The Dissertation Committee for Kathryn Anne Farrell
certifies that this is the approved version of the following dissertation:

Selection, Calibration, and Validation of Coarse-Grained Models of Atomistic Systems

Committee:

J. Tinsley Oden, Supervisor

Serge Prudhomme, Co-Supervisor

Ivo Babuska

Tan Bui-Thanh

Leszek Demkowicz

Ron Elber

Selection, Calibration, and Validation of Coarse-Grained Models of Atomistic Systems

by

Kathryn Anne Farrell, B.A., M.S.C.A.M.

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

Words cannot express my gratitude for the endless help and unfailing encouragement of those without whom I would not have accomplished this feat. I would like to thank my undergraduate advisor, Michael Holst, who believed in me before I did and offered me continuous advice and encouragement throughout my graduate career. I must also thank my committee, especially my advisors, J. Tinsley Oden and Serge Prudhomme, as well as Ivo Babuska, Leszek Demkowicz, Tan Bui-Thanh, Ron Elber, and former committee member, Peter Rossky, for their guidance and challenges that have enhanced the quality of this work. J. Tinsley Oden's contagious passion for predictive science, his vision and development of the discipline of computational science and engineering, and his insatiable appetite for learning have inspired and shaped this work. The many thought-provoking, theoretical discussions with Serge Prudhomme have, no doubt, had an unquantifiable impact on this work and my understanding and appreciation for this field. I would also like to extend my gratitude to Peter and Edith O'Donnell for the support they have given the Institute for Computational Engineering and Sciences and the CSEM Fellowship, from which I have benefited. Thanks to postdoctoral researcher, Danial Faghihi, for his contributions to this research, and to my fellow graduate students, especially Corey Bryant, Lindley Graham, Andrea Hawkins-Daarud, Michael Harmon, Jessica Meixner, Pearl Flath, and Talea Mayo for the many fruitful discussions and support throughout the past six years. I am indescribably grateful for the unconditional love and encouragement of Lauren Daiuto, Matthew Maupin, my parents, and the rest

of my family, to whom this work is dedicated.

Selection, Calibration, and Validation of Coarse-Grained Models of Atomistic Systems

Kathryn Anne Farrell, Ph.D.
The University of Texas at Austin, 2015

Supervisor: J. Tinsley Oden
Co-Supervisor: Serge Prudhomme

This dissertation examines the development of coarse-grained models of atomistic systems for the purpose of predicting target quantities of interest in the presence of uncertainties. It addresses fundamental questions in computational science and engineering concerning model selection, calibration, and validation processes that are used to construct predictive reduced order models through a unified Bayesian framework. This framework, enhanced with the concepts of information theory, sensitivity analysis, and Occam's Razor, provides a systematic means of constructing coarse-grained models suitable for use in a prediction scenario.

The novel application of a general framework of statistical calibration and validation to molecular systems is presented. Atomistic models, which themselves contain uncertainties, are treated as the ground truth and provide data for the Bayesian updating of model parameters. The open problem of the selection of appropriate coarse-grained models is addressed through the powerful notion of Bayesian model plausibility.

A new, adaptive algorithm for model validation is presented. The Occam-Plausibility ALgorithm (OPAL), so named for its adherence to Occam's Razor and the use of Bayesian model plausibilities, identifies, among a large set of models, the simplest model that passes the Bayesian validation tests, and may therefore be used to predict chosen quantities of interest. By discarding or ignoring unnecessarily complex models, this algorithm contains the potential to reduce computational expense with the systematic process of considering subsets of models, as well as the implementation of the prediction scenario with the simplest valid model.

An application to the construction of a coarse-grained system of polyethylene is given to demonstrate the implementation of molecular modeling techniques; the process of Bayesian selection, calibration, and validation of reduced-order models; and OPAL. The potential of the Bayesian framework for the process of coarse graining and of OPAL as a means of determining a computationally conservative valid model is illustrated on the polyethylene example.

Table of Contents

Chapter 1

Introduction

1.1 Introductory Remarks

This document contains a study of basic questions in computational science and engineering concerned with the use of coarse-grained (CG) molecular models to predict the behavior of atomistic systems. It addresses several long-standing issues in computer modeling and simulation of atomistic systems, including model selection, calibration, validation, and error estimation, all in the presence of uncertainties, and it also addresses the quantification of uncertainties in observables and quantities of interest, all associated with reduced-order models obtained by aggregating atoms into beads or super atoms or molecules. A general Bayesian setting is developed to address these questions, which, when augmented with concepts from information theory, such as information entropy, provides a unified framework for model selection, validation, and prediction under uncertainty. Throughout this investigation, for simplicity and specificity, attention is restricted to configurational energies of systems in thermodynamic equilibrium, with the understanding that the theory and methods developed are applicable to much more general systems.

The following developments are presented in this study:

1. *Validation of CG Models.* A general framework and processes for statistical calibration and validation of complex molecular models in the presence of

uncertainties are presented. Data is provided by the so-called ground-truth models of the all-atom (AA) system provided by hardened and calibrated molecular dynamics (MD) codes. These observational data are not deterministic and possess uncertainties themselves since they are generally drawn from distribution functions defined for the AA system. Also, examples and applications involve polymer systems as these have been the primary focus of our earlier work on molecular modeling. In many cases, an important component of the calibration and validation processes for such materials is the identification of Representation Polymer Chains (RPCs), which provide the basic subunits of more complex structures and which are, themselves, made up of smaller molecular units that constitute building blocks in what are used in so-called calibration scenarios. This systematic calibration of basic units and then validation of more complex sub-systems that depict in some sense key quantities of interest, constitutes the principal ideas underpinning the Bayesian-based model validation process.

2. *Model Selection.* A major open problem in the construction of CG molecular models has been the selection of appropriate force potentials and choosing model parameters for each such choice - a particularly challenging proposition when there are uncertainties in both the observational data and the model parameters themselves. A general approach to this problem is developed based on Bayesian methods of model selection, particularly using the concept of posterior model plausibilities. This idea, advanced in the recent paper [31], builds on extensions of the idea of Bayes' factors and the work of Beck and

Yuen [10]. A general adaptive algorithm that addresses and provides a means for resolving the elusive problem of model inadequacy (or "model form error" or "model bias") that has been a central issue in model validation for many years [1, 57, 58, 85, 86] is also developed.

3. *Parameter Sensitivities.* It is shown that the relative sensitivity of model outputs to variations in model parameters can be estimated using variance-based methods of sensitivity analysis, advocated, for example, by Saltelli *et al.* [103, 104]. By eliminating parameters that are judged to have little effect on key model outputs, dramatic reductions in model complexity and in the cost of assessing model plausibility and validity, and in computing quantities of interest can be achieved. It is also proposed that an added benefit of computing model sensitivities is their use in the checking the effectiveness of validation scenarios and tests. In particular, parameters that are judged to influence output values through sensitivity measures computed for the full prediction scenario should also influence results in validation scenarios, else the validation test is ill-designed.

4. *Adaptive Algorithms for Model Validation.* All of the methodologies described for statistical model calibration, validation, plausibility, parameter sensitivity, and uncertainty quantification are implemented in a general algorithm referred to as OPAL (Occam-Plausibility ALgorithm). The prefix "Occam" is in deference to *Occam's Razor*, the principle that states that among competing hypotheses, the simplest (the one with the fewest assumptions) should be selected. In OPAL, the simplest model is defined as the model with the

fewest parameters that also satisfies model validation criteria laid down by the modeler.

1.2 Coarse-Grained Models of Atomic Systems

Important advances in material science, biology, nanomanufacturing, drug design, and in many other fields, brought about by developments in computational modeling and simulation, high-performance computing, and experimental science, have dramatically expanded interest in the use of atomistic models to study a wide variety of physical phenomena and to analyze the behavior of many engineering, physical, biological, and medical systems. Although the universally accepted approach for modeling atomistic systems is to employ molecular mechanics simulations in the form of either molecular dynamics or Monte Carlo sampling, implemented using any of several well-documented and well-tested codes, the enormous size and complexity of systems of interest far exceed the capabilities of today's largest supercomputers or even those envisioned decades into the future. Coarse-grained (CG) models of atomistic systems, in which groups of atoms are aggregated into larger units to reduce the number of degrees of freedom, have been used for decades in significant technological and scientific applications. The development of a rigorous mathematical, physical, and statistical foundation for the process of coarse graining, including calibration, validation, and assessment of predictability of CG models, is, hence, a goal of great importance in computational science.

Advances in computer technologies inspired the development of numerical methods for molecular simulation, providing a tractable alternative to analytical

pursuits. Efforts to exploit the statistical nature of particular systems led to the first computer simulations of fluids in 1952 by Metropolis *et al.* [76], who developed a process for sampling the Boltzmann distribution, which, in turn, led to the advancement of the Monte Carlo method. Continuous potentials, used in molecular dynamics (MD) methods, recognize that the forces on each particle change as the positions of the particles and its neighbors change and were introduced by Rahman in 1964 [98]. These two contributions are among the most notable and influential in early molecular modeling and apply to both atomistic and coarse-grained implementations.

An extensive literature on coarse-grained methodologies exists, spanning over a half-century. Early versions of CG approximations appeared in the 1940s in the works of Paul J. Flory [34] and Maurice L. Huggins [43]. The united atom (UA) model, another predecessor to CG models, in which hydrogen atoms are ignored, has been used since the 1960s [38, 67]. Most MD codes include UA implementations due to the pervasiveness of these reduced order models.

In 1988, Robert L. McGreevy proposed the Reverse Monte Carlo (RMC) method for determining the potentials of CG models for condensed matter problems [74]. The method is a variation of the Metropolis-Hastings algorithm in which parameters are adjusted until those that yield the highest consistency with experimental data are obtained. However, it has been shown [12] that multiple models may agree with the data equally well while differing qualitatively in their predictions of key properties of interest. Furthermore, there are, for these methods in general, more parameters than observational data samples. Both of these issues

are common to many model problems in a wide variety of applications and may be addressed in the RMC framework by including additional constraints on the system in question [12]. In the present work, the Bayesian framework confronts the difficulty of model selection with the notion of model plausibility, and the presence of uncertainties, such as those introduced by lack of data, with the use of stochastic parameters, as discussed in Chapter 3. The RMC method continues to be further developed and used for a variety of model applications from crystalline materials to DNA [56, 70–72].

Furio Ercolessi and James B. Adams suggested the Force Matching (FM) method in 1994 [29], in which parameters are determined so that, as the name indicates, forces on the CG system match those in the AA system. The method was extended and applied to an assortment of applications, including condensed phase systems [45, 46, 48, 49] and biochemical applications [44, 47, 107, 120] by Sergei Izvekov, Gregory A. Voth, and their collaborators, who refer to the method as "multi-scale coarse-graining" (MS-CG) methods. As discussed in Chapter 2, CG particles are defined by groups of atoms. In MS-CG methods, coarse-grained parameters are chosen so that the sum of forces on each group of atoms is matched via least-squares minimization to the total force on the corresponding CG bead,

$$\boldsymbol{\theta}_{CG} = \underset{\theta}{\operatorname{argmin}} \left\| F_{AA} - F_{CG}\left(\boldsymbol{\theta}\right) \right\| .$$

The details of the implementation and numerical specifics of the method are given in [69] and its accuracy is assessed in [45]. It is argued [79, 81, 82] that MS-CG methods yield CG models that are "physically consistent" with underlying AA

models. In place of the notion of physical consistency defined in [81], the Bayesian setting for model construction includes the concept of model validation, in which a calibrated model is tested against experimental data. See Chapter 3 for details. Bayes' Rule has also been used in conjunction with MS-CG [68], but not for the purpose of stochastic model calibration or validation, such as that considered in this work. Instead, Bayes' Rule was used as a regularization tool for deterministic model calibration.

Dirk Reith and Florian Müller-Plathe have advocated the Iterative Boltzmann Method (IBM). Initially referred to as simply "a simplex method," IBM develops CG potentials by attempting to reproduce the radial distribution function, a function which characterizes the structure of an atomistic system [77]. This is done through iteration akin to a fixed-point algorithm, details of which can be found in [100]. IBM has successfully been applied to liquids, where only non-bonded terms need to be considered [77], and later to polymer systems in which bonded interactions must also be considered [99]. Other proponents of IBM are Wataru Shinoda, Russell DeVane, and Michael Klein, who have used the method to derive CG potentials for layer assembly, such as those for surfactants [108, 109] and zwitterionic lipids [110].

The relative entropy approach proposed by M. Scott Shell in 2008 is similar to IBM. However, instead of comparing radial distribution functions, Shell's method compares ensemble distribution functions [106], such as those described in Chapter 2. This information-theoretic approach was used by Di Wu and David Kofke [119] to compare molecular systems before being applied to coarse-graining by Shell and his

collaborators [18, 19]. The present work also uses information theory to compare CG models to their AA counterparts. However, Shell's framework uses relative entropy for the purpose of calibration only, while the present work uses the relative entropy as a measure for assessing validation of CG models.

What is often referred to as model validation in much of contemporary literature on CG methods amounts to qualitative and sometimes heuristic comparisons of deterministic predictions with data. Furthermore, few available approaches confront the issues of model selection and the presence of uncertainties, both in the data (*i.e.* in the AA model) and in the CG model. The Bayesian framework presented in this study provides a unified approach that includes model calibration, model selection, and model validation, all in the presence of uncertainties, for the purpose of building a predictive model of key quantities of interest.

1.3 The Bayesian Framework for Constructing Predictive Models

This study may be regarded as one on basic issues in predictive science which aims at extending existing approaches to the complex, discrete models encountered in predicting the behavior of atomic and molecular systems. The great majority of existing literature on determining parameters for CG models focuses on calibration processes, generally deterministic, and ignores the fundamental problem of model validation. Predictive science embodies the theory and methodology of predicting the behavior of physical systems or of predicting features of physical events in the presence of uncertainties. It entails the subjects of model calibration, model selection, model validation and verification, and uncertainty quantification.

At the heart of predictive science is the scientific method, whereby hypotheses are formulated, tested, and modified through comparison to observations or experimental measurements. Many attribute to the ancient work of Aristotle, collected in his *Organon* [3], the development of deductive logic, which provides the foundations for mathematics and inductive logic, and which forms a basis for scientific theory and discovery. In 1946, Richard Trelked Cox published a theorem and proof stating that under reasonable conditions on properties and measures of the plausibility of propositions, every natural extension of Aristotelian logic in the presence of uncertainties is Bayesian [22]. Many regard this work as laying the foundation of logical probability theory. Further discussion and examples were published in 1961 [23]. Debate on the validity of the theorem ensued with the publication of Joseph Halpern's counterexample to Cox's theorem [39, 40], later refuted by Stefan Arnborg and Gunnar Sjödin with the addition of "common sense" postulates to those presented by Cox [4, 5] and by Michael Hardy [41], who presented another approach to amending Cox's Theorem. A comprehensive overview of Cox's Theorem was published by Kevin van Horn in 2003 [113] and a self-proclaimed "trivial proof" was presented by Maurice Dupré and Frank Tipler in 2009 [28].

The debate over Bayes' Rule, however, began long before Cox's Theorem. Sharon Bertsch McGrayne details the long history of debate between Bayesian and frequentist probability theorists from its advent in 1763 when the rule was posthumously published by Richard Price [9], the executor of Bayes' estate. Some believe that the Reverend Thomas Bayes thought the cause-and-effect nature of the rule may unlock the proof (or disproof) of the existence of God [73] and, hence, delayed

its publication for decades. The use of Bayesian methods remains controversial, with recent publications highlighting its paradoxes (or "brittleness"), such as those by Houman Owhadi, Clint Scovel, and Tim Sullivan [94]. Recently, sufficient conditions for the convergence of Bayesian methods in the context of the Bernstein-von Mises Theorem (also called the "Bayesian Central Limit Theorem") for so-called misspecified parametric and nonparametric models were developed by B.J.K. Kleijn and A.W. van der Vaart [63, 64].

The powerful work, *Probability Theory: The Logic of Science* by Edwin T. Jaynes [53] provides exhaustive arguments and many examples that support the notion that "all science is inductive" and that the Bayesian setting provides a unified way to deal with scientific thought while naturally taking into full account inherent uncertainties. The present work subscribes fully to this philosophy.

1.4 Organization and Scope of this Study

As noted earlier, the physical systems of interest here are atomic systems in thermodynamic equilibrium, the behaviors of which are defined by contemporary MD simulations. The general goal is to develop theory and supporting numerical experiments underlying the selection, calibration, and validation of CG models of such systems in the presence of uncertainties. Following this Introduction, a review of theory governing the behavior of molecular models is given which includes the description of the statistical mechanics of the AA and the CG models. The construction of the AA-to-CG map, the choices that must be made in doing so, and the types of uncertainties that therefore arise are detailed. The technicalities of the

simulation of and the collection of data from the AA and CG systems are also given.

Chapter 3 is devoted to the broad problem of developing a Bayesian framework for predictive CG models. These general approaches for statistical model calibration and use of information theory are discussed together with the description of a powerful method for model selection, based on the idea of Bayesian model plausibilities, and an accompanying statistical validation method. The AA models are used to generate subsystem observational data that represents non-observable quantities of interest–the target goals of the prediction. Demonstrations of the theoretical results are given in the form of applications to typical molecular systems, particularly polyethylene. These concepts bring into the realm of molecular models the idea of the prediction pyramid, in which calibration scenarios involving basic model units are first used to get preliminary distributions of model parameters, then progressing up the pyramid to validation scenarios to update parameters and, finally, the full prediction is implemented by solving the forward problem to evaluate the quantity of interest.

In Chapter 4, a general adaptive algorithm for constructing valid CG models is presented: the Occam-Plausibility ALgorithm (OPAL). The idea is to implement Occam's Razor by finding the "simplest" valid model through sensitivity analysis, model ranking, and an iteration of calibration and validation tests. The results of numerical experiments are given to demonstrate the method with specific molecular structures, including polyethylene under axial extension. Once a CG model satisfying validation criteria is identified, macroscale models of materials can further be obtained through various homogenization techniques, which must be subjected

to calibration and validation, as well. These ideas are also discussed in Chapter 4. An illustrative example application to polyethylene is given in Chapter 5. Results relating frequentist and Bayesian ideas of model calibration under misspecification are presented in Chapter 6. Major conclusions of the work are collected in Chapter 7 together with recommendations for future work.

Chapter 2

Molecular Models

A universally-accepted approach to the study of the behavior of matter at atomistic scales is to employ the ideas of molecular dynamics (MD) in which the motion of systems of atoms is described by Newtonian mechanics and Hamiltonian dynamics. In parallel, Monte Carlo methods for modeling such systems, inspired by statistical thermodynamics provide, through the ergodic hypothesis, the basic tools that underpin much of modern materials science, chemistry, and engineering based on atomistic models. A number of hardened, verified, and well-documented MD codes have been in wide-spread use by the chemistry and materials science communities for decades and provide powerful machinery for studying atomistic systems.

However, in many significant applications, the atomic systems are of such an enormous size and complexity that it is impossible to solve the governing equations using the most advanced computers and computational tools available today. Thus, reduced models obtained by coarse graining must be used in the vast majority of investigations. The atomistic model, referred to here as the all-atom (AA) model is used as the "ground truth," the source of synthetic observational data to which results generated by the coarse-grained (CG) model are compared. This is the setting in which this study is done. The present chapter surveys the theory and

basic assumptions underlying a general class of AA models and the structures of various CG approximations.

2.1 All-Atom Models

Consider a system of n particles, the atoms in an atomistic model of a physical system. Invoking the Born-Oppenheimer approximation [14, 66, 87] allows an atom's position to be defined by a point mass at its center. Let the positions of the particles be given by coordinate vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$, and the corresponding momenta be given by the vectors $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$. A configuration is given by the collection $\mathbf{r}^n = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$, with $\mathbf{p}^n = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$. The set of possible configurations $\mathbf{r}^n$ and momenta $\mathbf{p}^n$ define a $6n-$dimensional phase space $\mathcal{H}_{AA}$. Classical mechanics in the form of Hamiltonian dynamics [36, 66, 116] are employed to describe the equations of motion of each atom,

$$\dot{\mathbf{r}}_i = \frac{\partial H}{\partial \mathbf{p}_i}, \qquad \dot{\mathbf{p}}_i = -\frac{\partial H}{\partial \mathbf{r}_i}, \qquad 1 \leq i \leq n, \tag{2.1}$$

where H is the Hamiltonian of the system, defined to be the sum of the kinetic energy K and potential energy V,

$$\begin{aligned} H(\mathbf{r}^n, \mathbf{p}^n) &= K(\mathbf{p}^n) + V(\mathbf{r}^n) && (2.2) \\ &= \sum_{i=1}^{n} \frac{\mathbf{p}_i \cdot \mathbf{p}_i}{2m_i} + V(\mathbf{r}^n), && (2.3) \end{aligned}$$

m_i being the mass of the i^{th} atom. The Hamiltonian describes the total energy of the n-body system.

The potential energy function, V, also referred to as simply a "potential" or a "force field," is at the heart of atomistic (and molecular) simulations. There

exist numerous commonly-used force fields, which are generally accepted as accurate in the MD literature. Of particular note are Assisted Model Building with Energy Refinement (AMBER) [20, 117], Chemistry at HARvard Macromolecular Mechanics (CHARMM) [16], and Optimized Potentials for Liquid Simulations (OPLS) [54, 55]. Here it will be assumed that the AA potential is described by the OPLS functional form,

$$V(\mathbf{r}^n) = V_{bond}(\mathbf{r}^n) + V_{angle}(\mathbf{r}^n) + V_{dihedral}(\mathbf{r}^n) + V_{nb}(\mathbf{r}^n), \tag{2.4}$$

where

$$V_{bond}(\mathbf{r}^n) = \sum_{i=1}^{n_b} k_{r,i}\left(r - r_{0,i}\right), \tag{2.5}$$

$$V_{angle}(\mathbf{r}^n) = \sum_{i=1}^{n_a} k_{\theta,i}\left(\theta - \theta_0\right), \tag{2.6}$$

$$\begin{aligned} V_{dihedral}(\mathbf{r}^n) = & \sum_{i=1}^{n_d} \frac{V_{1,i}}{2}\left(1 + \cos\left(\varphi\right)\right) + \frac{V_{2,i}}{2}\left(1 - \cos\left(2\varphi\right)\right) \\ & + \frac{V_{3,i}}{2}\left(1 + \cos\left(3\varphi\right)\right) + \frac{V_{4,i}}{2}\left(1 - \cos\left(4\varphi\right)\right), \end{aligned} \tag{2.7}$$

$$V_{nb}(\mathbf{r}^n) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{q_i q_j}{4\pi\epsilon_0 r} + 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] f_{ij}, \tag{2.8}$$

n_b, n_a and n_d are the number of bonds, angles, and dihedral interactions, respectively, in the AA system, it being understood that the arguments on the right hand sides of (2.5)-(2.8) are implicitly functions of the coordinates $\mathbf{r}^n$. Note that the non-bonded interactions, characterized by (2.8), are summed over all pairs of atoms in the system separated by three or more bonds, with $f_{ij} = 0.5$ if atoms i and j are separated by exactly three bonds and $f_{ij} = 1$ otherwise. While typical of most MD codes, many other functional forms of the force potentials may be used. The set

of parameters for the OPLS functional form of the AA model can be identified as $\boldsymbol{\theta}_{AA} = \{k_{r,i}, r_{0,i}, k_{\theta,i}, \theta_{0,i}, V_{1,i}, V_{2,i}, V_{3,i}, V_{4,i}, q_i, \epsilon_{ij}, \sigma_{ij}\}$; see Appendix A for a more detailed discussion.

The parameters of the AA model are deterministic and have known values, considered to represent reality. Virtual experiments may be performed on the AA system and measurements taken to produce data that are treated as ground truth. Details of how observational quantities are measured are given in Section 2.3. These quantities are used to determine prior information about the coarse-grained parameters, to inform and update our knowledge of CG parameter values in scenarios of varying complexity, and to confirm the validity of the CG model; these topics are discussed in detail in Chapter 3.

In subsequent applications, the number of particles n is fixed, the volume v in which the particles move is fixed, and the temperature T of the system is fixed, thus providing the basis for a canonical ensemble. In this case, the probability density of states is given by the Boltzmann distribution [66, 116],

$$\rho(\mathbf{r}^n, \mathbf{p}^n) = \frac{1}{Z} \exp\left(-\beta H\left(\mathbf{r}^n, \mathbf{p}^n\right)\right), \tag{2.9}$$

where Z is the partition function, which acts as a normalizing constant,

$$Z = \int_{\mathcal{H}_{AA}} \exp\left(-\beta H\left(\mathbf{r}^n, \mathbf{p}^n\right)\right)\, d\mathbf{r}^n d\mathbf{p}^n, \tag{2.10}$$

and is related to the Helmholtz free energy, A, according to

$$A = -\beta^{-1} \ln Z. \tag{2.11}$$

Here, $\beta = 1/k_B T$, k_B being Boltzmann's constant and T being the temperature [75, 116].

In general, the goal of constructing such an AA model is to compute certain thermodynamic properties of the system. These *quantities of interest* (or "QoIs") are defined by functions that map the phase space $\mathcal{H}_{AA}$ into the set of real numbers $\mathbb{R}$, called phase functions,

$$q : \mathcal{H}_{AA} \to \mathbb{R}, \quad q : (\mathbf{r}^n, \mathbf{p}^n) \mapsto q(\mathbf{r}^n, \mathbf{p}^n). \tag{2.12}$$

However, the value of this function at a single configuration is not of interest, as the probability of seeing any particular configuration is zero; instead, the average value is desired. This average may be a time average,

$$\bar{q} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau q\left(\mathbf{r}^n\left(t\right), \mathbf{p}^n\left(t\right)\right) \, dt, \tag{2.13}$$

or an ensemble average,

$$\langle q \rangle = \int_{\mathcal{H}_{AA}} q\left(\mathbf{r}^n, \mathbf{p}^n\right) \rho\left(\mathbf{r}^n, \mathbf{p}^n\right) \, d\mathbf{r}^n \, d\mathbf{p}^n, \tag{2.14}$$

ρ being the ensemble distribution function, such as that defined in (2.9). Thus the QoI, Q, may be defined as either

$$Q = \bar{q} \quad or \quad Q = \langle q \rangle . \tag{2.15}$$

Alternatively, due to the statistical behavior of atomistic systems in the ensemble setting, the QoI may actually be a distribution of values,

$$Q = \pi\left(q\right), \tag{2.16}$$

$\pi(\cdot)$ being a probability density.

Note that since the kinetic energy K is a quadratic function of the momenta, computation of (2.14) may be carried out analytically for phase functions that depend only on $\mathbf{p}^n$. Difficulty arises when the coordinates play a role. It is most often the case that phase functions of interest depend only on $\mathbf{r}^n$. Then averages of the form (2.14) in the canonical ensemble become

$$\begin{aligned}
\langle q \rangle &= \int_{\mathcal{H}_{AA}} q\left(\mathbf{r}^n\right) \frac{1}{Z} \exp\left(-\beta\left(K\left(\mathbf{p}^n\right)+V\left(\mathbf{r}^n\right)\right)\right)\, d\mathbf{r}^n d\mathbf{p}^n \\
&= \frac{\int_{\mathcal{H}_{AA}} q\left(\mathbf{r}^n\right) \exp\left(-\beta V\left(\mathbf{r}^n\right)\right)\, d\mathbf{r}^n \exp\left(-\beta K\left(\mathbf{p}^n\right)\right)\, d\mathbf{p}^n}{\int_{\mathcal{H}_{AA}} \exp\left(-\beta V\left(\mathbf{r}^n\right)\right)\, d\mathbf{r}^n \exp\left(-\beta K\left(\mathbf{p}^n\right)\right)\, d\mathbf{p}^n} \\
&= \frac{\int_{\mathcal{H}_{AA}} q\left(\mathbf{r}^n\right) \exp\left(-\beta V\left(\mathbf{r}^n\right)\right)\, d\mathbf{r}^n}{\int_{\mathcal{H}_{AA}} \exp\left(-\beta V\left(\mathbf{r}^n\right)\right)\, d\mathbf{r}^n}. \qquad (2.17)
\end{aligned}$$

Although momenta no longer play a role, the computation of this ensemble average remains challenging since, in general, analytical evaluation is not possible. A discussion of various numerical methods can be found in [36, 66], and details concerning their use in this work can be found in Section 2.3.

2.2 Coarse-Grained Models

The same atomistic system described above may be recast on a coarser scale using a so-called coarse-grained (CG) model. The system of n atoms described above may now be represented by N CG particles or "beads," where, in general, $N \ll n$. The theory behind simulating a physical system using a CG model is the same as that of the AA system. The positions of the coarse-grained particles are given by the coordinate vectors $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N$, and the momenta are given by the vectors $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N$, with $\mathbf{R}^N = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N\}$ and $\mathbf{P}^N = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N\}$.

The set of possible configurations and momenta of this mesoscale system defines a $6N-$dimensional phase space $\mathcal{H}_{CG}$ of the CG system. The particles move according to Hamilton's equations of motion,

$$\dot{\mathbf{R}}_I = \frac{\partial H_{CG}}{\partial \mathbf{P}_I}, \qquad \dot{\mathbf{P}}_I = -\frac{\partial H_{CG}}{\partial \mathbf{R}_I}, \qquad 1 \leq I \leq N, \tag{2.18}$$

where H_{CG} is the Hamiltonian of the coarse-grained system, defined to be

$$\begin{aligned} H_{CG}\left(\mathbf{R}^N, \mathbf{P}^N; \boldsymbol{\theta}\right) &= K_{CG}\left(\mathbf{P}^N\right) + V_{CG}\left(\mathbf{R}^N; \boldsymbol{\theta}\right) & (2.19) \\ &= \sum_{I=1}^{N} \frac{\mathbf{P}_I \cdot \mathbf{P}_I}{2M_I} + V_{CG}\left(\mathbf{R}^N; \boldsymbol{\theta}\right), & (2.20) \end{aligned}$$

with K_{CG} the kinetic energy of the system, depending only on the momentum $\mathbf{P}_I$ and mass M_I of each particle, and V_{CG} the potential energy of the system, depending only on the particle coordinates $\mathbf{R}^N$ and the parameters of the CG model, $\boldsymbol{\theta}$.

The potential, V_{CG}, dictates the behavior and motion of the coarse model. Not only is the functional form unknown, the parameters $\boldsymbol{\theta}$ are unknown, and they must be chosen to make the CG model deliver QoIs as close as possible to those of the AA model. That is, the difference between specific quantities measured in the CG system and those measured in the AA system must fall within a given tolerance. Measures of "closeness" are given in Section 3.5. *Determining V_{CG} so that the coarse-grained model accurately represents, and therefore may be used as a surrogate of, the full atomistic model is a main goal of this work.*

Phase functions analogous to those defined in the AA system may be defined in the CG system such that

$$q_{CG} : \mathcal{H}_{CG} \to \mathbb{R}, \quad q_{CG} : \left(\mathbf{R}^N, \mathbf{P}^N\right) \mapsto q_{CG}\left(\mathbf{R}^N, \mathbf{P}^N\right), \tag{2.21}$$

where q_{CG} represents the same observable quantity or property as q, but constrained to the CG system. Note that since the motion of the system, and therefore $\mathbf{R}^N$, depends on CG model parameters $\boldsymbol{\theta}$, q_{CG} implicitly depends on $\boldsymbol{\theta}$. To recognize this dependence, the time average is written

$$\bar{q}_{CG}(\boldsymbol{\theta}) = \lim_{\tau\to\infty} \int_0^\tau q_{CG}\left(\mathbf{R}^N(t), \mathbf{P}^N(t); \boldsymbol{\theta}\right) dt, \tag{2.22}$$

and the ensemble average is likewise denoted

$$\langle q(\boldsymbol{\theta})\rangle_{CG} = \int_{\mathcal{H}_{CG}} q_{CG}\left(\mathbf{R}^N, \mathbf{P}^N; \boldsymbol{\theta}\right) \rho_{CG}\left(\mathbf{R}^N, \mathbf{P}^N; \boldsymbol{\theta}\right) d\mathbf{R}^N d\mathbf{P}^N. \tag{2.23}$$

The representation of the probability distribution of states for the coarse-grained system in a canonical ensemble is, again, given by the Boltzmann distribution,

$$\rho_{CG}\left(\mathbf{R}^N, \mathbf{P}^N; \boldsymbol{\theta}\right) = \frac{1}{Z_{CG}(\boldsymbol{\theta})} \exp\left(-\beta H_{CG}\left(\mathbf{R}^N, \mathbf{P}^N; \boldsymbol{\theta}\right)\right), \tag{2.24}$$

where the partition function Z_{CG} is given for each parameter vector $\boldsymbol{\theta}$ by

$$Z_{CG}(\boldsymbol{\theta}) = \int_{\mathcal{H}_{CG}} \exp\left(-\beta H_{CG}\left(\mathbf{R}^N, \mathbf{P}^N; \boldsymbol{\theta}\right)\right) d\mathbf{R}^N d\mathbf{P}^N, \tag{2.25}$$

with β defined as before.

Thus, the calculation of the quantity of interest in the CG system, analogous to (2.15), is

$$Q_{CG} = \bar{q}_{CG}(\boldsymbol{\theta}) \quad or \quad Q_{CG} = \langle q(\boldsymbol{\theta})\rangle_{CG}. \tag{2.26}$$

As previously mentioned, the statistical nature of the behavior of particulate systems may give rise instead to a distribution of possible values of the QoI,

$$Q_{CG} = \pi_{CG}(q_{CG}(\boldsymbol{\theta})), \tag{2.27}$$

with $\pi_{CG}(\cdot)$ a probability distribution.

It should be noted that the quantity of interest (2.26) depends on the chosen definition of V_{CG}, which has its own parameters $\boldsymbol{\theta}$, and which depends on the choice of coarse-grained representation G of the all-atom system, as discussed in Section 2.2.1. That is,

$$Q_{CG} = Q_{CG}\left(G, \boldsymbol{\theta}\right). \tag{2.28}$$

2.2.1 Mapping the All-Atom System into the Coarse-Grained System

A fundamental question, at this point, is what are the relationships between the $\mathbf{r}^n$ and the $\mathbf{R}^N$? Noid *et al.*, point out that "each CG coordinate can be assigned a well-defined physical meaning in terms of the coordinates of the atomistic model" [81]. It is therefore convenient to regard the coordinates of each CG particle or bead, $\mathbf{R}_I$, as the image of a surjective map G of the AA coordinates $\mathbf{r}^n$ to the CG coordinates $\mathbf{R}^N$ [81, 82],

$$\mathbf{R}_I = G(\mathbf{r}^n), \quad I = 1, 2, \ldots, N. \tag{2.29}$$

Coarse-grained beads are idealizations of spheres (or ellipsoids) that envelop groups of atoms, and one can imagine that atoms "belong" to certain CG particles. Thus, for CG bead I, an index set $\mathcal{I}_I$ may be identified that contains indices of all atoms belonging to that bead. Then, the map G may be defined, for example, such that $\mathbf{R}_I$ is the center of mass of the atoms in the set $\mathcal{I}$,

$$\mathbf{R}_I = \frac{\sum_{i \in \mathcal{I}_I} m_i \mathbf{r}_i}{\sum_{i \in \mathcal{I}_I} m_i}, \tag{2.30}$$

or as the mean vector,

$$\mathbf{R}_I = \frac{1}{|\mathcal{I}_I|} \sum_{i \in \mathcal{I}_I} \mathbf{r}_i, \tag{2.31}$$

where m_i is the mass of atom i and $|\mathcal{I}_I|$ is the cardinality of the index set $\mathcal{I}_I$.

Clearly, the definition of the map G is largely heuristic and non-unique. The atom groups $\mathcal{I}_I$ may be defined based on "chemical intuition" about the structural, interactive, or reactive behavior of the all-atom system or according to a specification, *e.g.* such that every coarse-grained bead is charge-neutral. Some applications or scenarios may allow coarser mappings than others. Very large systems may mandate coarser mappings due to computational cost or limitations. Thus for each atomistic system, there are many appropriate representations G, as illustrated in Figure 2.1. A methodology for determining the *best* map G is presented in Chapter 4. Using this method, the optimal map G and the most favorable representation of CG potential V_{CG}, along with its accompanying parameters $\boldsymbol{\theta}$, are determined simultaneously.

2.3 Calculation of Observables Using Molecular Dynamics

The calculation of key observables and quantities of interest is the primary purpose of any model. This calculation is not just performed for the purposes of prediction; it is used for producing data for model calibration and validation in the Bayesian framework, discussed in Chapter 3, upon which the research presented here is based. In the case of virtual experimentation, as is the case here, observables and data need to be collected in both the ground truth or high-fidelity model (the AA model) and the surrogate or reduced-order model (the CG model) for the purposes

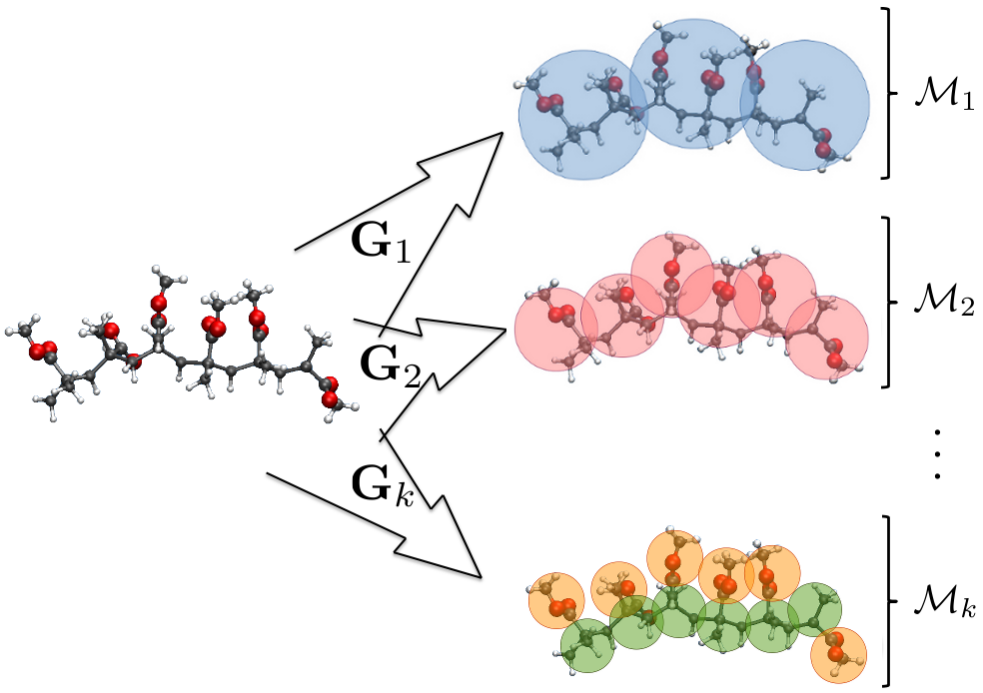

Figure 2.1: Illustration of various choices for coarse-grained mappings of an atomistic system. For each G_i, a set of possible parametric model classes may be defined, as discussed in Section 2.4.

of calibration and comparison. In this section, the methods used in this work for calculating observables in molecular systems are discussed. The details given use the notation from the AA system where convenient; application to the CG system is trivial.

Observables and quantities of interest are defined by phase functions q and q_{CG} in the AA and CG systems, respectively, and can take the form of a time average, an ensemble average, or a probability distribution, as previously noted in

(2.15)−(2.16) and (2.26)−(2.27). In any of these cases, the system is simulated using an MD code. In the present study, the AA and CG systems are simulated using a software package developed and maintained by Sandia National Laboratory, called Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [96]. A data file describing the atomistic or molecular system of interest is created; it contains the initial positions of the atoms as well as topological information, such as which atoms are bonded together and a tabulation of all angular and dihedral interactions. This file is read by an input script, in which the functional form of the potential energy function, V, and its parameter values are specified. This input script also contains information about the simulation itself: the ensemble, including the temperature or pressure specifications; equilibration steps according to specified time lengths; and external body forces, such as strain rate, applied force, and so on. Finally, the functional form of the phase function, q, is specified in the input script.

The initial positions and topological information included in the data file are read into the LAMMPS input file. Then an equilibration process is performed to ensure that the system is in thermodynamic equilibrium. A short period of Brownian dynamics is performed at a temperature, T_B, much higher than the desired value. This introduces a sufficient amount of randomness to the system to break any symmetries that may be falsely imposed by the specified initial configuration. This purpose is further served by simulating the system in an isobaric-isothermal ensemble, also called an NPT ensemble, in which the number of particles is fixed and a constant pressure P and temperature T_B are maintained. Periodic boundary conditions are used, allowing for the movement of molecules through the material.

The temperature is slowly lowered from elevated T_B to the desired temperature, T. This step, called quenching or annealing, preserves the randomness introduced by the NPT simulation, and brings the system to an approximate equilibrium configuration [59]. The criteria for what constitutes an equilibrium state vary depending on the imposed macroscopic conditions, *e.g.* the number of particles, volume, pressure, temperature, *etc.*, that define the ensemble. In the canonical ensemble, the Boltzmann distribution (2.9) describes a system in which the most likely states are those for which the total energy, given by the Hamiltonian, is minimized. Thus, simulations should also seek configurations that yield low potential energy.

Once the temperature of the system is quenched to the preferred temperature, T, simulation in a canonical ensemble is performed. Note that the volume, V, may have changed from that specified by the initial configuration contained in the datafile during the NPT simulation, but is now held at the volume produced by quenching. The duration of the canonical ensemble simulation may vary due to many factors, such as the size of the system, but should continue until the modeler is confident that residual effects from the quenching step are negligible. This can be thought of as a "burn-in" period for the simulation of a canonical ensemble.

Steps following equilibration depend on the modeling scenario. For simplicity, applications considered here are restricted to scenarios in which simulation in the canonical ensemble is continued following equilibration and in which the observable or quantity of interest is one that is measured in equilibrium. Recalling the intractability of analytically calculating (2.13) or (2.14), numerical calculation is re-

quired. Though many techniques have been developed for this purpose, an overview of which is given in [36, 66], Monte Carlo approximation is used here. Given M independent and identically distributed ($i.i.d$) samples $\{x_i\}_{i=1}^{M}$,

$$\int f(x)\,dx \approx \frac{1}{M}\sum_{i=1}^{M} f(x_i). \tag{2.32}$$

In particular, given $i.i.d$ samples of the phase function q at times $\{t_i\}$,

$$\bar{q} \approx \frac{1}{M}\sum_{i=1}^{M} q\left(\mathbf{r}^n\left(t_i\right), \mathbf{p}^n\left(t_i\right)\right), \tag{2.33}$$

or given $i.i.d$ samples at points of the phase space $\{(\mathbf{r}_i^n, \mathbf{p}_i^n)\}$,

$$\langle q \rangle \approx \frac{1}{M}\sum_{i=1}^{M} q\left(\mathbf{r}_i^n, \mathbf{p}_i^n\right)\rho\left(\mathbf{r}_i^n, \mathbf{p}_i^n\right). \tag{2.34}$$

Thus, by sampling the value of the phase function of the observable or quantity of interest throughout the MD simulation, the time and ensemble average may be approximated.

In LAMMPS, the Nosé-Hoover thermostat is used to enforce the temperature constraint [42, 84, 96]. Instead of integrating the equations of motion (2.1) forward in time, the Nosé-Hoover thermostat is enforced by sampling positions and velocities (momenta) from the canonical distribution function (2.9). This implementation is acceptable under the validity of the ergodic hypothesis, which states that given a sufficient length of time, the AA system will access all points in $\mathcal{H}_{AA}$, and may therefore be represented by statistical sampling, so that

$$\bar{q} = \langle q \rangle. \tag{2.35}$$

It can be shown [83] that if the samples $(\mathbf{r}_i^n, \mathbf{p}_i^n)$ are drawn according to a non-uniform distribution, in this case the Boltzmann distribution (2.9), the approximation (2.34) becomes

$$\langle q \rangle \approx \frac{1}{M} \sum_{i=1}^{M} q\left(\mathbf{r}_i^n, \mathbf{p}_i^n\right) \tag{2.36}$$

through a change of variables. Thus, for *i.i.d* samples $\{(\mathbf{r}_i^n, \mathbf{p}_i^n)\}$ taken from a LAMMPS simulation using the Nosé-Hoover thermostat, the calculation of the ensemble average of a phase function amounts to the arithmetic mean (2.36).

2.4 Uncertainties in the Coarse-Grained Model

The significant problems with coarse-grained approximations of atomistic systems are clear:

1. The choice of map G, and therefore the number of CG particles N, is not well-defined. If a given atomistic system may be represented by k different maps, the set

$$\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k\} \tag{2.37}$$

defines a collection of sets $\mathcal{M}_i$ of model classes that may be used to model the CG system yielded by the map G_i.

2. Once G is specified, the choice and structure of V_{CG} remains to be determined. For each map G_i, a set $\mathcal{M}_i$ of possible representations of V_{CG} can be identified. Each representation of V_{CG}, denoted $\mathcal{P}_{ij}$, has its own set of parameters $\boldsymbol{\theta}_{ij}$ and therefore is an infinite class of possible models. Thus, each $\mathcal{M}_i$, is the set of

possible models that may be used with map G_i,

$$\mathcal{M}_i = \{\mathcal{P}_{i1}(\boldsymbol{\theta}_{i1}), \mathcal{P}_{i2}(\boldsymbol{\theta}_{i2}), \ldots, \mathcal{P}_{im}(\boldsymbol{\theta}_{im})\}, \quad i = 1, 2, \ldots, k. \tag{2.38}$$

3. The parameters $\boldsymbol{\theta}_{ij}$ for each model $\mathcal{P}_{ij}$, corresponding to map G_i and representation j of V_{CG}, are unknown and posses uncertain (random) vectors. These parameters must be calibrated to agree with data provided by the AA system, which is, itself, stochastic.

Each step in the coarse graining process is met with choice and uncertainty. The methodology chosen here for the calibration and validation of models under uncertainty is the use of Bayes' Rule, discussed in the next chapter.

Chapter 3

Bayesian Framework for Predictive Models

The accuracy with which scientific theories depict observed phenomena, or more importantly, predict events not yet observed, is critical to the success of the scientific method. Exactly how reliable are the models of natural phenomena that are used to make critical predictions on the behavior of physical or engineered systems? This question is at the heart of all scientific discovery; implicitly it embraces the notion that observational data are to be obtained and that models, the mathematical transcriptions of theories, will be used to explain with sufficient accuracy the observational data and then be used to extrapolate to the prediction of unobserved events. As has been stressed repeatedly, the confounding difficulty in applying this fundamental scientific strategy is to cope with inevitable uncertainties in data, in the selection of appropriate models, in model parameters, and in assessing the adequacy

K. Farrell and J. T. Oden. Calibration and validation of coarse-grained models of atomic systems: Application to semiconductor manufacturing. *Computational Mechanics*, 54(1):3-19,2014. K. Farrell implemented Bayesian methods computationally. J. T. Oden supervised the work. Bayesian theories introduced for application to coarse-grained models were developed by both authors.

K. Farrell, J. T. Oden, and D. Faghihi. A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems. *Journal of Computational Physics*, 295:189-208, 2015. K. Farrell implemented Bayesian methods computationally. J. T. Oden supervised the work. D. Faghihi implemented the analysis of parameter sensitivities.

of the model itself.

The path to scientific discovery must include the acquisition of experimental or observational data and the interpretation of the causes of the observations through inductive hypotheses that take the form here of parametric mathematical models based on physical principles. To illustrate, consider as a starting point the abstract mathematical model,

$$A\left(\boldsymbol{\theta}, S; u\left(S, \boldsymbol{\theta}\right)\right) = 0, \tag{3.1}$$

where A is a collection of mathematical operations, arising from a mathematical characterization of scientific theories, $\boldsymbol{\theta}$ is a vector of model parameters, S is the scenario under which the model is implemented, and u is the solution of this so-called forward problem, which depends on S and $\boldsymbol{\theta}$. The scenario is generally understood to be the domain and initial and boundary conditions for which the model is implemented. For simplicity, and without loss of generality, it will be assumed that the scenarios are deterministically defined and are independent of $\boldsymbol{\theta}$. It is assumed that the solution u exists for each choice of scenario S and parameter vector $\boldsymbol{\theta}$ belonging to a parameter space Θ.

The fundamental questions arising in all such modeling processes are: 1) How can the parameters be chosen to represent our best knowledge of the physics in the presence of uncertainties? 2) How accurately does the model depict physical reality, particularly the QoI? and 3) How can the inevitable uncertainties in the simulation be quantified, particularly the uncertainty in the QoI? Answering these questions is the aim of this chapter.

3.1 Bayes' Rule

Throughout this investigation, uncertainties are dealt with in the context of probability theory. But which probability theory is most appropriate for the tasks at hand? This work subscribes to the logical theory of probability laid down by Richard T. Cox [22], expanded and implemented in the treatise of Edwin T. Jaynes [53], and debated extensively in statistics literature [4, 5, 28, 39, 40, 113]. There, the foundational conclusion is that, under reasonable "desiderata," every extension of Aristotelian logic in the presence of uncertainties is Bayesian. More specifically, according to Cox's Theorem, the product rule of logical plausibilities $p(\cdot) \in \mathbb{R}^+$ of propositions (or events) A and B conditioned on prior knowledge X is given by

$$p(A \wedge B|X) = p(A|X)p(B|AX), \tag{3.2}$$

where $A \wedge B$ is the conjunction of A and B (the proposition that both A and B are true), and $p(A|X)$ is the proposition that A is true, given that (or on condition that) X is true. The rule (3.2) is precisely Bayes' Rule. Owing to the commutativity of the operator $\wedge$, it is written more familiarly as

$$p(A|BX) = \frac{p(B|AX)p(A|X)}{p(B|X)}. \tag{3.3}$$

Traditionally, $p(A|X)$ describes prior information on A conditioned on X, $p(B|AX)$ is the likelihood of A conditioned on the premise that both B and X are true and are given, and $p(B|X)$, called the evidence, is a normalization factor. Hereafter, with proper scaling and with the sum rule ($p(A|X)+p(\bar{A}|X) = 1$, $\bar{A}$ being the negation of A) in force, the propositions in (3.3) can be replaced by probability

densities. Interpreted in the more common context of Kolmogorov probability theory [65], which is to be called upon when convenient, probability measures are assumed to be absolutely continuous with respect to the Lebesgue measure and, therefore, representable by probability densities.

Returning to the abstract model (3.1), the goal now is to use the model to match as best as possible observational data $\mathbf{y}$ for given $\boldsymbol{\theta}$, and, denoting the corresponding probability densities by $p \sim \pi$, an analogy to (3.3) is the equation

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})}, \tag{3.4}$$

where $\pi(\boldsymbol{\theta})$ is the prior probability density function (pdf) for parameters $\boldsymbol{\theta}$, $\pi(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood, the conditional probability of $\mathbf{y}$ given $\boldsymbol{\theta}$, $\pi(\mathbf{y})$ is the evidence, and $\pi(\boldsymbol{\theta}|\mathbf{y})$ the posterior probability density.

Bayes' Rule, (3.4), is the point-of-departure for statistical calibration and validation of models given a scenario S. It represents a recipe for updating model parameters: starting with prior information $\pi(\boldsymbol{\theta})$, the likelihood measures the probability of the data for given $\boldsymbol{\theta}$, the posterior being the Bayesian update for the model parameters.

3.2 The Prediction Pyramid

The processes leading to the prediction of a QoI are conceptualized as traversing a hypothetical pyramid, the prediction pyramid, shown in Figure 3.1. As discussed in [1, 6, 90–92], a sequence of calibration and then validation scenarios are designed to inform the mathematical model and build the modeler's confidence

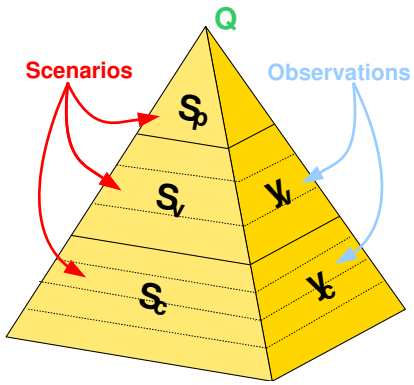

Figure 3.1: The prediction pyramid that illustrates the construction of predictive models [89–92]. An overview of this process is given in Section 3.2 and discussed in detail throughout this chapter.

in its ability to accurately predict the QoI. At the base of the pyramid is the calibration step, where the simplest, informative calibration scenario, S_c, yields calibration data $\mathbf{y}_c$. This data is used to update any prior information that may be known about the parameters. Progressing up the pyramid to more complicated validation scenarios, S_v, closer to the prediction scenario, the calibrated model is tested against validation data, $\mathbf{y}_v$. The implication is that as higher levels of the pyramid are reached, new information is used to update the model parameters, building upon knowledge gained in previous levels. In general, the accessibility of data at higher levels decreases, though this may not always be the case. After passing the validation tests, the model may be used to predict the unobservable QoI in the full prediction scenario at the top of the pyramid.

To illustrate, the prediction scenario shown in Figure 3.2 consists of a complex structure made up of polymeric material. This structure may be modeled

through the use of a representative volume element [7, 8], which therefore serves as a prediction scenario. Within this cube of material are long, interlocking polymer chains, which are thus deemed validation scenarios. Longer chains may be broken down into smaller chains that become the calibration scenarios.

It is imperative that all of the parameters $\boldsymbol{\theta}$ are used in the calibration scenarios so that they may be updated by observational measurements. If necessary, multiple calibration scenarios may be used to ensure that all parameters are included, as discussed in detail in [31]. Ideally, the validation stage should consist of multiple validation scenarios, each of increasing size and complexity, while simulta-

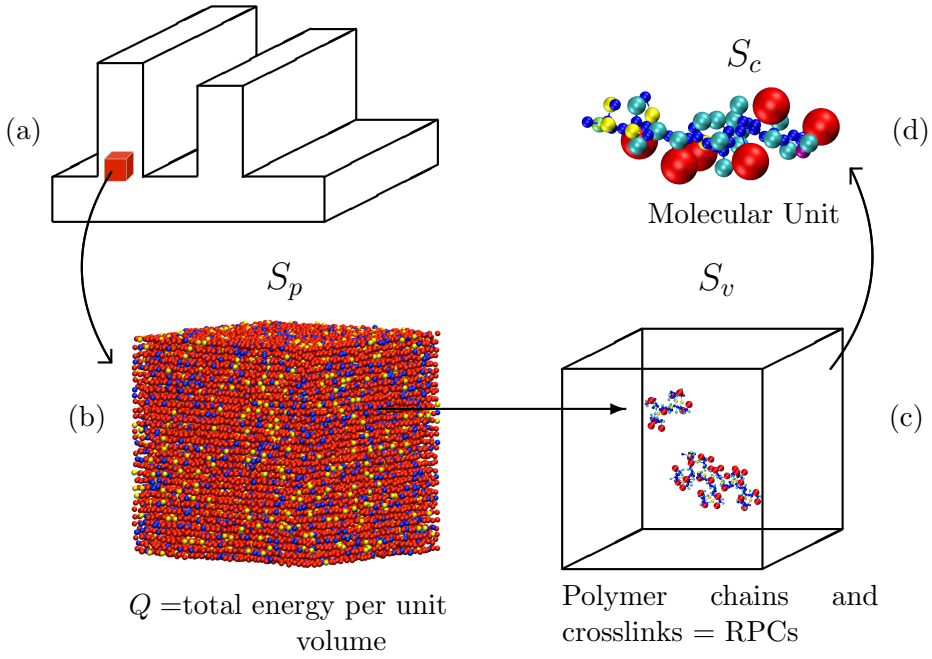

Figure 3.2: Illustration of the prediction, validation, and calibration scenarios for predictive models of polymer materials used in semiconductor manufacturing (see [31] for details): a) the desired pattern of material; b) one realization of a cube representing the representative volume element [7, 8]; c) validation chains and d) molecular units used in the calibration scenario

neously providing as much new information about the parameters as possible (see Section 3.5). Furthermore, the calibration and validation data against which the model will be compared should be related to the QoI. A fundamental tenant in the art and science of model validation is the notion that the validity of a model is only meaningful in connection with specific QoIs; a model may be deemed valid for predicting one QoI but invalid for another [92]. Heuristic arguments may be made for choosing $\mathbf{y}_c$ and $\mathbf{y}_v$, but a less arbitrary method is discussed in Section 4.2.

The calibration and validation scenarios are chosen as subsystems of the prediction scenario in an attempt to capture as many of the characteristics and behavior of S_p as possible. The data $\mathbf{y}_c$ and $\mathbf{y}_v$ are chosen to inform the parameters about the QoI. In this setting, traversing up the prediction pyramid through increasingly complex trials builds the modeler's confidence in the ability of the model to predict QoIs based on the accuracy with which the model can predict specific observables in validation scenarios.

3.3 Model Calibration

The outset of predictive modeling begins with calibration. As discussed previously, the calibration scenario, S_c, is the simplest possible scenario that is a subsystem of the prediction scenario and it is where data, $\mathbf{y}_c$, related to the QoI, are collected. In this scenario, the model parameters $\boldsymbol{\theta}$ are updated via Bayes' Rule, (3.4), with observational data $\mathbf{y} = \mathbf{y}_c$ such that

$$\pi\left(\boldsymbol{\theta}|\mathbf{y}_c\right) = \frac{\pi\left(\mathbf{y}_c|\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\pi\left(\mathbf{y}_c\right)}. \tag{3.5}$$

The likelihood distribution, $\pi(\mathbf{y}_c|\boldsymbol{\theta})$, characterizes how well the model, given the parameter choice $\boldsymbol{\theta}$, is able to reproduce the observed data. That is, according to Ronald A. Fisher [33] (who was, ironically, a staunch anti-Bayesian), "[t]he 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." The likelihood distribution is discussed in detail in Section 3.3.1.

The prior distribution, $\pi(\boldsymbol{\theta})$, captures all prior information that may be known about the parameters before calibration begins. To capture all prior information while simultaneously accounting for all possible uncertainties in this data, the maximum entropy approach of Claude E. Shannon [105] is used. Details are discussed in Section 3.3.2.

The evidence, $\pi(\mathbf{y}_c)$, is usually regarded as a constant whose purpose is merely to normalize the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{y}_c)$. However, the evidence can be used to calculate *the plausibility of a model* and can be used to select, from a group of models, the model that best fits the given data. This criteria for model selection is the basis of the present work. The evidence and its relationship to model selection are discussed in Sections 3.3.3 and 3.4.

3.3.1 The Likelihood Distribution

Observational data belong to a data set $\mathcal{Y}$, which can be endowed with mathematical structure, such as a metric topology. The observations $\mathbf{y} \in \mathcal{Y}$ are regarded as vectors of samples: $\mathbf{y} = \{y_1, y_2, \ldots, y_n\} \in \mathbb{R}^n$. The sample y_i is acquired in an attempt to measure a physical reality Ω_i at some point in the physical universe (*e.g.*

temperature or displacement or energy), but reality is seldom measurable exactly due to experimental noise ε_i, here assumed to be additive. Thus,

$$\Omega_i + \varepsilon_i = y_i, \quad 1 \leq i \leq n. \tag{3.6}$$

The observational error ε_i itself is uncertain; it is, in general, a random variable. The data $\mathcal{Y}$ are collected for a specific scenario S of the physical events of interest.

The adequacy with which the model can predict reality depends upon the closeness of the model predictions $d_i\left(u\left(S, \boldsymbol{\theta}\right)\right)$ to the physical reality Ω_i for scenario S, and are assumed to satisfy the relations,

$$\Omega_i = d_i\left(\boldsymbol{\theta}\right) + \eta_i, \quad 1 \leq i \leq n, \tag{3.7}$$

where $d_i\left(\boldsymbol{\theta}\right) \equiv d_i\left(u\left(S, \boldsymbol{\theta}\right)\right)$ is the parameter-to-observation map produced by the model and η_i represents the modeling error, also referred to as the model inadequacy or model form error. Eliminating Ω_i from (3.6) and (3.7) yields

$$y_i - d_i\left(\boldsymbol{\theta}\right) = \varepsilon_i + \eta_i, \quad 1 \leq i \leq n. \tag{3.8}$$

In much of the literature on statistical calibration of models in the presence of uncertainty, the model inadequacy η_i is ignored and the total discrepancy between the data and the model is attributed to experimental noise, the modeling error being treated separately or ignored completely.

The effects of the model are manifested in the likelihood probability $\pi(\mathbf{y}|\boldsymbol{\theta})$: if p is the probability density of the error $\varepsilon_i + \eta_i$ in (3.8), then

$$p\left(\varepsilon_i + \eta_i\right) = p\left(y_i - d_i\left(\boldsymbol{\theta}\right)\right) = \pi\left(y_i|\boldsymbol{\theta}\right), \tag{3.9}$$

and

$$\pi(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi\left(y_i|\boldsymbol{\theta}\right). \tag{3.10}$$

Note that is it often convenient to use the log-likelihood,

$$\log \pi\left(\mathbf{y}|\boldsymbol{\theta}\right) = \sum_{i=1}^{n} \log \pi\left(y_i|\boldsymbol{\theta}\right), \tag{3.11}$$

which has the same maximum as (3.10) due to the monotonicity of the logarithmic function.

Unless other information is known about the form of the error ε_i, it is common to assume that it is normally distributed about zero with variance σ^2. That is,

$$\pi(y_i|\boldsymbol{\theta}) \sim \mathcal{N}\left(0, \sigma^2\right). \tag{3.12}$$

In general, the variance σ^2 is unknown, but it may be added to the list of parameters $\boldsymbol{\theta}$. In this way, no unintentional bias is being added to the calibration process.

3.3.2 The Prior Distribution

Any and all prior information that is known about the parameters is characterized in the prior probability distribution, $\pi\left(\boldsymbol{\theta}\right)$. It must, however, be noted that this prior information contains its own uncertainties regarding the extrapolation of this information to new scenarios (*i.e.* the calibration scenarios) as well as the usual uncertainties introduced by experimental or measurement error.

Claude E. Shannon published a theorem in 1948 [105] proving that the only consistent measure, H, of the uncertainty content of a probability distribution, p satisfying four logical *desiderata* [53],

1. $H(p) \in \mathbb{R}$: the amount of uncertainty is a real number,

2. $H \in C^0$: H is a continuous function,

3. H is monotonic: the more possibilities from which there are to choose, the more uncertainty there is in the choice,

4. H is consistent: if there is more than one appropriate functional form, they all yield the same answer,

is the information entropy,

$$H(p) = -\sum_{i=1}^{m} p(\theta_i) \log p(\theta_i). \tag{3.13}$$

In the case that p is continuous,

$$H(p) = -\int p(\theta) \log p(\theta)\, d\theta. \tag{3.14}$$

This function is also called the Shannon or information entropy.

In a prior distribution, one seeks a distribution that captures any prior information that may be known about the model parameters, $\boldsymbol{\theta}$, while simultaneously acknowledging the fact that this information may also be uncertain, *i.e.* in its shape, range, *etc.* Thus, the proper choice for the prior distribution is the one that maximizes the information entropy (the uncertainty content) subject to constraints imposed by prior knowledge. This choice of prior is aptly named a "maximum entropy prior," and the method of deriving such priors was introduced by Edwin T. Jaynes in the 1950s [50–52].

To maximize the information entropy subject to constraints, Lagrange multipliers are used. Without loss of generality, let the prior probability distribution, p, be a discrete distribution, and for notational simplicity, let $p_i = p(\theta_i)$. At the very least, it must be that p is normalized such that,

$$\sum_{i=1}^{m} p_i = 1. \tag{3.15}$$

Relevant to this work are the additional constraints that the mean,

$$\sum_{i=1}^{m} p_i \theta_i = \langle \theta \rangle , \tag{3.16}$$

and possibly also the variance,

$$\sum_{i=1}^{m} p_i \left(\theta_i - \langle \theta \rangle\right)^2 = \sigma_\theta^2, \tag{3.17}$$

are known. In the case that only the mean of a parameter θ is known, the Lagrangian has the form,

$$\mathcal{L} = H(p) - (\lambda_0 - 1)\left(\sum_{i=1}^{m} p_i - 1\right) + \lambda_1 \left(\sum_{i=1}^{m} p_i \theta_i - \langle \theta \rangle\right). \tag{3.18}$$

If, in addition, the variance is known,

$$\begin{aligned} \mathcal{L} \;=\; & H(p) - (\lambda_0 - 1)\left(\sum_{i=1}^{m} p_i - 1\right) + \lambda_1 \left(\sum_{i=1}^{m} p_i \theta_i - \langle \theta \rangle\right) \\ & -\lambda_2 \left(\sum_{i=1}^{m} p_i \left(\theta_i - \langle \theta \rangle\right)^2 - \sigma_\theta^2\right). \end{aligned} \tag{3.19}$$

The Lagrangian $\mathcal{L}$ is then maximized with respect to p_i such that

$$\frac{\partial \mathcal{L}}{\partial p_i} = 0 \tag{3.20}$$

for each $i = 1, \ldots, m$. This equation, coupled with the constraints (3.15), (3.16), and, if applicable, (3.17) leads to a prior distribution whose information entropy is maximized. In the case that only the mean is known, the maximum entropy prior is an exponential distribution,

$$p_i = \frac{1}{\langle \theta \rangle} \exp\left(\frac{-\theta_i}{\langle \theta \rangle}\right), \tag{3.21}$$

and if the variance is also known, p is a Gaussian distribution,

$$p_i = \frac{1}{\sigma_\theta \sqrt{2\pi}} \exp\left(\frac{-\left(\theta_i - \langle \theta \rangle\right)^2}{2\sigma_\theta^2}\right). \tag{3.22}$$

These derivations and expressions extend trivially to the continuous case.

Unless prior information suggests otherwise, parameters are treated as independent random variables. Thus, maximum entropy prior distributions are derived for each parameter independently, and the prior distribution, $\pi(\theta)$, used in Bayes' Rule (3.5) is a concatenation of the prior distributions of all of the parameters.

3.3.3 The Evidence

The denominator of Bayes' Rule (3.5) is called the evidence. It is a marginalization of the likelihood distribution over the parameters $\boldsymbol{\theta}$,

$$\pi(\mathbf{y}_c) = \int \pi(\mathbf{y}_c|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{3.23}$$

and serves as a normalizing factor for the posterior parameter distribution. Given a single model, the parameters $\boldsymbol{\theta}$ may be calibrated to observational data via (3.5). In this setting, the evidence, $\pi(\mathbf{y}_c)$, is often overlooked, and one frequently sees the rule written as,

$$\pi(\boldsymbol{\theta}|\mathbf{y}_c) \propto \pi(\mathbf{y}_c|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}). \tag{3.24}$$

However, when given a choice of many possible models, the evidence plays a crucial role in model selection, discussed in the next section.

3.4 Model Plausibility and Model Selection

It is often the case that the abstract mathematical model A of (3.1) may be represented by more than one possible functional form. This situation was briefly discussed in Section 2.4 in the context of modeling the coarse-grained approximation of an atomistic system. Thus, consider a set of model classes,

$$\mathcal{M} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m\}, \tag{3.25}$$

each of which has its own parameters $\boldsymbol{\theta}_j$, $j = 1, \ldots m$ and may possibly be used to represent the behavior of the system of interest. For each model pair, $(\mathcal{P}_j, \boldsymbol{\theta}_j)$, the Bayesian rule may be written,

$$\pi\left(\boldsymbol{\theta}_j|\mathbf{y}, \mathcal{P}_j, \mathcal{M}\right) = \frac{\pi\left(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{P}_j, \mathcal{M}\right)\pi\left(\boldsymbol{\theta}_j|\mathcal{P}_j, \mathcal{M}\right)}{\pi\left(\mathbf{y}|\mathcal{P}_j, \mathcal{M}\right)}, \tag{3.26}$$

to update the parameters $\boldsymbol{\theta}_j$ with observational data $\mathbf{y}$. As before, $\pi\left(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{P}_j, \mathcal{M}\right)$ is the likelihood distribution, $\pi\left(\boldsymbol{\theta}_j|\mathcal{P}_j, \mathcal{M}\right)$ characterizes prior information regarding the model parameters $\boldsymbol{\theta}_j$, and the denominator, $\pi\left(\mathbf{y}|\mathcal{P}_j, \mathcal{M}\right)$, is the evidence of the model $\mathcal{P}_j$.

The model evidence, $\pi\left(\mathbf{y}|\mathcal{P}_j, \mathcal{M}\right)$, may be regarded as the likelihood in a higher form of Bayes' Rule,

$$\pi\left(\mathcal{P}_j|\mathbf{y}, \mathcal{M}\right) = \frac{\pi\left(\mathbf{y}|\mathcal{P}_j, \mathcal{M}\right)\pi\left(\mathcal{P}_j|\mathcal{M}\right)}{\pi\left(\mathbf{y}|\mathcal{M}\right)}, \tag{3.27}$$

that defines the *posterior model plausibility* of model $\mathcal{P}_j$ in the set $\mathcal{M}$ for given observational data $\mathbf{y}$. Here, $\pi(\mathcal{P}_j|\mathcal{M})$ is the prior plausibility that model $\mathcal{P}_j$ is true among those in the set $\mathcal{M}$, and the denominator $\pi(\mathbf{y}|\mathcal{M})$ is a normalization constant such that

$$\sum_{i=1}^{m} \pi(\mathcal{P}_j|\mathbf{y}, \mathcal{M}) = 1. \tag{3.28}$$

The posterior model plausibility is often referred to simply as the "plausibility of $\mathcal{P}_j$" and denoted

$$\rho_j = \pi(\mathcal{P}_j|\mathbf{y}, \mathcal{M}). \tag{3.29}$$

The plausibility (3.27) provides an immediate means to determine which model in the set $\mathcal{M}$ best fits the observed data . The model(s) in $\mathcal{M}$ with plausibilities ρ_j closest to unity are deemed the most plausible. In particular, if

$$\rho_j > \rho_k, \tag{3.30}$$

the model $\mathcal{P}_j$ is more plausible than model $\mathcal{P}_k$ for the given data $\mathbf{y}$. There is not necessarily a unique most plausible model within $\mathcal{M}$. The use of model plausibilities as a basis for model selection has been advocated in [10, 92, 93] and used successfully in selecting CG models in [31, 32].

It should be noted that the model plausibility balances model fitness to given data and model complexity. By defining the function

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}_j) = \ln\left(\pi(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{P}_j, \mathcal{M})\, \pi(\boldsymbol{\theta}_j|\mathcal{P}_j, \mathcal{M})\right), \tag{3.31}$$

where the index j has been dropped for notational simplicity, the marginalization

(3.23) may be rewritten as,

$$\pi\left(\mathbf{y}|\mathcal{P},\mathcal{M}\right)=\int\exp\left(l\left(\boldsymbol{\theta}\right)\right)d\boldsymbol{\theta}. \tag{3.32}$$

Following [95], $l\left(\boldsymbol{\theta}\right)$ may be expanded in a Taylor series about its maximum, $\boldsymbol{\theta}^{*}$,

$$l\left(\boldsymbol{\theta}\right)=l\left(\boldsymbol{\theta}^{*}\right)+\sum_{i=1}^{n}L_{i}\left(\boldsymbol{\theta}^{*}\right)\left(\theta_{i}-\theta_{i}^{*}\right)-\sum_{i=1}^{n}\sum_{j=1}^{n}H_{ij}\left(\boldsymbol{\theta}^{*}\right)\left(\theta_{i}-\theta_{i}^{*}\right)\left(\theta_{j}-\theta_{j}^{*}\right)+\ldots, \tag{3.33}$$

where $L_{i}\left(\boldsymbol{\theta}\right)$ is the ith component of the gradient of $l\left(\boldsymbol{\theta}\right)$,

$$L_{i}\left(\boldsymbol{\theta}\right)=\frac{\partial l\left(\boldsymbol{\theta}\right)}{\partial\theta_{i}}, \tag{3.34}$$

and $H_{ij}\left(\boldsymbol{\theta}\right)$ is the (i,j)th component of the negative Hessian, $\mathbf{H}\left(\boldsymbol{\theta}\right)$,

$$H_{ij}\left(\boldsymbol{\theta}\right)=-\frac{\partial^{2}l\left(\boldsymbol{\theta}\right)}{\partial\theta_{i}\partial\theta_{j}}, \tag{3.35}$$

and n is the dimension of $\boldsymbol{\theta}$.

Note that since $\boldsymbol{\theta}^{*}$ maximizes the likelihood, $L_{i}\left(\boldsymbol{\theta}^{*}\right)=0$. Neglecting higher order terms,

$$\pi\left(\mathbf{y}|\mathcal{P},\mathcal{M}\right)\approx\int\exp\left(l\left(\boldsymbol{\theta}^{*}\right)-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}H_{ij}\left(\boldsymbol{\theta}^{*}\right)\left(\theta_{i}-\theta_{i}^{*}\right)\left(\theta_{j}-\theta_{j}^{*}\right)\right)d\boldsymbol{\theta} \tag{3.36}$$

$$=\exp\left(\boldsymbol{\theta}^{*}\right)\int\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}H_{ij}\left(\boldsymbol{\theta}^{*}\right)\left(\theta_{i}-\theta_{i}^{*}\right)\left(\theta_{j}-\theta_{j}^{*}\right)\right)d\boldsymbol{\theta} \tag{3.37}$$

Recalling that $l\left(\boldsymbol{\theta}^{*}\right)=\ln\left(\pi\left(\mathbf{y}|\boldsymbol{\theta}^{*},\mathcal{P},\mathcal{M}\right)\pi\left(\boldsymbol{\theta}^{*}|\mathcal{P},\mathcal{M}\right)\right)$ and approximating the integral by Laplace's method of asymptotic expansion [13, 95],

$$\pi\left(\mathbf{y}|\mathcal{P},\mathcal{M}\right)=\pi\left(\mathbf{y}|\boldsymbol{\theta}^{*},\mathcal{P},\mathcal{M}\right)\pi\left(\boldsymbol{\theta}^{*}|\mathcal{P},\mathcal{M}\right)\frac{\left(2\pi\right)^{n/2}}{\sqrt{det\left|\mathbf{H}\left(\boldsymbol{\theta}^{*}\right|\right)}}. \tag{3.38}$$

The term $\pi\left(\mathbf{y}|\boldsymbol{\theta}^{*},\mathcal{P},\mathcal{M}\right)$ is the likelihood and thus quantifies model fitness by definition. Together, the remaining terms comprise the so-called Occam factor, which

penalizes larger numbers of parameters and models that are highly sensitive to changes in parameters, as discussed in [13, 95]. Should two models fit the observational data equally well, the "simpler" model will have higher plausibility. Thus the principle of Occam's Razor, discussed in detail in Chapter 4, is upheld in the computation of model plausibilities.

Model plausibilities are calculated at the calibration stage of the prediction pyramid, *i.e.* $\mathbf{y} = \mathbf{y}_c$. The model that is deemed the most plausible moves into the validation stage, discussed in the next section, where the model is tested for accuracy with respect to a new set of data, $\mathbf{y}_v$. In general, the validation scenario is far more complicated and computationally expensive than the calibration scenario, making the validation update of *every* model in the set $\mathcal{M}$ impractical; the validation update should be performed as few times as possible. If the calibration scenario is designed to be informative and relevant to the validation scenario and the prediction scenario, as discussed in Sections 3.2 and 4.2, it is reasonable to argue that the most plausible model has the best chance of passing the validation tests.

3.5 Model Validation

After Bayesian calibration is complete and the most plausible model is identified, the model is moved up the prediction pyramid into the validation stage. The validation scenario(s), S_v, correspond to more complex scenarios and are designed to test the model, based on the accuracy with which it is able to reproduce validation observables, which are chosen to reflect the QoI in the prediction scenario, as discussed in Section 3.2.

The calibration posterior from the model judged to be the most plausible, $\pi(\boldsymbol{\theta}|\mathbf{y}_c) = \pi(\boldsymbol{\theta}|\mathbf{y}_c, \mathcal{P}_j, \mathcal{M})$, becomes the prior parameter distribution in the validation scenario. With validation data, $\mathbf{y}_v$, produced in the validation scenario, the validation update is again given by Bayes' Rule,

$$\pi(\boldsymbol{\theta}|\mathbf{y}_v, \mathbf{y}_c) = \frac{\pi(\mathbf{y}_v|\boldsymbol{\theta}, \mathbf{y}_c)\,\pi(\boldsymbol{\theta}|\mathbf{y}_c)}{\pi(\mathbf{y}_v|\mathbf{y}_c)}, \tag{3.39}$$

where $\pi(\mathbf{y}_v|\boldsymbol{\theta}, \mathbf{y}_c)$ is the validation likelihood and is commonly assumed to be Gaussian for the same reasons discussed in Section 3.3.1, and $\pi(\mathbf{y}_v|\mathbf{y}_c)$ is the validation evidence and acts purely as a normalization constant. Note that although $\pi(\mathbf{y}_v|\mathbf{y}_c)$ may theoretically be used to calculate plausibility instead of $\pi(\mathbf{y}_c)$, this is not generally practical, as the validation scenario is far more complex and computationally expensive than the calibration scenario.

It is possible at this point to compute the information gain, which measures how much the parameters change with the addition of the validation data $\mathbf{y}_v$,

$$I(\pi(\boldsymbol{\theta}|\mathbf{y}_v, \mathbf{y}_c), \pi(\boldsymbol{\theta}|\mathbf{y}_c)) = D_{KL}(\pi(\boldsymbol{\theta}|\mathbf{y}_v, \mathbf{y}_c) \| \pi(\boldsymbol{\theta}|\mathbf{y}_c)). \tag{3.40}$$

Here, $D_{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence between two distributions, p and q, also called the relative entropy, defined to be

$$D_{KL}(p\|q) = \int p(\omega) \log \frac{p(\omega)}{q(\omega)} d\omega. \tag{3.41}$$

Clearly,

$$D_{KL}(p\|q) = -H(p) + H(p, q), \tag{3.42}$$

where

$$H(p, q) = -\int p(\omega) \log q(\omega)\, d\omega \tag{3.43}$$

is called the cross entropy. To prevent overfitting, the validation scenarios and experiments should be designed to produce the largest information gain, subject to the usual constraints of cost, complexity, and general feasibility.

The updated parameters $\pi\left(\boldsymbol{\theta}|\mathbf{y}_v, \mathbf{y}_c\right)$ determined by (3.39) are used to solve the forward problem for the validation observable so that the accuracy with which the model is able to predict the observable (which is related to the QoI in the prediction scenario) may be assessed. It is important to note that the parameters $\boldsymbol{\theta}$ are *random variables* distributed according to the validation posterior. Therefore, the observables produced by the model are also random variables. To solve the forward problem, the validation posterior, $\pi\left(\boldsymbol{\theta}|\mathbf{y}_v, \mathbf{y}_c\right)$, is stochastically sampled and a probability density function of the observable, $\pi\left(Q|\boldsymbol{\theta}\right) = \pi\left(Q|\boldsymbol{\theta}, \mathbf{y}_v, \mathbf{y}_c\right)$ is produced.

If the target observable is a probability distribution, $\pi\left(q\right)$, a natural measure of the similarity between the target and predicted distributions is the Kullback-Leibler divergence,

$$\gamma = D_{KL}\left(\pi\left(q\right) \| \pi\left(Q|\boldsymbol{\theta}\right)\right) = \int \pi\left(q\left(\omega\right)\right) \log \frac{\pi\left(q\left(\omega\right)\right)}{\pi\left(Q\left(\omega\right)|\boldsymbol{\theta}\right)} d\omega. \tag{3.44}$$

If, instead, the target observable is a scalar, q, the corresponding prediction may be found by taking the expected value,

$$Q = \mathbb{E}_{\pi_v}\left[\pi\left(Q|\boldsymbol{\theta}\right)\right] = \int_{\Theta} \pi\left(Q|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}|\mathbf{y}_v, \mathbf{y}_c\right) d\boldsymbol{\theta}. \tag{3.45}$$

Then the results may be compared in a Euclidean metric,

$$\gamma = \left|q - Q\right|. \tag{3.46}$$

At this point, an often subjective decision must be made as to whether or not the accuracy of the model, γ, measured by the D_{KL} pseudo-distance defined in (3.44) or the Euclidean metric (3.46), is sufficiently small to declare the model "valid" (or, more accurately, "not invalid," as all models are technically wrong). For a preset tolerance, γ_{tol}, the model is declared valid if

$$\gamma \leq \gamma_{tol}. \tag{3.47}$$

Many other metrics could be chosen. The major point is that the modeler is obliged to choose a metric (or pseudo-metric) and a tolerance to give meaning to the validation process. The choice of an appropriate metric and tolerance may be purely based on the subjective judgment and experience of the modeler, or it may result from more elaborate considerations based, *e.g.* on information theory. If this condition is not satisfied, the model is declared invalid.

One of several ensuing steps may be taken at the modeler's discretion. For example, a new set of possible model classes,

$$\mathcal{M}^* = \{\mathcal{P}_1^*, \mathcal{P}_2^*, \ldots, \mathcal{P}_l^*\}, \tag{3.48}$$

may be defined, the calibration of each performed, and the plausibilities calculated, yielding a new most-plausible model subject to validation tests. Another highly debated option may be to move the model from $\mathcal{M}$ with the next-highest plausibility into the validation stage. Although the calibration scenario should be designed such that the most plausible model has the highest chance of passing the validation test, this is not guaranteed, and it may be that one of the discarded models is valid while the most plausible is not. The present work proposes an algorithm, discussed and

outlined in Chapter 4, through which the next step in the case of model invalidity is defined.

It should be noted that it is possible to define a sequence of validation experiments, with respective scenarios $S_{v,1}, S_{v,2}, \ldots$, each of which is more complex and closer to the prediction scenario than the last. In each scenario, the parameter distributions are updated via Bayes' Rule,

$$\begin{aligned} \pi\left(\boldsymbol{\theta}|\mathbf{y}_{v,k}, \mathbf{y}_{v,k-1}, \ldots, \mathbf{y}_{v,1}, \mathbf{y}_c\right) &= \pi\left(\mathbf{y}_{v,k}|\boldsymbol{\theta}, \mathbf{y}_{v,k-1}, \ldots, \mathbf{y}_{v,1}, \mathbf{y}_c\right) \\ &\quad \times \frac{\pi\left(\boldsymbol{\theta}|\mathbf{y}_{v,k-1}, \ldots, \mathbf{y}_{v,1}, \mathbf{y}_c\right)}{\pi\left(\mathbf{y}_{v,k}|\mathbf{y}_{v,k-1}, \ldots, \mathbf{y}_{v,1}, \mathbf{y}_c\right)} \end{aligned} \tag{3.49}$$

and the forward problem may be solved, yielding the stochastic predicted observable $\pi\left(Q_k|\boldsymbol{\theta}_k\right) = \pi\left(Q_k|\boldsymbol{\theta}, \mathbf{y}_{v,k}, \mathbf{y}_{v,k-1}, \ldots, \mathbf{y}_{v,1}, \mathbf{y}_c\right)$. This may then be compared to the target observable, either of the form $\pi\left(q_k\right)$ or q_k, which may change for each level of validation. As before, if the target observable is characterized by a probability distribution, (3.44) is used to calculate γ_k, whereas if the target observable is a scalar, (3.45) and (3.46) are employed.

As the sequence of validation scenarios is carried out, with $\gamma_k \leq \gamma_{tol}$, the modeler should become more confident in the model's ability to accurately predict the QoI in the prediction scenario. Once a sufficient level of confidence is attained, the "valid" model may be used in the prediction scenario. The stochastic nature of the validation posterior distribution of the parameters should be exploited in the calculation of the QoI. For example, a Monte Carlo method may be used to draw samples of the parameters, each of which is used in an implementation of the forward problem to calculate the QoI in the prediction scenario S_p.

3.6 Implementation

In this work, Bayes' Rule is implemented in the software package QUESO [30], which uses a variety of Monte Carlo sampling techniques. Sample parameter vectors are taken from the prior distribution, and, for each sample, the forward problem is run to obtain $d_i(\boldsymbol{\theta})$. This value is compared to the observational data according to (3.9). If the computed likelihood value is higher than that of the last sample, the parameter is accepted, otherwise it is rejected and a new sample is drawn. This continues until a set of parameter vectors of pre-determined size is obtained.

The main question that arises at this point is: how can one determine the number of samples necessary to obtain the correct posterior distribution? The truth is that this cannot be determined; however, there are steps that may be taken to increase the confidence the modeler has that the posterior produced by QUESO is a good approximation to the true posterior distribution. It bodes well if, when the number of Monte Carlo samples is increased, the posterior does not change appreciably. One challenge is to choose a rigorous measure of closeness of probability densities to judge convergence. Again, the Kullback-Leibler divergence (3.41) is called upon to measure the similarity between the posteriors yielded when different numbers of Monte Carlo samples are used.

The posterior distributions produced by QUESO are clearly discrete, while the Kullback-Leibler divergence (3.41) is defined for continuous probability distribution functions. The method used in this work for calculating the discrete D_{KL}-distance was proposed in [115], in which k-th nearest-neighbor distances are com-

pared between the two distributions being analyzed. Here, the posterior distributions are considered sufficiently "converged" if the discrete D_{KL}-distance, $\bar{D}_{KL}$, is less than a tolerance. In particular, let π_i be the posterior parameter distribution produced by using i Monte Carlo samples, and π_j is that corresponding to j samples, where $j > i$. Then, if the gain in information, measured by $\bar{D}_{KL}$, satisfies a tolerance related to the dimension of the parameter space Θ,

$$\bar{D}_{KL}\left(\pi_{i+1} \| \pi_i\right) \leq \gamma_{post} = 0.5 \dim\left(\Theta\right), \tag{3.50}$$

the posterior, π_i, is considered converged. Past explorations of this topic have determined this to be an appropriate tolerance.

In general, a relatively small number of Monte Carlo samples, *i.e.* 5000 samples, is specified in the initial implementation of Bayes' Rule in QUESO. In the next implementation, this number is increased to, for example, 10000 samples. If

$$\bar{D}_{KL}(\pi_{10000} \| \pi_{5000}) \leq \gamma_{post}, \tag{3.51}$$

the posterior is considered to be converged. Otherwise, the number of Monte Carlo samples is increased incrementally by, for example, 5000 samples, until the condition (3.50) is met.

The converged posterior is used to calculate the plausibility and is used as the prior in the validation update, if applicable. In this work, the evidence is calculated by QUESO during the Bayesian update, from which normalized plausibilities may be calculated.

Chapter 4

The Occam-Plausibility Algorithm

The previous chapter addressed the process by which the reliability of a model may be assessed in the presence of inevitable uncertainties. Incomplete knowledge of the physical world may be compensated for through the use of probability distributions, Bayes' Rule, and information theory. The present chapter addresses a more specific goal: finding the *simplest* valid model. As the thirteenth century Franciscan monk, William of Occam, stated, "Entities must not be multiplied beyond necessity" [17]. The balance of necessity and simplicity has come to be known as Occam's Razor. In this chapter, the framework developed in previous chapters is used to construct a systematic, adaptive paradigm to identify the simplest model that is valid for the purpose of predicting quantities of interest. This process is referred to as the Occam-Plausibility ALgorithm (OPAL). An illustrative flowchart of the algorithm is shown in Figure 4.1, and an example application to constructing a coarse-grained model of polyethylene is detailed in the next chapter.

K. Farrell, J. T. Oden, and D. Faghihi. A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems. *Journal of Computational Physics*, 295:189-208, 2015. K. Farrell developed OPAL and introduced the new interpretation and use of parameter sensitivity analysis as a method of confirming scenario-observable pairs. J. T. Oden supervised the work. D. Faghihi implemented the analysis of parameter sensitivities.

4.1 Step 1: Initialization

As discussed in the previous chapter, it is often the case that the modeler has several choices of possible mathematical models representing the physical system of interest. OPAL begins by identifying a set $\mathcal{M}$ of possible model classes, each with its own set of parameters,

$$\mathcal{M} = \{\mathcal{P}_1(\boldsymbol{\theta}_1), \mathcal{P}_2(\boldsymbol{\theta}_2), \ldots \mathcal{P}_m(\boldsymbol{\theta}_m)\}. \tag{4.1}$$

In general, it is possible (even desirable) that the models within $\mathcal{M}$ are closely related such that two or more models have some of the same parameters. For example, one choice of model may consider a molecular system represented only by harmonic bonds, while another may include both harmonic bonds and angular interactions.

4.2 Step 2: Sensitivity Analysis

It is often the case that one or more model parameters do not affect the specified QoI. Changes in the parameter values do not appreciably change the observable to be measured; that is, the QoI is not *sensitive* to these parameters. The technology associated with quantifying the sensitivity of the model output to changes in model parameters is referred to as *sensitivity analysis*. One common approach to sensitivity analysis is to use scatterplots. In this approach, the parameter space of the model is sampled using, for example, Monte Carlo sampling. For each sampled parameter vector, the desired observable or output is computed. After many samples have been taken, a scatterplot for each parameter, with the parameter value on one axis and the model output on the other, may be produced. While it is often

easy to see which parameters are influential (see, *e.g.* Figure 5.6), for models with many parameters, checking tens or hundreds of scatterplots may be impractical.

Another approach, employed in the present work and developed in the work of Saltelli, *et al.* [101, 102, 104], is to compute variance-based sensitivity indices. In this approach, the variance of the "output," *i.e.* the observable in the calibration or validation scenarios or the QoI in the prediction scenario, is decomposed into the contributions of the variance from each of the parameters. As discussed in Appendix B, by taking Monte Carlo samples of the parameter space, the total sensitivity index, S_{T_i}, for parameter $\boldsymbol{\theta}_i$ may be computed,

$$S_{T_i} = \frac{\mathbb{E}\left(V\left(Y|\boldsymbol{\theta}_{\sim i}\right)\right)}{V(Y)}, \tag{4.2}$$

where $Y = Y(\theta_1, \theta_2, \ldots, \theta_k)$ is the output, $\boldsymbol{\theta}_{\sim i}$ is the vector of all parameters except θ_i, $V(Y)$ is the variance of Y, and $\mathbb{E}\left(V\left(Y|\boldsymbol{\theta}_{\sim i}\right)\right)$ is the expected value with respect to the joint distribution $\pi\left(\boldsymbol{\theta}_{\sim i}\right)$ of the variance that remains in Y when the parameters $\boldsymbol{\theta}_{\sim i}$ are fixed. This quantity measures the total contribution from parameter θ_i to the variance in the output Y. Smaller values of S_{T_i} indicate that θ_i can be fixed at any value within its range of variability without appreciably affecting the output, while higher values of S_{T_i} imply that parameter θ_i is important to the model.

It is asserted in this work that sensitivity analysis should be performed in each of the calibration, validation, and prediction scenarios. Sensitivity analysis not only indicates which parameters the observable or QoI are sensitive to, they indicate which parameters are informed by the observables chosen for the calibration and validation scenarios. Parameters insensitive to observables will not be updated

effectively during calibration or validation. Thus sensitivity analysis may also be used to determine and justify the choice of observational quantity in the calibration and validation scenarios.

Furthermore, if the models in the set $\mathcal{M}$ are closely related such that similar interactions are included, sensitivity analysis may reveal redundant mathematical representations or correlations, and, subsequently, the set of model classes may be reduced. Consider the example briefly discussed in Section 4.1. Should sensitivity analysis show that the QoI and other observables are insensitive to the parameters for the angular interactions, the model including both bonded and angular interactions may be removed from the set of model classes. The set of model classes resulting from sensitivity analysis is denoted,

$$\bar{\mathcal{M}} = \left\{ \bar{\mathcal{P}}_1 \left(\bar{\boldsymbol{\theta}}_1 \right), \bar{\mathcal{P}}_2 \left(\bar{\boldsymbol{\theta}}_2 \right), \ldots \bar{\mathcal{P}}_l \left(\bar{\boldsymbol{\theta}}_l \right) \right\}, \quad l \leq m, \tag{4.3}$$

where it may be the case that no parameters, and, therefore, no models, can be eliminated and $\bar{\mathcal{M}} = \mathcal{M}$.

4.3 Step 3: The Occam Step

The models in the set $\bar{\mathcal{M}}$ resulting from sensitivity analysis may be separated into categories based on their complexity. One may interpret Occam's Razor in an alternative statement: among competing theories that lead to the same prediction, the one that relies on the fewest assumptions is the best. In the context of choosing a single model among a set of models, the simplest valid model is the best choice. The notion of the "simplicity" of a model is largely heuristic and exactly how to

quantify simplicity has been debated for centuries. Because the computational effort in calibrating most models is strongly dependent on the number of parameters in the model (*e.g.* the larger the number of parameters, the more computing time generally needed to generate posteriors), the number of parameters in a model is chosen as a measure of its "simplicity."

With this convention in mind, a partition and re-ordering of the models in the set $\bar{\mathfrak{M}}$ of (4.3) into subsets of C_k^*, $k = 1, 2, \ldots, N^*$, is introduced, where each model in C_k^* has the same number, n_k^*, of parameters, with $n_1^* < n_2^* < \ldots, n_{N^*}^*$. That is, $\bar{\mathfrak{M}}$ is partitioned into subsets ordered according to simplicity, beginning with the set of simplest models, C_1^* and progressing consecutively to models of less simplicity (more "complexity"):

$$\bar{\mathfrak{M}} = \{C_1^*, C_2^*, \ldots, C_{N^*}^*\}\,. \tag{4.4}$$

The labels on these subsets are referred to as *Occam categories*. Thus, models in C_1^* are models of Occam category 1, those in C_2^* are of Occam category 2, *etc.*

Two overriding questions emerge at this point. Firstly, which model(s) within each category are the most plausible, given the calibration data $\mathbf{y}_c$? Secondly, are any of the models valid for predicting the S_v observables? One or more models will be the most plausible within each category, but it may be that none are valid relative to the designated tests and tolerances for model validation.

4.4 Step 4: The Calibration Step

Beginning with $k = 1$, the set

$$\mathcal{M}^* = C_k^* = \{\mathcal{P}_1^*(\boldsymbol{\theta}_1^*), \ldots, \mathcal{P}_{l_k}^*(\boldsymbol{\theta}_{l_k}^*)\} \tag{4.5}$$

is defined, all of the models in $\mathcal{M}^*$ are calibrated via Bayes' Rule (3.5), and the plausibility (3.27) of each model is calculated, following the processes detailed in Sections 3.3 and 3.4. The model plausible model $\mathcal{P}_j^*$ is identified such that

$$\rho_i^* \leq \rho_j^*, \quad i = 1, \ldots, k, \quad i \neq j, \tag{4.6}$$

and is moved into the next step of OPAL, where it will be subjected to validation test(s).

4.5 Step 5: The Validation Step

According to the philosophy and methodology laid down in Chapter 3, calibrated models are tested in a validation scenario or a series of validation scenarios. The parameters $\boldsymbol{\theta}_j^*$ of the most plausible model $\mathcal{P}_j^*$ are updated with Bayes' rule (3.39). As detailed in Section 3.5, the updated parameters are used in a forward problem, and the accuracy with which the model is able to predict the validation observable may be assessed via (3.44) and (3.46), with validity determined by the relation (3.47). If the model $\mathcal{P}_j^*$ is deemed valid, it is suitable for use in the prediction scenario, and the algorithm moves into the prediction stage; otherwise, the model is invalid, and the algorithm proceeds to the Iterative Occam Step. The process is thus terminated when the simplest valid model is identified, while many other valid models may exist in higher Occam categories.

4.6 Step 6: Iteration

4.6.1 The Iterative Occam Step

In the case that model $\mathcal{P}_j^*$ is shown to be invalid, $\mathcal{M}^*$ is redefined so as to contain models of the next category, and Steps 4 and 5 (and possibly Step 6) are repeated. At the modeler's discretion, an additional constraint may be imposed mandating that the models within the new $\mathcal{M}^*$ contain all of the parameters present in the current most plausible model $\mathcal{P}_j^*$. With this specification, model improvements are more akin to hierarchical model inadequacy approaches in which the most plausible model is built upon and improved instead of discarded when invalidated.

4.6.2 Redefining $\mathcal{M}$

Once the highest Occam category is reached, iteration of OPAL entails redefining the set of models chosen in Step 1. This may be done by identifying new possible functional forms to represent the system being studied or by refining the reduced-order representation. For example, when applied to coarse graining, the CG map itself may be redefined to contain fewer atoms per bead and the set of possible model classes, $\mathcal{M}$, may then be redefined. After the new set $\mathcal{M}$ has been identified, Step 2 through Step 5 and, if applicable, Step 6, are repeated.

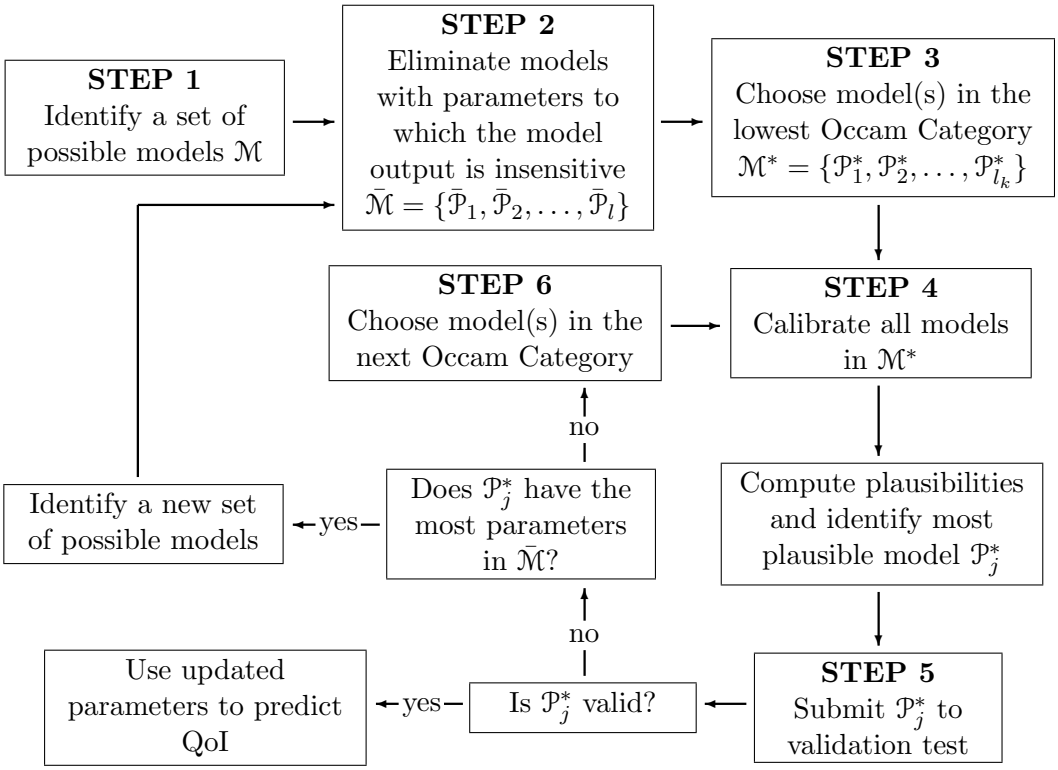

Figure 4.1: Flowchart illustrating the Occam-Plausibility Algorithm. Beginning with the identification of a set of possible model classes $\mathcal{M}$, iteration through this algorithm yields the simplest model that passes a Bayesian validation test. A sensitivity analysis study may reveal that one or more models may be eliminated as possible representations of the system being studied. The remaining models are divided into categories based on the number of parameters they contain, and the calibration and plausibility calculations begin in the category with the simplest models. Should the most plausible model be rendered invalid, the algorithm moves into the next category of models until the simplest valid model is identified.

Chapter 5

An Example Application

This chapter discusses the construction of a coarse-grained model of polyethylene using the Bayesian framework of model calibration, selection, and validation, and the Occam-Plausibility Algorithm. Consider, as a prediction scenario, a cube of polyethylene, containing 25 chains of $C_{80}H_{162}$ and constrained by periodic boundary conditions. The cube is simulated in a canonical ensemble with temperature $T = 300K$, and the QoI is taken to be the distribution of the potential energy. The all-atom system of S_p is shown in Figure 5.1.

5.1 OPAL Step 1: Initialization

5.1.1 The Coarse-Grained Map

In Chapter 2, the choices and uncertainties that arise from defining the coarse-grained (CG) mapping of an atomistic system were discussed. In particular, the number of beads or, equivalently, the number of atoms per bead, must be defined.

K. Farrell, J. T. Oden, and D. Faghihi. A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems. *Journal of Computational Physics*, 295:189-208, 2015. K. Farrell developed OPAL, introduced the new interpretation and use of parameter sensitivity analysis as a method of confirming scenario-observable pairs, and implemented Bayesian methods computationally. J. T. Oden supervised the work. D. Faghihi implemented the analysis of parameter sensitivities.

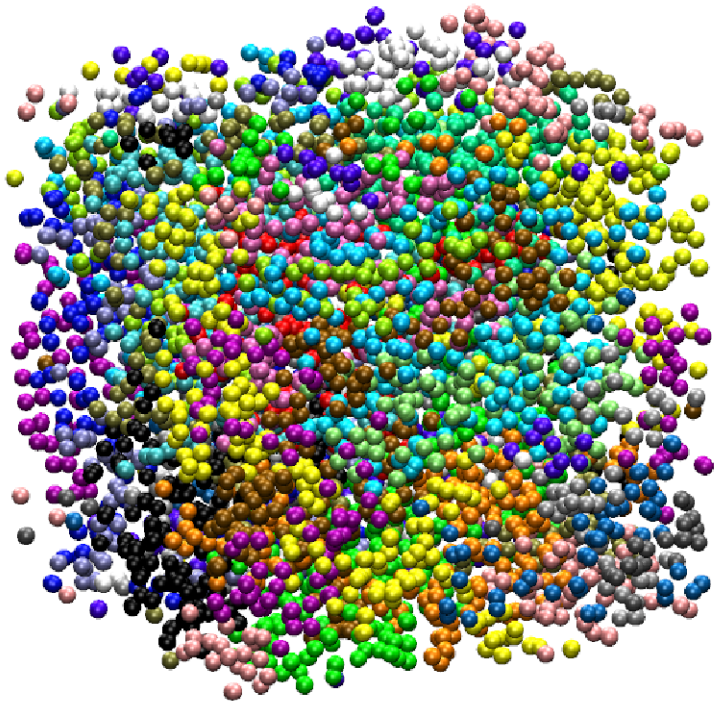

Figure 5.1: Prediction scenario of polyethylene, a cube containing 25 chains of $C_{80}H_{162}$ and constrained by periodic boundary conditions, with each color corresponding to a different chain within the cube.

When considering polyethylene, if each CG bead is defined to contain a single carbon atom and accompanying hydrogen atoms, the CG system is called the "united atom" representation and is a very common choice of mapping. One may also define the CG map, G, such that CG particles contain two or three carbon atoms and their bonded hydrogen atoms. The carbon atoms in polyethylene are serially bonded, as shown in Figure 5.2. Thus, in coarser mappings for which four or more carbon atoms are assigned to each bead, CG beads are ellipsoids rather than spheres, and additional degrees of freedom are introduced to account for the asymmetry of the

particles. Illustrations of possible CG mappings for polyethylene are given in Figure 5.3. In the present work, each CG bead is defined to represent two carbon atoms and their accompanying hydrogen atoms. The map G is defined to map the coordinates of the atoms in each bead to the center of mass of the aggregation. Thus $\mathbf{R}_I$ is defined by (2.30).

5.1.2 The Model Classes

Once the CG map, G, has been specified, the set $\mathcal{M}$ of possible model classes is the set of possible representations of the CG potential V_{CG} (see Section 2.2). In the present work, it is assumed that V_{CG} takes on the same functional form as the AA system defined in (2.4)-(2.8). Note that with the current map G, each CG bead is charge-neutral, so V_{CG} has no Coulomb term. In addition, the Lennard-Jones 9-6 potential,

$$V_{nb}\left(\mathbf{R}^N\right) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{R_{ij}} \right)^9 - \left(\frac{\sigma_{ij}}{R_{ij}} \right)^6 \right] f_{ij}, \tag{5.1}$$

may be considered as an alternative to the Lennard-Jones 12-6 potential in (5.6). The Lennard-Jones 9-6 potential is a softer potential to account for the empty space that results from grouping atoms into a single spherical CG bead.

The QoI in this example will be the potential energy, so an additive constant parameter, A, should be added to V_{CG} such that

$$V(\mathbf{R}^N) = V_{bond}(\mathbf{R}^N) + V_{angle}(\mathbf{R}^N) + V_{dihedral}(\mathbf{R}^N) + V_{nb}(\mathbf{R}^N) + A, \tag{5.2}$$

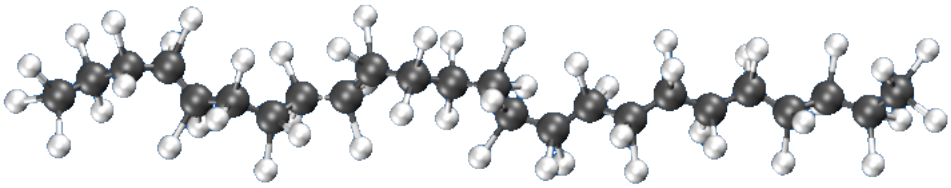

Figure 5.2: Short segment of an all-atom representation of polyethylene. Carbon atoms are shown in grey and are serially bonded, while hydrogen atoms are shown in white.

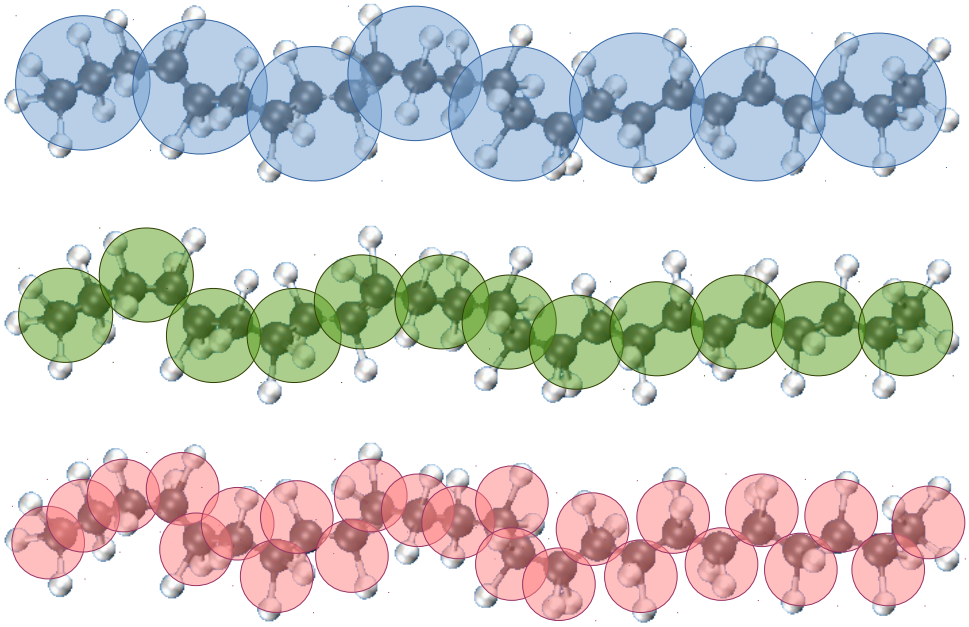

Figure 5.3: Examples of possible coarse-grained mappings of polyethylene. Reasonable mappings may contain as many as three serially bonded carbon atoms and their accompanying hydrogen atoms (top). In this example, G maps two carbon atoms and their hydrogen atoms into a single CG bead, which is shown in the middle figure. The united atom mapping (bottom) is also commonly used.

where

$$V_{bond}\left(\mathbf{R}^N\right) = \sum_{i=1}^{N_b} K_{R,i}\left(R_i - R_{0,i}\right), \tag{5.3}$$

$$V_{angle}\left(\mathbf{R}^N\right) = \sum_{i=1}^{N_a} K_{\theta,i}\left(\theta_i - \theta_{0,i}\right), \tag{5.4}$$

$$\begin{aligned} V_{dihedral}\left(\mathbf{R}^N\right) = & \sum_{i=1}^{N_d} \frac{V_{1,i}}{2}\left(1+\cos\left(\varphi_i\right)\right) + \frac{V_{2,i}}{2}\left(1-\cos\left(2\varphi_i\right)\right) \\ & + \frac{V_{3,i}}{2}\left(1+\cos\left(3\varphi_i\right)\right) + \frac{V_{4,i}}{2}\left(1-\cos\left(4\varphi_i\right)\right), \end{aligned} \tag{5.5}$$

and $V_{nb}\left(\mathbf{R}^N\right)$ is a Lennard-Jones 12-6 potential,

$$V_{nb}\left(\mathbf{R}^N\right) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right] f_{ij}, \tag{5.6}$$

or the Lennard-Jones 9-6 given in Equation (5.1). The number of bonds, angles, and dihedral interactions in the CG system, are respectively, N_b, N_a and N_d. See Appendix A for details. Within a single CG bead, the minimum energy configuration of the contained atoms yields a nonzero potential energy. Thus the parameter A represents the potential energy lost due to the coarse graining process and can be thought of as the internal bead energy.

The set $\mathcal{M}$ is created by tabulating possible combinations of types of interactions. It therefore has 23 possible model classes, each with its own parameter vector that depends on the interactions included, and all of which include the parameter A. With the current map G, all of the coarse-grained particles are of the same type. Therefore, there is only one type of bonded interaction, yielding two parameters, K_R and R_0; one type of angular interaction, also yielding two parameters, K_θ and θ_0; one type of dihedral interaction, which requires four parameters, V_1, V_2, V_3, and

V_4; and one type of non-bonded interaction with parameters ϵ and σ. Table 5.1 contains a complete list of the possible models.

5.2 OPAL Step 2: Sensitivity Analysis

It has been discussed previously that sensitivity analysis may be used to eliminate parameters to which the target quantity of interest is insensitive. In Section 4.2, it was argued that sensitivity analysis may also be used to support the choice of $(S_c, \mathbf{y}_c)$ and $(S_{vi}, \mathbf{y}_{vi})$ pairs. The calibration and validation scenarios are considered acceptable for use in the Bayesian framework if the sensitivity analysis results in the calibration and validation scenarios are similar to those in the prediction scenario, since this indicates which parameters will be informed by the chosen observable data. The observable Y in the prediction scenario is necessarily the QoI, while in the calibration and validation scenarios, Y is defined to be the same function that is used to calculate $d_i(\boldsymbol{\theta})$ in the likelihood (3.9). However, in general, Y is taken to be an average value, *e.g.* the ensemble average in molecular systems, while $d_i(\boldsymbol{\theta})$ are individual samples of the function value.

Recall from Section 3.2 that the validation and calibration scenarios are subsets of the prediction scenario, designed to capture as many of its characteristics and behavior as possible, and that the data collected in these scenarios are chosen to inform the parameters about the quantity of interest. As the prediction scenario is a cube of polyethylene, containing 25 chains of length 80 carbon atoms each, two validation scenarios are identified: a smaller cube, consisting of four polyethylene chains of $C_{80}H_{162}$, which is denoted S_{v2}, and a system of two polyethylene chains

Model	Bonds	Angles	Dihedrals	LJ 12-6	LJ 9-6	Params	Cat.
$\mathcal{P}_1$	✓					3	1
$\mathcal{P}_2$		✓				3	
$\mathcal{P}_3$				✓		3	
$\mathcal{P}_4$					✓	3	
$\mathcal{P}_5$	✓	✓				5	2
$\mathcal{P}_6$	✓			✓		5	
$\mathcal{P}_7$	✓				✓	5	
$\mathcal{P}_8$		✓		✓		5	
$\mathcal{P}_9$		✓			✓	5	
$\mathcal{P}_{10}$			✓			5	
$\mathcal{P}_{11}$	✓	✓		✓		7	3
$\mathcal{P}_{12}$	✓	✓			✓	7	
$\mathcal{P}_{13}$	✓		✓			7	
$\mathcal{P}_{14}$		✓	✓			7	
$\mathcal{P}_{15}$			✓	✓		7	
$\mathcal{P}_{16}$			✓		✓	7	
$\mathcal{P}_{17}$	✓	✓	✓			9	4
$\mathcal{P}_{18}$	✓		✓	✓		9	
$\mathcal{P}_{19}$	✓		✓		✓	9	
$\mathcal{P}_{20}$		✓	✓	✓		9	
$\mathcal{P}_{21}$		✓	✓		✓	9	
$\mathcal{P}_{22}$	✓	✓	✓	✓		11	5
$\mathcal{P}_{23}$	✓	✓	✓		✓	11	

Table 5.1: Table of possible CG models created by all combinations of interaction potentials that may be included in the representation of the potential energy. The listed Occam categories are determined by the number of parameters in the model.

of $C_{80}H_{162}$, denoted S_{v1}. Decomposing the system further yields a single chain of $C_{80}H_{162}$ as the calibration scenario. These scenarios are shown in Figure 5.4. In this example, the QoI and, hence, the output Y for the sensitivity analysis are taken to be the potential energy of the atomistic-molecular system defined by S_p; therefore, in the calibration and validation scenarios, the observables to be used in the Bayesian updates and the output to be used in the sensitivity analysis are chosen to be the potential energy as well.

To perform the sensitivity analysis, the CG systems in each scenario are simulated using (5.2). The parameter vector, $\boldsymbol{\theta} = \{K_R, R_0, K_\theta, \theta_0, V_1, V_2, V_3, V_4, \epsilon, \sigma\}$ captures all of the parameters from the models in the set $\mathcal{M}$, described above and tabulated in Table 5.1. A random sample of this parameter vector is drawn from the ranges,

$$\begin{array}{ll} r_0 \sim \mathcal{U}(0.1, 5); & K_r \sim \mathcal{U}(0.1, 90); \\ \theta_0 \sim \mathcal{U}(80, 180); & K_\theta \sim \mathcal{U}(0, 15); \\ V_1 \sim \mathcal{U}(-1, 1); & V_2 \sim \mathcal{U}(-1, 1); \\ V_2 \sim \mathcal{U}(-1, 1); & V_4 \sim \mathcal{U}(-1, 1); \\ \sigma \sim \mathcal{U}(0.05, 7); & \epsilon \sim \mathcal{U}(0.01, 5). \end{array} \tag{5.7}$$

Then the CG system in the scenario currently being studied (S_c, S_{vi}, or S_p) is simulated, and the desired observable is calculated, following details given in Section 2.3. The total sensitivity indices for each of the ten parameters in each of the defined scenarios are computed according to (4.2), using the methods described in Appendix B.

In the prediction scenario, 20000 samples of the parameter space (5.7) were drawn. For each sample, $\boldsymbol{\theta}^j$, the CG system of 25 chains of length 80 carbon atoms contained within the cube defining S_p is simulated according to the details con-

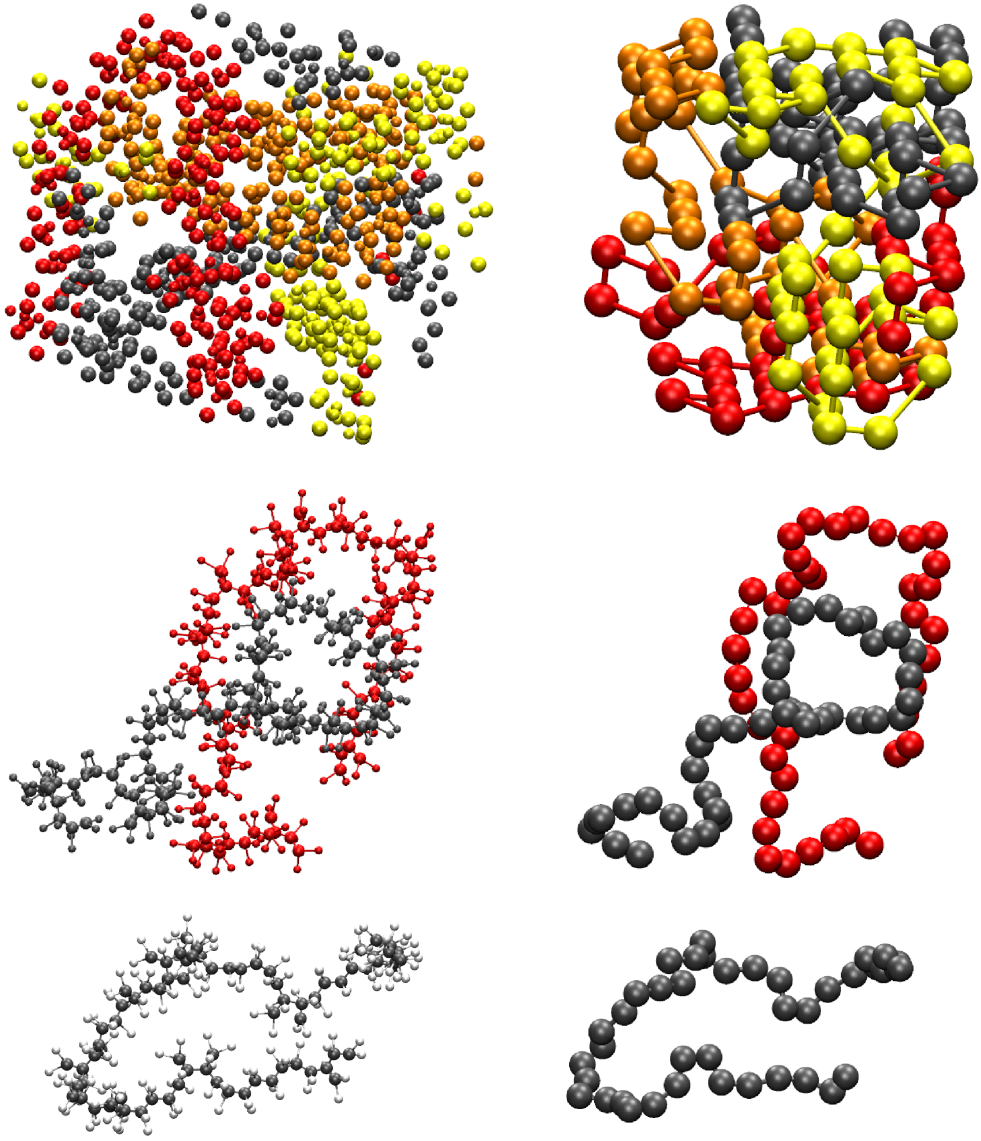

Figure 5.4: Illustrations of scenarios to be used in the calibration and validation of polyethylene, both in the AA (left) and CG (right) systems. A subsystem of the prediction scenario is a small cube (top), contains four chains of $C_{80}H_{162}$, each colored differently, and comprises S_{v2}. Further decomposition of the system yields two chains, creating S_{v1} (middle), and the calibration scenario, consisting of a single chain of $C_{80}H_{162}$, is shown in the bottom row.

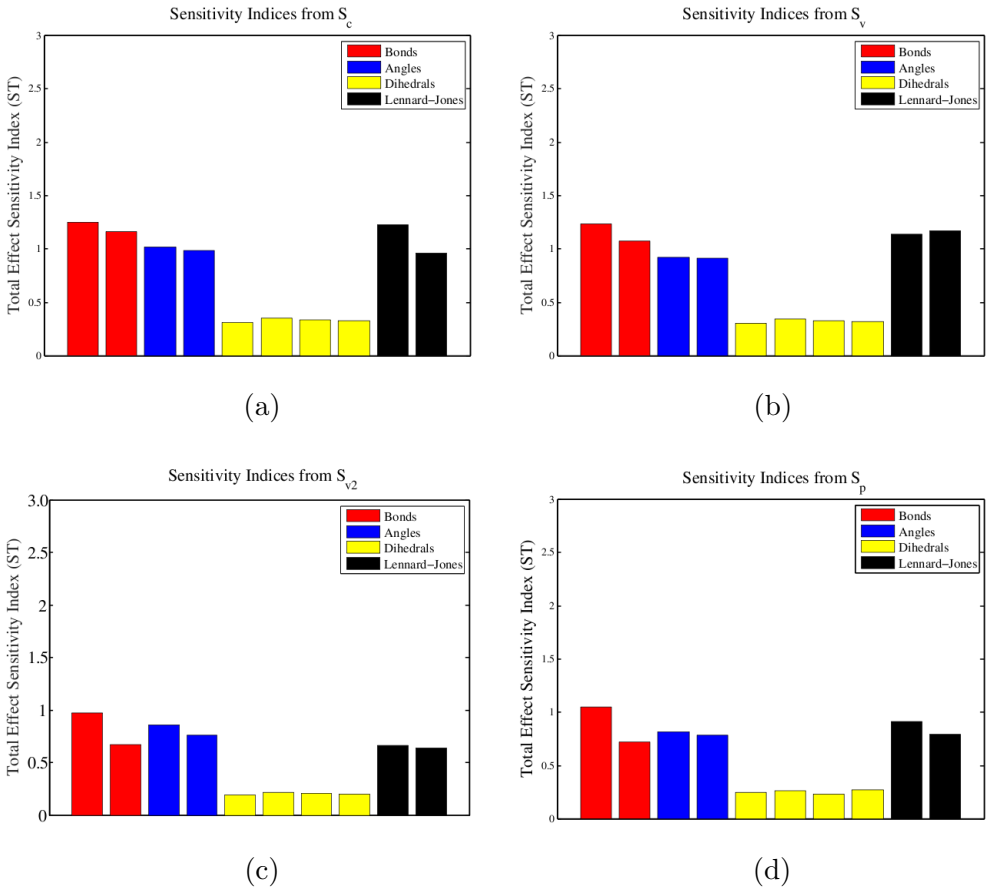

Figure 5.5: Sensitivity indices for each of the scenarios considered in this example. The calibration scenario (a), which contains a single chain of $C_{80}H_{162}$, the first validation scenario (b), consisting of two chains of $C_{80}H_{162}$, the second validation scenario (c), in which four chains of $C_{80}H_{162}$ are considered, and the prediction scenario (d), consisting of 25 chains of $C_{80}H_{162}$ restrained to a cube with periodic boundary conditions. All sensitivity results illustrate that dihedral interactions are not influential on the value of the output function, indicating not only that the scenario-observable pairs are appropriate, but that the parameters will be informed by the observables.

tained in Section 2.3. The ensemble average of the potential energy in the canonical ensemble,

$$Y_j = \left\langle V_{CG}\left(\boldsymbol{\theta}^j\right)\right\rangle_{S_p}, \tag{5.8}$$

is calculated by (2.23). The sensitivity indices for the prediction scenario are shown in Figure 5.5. It can be seen that the dihedral parameters, V_1, V_2, V_3, and V_4 do not strongly influence the potential energy.

In a similar way, for each sample $\boldsymbol{\theta}^j$ of the parameter space (5.7), the validation scenario, S_v, and the calibration scenario, S_c, are simulated in a canonical ensemble, again according to the steps outlined in Section 2.3. The ensemble average of the potential energies,

$$Y_j = \left\langle V_{CG}\left(\boldsymbol{\theta}^j\right)\right\rangle_{S_v}, \quad Y_j = \left\langle V_{CG}\left(\boldsymbol{\theta}^j\right)\right\rangle_{S_c} \tag{5.9}$$

in the validation and calibration scenarios, respectively, are calculated. The sensitivity indices for the validation scenarios are shown in Figure 5.5, and those for the calibration scenario are given in Figure 5.5. In each case, as in the prediction scenario, it can be seen that the potential energy is not sensitive to the dihedral parameters, V_1, V_2, V_3, and V_4.

Comparing the sensitivity indices from each of the scenarios in question, it can be seen that the dihedral parameters do not strongly influence the potential energy. Therefore, the models $\mathcal{P}_i$ in the set $\mathcal{M}$ that contain dihedral parameters may be eliminated from the set of possible model classes. The resulting set, $\bar{\mathcal{M}}$, is given in Table 5.2. Furthermore, it can be argued that the chosen observable, *i.e.* the potential energy, will inform the remaining parameters, R_0, K_R, θ_0, K_θ, σ,

and ϵ in the Bayesian updates during calibration and validation. The qualitative agreement between the results yielded in the three scenarios supports the choice of scenario-observable pairs.

Note that for each sampled parameter vector $\boldsymbol{\theta}^j$, an ordered pair, $\left(\theta_i^j, Y_j\right)$, may be defined for each parameter, θ_i, in the parameter vector. After many sample parameter vectors have been selected and the corresponding output is calculated, a scatterplot for each parameter is produced. Figures 5.6 and 5.7 show the scatterplots produced in the calibration scenario. A strong correlation between the output and the bonded parameters, for example, can be seen, while it is clear the the dihedral parameters do not strongly influence the value of the potential energy. These scatterplots, therefore, agree with the sensitivity indices that the dihedral parameters may be eliminated as parameters.

Model	**Bonds**	**Angles**	**Dihedrals**	**LJ 12-6**	**LJ 9-6**	**Params**	**Cat.**
$\bar{\mathcal{P}}_1$	✓					3	1
$\bar{\mathcal{P}}_2$		✓				3	
$\bar{\mathcal{P}}_3$				✓		3	
$\bar{\mathcal{P}}_4$					✓	3	
$\bar{\mathcal{P}}_5$	✓	✓				5	2
$\bar{\mathcal{P}}_6$	✓			✓		5	
$\bar{\mathcal{P}}_7$	✓				✓	5	
$\bar{\mathcal{P}}_8$		✓		✓		5	
$\bar{\mathcal{P}}_9$		✓			✓	5	
$\bar{\mathcal{P}}_{10}$	✓	✓		✓		7	3
$\bar{\mathcal{P}}_{11}$	✓	✓			✓	7	

Table 5.2: Table of models that remain after the sensitivity analysis is performed. Bonded, angular, and Lennard-Jones interactions were shown to affect the potential energy in each of the calibration, validation, and prediction scenarios.

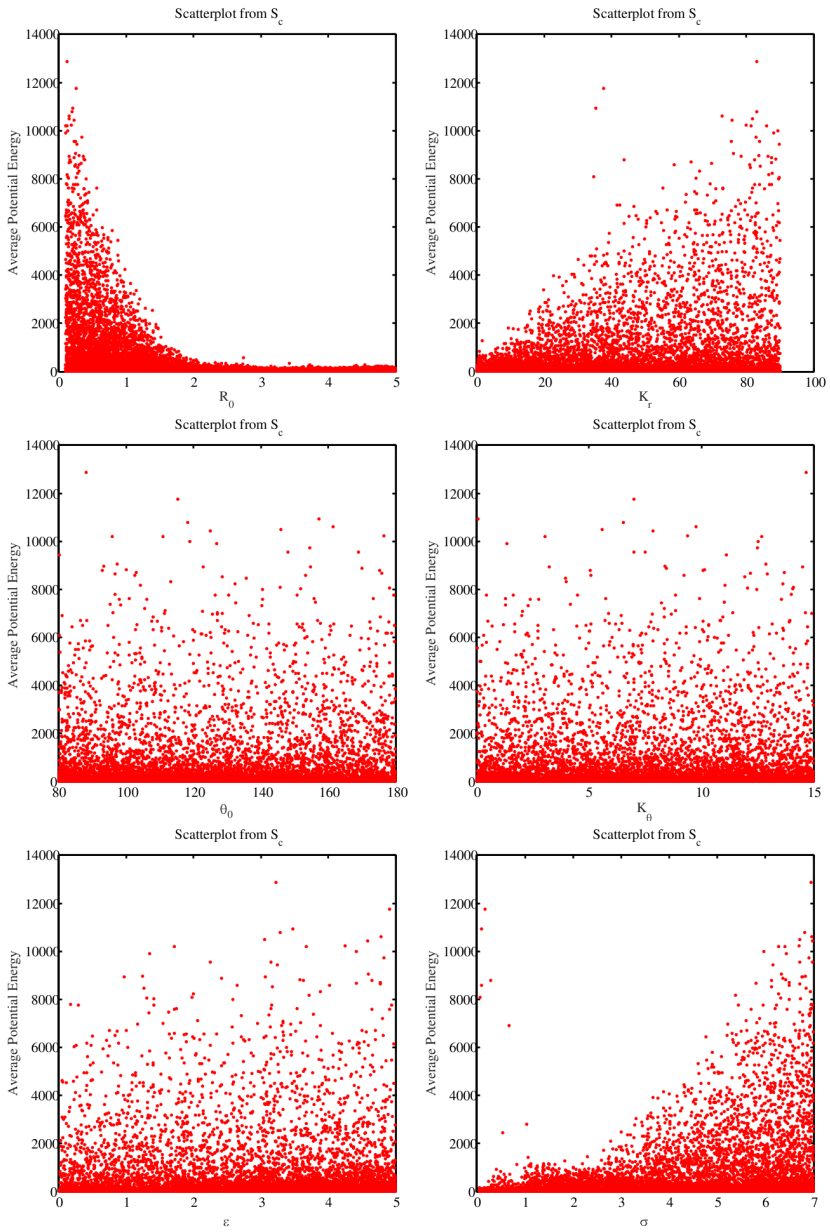

Figure 5.6: Scatterplots produced in the calibration scenario for the bonded, angular, and Lennard-Jones parameters. Each point in the scatterplot corresponds to a sample of the parameter space and the resulting calculation of the output function.

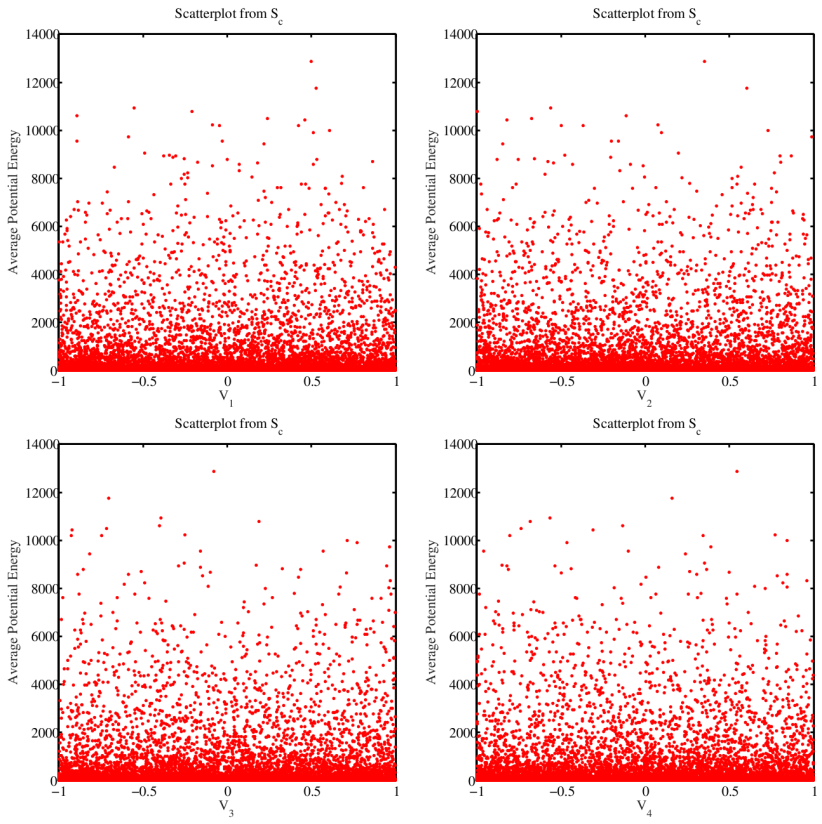

Figure 5.7: Scatterplots produced in the calibration scenario for the dihedral parameters. Each point in the scatterplot corresponds to a sample of the parameter space and the resulting calculation of the output function.

To support the claim that dihedral parameters are not influential to the potential energy, a polyethylene chain is simulated for three cases: with all ten parameters, with dihedrals excluded, and with angular interactions excluded. The results are shown in Figure 5.8. The graphs in the two former cases are nearly identical while the graph produced by the latter case is markedly different, further justifying the exclusion of dihedral interactions.

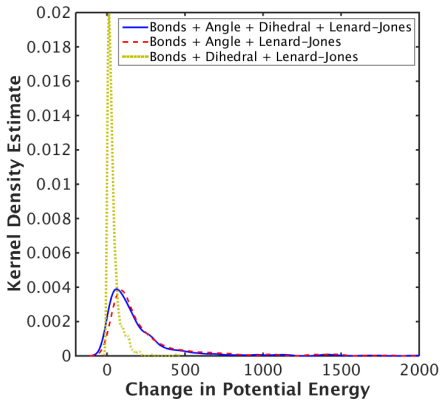

Figure 5.8: Kernel density estimates produced by a deformation of a polyethylene chain with three different combinations of intramolecular interactions. In one case, all four types of interactions are included: bonded, angular, dihedral, and non-bonded. In the case where dihedral interactions are excluded, almost no change in the kernel density estimate can be seen, while in the case where angular interactions are excluded, a drastic difference in kernel density estimates is clear, confirming that dihedral interactions do not strongly influence the potential energy of polyethylene.

5.3 OPAL Step 3: The Occam Step

The model classes in the set $\bar{\mathcal{M}}$ shown in Table 5.2 may be separated into categories according to the number of parameters in each model. Category 1 models each depend on three parameters, and the set of these may be written,

$$C_1^* = \left\{ \mathcal{P}_1^* = \bar{\mathcal{P}}_1, \mathcal{P}_2^* = \bar{\mathcal{P}}_2, \mathcal{P}_3^* = \bar{\mathcal{P}}_3, \mathcal{P}_4^* = \bar{\mathcal{P}}_4 \right\}. \tag{5.10}$$

Model classes in Category 2 each contain five parameters, yielding,

$$C_2^* = \left\{ \mathcal{P}_1^* = \bar{\mathcal{P}}_5, \mathcal{P}_2^* = \bar{\mathcal{P}}_6, \mathcal{P}_3^* = \bar{\mathcal{P}}_7, \mathcal{P}_4^* = \bar{\mathcal{P}}_8, \mathcal{P}_5^* = \bar{\mathcal{P}}_9 \right\}, \tag{5.11}$$

and Category 3 model classes have seven parameters each, so that

$$C_3^* = \left\{ \mathcal{P}_1^* = \bar{\mathcal{P}}_{10}, \mathcal{P}_2^* = \bar{\mathcal{P}}_{11} \right\}. \tag{5.12}$$

The steps of Bayesian model calibration and validation begin with Category 1.

5.4 OPAL Step 4: Calibration

In this step of OPAL, the models in a single Occam Category C_k^* are calibrated according to the process outlined in Section 3.3. The Bayesian update requires the definition of two probability distributions: the likelihood and the prior. The product of these two not only results in the posterior distribution to be used, if applicable, in the validation scenario, but is also used to calculate the evidence, which is then used to calculate posterior model plausibilities.

5.4.1 The Likelihood Distribution

It is assumed here that the likelihood is a normal distribution, centered at zero with unknown variance σ^2,

$$\pi\left(y_i|\boldsymbol{\theta}\right) = p\left(y_i - d_i\left(\boldsymbol{\theta}\right)\right) \sim \mathcal{N}\left(0, \sigma^2\right), \tag{5.13}$$

where y_i are the observed data from the AA system in the calibration scenario, and $d_i(\boldsymbol{\theta})$ are the data observed in the CG system as it is being simulated with parameters $\boldsymbol{\theta}$. As the variance is unknown, σ^2 is added to the list of calibration parameters. The total likelihood is given by (3.10).

In general, calibration data is collected from experiments or high-fidelity virtual experiments simulated in the calibration scenario. In this example, a molecule of $C_{80}H_{162}$ is simulated in a canonical ensemble. Every 100 timesteps, the atomistic configuration and the potential energy are recorded. Note that there should be a number of timesteps between samples to ensure that they are independent. Thus

the calibration data is the vector of potential energies at sample configurations ω_i,

$$\mathbf{y}_c = \{y_i\} = \{V(\omega_i)\}, \tag{5.14}$$

where $V(\omega_i)$ is calculated by (2.4).

The corresponding CG data is determined, for parameter vector $\boldsymbol{\theta}$, by placing the CG molecule into configuration ω_i and calculating the potential energy. That is,

$$d_i(\boldsymbol{\theta}) = V_{CG}(G(\omega_i); \boldsymbol{\theta}), \tag{5.15}$$

where G is the coarse-grained map. The potential energy V_{CG} is calculated by (5.2), where only the interactions included in the model are computed.

5.4.2 The Prior Distributions

The methodology of deriving maximum entropy prior distributions was described in Section 3.3.2. In the present example, parameters are derived independently of one another, taking the form (3.21) in the case that only the mean of the parameter, $\langle\theta\rangle$, is known, and (3.22) in the case that the variance of the parameter, σ_θ^2 is additionally known. In the case of molecular systems, the mean and, occasionally, the variance of the parameters may be gleaned from the atomistic system by considering a simplified scenario, for example, the calibration scenario.

The AA system is simulated in a canonical ensemble and samples of the configuration, ω_i, are recorded periodically. For each ω_i, the distance between two bonded CG sites, j and k, having coordinates $\mathbf{R}_j$ and $\mathbf{R}_k$, respectively, can be

measured,

$$R(\omega_i) = \|\mathbf{R}_j - \mathbf{R}_k\|. \tag{5.16}$$

The average of these distances, taken over all bonds and all sampled configurations, yields an average bond length,

$$\langle R_0 \rangle = \frac{1}{n \cdot N_b} \sum_{i=1}^{n} \sum_{l=1}^{N_b} R_l(\omega_i), \tag{5.17}$$

n being the number of configuration samples and N_b being the total number of bonds in the system. Furthermore, the variances in the observed bond lengths can be computed,

$$\sigma^2_{R_0} = \frac{1}{n \cdot N_b} \sum_{i=1}^{n} \sum_{l=1}^{N_b} \left(R_l(\omega_i) - \langle R_0 \rangle\right)^2. \tag{5.18}$$

The mean, $\langle R_0 \rangle$, and variance, $\sigma^2_{R_0}$, may be used as prior information regarding the parameter R_0, making the prior distribution for R_0 a Gaussian distribution.

Similarly, for each ω_i, the angle between any three CG sites can be measured. The average angle value can therefore be taken over all configurations and all angles in the system,

$$\langle \theta_0 \rangle = \frac{1}{n \cdot N_a} \sum_{i=1}^{n} \sum_{l=1}^{N_a} \theta_l(\omega_i), \tag{5.19}$$

where N_a is the total number of angles in the system and $\theta_l(\omega_i)$ is the measure of the l-th angle of configuration ω_i. The variance in the observed angle measures may then be calculated,

$$\sigma^2_{\theta_0} = \frac{1}{n \cdot N_a} \sum_{i=1}^{n} \sum_{l=1}^{N_a} \left(\theta_l(\omega_i) - \langle \theta \rangle\right)^2. \tag{5.20}$$

With the mean and variance known, the maximum entropy prior distribution for θ_0 is a Gaussian distribution.

The Equipartition Theorem states that each independent quadratic term in the energy has a mean value of $k_BT/2$, where k_B is Boltzmann's constant and T is the temperature [116]. In the present example, both the bonded and angular potential energies are quadratic, leading to, as mentioned in [118],

$$\left\langle K_R \left(R - R_0\right)^2 \right\rangle = \frac{k_B T}{2} \tag{5.21}$$

and

$$\left\langle K_\theta \left(\theta - \theta_0\right)^2 \right\rangle = \frac{k_B T}{2}. \tag{5.22}$$

Due to the assumed independence of K_R and R_0,

$$\langle K_R \rangle = \frac{k_B T}{2\left\langle \left(R - R_0\right)^2 \right\rangle} = \frac{k_B T}{2\sigma_{R_0}^2} \tag{5.23}$$

and, similarly,

$$\langle K_\theta \rangle = \frac{k_B T}{2\left\langle \left(\theta - \theta_0\right)^2 \right\rangle} = \frac{k_B T}{2\sigma_{\theta_0}^2}. \tag{5.24}$$

Thus, the mean value of the spring coefficients K_R and K_θ are inversely related to the variance in the equilibrium values R_0 and θ_0, respectively. Since no further information can be extracted, the prior distributions for K_R and K_θ are exponential distributions.

To derive the Lennard-Jones parameters, the distance (5.16) is computed for all pairs of like-particles in the CG system, which, in this example, is all of the particles. With n sampled configurations and N_{nb} total non-bonded interactions, this creates a set of $n \cdot N_{nb}$ distances which are sorted into bins, as in a histogram. This information can be used to approximate the 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 [66]. The algorithm used to approximate $g(R)$ in this work can be found in [36].

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

$$U(R) = -k_B T \ln g(R), \tag{5.25}$$

which characterizes how the energy of the system changes as a function of the distance between two particles [60, 66]. Since the radial distribution function has a single maximum [66], 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 Lennard-Jones 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 Lennard-Jones well-depth parameter ϵ, to be used in a prior given by (3.21).

The average distance between CG beads is measured to be 2.5781 Angstroms, with a variance of 6.4042×10^{-3} and the average angle measure between groups of three CG sites is 135.7104 degrees with variance 524.6647. A list of distances between any two pairs of CG beads may be compiled and used to approximate the potential of mean force (5.25). The minimum occurs at 2.5007 Angstroms, has a depth of 1.3871 units, and the approximate width of this well yields a variance of 4.505×10^{-3}. The resulting prior distributions are shown in Figure 5.9.

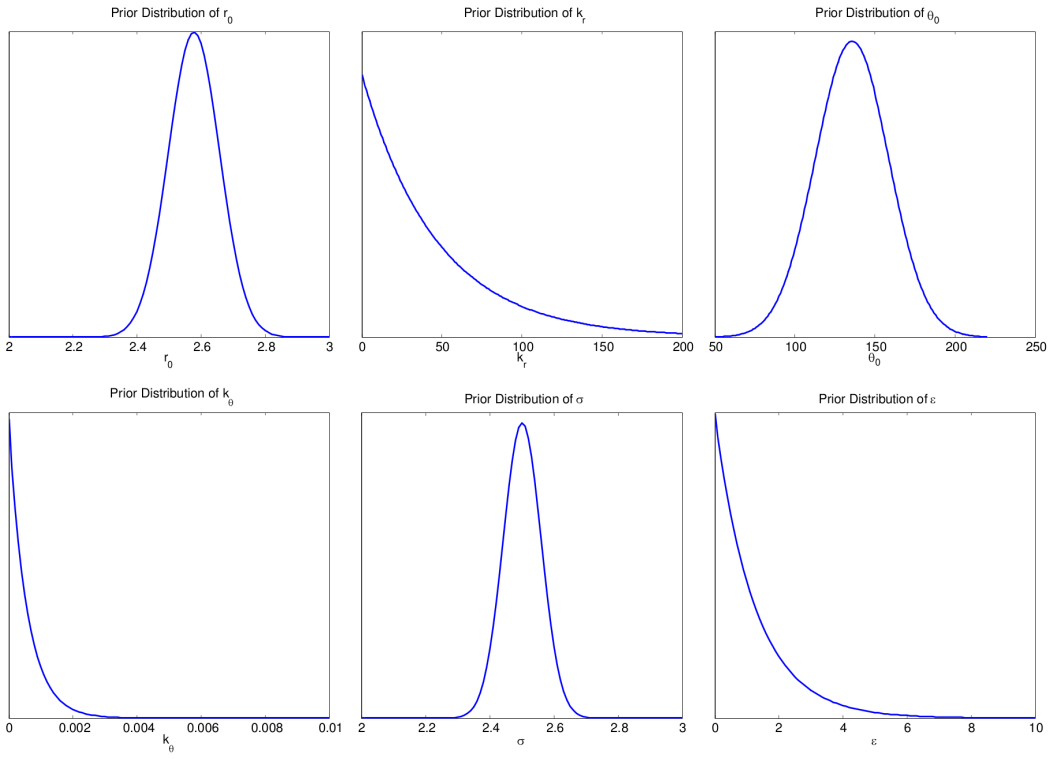

Figure 5.9: Maximum entropy prior distributions for polyethylene, yielded by simulating a simplified scenario and deriving means and variances using methodologies described in this section. These priors capture all known prior information and will be updated in the calibration scenario via Bayes' Rule.

5.4.3 C_1^*: Category 1

In the first iteration of OPAL, only the models in C_1^* are calibrated using QUESO, as described in Section 3.6. For each model, an initial amount of 5000 Monte Carlo samples is used, and then a second QUESO implementation is run with 10000 samples. The discrete Kullback-Leibler distance between these two posterior distributions, $\bar{D}_{KL}\left(\pi_{10000} \| \pi_{5000}\right)$, is calculated according to the method described in [115]. If this value is less than the preset tolerance (3.50) the posterior is considered

Model	Bonds	Angles	LJ 12-6	LJ 9-6	Params	Plausibility
$\mathcal{P}_1^*$	✓				3	1
$\mathcal{P}_2^*$		✓			3	0
$\mathcal{P}_3^*$			✓		3	0
$\mathcal{P}_4^*$				✓	3	0

Table 5.3: Table of plausibilities calculated for the models in the lowest Occam category, produced during model calibration against data from a single chain of $C_{80}H_{162}$. Values shown as zero are approximate.

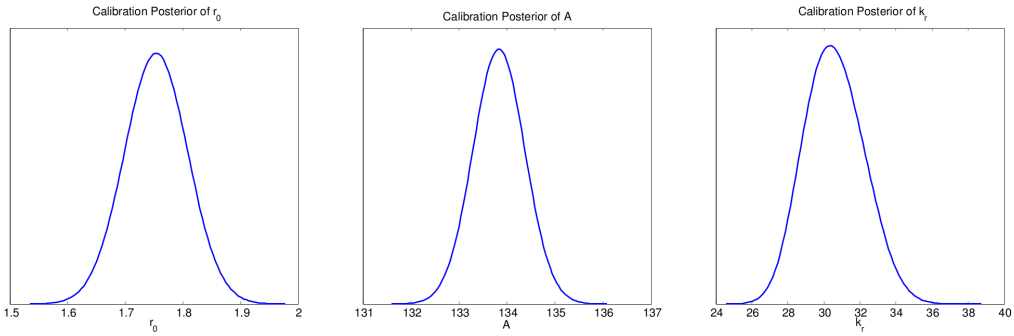

Figure 5.10: Kernel density estimates of the posterior distribution functions for the parameters of model $\mathcal{P}_1^*$ produced by the calibration Bayesian update, in which a chain of $C_{80}H_{162}$ is simulated in a canonical ensemble. These distributions are the prior distributions in the validation update.

to be converged. Otherwise, the number of Monte Carlo samples is increased in increments of 5000 until the tolerance is met. The converged posteriors for $\mathcal{P}_1^*$ are shown in Figure 5.10. The plausibilities, calculated by QUESO, as mentioned in Section 3.6, are given in Table 5.3. It is clear that model class $\mathcal{P}_1^*$, which includes only bonded interactions and the internal bead energy A, is the most plausible, but is it valid?

5.5 OPAL Step 5: The Validation Step

The model found to be the most plausible in Step 4 of OPAL is submitted to a validation test, or a series of validation tests, in accordance with the framework laid down in Chapter 3. The prior in the Bayesian validation update is the posterior from the calibration update, as discussed in Section 3.5. The Bayesian update is performed with a Gaussian likelihood (5.13), where in this step of OPAL, the observational data, y_i, and model data, $d_i(\boldsymbol{\theta})$, are taken in the validation scenario. As a validation scenario, two chains of $C_{80}H_{162}$ are simulated in a canonical ensemble, and the atomistic configuration and potential energy are recorded every 100 timesteps. The validation data is therefore a vector of potential energies at sample configurations ω_i,

$$\mathbf{y}_v = \{y_i\} = \{V(\omega_i)\}, \tag{5.26}$$

where $V(\omega_i)$ is calculated by (2.4). As in the calibration scenario, the CG data is determined, for each parameter vector $\boldsymbol{\theta}$, by placing the CG molecule into configuration ω_i and calculating the potential energy. That is,

$$d_i(\boldsymbol{\theta}) = V_{CG}(G(\omega_i); \boldsymbol{\theta}), \tag{5.27}$$

where G is the CG map, and potential energy V_{CG} is calculated via (5.2), where only the interactions included in the model contribute to the computation.

Meaning is given to the validation scenario when a tolerance on the difference between the target and predicted observables is defined. Adhering to the process laid down in Chapter 3, if the target observable is the distribution of the potential energy, the similarity between between the target (AA) and predicted (CG) distributions

may be measured by the Kullback-Leibler divergence (3.44). If, instead, the target observable is the ensemble average of the potential energy, the accuracy of the CG model is measured by the Euclidean distance (3.46) between the ensemble averages (2.14) and (2.23). In this work, the tolerance for the D_{KL}-distance is defined to be

$$\gamma_{tol,D_{KL}} = 0.15\sigma_V^2 \mathcal{O}\left(\mathbb{E}[\pi\left(V\right)]\right), \tag{5.28}$$

where σ_V^2 is the variance of the target observable and $\mathcal{O}\left(\mathbb{E}\left[\pi\left(V\right)\right]\right)$ is the order of magnitude of the expected value of the target observable. This definition captures the intuitive comparison of the location and spread of the distributions in question. The tolerance for the Euclidean metric is defined be

$$\gamma_{tol,Euc} = 0.1\left\langle V\right\rangle_{AA}, \tag{5.29}$$

so that the predicted ensemble average must be within 10% of the target ensemble average.

5.5.1 C_1^*: Category 1

Model $\mathcal{P}_1^*$ was shown to be the most plausible model in Section 5.4.3 and is therefore subjected to validation tests. The validation update is again performed using QUESO, with an initial amount of 5000 Monte Carlo samples, followed by an implementation using 10000 samples. As before, if the discrete Kullback-Leibler divergence, $\bar{D}_{KL}\left(\pi_{10000}\|\pi_{5000}\right)$, satisfies the tolerance (3.50), the posterior is considered to be converged, otherwise the number of Monte Carlo samples is increased. The converged posterior distributions for the validation update are given in Figure 5.11.

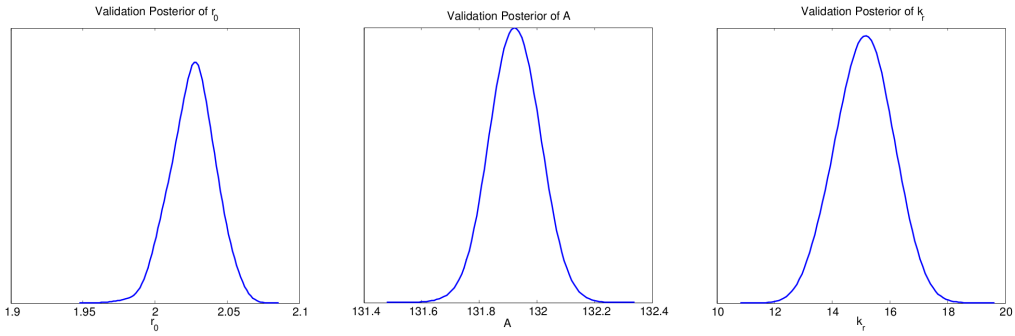

Figure 5.11: Kernel density estimates of the posterior distribution functions for the parameters of model $\mathcal{P}_1^*$ produced by the validation Bayesian update, in which two chains of $C_{80}H_{162}$ are simulated in a canonical ensemble.

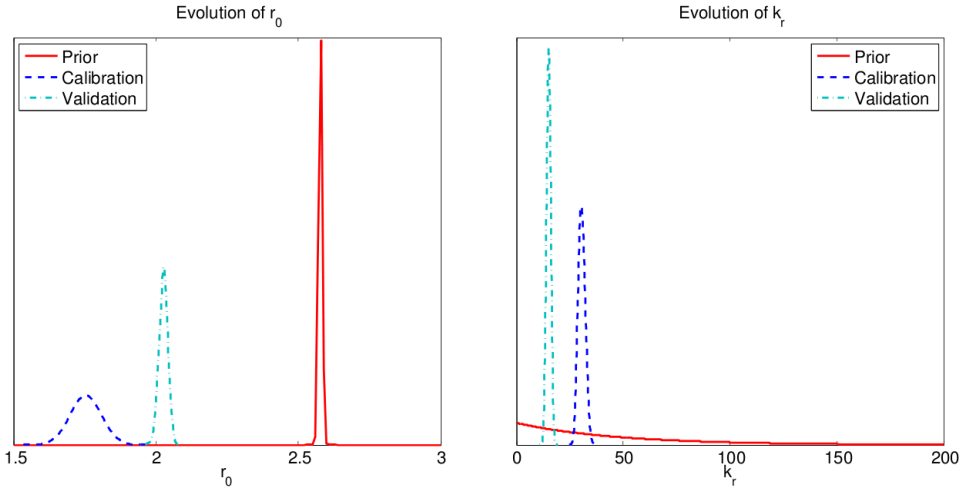

Figure 5.12: Figures of the prior and posterior probability distribution functions of model class $\mathcal{P}_1^*$ from the calibration and validation scenarios. Information about the behavior of the system is supplied to the parameters through Bayes' rule, and it can be seen that the distributions change as information is updated.

It has been discussed repeatedly in this work that information about the behavior of the system is added to the parameters through the calibration and validation process. This is evidenced by the evolution in the parameter distributions, shown in Figure 5.12.

The parameters are stochastically sampled from the converged posterior distribution and used in the forward problem of the validation scenario. The potential energy is measured every 100 timesteps throughout the simulation; the distributions of the potential energies produced by the AA model (the "truth") and that of the

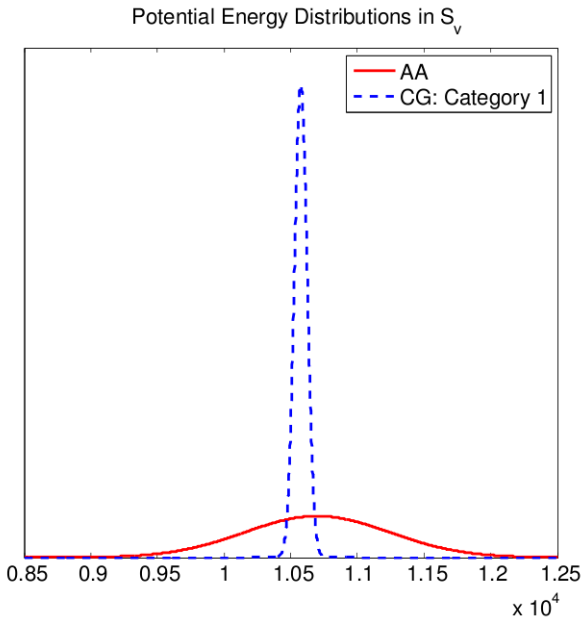

Figure 5.13: Distributions of the potential energies yielded by the AA and CG models, where the CG model is $\mathcal{P}_1^*$ from Category 1. Simulations were run in a canonical ensemble and consisted of two chains of $C_{80}H_{162}$. It is clear from this illustration that $\mathcal{P}_1^*$ should be deemed not invalid for predicting the potential energy of polyethylene.

CG model, denoted $\pi(V)$ and $\pi(V_{CG}(\boldsymbol{\theta}))$, respectively, are given in Figure 5.13. It is qualitatively clear that this C_1^* model should be deemed not invalid for predicting the potential energy of a system of polyethylene. Using the full distributions,

$$\gamma_{D_{KL}} = D_{KL}(\pi(V) \| \pi(V_{CG}(\boldsymbol{\theta}))) = 0.0622\sigma_V^2 \mathcal{O}(\mathbb{E}[\pi(V)]), \tag{5.30}$$

and using the ensemble averages,

$$\gamma_{Euc} = |\langle V \rangle_{AA} - \langle V_{CG}(\boldsymbol{\theta}) \rangle_{CG}| = 0.0118 \langle V \rangle_{AA}. \tag{5.31}$$

Thus, $\gamma_{D_{KL}} \leq \gamma_{tol,D_{KL}}$ and $\gamma_{Euc} \leq \gamma_{tol,Euc}$, rendering $\mathcal{P}_1^*$ not invalid for predicting the potential energy of polyethylene.

As a second validation test, consider a small cube of polyethylene, consisting of four chains of $C_{80}H_{162}$. Without updating the parameters, the system is simulated in a canonical ensemble in both the AA and CG systems and the potential energies are sampled, as before. The resulting distributions are shown in Figure 5.14, where it can be seen that model class $\mathcal{P}_1^*$ continues to accurately predict the potential energy. Specifically,

$$\gamma_{D_{KL}} = D_{KL}(\pi(V) \| \pi(V_{CG}(\boldsymbol{\theta}))) = 0.0826\sigma_V^2 \mathcal{O}(\mathbb{E}[\pi(V)]), \tag{5.32}$$

and,

$$\gamma_{Euc} = |\langle V \rangle_{AA} - \langle V_{CG}(\boldsymbol{\theta}) \rangle_{CG}| = 0.0181 \langle V \rangle_{AA}, \tag{5.33}$$

both of which are well below the specified validity criteria. This presents further evidence and increases one's confidence that the model parameters produced by the first validation update will yield an accurate approximation of the QoI in the prediction scenario.

Theoretically, the QoI is never observed; once it is measured and compared to model output it becomes another validation scenario. However, for the sake of illustration, this tenant rule will be broken. The AA and CG systems of the prediction scenario, a cube of polyethylene consisting of 25 chains of $C_{80}H_{162}$, are simulated in a canonical ensemble. The distributions of sampled potential energies for each system are shown in Figure 5.15, where it is seen that model $\mathcal{P}_1^*$ satisfies

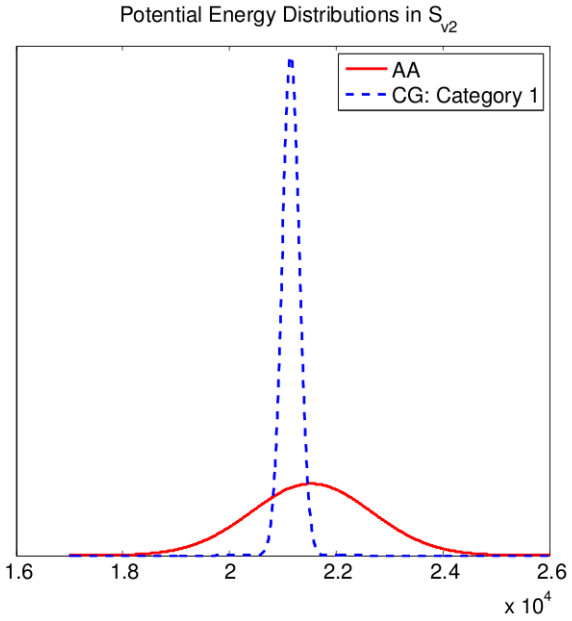

Figure 5.14: Distributions of the potential energies yielded by the AA and CG models, where the CG model is $\mathcal{P}_1^*$ from Category 1. Simulations were run in a canonical ensemble and consisted of four chains of $C_{80}H_{162}$ contained in a cube enforced with periodic boundary conditions. The fact that $\mathcal{P}_1^*$ is able to reproduce the potential energy accurately without further updating the parameters is evidence that it should be accepted as a reliable estimator for the QoI in the prediction scenario.

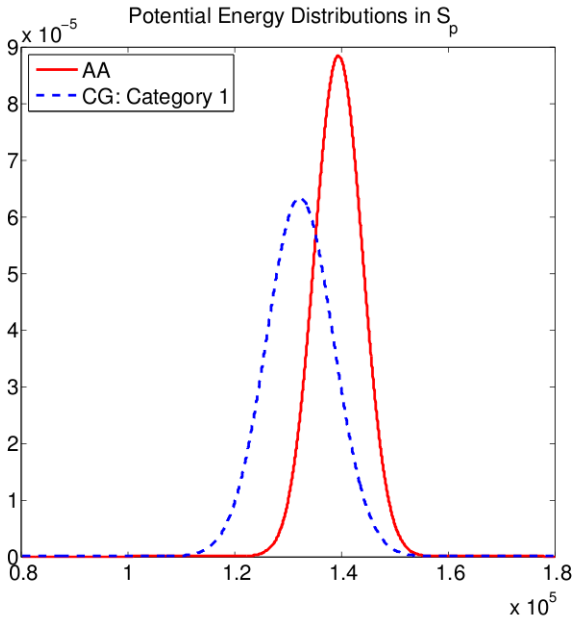

Figure 5.15: Distributions of the potential energies yielded by the AA and CG models in the prediction scenario, where the CG model is $\mathcal{P}_1^*$ from Category 1. Simulations were run in a canonical ensemble and consisted of 25 chains of $C_{80}H_{162}$ contained in a cube enforced with periodic boundary conditions. It can be seen that the CG model accurately reproduces the potential energy, validating the Bayesian framework for model selection, calibration, and validation.

both the D_{KL}-distance and Euclidean metric tolerances, with

$$\gamma_{D_{KL}} = D_{KL}\left(\pi\left(V\right)\|\pi\left(V_{CG}\left(\boldsymbol{\theta}\right)\right)\right) = 0.1143\sigma_V^2\mathcal{O}\left(\mathbb{E}[\pi\left(V\right)]\right), \tag{5.34}$$

and,

$$\gamma_{Euc} = \left|\langle V\rangle_{AA} - \langle V_{CG}\left(\boldsymbol{\theta}\right)\rangle_{CG}\right| = 0.0448\left\langle V\right\rangle_{AA}. \tag{5.35}$$

Model	Bonds	Angles	LJ 12-6	LJ 9-6	Params	Plausibility
$\mathcal{P}_1^*$	✓	✓			5	0
$\mathcal{P}_2^*$	✓		✓		5	1
$\mathcal{P}_3^*$	✓			✓	5	0
$\mathcal{P}_4^*$		✓	✓		5	0
$\mathcal{P}_5^*$		✓		✓	5	0

Table 5.4: Table of plausibilities calculated for the models in the Occam category C_2^*, produced during model calibration against data from a single chain of $C_{80}H_{162}$. Values shown as zero are approximate.

5.6 OPAL Step 6: Iteration

The validation tolerances (5.28) and (5.29) may be tightened such that the C_1^* model, $\mathcal{P}_1^*$, is rendered invalid. For example, define

$$\gamma_{tol,D_{KL}} = 0.06\sigma_V^2 \mathcal{O}\left(\mathbb{E}[\pi\left(V\right)]\right). \tag{5.36}$$

Then, $\mathcal{P}_1^*$ is invalid and OPAL moves into the iteration step, where the next category of models, C_2^*, is considered. Category 2 models each have five parameters, as shown in Table 5.2. The prior distributions for these parameters are the same as those previously described and shown in Figure 5.9, and the calibration data is the same as that used for the models in the first Occam category. The Bayesian update for each of the five models in C_2^* is again performed using QUESO. Once the posterior distributions for each model have converged, the normalized plausibilities may be calculated. These plausibilities are given in Table 5.4, and the converged posterior parameter distributions for $\mathcal{P}_2^*$, the most plausible model, are shown in Figure 5.16. Model class $\mathcal{P}_2^*$, containing bonded and Lennard-Jones interactions, is therefore passed into the validation step of OPAL.

The validation stage is conducted in the same way as described before. Data is collected from the AA system, which contains two chains of $C_{80}H_{162}$, and is used to update the parameters using Bayes' Rule, which is implemented using QUESO. The converged parameters, given in Figure 5.17, are stochastically sampled and used in a simulation of S_{v1}, during which samples of the potential energy are taken. The distribution of potential energy samples, taken from the AA system, as well as the CG system, represented by the new CG model $\mathcal{P}_2^*$, are shown in Figure 5.18. For comparison, the CG distribution produced by the most-plausible C_1^* model has

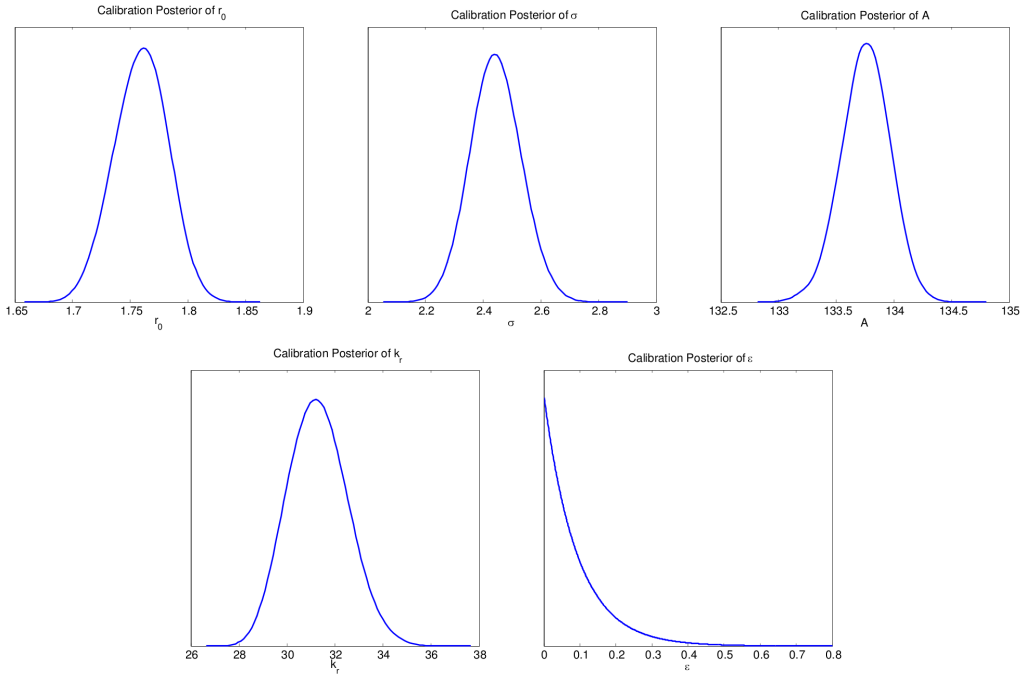

Figure 5.16: Kernel density estimates of the posterior distribution functions for the parameters of model class $\mathcal{P}_2^*$ yielded by the calibration Bayesian update, in which a chain of $C_{80}H_{162}$ is simulated in a canonical ensemble. These calibration posteriors become the validation priors in the validation update.

also been included. Although very similar, the distribution yielded by $\mathcal{P}_2^*$ is slightly closer to the AA distribution, both in the D_{KL}-distance,

$$\gamma_{D_{KL}} = D_{KL}\left(\pi\left(V\right)\|\pi\left(V_{CG}\left(\boldsymbol{\theta}\right)\right)\right) = 0.044\sigma_V^2\mathcal{O}\left(\mathbb{E}[\pi\left(V\right)]\right), \tag{5.37}$$

and the Euclidean metric,

$$\gamma_{Euc} = \left|\langle V\rangle_{AA} - \langle V_{CG}\left(\boldsymbol{\theta}\right)\rangle_{CG}\right| = 0.0115\,\langle V\rangle_{AA}\,. \tag{5.38}$$

Model $\mathcal{P}_2^*$ is therefore not invalid by the new validation tolerance (5.36).

As before, a second validation test is implemented, in which four chains of

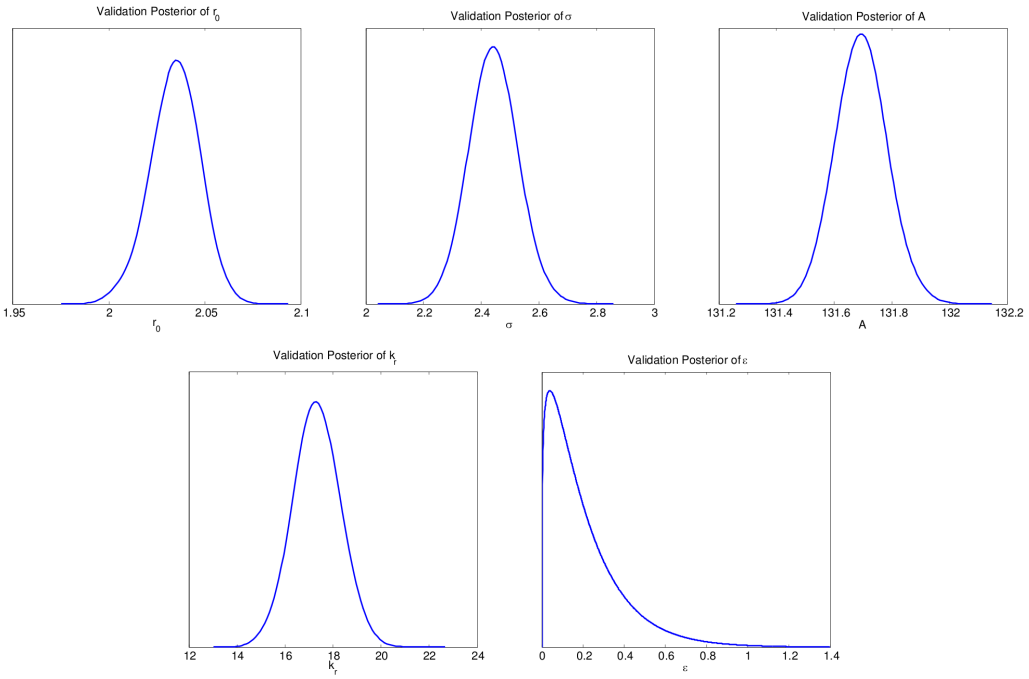

Figure 5.17: Kernel density estimates of the posterior distribution functions for the parameters of model class $\mathcal{P}_2^*$ yielded by the validation Bayesian update, in which two chains of $C_{80}H_{162}$ is simulated in a canonical ensemble.

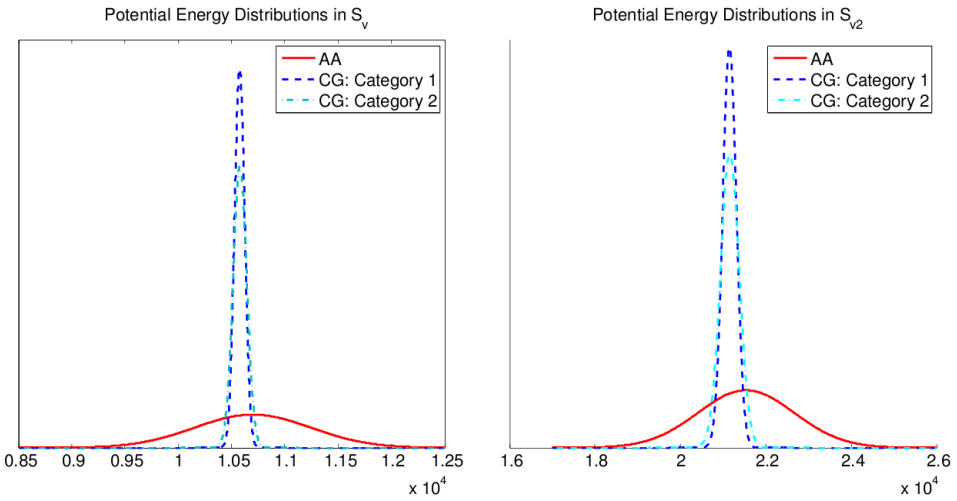

Figure 5.18: Distributions of the potential energies yielded by the AA and CG models in each of the two validation scenarios, where one CG model is $\mathcal{P}_1^*$ from Category 1 and the other is $\mathcal{P}_2^*$ from Category 2. Simulations were run in a canonical ensemble. It is clear from this illustration that $\mathcal{P}_2^*$ should be deemed not invalid for predicting the potential energy of polyethylene.

$C_{80}H_{162}$ are simulated in a canonical ensemble, the model parameters being stochastically sampled from the validation posterior yielded from S_{v1}. The resulting distribution of potential energy values, as well as those from the AA model and $\mathcal{P}_1^* \in C_1^*$, are shown in Figure 5.18. Again, the distributions produced by the two CG models are nearly identical, but the results from $\mathcal{P}_2^*$ are slightly better,

$$\gamma_{D_{KL}} = D_{KL}\left(\pi\left(V\right) \| \pi\left(V_{CG}\left(\boldsymbol{\theta}\right)\right)\right) = 0.0587\sigma_V^2 \mathcal{O}\left(\mathbb{E}[\pi\left(V\right)]\right), \tag{5.39}$$

and,

$$\gamma_{Euc} = \left|\langle V \rangle_{AA} - \langle V_{CG}\left(\boldsymbol{\theta}\right)\rangle_{CG}\right| = 0.0178\,\langle V \rangle_{AA}\,. \tag{5.40}$$

From the demonstration that $\mathcal{P}_2^*$ is able to meet the validation criteria, yet again, it is evident that this model should be able to accurately predict the QoI in the

prediction scenario. This is confirmed by a simulation of the prediction scenario using $\mathcal{P}_2^*$, the results of which are shown in Figure 5.19, where the distributions of the potential energy, sampled from the AA and CG systems are shown. As with the first iteration of OPAL, the CG model satisfies both the D_{KL}-distance and Euclidean metric tolerances, with

$$\gamma_{D_{KL}} = D_{KL}\left(\pi\left(V\right) \| \pi\left(V_{CG}\left(\boldsymbol{\theta}\right)\right)\right) = 0.1060\sigma_V^2 \mathcal{O}\left(\mathbb{E}[\pi\left(V\right)]\right), \tag{5.41}$$

and,

$$\gamma_{Euc} = \left|\langle V \rangle_{AA} - \langle V_{CG}\left(\boldsymbol{\theta}\right)\rangle_{CG}\right| = 0.0445 \langle V \rangle_{AA}. \tag{5.42}$$

Again, the model constructed in the second iteration of OPAL is slightly closer to the truth than the Category 1 model, $\mathcal{P}_1^*$.

This example application to polyethylene illustrates a successful implementation of OPAL for the purposes of constructing the simplest, valid CG model. The computational benefits of this algorithm are two-fold. By eliminating models based on sensitivity analysis and separating the set of possible models into smaller categories, the number of models that require calibration may be reduced, thus possibly lowering the cost of the inverse analysis. By starting the calibration process with the simplest models, the forward problem will be as inexpensive as possible.

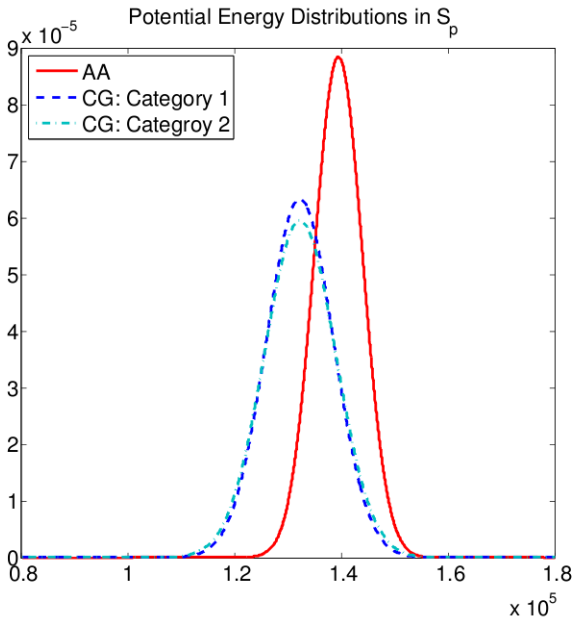

Figure 5.19: Distributions of the potential energies yielded by the AA and CG models in the prediction scenario, where the CG model $\mathcal{P}_1^*$ is from Category 1 and the CG model $\mathcal{P}_2^*$ is from Category 2. Simulations were run in a canonical ensemble and consisted of 25 chains of $C_{80}H_{162}$ contained in a cube enforced with periodic boundary conditions. It can be seen that the CG model accurately reproduces the potential energy, validating the Bayesian framework for model selection, calibration, and validation.

Chapter 6

Other Theoretical Results

Throughout this work, the prevalent goal has been finding a model that is "good enough" in some measure. The Bayesian framework advocated here is designed to build predictive models in which the modeler has confidence while simultaneously acknowledging that the truth is never actually known. As the Bayesian statistician George E. P. Box once wrote, "all models are wrong, but some are useful" [15]. Box was, interestingly enough, the son-in-law of frequentist Ronald A. Fisher. Much of statistics literature on parametric models addresses the concept of model specificity. If there exists a specific parameter vector $\boldsymbol{\theta}^*$ that the model $\mathcal{P}$ maps into the true observational data, the model is said to be *well-specified*; if no such parameter exists, the model is said to be *misspecified*.

Consider a space $\mathcal{Y}$ of physical observations (for example, observables samples from the AA model) and a set $\mathbb{M}(\mathcal{Y})$ of probability measures μ on $\mathcal{Y}$. As always, a target quantity of interest $Q : \mathbb{M} \to \mathbb{R}$ is selected. A particular measure, μ^*, is sought, from which the "true" value of the QoI, μ^*, may be computed. The goal is to predict $Q(\mu^*)$ using a parametric model, $\mathcal{P} : \Theta \to \mathbb{M}(Y)$, where Θ is the space of parameters. As noted, if there exists a parameter $\boldsymbol{\theta}^* \in \Theta$ such that $\mathcal{P}(\boldsymbol{\theta}^*) = \mu^*$, the model is well-specified; otherwise if $\mu^* \notin \mathcal{P}(\Theta)$, the model is misspecified [88].

Bayesian frameworks provide a general setting for the analysis of such mod-

els, whether they are well-specified to misspecified, and their predictive capabilities in the presence of uncertainties. Let the observational data $\mathbf{y}$ be a vector of independent and identically distributed ($i.i.d$) samples from μ^* such that $\mathbf{y}^n = \{y_1, y_2, \ldots, y_n\} \in \mathcal{Y}^n \subset \mathcal{Y}$, and let $\mathcal{M}$ be a set of m parametric models, each with its own parameter space,

$$\mathcal{M} = \{\mathcal{P}_1(\boldsymbol{\theta}_1), \mathcal{P}_2(\boldsymbol{\theta}_2), \ldots, \mathcal{P}_m(\boldsymbol{\theta}_m)\}, \quad \boldsymbol{\theta}_k \in \Theta_k, \quad 1 \leq k \leq m. \tag{6.1}$$

As discussed in Chapter 3, prior information regarding parameters $\boldsymbol{\theta}_i$ may be collected and characterized by prior probability density functions, $\pi(\boldsymbol{\theta}_i|\mathcal{P}_i, \mathcal{M})$ and updated via Bayes' Rule,

$$\pi(\boldsymbol{\theta}_i|\mathbf{y}^n, \mathcal{P}_i, \mathcal{M}) = \frac{\pi(\mathbf{y}^n|\boldsymbol{\theta}_i, \mathcal{P}_i, \mathcal{M})\,\pi(\boldsymbol{\theta}_i|\mathcal{P}_i, \mathcal{M})}{\pi(\mathbf{y}^n|\mathcal{P}_i, \mathcal{M})}. \tag{6.2}$$

Recall that the denominator is a marginalization of the numerator over the parameters and becomes the likelihood in the model plausibility,

$$\rho_i = \pi(\mathcal{P}_i|\mathbf{y}^n, \mathcal{M}) = \frac{\pi(\mathbf{y}^n|\mathcal{P}_i, \mathcal{M})\,\pi(\mathcal{P}_i|\mathcal{M})}{\pi(\mathbf{y}^n|\mathcal{M})}. \tag{6.3}$$

The model with plausibility closest to unity is deemed the most plausible model in the set $\mathcal{M}$.

6.1 Model Misspecification

Suppose that the model $\mathcal{P}$ (or $\mathcal{P}_i$) is misspecified, *i.e.* $\mu^* \notin \mathcal{P}(\Theta)$. Suppose further that μ^* is absolutely continuous with respect to the Lebesgue measure and that $g(\mathbf{y})$ is the probability density associated with μ^*. Then the best approximation

to g in $\mathcal{P}(\Theta)$ is the model with the parameter

$$\boldsymbol{\theta}^\dagger = \underset{\Theta}{\operatorname{argmin}}\, D_{KL}(g(\mathbf{y}) \| \pi(\mathbf{y}|\boldsymbol{\theta}, \mathcal{P}, \mathcal{M})), \tag{6.4}$$

where $D_{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence,

$$D_{KL}(p(\mathbf{y})\|q(\mathbf{y})) = \int_{\mathcal{Y}} p(\mathbf{y}) \log \frac{p(\mathbf{y})}{q(\mathbf{y})}\, d\mathbf{y}. \tag{6.5}$$

The parameter $\boldsymbol{\theta}^\dagger$ yields a probability measure, $\mu^\dagger = \mathcal{P}(\boldsymbol{\theta}^\dagger)$, that is as close as possible to μ^* in the D_{KL} pseudo-measure.

It is easily shown that $\boldsymbol{\theta}^\dagger$ is the maximum likelihood estimate, *i.e.* it maximizes the expected value of the log-likelihood relative to the true density, g. To see this, expand (6.4) to obtain

$$\boldsymbol{\theta}^\dagger = \underset{\Theta}{\operatorname{argmin}} \left[\int_{\mathcal{Y}} g(\mathbf{y}) \log g(\mathbf{y}) d\mathbf{y} - \int_{\mathcal{Y}} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta}, \mathcal{P}, \mathcal{M}) d\mathbf{y} \right]. \tag{6.6}$$

It can be seen that the negative self-entropy,

$$\int_{\mathcal{Y}} g(\mathbf{y}) \log g(\mathbf{y}) d\mathbf{y}, \tag{6.7}$$

does not depend on $\boldsymbol{\theta}$, and therefore does not affect the optimization. Then,

$$\begin{aligned} \boldsymbol{\theta}^\dagger &= \underset{\Theta}{\operatorname{argmin}} \left[- \int_{\mathcal{Y}} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta}, \mathcal{P}, \mathcal{M}) d\mathbf{y} \right] \\ &= \underset{\Theta}{\operatorname{argmax}} \int_{\mathcal{Y}} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta}, \mathcal{P}, \mathcal{M}) d\mathbf{y} \\ &= \underset{\Theta}{\operatorname{argmax}}\, \mathbb{E}_g \left[\log \pi(\mathbf{y}|\boldsymbol{\theta}, \mathcal{P}, \mathcal{M}) \right]. \end{aligned} \tag{6.8}$$

Thus $\boldsymbol{\theta}^\dagger$ maximizes the likelihood distribution function relative to the true density, g.

The parameter $\boldsymbol{\theta}^\dagger$ is of fundamental significance in the theories of the asymptotic behavior of parameter estimates of both Bayesian and frequentist statistics. Under suitable smoothness assumptions, as more data becomes available, the posterior density characterized by $\pi(\boldsymbol{\theta}^n|\mathbf{y}^n, \mathcal{P}, \mathcal{M})$ through Bayes's Rule converges in total variation to a normal distribution centered at $\boldsymbol{\theta}^\dagger$,

$$\pi\left(\boldsymbol{\theta}^n|\mathbf{y}^n, \mathcal{P}, \mathcal{M}\right) \to \mathcal{N}\left(\boldsymbol{\theta}^\dagger, V^{-1}\left(\boldsymbol{\theta}^\dagger\right)\right), \tag{6.9}$$

where the covariance matrix is given by [35, 64]

$$V_{ij}(\boldsymbol{\theta}^\dagger) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} D_{KL}\left(g\left(\mathbf{y}\right) \| \pi\left(\mathbf{y}|\boldsymbol{\theta}\right)\right). \tag{6.10}$$

This result is referred to as the Bernstein-von Mises Theorem for misspecified models (see *e.g.* [11, 64, 114]). Note that if $\mathcal{P}$ is well-specified, the Bayesian Central Limit Theorem, under suitable conditions [61, 62, 80], asserts that the posterior instead converges, in total variation, to a normal distribution centered at $\boldsymbol{\theta}^*$,

$$\pi\left(\boldsymbol{\theta}^n|\mathbf{y}^n, \mathcal{P}, \mathcal{M}\right) \to \mathcal{N}\left(\boldsymbol{\theta}^*, I^{-1}\left(\boldsymbol{\theta}^*\right)\right), \tag{6.11}$$

with covariance given by the Fisher information matrix,

$$I_{ij}\left(\boldsymbol{\theta}^*\right) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln \pi\left(\mathbf{y}|\boldsymbol{\theta}\right). \tag{6.12}$$

6.2 Plausibility-D_{KL} Theory

Suppose that the modeler has a choice between two misspecified models, $\mathcal{P}_1$ and $\mathcal{P}_2$. These models may be compared in the Bayesian setting through the concept of model plausibility: if $\mathcal{P}_1$ is more plausible than $\mathcal{P}_2$, $\rho_1 > \rho_2$, with plausibilities

calculated using (6.3); see Section 3.4 for details. When, instead, a maximum likelihood approach is used, the model that yields a probability measure closer to g is considered a "better" model, *i.e.* if

$$D_{KL}\left(g\left(\mathbf{y}\right)\|\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger,\mathcal{P}_1,\mathcal{M}\right)\right)\leq D_{KL}\left(g\left(\mathbf{y}\right)\|\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger,\mathcal{P}_2,\mathcal{M}\right)\right). \tag{6.13}$$

The theorems presented in this chapter are also given in [88] and define the relationship between these two notions of model comparison.

Bayesian and frequentist methods fundamentally differ in the way they view the model parameters. Bayesian methods consider parameters to be stochastic, characterized by probability density functions, while frequentist approaches seek a single, deterministic parameter value. To bridge this gap in methodology, it may be noted that considering a parameter vector to be deterministic is equivalent to assigning it a delta function as a probability distribution function.

To achieve a delta function posterior, the prior must necessarily be a delta function. Suppose, for example, the deterministic (*e.g.* maximum likelihood) parameter is $\boldsymbol{\theta}_i^\dagger$ for model $\mathcal{P}_i$. Then

$$\pi\left(\boldsymbol{\theta}_i^\dagger|\mathbf{y},\mathcal{P}_i,\mathcal{M}\right)=\pi\left(\boldsymbol{\theta}_i^\dagger|\mathcal{P}_i,\mathcal{M}\right)=\delta\left(\boldsymbol{\theta}_i-\boldsymbol{\theta}_i^\dagger\right), \tag{6.14}$$

and the evidence (6.3) becomes

$$\begin{aligned}\pi\left(\mathbf{y}|\mathcal{P}_i,\mathcal{M}\right) &= \int\pi\left(\mathbf{y}|\boldsymbol{\theta}_i,\mathcal{P}_i,\mathcal{M}\right)\delta\left(\boldsymbol{\theta}_i-\boldsymbol{\theta}_i^\dagger\right)d\boldsymbol{\theta}_i &(6.15)\\ &= \pi\left(\mathbf{y}|\boldsymbol{\theta}_i^\dagger,\mathcal{P}_i,\mathcal{M}\right). &(6.16)\end{aligned}$$

The ratio of posterior model plausibilities then becomes

$$\frac{\rho_1}{\rho_2} = \frac{\pi\left(\mathbf{y}|\mathcal{P}_1,\mathcal{M}\right)\pi\left(\mathcal{P}_1|\mathcal{M}\right)}{\pi\left(\mathbf{y}|\mathcal{P}_2,\mathcal{M}\right)\pi\left(\mathcal{P}_2|\mathcal{M}\right)} \tag{6.17}$$

$$= \frac{\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger,\mathcal{P}_1,\mathcal{M}\right)\pi\left(\mathcal{P}_1|\mathcal{M}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger,\mathcal{P}_2,\mathcal{M}\right)\pi\left(\mathcal{P}_2|\mathcal{M}\right)} \tag{6.18}$$

$$= \frac{\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger,\mathcal{P}_1,\mathcal{M}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger,\mathcal{P}_2,\mathcal{M}\right)}O_{12}, \tag{6.19}$$

where O_{12} is the ratio of prior odds and is often assumed to be one. With these tools in hand, the following theorems may be presented.

Theorem 6.2.1. *Under the condition that the prior distribution for $\mathcal{P}_1$ is a delta function centered at the corresponding maximum likelihood parameter, $\boldsymbol{\theta}_1^\dagger$, and the prior distribution for $\mathcal{P}_2$ is a delta function centered at the maximum likelihood parameter $\boldsymbol{\theta}_2^\dagger$, if model $\mathcal{P}_1$ is more plausible than model $\mathcal{P}_2$ and $O_{12} \leq 1$, then*

$$D_{KL}\left(g\left(\boldsymbol{y}\right)\|\pi\left(\boldsymbol{y}|\boldsymbol{\theta}_1^\dagger,\mathcal{P}_1,\mathcal{M}\right)\right) \leq D_{KL}\left(g\left(\boldsymbol{y}\right)\|\pi\left(\boldsymbol{y}|\boldsymbol{\theta}_2^\dagger,\mathcal{P}_2,\mathcal{M}\right)\right). \tag{6.20}$$

Proof. Given that the prior distributions for $\mathcal{P}_1$ an $\mathcal{P}_2$ are defined as delta functions centered at their respective maximum likelihood parameters, a Bayesian update with data $\mathbf{y}$ yields model evidences according to (6.16) for $i = 1, 2$, as well as the ratio of plausibilities (6.19). Then, if model $\mathcal{P}_1$ is more plausible than model $\mathcal{P}_2$,

$$1 < \frac{\rho_1}{\rho_2} = \frac{\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger,\mathcal{P}_1,\mathcal{M}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger,\mathcal{P}_2,\mathcal{M}\right)}O_{12} \leq \frac{\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger,\mathcal{P}_1,\mathcal{M}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger,\mathcal{P}_2,\mathcal{M}\right)}. \tag{6.21}$$

Since logarithms are monotonic, this implies

$$0 < \log\frac{\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger,\mathcal{P}_1,\mathcal{M}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger,\mathcal{P}_2,\mathcal{M}\right)}, \tag{6.22}$$

and, due to the fact that the density g is non-negative for every value of $\mathbf{y}$,

$$0 < g\left(\mathbf{y}\right) \log \frac{\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M}\right)}. \tag{6.23}$$

Consequently,

$$0 < \int g\left(\mathbf{y}\right) \log \frac{\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M}\right)} d\mathbf{y}, \tag{6.24}$$

which leads to the inequality,

$$0 < \int g\left(\mathbf{y}\right) \log \frac{g\left(\mathbf{y}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M}\right)} d\mathbf{y} - \int g\left(\mathbf{y}\right) \log \frac{g\left(\mathbf{y}\right)}{\pi\left(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M}\right)} d\mathbf{y}, \tag{6.25}$$

By moving the second integral to the left-hand side of the inequality, the desired result holds. □

This theorem demonstrates that if $\mathcal{P}_1$, with maximum likelihood parameter $\boldsymbol{\theta}_1^\dagger$, is "better" than $\mathcal{P}_2$, with maximum likelihood parameter $\boldsymbol{\theta}_2^\dagger$, in the Bayesian sense, it is also a "better" deterministic model. However, the reverse implication requires much stronger conditions.

Theorem 6.2.2. *Suppose that*

$$D_{KL}\left(g\left(\boldsymbol{y}\right) \| \pi\left(\boldsymbol{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M}\right)\right) \leq D_{KL}\left(g\left(\boldsymbol{y}\right) \| \pi\left(\boldsymbol{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M}\right)\right). \tag{6.26}$$

and the observation space, $\mathcal{Y}^n$, is finite. If the ratio of likelihood distributions,

$$\frac{\pi\left(\boldsymbol{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M}\right)}{\pi\left(\boldsymbol{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M}\right)}, \tag{6.27}$$

is continuous, then there exists a $\bar{\boldsymbol{y}} \in \mathcal{Y}^n$ such that $\mathcal{P}_1$ is more plausible than model $\mathcal{P}_2$, given that $O_{12} \geq 1$, assuming that the prior distributions in the Bayesian updates are delta functions centered at $\boldsymbol{\theta}_1^\dagger$ and $\boldsymbol{\theta}_2^\dagger$ for $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively.

Proof. Consider the equivalent form of the assertion,

$$\int_{\mathcal{Y}^n} g(\mathbf{y}) \log \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}\, d\mathbf{y} < 0. \tag{6.28}$$

For this inequality to hold, the relationship

$$\frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 1 \tag{6.29}$$

does not necessarily need to be true for *every* point $\mathbf{y} \in \mathcal{Y}^n$. However, due to the continuity of the integrand, the Mean Value Theorem [21] guarantees the existence of $\bar{\mathbf{y}} \in \mathcal{Y}^n$ such that

$$\int_{\mathcal{Y}^n} g(\mathbf{y}) \log \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}\, d\mathbf{y} = |\mathcal{Y}^n|\, g(\bar{\mathbf{y}}) \log \frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}. \tag{6.30}$$

Combining (6.28)-(6.30),

$$|\mathcal{Y}^n|\, g(\bar{\mathbf{y}}) \log \frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 0. \tag{6.31}$$

Since $|\mathcal{Y}^n| > 0$ and $g(\mathbf{y}) > 0$,

$$\log \frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 0. \tag{6.32}$$

Thus

$$\frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 1, \tag{6.33}$$

or, equivalently,

$$\frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})} > 1. \tag{6.34}$$

If $O_{12} \geq 1$,

$$\frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})} O_{12} > 1. \tag{6.35}$$

The left-hand side is the ratio of posterior model plausibilities (6.19). Thus model $\mathcal{P}_1$ is more plausible than $\mathcal{P}_2$ for given data $\bar{\mathbf{y}}$. □

These theorems characterize the relationship between the Bayesian and frequentist notions of model comparison. It has been demonstrated that the most plausible model and the deterministic (maximum likelihood) model in which parameters minimize the D_{KL}-distance between the model output and the so-called truth parameters are, under stated assumptions, equivalent. Whether the resulting "best" model is valid for the intended purpose depends on the severity of the validation criteria, as laid down in Chapter 3.

Chapter 7

Conclusions

This study has explored fundamental questions in computational science and engineering, regarding the construction of coarse-grained models of atomistic systems. The issues addressed herein have included model selection, calibration, and validation in the presence of uncertainties in observables, model reduction, and quantities of interest, all through the use of a unified framework with foundations in Bayes' Rule, information theory, and sensitivity analysis. Bayesian statistical inverse analysis provides a powerful approach to the calibration and validation of coarse-grained models of atomistic systems that inherently copes with uncertainties in the data and model selection. Posterior model plausibilities, provided by Bayes' Rule, are employed to select the model that best fits the given data. The data, provided by the "ground truth" all-atom model, is used to inform the parameters through increasingly complex model scenarios. As the model is tested and updated, confidence in its ability to predict predetermined quantities of interest is increased.

Sensitivity analysis is not only used to eliminate models with parameters to which the QoI is insensitive, thereby dramatically reducing model complexity and the cost of assessing model plausibility, it is also used to check the effectiveness of the scenario-data pairs in the calibration and validation scenarios. A parameter's calculated sensitivity indicates its contribution to the quantity being observed

and, simultaneously, indicates how much it is informed by the inverse analysis. By mandating that the sensitivity results in the calibration, validation, and prediction scenarios agree on the relative importance of the parameters, both the contribution of the parameters to the quantity of the interest and the ability of Bayes' Rule to update the parameters for given data are accounted for.

Bayes' Rule, when coupled with sensitivity analysis and an appropriate interpretation of Occam's Razor, lends itself to a new, adaptive algorithm, OPAL, the Occam-Plausibility ALgorithm, that provides a systematic means of constructing the simplest model suitable for use in a prediction scenario. After a set of possible model classes is identified, a sensitivity analysis is performed to determine which parameters are insignificant. Eliminating these unnecessarily complex models reduces the computational cost of calculating model plausibilities. By separating the remaining models into categories according to complexity, OPAL ensures that the simplest valid model will be chosen for use in the prediction scenario. In this way, OPAL provides a systematic approach to determining and reducing model inadequacy to within preset tolerances.

In future work, it is hoped that the processes developed in the present study can be extended to multiscale applications of material science. Just as the all-atom model was used as data in the calibration and validation of coarse-grained models, the validated coarse-grained model may be used as calibration and validation data of macro-scale models. By using OPAL, the simplest valid continuum model may be identified for nanoscale materials. Properties such as average lattice deformations, tensile strength, or elastic modulus are possible quantities of interest. Implemen-

tation of OPAL can be computationally expensive for some choices of QoI, despite the reduced cost, indicating that additional research on improved parallel molecular dynamics or Monte-Carlo algorithms is needed.

Appendices

Appendix A

Empirical Force Fields

This appendix is included to provide a broad introductory overview of the most common force field used in popular MD codes. These form the basis for generating parametric model classes for CG systems as well as producing synthetic data from AA systems such as those introduced in Chapter 2.

The function, V, that describes the potential energy of an atomistic or particulate system can be written as the sum of pair and multi-body interactions of varying complexity. Given a configuration $\mathbf{r}^n$,

$$V(\mathbf{r}^n) = V_{bond}(\mathbf{r}^n) + V_{angle}(\mathbf{r}^n) + V_{dihedral}(\mathbf{r}^n) + V_{nb}(\mathbf{r}^n). \tag{A.1}$$

The bonded pair potential, V_{bond}, is the total potential energy due to N_b bonds,

$$V_{bond} = \sum_{i=1}^{N_b} E_{r,i}, \tag{A.2}$$

where $E_{r,i}$ is the potential energy contained in bond i. Bonds are most often represented as harmonic springs, thus

$$E_r = \frac{1}{2} k_r (r - r_0)^2, \tag{A.3}$$

k_r being the spring constant, r_0 the equilibrium bond length, and r the instantaneous distance between the bonded atoms. The parameters in this expression are k_r and r_0.

Three sequentially bonded atoms or particles define an angle whose potential may also be described by a harmonic spring,

$$E_\theta = \frac{1}{2} k_\theta (\theta - \theta_0)^2. \tag{A.4}$$

Here, as in the harmonic bond, k_θ is the spring constant, θ_0 is the equilibrium angle, and θ is the instantaneous angle created by the three particles. As before, k_θ and θ_0 are the parameters. The total contribution to the potential energy of N_a angular interactions is thus

$$V_{angle} = \sum_{i=1}^{N_a} E_{\theta,i}. \tag{A.5}$$

Torsional interactions are defined by four atoms that are sequentially bonded. The dihedral angle φ is the angle between the first and third bond, and the total potential energy contained in N_d dihedral interactions is given by

$$V_{dihedral} = \sum_{i=1}^{N_d} E_{\varphi,i}, \tag{A.6}$$

where $E_{\varphi,i}$ is the contribution of torsion i, which is almost always represented as a cosine expansion. For example, in AMBER [20, 117],

$$E_\varphi = \sum_n \frac{V_n}{2} \left[1 + \cos(n\varphi - \gamma)\right], \tag{A.7}$$

where n is a parameter that varies depending on the atoms or particles involved, V_n is a scalar coefficient, and γ is a phase offset that changes the location of the extrema of the potential. By contrast, the OPLS expression of torsional interactions always contains four terms [54, 55],

$$E_\varphi = \frac{V_1}{2}(1+\cos(\varphi)) + \frac{V_2}{2}(1-\cos(2\varphi)) + \frac{V_3}{2}(1+\cos(3\varphi)) + \frac{V_4}{2}(1-\cos(4\varphi)), \tag{A.8}$$

though one or more of the coefficients V_n are often set to zero. Note that this form can be rewritten in the form (A.7) by setting $\gamma = 0$ for $n = 1, 3$ and $\gamma = 180$ for $n = 2, 4$.

It is not uncommon that an additional four-body potential also contributes to the total potential $V(\mathbf{r}^n)$. Improper torsions in, for example, the CHARMM force field are included to maintain chirality, a property of asymmetry, of bonded atoms about a heavy atom. This potential is harmonic [16],

$$E_\omega = \frac{1}{2} k_\omega (\omega - \omega_0)^2, \tag{A.9}$$

where k_ω is the spring constant and ω and ω_0 are the instantaneous and equilibrium improper angles, respectively. Neither AMBER nor OPLS include this type of torsion in their force fields. As was the case with the previous interaction types, the total energy from improper torsion interactions is given by a summation over all N_t instances,

$$V_{torsion} = \sum_{i=1}^{N_t} E_{\omega,i}. \tag{A.10}$$

The non-bonded potential, $V_{non-bond}$ is the total potential energy due to pair interactions between particles on different molecules or between particles separated by at least three bonds, such as van der Waals and electrostatic interactions. Electrostatics are represented by a Coulomb potential,

$$E_{elec} = \frac{q_i q_j}{4\pi\epsilon_0 r}. \tag{A.11}$$

The parameters q_i and q_j are the charges on the two interacting atoms and r is the instantaneous distance between them. The permittivity constant, ϵ_0, has the value

$$\epsilon_0 = 8.85 \times 10^{-12} \ \left(\mathrm{Nm}^2/\mathrm{C}^2\right)^{-1}. \tag{A.12}$$

Capturing attractive and repulsive forces between atoms or particles other than those due to bonded or electrostatic interactions, van der Waals interactions are most often represented by a Lennard-Jones potential,

$$E_{lj} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right], \tag{A.13}$$

where r is again the instantaneous distance between the two particles, ϵ is the well-depth, and σ is the collision parameter and the distance at which the potential is zero. At times, a softer potential, the Lennard-Jones 9-6 potential,

$$E_{lj} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{9} - \left(\frac{\sigma}{r} \right)^{6} \right], \tag{A.14}$$

may be desired, but this is less common than (A.13). In either representation, ε and σ are the model parameters.

The total energy from non-bonded interactions is summed over all pairs of n atoms separated by three or more bonds,

$$V_{nb} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[E_{elec} + E_{lj} \right] f. \tag{A.15}$$

In OPLS, $f = 0.5$ for atoms separated by exactly three bonds and 1.0 otherwise, but it is not unusual to specify $f = 1.0$ for all non-bonded interactions. The combining rules $\sigma_{ij} = (\sigma_{ii}\sigma_{jj})^{1/2}$ and $\varepsilon_{ij} = (\varepsilon_{ii}\varepsilon_{jj})^{1/2}$ are the most common.

It is important to note that the functional forms discussed above are widely accepted. Parameter values in popular codes such as AMBER, CHARMM, and OPLS are widely trusted and used in a variety of applications. However, other potentials may be used if considered more appropriate. The embedded atom model [24, 25] or the Morse potential [78], for example, are among popular alternatives.

The values of the parameters are defined by the atoms involved in the interaction, and the general environment in which these atoms exist. The bond between a carbon atom and a hydrogen atom, for example, is much weaker than the bond between two carbon atoms, which can be further distinguished based on whether these two carbons are part of a benzene ring or t-butyl component. This is reflected in the values of the spring coefficients k_r. Values can generally be found in parameter tables or in the literature [16, 20, 54, 55, 117]

Appendix B

Sensitivity Analysis

Let $Y = f(\theta_1, \theta_2, \ldots, \theta_k)$ be the output function to be observed in any given model scenario such that f is a square-integrable function over the k-dimensional unit hypercube,

$$\Omega^k = (\boldsymbol{\theta} | 0 \leq \theta_i \leq 1, i = 1, \ldots, k) \,. \tag{B.1}$$

Using the Hoeffding decomposition, the Russian mathematician Ilya M. Sobol' [111, 112] illustrated that f may be expanded as

$$f = f_0 + \sum_i f_i + \sum_i \sum_{j>i} f_{ij} + \ldots + f_{12\ldots k}, \tag{B.2}$$

where each term is a function of the indices, for example, $f_i = f_i(\theta_i)$, $f_{ij} = f_{ij}(\theta_i, \theta_j)$, *etc.* In particular, each function in the decomposition is an expected value,

$$f_0 = \mathbb{E}(Y), \tag{B.3}$$

$$f_i = \mathbb{E}\left(Y | \theta_i\right) - f_0, \tag{B.4}$$

$$f_{ij} = \mathbb{E}\left(Y | \theta_i, \theta_j\right) - f_i - f_j - f_0. \tag{B.5}$$

By square integrating each term in the decomposition (B.2), the ANOVA-HDMR decomposition may be written [104],

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + V_{12\ldots k}, \tag{B.6}$$

where $V(Y)$ is the variance of Y, $V_i = V(f_i(\theta_i))$ is the variance in Y due to parameter θ_i, $V_{ij} = V(f_{ij}(\theta_i, \theta_j))$ is the variance in Y due to the interaction of parameters θ_i and θ_j, and so on.

Alternatively, the variance $V(Y)$ may be expressed as a function of all of the parameters *except* θ_i,

$$V(Y) = V\left(\mathbb{E}\left(Y|\boldsymbol{\theta}_{\sim i}\right)\right) + \mathbb{E}\left(V\left(Y|\boldsymbol{\theta}_{\sim i}\right)\right). \tag{B.7}$$

The quantity $V(Y) - V\left(\mathbb{E}\left(Y|\boldsymbol{\theta}_{\sim i}\right)\right) = \mathbb{E}\left(V\left(Y|\boldsymbol{\theta}_{\sim i}\right)\right)$ is the expected value with respect to the joint distribution $\pi\left(\boldsymbol{\theta}_{\sim i}\right)$ of the variance that remains in Y when the parameters $\boldsymbol{\theta}_{\sim i}$ are fixed. Dividing both sides by $V(Y)$ yields two expressions for a variance-based sensitivity measure, the so-called *total effect index*,

$$S_{T_i} = \frac{\mathbb{E}\left(V\left(Y|\boldsymbol{\theta}_{\sim i}\right)\right)}{V(Y)} = 1 - \frac{V\left(\mathbb{E}\left(Y|\boldsymbol{\theta}_{\sim i}\right)\right)}{V(Y)}. \tag{B.8}$$

This quantity measures how much of the variance in Y may be attributed to the variability of parameter θ_i. Parameters with small values of S_{T_i} do not contribute much to the output Y and may, therefore, be fixed at a single value.

Sensitivity indices in this work were computed following the work of Saltelli [101, 102, 104], in which $2N$ Monte Carlo samples of the parameter space are taken. These samples are used to generate two $N \times k$ matrices, $\mathbf{A}$ and $\mathbf{B}$, where each row is a sample point of the k-dimensional parameter space. That is,

$$\mathbf{A} = \begin{bmatrix} \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_i^{(1)} & \dots & \theta_k^{(1)} \\ \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_i^{(2)} & \dots & \theta_k^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_1^{(N)} & \theta_2^{(N)} & \dots & \theta_i^{(N)} & \dots & \theta_k^{(N)} \end{bmatrix}, \tag{B.9}$$

and

$$\mathbf{B} = \begin{bmatrix} \theta_1^{(N+1)} & \theta_2^{(N+1)} & \dots & \theta_i^{(N+1)} & \dots & \theta_k^{(N+1)} \\ \theta_1^{(N+2)} & \theta_2^{(N+2)} & \dots & \theta_i^{(N+2)} & \dots & \theta_k^{(N+2)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_1^{(2N)} & \theta_2^{(2N)} & \dots & \theta_i^{(2N)} & \dots & \theta_k^{(2N)} \end{bmatrix}, \tag{B.10}$$

where $\theta_i^{(j)}$ denotes the j-th Monte Carlo sample of the i-th component of the parameter vector $\boldsymbol{\theta}$. A third matrix $\mathbf{D}_i$ may be defined such that all of the columns come from the matrix $\mathbf{A}$ with the exception of the i-th column, which comes from $\mathbf{B}$,

$$\mathbf{D}_i = \begin{bmatrix} \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_i^{(N+1)} & \dots & \theta_k^{(1)} \\ \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_i^{(N+2)} & \dots & \theta_k^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_1^{(N)} & \theta_2^{(N)} & \dots & \theta_i^{(2N)} & \dots & \theta_k^{(N)} \end{bmatrix}. \tag{B.11}$$

The model output values are computed for each row of the three matrices, creating three $N \times 1$ vectors,

$$\mathbf{Y}_A = f(\mathbf{A}), \quad \mathbf{Y}_B = f(\mathbf{B}), \quad \mathbf{Y}_{D_i} = f(\mathbf{D}_i). \tag{B.12}$$

The total effect sensitivity index may then be approximated by

$$S_{T_i} = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} \mathbf{Y}_A^{(j)} \mathbf{Y}_{D_i}^{(j)} - \left(\frac{1}{N}\sum_{j=1}^{N} \mathbf{Y}_A^{(j)}\right)^2}{\frac{1}{N}\sum_{j=1}^{N} \mathbf{Y}_A^{(j)} \mathbf{Y}_A^{(j)} - \left(\frac{1}{N}\sum_{j=1}^{N} \mathbf{Y}_A^{(j)}\right)^2}. \tag{B.13}$$

This approximation method is used in all of the sensitivity calculations in this work.

Bibliography

[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] M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Oxford science publications. Clarendon press, 1987.

[3] Aristotle. *The Organon: The works of Aristotle on Logic*. Edited by Roger Bishop Jones. Translated by E. M. Edghill, A. J. Jenkinson, G. R. G. Mure, and W. A. Pickard-Cambridge. CreateSpace, Charleston, 2012.

[4] S. Arnborg and G. Sjödin. A note on foundations of Bayesianism. *Preprint: Nada, KTH*, 1999.

[5] S. Arnborg and G. Sjödin. Bayes rules in finite models. 2000.

[6] 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.

[7] 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.

[8] 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.

[9] T. Bayes. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions*, 53:370–418, 1763.

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

[11] S. N. Bernstein. *Theory of Probability*. Moscow, 1917.

[12] P. Biswas, R. Atta-Fynn, and D. Drabold. Reverse Monte Carlo modeling of amorphous silicon. *Physical Review B*, 69(19):195207, 2004.

[13] N. Bleistein and R. Handelsman. *Asymptotic Expansions of Integrals*. Dover Books on Mathematics Series. Dover Publications, 1975.

[14] M. Born, R. Blin-Stoyle, and J. Radcliffe. *Atomic Physics*. Dover books on physics and chemistry. Dover Publications, 1989.

[15] G. Box and N. Draper. *Empirical model-building and response surfaces*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1987.

[16] 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.

[17] T. V. Carey. Parsimony, in as few words as possible. *Philosophy Now*, 81:6–8, 2010.

[18] 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.

[19] 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.

[20] 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.

[21] R. Courant. *Differential and Integral Calculus*. Wiley Classics Library. Wiley, 2011.

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

[23] R. T. Cox. *The algebra of probable inference.* Johns Hopkins University Press, Baltimore, 1961.

[24] M. S. Daw and M. I. Baskes. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Physical Review B*, 29(12):6443, 1984.

[25] M. S. Daw, S. M. Foiles, and M. I. Baskes. The embedded-atom method: a review of theory and applications. *Materials Science Reports*, 9(7):251–310, 1993.

[26] L. G. Dunfield, A. W. Burgess, and H. A. Scheraga. Energy parameters in polypeptides. 8. empirical potential energy algorithm for the conformational analysis of large molecules. *The Journal of Physical Chemistry*, 82(24):2609–2616, 1978.

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

[28] M. J. Dupré and F. J. Tipler. New axioms for rigorous Bayesian probability. *Bayesian Analysis*, 4(3):599–606, 2009.

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

[30] K. C. Estacio-Hiroms and E. E. Prudencio. *Quantification of Uncertainty for Estimation, Simulation, and Optimization (QUESO), User's Manual.* Center

for Predictive Engineering and Computational Sciences (PECOS), Institute for Computational and Engineering Sciences (ICES), The University of Texas at Austin, Austin, TX 78712, USA, version 0.45.3 edition, October 2012.

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

[32] K. Farrell, J. T. Oden, and D. Faghihi. A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems. *Journal of Computational Physics*, 295:189–208, 2015.

[33] 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.

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

[35] D. A. Freedman. On the so-called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, 34:299–302, 2006.

[36] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Press, 2 edition, 2002.

[37] B. Gelin and M. Karplus. Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochemistry*, 18(7):1256–1268, 1979.

[38] K. D. Gibson and H. A. Scheraga. Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease s-peptide. *Proceedings of the National Academy of Sciences of the United States of America*, 58(2):420, 1967.

[39] J. Y. Halpern. A counterexample to theorems of Cox and Fine. *Journal of Artificial Intelligence Research*, 10:67–85, February 1999.

[40] J. Y. Halpern. Cox's theorem revisited. *Journal of Artificial Intelligence Research*, 11:429–435, December 1999.

[41] M. Hardy. Scaled Boolean algebras. *Advances in Applied Mathematics*, 29(2):243–292, 2002.

[42] W. G. Hoover. Canonical dynamics: equilibrium phase-space distributions. *Physical Review A*, 31(3):1695, 1985.

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

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

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[51] E. T. Jaynes. Information theory and statistical mechanics. II. *Physical review*, 108(2):171, 1957.

[52] E. T. Jaynes. Prior probabilities. *Systems Science and Cybernetics, IEEE Transactions on*, 4(3):227–241, 1968.

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

[54] 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.

[55] 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.

[56] D. Keen, M. Tucker, and M. Dove. Reverse Monte Carlo modelling of crystalline disorder. *Journal of Physics: Condensed Matter*, 17(5):S15, 2005.

[57] 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.

[58] 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.

[59] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, et al. Optimization by simmulated annealing. *science*, 220(4598):671–680, 1983.

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

[61] B. J. K. Kleijn. *Bayesian Asymptotics Under Misspecification*. PhD thesis, Free University Amsterdam, 2004.

[62] B. J. K. Kleijn and A. W. van der Vaart. The asymptotics of misspecified Bayesian statistics. In *Proceedings of the 24th European Meeting of Statisticians, (eds. T. Mikosch and M. Janzura)*,, 2002.

[63] B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, pages 837–877, 2006.

[64] B. J. K. Kleijn and A. W. van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

[65] A. N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea Publishing Co., 1950.

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

[67] M. Levitt and S. Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *Journal of molecular biology*, 46(2):269–279, 1969.

[68] 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.

[69] L. Lu, S. Izvekov, A. Das, H. C. Andersen, and G. A. Voth. Efficient, regularized, and scalable algorithms for multiscale coarse-graining. *Journal of Chemical Theory and Computation*, 6:954–965, 2010.

[70] 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.

[71] 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.

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

[73] 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.

[74] 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.

[75] D. McQuarrie. *Statistical Thermodynamics.* Harper's chemistry series. University Science Books, 1973.

[76] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[77] H. Meyer, O. Biermann, R. Faller, D. Reith, and F. Müller-Plathe. Coarse graining of nonbonded inter-particle potentials using automatic simplex optimization to fit structural properties. *The Journal of Chemical Physics*, 113(15):6264–6275, 2000.

[78] P. M. Morse. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Physical Review*, 34(1):57, 1929.

[79] 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.

[80] R. Nickl. Statistical theory, June 2012. Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge.

[81] 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.

[82] 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.

[83] K. Nordlund. Basics of Monte Carlo simulations. Lecture Notes, 2006.

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

[85] W. L. Oberkampf and C. J. Roy. *Verification and validation in scientific computing*. Cambridge University Press, 2010.

[86] W. L. Oberkampf, T. G. Trucano, and C. Hirsch. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*, 57(5):345–384, 2004.

[87] J. Oden. *An Introduction to Mathematical Modeling: A Course in Mechanics*. Wiley Series in Computational Mechanics. Wiley, 2012.

[88] J. T. Oden, K. Farrell, and D. Faghihi. Estimation of error in observables of coarse-grained models of atomic systems. *Advanced Modeling and Simulation in Engineering Sciences, To Appear*.

[89] 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.

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

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

[92] 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.

[93] 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.

[94] H. Owhadi, C. Scovel, and T. Sullivan. Bayesian brittleness: Why no Bayesian model is "good enough". *arXiv preprint arXiv:1304.6772*, 2013.

[95] C. Papadimitriou, J. L. Beck, and L. S. Katafygiotis. Asymptotic expansions for reliability and moments of uncertain systems. *Journal of Engineering Mechanics*, 123(12):1219–1229, 1997.

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

[97] R. Pozzo. *The Impact of Aristotelianism on Modern Philosophy*. Studies in philosophy and the history of philosophy. Catholic University of America Press, 2004.

[98] A. Rahman. Correlations in the motion of atoms in liquid argon. *Physical Review*, 136(2A):A405, 1964.

[99] D. Reith, H. Meyer, and F. Müller-Plathe. Mapping atomistic to coarse-grained polymer models using automatic simplex optimization to fit structural properties. *Macromolecules*, 34(7):2335–2345, 2001.

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

[101] A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280 – 297, 2002.

[102] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2):259 – 270, 2010.

[103] A. Saltelli, K. Chan, and E. Scott. *Sensitivity Analysis*. Number 2008 in Wiley paperback series. Wiley, 2009.

[104] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis: The Primer*. Wiley, 2008.

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

[106] M. S. Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *Journal of Chemical Physics*, 129(144):108, 2008.

[107] 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.

[108] W. Shinoda, R. Devane, and M. Klein. Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. *Molecular Simulation*, 33(1-2):27–36, 2007.

[109] W. Shinoda, R. DeVane, and M. L. Klein. Coarse-grained molecular modeling of non-ionic surfactant self-assembly. *Soft Matter*, 4(12):2454–2462, 2008.

[110] 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):6836–6849, 2010.

[111] I. M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 2:112–118, 1990.

[112] I. M. Sobol'. Sensitivity analysis for non-linear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414, 1993.

[113] K. S. Van Horn. Constructing a logic of plausible inference: A guide to Cox's Theorem. *International Journal of Approximate Reasoning*, 34(1):3–24, 2003.

[114] R. von Mises. *Wahrscheinlichkeitsrechnung*. Springer, 1931.

[115] 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.

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

[117] 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.

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

[119] D. Wu and D. A. Kofke. Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation. *The Journal of chemical physics*, 123(5):054103, 2005.

[120] 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.