frenkel. *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1 of *Computational science*. Academic Press, 2 edition, 2001.

[20] L. Devroye. A course in density estimation. *Progress is Probability and Statistics*, 14, 1987.

[21] M. D. Dickey, R. L. Burns, E. Kim, S. C. Johnson, N. A. Stacey, and C. G. Willson. Study of the kinetics of Step and Flash Imprint Lithography photopolymerization. *AlChE Journal*, 51(9):2547–2555, 2005.

[22] M. D. Dickey and C. G. Willson. Kinetic parameters for Step and Flash Imprint Lithography photopolymerization. *AlChE Journal*, 52(2):777–784, 2006.

[23] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 45–97, 1995.

[24] F. Ercolessi and J. B. Adams. Interatomic potentials from first-principles calculations: The force-matching method. *Europhysics Letters*, 26(8):583, 1994.

[25] K. Farrell and J. T. Oden. Statistical calibration and validation methods of coarse-grained and macro models of atomic systems. *ICES REPORT 12-45*, 2012.

[26] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transcations of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

[27] P. J. Flory. Thermodynamics of high polymer solutions. *The Journal of Computational Physics*, 10(1):51–61, 1942.

[28] S. Geisser. *Predictive Inference: An Introduction*. Chapman & Hall, New York, 1993.

[29] J. Y. Halpern. A counterexample to theorems of cox and fine. *Journal of Artificial Intelligence Research*, 10:67–85, 1999.

[30] J. Y. Halpern. Cox's theorem revisited (technical addendum). *Journal of Artificial Intelligence Research*, 11:429–435, 1999.

[31] D. Higdon. Brittleness again. In *Xi'an's Og*. xianblog.wordpress.com/2013/09/11/bayesianbrittleness, 2013.

[32] K. S. V. Horn. Constructing a logic of plausible inference: a guide to coxs theorem. *International Journal of Approximate Reasoning*, 34(1):3 – 24, 2003.

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

[34] IUPAC. *Compendium of Chemical Terminology, (the "Gold Book")*. Blackwell Scientific Publications, 2nd edition, 1997. Compiled by A. D. McNaught and A. Wilkinson.

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

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

[37] S. Izvekov, M. Parrienllo, 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):10896–10913, 2004.

[38] S. Izvekov and G. A. Voth. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry B*, 109(7):2469–2473, 2005.

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

[40] 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, 2006.

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

[42] H. Jeffreys. *The Theory of Probability*. Oxford University Press, 4 edition, 1988.

[43] X. Jiang and S. Mahadevan. Bayesian cross-entropy methodology for optimal design of validation experiments. *Measurement Science and Technology*, 17:1895–1908, 2006.

[44] X. Jiang and S. Mahadevan. Bayesian validation assessment of multivariate computational models. *Journal of Applied Statistics*, 15(1):49–65, January 2008.

[45] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[46] W. L. Jorgensen and J. Tirado-Rives. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.

[47] M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87:1–13, 2000.

[48] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(3):pp. 425–464, 2001.

[49] J. G. Kirkwood. Statistical mechanics of liquid solutions. *Chemical Reviews*, 19(3):275–307, 1936.

[50] A. R. Leach. *Molecular Modeling: Principles and Applications.* Pearson Education Limited, Prentice Hall, Harlow, 2nd edition, 2001.

[51] E. E. Leamer. *Specification searches: ad hoc inference with nonexperimental data.* Wiley, New York.

[52] S. Lehti and V. Karimäki. Computing methods in high energy physics. Lecture Notes, Helsinki Institute of Physics, University of Helsinki.

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

[54] M. Loève. *Probability Theory.* Springer-Verlag, 1977.

[55] 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.

[56] 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.

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

[58] S. B. McGrayne. *The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy.* Yale University Press, 2011.

[59] 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.

[60] 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.

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

[62] W. G. Noid, J.-W. 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*, 128:244114, 2008.

[63] W. G. Noid, P. Liu, Y. Wang, J.-W. 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*, 128:244115, 2008.

[64] S. Nosé. Constant temperature molecular dynamics. *Journal of Chemical Physics*, 81(1), 1984.

[65] J. T. Oden, A. Hawkins, and S. Prudhomme. General diffuse-interface theories and an approach to predictive tumor growth modeling. *Mathematical Models and Methods of Applied Science*, 20(3):1–41, 2010.

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

[67] J. T. Oden, R. Moser, and O. Ghattas. Computer predictions with quantified uncertainty, Part II. *SIAM News*, 43(10), December 2010.

[68] 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 (in review)*. Published as a preprint as ICES Report 13-12, ICES, 2013.

[69] J. T. Oden, E. E. Prudencio, and A. Hawkins-Daarud. Selection and assessment of phenomenological models of tumor growth. *Mathematical Models and Methods in Applied Sciences*, 23(7):1309–1338, 2013.

[70] J. T. Oden and S. Prudhomme. Estimation of modeling error in computational mechanics. *Journal of Computational Physics*, 182(2):496–515, 2002.

[71] J. T. Oden, S. Prudhomme, A. Romkes, and P. T. Bauman. Multiscale modeling of physical phenomena: Adaptive control of models. *SIAM Journal on Scientific Computing*, 28(6):2359–2389, 2006.

[72] H. Owhadi and C. Scovel. Brittleness of bayesian inference and new selberg formulas. *ArXiv e-prints*, Apr. 2013.

[73] H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz. Optimal uncertainty quantification. *ArXiv e-prints*, Sept. 2010.

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

[75] E. Prudencio and K. Schulz. The parallel c++ statistical library queso: Quantification of uncertainty for estimation, simulation and optimization. In M. Alexander, P. DAmbra, A. Belloum, G. Bosilca, M. Cannataro, M. Danelutto, B. Martino, M. Gerndt, E. Jeannot, R. Namyst, J. Roman, S. Scott, J. Traff, G. Valle, and J. Weidendorfer, editors, *Euro-Par 2011: Parallel Processing Workshops*, volume 7155 of *Lecture Notes in Computer Science*, pages 398–407. Springer Berlin Heidelberg, 2012.

[76] A. E. Raftery, D. Madigan, and C. T. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian statistics*, 5:323–349, 1996.

[77] 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.

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

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

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

[81] Q. Shi, S. Izvekov, and G. A. Voth. Mixed atomistic and coarse-grained molecular dynamics: Simulation of membrance-bound ion channel. *The Journal of Physical Chemistry B*, 110(31):15045–15048, 2006.

[82] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, UK, 1986.

[83] B. Smit, P. A. J. Hilbers, K. Esselink, L. A. M. Rupart, N. M. van Os, and A. G. Schlijper. Computer simulations of water/oil interface in the presence of micelles. *Nature*, 348:624–625, 1990.

[84] 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), May 2009.

[85] J. H. Weiner. *Statistical Mechanics of Elasticity*. Dover, Mineola, N.Y., 2002.

[86] P. K. Weiner and P. A. Kollman. AMBER: Assisted Model Building with Energy Refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry*, 2(3):287–303, 1981.

[87] E. Wright. A Bayesian framework for calibration and uncertainty quantification of coarse-grained atomistic models. *ICES REPORT*, April 2013.

[88] J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth. Coarse-grained peptide modeling using a systematic multiscale approach. *Biophysical Journal*, 92(12):4289–4303, 2007.

Appendix

A Derivation of All-Atom Data for Maximum Entropy Priors

The type of information that must be collected from the AA system to inform the CG parameter priors depends on how the parameter appears in the model. In this appendix, we discuss how statistical mechanics is used to determine what type of information we can infer about the CG parameters and how this information is collected from the AA system. Specifically, the equilibrium bond distance parameter, R_0, the bond spring constant, k_R, the equilibrium angle parameter, θ_0, the angle spring constant, k_θ, and the Lennard-Jones radius, σ, and well depth, ϵ, are examined.

A.1 Harmonic Bond Parameters

Consider the CG parameter that describes the equilibrium bond length of one particular type of bond (such as, for example, the R1−L bond in the SFIL system). For each sample configuration of the AA system, ω_i, the distance between the CG sites in the AA system can be measured. We have chosen our AA-to-CG map such that the CG sites correspond to specific atoms contained in each CG bead.

In the case of the SFIL system, the central silicon atom in the R1 bead and the carbon atom in the L bead correspond to the CG sites for R1 and L, respectively. The distance between these two CG sites is then

$$R_{0,R1-L}(G(\omega_i)) = r_{Si-C}(\omega_i) = \|\mathbf{r}_{Si}(\omega_i) - \mathbf{r}_C(\omega_i)\|. \tag{A.1}$$

This distance can be considered to be a sample distance of the CG bond length. The average of these distances can be considered to be the average equilibrium length of the bond in the CG system,

$$\langle R_0 \rangle = \frac{1}{n} \sum_{i=1}^{n} R_{0,i}, \tag{A.2}$$

where for notational simplicity, $R_0 = R_{0,R1-L}$ and $R_{0,i} = R_{0,R1-L}(G(\omega_i))$. Furthermore, the variance in the observed length can be computed,

$$\sigma^2_{R_0} = \frac{1}{n}\sum_{i=1}^{n}\left(R_{0,i} - \langle R_0\rangle\right)^2 . \tag{A.3}$$

Therefore, since the approximate mean and variance of the bond length can be extracted, the prior distribution for any equilibrium bond length parameter is a Gaussian.

For non-structural parameters, such as the spring constant k_R, the situation is more theoretical. The mean values for the spring coefficients of these interactions can be derived in the same way as the Equipartition Theorem of statistical mechanics, as mentioned in [87], the proof of which is contained in [85].

The Equipartition Theorem states that each quadratic term in the energy with positive coefficient contributes $k_BT/2$ to the mean energy, where, as before, k_B is Boltzmann's constant, T is the temperature, and the mean is taken with respect to the canonical distribution function. That is,

$$\left\langle \sum_{i=1}^{r} a_i y_i^2 \right\rangle = \frac{rk_BT}{2}, \tag{A.4}$$

where a_i is the positive constant and y_i represents a structural variable, such as coordinates.

Recall that the bonded and angular interactions are represented by harmonic springs. Therefore, each bond is a quadratic term with positive coefficient. For each bond, we can write

$$\left\langle k_R(R-R_0)^2 \right\rangle = \frac{k_BT}{2}. \tag{A.5}$$

Although the random variable k_R and the possible bond lengths $(R-R_0)^2$ are not completely independent, we let

$$\langle k_R\rangle = \frac{k_BT}{2\left\langle (R-R_0)^2\right\rangle}. \tag{A.6}$$

Thus, the mean value of the spring coefficient is roughly inversely related to the variance of the bond length,

$$\langle k_R\rangle = \frac{k_BT}{2\sigma^2_{R_0}}. \tag{A.7}$$

As no further information can be extracted, the prior distributions for bond spring coefficient parameters are given by (10).

A.2 Harmonic Spring Parameters

Harmonic interactions are also used to represent harmonic springs, thus similar arguments to those above can be used to derive prior information regarding angular parameters. Consider as an example the angle between an R1−L bond and an L−L bond. This angle can be measured by

$$\theta_{0,R1-L-L}(\omega_i) = \cos^{-1}\left(\frac{(\mathbf{r}_{Si}(\omega_i) - \mathbf{r}_{C,1}(\omega_i))^T(\mathbf{r}_{C,2}(\omega_i) - \mathbf{r}_{C,1}(\omega_i))}{\|\mathbf{r}_{Si}(\omega_i) - \mathbf{r}_{C,1}(\omega_i)\|\|\mathbf{r}_{C,2}(\omega_i) - \mathbf{r}_{C,1}(\omega_i)\|}\right), \tag{A.8}$$

where $\mathbf{r}_{Si}$, $\mathbf{r}_{C,1}$, and $\mathbf{r}_{C,2}$ are the coordinate vectors of the CG sites associated with R1, the cental L and the end L beads, respectively. Angular analogies to Eqs. (A.2) - (A.3) follow, yielding prior information regarding the average and variance of the equilibrium angle distance. Furthermore, the mean values for the spring coefficients of angular interactions can be derived from the Equipartition Theorem, as discussed above, where the angular analogy to (A.7) is used.

A.3 Lennard-Jones Parameters

In 2003, Reith *et. al* [77] derived a coarse-graining method called iterative Boltzmann inversion and demonstrated its accuracy for a Lennard-Jones fluid. The general idea for extracting information about Lennard-Jones parameters is therefore borrowed from their procedure, as described in [87].

The process begins by building a radial distribution function $g(R)$, which describes the probability of finding a particle a distance R from a given particle, as compared to the ideal gas distribution [50]. In the OPLS functional form, the LJ parameters are defined between pairs of similar particles. These parameters are geometrically averaged when a non-similar pair of particles is considered. Therefore, for each sample configuration of the AA system, ω_i, the distance between pairs of like-particles is measured. Details on how this statistical data is compared to the ideal gas distribution are given in [19, 50].

It can be shown that the radial distribution function is related to the potential of mean force,

$$U(R) = -k_B T \ln g(R), \tag{A.9}$$

which examines how the energy of the system changes as a function of the distance between two particles [49, 50]. Since the radial distribution function has a single maximum [50], the potential of mean force has a single, well-defined minimum, say R^*. Furthermore, this minimum is locally Gaussian. Therefore, we can take R^* to be the mean of the LJ radius parameter σ, and the local variance about this minimum to be the variance in its prior Gaussian distribution. The depth of the well in the potential of mean force, $U(R^*)$, is taken as the mean for the LJ well-depth parameter ϵ, to be used in a prior given by (10).