HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences
have examined a dissertation entitled:

"Machine Learned Coarse Grained Force Fields for Dimensionality Reduction in Computational Materials Design"

presented by: Blake R. Duschatko

Signature _____
Typed name: Professor Boris Kozinsky

Signature _____
Typed name: Professor Efthimios Kaxiras

Signature _____
Typed name: Professor Michael P. Brenner

March 20, 2024

Machine Learned Coarse Grained Force Fields for Dimensionality Reduction in Computational Materials Design

A DISSERTATION PRESENTED
BY
BLAKE R. DUSCHATKO
TO
THE DEPARTMENT OF JOHN A. PAULSON SCHOOL OF ENGINEERING AND APPLIED SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
APPLIED PHYSICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MARCH 2024

Thesis advisor: Professor Boris Kozinsky Blake R. Duschatko

Machine Learned Coarse Grained Force Fields for Dimensionality Reduction in Computational Materials Design

Abstract

Computational materials science enables unique insight into atomistic processes that cannot be probed with current experimental apparatuses. For decades, the field has offered new understanding in a variety of material domains, ranging from polymers, proteins, and biophysics to battery materials, solvents, and many others. Of particular interest are molecular dynamics studies that can be used to analyze the kinetic and thermodynamic behavior of such systems.

Machine learning has shown to be a valuable tool in modeling the potential energy surface required for these methods. Despite significant progress in both the hardware, software, and theoretical capabilities of molecular dynamics approaches and machine learning, studying atomistic processes with long spatial and temporal scales at full atomistic resolution becomes prohibitive. To this end, coarse graining is a crucial alternative that enables the study of these systems with more fine tuned control than can be achieved in experimental setups, while also retaining a higher degree of spatial and temporal fidelity than would be possible experimentally.

In this dissertation, I will introduce a flexible Bayesian force field approach to design coarse grained free energy models. These novel methods enable an automated approach to the data collection process, while most importantly allowing for highly transferable models. I will further demonstrate how new approaches that utilize the integration of physics principles can provide more accurate and robust machine learning models for coarse graining applications. Finally, I will discuss ongoing and future developments, applications, and considerations for the future of computational materials exploration using these scalable and accurate methodological advancements.

Contents

Listing of figures

To my father, for without whom I would not be where I am today. For we can love completely, without complete understanding.

Acknowledgments

I WILL BE FOREVER GRATEFUL to all of the wonderful people who contributed to making my PhD journey possible. Thanks is due, first and foremost, to my advisor, Boris, for his kindness, insightful conversations, and support of my ideas. My first year as a graduate student came with many obstacles, and the impact of COVID-19 cannot be understated. Boris has always been a supporter of my growth, well-being, and development as a scientist. For this, I am extremely grateful.

To the Materials Intelligence Research group as a whole, I give thanks for the scientific discussions that have shaped my work and helped me learn more than I could without them. My sincerest thanks is extended to a particular number of individuals in the group; Jenny Coulter has shared in difficult experiences throughout our graduate careers. As a brilliant and kind person, she has provided support and knowledge on countless occasions. I would like to thank as well Kyle Bystrom and Cameron Owen, for their continuous lack of judgement and an endless effort to accept me. Although my time working with these individuals was short, I thank Zac Goodwin and Julia Yang, for giving new life to my work by taking a unique interest in it. Thanks is due to Nicola Molinari who has provided career guidance throughout my PhD work. His input has surely helped shape the way I have designed my future career. Finally, an alumnus of the group, I thank Steven Torissi. The way you think about science is an inspiration, and the way you communicate ideas to the greater community has been a large factor in my approach to teaching. Moreover, your kindness and support has helped in more ways than you know.

Within the greater Harvard community, I would like to thank Seth Avakian, for providing emotional guidance in times of difficulty throughout my PhD. While our paths did not cross as frequently as I may have liked, I hold dearly the memories of our conversations.

I would not be where I am today without the unwavering support of my family and friends. To all of you, I express my deepest gratitude. Thank you to my mother, Dottie, for never losing patience with my frequent absence and being my biggest fan at every stage. Thank you to my brother, Logan, for always making such a sincere effort to support my accomplishments. Lastly, thank you to my dad, Wayne, and my younger brothers, Cooper and Daniel, for always taking an interest in my work, and for their unconditional love. To my friends, I thank you for the opportunities to rest my mind in your company; Caitlin, Kyle, MacKenzie, Jordan, Geoff, Amelia, Sarah, Deirdre, Julian, Mary, Brynne, Abby, Bryce, Jasmine, Cory, Joe, Addy and Gabby. Particular thanks is due to Noah Bice, who has not only been an emotionally grounding best friend, but a brilliant scientific peer that I

have had the benefit of growing alongside.

Last but not least, I would like to thank my wonderful partner, Shelby. You are more deserving of thanks than I can put in a single paragraph, but I hope you can understand the depth of my love and gratitude for all you have done. Thank you for putting up with countless late nights, hours of my scientific rambling, and the highs and many lows of my degree. In moments of joy you have rejoiced with me, while being a guiding light in moments of pain. You have accepted me, supported and encouraged me at every stage of growth in this process, both academically and as a person. For this, I owe much of who I am to you, and my appreciation of you cannot be understated. To you, I promise to return the kindness, love and support by forever helping you pursue your goals and dreams, in the same way that you have always supported mine.

An expert is a person who has made all the mistakes that can be made in a very narrow field.

Niels Bohr

0 Introduction

Understanding the behavior of materials is of central importance to the development of new technologies. In order to tackle problems such as clean energy and the discovery of novel drugs, it is essential to have a deep understanding of a variety of properties, ranging from thermodynamic quantities such as solubility and phase transition curves, to kinetic quantities including diffusion and thermal conductivity.

While experimental methods are the final arbiter in definitively stating how materials behave, existing apparatuses are limited in time- and length-scales resolutions they can access. Moreover, many high resolution techniques require removing the system from its native state, such as in cryogenic electron microscopy[4]. In addition, rapidly screening large quantities of materials is impractical for experimental techniques.

On the other hand, computational tools allow for much finer control over the system state, enabling precise control of temperature and pressure conditions and removing undesired effects from external sources. With the comparative compactness and scalability of CPU and GPU computing architectures, these ideas can be leveraged to employ simulations in massively parallel settings. As a result, computational approaches are able to remove more unknowns in an experimental set up and more rapidly investigate the origins of material behavior.

From this lens, atomistic modeling in particular has enjoyed a rich history of innovation and discovery[5,6,7,1,8,9,10,11]. Specifically, atomistic models have long been used to study the phase behavior[12], structure[13,7,1,8,9,10,11], and a vast array of other properties of a wide range of materials. More importantly, atomistic modeling has complemented experimental efforts in a positive feedback loop, wherein computation has been used to better understand atomic processes that lead to observable behavior in experiments that are otherwise poorly understood[14,15,16]. In some exceptional cases, computational works have even been used to discover new materials altogether[17].

A central component that enables these studies lies in developing an understanding of the system's potential energy surface. The potential energy is ubiquitous in these approaches, as it can be used in molecular dynamics (MD) to integrate Newton's equations of motion. This gives direct access to the kinetic and thermodynamic behavior of a system, akin to approaches taken in experimental settings. Alternatively, sampling techniques such as Monte Carlo (MC) methods are able to use the potential energy to sample the thermodynamic Boltzmann distribution. As such, accurate modeling of this function and its derivatives (i.e. atomic forces) is a key focus of computational

materials science research.

While we often regard atoms as classical point particles, the origin of atomic forces is fundamentally quantum mechanical in nature. It is often assumed that nuclei move on the Born-Oppenheimer (BO) potential energy surface, wherein electronic degrees of freedom instantaneously relax to their ground state, setting up an external potential in which nuclei move in. The Hellman-Feynman theorem provides a direct relationship between the ground state of the electrons to the potential energy and atomic forces experienced by atoms, enabling *ab initio* MD studies to be performed. Although solving Schrodinger's equation at every step of an MD trajectory is possible, at least in an approximate way through density functional theory (DFT), these approaches are extremely expensive.

One alternative solution is to approximate molecular forces using classical potentials. Chemical intuition informs us that many bonded systems are quadratic in nature[6,5], for example, and simple functional forms can similarly be employed to model other two-, three- and four-body interactions of atoms[13]. While these interatomic potentials (IPs) are quite fast, they remain limited in accuracy compared to *ab initio* approaches[2,1,18,19]. More recently, advances in machine learning (ML) have aimed at modeling the BO surface more efficiently[1,2,8,9,20,11,10,21,22,23,12]. Leveraging DFT-based atomic force and energy data, machine learning trains models that are orders of magnitudes faster than DFT while being able to maintain a high degree of accuracy and interpolate within the space of data. While MD in this framework is slower than classical IPs, the gained accuracy is significant.

Despite tremendous success in these directions, limitations remain in the ability of atomistic models to probe important physical processes. Classical force fields are, in most realistic cases, limited to the study of systems on the order of millions of atoms at most, and at timescales of microseconds. Moreover, exceeding these limits simultaneously is particularly challenging. While exceptions to these limits have been demonstrated, in all practical terms it is infeasible to acquire the resources needed to achieve such simulations. Similarly, ML based *ab initio* simulations suffer from the same performance bounds.

In response to such limitations, coarse graining (CG) approaches aim to model systems at a reduced resolution that allows for computational efficiency while still providing fine control of variables that is absent in experimental set ups[24,25,26,27,28,29,30,31,32,33,34]. The fundamental idea underlying these approaches is that a) certain degrees of freedom are uninteresting to large length- and time-timescale motions, such as heavy atom - hydrogen bonds, and b) the timesteps required for numerically stable simulation are limited by these unimportant fast degrees of freedom. By removing this information, CG methods promise to allow simulation of entirely new length- and time- scales.

However, two competing approaches to coarse graining have brought with them important trade-offs. One idealogical principle suggests that, much like classical force fields, coarse grained potentials should be constructed with common molecular mechanic functional forms, with parameters tuned to reproduce experimental observations[24,31,30,29,35]. While promising in principle and extremely fast, these approaches limit the general transferability of potentials to new system classes and have no assurances about model behavior on properties outside of the training set. Alternatively, it is possible to take a "bottom up" approach, wherein the CG model is built from an underlying atomistic model. In this approach, the thermodynamic properties of the target system are entirely preserved[32,33,36,37]. Though this is a highly appealing aspect of such models, it comes at the expense of requiring far more complex functional forms. The growing success of ML in materials science, as well as significant advances in computational power, has recently motivated a new wave of work in developing better bottom up CG models.

In this thesis, I will start by providing the necessary background to understand ML in the context of atomistic modeling. I will then introduce the background, successes, and limitations of existing CG approaches in order to motivate the integration of ML techniques into CG modeling. Further, recent advances in the use of ML for CG modeling will be reviewed. The main body of work of my PhD is presented herein, starting from a novel Bayesian implementation of principled coarse grained potentials. I demonstrate how these new developments enable on-the-fly adaptation

of models across molecular systems while providing far greater efficiency to their all-atom counterparts. This work further illuminates issues that have begun to show up in the atomistic and CG communities regarding the insufficiency of aggregate force-error metrics for understanding model performance. Important new links between Bayesian uncertainty and model predictions will thus be presented.

Subsequently, we aim to address the inaccuracies of many CG models in structural property prediction with new physics-informed approaches to the design of coarse grained models. This work serves as the basis for a fundamental shift in the CG research community. Where models have traditionally been trained with force information alone, we propose the integration of a broad range of thermodynamic properties into model training that can further improve data efficiency while leading effortlessly to models that are more accurate over an array of different metrics. This in turn inspires a wide range of new problems to be addressed, such as the integration of energetic information for more accurate modeling of solubility and phase transformations at the CG level. I will end these discussions with important examples of applications and theoretical developments that remain active research, as well as an overview of where the field stands in relation to atomistic modeling.

A model must be wrong, in some respects, else it would be the thing itself. The trick is to see where it is right.

Henry Bent

1

Machine Learning and Molecular Modeling

1.1 Molecular Dynamics

While many computational approaches exist for studying material properties, the focus of the work in this thesis will rely primarily on the conclusions drawn from molecular dynamics (MD). The central purpose of MD is to solve Newton's equations of motion by integrating in time. However, it is important to rigorously link experimental observation with computational methods in order

for them to be of any meaningful significance. The link between simulation, statistical mechanics, and experiment will prove to be a central theme in the design of coarse grained potentials in later chapters.

In a similar manner that physicists connect experimental approaches to statistical mechanics, so too is the aim of molecular dynamics approaches. The fundamental assumption in both is the idea of ergodicity. Specifically, statistical mechanics proposes that macroscopic material properties can be understood as the statistical behavior of microscopic atomic states. Suppose that a system of n atoms, with Cartesian coordinates given by r^n and momenta p^n has interactions governed by the Hamiltonian

$$H(r^n, p^n) = \sum_i^n \frac{p_i^2}{2m_i} + U(r^n) \tag{1.1}$$

Here, m_i is the mass of each atom and U the potential energy. A given microstate of the system, defined by a specific point in the phase space of r^n and p^n, is said to occur with probability density equal to the Boltzmann weight of the state, i.e.

$$\mathcal{P}(r^n, p^n) = \frac{\exp(-\beta(\sum_i^n \frac{p_i^2}{2m_i} + U(r^n, p^n)))}{Z} \tag{1.2}$$

where Z is the normalizing partition function, $\beta = 1/kT$ and k is the Boltzmann constant. Given a function of this phase space, $\mathcal{A}(r^n, p^n)$, one can integrate over the density to compute the expected value

$$\bar{\mathcal{A}} = \int dr^n \, dp^n \mathcal{A}(r^n, p^n)\mathcal{P}(r^n, p^n) \tag{1.3}$$

In practice, what experiments actually measure is a quantity averaged over time

$$\bar{\mathcal{A}}_\tau = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau dt \, \mathcal{A}(t) \tag{1.4}$$

The ergodic hypothesis assumes that all available microstates of a system will be explored over a sufficiently long period of time, and therefore a time average and an ensemble average are equivalent. Mathematically, this enables the comparison of predictions of statistical mechanics, experiments and MD by assuming the equality

$$\bar{\mathcal{A}} = \bar{\mathcal{A}}_\tau \tag{1.5}$$

We note the ergodic hypothesis is not in general true. Nonetheless, MD work presumes that the ergodic hypothesis is true in practice. In this way, rather than as an investigator of explicit dynamics, MD can be seen in a more abstract sense as a sampler of the Boltzmann distribution. This will be a crucial point in CG modeling.

1.2 Interatomic Potentials

It is evident in the presentation of MD that a central quantity is the potential energy. In order to understand the behavior of materials, it is essential to understand how the constituent particles interact. Should we have a potential energy surface, U, that is a function of the Cartesian coordinates of n nuclei, $\vec{r}^n$, Newton's equations of motion tell us how such a system evolves in time:

$$m_i a_i = -\frac{\partial U(r^n)}{\partial r_i} \tag{1.6}$$

Clearly, our understanding of materials systems hinges on the accuracy of the potential energy function.

At the most fundamental level, the interactions between electrons and nuclei are quantum mechanical in nature. As such, it is normal to first approach the problem by solving the Schrodinger equation for the electronic Hamiltonian. If n electrons have coordinates r^n and N nuclei have coor-

dinates R^N, then the Hamiltonian acting on the electrons only is given by

$$H = -\sum_i^n \frac{\hbar^2}{2m_i}\nabla_i^2 - \sum_i^n \sum_j^N \frac{e^2 Z_j}{|r_i - R_j|} + \sum_{i \neq j} \frac{e^2}{|r_i - r_j|} \tag{1.7}$$

where Z_j is the nuclear charge. Hidden in this formalism is the assumption that nuclei move far slower than the electrons themselves, such that the wavefunction of electrons and nuclei can be factored as $\psi(r^n, R^N) = \psi_e(r^n; R^N)\psi_N(R^N)$. In other words, the nuclear positions appear parametrically in the motion of the electrons, while the nuclear wavefunction evolves in response to the relaxed ground state of the electrons. From this approximation, known as the Born-Oppenheimer approximation, the electrons can be assumed to reach their ground state instantaneously, and their explicit treatment can be removed by allowing the nuclei to move on the ground state energy surface of the electrons parameterized by the nuclear coordinates, $U(R^N)$.

The goal in these methods is to use computational techniques such as coupled cluster theory or DFT to solve for the electronic ground state, $\psi_0(r^n; R^N)$. From this ground state, the Schrodinger equation provides the ground state energy

$$U(R^N) = \langle \psi_0(r^n; R^N) | H | \psi_0(r^n; R^N) \rangle \tag{1.8}$$

It's no surprise that this approach is extremely computationally demanding. Even the more efficient of quantum methods, DFT, scale roughly as the cube of the number of electrons in the system, effectively limiting calculations to hundreds of atoms. In addition to poor scaling, the long computational times in ground state calculations mean that MD simulations of atomic motion, based on *ab initio* calculations, are often limited to picosecond to nanosecond timescales.

Prior to advances in machine learning, the standard approach to improving simulation efficiency was to instead rely on empirical observation and chemical intuition. In particular, one may imag-

ine that molecules involve harmonic bond potentials, harmonic bond angle potentials, dihedral interactions describing bond rotations, as well as intermolecular effects such as van der Waals and electrostatic interactions. A common molecular mechanics function may take the form

$$\begin{aligned} U(R^N) = &\sum_{\text{bonds}} k^b (R_b - R_b^0)^2 + \sum_{\text{angles}} k_i^a (\theta_a - \theta_a^0)^2 \\ &+ \sum_{\text{dihedrals}} \sum_{n=1}^{4} \frac{V_n^d}{2} (1 + (-1)^{n+1} \cos(n\varphi - \varphi_n^0)) \\ &+ \sum_{i>j} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{Z_i Z_j e^2}{4\pi\varepsilon_0 R_{ij}} \right) \end{aligned} \tag{1.9}$$

In this equation, we provide parameters for the bond strengths, lengths, angles, as well as the dihedral, or torsional, parameters for bond rotations. Additionally, we provide the nuclear charges for the Coulomb interaction as well as coefficients for the so-called Lennard Jones interaction, which captures long range dipole interaction effects as well as short-ranged repulsive effects. All parameters can be tuned so that the systems of interest reproduce experimental observables, such as hydration free energies, densities, and intramolecular structures through the MD methods presented previously. This has seen great success in many available force fields.

1.3 Machine Learning for Atomistic Models

Despite the success of classical approaches to developing interatomic potentials, qualitative agreement with experiment is often the best that can be hoped for. Clearly it is desirable to make more accurate predictions with computation. In general, the functional form of the *ab-initio* based potential presented above is unknown and highly complex. Inspired by the success of machine learning in other problem domains, the early 2000's began to see crucial developments in the application of ML to materials science. Here, the goal is to model the BO potential energy surface of nuclei and its

corresponding spatial gradients for subsequent use in MD studies.

The early work of Behler and Paranillo set the stage for these advances [20]. From this launch point, rapid innovation has followed. Most recently, state of the art Gaussian process (GP) methods [10,1,2] and neural network (NN) approaches [9,8,11] have enabled tremendous studies at massive length scales with high accuracy [23,12]. The training of these models is enabled by the Hellman-Feynamn theorem, giving access to nuclear forces in the ground state electron configuration. These forces, $\vec{F}^n$, and energies, E, are often compiled into a loss function such that the parameters of an NN can be minimized to produce a sufficient approximation to the underlying potential. Specifically, given a set of n_t frames of *ab initio* data, for an ML model of the potential energy surface, $\tilde{E}$, as a function of atomistic coordinates, $\vec{r}^n$, the functional

$$L(r^n, \theta) = \sum_t^{n_t} \left(\alpha_E \left| E_{n_t} - \tilde{E}_{n_t} \right|^2 + \alpha_F \sum_i^n \left| F_{i,t} + \frac{\partial \tilde{E}}{\partial r_{i,t}} \right|^2 \right) \tag{1.10}$$

can be minimized with respect to model parameters, θ. Here, the sum over t and i represent sums over system frames in a DFT calculation and all atoms within that frame, respectively. The alpha coefficients are hyper parameters that can be tuned. The global minimum is that model for which the energy and forces of the training set are well reproduced. Based on this fundamental idea, different NN architectures give rise to both new advances and new problems that are active areas of on going research.

Alternatively, in a GP framework, the output of a model is related to force and energy labels via a weighting scheme that depends on a kernel function. This kernel function describes the similarity between two inputs, with the assumption being that test inputs that are close to training inputs should have outputs that are weighted more heavily to those of that training output. The mathematics of Gaussian processes will be explored more thoroughly in subsequent chapters.

Despite the tremendous success of MLFF's in material modeling, important physical domains re-

main relatively inaccessible. Coarse graining techniques have a long history in addressing these short comings in the context of accelerating classical potentials, but only recently has ML been integrated into these approaches to target the ultimate goal - a low resolution model with the accuracy of DFT. In the following chapter, I will provide an overview of coarse graining techniques in the context of classical potentials, and describe the ways in which coarse graining challenges necessitate a new ML perspective.

Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics.

David Goodstein

2

The History and Theory of Coarse Graining

The contents of this chapter closely follow the work:

B. R. Duschatko *et al*, *npj Comput. Mater.* **10** (1), 9 (2024).

Direct quotes from the text are used where appropriate, and additional detail is added for completeness.

2.1 Limitations of Atomistic Models

There are substantial limitations on the applicability of MD to situations where the behavior at long length and time scales is relevant, such as many biological processes. For example, even the fastest of classical interatomic potentials, those of Lennard-Jones (LJ) fluids, have severe limitations in these regimes. On CPU hardware, we can get an approximate sense of the existing bounds on these methods. In Fig. 2.1, using a single CPU with with a 10 A cutoff and 1 kg/m^3 density, and assuming each particle is the mass of carbon, we examine the expected wall-time for a variety of system sizes across a range of desired simulation times. Note that these times are approximations, in that different processors and different processor configurations will give variations in performance. Moreover, advances in GPU computing can significantly accelerate these values. However, even in these cases, it is clear that biological scales remain out of reach.

The underlying reasons for these deficiencies are two fold. Fast degrees of freedom in the system prohibit the use of larger integration time steps while creating a more rugged energy landscape that typically slows important structural changes[24,34]. Exploring a rough energy landscape with small time steps becomes unfavorable when coarser resolution is of interest. In addition, computing forces and updating the configuration state of every degree of freedom (DOF) at each time step in an all-atom (AA) system can be a computational burden. For a variety of problems, fast motions such as those of hydrogen vibrations do not play a significant role in long length- and time-scale properties, making it unnecessary to track each DOF.

In the context of biological systems, understanding protein folding pathways and the bulk prop-

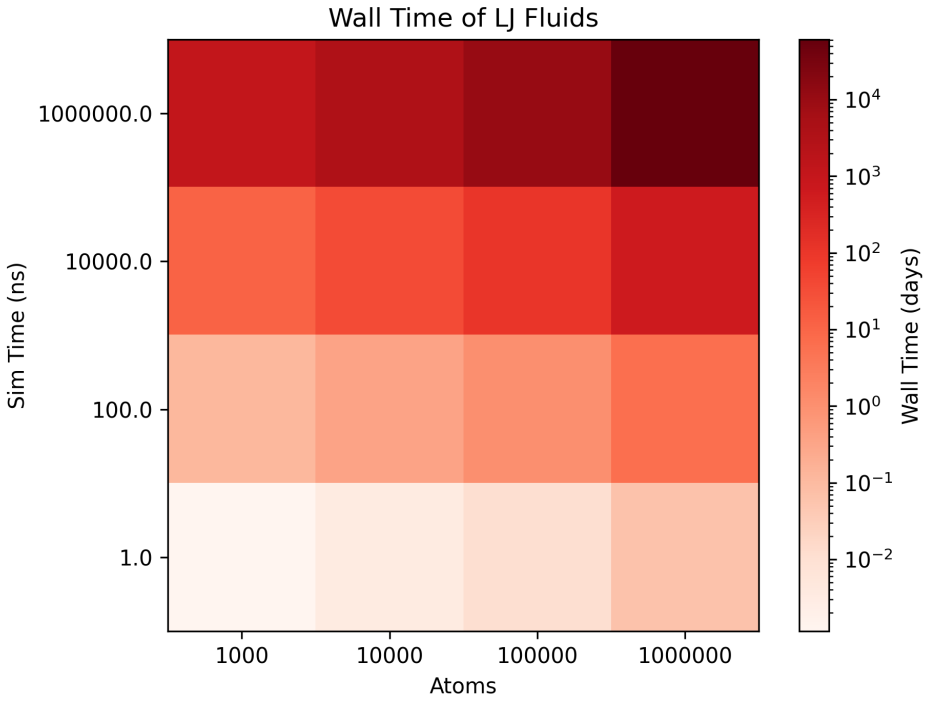

Figure 2.1: Wall Time of Lennard Jones Fluids. The simulation of Lennard-Jones liquids on single CPU's for different amounts of time and number of atoms is shown. Values are computed by running short simulations of the LJ system with different numbers of atoms and extrapolating the performance to long time scales.

erties of lipid bilayer membranes often require simulations of physical timescales on the order of microseconds or longer, while also typically modeling tens or hundreds of thousands of atoms[7,38]. In the most cutting-edge applications of MD, on the other hand, it is not typically feasible to exceed these two limits simultaneously. Similarly, at the all-atom scale protein-protein interactions, or more generally polymer-polymer interactions, may not depend on all degrees of freedom. Accounting for the effects of solvents adds additional computational complexity. In such scenarios as described above, a variety of coarse grained (CG) techniques are often used to probe the system at longer time scales and lower spatial resolution [39,40,41,42,34,43,29,24,28,25,26,27,31,30,44,29,35,45,46,47,48,49,50].

Of central importance to modeling the thermodynamics of systems with CG approaches is identifying the reduced degrees of freedom, or CG beads, and determining the interactions between them. To this end, two primary approaches are taken. Top-down methods are parametric models tuned to reproduce experimental observations, while bottom-up methods are built upon an underlying all-atom description of the system. Top-down coarse grained force fields are well established with widespread use across different applications [31,24,28,26,27,25,29,30,35,45,34], while the development of bottom-up coarse grained models has seen substantial activity within the past decade or so [32,51,52,53,33,54,47,55,56,57,58,59,60,21,61]. Although top-down methods are appealing for many reasons including thermodynamic transferability and speed, they fail in many regards to be more than simply qualitative in nature. On the other hand, despite acquiring explicit dependence on thermodynamic state points, bottom-up approaches can rigorously reproduce the statistical behavior of the system of interest by targeting the many-body coordinate dependent free energy, often called the potential of mean force (PMF), represented by a reduced set of coordinates [32,33]. In the following, we will compare and contrast these two approaches, weighing their relative pros and cons.

2.2 Top Down Coarse Graining

One approach to reducing the resolution of atomistic models is to design a model such that certain experimentally observed properties of a set of systems are well reproduced in a MD simulation. Such approaches are commonly referred to as "top down" coarse graining. In practice, this is not so different to the techniques employed for the design of classical atomistic force fields. However, the design space in CG models is far more expansive. While atomistic systems have well defined interacting point particles, the choice of *which* degrees of freedom to ignore is non-trivial. The resulting CG particles, or beads, should reflect in some sense the behavior of their constituent parts, and as such, this choice has a significant impact on the optimal parameters of the corresponding force field.

Regardless, such approaches have seen widespread use and success[24,25,26,27,28,35,31,30]. The functional form of these new potentials is most commonly in resemblance to the typical molecular model given by equation 1.9. As a result, the CG models exhibit far greater speed that has enabled the study of a broader class of biological systems at increased length- and time-scales[34].

While highly efficient, top-down CG approaches are far from perfect. The simple analytical form of the potential limits the achievable accuracy of the models, while targeting a subset of experimental properties prohibits confidence that a model will capture behavior outside of the training domain. Further, due to the increased design space of CG particle types, transferability across systems is limited.

2.3 Bottom Up Coarse Graining

Where top down methods fail to capture true material behavior with high confidence, bottom up CG approaches aim to resolve this issue. We recall that MD can be viewed as a sampler the Boltzmann distribution. From this perspective, one might imagine constructing a potential such that the resulting, low dimensional MD simulation appropriately samples the same distribution. In fact, this

is precisely the approach of bottom-up models, whose name is acquired from the fact that they are built in reference to a more fundamental atomistic system description. We review the mathematical details, originally introduced in Ref. [32], below.

Suppose we have an all-atom model consisting of n atoms at a set of positions $r^n = \{r_1, r_2, \ldots, r_n\}$ and governed by a potential $u(r^n)$. We consider a mapping of the AA system to N coarse grained sites, given by $R^N = \{R_1, R_2, \ldots, R_N\}$, of the form

$$R_j = M_j(r^n) = \sum_{i=1}^{n} c_{ij} r_i \tag{2.1}$$

where j denotes the index of the coarse grained site and i the index of each atom. The mapping coefficients c_{ij} must satisfy $\sum_i c_{ij} = 1$ to ensure translation invariance of the resulting model [32].

The above information is enough to define the Boltzmann probability distribution of the AA system within the canonical ensemble, so that

$$p(r^n) = \frac{\exp(-u(r^n)/k_B T)}{Z_r} \tag{2.2}$$

with Boltzmann constant k_B, temperature T, and partition function Z_r. Let the coarse grained system be governed by its own potential, $U(R^N)$, so that we can define a Boltzmann probability for the coarse grained sites at the same state points as

$$P(R^N) = \frac{\exp(-U(R^N)/k_B T)}{Z_R} \tag{2.3}$$

For a consistent coarse grained model, we would like the probability of sampling a given coarse grained configuration to be the same as if we had sampled the sites from the AA system. In other

words,

$$P(R^N) = \int p(r^n) \prod_{j=1}^{N} \delta(M_j(r^n) - R_j) dr^n \tag{2.4}$$

This requirement defines the potential of mean force (PMF), $U(R^N)$, as a coordinate-dependent free energy surface in terms of an underlying AA model. In particular,

$$\exp(-U(R^N)/k_B T) \propto \int p(r^n) \prod_{j=1}^{N} \delta(M_j(r^n) - R_j) dr^n \tag{2.5}$$

The PMF allows us to define a force on each coarse grained site as a gradient of the free energy. This force captures not only energetic effects, but also the entropic contributions to the free energy arising from the degrees of freedom being integrated out.

Formally, for mappings that take a group of atoms to their center of mass, and for which no atom belongs to more than one coarse grained site, the mean force on each site is [32]

$$F_j(R^N) = \left\langle \sum_{i \in S_j} f_i \right\rangle_{R^N} \tag{2.6}$$

where S_j is the set of atoms involved in the mapping to site j, f_i is the atomistic force on atom i, and the average is a weighted ensemble average defined as

$$\langle g(r^n) \rangle_{R^N} = \frac{\int g(r^n) e^{-u(r^n)/k_B T} \prod_{j=1}^{N} \delta(M_j(r^n) - R_j) dr^n}{\int e^{-u(r^n)/k_B T} \prod_{j=1}^{N} \delta(M_j(r^n) - R_j) dr^n} \tag{2.7}$$

for an arbitrary function g of the atomistic coordinates. Note that it is not a requirement that every atom have a non-zero weight in at least one mapping to a coarse grained site.

Inside every non-Bayesian there is a Bayesian struggling

to get out.

Dennis Lindley

3

Gaussian Processes for Coarse Graining

The contents of this chapter closely follow the work:

B. R. Duschatko *et al*, *npj Comput. Mater.* **10** (1), 9 (2024).

Direct quotes from the text are used where appropriate, and additional detail is added for completeness.

The inherent complexity of the PMF has limited, in many cases, the utility of non-ML approaches. In these settings, once again simple parametric equations are often employed. Even more flexible spline-based models have proven to be of limited accuracy, showing difficulties in capturing structural material properties for example. While still of great value, these models leave much to be desired.

Therefore, the wide array of problems that are faced in CG modeling motivate a new perspective. For instance, in exploratory applications, such as the discovery of structure of large proteins or rapid screening of soft materials, training CG models with a set of relevant configurations may become infeasible. Some configurations may never be sampled due to their rarity or the time it takes for a system to evolve into these states. For this reason, current approaches that do not use uncertainty metrics rely heavily on substantial *a priori* knowledge of the target behavior of the all-atom system, posing a limitation on their potential for wide spread applicability.

Moreover, it becomes increasingly difficult to assess the quality of models by estimating the true error on the test set, since this requires long constrained dynamics trajectories to obtain well-converged PMF data. In addition, traditional aggregate metrics such as mean absolute errors of forces may not capture subtle deficiencies in a model. Also, bottom-up CG models are highly dependent on the configuration and chemical make up of the all-atom system used for training, making the transferability of these CG models difficult to anticipate.

Due to these complications and need for greater accuracy, machine learned (ML) forces fields have recently gained traction over empirically parameterized classical potentials. In particular, the success of all-atom machine learned force fields as surrogate models for *ab initio* molecular dynamics [20,62,10,63,64,65,66,11,1,2,3,8,9,67,68,69] has resulted in increased interest in applying similar techniques to modeling the PMF [47,55,56,57,58,59,60,21,61]. These approaches typically make use of regression over long all-atom trajectories via a multiscale coarse graining / force matching technique that reproduces the all-atom PMF in the limit of sufficient sampling of the canonical ensemble [32]. In addition,

machine learning has targeted related problems of choosing an optimal low-resolution representation of atomic systems[59] and more broadly the problem of reconstructing atomic details from CG models[70,71,72]. So far, most ML approaches to CG models were based on neural networks (NN), which possess a number of benefits. For example, they serve as highly flexible representations of complex functions that can be trained on large training sets. Kernel-based methods have also been proposed[60,21,61]. However, all ML-based CG models of the PMF to date lack a measure of uncertainty through Bayesian techniques or even through neural network ensembles.

Compared to NNs, kernel-based Bayesian regression methods, such as Gaussian process (GP) regression, provide access to predictive uncertainty and have been demonstrated to efficiently select sufficiently representative training sets via active learning in all-atom settings[1,2,66]. One limitation is the increase of the computational cost of the training and inference tasks with the size of the training set. However, in many cases the full GP can be mapped onto an exact model for predicting both the mean and uncertainty with a cost that is independent of the training set size. Application of these recent methods have been demonstrated to simulations of complex heterogeneous systems at record speed and size, reaching 500 billion atoms [3,23,12]. In the context of coarse graining, active learning frameworks have seen less activity, and the existing literature on these methods is limited to non-bottom-up approaches[73].

In this chapter, we present an active learning regression framework based on principled Bayesian uncertainty inherent to sparse Gaussian processes (SGP's) for autonomous development of coarse grained force field models. First, we demonstrate the ability of Gaussian process regression to learn the coarse grained PMF on-the-fly, thereby reducing the guess work typically needed to construct the training set. In practice, this allows the model to discover unknown configurations that may appear at long timescales by bypassing the more predictable motions of faster degrees of freedom. We emphasize that this differs from current approaches that rely on all-atom simulations to sufficiently explore the full configuration space on their own before removing any degrees of freedom. Second,

we show that uncertainty-aware active learning enables the development of more transferable coarse grained models. In particular, we demonstrate how uncertainty, along with the locality, allows models of one molecular system to be transferred to a new system by updating the training database. In addition, we explore the implications of the design choice of how to label the species of CG beads on model accuracy and transferability, which has an important impact on the final speed of the model. Third, we find that uncertainty allows for a rapid and direct assessment of model robustness and limitations where traditional metrics such as force errors are insufficient, thereby accelerating deployment and facilitating automatic refinement.

3.1 Gaussian Process Regression

We shall first present, briefly, the mathematics of GPs. Suppose we have a model, f, that depends on some input, x. The function depends on some set of parameters, w. We suppose that the functional relationship between f and w is predetermined, for example it could be that

$$f(x; w) = w^T x \tag{3.1}$$

We observe, say experimentally, that the input x generates an output, y, and we wish to model this relationship with the function, $f(x; w)$. We begin by assuming that, given a specific set of parameters, our function prediction differs from the target value by a value ε which is Gaussian distributed according to $\mathcal{N}(0, \sigma)$. This may be written as

$$y = f(x; w) + \varepsilon \tag{3.2}$$

The epsilon parameter is called the noise, and it represents the fact that a) our data might be inherently noisy, and b) a given set of parameters to our function f may, for many reasons, not predict y

exactly. This could be due to insufficiently expressive basis sets, over fitting to other data, etc. This choice supposes that the functional form we have defined is somewhat likely to give the correct result given the right hyperparameters, i.e. the functional form is indeed capable of modeling what we are asking it to. If we choose the noise to be infinity, we are saying that we have absolutely no belief that our choice of functional form will predict y correctly or not. If we choose a very small value for the noise, or a different distribution altogether, we may be saying we expect our model choice should very closely match our observations.

In all of the following, ε therefore becomes a hyperparameter that we must decide how to set. If we pick too small of a value, then given a functional form f, we will tend to predict target values that might over fit to a data point. If we choose too large of a value, our model may not realistically reflect the process we want to model.

Note, however, that this implies a probability distribution. In particular,

$$p(y|w,x) \sim \mathcal{N}\big(y - f(x;w), \sigma\big) \tag{3.3}$$

represents the probability that our chosen function with parameters, w, given some input x, predicts the target value y. Said another way, this is the probability that the observed data is described by a model, f with parameters w. For a single data point, the parameters that give f centered around the value y are most likely. However, for a set of independent observations, $\{y_i\}$, the probability that a given set of parameters describes that particular set of observations is

$$p(\{y_i\}|w,\{x_i\}) \sim \prod_i^N \mathcal{N}(y_i - f(x_i;w),\sigma) \tag{3.4}$$

The primary quantity of interest, however, is not the above likelihood but rather, the posterior distribution. Specifically, we want to know what the probability is that a given set of data results in a

model with parameters, w. This is given by using Bayes' rule:

$$P(A|B)P(B) = P(B|A)P(A) \tag{3.5}$$

resulting in

$$p(w|\{y_i\},\{x_i\}) \sim p(w|x)\prod_i^N \mathcal{N}(y_i - f(x_i;w),0) \tag{3.6}$$

In this expression, $p(w)$ represents our belief on what the function should look like, prior to having seen any data. At this stage we can bake a lot of physical intuition into our models. We may let the distribution be

$$p(w|\{x_i\}) \sim \mathcal{N}(0,\Sigma(\{x_i\}) \tag{3.7}$$

where Σ is the covariance matrix between given inputs. We might expect, for example, that two similar inputs should give similar outputs, i.e. the function is not rapidly varying. The zero mean assumptions supposes we do not have any inclination what the predicted value should be before seeing data. However, there is no reason why we could not, for example, choose the mean prediction to be of the traditional molecular mechanics form.

In an application setting, we don't want to stop at knowing this distribution. What we would really like is to be able to predict the distribution of model values, y^*, given some new input, x^*. In order to do this, we could stop and maximize the likelihood function in equation 3.1. However, likelihood optimization can be severely prone to overfitting. Instead, we want to take a Bayesian approach and predict the posterior distribution:

$$p(y^*|x^*,\{y_i\}) = \int dw\, p(y^*|x^*,w)p(w|\{y_i\},\{x_i\}) \tag{3.8}$$

By doing so, we have access to not only predictions of the forces and energies of the CG units, but

additional access to principled uncertainty, which makes this approach appealing.

3.2 Sparse Gaussian Processes

Computation of the above expressions are computationally demanding and involve the inversion of $N \times N$ matrices, where N is the number of training data points. The inversion scales as N^3, and this is obviously problematic for large data sets.

It is more appropriate in this analysis to take a functional view of the original Gaussian process. If we have predicted a function value, f_i, for each input, x_i, and we wish to predict a function value, f^*, at a different input x^*, we know the function outputs are jointly distributed, per the definition of a Gaussian process. As such, we can specify a prior distribution on the expected form of these function values

$$p(\vec{f}, f^* | \vec{x}, x^*) \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} \Sigma(\vec{x}, \vec{x}) & \Sigma(\vec{x}, x^*) \\ \Sigma(x^*, \vec{x}) & \Sigma(x^*, x^*) \end{bmatrix}\right) \tag{3.9}$$

We can still specify the probability that a specific function value, f, produces an observation, y as before

$$p(y|f) \sim \mathcal{N}(y - f, \sigma) \tag{3.10}$$

We may again, then, condition the Gaussian process on the joint probability distribution of observation using Bayes' rule, giving

$$p(\vec{f}, f^* | \vec{y}, \vec{x}, x^*) = \frac{p(\vec{f}, f^* | \vec{x}, x^*) p(\vec{y} | \vec{f}, \vec{x})}{p(\vec{y})} \tag{3.11}$$

Note that this is a distribution of how likely it is that our data is explained by a function who's values are $\vec{f}$ at $\vec{x}$. We can compute the probability that the output should be f^* for an input x^* by marginalizing over all possible function outputs, $\vec{f}$, that could be produced by our data (the ana-

logue of integrating over possible function parameters)

$$p(f^*|\vec{y}) = \int p(\vec{f},f^*|\vec{y})d\vec{f} = \frac{1}{p(\vec{y})}\int p(\vec{y}|\vec{f})p(\vec{f},f^*)d\vec{f} \tag{3.12}$$

Again, the scaling here is unfavorable. An approach to circumvent this is to intrudce pseudo-inputs, or inducing points.

The argument for how to reduce the complexity of this problem was introduced in Ref.[74], which we review here. In addition to the training inputs and test inputs, suppose we additionally have a set of pseudo-input locations, $\vec{x}_u$. These pseudo-inputs have corresponding function values, $\vec{f}_u$. The joint prior previously written for the training and test values can be written in terms of the marginalization of the pseudo-inputs that are jointly distributed with all others, namely

$$p(\vec{f},f^*) = \int p(\vec{f},f^*,\vec{f}_u)d\vec{f}_u = \int p(\vec{f},f^*|\vec{f}_u)p(\vec{f}_u)d\vec{f}_u \tag{3.13}$$

The inducing outputs are normally distributed with zero mean. The covariance function specified in the prior before remains the same for the inducing points here. The key to sparse Gaussian processes is assuming that, rather than being fully jointly distributed, the training and test outputs are only jointly distributed conditioned on the inducing points

$$p(\vec{f},f^*) \approx \int q(\vec{f}|\vec{f}_u)q(f^*|\vec{f}_u)p(\vec{f}_u)d\vec{f}_u \tag{3.14}$$

The central choice in this framework is the choice of the distribution q, which leads to many different approximations with different pros and cons. However, the acceleration afforded by these approximations is profound and of great use.

3.3 Descriptors of Local Environments

We assume that the PMF can be modeled with a purely local function. In particular, given a set of coarse grained coordinates $R^N = \{R_1, R_2, \ldots, R_N\}$ with chemical-type identities $\{s_1, s_2, \ldots, s_N\}$, the PMF is given by a sum of local free energy contributions from the CG sites in the form

$$W(R_1, R_2, \ldots, R_N; s_1, s_2, \ldots, s_N) = \sum_i^N w(s_i, \rho_i) \tag{3.15}$$

where ρ_i is some description of the local environment of CG site i. In particular, the environment is the set of distance vectors between neighboring sites, expressed as a local spatial density, such that $j \neq i$, within a cutoff radius of $R_{\text{cut}}^{(s_i,s_j)}$ that can, in general, be species dependent. Formally,

$$|\rho_i\rangle = \sum_{j \in \varepsilon_i} |R_{ij}\rangle \otimes |s_j\rangle \tag{3.16}$$

where

$$\varepsilon_i = \{(R_{ij}, s_j) | \, |R_{ij}| < R_{\text{cut}}^{s_i,s_j}\} \tag{3.17}$$

Following the Atomic Cluster Expansion approach introduced by Drautz[22], the local environment of a site can be projected onto single-particle basis functions

$$c_{isn\ell m} = \langle \varphi_{n\ell m} | \rho_i \rangle = \sum_{j \in \varepsilon_i} \delta_{s_i,s_j} \varphi_{n\ell m}(R_{ij}) \tag{3.18}$$

where we choose for φ a decomposition into radial and spherical harmonic components with cutoff function, $\tilde{c}$,

$$\varphi_{n\ell m}(R_{ij}) = \tilde{R}_n(|R_{ij}|) Y_\ell^m(\hat{R}_{ij}) \tilde{c}(|R_{ij}|, R_{\text{cut}}^{s_i,s_j}) \tag{3.19}$$

A rotationally invariant descriptor can be constructed by invoking the sum rule of spherical har-

monics, giving a descriptor

$$d_{is_1s_2n_1n_2\ell} = \sum_{m=-\ell}^{\ell} c_{is_1n_1\ell m} c_{is_2n_2\ell m} \tag{3.20}$$

This is shown schematically in Fig. 3.1.

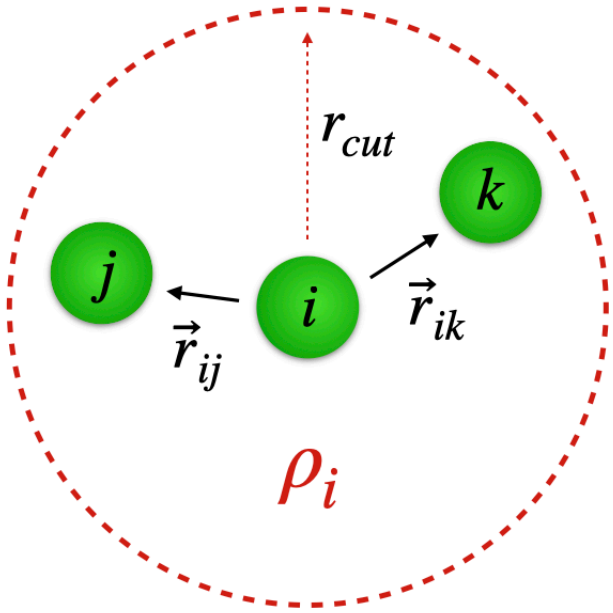

Figure 3.1: Schematic of the ACE descriptor. The local environment of CG site i is denoted ρ_i, and pairwise vectors to neighboring sites are given by $\vec{r}_{ij}$. The cutoff, shown as a red dotted line, has a radius of r_{cut}.

3.3.1 Sparse Gaussian Processes for Coarse Graining

We round out our discussion of SGP's by discussing their more specific application to CG problems. For a complete description of the system input, we must define our basis functions and hyperparameters. In this work we use Chebyshev polynomials for the radial basis and spherical harmonics for the angular basis in the descriptor above. The parameters n, ℓ, and r_{cut} are hyperparameters that we manually specify. To do so, we maximize the marginal log-likelihood (see Appendix A), and use the values $n = 12$, $\ell = 5$, and $r_{\text{cut}} = 4.5$ Å, independent of chemical type.

To complete the description of the sparse Gaussian process, we must define a kernel that compares the local environment descriptors. As in Ref. [2], we choose a kernel that resembles the smooth overlap of atomic potentials (SOAP) kernel [67,68,69] and takes the form

$$k(d_i, d_s) = \sigma^2 \left(\frac{d_1 \cdot d_2}{d_1 d_2} \right)^{\xi} \tag{3.21}$$

Here, the hyperparameter, σ, is optimized by maximizing the marginal log-likelihood during on-the-fly training, and ξ is a chosen parameter that can be used to increase the body-order of the model. Here, we take $\xi = 2$.

For a set of sparse coarse grained environments, S, that is a subset of a larger training set, F, a prediction of the local free energy on a new environment ρ_i can be cast as a sum over the sparse points:

$$\varepsilon(\rho_i) = \sum_{s \in S}^{N_s} k(d_i, d_s) \alpha_s \tag{3.22}$$

where

$$\alpha = (\sigma_n^{-1} K_{\mathrm{SF}} K_{\mathrm{FS}} + K_{\mathrm{SS}})^{-1} K_{\mathrm{SF}} y \tag{3.23}$$

where y is the vector of training force labels, K_{SF} the matrix of kernel values between the sparse set and training set, and K_{SS} the matrix of kernel values between points in the sparse set alone. The noise hyperparemter, σ_n, quantifies the inherent uncertainty present in the training labels and is another hyperparameter that is tuned via maximizing the marginal log-likelihood.

Such mean predictions with the sparse Gaussian process also allow for posterior predictive distribution variances. For SGP's, computing the variance requires approximate methods, where we choose in this work to use the Deterministic Training Conditional (DTC) approximation [74]. As in Ref. [2], we use a further simplified form that is the predictive variance of a fictitious Gaussian process trained on local free energies of the sparse environments alone. The resulting uncertainty on local

free energies is

$$\tilde{V}(\varepsilon) = \frac{k_{\varepsilon\varepsilon} - k_{\varepsilon S} K_{SS}^{-1} k_{S\varepsilon}}{\sigma^2} \tag{3.24}$$

Lying between 0 and 1, this form gives us a unitless measure in defining uncertainty thresholds during on-the-fly training. We find a relative tolerance of 0.02 to perform well. This is found to be stable from our empirical observations and is consistent with similar values used in Ref. [2]

The resulting models, following the on-the-fly training trajectory, can be simplified such that the summation over sparse points in the predictive distribution can be computed once and used for all future predictions [2]. This simplification allows for efficient inference upon deployment of the SGPs and is expanded upon in Appendix A.

3.4 Uncertainty and Active Learning

Bottom-up training of coarse grained models is typically done by regressing the PMF derivatives to time averages of forces, the precise form of which is given by equations (2.5) and (2.6). A common approach is to utilize instantaneous forces in a long unconstrained molecular dynamics trajectory to minimize appropriate functionals that reproduce the PMF [32,55]. This approach requires care in dealing with two implicit timescales in the problem. For one, the simulation times of the fast degrees of freedom must be long enough to ensure that the sampled all-atom configurations are not highly correlated. Also, collecting sufficient training data requires simulating long enough time scales in order to visit a full range of coarse grained configurations. Even so, depending on the shape of the PMF, some regions in CG configuration space separated by barriers may not be visited during the training simulation. Principled quantitative uncertainty provides a rigorous way of identifying configurations that lie within and outside the training set.

Here, we introduce an on-the-fly CG workflow implemented in the FLARE framework [1,2], depicted in Fig. 3.2, that utilizes the predictive Bayesian uncertainty of SGPs. The goal of this FLARE-

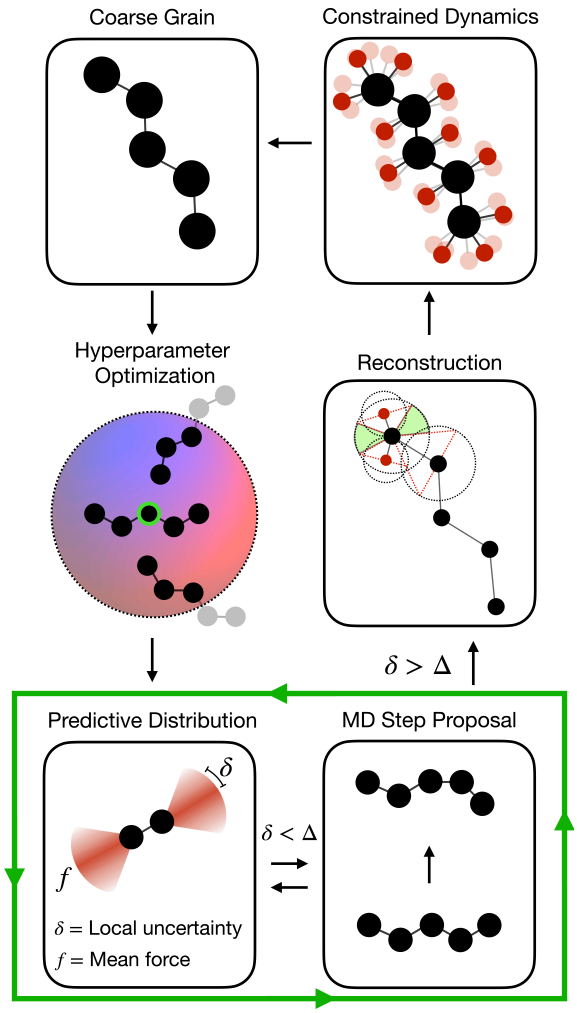

Figure 3.2: A Schematic Representation of the Active Learning Workflow. FLARE-CG is an extension of the Fast Learning of Atomic Rare Events software [1,2,3]. Here we schematically demonstrate the workflow of the on-the-fly active learning training loop. An initial all-atom frame is run under constrained dynamics and coarse grained to obtain initial force labels. A select number of sparse environments are randomly added to the training set of the Gaussian process. The construction of model descriptors is performed, and the hyperparameters of the SGP updated by maximizing the log marginal likelihood. The model proposes a molecular dynamics step, along with force and local free energy uncertainties. If all local free energy uncertainties are below a tolerance threshold, the step is accepted. Otherwise, reconstruction is performed in order to collect more constrained dynamics training data. To perform reconstruction, we construct excluded volumes around already placed atomic centers in the system. New atom placements are proposed by randomly drawing an azimuthal and polar angle pair, which are subsequently accepted if they do not lie within regions of overlap between existing excluded volumes.

CG approach is to automate the collection of the PMF labels by deriving a decision threshold of data acquisition from the uncertainty associated with every prediction during a CG MD simulation. This is implemented by directly comparing each local CG environment for which predictions are made with those in the training set using a pre-defined kernel function operating on geometric descriptors of local configuration environments. Specifically, we consider the local environment of CG sites within a defined cutoff radius from the central site. Many-body symmetry preserving descriptors based on the atomic cluster expansion (ACE) are constructed for this purpose, with different hyperparameters allowing for different levels of expressiveness [2,2,2].

The workflow begins from an initial coarse grained structural configuration, for which a constrained dynamics trajectory is performed to acquire force labels for training. Note that the acquired forces are the gradients of, and therefore directly related to, the PMF that we are trying to learn (see Methods A). Subsequently, for each configuration, the predicted models local PMF uncertainties are used to decide whether to evolve the system forward in time. If the uncertainty of the model on the new configuration is above a user-defined threshold, more constrained dynamics data is collected to augment the Gaussian process model database; otherwise, the CG MD step is accepted, and the system evolves forward by a time step. This process is repeated at every step of the CG MD trajectory, forming an autonomous active learning loop. Eventually, after the model no longer makes frequent calls to the all-atom baseline, the trained SGP model's explicit dependence on the training set size can be eliminated by mapping it onto an exactly equivalent and much faster parametric model [2].

During the active learning process, the model hyperparameters are adaptively optimized by maximizing the log-likelihood in response to new training data. By maintaining well calibrated uncertainties, models are then capable of discovering new configurations on their own, circumventing a major problem of potentially missing unknown structures in a model's training set. Further, this methodology allows us to rely on the all-atom models only to remove fast degrees of freedom and

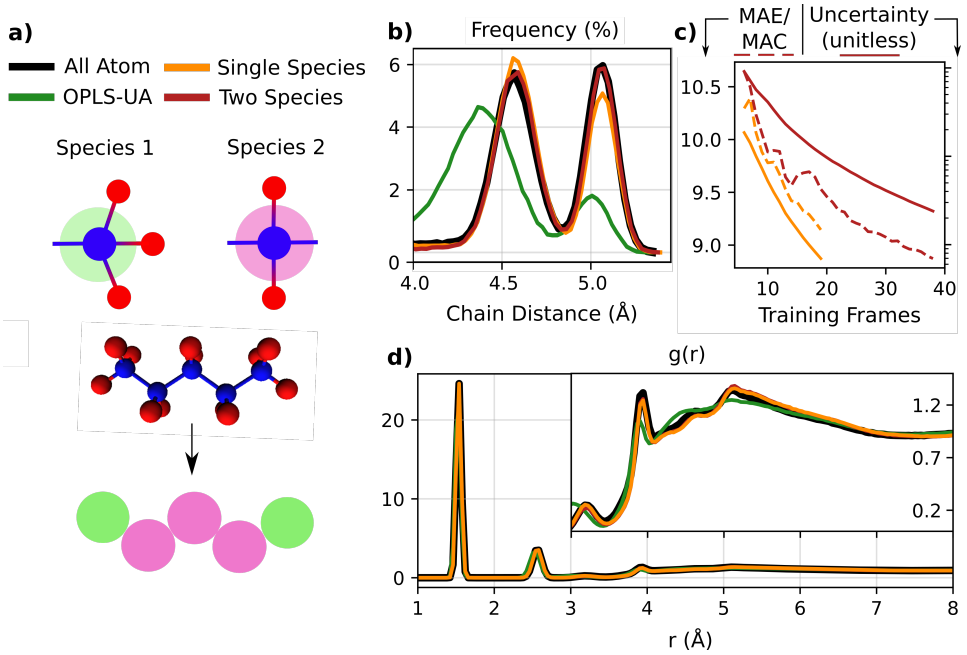

Figure 3.3: Demonstration of Active Learning to Coarse Grained Pentane. a) atoms in a pentane liquid are mapped directly to their carbon sites, integrating out hydrogen degrees of freedom. In two experiments, the beads are treated as either one or two different species, based on the underlying bond topology. b) the end-to-end chain distance distribution of single and two species models are compared to the all-atom training baseline, as well as a common coarse grained force field for hydrocarbon liquids, OPLS-UA. c) the learning rate for two and single species models is reported by showing the mean absolute force error on a test set as a relative percentage to the mean absolute force component in each test frame. We also report the mean free energy uncertainty of the model on the test set, which is a unitless quantity. d) the full carbon-carbon radial distribution function for single species, two species, all atom, and OPLS-UA simulations. The inset shows a zoomed-in view of the long-range structure.

obtain CG force training labels, so that the slow degrees of freedom are evolved directly using the learned low-resolution CG surrogate model. We illustrate the performance of SGPs trained on-the-fly for a pentane liquid structure consisting of 70 molecules, with the hydrogen degrees of freedom integrated out (Fig. 3.3).

The all-atom to CG mapping sequence is shown schematically in Fig. 3.3a. Our approach requires a prior definition of CG sites. Although recent works have considered automating the CG site selection process [59,72,70,75,76,71], more work is needed to integrate these approaches with our proposed framework. We instead rely on chemical intuition to choose the CG mapping, where for the

pentane system we choose to map the all-atom system to carbon sites. In this example, the hydrogen atoms are integrated out, effectively setting to zero their weight in the mapping function, defined in equation (2.1) of Methods.

Existing empirical and ML CG models often label coarse grained species by their underlying all-atom composition and bond topology, suggesting the end carbon atoms of pentane should be treated differently than those on the interior of the chain [55,24]. In particular, CG carbons can be treated as different species types in the descriptor equation (3.20) depending on whether they are at the end of the molecular chain. Alternatively, all carbons can be treated as the same species type, and we can rely on the ML model to correctly learn CG forces from the geometry of the local environment structure.

We investigate the impact of this choice by exploring the performance of models of pentane with and without explicit labeling of end carbons as a different species in the descriptor. Hydrocarbon systems such as pentane liquid are a convenient test case, as many top-down empirical models are parameterized for them. In the following, we compare the accuracy of our ML CG models, whose all atom baseline is the OPLS force field, to an empirical CG force field with the same CG site mapping, OPLS-UA [13,31]. Both the OPLS parameter set and its united atom variant were designed to capture the densities and heats of vaporization of organic liquids. While of the same molecular form, the OPLS-UA force field is further refined such that groups consisting of hydrogen atoms bound to a carbon atom are treated as a single interaction site.

For the ML CG approach, we find in Fig. 3.3c that the single-species and the two-species models provide similar force accuracy as well as the reproduction of structural properties compared to the all-atom baseline. Such comparisons are valuable because the less complex single-species models have shorter inference times. This arises from the scaling of the dimensions of ACE descriptors and kernel matrices with increasing number of species, associated with more computationally heavy linear algebra computations at inference. We are not claiming, however, that this finding for pentane

is a general result expected to hold for other molecular systems.

Similarly, comparing to each other the learning of on-the-fly models for both single- and two-species realizations in Fig. 3.3c, we find both types of models achieve comparable force accuracy. We note that the predictive uncertainties of each model differ between single- and two- species as a result of having inherently different model complexity. In practice this means that single-species models will tend to make fewer calls to the reference method (constrained all-atom dynamics). In this example, we use a predictive uncertainty value threshold, defined in equation (3.24) of Methods, of 0.02 for the two species case and 0.01 for the single species case to further highlight this contrast in behavior with respect to uncertainty. The final mean absolute force error, as a percentage of the mean absolute force component of the model predictions, on a test set lies around 9% for both single and two species models. For comparison, we also compute the error in forces predicted by the OPLS-UA force field on the same test set and find an error of 30%.

To emphasize that uncertainties are indeed predictive, in Fig. 3.3c we show the (unitless) mean local PMF uncertainty given by equation (3.24) of Methods in each test set frame. A frame refers to a structure snapshot together with atomic force information. As a function of training set size, the uncertainty correlates directly with the trends followed by the mean absolute force errors on the test set. Even though the quantitative values of uncertainty and force test-set error do not agree, the crucial point is that the same trend behavior persists.

Fig. 3.3b and Fig 3.3d demonstrate that the inter- and intra-molecular properties, respectively, of the CG carbon sites match the behavior of the all-atom carbon atoms with high fidelity. The full carbon-carbon radial distribution functions are well reproduced, as are the end-to-end molecule chain distance. This distance is defined as the linear separation between the two carbons on the ends of each molecule.

In addition to pair distributions, more complex structural correlations have been suggested for assessing the fidelity of coarse grained models [77,78,79,80,81]. For example, it is possible for CG models to

reproduce a set of distribution functions such as those shown, but the relative sampling between a set of states may still be inconsistent[78]. To further establish the strength of our method, we analyze the possible correlated states for pentane.

As in Fig. 3.4, pentane bond and angle distributions are unimodal, and therefore do not meaningfully contribute to understanding correlations between structural states. Instead, we consider the trans (T) and gauche (G) dihedral angle regimes, akin to Ref.[78]. With two dihedral angles, we label the structural state pairings of pentane as TT, GG, and TG. We show in Fig. 3.4 that both the single and two species CG models capture the relative sampling of these sates quite well compared to the all-atom baseline.

For completeness, we compare the accuracy of structural property predictions of our coarse grained models to an existing empirical (non-ML) model, OPLS-UA[31], in both Fig. 3.3 and Fig. 3.4. We find far greater fidelity in the reproduction of structural properties with our SGPs compared to this empirical model, as well as the relative sampling of dihedral angle pairs. In particular, we emphasize that the OPLS-UA model describes two, three, and four body interactions that are typical in classical force fields. Despite this, the simple functional form of such models is insufficient for capturing structural correlations. This further motivates the use of ML approaches to PMF modeling.

3.5 Structural Transferability

Bottom-up CG models are typically developed to preserve the partition function of the Boltzmann distribution maintained by thermostats in the all-atom MD simulation, thereby ensuring consistency with the AA thermodynamics. However, this implicitly assumes that the model will be specific to both the chemical make up of the all-atom system as well as the thermodynamic state points. Particularly in this setting, where models are specific to a given system, predictive model uncertainty is helpful in quantifying the distance of local CG configurations in the test set from those in the

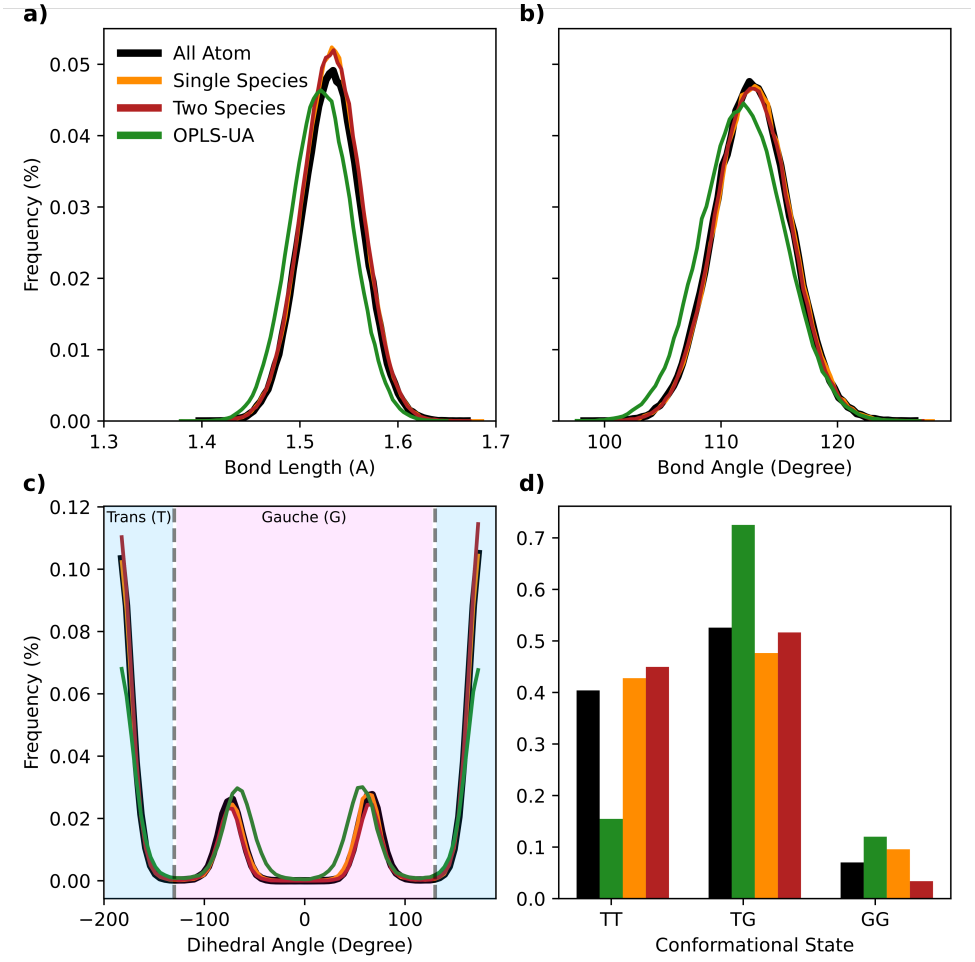

Figure 3.4: Structural Correlations Captured via Active Learning. a) the distribution of bond lengths in pentane molecules for single species, two species, and OPLS-UA models relative to the all-atom baseline. b) bond angle distributions of the same set of models. c) dihedral angle distributions of the same set of models. Vertical lines indicate the separation between trans and gauche dihedral conformations. d) the relative sampling of the three dihedral pair states for pentane, trans-trans (TT), trans-gauche (TG) and gauche-gauche (GG).

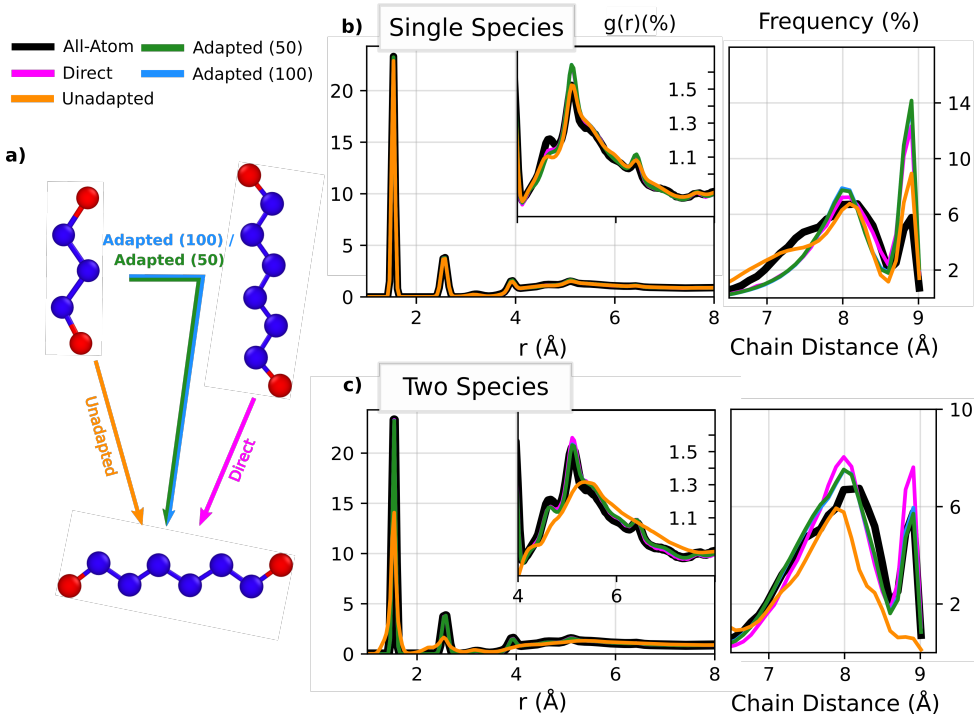

Figure 3.5: Transferable Coarse Graining Enabled by Active Learning. a) a schematic representation of the computational experiment considered. Adapted (50 and 100) pentane models are those that have seen extra octane data, generated on the fly for 50,000 and 100,000 steps, respectively. Unadapted models are pentane models deployed directly on octane with no additional training. b) the carbon-carbon radial distribution function and the end-to-end chain distance frequency are shown for single species models. The inset shows the long-range structure of the liquid c) the carbon-carbon radial distribution function and the end-to-end chain distance frequency are shown for two species models, with the inset showing the long-range structure.

training set. This enables us to systematically adjust and improve CG models.

To examine the degree to which SGP CG models trained on one system can transferred to another system, we consider several ways in which a model trained on the pentane liquid can be applied and adapted to an octane liquid, as summarized graphically on Fig. 3.5a. Direct CG models of octane are trained on all-atom octane data on-the-fly using active learning. Unadapted CG models are trained on-the-fly on the pentane system for 100,000 time steps and then deployed directly on octane. Here, the SGP evolves the system in CG space over time for 100,000 steps, checking the uncertainty against the user defined threshold at each step, before being deployed on octane. Note

that 100,000 does not refer to the number of AA constrained dynamics steps used when collecting more data. The models labeled as adapted (50) and adapted (100) have identical pentane training data, but their training sets are subsequently augmented on-the-fly for 50k and 100k time steps with octane training data, respectively, via the same procedure defined for the pentane on-the-fly loop. We also compare the performance of trained CG pentane models to the underlying AA results for octane. The goal of this experiment is to determine whether or not uncertainty can enable models previously trained on other systems to be extended to a new system more efficiently than starting over and without loss of accuracy. The results of each model type shown in Fig. 3.5 and Table 3.1 is the average of 40 independently trained models.

We find in Fig. 3.5c for the two species models that the inter- and intra-molecular structure of the carbon atoms is reproduced substantially better with the adapted models than their unadapted counterparts. This is further reflected in the end-to-end chain distance RMSE values given in Table 3.1. Note, however, that the unadapted single-species models shown in Fig. 3.5b give higher fidelity radial distribution functions compared to the unadapted two species models while using far less training data on pentane (see Table 3.1). This result can be attributed to the higher dimensionality of the two-species descriptors, which makes the two-species model more sensitive to changes in local environments. In the single-species case, for example, local environments near the end of the hydrocarbon chains, look highly similar to those of octane. For the two-species case, this is not true as the local environments, within the 4.5 Å cutoff, of octane carbon sites will almost never contain both ends of the chain, which is likely to happen in pentane. As a result, to a single species kernel, octane looks more similar to pentane and it is able to extrapolate with a limited amount of data.

To emphasize the utility of active learning, in particular in the context of designing transferable models, we note the stark contrast in accuracy between adapted models trained with and without using uncertainty. In Fig. 3.6, we examine difficulties that arise in modeling the end-to-end chain lengths of single and two species approaches. Here, the non-active learning based models are trained

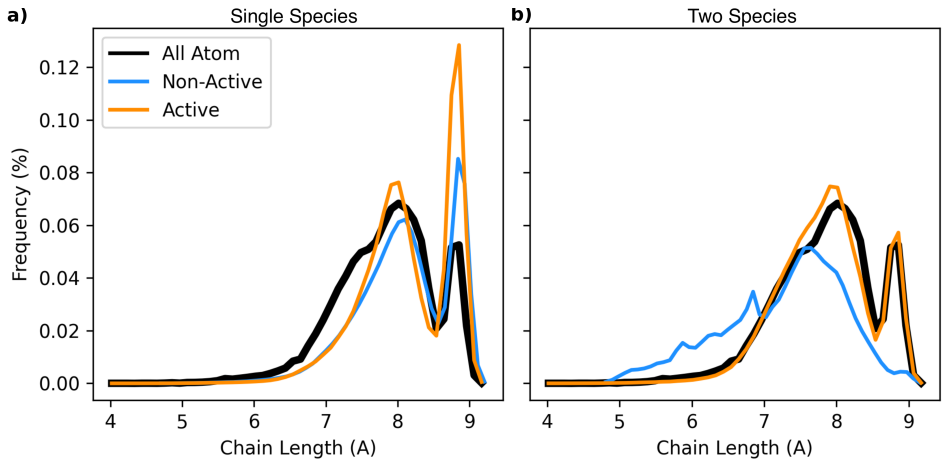

Figure 3.6: Comparison of Transferability With and Without Active Learning. We show the end-to-end chain length distribution for single and two species CG models trained with and without active learning. a) compared to the all-atom baseline, single species models are shown. Each CG curve is an average over the results of 40 models. The non-active learned models has seen one additional frame of octane data, while the active learning models see on average 1.7 frames of data. b) the same data is shown but for two species models. In this case, the non-active learned models each see 15 frames of octane data, while the active learning models see on average 15.2 additional frames.

	GP Model	Pentane Frames	Octane Frames	Force MAE/MAC (%)	Population Error (%)
Single Species	Unadapted Pentane	11.45	0	15.29	34.98
	Adapted (50) Pentane	11.45	1.77	10.10	53.71
	Adapted (100) Pentane	11.45	2.02	10.10	52.09
	Direct	0	12.1	10.18	49.72
Two Species	Unadapted Pentane	42.42	0	78.18	597.05
	Adapted (50) Pentane	42.42	15.26	10.14	9.84
	Adapted (100) Pentane	42.42	17.52	10.04	5.85
	Direct	0	54.73	10.20	13.86

Table 3.1: Accuracy and Efficiency of Transferable Coarse Grained Models. The average number of all-atom LAMMPS calls over a set of 40 models, as well as the average total number of training frames in the GP, is reported for both single- and two-species models. In addition, we report the mean absolute force error on a test set, averaged over 10 models, as a percentage of the mean absolute force component (MAC) in each test frame, as well as the population error of the averaged chain distance distribution from the all-atom ground truth, defined in equation 3.25

from randomly selected octane data, with a single frame being used for the single species models and 15 frames for the two species, in line with the amount of data collected by on-the-fly models reported in Table 3.1. Each curve is the average over an ensemble of 40 models.

A relatively small amount of octane data is added while training these adapted models, and it is therefore crucial that the added data be maximally informative. In this regard, two-species models are more susceptible to poor performance when the available data is not representative of the differences between the original (pentane) system and the new (octane) system. Being a less descriptive model, the single species examples do not suffer from this to the same degree. It is quite evident that uncertainty based active learning enables accurate transferability of coarse grained models where non-active learning could not. For completeness, we examine the learned distributions of active and non-active learning based models in the Supplementary Discussion, along with a detailed discussion on the structural correlations for this system.

To highlight the benefits of using an adapted model on systems outside of the training set as opposed to starting from scratch, we compare the computational cost of AA constrained dynamics reference computations as well as the resulting force errors and structural properties of adapted models compared to training an octane model from scratch (Table 3.1). Assuming the pentane data is already available, we find that by using an SGP containing data from a chemically similar system, far fewer AA reference calls to the new system are needed compared to starting from scratch. In both the single- and two-species cases, the force accuracy of adapted models with fewer active learning calls to the octane reference AA constrained dynamics method is higher than in direct models trained on octane from scratch which also requests more data in active learning. Additionally, over the course of the 50,000 and 100,000 on-the-fly training steps, we report in Table 3.1 the average number of frames in which the models request more octane data. We see explicitly that in all cases, models starting from pentane request fewer frames of data for the new system. This is a crucial point, as the reduction in required constrained dynamics calls of adapted models significantly

reduces the computational cost of model training. We observe that the single- and two-species models display comparable force errors despite their disagreement in the reproduction of chain distance distributions, which we explore further in the next section.

3.6 Force-Energy Correspondence

In this section we aim to examine the correlation between the force error and error in structural properties determined by the PMF. Specifically, we focus our attention on the end-to-end chain distance distribution. Physically, the multi-modal distributions that appear in pentane and octane systems arise from the PMF minima at two values of dihedral angles along the chain. For shorter chains, fewer pairs of dihedral angles exist that are energetically favorable, whereas longer chains effectively become more flexible to bending.

To quantify the variation between a model and the AA system in the end-to-end chain distance distribution, we define the population error as

$$100 \times \left| \frac{p_1^{cg}/p_2^{cg}}{p_1^{aa}/p_2^{aa}} - 1 \right| \tag{3.25}$$

where $p_{1,2}$ are the populations (i.e. frequencies of sampling end-to-end chain distance values) of the first and second PMF basins of the AA or CG system.

With this definition, we can motivate this discussion by considering the relationship between mean force errors and population errors for the octane models discussed in the previous section. The mean force errors of the adapted single-species models are quite similar to the two-species models, while the population error between the single- and two-species approaches differ substantially (see Table. 3.1). The population error of the adapted (100) single- and two-species models differ by nearly a factor of 10. This result points to a clear disconnect between force and PMF errors.

For the remainder of this section, we focus our attention on the case of pentane specifically as it

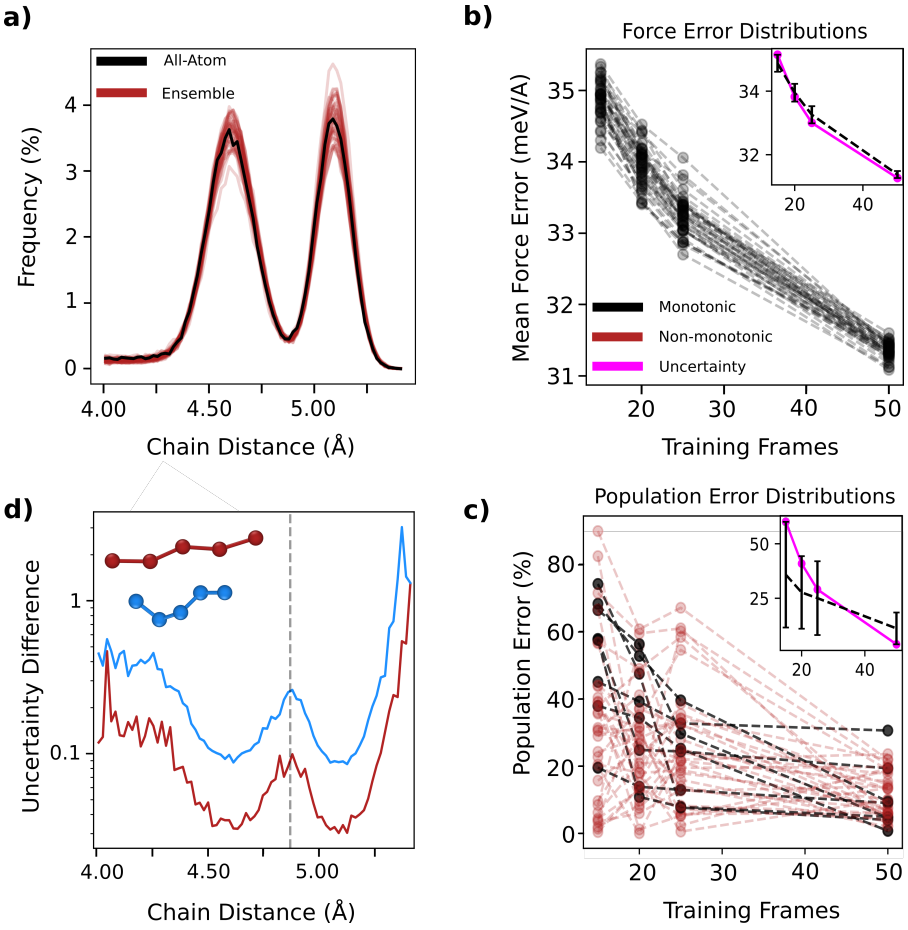

Figure 3.7: Insufficiency in Relative Energies Captured by Uncertainty. a) the end-to-end chain distance distribution of an ensemble of 40 pentane models, each trained with 50 frames of data b) the distribution of mean absolute force errors and c) the population errors defined in Eqn. 3.26 for the pentane ensemble as a function of training set size. Red lines correspond to non-monotonic trends, while black are monotonic. The insets show the mean and standard deviation of the computed property (black), as well as the uncertainty defined in equation 3.26 averaged over models in the ensemble and over molecules within the transition region (magenta). The uncertainty does not share an axis with the force or population errors, but is simply a unitless quantity as described in the Methods d) the effect of adding different types of data is shown. The carbon sites of a transition state molecule (blue) and unphysically stretched molecule (red) are added as new data to a model having 50 frames of data. The resulting change in molecular uncertainty from the baseline model is shown as a function of chain distance.

has a much more bimodal distribution in chain distances compared to octane, corresponding to a more challenging energy landscape in which to study population ratio errors. Even longer chains do not display such bimodal distributions at all (see Supplementary Figure 4). To this end, in Fig. 3.7a we show the end-to-end chain distance distributions for a set of 40 pentane models, each trained with subsets of 50 training frames and 50 sparse points per frame, drawn randomly from the same set of AA constrained dynamics data. It is clear that models vary widely in predicting end-to-end pentane chain distance distributions, despite all models exhibiting low errors on forces. Quantitatively, we show in Fig. 3.7b the learning curves for the set of 40 models along with their corresponding population errors defined by equation (3.25). Black and red lines correspond to monotonically and non-monotonically decreasing quantities, respectively. Each model in the ensemble exhibits monotonic decrease in the force error with more force training data. At the same time, many models show non-monotonic evolution in the population error as a function of the training set size.

These trends can be understood by noting that pentane chain distance distributions are characterized by the difference in PMF values between the two basins. However, in our training, only PMF derivatives (forces) are used, which only implicitly constrain the model's estimate of PMF. As a result, in a test set with minimal representation of transition structures and high sampling of equilibrium ones, we expect that mean force errors will decrease with added data more rapidly than the PMF errors. Thus, improving the fidelity with which these models reproduce distributions of structural properties requires a large amount of force data, especially in the rarely sampled transition region. We note that despite the variability across models in the ensemble, the mean population error as a function of training set size indeed decreases, as does the mean force error. The insets of Fig. 3.7b show the mean force error and population error of the ensemble of models, along with the standard deviation represented with error bars.

The overlap in the distribution of force errors is far smaller than that of the population errors. In each step of the learning process, a model's PMF predictions are less constrained than the force

predictions. As a result, there are effectively more learning pathways that quantities arising from the PMF, such as population errors considered here, can take towards a converged value. Because of the substantial overlap in distributions at different stages of training, many of these pathways are not necessarily monotonically decreasing functions.

One possible solution to minimize this variability would be to include PMF labels in the training set. However, free energies are difficult to compute. In the absence of such PMF labels, however, we argue that by training with force labels alone, the local PMF uncertainties can still be meaningfully interpreted, and that the uncertainties are able to capture useful information regarding the impact force data will have on the model's PMF predictions.

To connect uncertainty more directly to the performance of observable properties, we explore the local PMF uncertainties of CG environments and their relationship to equation (3.25). First, we define the molecular uncertainty as the average local PMF uncertainty on CG sites, i, belonging to molecules with an end-to-end chain distance L. Formally, this is given by

$$\sigma(L) = \frac{1}{N}\sqrt{\sum_{t}\sum_{m_t(L)}\sum_{i\in m_t(L)} \delta\varepsilon_i^2} \tag{3.26}$$

where $\delta\varepsilon_i^2$ is the local PMF uncertainty of CG site i belonging to a molecule m of length L, evaluated on a frame in a test set t. The average is performed over all molecules whose end-to-end chain distance lies within a range $L \pm \Delta L$ in the test set.

In addition to the mean and standard deviation of force errors and population errors shown in Fig. 3.7b, we plot the molecular PMF uncertainty in equation (3.26) as a function of the training set size for a single model averaged over molecules within the transition state region. The molecular uncertainty follows the monotonic trend of the true population error averaged over the ensemble of models, while individual models do not necessarily follow such monotonic trends in population errors, as mentioned above. We suggest that, due to the wide variation across models, the local PMF

uncertainty for a single model does not necessarily predict the true error of populations on a single model, but rather correlates with the expected average error of an ensemble of models.

We also find that force training labels alone constrain PMF predictions in a meaningful way. In particular, the usefulness of the force data that is added is also reflected in the local PMF uncertainties, supporting its use as a metric in on-the-fly training. To demonstrate this, we examine the molecular uncertainty defined by equation (3.26) as a function of chain distance for two different scenarios. Starting from a baseline model trained with 50 frames of pentane force data, we compute $\sigma(L)$ on an independent test set. Subsequently, we consider the addition of force labels in a molecule whose chain distance lies within the transition region ($\sim$ 4.8 Å, model A) compared to the addition of sites in a stretched molecule of chain distance $\sim$ 6.3 Å, model B. The stretched configuration in model B is highly energetically unfavorable, and we expect that this data would have a less meaningful impact on the local PMF uncertainties of the model within the more well-explored regions of phase space.

In Fig. 3.7c, we plot differences of molecular uncertainties, $\sigma(L) - \sigma_{A,B}(L)$, between the baseline model and models A and B as a function of chain distance. Indeed, the transition state model shows a sharper decrease in uncertainty overall compared to the stretched model, directly indicating that the local PMF uncertainties are capable of identifying how force data will constrain the corresponding PMF predictions.

3.7 Efficiency and Stability of CG Models

The benefits of coarse graining are two fold. First and foremost, the reduction of the number of degrees of freedom tracked throughout the simulation reduces the computational cost of the simulation on a per step basis. Once the trained SGP's are mapped, pentane and octane models are equilibrated in LAMMPS with the learned force field for 250,000 steps and a timestep of 1 fs at

constant temperature. The damping parameter of the Nose-Hoover thremostat is set to 100 times the timestep. Production runs to acquire radial distribution functions are run for 400,000 steps, sampling frames every 400 steps.

We demonstrate also the efficiency that is gained over all-atom simulations using ML CG models. While we have used classical force fields in this work as our baseline all-atom reference, this need not be the case. Because classical force fields are generally quite fast, and not nearly as accurate as ML models, comparing the Gaussian process models to them would not be a fair comparison. Instead, we demonstrate the efficiency of the single and two species CG models with respect to an all-atom two species Gaussian process model of the same system. We show this for the pentane systems, but the octane models follow the same trend as their densities are highly similar. In Supplementary Figure 3.8, we can see clearly the substantial gain in computational speed with coarse graining. In particular, the single species models which have shown to be quite accurate are faster by a factor of 30-40.

In addition, removing fast degrees of freedom allows for greater timesteps to be used to enhance sampling efficiency. Here we show the stability of our models with respect to larger integration timesteps. We perform the analysis for the pentane single-species liquid system, but the same conclusions hold for octane where the densities and interactions are highly similar.

All-atom simulations of hydrocarbons would typically use timesteps between 0.5-1 fs. The results previously presented for the coarse grained models are all obtained with a 1 fs integration step in order to ensure fair comparison. Supplementary Figure 3.9 shows that in fact, for timestep sizes between 2-5 fs, the sampling of the pentane models with CG models remains stable, and a larger integration step could be used if desired. On the other hand, we find that above 2 fs timesteps for the all atom model, simulations no longer remain stable. Further, this argument is strengthened by considering the energy drift in NVE simulations, shown in Supplementary Figure 3.10. Here, the 2 fs timestep all-atom simulation already begins to show appreciable energy drift, while the CG

models do not.

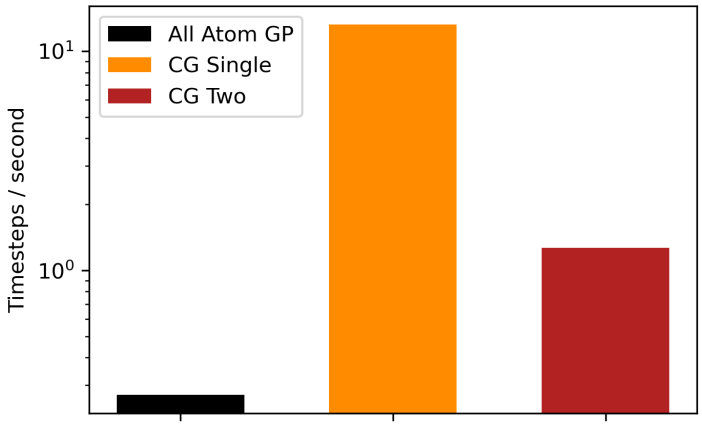

Figure 3.8: Performance comparison between CG and AA models. The number of timesteps per second is shown for single species CG, two species CG, and two species all-atom models. These computations are all performed with a single CPU on an Intel Icelake node. A custom compilation of LAMMPS using the flare pair style was used to run mapped GP models.

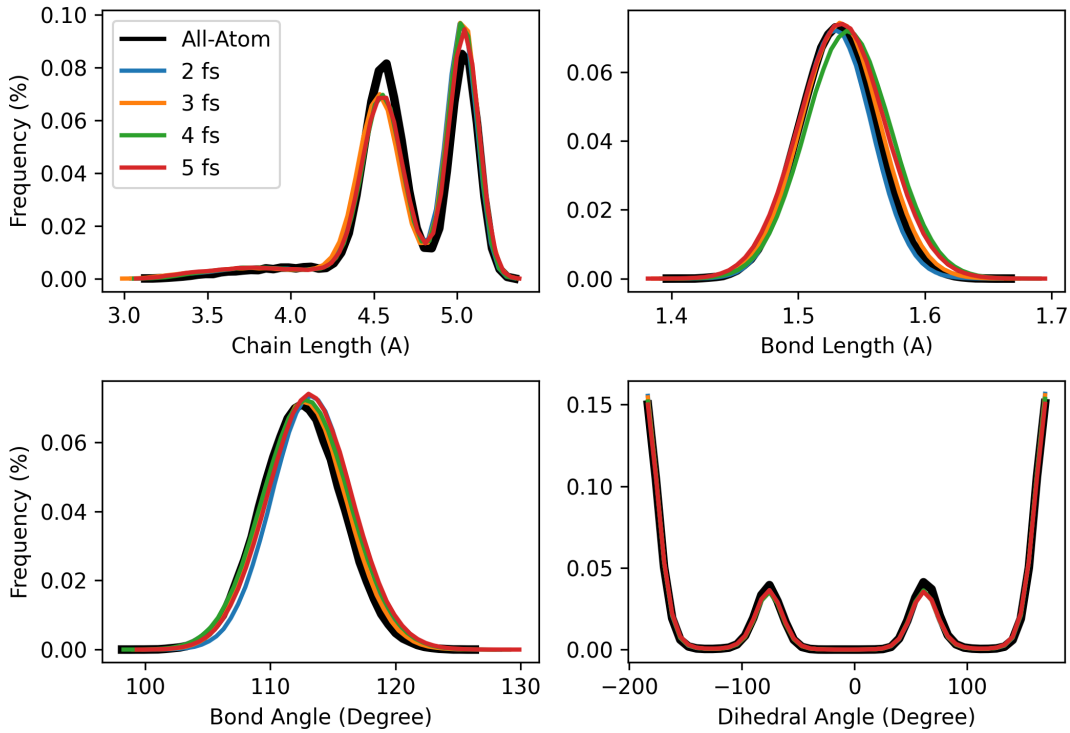

Figure 3.9: Radial Distribution Functions of Large-Timestep CG Models. Distributions are shown for single-species Gaussian process models of pentane using various time steps. All models are the same as those in the main text, with the mapped simulations using different integration steps.

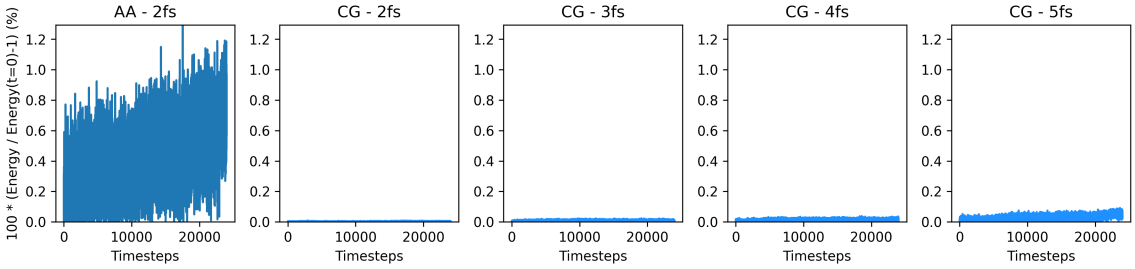

Figure 3.10: Energy drift of CG models with different timesteps. Energy drift for simulations performed in the NVE ensemble. All models were run for 1.2 million time steps. The all-atom model used here is the OPLS model for efficiency, rather than the Gaussian process.

4

Physics Informed Coarse Graining

The conclusions drawn in this chapter closely follow the conclusions reached in the following works. As such, much of the language is shared.

B. R. Duschatko *et al*, *arXiv preprint* arXiv:2405.19386 (2024)

Molecular dynamics (MD) is a vital tool in computational materials design that allows one to

probe the thermodynamic and kinetic properties of matter at atomistic resolution. Of central importance to MD is the accuracy of the force fields used to capture interactions between atoms. While classical empirical force fields derived from experimental observation have dominated the field for many decades, due to the prohibitive computational burden of *ab initio* methods, progress in state-of-the-art machine learning methods has helped to increase the speed and size of simulations while reaching *ab initio* accuracy[22,10,11,8,9,1]. These developments have opened unprecedented capabilities in the modeling of phase transitions, surface reconstruction, catalytic reactions, and biomolecular conformational changes[12,23,82,2,83,84,16,85,86,87,88,89].

Existing atomistic machine learning force field (MLFF) methods have been shown to very effective at regression of the atomistic potential energy surface (PES) of the reference quantum mechanical model. However, because they include all of the atomistic degrees of freedom (DOFs), they require small integration timesteps to resolve fast atomic motion, while accurately computing forces on all DOFs. This limitation is particularly significant for describing soft matter phenomena characterized by a wide range of time scales, such as in dynamics of liquid crystal ordering, polymer reptation, and protein conformation. Coarse graining (CG) methods have long been used to address this issue by integrating out fast DOFs, allowing for larger timesteps and fewer force calculations. Most widely used are coarse grained force fields (CGFFs) with simple fixed functional forms parameterized "top-down" to reproduce experimental observations[31,24,30,45,34], while "bottom-up" CGFFs are derived from a fixed microscopic atomistic model[32,33]. The latter approach is appealing since there is a rigorous way to maintain thermodynamic consistency of the statistical ensembles of the two scales of description. In the bottom-up CG model, the effective energy function of the coarse-grained coordinates is the potential of mean force (PMF)[32,33,36], or the free energy of the constrained CG system, which accounts for the entropy of the all-atom (AA) fine-grain configurations.

Several approaches have been proposed for bottom-up coarse graining. For example, the itera-

tive Boltzmann inversion (IBI) method is a relatively simple approach that allows models to self-consistently target radial distribution functions (RDF) of particular systems[36,37]. Alternatively, relative entropy minimization (REM)[33,90] has proven to aid in the task of learning accurate structural distributions that correctly capture correlated behavior. However, both approaches suffer from various limitations. In particular, the IBI approach is limited in its ability to reproduce accurate structural correlations outside of the RDF it is trained on, while the REM approach requires comparatively expensive training due to the need of multiple molecular dynamics (MD) simulations to self-consistently converge the potential. It is also possible, however, to learn the PMF by regressing its gradients to match Boltzmann averages of forces on the CG sites[32,91]. Compared to IBI methods, these force-matching approaches are capable of capturing a wider array of structural and thermodynamic properties while being significantly faster to train than REM methods.

Despite their appeal, bottom-up CG approaches often require complex functional forms to capture many body interactions. As such, machine learning (ML) approaches have emerged as a promising resolution to this issue[55,60,58,59,91,90]. These ML CGFFs have been shown to be capable of accurate modeling of protein hydration free energies[55], joint learning of interactions and inverse CG to AA mappings[59], as well as on-the-fly learning and transferability across structural spaces[91]. Moreover, ML approaches present a more promising means of accurately capturing higher-order structural correlations that had previously been limited by the simple functional forms of interactions[90].

In the context of the current state-of-the-art in ML CGFF methods, and motivated by the recognized limitations in the field[92], we propose a method that opens new research avenues for exploring some of the outstanding problems in coarse graining.

The first challenge is thermodynamic representability, or correct description by the CG model of thermodynamic properties such as pressures, potential energies, or entropies, in accord with the AA reference. Existing methods rely on learning thermodynamic properties independently from the

free energy functions, such as learning separately the mean potential energy [93], and therefore do not enforce exact consistency between the AA and CG models. The method formulated herein offers a rigorous, consistent way of learning the PMF and other thermodynamic properties subject to exact constraints.

Another capability missing in existing CGFF models is multimodal learning, or the ability of the model to efficiently utilize multiple types of available training data from AA simulations. Current bottom-up coarse graining methods typically only use average AA forces for training, and this results in poor sample efficiency and accuracy of the free energy [91,90,78]. While some previous models were designed to learn from additional thermodynamic properties such as densities[94,95], the PMF remains fixed in the learning procedure as additional variables are incorporated and is hence not improved. In contrast, our proposed formalism can efficiently utilize multiple types of available training data, readily available from fine-grained AA sampling simulations, in a manner that simultaneously improves the accuracy of the PMF and related thermodynamic properties, due to the consistency enforced among them.

Lastly, we note that our new methodology introduces a new direction in exploring a wide range of response properties and the possibility of consistently simulating systems under the influence of external fields. Notably, both AA and CG models have lacked this capability so far. Our approach addresses this challenge by endowing differentiable models with arbitrary parameter inputs that can be used to learn and predict CG-level response properties of any order. The unified differential framework exhibits significantly improved learning efficiency and accuracy. Our work also extends recent works on Sobolev learning of generalized ground-state response properties [96,97] to the case of free energies with arbitrary parameter and temperature dependence.

In this chapter, we present the framework in general terms and subsequently focus on a representative example to illustrate its benefit. Specifically, we include the mean AA potential energies as an additional previously unused learning target for CG free energies. We show that such energy-

informed CG models more accurately capture free energy differences and interaction correlations with far less data, specifically in scenarios with multiple free energy minima. We demonstrate these principles first on a simple model system using a kernel-based approach and further with an equivariant neural network CG representation of free energy of hexane, highlighting how accurately complex structural correlations are preserved, and confirming the utility of our formalism.

4.1 Physics Informed Force-Matching

Let us consider a system with microscopic atomistic (AA) potential energy $U(r, \lambda_a)$ that depends in a general nonlinear way on the set of n atomic coordinates $r \equiv \{r_i\} \in \mathbb{R}^{3n}$, and arbitrary parameters λ_a. The parameters λ_a can be local or global and include the Bravais lattice vectors (or volume) and any generalized forces, such as electric field, magnetic field, electrostatic or chemical potential, etc. The conjugate properties $\partial U/\partial\lambda_a$ are correspondingly stress (or pressure), polarization, magnetization, charge or particle number. This generalization can be applied to lower-dimensional coarse grained representations that can be used to accelerate calculations of thermodynamic properties. This coarse-graining is achieved by integrating over r, the n coordinate degrees of freedom (DOFs) of the microscopic AA system, at a particular value of inverse temperature $\beta = 1/k_B T$ to obtain the free energy $W(R, \lambda_a, \beta)$ of the coarse-grained (CG) system as a function of parameters λ_a and the N CG DOF coordinates $R \equiv \{R_I\} \in \mathbb{R}^{3N}$ defined by the $\mathbb{R}^{3n} \to \mathbb{R}^{3N}$ mapping $R_I = M_I(r)$ for each CG DOF I. In macroscopic thermodynamics, all atomistic DOFs are integrated out fully, with the free energy depending only on global parameters such as volume and temperature. In the context of molecular coarse-graining, the set of coordinates R typically correspond to positions of CG beads, where the map and the free energy W is referred to as the potential of mean force (PMF). We consider the latter case in this work, but the formalism generally applies to any type of dimensionality reduction, including the context of collective variables for enhanced sampling. We implement our

differentiable thermodynamics framework in the context of bottom-up coarse graining, capitalizing on its rigorous statistical consistency between the AA and the CG models. Thermodynamic consistency requires that the partition function, and therefore statistical weights, are preserved, thereby defining the free energy W.

$$Z = e^{-\beta W(R,\lambda_a,\beta)} = \int d^n r e^{-\beta U(r,\lambda_a)} \delta^N(M_I(r) - R_I) \tag{4.1}$$

The function $W(R, \lambda_a, \beta)$ determines all thermodynamic and response properties of the CG system, and it is our goal to learn it, assuming we have the knowledge of the underlying AA potential energy function $U(r, \lambda_a)$. Herein we follow the common choice of a linear mapping of AA to CG coordinates $M_I(r)$ that defines the coordinate of a CG unit in terms of a collection of AA coordinates $R_I = M_I(r) = \sum_j c_{Ij} r_j$, where c_{Ij} are the mapping coefficients between AA coordinates and the CG unit I.

Directly using Eq. 4.1 to compute the value of W for every CG configuration is well known to be intractable due to the high dimensionality of the integral. In a class of CG methods referred to as force matching, the potential of mean force (PMF) is learned using its derivatives with respect to CG coordinates R_I with a loss function consisting of the mean squared residual between the all-atom instantaneous forces on CG sites and the CG forces that are gradients of the PMF[32]. Our approach generalizes this approach by including in the learning objective an expanded set of derivatives the CG free energy function W. To this end, we expand the dimensionless free energy $\beta W = -\ln Z$ with a Taylor series of its parameters and use its various differential coefficients in the

learning task.

$$\beta W(R+\Delta R,\lambda_a+\Delta\lambda_a,\beta+\Delta\beta)=\beta W(R,\lambda_a,\beta)+\frac{\partial(\beta W)}{\partial\beta}\Delta\beta+\beta\frac{\partial W}{\partial R_I}\Delta R_I+\beta\frac{\partial W}{\partial\lambda_a}\Delta\lambda_a \\ +\frac{1}{2}\beta\frac{\partial^2 W}{\partial\lambda_a\partial\lambda_b}\Delta\lambda_a\Delta\lambda_b+\frac{1}{2}\frac{\partial^2(\beta W)}{\partial\lambda_a\partial\beta}\Delta\lambda_a\Delta\beta+\cdots \tag{4.2}$$

where we use the summation convention. The key to our approach is that derivatives of the free energy, which are the coefficients of this expansion, are ensemble averages of the system's response properties that are readily obtained from microscopic constrained dynamics or Monte-Carlo sampling computations driven by the known AA energy function $U(r,\lambda_a)$.

Similarly to training conventional atomistic MLFFs, one approach to fitting the free energy is to train via force labels. The mean force is an ensemble average of atomistic forces constrained by each CG configuration, given by

$$F_I(R)=-\frac{\partial W}{\partial R_I}=-\left\langle\frac{\partial U}{\partial R_I}\right\rangle_{R,\beta}=\frac{1}{Z}\int dr^n\Big(\sum_{i\in I}f_i\Big)e^{-\beta U(r)}\delta^N(R_I-M_I(r)) \tag{4.3}$$

where f_i are AA forces obtained from AA simulations using constrained sampling, and the ensemble average is taken only over AA configurations r_i that map to CG coordinates R_I. If $\tilde{W}$ is the learned model of the true free energy W, the commonly used force-matching loss function takes the form

$$\mathcal{L}_{MF}=\sum_{t}^{T}\sum_{I}^{N}\Big|F_I(R_t)+\frac{\partial\tilde{W}(R_t)}{\partial R_{I,t}}\Big|^2 \tag{4.4}$$

where t indexes the timeframe of a given training configuration, T is the number of CG resolution training frames, I indexes the CG coordinate, N is the number of CG degrees if freedom. We discuss in Methods alternative formulations of the loss function that do not require constrained AA dynamics for training label generation.

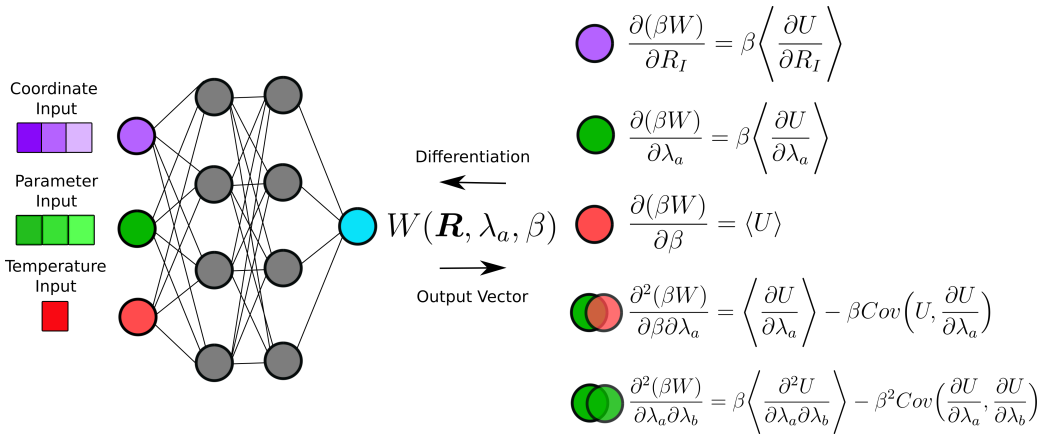

Figure 4.1: Thermodynamically Informed Neural Network Framework for Free Energy Models. In addition to system coordinates (purple), new inputs can be introduced into machine learning models, such as temperature (red) or other global parameters (green), including external fields. The resulting free energy output can be differentiated with respect to these parameters, giving access to new observables and field responses at any order.

The fully general expression of Eq. 4.2 indicates many possibilities to utilize additional properties for training. For instance, a derivative with respect to a parameter λ_a is the linear response coefficient

$$\frac{\partial \beta W}{\partial \lambda_a} = \beta \left\langle \frac{\partial U(r, \lambda_a))}{\partial \lambda_a} \right\rangle_{R,\beta} \tag{4.5}$$

that relates to the pressure of the CG system if the λ_a is the volume, or polarization if λ_a is the electric field. Other training targets are made clear from the fact that derivatives of the dimensionless free energy βW with respect to β are the statistical cumulants of the microscopic potential energy U, since βW is the cumulant generating function for the potential energy as the underlying random variable

$$\frac{\partial}{\partial \beta} \beta W(R, \lambda_a, \beta) = \langle U(r, \lambda_a) \rangle \tag{4.6}$$

$$\frac{\partial^2}{\partial \beta^2} \beta W(R, \lambda_a, \beta) = \left\langle (U - \langle U \rangle)^2 \right\rangle = Var(U) \tag{4.7}$$

We can also include mixed derivatives related to temperature dependent response quantities of the

CG system, noting that the ensemble average operation does not commute with the differentiation operation with respect to parameters of the PMF, e.g.

$$\frac{\partial^2 \beta W}{\partial \lambda_a \partial \lambda_b} = \beta \left\langle \frac{\partial^2 U}{\partial \lambda_a \partial \lambda_b} \right\rangle - \beta^2 \, Cov \left(\frac{\partial U}{\partial \lambda_a}, \frac{\partial U}{\partial \lambda_b} \right) \tag{4.8}$$

$$\frac{\partial^2 \beta W}{\partial \lambda_a \partial \beta} = \left\langle \frac{\partial U}{\partial \lambda_a} \right\rangle - \beta \, Cov \left(U, \frac{\partial U}{\partial \lambda_a} \right) \tag{4.9}$$

The scheme, which we call a "thermodynamically informed neural network", is depicted in Fig. 4.1 and applies generally to other methods such as Gaussian process regression. This differentiability enables access to an extensive range of previously inaccessible physical quantities, both for prediction and for model training. We derive inspiration from the idea of Sobolev training[98,99,100], which is beneficial if the training data for the derivatives is computationally cheap to obtain, which is the case for molecular simulations. Learning can be accomplished by including various combinations of available labels and target properties into the loss function

$$\mathcal{L}_{constr} = \sum_p \gamma_p \sum_t^{N_t} \left| \mathcal{D}_p \left[\beta W(R_t, \lambda_a, \beta) \right] - \left\langle \mathcal{K}_p \left[\beta U(r, \lambda_a) \right] \right\rangle_{R_t, \beta} \right|^2 \tag{4.10}$$

where $\mathcal{D}_p$ are the various derivative operators acting on βW on the left hand sides of equations above, $\mathcal{K}_p$ are the operators in the ensemble averages on the right hand sides, and γ_p are weights for the various loss terms. The sums in the loss function run over the number of training frames N_T and the number of CG sites N. Crucially, when physical quantities are obtained as derivatives of the free energy, they exactly satisfy the correct physical symmetries and conservation laws. This is analogous to ensuring energy conservation when forces are learned as gradients of the energy in standard AA MLFFs. Such conservation laws are not enforced exactly if physical quantities are learned directly with separate dedicated models. An immediate implication for molecular coarse graining

is that these differential relations provide many more possibilities to train a thermodynamically consistent CG free energy model (Eq. 4.2) in a bottom-up fashion. The training labels for the absolute value of the free energy is intractable to obtain due to the difficulty in summing over all states; however, derivatives corresponding to response properties (e.g. Eqs. 4.5, 4.8), are readily obtained from microscopic sampling simulations and can help learn parameter dependent properties, such as stress and polarization, for the finite-temperature CG system. Even more directly important is the ability provided by Eqs. 4.6 and 4.7 to use AA potential energy cumulants for learning the PMF. This addresses the significant difficulty of predicting relative free energies, particularly in free energy landscapes with multiple minima. Because force-matching traditionally relies on force training data alone (Eq. 4.3), it can be exceedingly difficult to capture relative free energies of the minima as well as of the large-barrier transition states. In the limit that we can use for training the force labels over a dense set of Cartesian coordinates of the CG sites, at fixed β, the PMF is determined up to a constant (since only gradients are used). In practice, however, there are often gaps in the training set due to rare occurrence of some configurations. Therefore, the ability to train PMF models using AA potential energies is very valuable. Implementation of this formalism simply requires that the PMF models be made explicitly dependent on temperature so that appropriate derivatives can be taken and used in the loss function. In the following, we provide two demonstrations of the value of our formalism and using additional training labels: one utilizing the loss function in Eq. B.1 using the Allegro model[9], and another using the sparse Gaussian process (SGP) CG framework based on FLARE[91].

4.2 Energy-Informed Free Energy Models

We first illustrate the proposed augmented free energy training by combining force-matching Eq. 4.3 with mean AA potential energy Eqs. 4.6 within a Gaussian process (GP) regression context.

The concept of a loss function in GPs is well understood and discussed in detail in literature [101]. To employ these ideas in the sparse Gaussian process (SGP) approach, we must additionally define covariance relationships to jointly train against and predict both forces and mean potential energies. Specifically, we use the framework introduced in our earlier work based on the FLARE framework [91] wherein the PMF is written as a sum of local contributions as

$$W(R, \lambda_a, \beta) = \sum_{I}^{N} w_I(R, \lambda_a, \beta) \tag{4.11}$$

The covariance between local free energy contributions, w, is given in this case by a dot product kernel function between two descriptors

$$\mathrm{cov}(w_I(d_I), w_J(d_J)) = k(d_I, d_J) = \sigma^2 \left(d_I \cdot d_J \right)^{\xi} \tag{4.12}$$

where ξ is the kernel power, σ the signal hyper parameter, and the descriptors $d_I(R, \lambda_a, \beta)$ are functions of the coordinates, parameters, and temperature. The detail of their implementation are given in Methods. The bi-linearity of the covariance allows us to specify the kernel elements between local free energies and total mean potential energies as

$$\mathrm{cov}(\langle U \rangle, w_J(d_J)) = \frac{\partial}{\partial \beta} \left(\beta \sum_{I} k(d_I, d_J) \right) \tag{4.13}$$

where I labels each CG site. Expressions for the covariance between mean potential energies and forces can be similarly derived. By incorporating the new properties into the kernel of the SGP, one gains not only the ability to train against additional targets, but also a means of predicting principled quantitative uncertainties on predictions of these properties, which have proven to be quite useful in previous work [91,2].

We first illustrate the utility of the SGP approach by considering a simple low-dimensional model

previously used to examine coarse graining strategies [55]. The fine-grain potential energy is given by

$$\beta V(x,y) = \frac{1}{50}(x-4)(x-2)(x+2)(x+3)+\frac{y^2}{20}+\frac{\sin(3(x+5)(y-6))}{25}+\left(\frac{x}{2}\right)^3\left(\frac{y}{\sqrt{170}}\right)^2 \quad (4.14)$$

The CG coordinate is taken to be the value x, with integration over y serving as the coarse graining. Integration for computing the free energy, as well as the mean fine grained potential energy, is performed numerically, and the coefficients are chosen to provide numerically stable solutions. For this example we set $\beta = 1$.

In the following, we demonstrate that learning against average fine-grain potential energies in addition to forces improves the PMF model accuracy, and verify that the use of more complex and expressive temperature-dependent descriptors are not the sole reason for improved performance. To examine the impact of including the mean potential energy in the kernel on helping to capture relative free energy differences, we consider models in two regimes, one in a large data regime and the other in a low data regime. The model free energy introduced in Eq. 4.14 has two energy minima separated by an energy barrier. The SGP's are trained with data collected only from these two minima. In the low data regime, two training points are randomly sampled from each basin, while in the large data regime we sample four (with a total of 8). We further define three types of models: a) models that have no temperature dependence in the descriptor and hence no energy information in the kernel matrix, b) models that have temperature dependence but no energy information in the kernel matrix, trained only on forces, and c) have both temperature dependent descriptors and energy-dependent kernels, trained on both force and energy labels. For the sake of brevity, we will from here on denote models with energy-dependent kernels as "energy labeled models."

The results, depicted in Fig. 4.2, make clear the benefit of using fine-grained energy labels for training the PMF. We note that we are only interested in the relative energy differences between the basins, and in this figure the zero of the free energy is taken to be at the right basin minimum.

As such, the models are made to agree at this location. In the low-data regime, depicted in the first panel of subplots a-c, the energy labeled models capture relative energy difference with much higher accuracy. Moreover, we observe that energy labeled models are additionally more certain in domains outside of the training set, as depicted by the standard deviation of the model predictions. Further, we note the improvement of the temperature dependent model in the absence of energy labels in relation to its temperature independent counterpart. The increased expressiveness of the temperature dependent descriptors allows for a more descriptive model, thereby improving the achievable accuracy. Nonetheless, it is clear that the energy labels are the most significant source of improvement on the attainable accuracy compared to the force-only models.

As a more realistic test case, we shift our attention to the coarse graining of small molecules, relevant in the practical context of liquid solvents, or the design of low-dimensional PMFs for implicitly solvated small proteins. To illustrate the approach in the context of neural network CGFFs, we apply our approach to a coarse grained representation of a single hexane molecule. Hydrocarbons have been frequently explored in the CG methods and applications literature [78,91,54,80,81]. In particular, capturing relative energy differences and accurate sampling of correlated interactions in these molecules has been identified as an important outstanding challenge[91,78,80]. To explore the impact of our temperature-dependent energy-informed approach on the learning of structural correlations in real molecules, we consider a 4-site AA to CG mapping of hexane, depicted in Fig. 4.3a. We use this case to demonstrate our approach, and specifically the loss function proposed in Eq. B.1, with an equivariant neural network model of the free energy, based on the Allegro architecture[9]. We examine the performance of two model types, those with and without training on AA potential energy labels. As metrics of model accuracy, we compare bond length, bond angle, and dihedral angle distributions, obtained with CG and reference AA data. Specifically, the bond-length distribution is defined as the pairwise distance between nearest-neighbor CG sites shown in Fig. 4.3a, the bond angle distribution is between the first three and last three CG sites in the molecule, and the dihedral

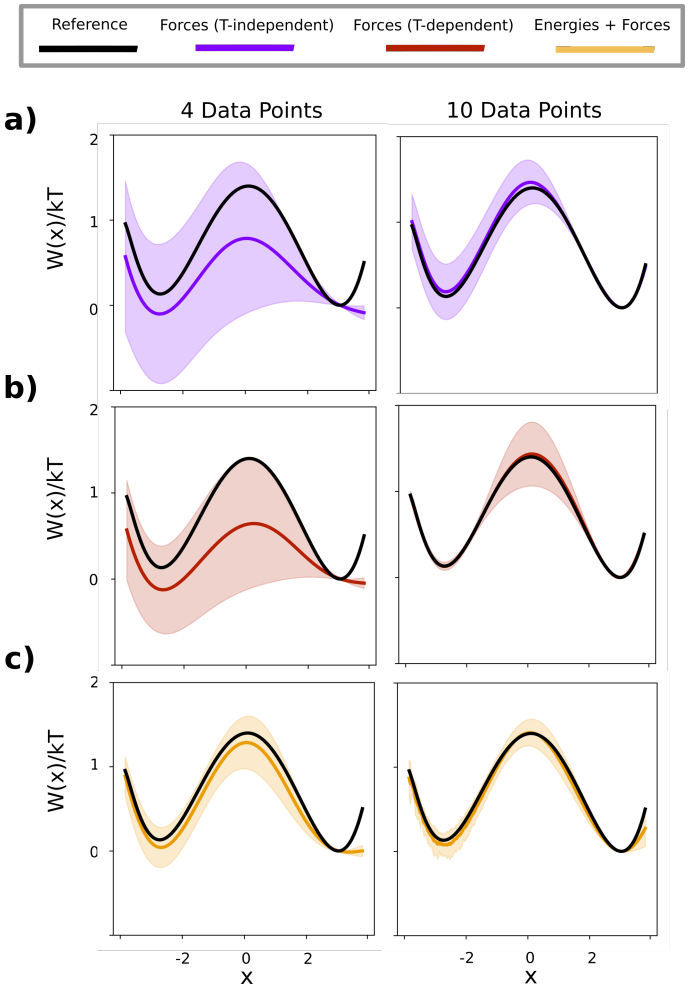

Figure 4.2: Model system of energy-informed CG learning. Shown are the mean (solid lines) and standard deviations (shaded regions) of ensembles of 10 models in two different regimes. In column 1, models contain 4 data points, with 2 sampled from each basin. In column 2, models contain 10 data points, with 5 sampled from each basin. a) The mean and standard deviation of models with force only data and temperature-independent descriptors. b) The mean and standard deviation of models with force only data and temperature-dependent descriptors. c) The mean and standard deviation of models with force and energy data, as well as temperature-dependent descriptors.

distribution is for the single dihedral angle in the coarse grained molecule. Our models are trained on 200,000 time frames (configurations) of AA data obtained from NVT simulations at 250K using the OPLS force field [13].

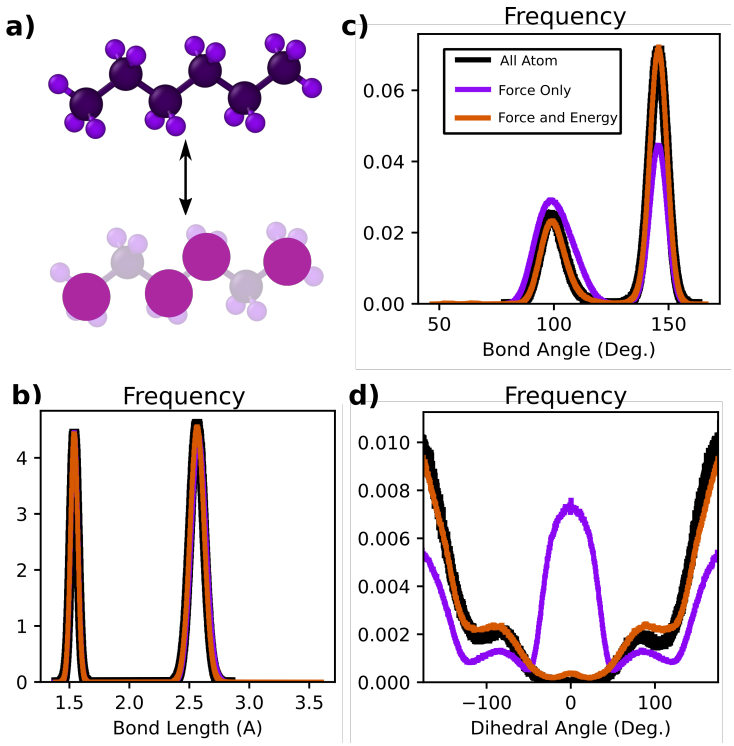

Figure 4.3: Comparison of Structural Distributions With and Without Potential Energies. a) A hexane molecule is mapped to its most interior and exterior carbons for a 4 site CG model. b) the bond length distribution of energy-labeled models (orange) and force-only models (purple) compared to the all-atom baseline (black). c) the bond angle distribution between the two pairs of three consecutive CG sites. d) the dihedral angle distribution for the four CG sites.

As seen in Fig. 4.3b-d, the models containing energy labels are far superior. In particular, Fig. 4.3c and d show that the bond angle and dihedral angle distributions of the energy labeled model match the all-atom baseline with much greater fidelity. Further, we see in Fig. 4.3d that the high-energy states near zero degrees, which in the AA model is very sparsely sampled, are significantly oversampled by the non-energy labeled CG model.

We note that the two Allegro models, both with and without energy labels, are trained in a rel-

atively low data regime compared to other CG NN models in the literature developed for small molecules. For example, CG NN models of alanine dipeptide have in previous works required upwards of a million frames of AA data to train[55]. Equivariant models like Allegro and NequIP have been widely shown to have higher data efficiency in learning compared to other NNs[9,8,102]. At the same time, in the data limited regime, the addition of energetic information leads to even higher data efficiency and much more accurate models.

It is well known that force-matching based CG models can miss key features in the relative values of correlated structural distributions[77,78,79,80,81]. To this end, we additionally examine the free energy surface (FES) of the hexane molecule, defined as follows. We define the FES in this case to describe the relative probabilities of the 12 distinguishable unique structural states of the molecule at CG resolution. By choosing a 4-site CG mapping in Fig. 4.3a, we partition the bond-angle and dihedral angle distributions into 3 and 4 domains, respectively. This is illustrated in Fig. 4.4a. The sampling of these states for a baseline OPLS AA model is depicted in Fig. 4.4a. Further supporting the results in Fig. 4.3, we see that the FES generated by the energy-labeled CG PMF model (Fig. 4.4c) is significantly more accurate than the force-only model (Fig. 4.4b).

The totality of these improvements is summarized and further quantified in Table 4.1 by considering a variety of error metrics. Moreover, we demonstrate the improved learning rate of energy-labeled models by comparing the performance of models with 100,000 and 200,000 data frames. We note that the noisy validation force loss in Eq. B.1 of the Methods is effectively the same across models. This further emphasizes the fact that mean force errors are not a sufficient metric of quality of models for systems with distinct minima states, in problems where occupancy distributions are of interest. In all other error metrics shown with respect to the FES and intra-molecular distributions, the energy labeled models perform better. Specifically, the mean absolute error (MAE) values in all structural distributions are lower for the energy labeled models. Further, we consider which of the 12 FES bins, depicted in Fig. 4.4, have the highest error and lowest error. In both cases, these two

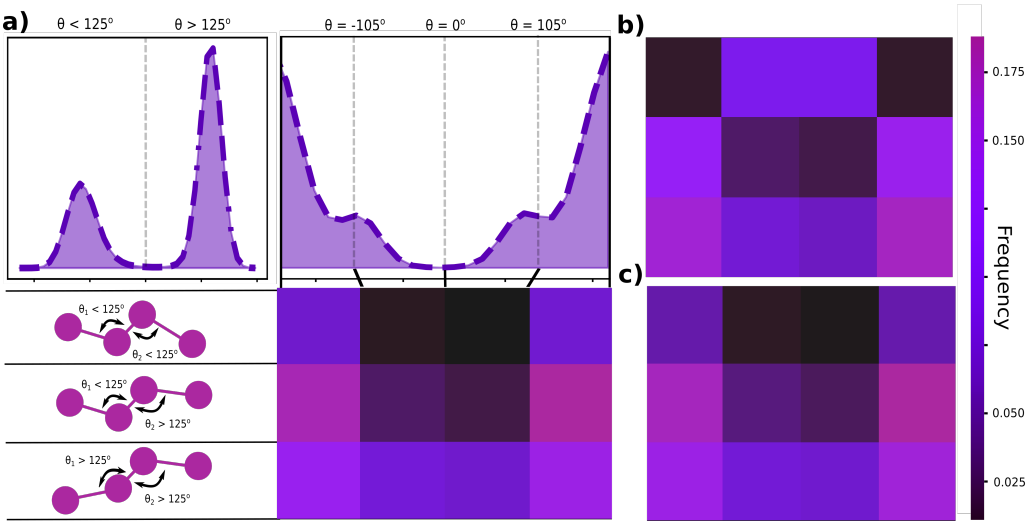

Figure 4.4: Comparison of Structural Correlations With and Without Potential Energies. a) All-atom angular distribution of the molecule corresponding to the CG mapping is shown, with two peaks on either side of 125°. The CG molecule can adopt three different angular conformations corresponding to whether each of its two angles is above or below 125°. The dihedral angle distribution of the all-atom model is also shown and separated into four distinct regions. The relative sampling of the 12 unique correlated dihedral states as measured by an all-atom MD simulation is shown. b) Relative sampling of an optimized Allegro model that is temperature dependent but has seen no energy labels. f) Relative sampling of an optimized Allegro model that is temperature dependent and has seen both force and energy labels. a-c share the color bar shown.

error metrics are higher when energy labels are absent. In addition, the energy-labeled models having been trained on only 100,000 frames of data are more accurate than even the force-only models in the 200,000 data frame regime. In summary, the results on the hexane system indicate that already just one additional training target of our proposed approach, the energy labeling using mean AA potential energies, improves many facets of the PMF learning process with already available AA data, with no additional computational burden, and even requires fewer training data overall.

Metric	Force Only		Forces and Energies	
Data	100,000	200,000	100,000	200,000
Force Loss $(meV/A)^2$	36.916	36.869	36.969	36.867
Bond Length MAE	0.0723	0.009	0.003	0.004
Bond Angle MAE	0.0042	0.0005	0.0005	0.0003
Dihedral MAE	0.0054	0.003	0.002	0.0003
FES MAE	0.0874	0.035	0.029	0.006
FES MAE Max	0.197	0.088	0.052	0.016
FES MAE Min	0.0121	0.001	0.006	0.001

Table 4.1: The table shows a variety of different error metrics for Allegro models, trained on hexane, both with and without energy labels. In both cases, results are given for low-data amounts (100,000 frames) and moderate data amounts (200,000 frames). Considered quantities include the noisy validation force loss given by Eq. B.1, the MAE of the normalized bond length distribution, as well as the bond angle and dihedral angle distributions. In addition, we give the MAE of over the 12-basin free energy surface, including the maximum and minimum single-bin errors.

5

Ongoing Opportunities in Coarse Graining

5.1 Solubility of Ionic Liquids

Coarse graining in the context of biological applications has seem tremendous use over the years. Having naturally long length- and time-scales, it is perhaps most obvious to apply dimensionality reduction to these problems. However, coarse graining need not be limited to this domain. In particular, one can consider coarse graining as a means to simplify sampling problems. In Widom

insertion approaches to computing solubility, one must compute the excess chemical potential as

$$\mu_{ex} = -kT\ln\left(\frac{1}{Z}\int dr^n e^{-\beta u(r^{n+1})}\right) \tag{5.1}$$

over billions of insertions for converged results. Here, the integral is a Boltzmann ensemble average over the configuration space of solvent atoms, measuring the Boltzmann weight of the potential with the solute atom inserted. Moreover, the error of the method goes exponentially as the error in energy. This has recently motivated the use of MLFF's for solubility calculations. However, their relative speed presents a significant bottleneck.

Instead, we consider that as a free-energy based problem, solubility can be computed at CG scales instead. To see that this is so, note that we can write the chemical potential as

$$\mu_{ex} = -kT\ln\left(\frac{Z(N+1,V,T)}{Z(N,V,T)}\right) \tag{5.2}$$

where Z is the partition function. We recall that the coarse grained PMF is defined as

$$W(R^N,V,T) = -kT\ln\left(\int dr^n e^{-\beta u(r^n)}\delta(R^N - M^N(r^n))\right) - kT\ln(C(V,T)) \tag{5.3}$$

where C is a constant term that can only depend on the state points, not the coordinates. Thus the probability of a CG configuration is given by

$$p(R^N) = \frac{C(V,T)}{Z'}e^{-\beta W(R^N)} \tag{5.4}$$

where Z' is the partition function of the CG system. However, note that we may also write

$$p(R^N) = \frac{1}{Z}\int dr^n e^{-\beta u(r^n)}\delta(R^N - M^N(r^n)) \tag{5.5}$$

with Z the partition function of the all-atom system. Equating the two expressions gives

$$\frac{1}{Z} = \frac{C(V,T)}{Z'} \tag{5.6}$$

It then follows directly that, due to C containing no dependence on particle number, the following equality holds

$$\frac{Z(N+1,V,T)}{Z(N,V,T)} = \frac{Z'(N+1,V,T)}{Z'(N,V,T)} \tag{5.7}$$

Thus we may safely use coarse graining to compute the chemical potential of a solute at reduced resolution. With this established, we are actively investigating, in the context of ionic liquids, the following questions:

- Does a CG energy surface lead to more accepted insertions, thereby improving convergence speed due to being a smoother energy landscape?
- Does a CG model reduce the time required to produce diverse sets of initial system frames in which to perform Widom insertions?
- An atomistic ML model will accrue some error with respect to DFT; how severe is the error subsequently accrued by a CG model in learning the ML model as a baseline?

By answering these questions, we hope to bring greater efficiency and accuracy to the problem of solubility calculations.

5.2 Multi-scale Active Learning

An important future direction that becomes an engineering challenge more than a physics problem, but essential none the less, is the efficient connection of CG methods to *ab initio* approaches. Specifically, the accuracy afforded by atomistic ML FF's from DFT data is highly appealing, yet for

long time scale problems become a more significant barrier than with classical potentials. While active learning in both atomistic and CG contexts has shown great promise, the connection of the two is non trivial.

To this end, it is valuable to consider an infrastructure where multiple scales can be traversed. Some of the obstacles that make CG active learning so challenging remain the problem of reconstruction and computational efficiency of constrained dynamics. However, leveraging modern computational infrastructures will play a key role in making this possible. Specifically, constrained dynamics is ammenable to parallelization. Our code infrastructure has implemented this already, and it is an important step forward. Moreover, recent advances in reconstruction techniques would benefit from the vast amount of data generated during constrained dynamics. A positive feedback loop that allows both an active learning FF and a reconstruction algorithm to improve simultaneously is a natural extension of existing tools.

Finally, one can imagine the processes running constrained dynamics being extended to their own atomistic active learning trajectories. Indeed, all of the tools needed for such an extension are in place, but need an effort of engineering to bring them together. In these ways, a clear path forward is obvious provided the right approach to engineering the problem is taken. Constructing a CG active learning routine that utilizes active learning of atomistic processes during constrained dynamics is sure to enable *ab initio* accurate CG models at far longer time scales than previously attainable.

6

Conclusion

The conclusions drawn in this chapter closely follow the conclusions reached in the following works. As such, much of the language is shared.

B. R. Duschatko *et al*, *npj Comput. Mater.* **10** (1), 9 (2024).

B. R. Duschatko *et al*, *arXiv preprint* arXiv:2405.19386 (2024)

Modern advances in computing power have inspired new waves of innovation in computational modeling. Combined with the expressiveness of machine learning architectures, materials science is set in a uniquely exciting position. In a way that was previously perhaps unimaginable, computation is approaching a comparable utility to experimental efforts. Better potentials allow us to probe materials more accurately and at larger scales than was previously possible, starting from first principles and circumventing many of the accuracy limiting approximations that existed before.

All-atom modeling techniques on their own continue to experience tremendous growth and push the boundaries of what we are capable of studying. The back-end software of machine learning potentials is continuously being improved and expanded, and the theory underlying our approaches to the design of inter-atomic potentials are maturing steadily. Regular, DFT-accurate simulations of millions of atoms is on the horizon of computational capabilities.

Further, developments within the coarse graining community serve to enhance the scope of all-atom methods by making even longer length and time scales attainable. In particular, the developments presented herein for coarse graining offer a significant improvement upon state of the art techniques. First, we have explored the utility of uncertainty aware machine learning models for coarse graining. The principled Bayesian uncertainty measure of Gaussian processes enables an on-the-fly active learning scheme for CG models that allows for automating the creation of training sets. A key aspect of the on-the-fly ML CG method is that it overcomes the time- and length-scale limitations of AA models by integrating the AA system in time only over fast degrees of freedom. In addition, by collecting training data only when necessary during MD, we eliminate the need for specifying *a priori* which configurations our model will need to be trained on. Instead, as the algorithm explores configurations, it automatically decides if they are new enough to be added to the training set.

In addition, computational speed is a key metric in CG force fields. While fast classical all-atom force fields are used as a reference in this work, this need not be the case. For example, one could

target ML models based on *ab initio* data that would be far more accurate. To this end, we quantified the efficiency gained by using ML models at a coarse grained resolution as opposed to the full all-atom alternative. We find that these are even more efficient when taking into account the largest stable timesteps.

Further to the point of designing more efficient CG models, the analysis we have done on the performance of single- versus two-species models is valuable. In this Gaussian process framework, there is unfavorable scaling with the number of species. As such, a single-species representation equates to a much faster model. Systems where this choice does not lead to significant accuracy loss in pair-wise and correlated distributions, such as the pentane system considered, will benefit greatly from improved model efficiency.

Moreover, the on-the-fly framework enables models to be transferred across molecular systems in a principled way. By the locality assumption inherent in the structural descriptors, for local environments that are similar across systems forces will be predicted to be similar, at a given thermodynamic state. Similarly, environments that differ will result in small kernel values and contribute proportionally less to the prediction of forces on new environments, and the framework will request more data.

In practice, this approach could be used to develop CG models for common polymer backbones that could then be efficiently adapted to systems with differing functionalization. In this case, only data around new functional groups would be required, as opposed to starting from scratch. This would allow for more rapid development of coarse grained models in materials screening settings. We emphasize that there is a tradeoff in single- and two-species representations for transferable models. In the octane example considered in this work, there is an ease of direct transferability of single-species models without adaption. On the other hand, the achievable accuracy upon the addition of more data of the more descriptive two-species model may be more desirable. In general, it is difficult to anticipate the extent of this complex tradeoff.

The issue of model variance is further addressed herein as arising from limited direct PMF information. By training on time-averaged forces only, obtained from unbiased MD configurations, PMF in transition regions is difficult to capture. What's more, examining force errors alone can seem to suggest that models should always improve with more data. In fact, we show that over an ensemble of models, average force errors indeed decrease monotonically with more data, while PMF errors do not necessarily decrease for each model. As a result, properties arising from PMF values, such as population ratios of stable configurations, can vary significantly and converge slowly in the training process. We find only that the average over a set of models will improve such property predictions. We demonstrate that despite the lack of direct PMF labels in the training set, providing forces still imparts meaningful PMF information to the model that is reflected in the monotonic behaviour of uncertainties. This allows a better understanding of model robustness where traditional metrics such as force errors are insufficient.

We note that the molecular local PMF uncertainty plotted in Figs. 3.7b,c does not correspond quantitatively to the mean absolute force error or population error. While some work has been done on understanding how to more concretely link uncertainty and performance [103], it is not well understood how we can, for example, translate specific values of uncertainty directly to the variance of observed structural properties over an ensemble of models. This remains an open question and demands systematic investigation even for force-fields in all-atom simulations. Moreover, the particular form in which we analyze molecular uncertainty in equation (3.26) is a choice. Other functional forms of molecular uncertainty can be conceptualized that may or may not provide deeper physical insight. This will require careful considerations in future works.

Finally, we note that a particular challenge in making the proposed method more generally applicable is the reconstruction of all-atom configurations to enable constrained dynamics during the active learning loop. In order to seamlessly go between the all-atom and coarse grained representations, a scheme for recovering lost degrees of freedom must be designed. In many cases this is

done with techniques designed for specific systems (as we have utilized in this work), or brute force methods such as a multi-stage compression and expansion of the system box to equilibrate replaced degrees of freedom. Recent works have begun to examine this problem from the perspective of machine learning[72,75,70,76,71], but such approaches require pre-selected training sets to learning the CG mapping that would require additional work to reconcile with our active learning scheme. Doing so will enable a much broader study of materials at a variety of resolutions.

The second primary contribution to the field is the establishment of a rigorous theoretical framework for improved learning of high-dimensional free energy models, such as the PMF in the context of coarse graining.

A primary goal of coarse graining approaches lies in the accurate estimation of free energies as functions of coordinates of reduced degrees of freedom. Estimation of free energy functions is a long-standing challenge in statistical mechanics. In abstract statistical sense, our work provides a general framework for constructing differentiable models to learn high-dimensional cumulant generating functions. In the conceptual vein of Sobolev training, we train the models on derivatives of the target function, which are the cumulants of the underlying random variables, that we obtain by statistical sampling. We note the distinction from the idea of physically informed neural networks (PINNs)[104], where differential expressions involving only the output are used as additional regularization terms in the loss function. Instead, we augment the model inputs with parameter inputs and use exact differential relations to allow for additional training targets, improving the model transferability, accuracy and training efficiency. Our approach also builds on the elements of the classic CALPHAD method, where phase diagrams and equations of state are determined by learning the macroscopic free energy as a low-dimensional polynomial function using its observable derivatives[105]. Neural network models for the free energy were also proposed for learning the macroscopic equations of state[106].

Specifically in the context of molecular coarse graining, we provide a way to learn the free energy

by using physically observable thermodynamic and response properties in a unified thermodynamically consistent manner by identifying these statistical cumulants with derivatives of the free energy. The advantage of this approach is the simplicity of its implementation, where existing CG models are endowed with addition explicit inputs on parameters, specifically including the temperature and modifying the loss function with matching observable - derivative pairs. This is the first method to date, to our knowledge, that utilizes potential energies for training in such a way and that enables the values (not only gradients) of the model PMF to be directly improved through the new training information.

Further, the resulting models require no additional computational overhead, since atomistic constrained dynamics simulations produce trajectories that contain force training labels along with total potential energies, from which a variety of statistical cumulants can be estimated at no additional computational cost. This framework can be seamlessly implemented into existing learning architectures, and is shown to produce more accurate PMFs with a higher learning efficiency. As demonstrations of the improvements at fixed temperature, we have considered a low-dimensional toy model example, as well as a realistic molecular coarse graining procedure for the hexane molecule.

We note that additional computational overhead may be incurred in neural network implementations of the Sobolev training strategy, as a result of needing multiple passes of backpropagation for higher order derivatives. However, we expect that the statistical sampling of AA labels will dominate the computational cost nonetheless. An issue to be resolved for unconstrained sampling is that the loss function over noisy labels in Eq. B.1 of Methods is easy to implement only up to first derivatives (our demonstration includes the use of forces and potential energies). If using the variance of the potential energy (Eq. 4.7), it is necessary to estimate the mean potential energy for each CG configuration, prior to formulating a noisy unconstrained estimation of the variance. Our formalism applies straightforwardly at any order when using constrained sampling to estimate ensemble averages, but efficient implementation of such estimates using noisy unconstrained labels for higher

order derivatives should be addressed in future work.

As an extension of this multi-modal learning, we recognize that our method also offers a promising avenue for improving the state point transferability of CG models. For example, while CG models have been trained before against multi-temperature force data with an explicit temperature input parameter [57], our proposed inclusion of cumulants of the AA potential energy data in the training labels brings significant additional information via the PMF's temperature dependence via its gradients. As a result, this is a promising new means of simulating systems across different phases with improved model transferability. More generally, our model formulation allows for the learning of the PMF dependence on any thermodynamic state point directly through an expanded set of ensemble observables. Demonstration of this capability is left to a future investigation.

We note some limitations of the present analysis as opportunities for future work. For example, we should expect that, in the case of potential energy labeling at fixed temperature and volume, the energy labels will be maximally helpful to the free energy when the all-atom constrained entropic contribution is close to zero, i.e. when the atomistic potential energy labels are most informative. While any additional information from the energy labels will help, the extent to which energies help in learning the PMF should be more closely explored, particularly for large systems where the total potential energy label may contain limited information.

Broadly, for a given mapping between the fine and coarse grained representations, we present a unified differential framework for connecting the two scales in a rigorous consistent way and for capturing arbitrarily complex nonlinear and coupled dependencies of coarse-grained free energy on system parameters. Our method can be readily generalized to include additional thermodynamic observables as targets, such as stresses, polarization, and magnetization. This can be accomplished by including as inputs to the CG model and differentiating it with respect to variables such as strain, electric field, and magnetic field, respectively, and matching in the loss function with the corresponding AA observables. This framework enables the development of coarse grained

temperature-dependent models in the presence of any set of generalized forces acting in a variety of thermodynamic ensembles. Examples of such future work might include exploring temperature and pressure-dependent phase transitions, coupled electro-mechanial response, and extensions to other thermodynamic ensembles such as the grand canonical ensemble. Examples to be explored include higher-order derivatives, corresponding to learning from and predicting e.g. heat capacity and compressibility. In general, this framework will enable a better understanding of what aspects of coarse-grained models can be improved by training with different combinations of all-atom ensemble averages.

With such novel developments placed on firm footing, the field is situated in an exciting position. New problems can be addressed at higher levels of accuracy than previously possible. The future of computational materials design will benefit from the continued development of coarse graining procedures to enable full multi-scale solutions to real world problems.

A

Supplementary Materials for Chapter 3

A.1 Mapped Sparse Gaussian Processes

A great deal of work has been done in accelerating Gaussian process models. For example, the mapping procedure for two- and three-body GP's has been extensively discussed in previous work [3]. While these methods are perfectly amenable to CG approaches, we have focused in this work on SGP's. To this end, we present a short review of the methods introduced in Ref. [2] for accelerating

SGP's.

Provided we have a normalized descriptor $\tilde{d}_i = d_i/d_i$ and a kernel that is a normalized dot product of the form

$$k(\tilde{d}_i, \tilde{d}_s) = \sigma^2(\tilde{d}_i \cdot \tilde{d}_s)^\xi \tag{A.1}$$

it can be shown that the sum of local energy predictions over sparse environments factors as

$$\varepsilon(\rho_i) = \sigma^2 \sum_s (\tilde{d}_i \cdot \tilde{d}_s)^\xi \alpha_s \tag{A.2}$$

$$= \sum_{m_1,\ldots,m_\xi} \tilde{d}_{im_1} \cdots \tilde{d}_{im_\xi} \Big(\sigma^2 \sum_s \tilde{d}_{sm_1} \cdots \tilde{d}_{sm_\xi} \alpha_s \Big) \tag{A.3}$$

where the index m represents an expansion based on the kernel power. When $\xi = 1$, the term in brackets is a vector, for $\xi = 2$ a rank 2 tensor, and so on. Crucially, these terms in brackets are the only ones depending on the training set sparse points and do not depend on the environment being predicted on. It can therefore be computed once and for all and used in subsequent computations without recalculating. This effectively removes the training size dependence.

A.2 Mapped Sparse Gaussian Process Simulations

Once the trained SGP's are mapped, pentane and octane models are equilibrated in LAMMPS with the learned force field for 250,000 steps and a timestep of 1 fs at constant temperature. The damping parameter of the Nose-Hoover thremostat is set to 100 times the timestep. Production runs to acquire radial distribution functions are run for 400,000 steps, sampling frames every 400 steps.

We demonstrate also the efficiency that is gained over all-atom simulations using ML CG models. While we have used classical force fields in this work as our baseline all-atom reference, this need not be the case. Because classical force fields are generally quite fast, and not nearly as accurate as ML models, comparing the Gaussian process models to them would not be a fair comparison. In-

stead, we demonstrate the efficiency of the single and two species CG models with respect to an all-atom two species Gaussian process model of the same system. We show this for the pentane systems, but the octane models follow the same trend as their densities are highly similar. In Supplementary Figure 3.8, we can see clearly the substantial gain in computational speed with coarse graining. In particular, the single species models which have shown to be quite accurate are faster by a factor of 30-40.

A.3 ASE On-the-Fly Parameters

The on-the-fly molecular dynamics loop is run with ASE [107]. The NPT ensemble is used for the Nose-Hoover thermostat implementation, but with the barostat turned off and kept at constant volume. A timestep of 1 fs is used, with a damping parameter for the thermostat equal to 100 times the timestep. On-the-fly trajectories are performed for 100,000 steps within ASE's molecular dynamics engine prior to being mapped and run in LAMMPS.

A.4 Constrained Dynamics Parameters

After reconstructing the all-atom representation for them coarse grained representation (as well as when collecting fixed training and test data for which no reconstruction is necessary), the free degrees of freedom are allowed to evolve at a constant 250K temperature. For the on-the-fly runs, we use a 0.5 fs timestep during constrained dynamics and find that averaging forces over 10,000 frames sampled 200 steps apart is sufficient for mean-force convergence (see Supplementary Figure A.1). This trend can be verified both by looking at the standard error of the mean of a selection of force components, as well as the marginal log likelihood of an SGP trained on a limited amount of data as a function of mean-force sampling. Once sufficient sampling is reached, the model noise becomes dominant over the data noise, and further constrained dynamics will not benefit the models perfor-

mance.

A.4.1 Hyperparameter Selection

As in Ref. [2], we maximize the log marginal likelihood of the sparse Gaussian processes in order to choose optimal hyperparameters. We optimize the likelihood with respect to the kernel power, ξ, basis expansion parameters n and ℓ, as well as the environment cutoff radius for a pentane model (see Supplementary Figure A.2). For consistency of comparison, we use the optimized parameters for the two species model for all single species models. The octane models use the same parameters. While the cutoff radius is optimized by the likelihood at 4.2 Å, we find better overall performance of each given model when using a cutoff of 4.5 Å.

A.4.2 Structure of n-alkane Chains

In the main text we explore the behavior of an ensemble of SGP models in the presence of nearby local free energy minima. Supplementary Figure A.3 motivates the study of pentane by examining the end to end chain distance distribution of increasingly long n-alkane chains.

A.5 OPLS Force Field Parameters

In order to generate force field parameters for both pentane and octane, the LigParGen [108,109] server is used. Interactions between different species are defined using a geometric mixing rule for the parameters. The pentane system, consisting of 70 molecules, and the octane system consisting of 40 molecules, are placed around the edges of a box to begin. The fire minimization scheme implemented in LAMMPS [110] is first performed, with a time step of 0.1 fs, with an energy and force tolerance of 1e-7 and 1e-9, respectively. Random velocities are then drawn for all atoms for a Boltzmann velocity distribution with temperature 62.5K.

For the pentane liquid, using a timestep of 0.05 fs, 500,000 timesteps are run in the isothermal-isobaric ensemble from a temperature of 62.5K to 250K, and from a pressure of 0 to 1 atmosphere. The damping parameters for the thermostate and barostat are set to 100 and 1,000 times the timestep, respectively.

Another isothermal-isobaric ensemble segment is performed, this time with a timestep of 0.5 fs for 800,000 steps at constant temperature and pressure. Finally, 200,000 steps are made with a 1 fs timestep at a constant temperature of 250K. The resulting equilibrated structure is used as the starting point for all subsequent simulations, including constrained dynamics, initial on-the-fly frames, and the starting point for all mapped simulations.

For the octane liquid, the first isothermal-isobaric stage is performed with a 0.01 fs timestep for 1,000,000 steps, from a temperature of 25K to 300K, and from a pressure of 1 to 151 atmospheres. The damping parameters are the same as for pentane. In the second isothermal-isobaric stage, a 0.5 fs timestep is used over 800,000 steps, from a temperature 300K to 250K, and from 151 to 1 atmosphere of pressure. A third isothermal-isobaric stage is run with a 0.5 fs timestep for 200,000 steps at constant temperature and pressure. Finally, an isothermal stage with a timestep of 1 fs is run for 200,000 steps at constant temperature.

To compare our coarse grained models against a common baseline for n-alkane systems, we run a simulation of the same pentane system with the parameter set of the OPLS United Atom approach [31,111]. In particular, the potential energy takes the form

$$\begin{aligned} E(r^N) = &\sum_{\text{bonds}} K_b(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 \\ &+ \sum_{\text{dihedrals}} \sum_{k=1}^{4} \frac{V_k}{2}(1 + (-1)^{k+1}\cos(k\varphi)) \\ &+ \sum_{i \neq j} 4\varepsilon \left(\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6} \right) \end{aligned} \tag{A.4}$$

The groupings in the OPLS-UA formalism chosen here are electrically neutral and give no Coulomb contributions. The system is broken into two species, the CH_3 ground and the CH_2 group. The longest values reported in [111] are for the butane groups, and we apply these parameters (see Supplementary Tables A.1-A.4) to the pentane system.

Atom Type	α (CH_α)	σ (Å)	ε ($\frac{\text{kcal}}{\text{mol}}$)
1	3	3.905	0.175
2	2	3.905	0.118

Table A.1: The OPLS United Atom vander-waals parameters used to simulate pentane.

Bond Type	Atom 1	Atom 2	K_b ($\frac{\text{kcal}}{\text{mol}}$)	r_0 (Å)
1	2	1	260.0	1.526
2	2	2	260.0	1.526

Table A.2: The OPLS United Atom bond parameters used to simulate pentane.

Angle Type	Atom 1	Atom 2	Atom 3	k_θ ($\frac{\text{kcal}}{\text{mol}}$)	θ_0 (Degrees)
1	1	2	2	63.0	112.4
2	2	2	2	63.0	112.4

Table A.3: The OPLS United Atom bond angle parameters used to simulate pentane.

Dihedral Type	Atom 1	Atom 2	Atom 3	Atom 4	V_1 ($\frac{\text{kcal}}{\text{mol}}$)	V_2 ($\frac{\text{kcal}}{\text{mol}}$)	V_3 ($\frac{\text{kcal}}{\text{mol}}$)	V_4 ($\frac{\text{kcal}}{\text{mol}}$)
1	1	2	2	2	0.0	0.0	2.0	0.0

Table A.4: The OPLS United Atom dihedral angle parameters used to simulate pentane.

A.6 Figures

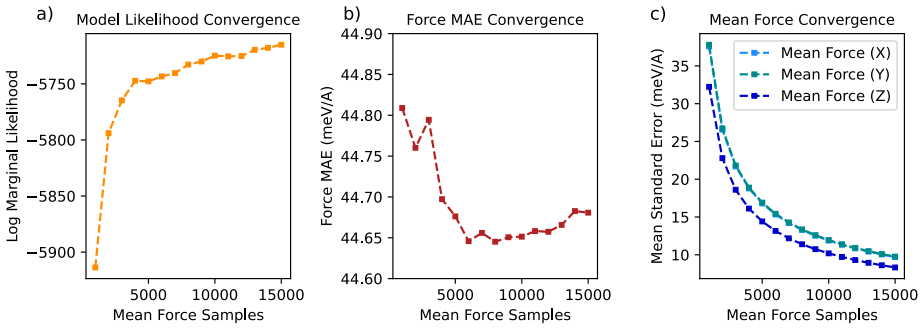

Figure A.1: Optimization of Mean Force Convergence. The marginal likelihood (a), mean absolute force error on a test set (b) and the average standard error of the mean of PMF derivatives, i.e. forces, (c) are shown as a function of constrained dynamics sampling. The x-axis indicates the number of uncorrelated frames in which data was averaged over to provide force labels to the corresponding model.

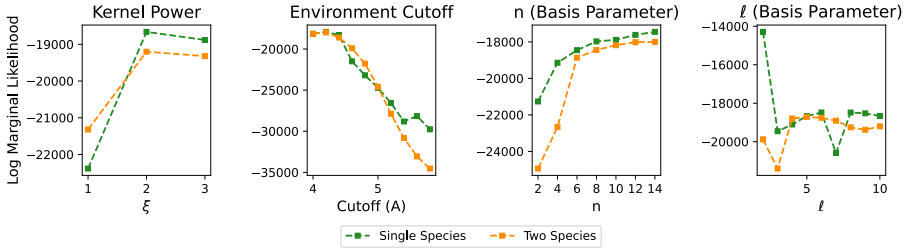

Figure A.2: Gaussian process hyperparameter optimization. The four hyperparameters for the sparse Gaussian processes are optimized by maximizing the log marginal likelihood. For each parameter sweep, the values are fixed at $\xi = 2$, $r_{\text{cut}} = 4.5$ Å, $n = 5$ and $\ell = 12$ for those not being varied.

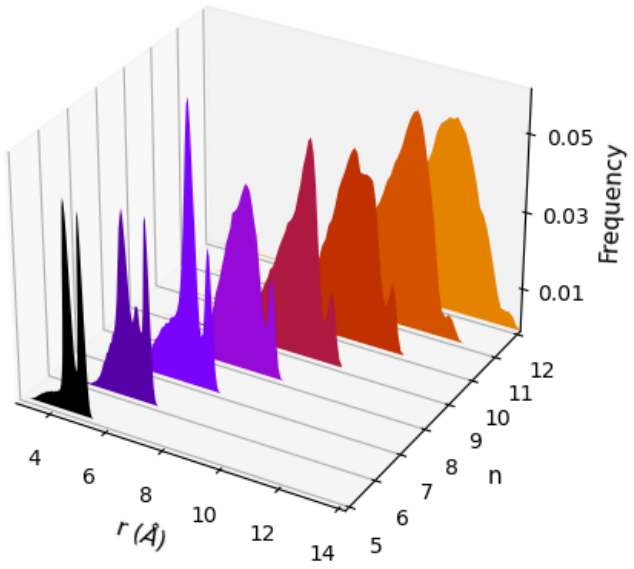

Figure A.3: Radial distribution functions of hydrocarbons. The end-to-end chain distance distributions for n-alkanes from pentane to dodecane. At longer chain distances, the bimodal distribution begins to disappear. This is the result of many more dihedral energy minima along the chain.

B

Supplementary Materials for Chapter 4

B.1 Methods

The loss function discussed in the main text targeting constrained dynamics labels is not the only choice of a thermodynamically consistent loss function. In place of fitting to these constrained force labels as in Eq. 4.4, it is common in force-matching schemes to minimize a noisy loss function using unconstrained estimates of the forces on CG units. In this case, the framework proposed in this

work for including arbitrary thermodynamic properties can be extended to these schemes using the relationships between free energy and other thermodynamic properties, via the the loss function

$$\mathcal{L}_{noisy} = \sum_m \gamma_m \sum_t^{N_t} \left| \mathcal{D}_m \left[\beta W(M(r_t), \lambda_a, \beta)\right] - \mathcal{K}_m \left[\beta U(r_t, \lambda_a)\right] \right|^2 \tag{B.1}$$

where the frame index t now indexes instantaneous atomistic frames, and not constrained CG values. We emphasize the subtle but important difference in this expression compared to Eq. 4.4: The loss function in Eq. 4.4 regresses the properties of the PMF to their thermodynamic averages, while Eq. B.1 regresses the properties of the PMF to the noisy, instantaneous atomistic predictions of the property. It is not immediately obvious that the two loss functions produce the same minimum. We provide a derivation of this equivalence in the SI.

The implementation of this loss function requires the model to be explicitly dependent on temperature. To accomplish this, we modify the Allegro architecture to introduce the temperature as part of the input node features. The original Allegro model uses one-hot embedding of the atom type as the node feature: $\mathbf{h}_i = \text{1Hot}(Z_i)$, where Z_i is the discrete type of node i (chemical species in all-atom potentials). We augment $\mathbf{h}_i$ by concatenating a temperature embedding with a Gaussian basis: $\mathbf{h}_i = \text{1Hot}(Z_i) \parallel B(\beta)$, where $\parallel$ denotes concatenation, B is the Gaussian basis function, and $\beta = 1/k_B T$. A similar encoding was described in an earlier work[57], which however did not consider the use of thermodynamic differential relationships to augment the training targets, as proposed here. In our case, we compute the derivative $\partial \tilde{W}(\mathbf{R}^N, \beta)/\partial\beta$ through backpropagation and use the result in a generalized loss function.

Similarly, Gaussian process models also require temperature dependence in order to leverage additional training targets. For the Gaussian process model, we choose to define descriptors as modified variants of those introduced in our earlier work[91] based on the Atomic Cluster Expansion (ACE)[22]. Specifically, in order to leverage thermodynamic relationships of the PMF, we introduce

temperature dependence into the SGP by adding temperature dependent embedding functions in the descriptors. The non-temperature dependent SGPs are constructed in the FLARE MLFF with a descriptor of the form

$$d_{is_1s_2n_1n_2\ell} = \sum_m c_{is_1n_1\ell m} c_{is_2n_2\ell m} \tag{B.2}$$

where the ACE atomic base is given by

$$c_{isn\ell m}(R) = \sum_{j \in \rho_i} R_n(|R_{ij}|) Y_\ell^m(\hat{R}_{ij}) \delta_{s,s_j}$$

while for the proposed temperature-dependent models, we introduce an expansion in a basis of functions of temperature

$$c_{isn\ell m}(R, \beta) = \sum_\gamma a_\gamma^{n\ell} \sum_{j \in \rho_i} R_n(|R_{ij}|) Y_\ell^m(\hat{R}_{ij}) \delta_{s,s_j} \Gamma_\gamma(\beta)$$

with new coefficients, a, that can couple to the radial and angular basis functions. In this work, we take the basis functions, Γ, to be the Chebyshev polynomials. With this change, the descriptor in Eq. B.2 becomes a temperature and parameter dependent descriptor such as those appearing in the main text.

B.2 Computational Details

All production simulations were performed in the LAMMPS software package [110], with atomistic models utilizing parameters from the OPLS force field [13]. The parameter files were generated using the LigParGen server [108,109]. In order to train the Gaussian process models, a modified version of FLARE was developed to include new descriptors and kernels. For production CG simulations, custom LAMMPS software was developed for fast implementation of the NN and SGP models as pair styles.

In order to generate structural distributions for model validation, we run NVT simulations of each AA model for 2 nanoseconds, starting from 343 different initial velocities and equilibrating for 200 picoseconds. We use a timestep of $dt = 0.5$ fs with a damping constant for a Langevin thermostat of $100 * dt$.

In the SGP models, we construct a set of 10 models having been trained on different random samplings of the data described in the main text. Each model is then optimized independently by minimizing the force mean squared error (MSE) between model predictions and the ground truth values over the full domain of the PMF shown in Fig. 4.2. The ground truth values are computed via numerical integration of the PMF. Further details of the optimization procedure are provided in the SI.

In the NN models, a temperature embedding was included in the architecture for both the force-only case and the case including atomistic potential energies. The energy dependent models were trained by optimizing a joint loss function as in Eq. B.1 that includes a noisy force contribution as well as a noisy potential energy contribution. The relative weighting of these terms was taken to be a hyperparameter. The hyperparameters for each model type were chosen by minimizing the total validation loss of the models. In particular, the energy-labeled models included an energy-loss term in their validation metric, as in Eq. B.1. The details of parameter sweeps, as well as a sample input

file, are expanded upon in the SI.

B.3 Regression with Thermodynamic Properties in the Force Matching Framework

The multi-scale coarse graining (CG) methodology established by Noid et. al.[32] includes a proof that the minimum of the functional

$$\chi^2_{FM}[W] = \frac{1}{3N}\left\langle \sum_{I=1}^{N} \left| f_I(r) + \frac{\partial W(M_I(r))}{\partial R_I} \right|^2 \right\rangle_r$$

can only be the true potential of mean force, up to an additive constant. Here, W is an arbitrary model of the PMF, $\tilde{W}$. Here, $r \in \mathbb{R}^{3n}$ are the atomistic coordinates, $R \in \mathbb{R}^{3N}$ are the CG coordinates. The ensemble average is over the Boltzmann distribution in the all-atom configuration space $\{r\}$, and the sum is over all coarse grained sites. $f_I \in \mathbb{R}^3$ are the instantaneous forces acting on the CG sites, which, for a center of mass mapping, is simply the sum of atomic forces. The mapping between atomistic and CG coordinates is given by the linear mapping function $M_I = \sum_i c_{iI} r_i$. Note that, while translational invariance of the CG model can be maintained by a more general surjective equivariant mapping, the proof in Ref.[32] relies on an assumption of linearity. An important consequence of this proof is that minimizing the approximation of this functional

$$\chi^2[F^{MS}] = \frac{1}{3n_t N} \sum_{t=1}^{n_t} \sum_{I=1}^{N} \left| f_I(r_t) - F_I^{MS}(M_I(r_t)) \right|^2 \tag{B.3}$$

where $F^{MS} = -\frac{\partial W(M_I(r))}{\partial R_I}$, provides a means for estimating the potential of mean force. Note that this is an approximation when the Boltzmann distribution is only partially sampled, but will become exact as $n_t \to \infty$. For a finite n_t, this differs from the previous expression in that the first implies a complete ensemble average, while the later represents an approximate sampling of the ensemble. Here, the first sum is over time frames of, for example, a molecular dynamics simulation

or another sampling technique. We propose in this work that additional terms may be added to the loss function in Eq. B.3. In order to do so, we use a more general dependence of the PMF on inverse temperature, β, and other parameters such as external fields, λ_a. Suppose then that a thermodynamic property of the system, $\mathcal{P}$, is computed by an operation $\mathcal{K}_p$ acting on the atomistic potential energy times β, integrated over the Boltzmann distribution. The property in question could be, for example, the mean potential energy. In this case, the operator is simply $\mathcal{K}_p = 1/\beta$ and $\mathcal{P}(r) = U(r, \lambda_a)$, so

$$\langle U(r, \lambda_a) \rangle = \langle \mathcal{K}_p[\beta U(r, \lambda_a)] \rangle = \langle \mathcal{P}(r) \rangle$$

An ensemble average of $\mathcal{P}$ can equivalently be taken and related to the PMF by an operator $\mathcal{D}_p$ acting on the true PMF, $\tilde{W}$. Such operations can include higher order mixed derivatives of the PMF with respect to its parameters. In the example of mean potential energy, $\mathcal{D}_p = \partial/\partial\beta$ such that

$$\langle U(r, \lambda_a) \rangle = \frac{\partial(\beta \tilde{W})}{\partial \beta} = \mathcal{D}_p[\beta \tilde{W}]$$

Formally, we write the ensemble average of a property in terms of operators as

$$\langle \mathcal{P}(r) \rangle_R = \Big\langle \mathcal{K}_p[\beta U(r, \lambda_a)] \Big\rangle_R = \mathcal{D}_p[\beta \tilde{W}(R, \beta, \lambda_a)] \tag{B.4}$$

We claim that the following property-matching residual may be additionally minimized to constrain the PMF further, such that our estimate of the PMF, W, produces the same value of the all-atom property $\mathcal{P}$:

$$\chi^2_{PM}[W] = \left\langle \left| \mathcal{P}(r) - \mathcal{D}_p[\beta W(M(r), \beta, \lambda_a)] \right|^2 \right\rangle_r$$

where the brackets indicate a weighted average over the Boltzmann distribution in the full all-atom configuration space. In order to rigorously show this, define $\Delta = \beta W - \beta \tilde{W}$ where $\tilde{W}$ is the true

PMF, W is a model of the PMF:

$$\begin{aligned}\chi^2_{PM}[W] &= \left\langle \left| \mathcal{P}(r) - \mathcal{D}_p[\beta W(M(r), \beta, \lambda_a)] \right|^2 \right\rangle \\ &= \left\langle \left| \mathcal{P}(r) - \mathcal{D}_p[\Delta(M(r)) + \beta \tilde{W}(M(r))] \right|^2 \right\rangle\end{aligned}$$

Here we have dropped the explicit dependence on temperature and parameters for brevity. In the proof that follows, we assume that the operator $\mathcal{D}_p$ is linear, because thermodynamic properties we are concerned with are derivatives of the free energy. Expanding the terms in the ensemble average, we have

$$\begin{aligned}\chi^2_{PM}[W] = \Big\langle & \mathcal{P}^2(r) - 2\mathcal{P}(r)\mathcal{D}_p\big[\Delta(M(r))\big] - 2\mathcal{P}(r)\mathcal{D}_p\big[\beta\tilde{W}(M(r))\big] + \mathcal{D}_p[\Delta(M(r))]^2 \\ & + 2\mathcal{D}_p\big[\Delta(M(r))\big]\mathcal{D}_p\big[\beta\tilde{W}(M(r))\big] + D_p\big[\beta\tilde{W}(M(r))\big]^2 \Big\rangle \end{aligned} \tag{B.5}$$

The minimization is to be done over the parameters of the model, W, and terms that do not depend on these parameters can be treated as constants. In particular, the first, third, and sixth term can be grouped into a constant, C. We then have the simplified expression

$$\chi^2_{PM}[W] = C - \left\langle 2\mathcal{P}(r)\mathcal{D}_p\big[\Delta(M(r))\big] - \mathcal{D}_p[\Delta(M(r))]^2 - 2\mathcal{D}_p\big[\Delta(M(r))\big]\mathcal{D}_p\big[\beta\tilde{W}(M(r))\big] \right\rangle \tag{B.6}$$

We may now use the fact that the surjective function M maps every atomistic point r to some CG coordinate value R. This fact allows us to partition the integration and obtain the following two

results. First,

$$\begin{aligned}\langle \mathcal{P}(r)\mathcal{D}_p[\Delta(M(r))]\rangle &= \frac{1}{Z}\int dr\mathcal{P}(r)\mathcal{D}_p[\Delta(M(r))]e^{-\beta U(r)} \\ &= \frac{1}{Z}\int dR\int dr\mathcal{P}(r)\mathcal{D}_p[\Delta(M(r))]e^{-\beta U(r)}\delta(M(r)-R) \\ &= \frac{1}{Z}\int dR\mathcal{D}_p[\Delta(R)]\int dr\mathcal{P}(r)e^{-\beta U(r)}\delta(M(r)-R) \\ &= \frac{1}{Z}\int dR\mathcal{D}_p[\Delta(R)]Z(R)\langle \mathcal{P}(r)\rangle_R\end{aligned}$$

where we have defined the CG-coordinate dependent partition function:

$$Z(R) = \int dre^{-\beta U(r)}\delta(M(r)-R)$$

such that the full partition function is

$$Z = \int dre^{-\beta U(r)} = \int dR\int dre^{-\beta U(r)}\delta(M(r)-R) = \int dRZ(R)$$

Second, we can see the following:

$$\begin{aligned}\langle \mathcal{D}_p[\beta\tilde{W}]\mathcal{D}_p[\Delta(M(r))]\rangle &= \frac{1}{Z}\int dr\mathcal{D}_p[\beta\tilde{W}(M(r))]\mathcal{D}_p[\Delta(M(r))]e^{-\beta U(r)} \\ &= \frac{1}{Z}\int dR\int dr\mathcal{D}_p[\beta\tilde{W}(M(r))]\mathcal{D}_p[\Delta(M(r))]e^{-\beta U(r)}\delta(M(r)-R) \\ &= \frac{1}{Z}\int dR\mathcal{D}_p[\beta\tilde{W}(R)]\mathcal{D}_p[\Delta(R)]\int dre^{-\beta U(r)}\delta(M(r)-R) \\ &= \frac{1}{Z}\int dR\mathcal{D}_p[\Delta(R)]Z(R)\langle P(r)\rangle_R\end{aligned}$$

where in the last step we have used Eq. B.4. With this result, Eq. B.6 reduces to

$$\chi^2_{PM}[W] = C + \left\langle \mathcal{D}_p[\Delta(M(r))]^2\right\rangle_r \tag{B.7}$$

Because the second term is the only one that can be minimized, and because it is strictly non-negative, it is clear that the minimum of the loss function occurs at

$$\mathcal{D}_p[\Delta(\mathcal{M}(r))] = \mathcal{D}_p[\beta W] - \mathcal{D}_p[\beta \tilde{W}] = 0 \tag{B.8}$$

whose solution is where the target property of the model PMF, W, is equal to that of the true PMF, $\tilde{W}$.

B.4 Model Parameter Optimization - Toy Model

For each model of the toy system, we compute the force MSE over the entire domain of structures. A large grid search is performed, scanning the following values: $n \in [2, 12]$, $\ell \in [2, 6]$, $\xi \in [1, 2]$. During this part of the optimization, the parameters t and $\tilde{T}$ are both fixed to 4. After this stage, the remaining parameters are varied: $t \in [1, 6]$ and $\tilde{T} \in [2, 6]$.

The embedding coefficients are determined by a scan over 100,000 random samplings of the embedding coefficient values, after all other parameters have been determined. The continuous hyperparameters, σ, σ_f and σ_a were optimized periodically through the training process by maximizing the log likelihood of the model.

B.5 Model Parameter Optimization - Hexane

In order to ensure the best model is selected for both the energy labeled and non-energy labeled cases, we do an exhaustive sweep of hyperparemeters in Allegro. In particular, we tune: the batch size, the learning rate, and the relative weights of forces and energies in the loss function while training on 100,000, 200,000 and 300,000 frames of unconstrained data. For each set of hyperparameters, we first evaluate the best model based on the validation loss. The validation loss is computed over 200,000 independent structures. For force-only models, the validation loss is the mean squared residual over the noisy force labels, while for the energy-labeled cases, the validation loss additionally includes the mean squared energy residual term.

Then, the optimal model is equilibrated and run for 2ns over 343 molecules to determine the models sampling of bond lengths, angles, dihedral angles, and the free energy surface defined in the main text. The best performing models based on these metrics are shown in the main text. In Table B.1 we show the results of this parameter search in regards to the different error metrics considered during simulation. The optimal parameters for force-only and energy training differ between each other and between different data amounts.

Data Size	100000	100000	200000	200000	300000	300000
LR	0.001	0.001	0.001	0.001	0.001	0.001
Batch	250	500	570	500	750	500
Energy Weight	0	10	0	1	0	25
Force Weight	10	1	100	1	100	10
Force Loss	36.9164	36.9688	36.8691	36.8673	36.8445	36.8420
Bond MAE	0.0723	0.0030	0.0090	0.0041	0.0129	0.0044
Angle MAE	0.0042	0.0011	0.0005	0.0003	0.0004	0.0002
Dihedral MAE	0.0054	0.0023	0.0028	0.0003	0.0028	0.0007
FES	0.0874	0.0297	0.0353	0.0059	0.0390	0.0103
FES MAE Min	0.0121	0.0058	0.0007	0.0005	0.0008	0.0012
FES MAE Max	0.1970	0.0516	0.0875	0.0156	0.0825	0.0187

Table B.1: The table shows a variety of different error metrics for Allegro models, trained on hexane, both with and without energy labels. Results are given as a function of data size, and the best performing model for each data amount and each model type (energy labeled versus non-energy labeled) are shown. This represents a selection of the parameters that were swept over in model optimization, serving to highlight the differing hyper parameters found for each model. The parameters shown are the learning rate (LR), batch size, energy weight, and force weight in the loss function.

B.6 Training Data Generation

The data for the toy model considered in this work can be generated via numerical integration. Here we use the trapezoid rule techniques to collect force, energy, and free energy information at regular intervals.

For collecting unconstrained hexane data, structures and energies are saved every 100fs for a total of 500,000 structures. Molecular dynamics for this purpose was performed with a Nose Hoover thermostat and 1fs timestep, using the LAMMPS MD code [110]. Data files with OPLS parameters were generated using the LigParGen server [108,109].

B.7 GP Instabilities

In our work we encountered two sources of numerical instability, one of which is illuminated by the toy model. First, the choice of kernel function in GP implementations has implications for the resulting kernel matrices. In the highly simplified example with two CG sites that we examined for the toy model, a normalized dot product kernel

$$k(d_i, d_j) = \sigma\Big(\frac{d_i \cdot d_j}{d_i d_j}\Big)^{\xi} \tag{B.9}$$

leads to a force-energy kernel that is zero for identical descriptors. In this expression, σ is the kernel signal parameter and ξ the power of the dot product. This effectively makes it more challenging for the model to learn. We note that when the normalization is removed, this issue goes away. Therefore, for the toy model we use an un-normalized dot product kernel.

Second, we observe that ML models are prone to exploding when they encounter states of high uncertainty. While this is not an issue for the toy model, whose phase-space is quite small, this becomes problematic for the single-molecule hexane systems considered in this work. Previous active learning efforts have found stability in models of relatively large liquids[91], but the single molecule limit is a special case. Unlike in liquid systems, each frame of training data provides less force information, and it becomes more challenging to capture important transition state information.

To counter the issue of phase space exploration and ensuing instability, we could turn to on-the-fly active learning to generate more diverse and informative data for the hexane molecule. However, we note that this would require the use of a normalized kernel to define a relative uncertainty threshold. It is unclear whether or not the numerical issues that present in the 2 CG site case persist in the 4 CG site case encountered by the hexane models, and if so then to what extent. This motivates future work on better understanding stability. As such, we have focused on the use of neural network

approaches for real systems.

B.8 Model Simulation Details

Simulations of the hexane molecule are performed in LAMMPS via in-house software for the Allegro force field. Appropriate modifications are made to incorporate the temperature parameter. For each model, a 2ns simulation was run after a 200ps equilibration period, using a 0.5 fs timestep. While note strictly necessary, a smaller timestep than the all-atom case was used to ensure stability. A Langevin thermostat at 250K with a damping parameter of 100 femtoseconds was used, and samples of positions were collected every 200 steps. Each model was run with different starting velocities to enhance sample efficiency.

References

[1] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Comput. Mater.*, 6, 2020.

[2] Jonathan Vandermause, Yu Xie, Jin Soo Lim, Cameron J. Owen, and Boris Kozinsky. Active learning of reactive bayesian force fields applied to heterogeneous catalysis dynamics of h/pt. *Nat. Comm.*, 13, 2022.

[3] Yu Xie, Jonathan Vandermause, Lixin Sun, Andrea Cepellotti, and Boris Kozinsky. Fast bayesian force fields from active learning: Study of inter-dimensional transformation of stanene. *Npj Comput. Mater.*, 7, 2021.

[4] J. Dubochet and A. W. McDowall. Vitrification of pure water for electron microscopy. *Journal of Microscopy*, 124, 1981.

[5] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2002.

[6] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 1987.

[7] Å A. Skjevik, Benjamin D. Madej, Callum J. Dickson, Knut Teigen, Ross C. Walker, and Ian R. Gould. All-atom lipid bilayer self-assembly with the amber and charmm lipid force fields. *Chem. Commun.*, 51:4402, 2015.

[8] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.*, 13:1–11, 2022.

[9] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.*, 14:579, 2023.

[10] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104, 2010.

[11] K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller. Schnet - a deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148, 2018.

[12] Yu Xie, Jonathan Vandermause, Senja Ramakers, Nakib H. Protik, Anders Johansson, and Boris Kozinsky. Uncertainty-aware molecular dynamics from bayesian active learning: Phase transformations and thermal transport in sic. *Npj Comput. Mater.*, 9:36, 2023.

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

[14] Cameron J. Owen, Yu Xie, Anders Johansson, Lixin Sun, and Boris Kozinsky. Stability, mechanisms and kinetics of emergence of au surface reconstructions using bayesian force fields. *arXiv preprint arXiv:2308.07311*, 2023.

[15] Cameron J. Owen, Nicholas Marcella, Yu Xie, Jonathan Vandermause, Anatoly I. Frenkel, Ralph G. Nuzzo, and Boris Kozinsky. Unraveling the catalytic effect of hydrogen adsorption on pt nanoparticle shape-change. *arXiv preprint arXiv:2368.00901*, 2023.

[16] Jonathan Vandermause, Anders Johansson, Yucong Miao, Joost J. Vlassak, and Boris Kozinsky. Phase discovery with active learning: Application to structural phase transitions in equiatomic niti. *arXiv preprint arXiv:2401.05568*, 2024.

[17] Georgy Samsonidze and Boris Kozinsky. Accelerated screening of thermoelectric materials by first-principles computations of electron-phonon scattering. *Advanced Energy Materials*, 8:1800246, 2018.

[18] Saeed Arabha and Ali Rajabpour. Thermo-mechanical properties of nitrogenated holey graphene: A comparison of machine-learning-based and classical interatomic potentials. *International Journal of Heat and Mass Transfer*, 178, 2021.

[19] Chaoying Wang, Zhenqing Wang, and Qingyuan Meng. Comparative study of the empirical interatomic potentials and density-functional simulations of divacancy and hexavacancy in silicon. *Physica B: Condensed Matter*, 406, 2011.

[20] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, 2007.

[21] Jiang Wang, Stefan Chmiela, Klaus-Robert Müller, Frank Noè, and Cecilia Clementi. Ensemble learning of coarse-grained molecular dynamics force fields with a kernel approach. *J. Chem. Phys*, 152:194106, 2020.

[22] Ralf Drautz. Atomic cluster expansion for accurate and transferable iteratomic potentials. *Phys. Rev. B*, 99:014104, 2019.

[23] Anders Johansson, Yu Xie, Cameron J. Owen, (David) Lim Jin Soo, Lixin Sun, Jonathan Vandermause, and Boris Kozinsky. Micron-scale heterogeneous catalysis with bayesian force fields from first principles and active learning. *Preprint at https://arxiv.org/abs/2204.12573*, 2022.

[24] Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. De Vries. The martini force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem.*, 111:7812–7824, 2007.

[25] Cesar A. López, Andrzej J. Rzepiela, Alex H. de Vries, Lubbert Dijkhuizen, Philippe H. H‸unenberger, and Siewert J. Marrink. Martini coarse-grained force field: Extension to carbohydrates. *J. Chem. Theory Comput.*, 5:3195–3210, 2009.

[26] Djurre H. de Jong, Gurpreet Singh, W. F. Drew Bennett, Clement Arnarez, Tsjerk A. Wassenaar, Lars V. Sch‸afer, Xavier Periole, D. Peter Tieleman, and Siewert J. Marrink. Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.*, 9:687–697, 2013.

[27] Semen O. Yesylevskyy, Lars V. Sch'afer, Durba Sengupta, and Siewert J. Marrink. Polarizable water model for the coarse-grained martini force field. *PLoS Comput. Biol.*, 6, 2010.

[28] Luca Monticelli, Senthil K. Kandasamy, Xavier Periole, Ronald G. Larson, D. Peter Tieleman, and Siewert-Jan Marrink. The martini coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.

[29] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Adv. Prot. Chem.*, 66:27–85, 2003.

[30] Andrzej Koliński. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, 51:349–71, 2004.

[31] D. L. Beveridge and W. L. Jorgensen. The opls potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. *Annu. Rev. Biophys. Bioeng.*, 110:18, 1988.

[32] W. G. Noid, Jhih-Wei Chu, Gary S. Ayton, Vinod Krishna, Sergei Izvekov, Gregory A. Voth, Avisek Das, and Hans C. Andersen. The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128:244114, 2008.

[33] M. Scott Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.*, 129:144108, 2008.

[34] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chem. Rev.*, 116:7898–7936, 2016.

[35] Carol A. Rohl, Charlie E. M. Strauss, Kira M. S. Misura, and David Baker. Protein structure prediction using rosetta. *Num. Comp. Methods*, 383:66–93, 2004.

[36] D. Reith, M. Putz, and F. Muller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *J Comput. Chem.*, 24:1624–1636, 2003.

[37] Timothy C. Moore, Christopher R. Iacovella, and Clare McCabe. Derivation of coarse-grained potentials via multistate iterative boltzmann inversion. *J. Phys. Chem.*, 140:224104, 2014.

[38] Sun Young Woo and Hwankyu Lee. All-atom simulations and free energy calculations of coiled-coil peptides with lipid bilayers: binding strength, structural transition, and effect on lipid dynamics. *Scientific Reports*, 6:22299, 2016.

[39] M. Bonomi, A. Barducci, and M. Parrinello. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *J. Comput. Chem.*, 30(11):1615–1621, 2009. 00220.

[40] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, 1977.

[41] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(5):826–843, September 2011.

[42] Marc Souaille and Benoıt Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.*, 135(1):40–57, 2001.

[43] ŁL Baran and W Rżysko. Application of a coarse-grained model for the design of complex supramolecular networks. *Mol. Syst. Des. Eng.,*, 5:484–492, 2020.

[44] Marcus G. Martin and J. Ilja Siepmann. Transferable potentials for phase equilibria. 1. united-atom description of n-alkanes. *J. Phys. Chem. B*, 102:2569–2577, 1998.

[45] Adam Liwo, Maciej Baranowski, Cezary Czaplewski, Ewa Gołaś, Yi He, Dawid Jagieła, Paweł Krupa, Maciej Maciejczyk, Mariusz Makowski, Magdalena A Mozolewska, Andrei Niadzvedtski, Stanisław Ołdziej, Harold A Scheraga, Adam K Sieradzan, Rafał Ślusarz, Tomasz Wirecki, Yanping Yin, and Bartłomiej Zaborowski. A unified coarse-grained model of biological macromolecules based on mean-field multipole–multipole interactions. *J Mol. Model.*, 20, 2014.

[46] W. G. Noid. Perspective: coarse-grained models for biomolecular systems. *J. Chem. Phys.*, 139:90901, 2013.

[47] Yaoyi Chen, Andreas Krämer, Nicholas E. Charron, Brooke E. Husic, Cecilia Clementi, and Frank Noé. Machine learning implicit solvation for molecular dynamics. *J. Chem. Phys.*, 155:084101, 2021.

[48] Biswaroop Mukherjee, Luigi Delle Site, Kurt Kremer, and Christine Peter. Derivation of coarse grained models for multiscale simulation of liquid crystalline phase transitions. *J. Phys. Chem. B*, 116, 2012.

[49] Anton V Sinitskiy and Gregory A Voth. Quantum mechanics/coarse-grained molecular mechanics (QM/CG-MM). *J. Chem. Phys.*, 148:014102, 2018.

[50] Alexander V Mironenko and Gregory A Voth. Density functional theory-based quantum mechanics/coarse-grained molecular mechanics: Theory and implementation. *J. Chem. Theory Comput.*, 16:6329–6342, 2020.

[51] Ronald D. Hills Jr., Lanyuan Lu, and Gregory A. Voth. Multiscale coarse-graining of the protein energy landscape. *PLoS Comput. Biol.*, 6, 2010.

[52] W. G. Noid, Jhih-Wei Chu, Gary S. Ayton, and Gregory A. Voth. Multiscale coarse-graining and structural correlations: Connections to liquid-state theory. *J. Phys. Chem. B*, 111:4116–4127, 2007.

[53] Luca Larini, Lanyuan Lu, and Gregory A. Voth. The multiscale coarse-graining method. vi. implementation of three-body coarse-grained potentials. *J. Chem. Phys.*, 132:164107, 2010.

[54] Segei Izvekov and Gregory A. Voth. Multiscale coarse graining of liquid-state systems. *J. Chem. Phys.*, 123:134105, 2005.

[55] Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charron, Gianni De Fabritiis, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.*, 5:755–767, 2019.

[56] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deepcg: Constructing coarse-grained models via deep neural networks. *J. Chem. Phys.*, 149:034101, 2018.

[57] Jurgis Ruza, Wujie Wang, Daniel Schwalbe-Koda, Simon Axelrod, William H. Harris, and Rafael Gómez-Bombarelli. Temperature-transferable coarse-graining of ionic liquids with dual graph convolutional neural networks. *J. Chem. Phys.*, 153:164501, 2020.

[58] Brooke E. Husic, Nicholas E. Charron, Dominik Lemm, Jiang Wang, Adrià Pérez, Andreas Krämer, Yaoyi Chen, Simon Olsson, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.*, 153:194101, 2020.

[59] Wujie Wang and Rafael Gómez-Bombarelli. Coarse-graining auto-encoders for molecular dynamics. *Npj Comput. Mater.*, 5, 2019.

[60] S. T. John and Gábor Csányi. Many-body coarse-grained interactions using gaussian approximation potentials. *J. Phys. Chem. B*, 121:10934–10949, 2017.

[61] Christoph Scherer, René Scheid, Denis Andrienko, and Tristan Bereau. Kernel-based machine learning for efficient simulations of molecular liquids. *J. Chem. Theory Comput.*, 16:3194–3204, 2020.

[62] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134, 2011.

[63] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87, 2013.

[64] Zhenwei Li, James R. Kermode, and Alessandro De Vita. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.*, 114, 2015.

[65] Aldo Glielmo, Peter Sollich, and Alessandro De Vita. Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B*, 95, 2017.

[66] Aldo Glielmo, Claudio Zeni, and Alessandro De Vita. Efficient nonparametric n -body force fields from machine learning. *Phys. Rev. B*, 97, 2018.

[67] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104(13):136403, 2010.

[68] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):184115, 2013.

[69] Albert P Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.*, 115(16):1051–1057, 2015.

[70] Junhui Peng, Chuang Yuan, Rongsheng Ma, and Zhiyong Zhang. Backmapping from multiresolution coarse-grained models to atomic structures of large biomolecules by restrained molecular dynamics simulations using bayesian inference. *J. Chem. Theory Comput.*, 15:3344–3353, 2019.

[71] Wei Li, Craig Burkhart, Patrycja Polińska, Vagelis Harmandaris, and Manolis Doxastakis. Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. *J. Chem. Phys.*, 153, 2020.

[72] Wujie Wang, Minkai Xu, Chen Cai, Benjamin Kurt Miller, Tess Smidt, Yusu Wang, Jian Tang, and Rafael Gómez-Bombarelli. Generative coarse-graining of molecular conformations. *Preprint at http://arXiv.org/abs/2201.12176*, 2022.

[73] Troy D. Loeffler, Tarak K. Patra, Henry Chan, and Subramanian K. R. S. Sankaranarayanan. Active learning a coarse-grained neural network model for bulk water from sparse training data. *Mol. Syst. Des. Eng.*, 5:902–910, 2020.

[74] Joaquin Candela-Quiñonero and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, 2005.

[75] Li-Jun Chen, Hu-Jun Qian, Zhong-Yuan Lu, Ze-Sheng Li, and Chia-Chung Sun. An automatic coarse-graining and fine-graining simulation method: Application on polyethylene. *J. Phys. Chem. B*, 110:24093–24100, 2006.

[76] Simon Hunkler, Tobias Lemke, Christine Peter, and Oleksandra Kukharenko. Backmapping based sampling: Coarse grained free energy landscapes as a guideline for atomistic exploration. *J. Chem. Phys.*, 151, 2019.

[77] J. W. Mullinax and W. G. Noid. A generalized-yvon-born-green theory for determining coarse-grained interaction potentials. *J. Phys. Chem. C*, 114:5661–5674, 2010.

[78] Joseph F. Rudzinski and W. G. Noid. Investigation of coarse-grained mappings via an iterative generalized yvon-born-green method. *J. Phys. Chem. B*, 118:8295–8312, 2014.

[79] Christoph Scherer and Denis Andrienko. Understanding three-body contriutions to coarse-grained force fields. *Phys. Chem. Chem. Phys.*, 20:22387–22394, 2018.

[80] Lanyuan Lu, James F. Dama, and Gregory A. Voth. Fitting coarse-grained distribution functions through an iterative force-matching method. *J. Chem. Phys.*, 139:121906, 2013.

[81] Joseph F. Rudzinski and W. G. Noid. The role of many-body correlations in determining potentials for coarse-grained models of equilibrium structure. *J. Phys. Chem. B*, 116:8621–8635, 2018.

[82] Jin Soo Lim, Jonathan Vandermause, Matthijs A. van Spronsen, Albert Musaelian, Yu Xie, Lixin Sun, Christopher R. O'Connor, TObias Egle, Nicola Molinari, Jcaob Florian, Kaining Duanmu, Robert J. Madix, Philippe Sautet, Cynthia M. Friend, and Boris Kozinsky. Evolution of metastable structures at bimetallic surfaces from microscopy and machine-learning molecular dynamics. *J. Am. Chem. Soc.*, 142:15907–15916, 2020.

[83] Cameron J. Owen, Nickolas Marcella, Jonathan Vandermause, Anatoly I. Frenkel, Ralph G. Nuzzo, and Boris Kozinsky. Unraveling the catalytic effect of hydrogen adsorption on pt nanoparticle shape-change. *arXiv preprint arXiv:2306.00901*, 2023.

[84] Cameron J. Owen, Yu Xie, Anders Johansson, Lixin Sun, and Boris Kozinsky. Stability, mechanisms and kinetics of emergence of au surface reconstructions using bayesian force fields. *arXiv preprint arXiv:2308.07311*, 2023.

[85] Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, , and Klaus-Robert Müller. Machine learning force fields. *Chem. Rev.*, 121:10142–10186, 2021.

[86] Shiru Wu, Xiaowei Yang, Xun Zhao, Zhipu Li, Min Lu, Xiaoji Xie, and Jiaxu Yan. Applications and advances in machine learning force fields. *J. Chem. Inf. Model.*, 63:6972–6985, 2023.

[87] Frank Noe, Gianni De Fabritiis, and Cecilia Clementi. Machine learning for protein folding and dynamics. *Current Opinion in Structural Biology*, 60:77–84, 2020.

[88] Silvan Kaser, Oliver T. Unke, and Markus Meuwly. Reactive dynamics and spectroscopy of hydrogen transfer from neural network-based reactive potential energy surfaces. *New J. Phys.*, 22:055002, 2020.

[89] Weile Jia, Han Wang, Mohan Chen, Denghui Lu, Lin Lin, Roberto Car, Weinan E, and Linfeng Zhang. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press, 2020.

[90] Stephan Thaler, Maximilian Stupp, and Julija Zavadlav. Deep coarse-grained potentials via relative entropy minimization. *J. Chem. Phys*, 157, 2022.

[91] Blake R. Duschatko, Jonathan Vandermause, Nicola Molinari, and Boris Kozinsky. Uncertainty driven active learning of coarse grained free energy models. *npj Comput. Mater*, 10:9, 2024.

[92] Jaehyeok Jin, Alexander J. Pak, Aleksander E. P. Durumeric, Timothy D. Loose, and Gregory A. Voth. Bottom-up coarse-graining: Principles and perspectives. *J. Chem. Phys.*, 18:5759–5791, 2022.

[93] Kathryn M. Lebold and W. G. Noid. Dual-potential approach for coarse-grained implicit solvent models with accurate, internally consistent energetics and predictive transferability. *J. Chem. Phys.*, 151, 2019.

[94] Avisek Das and Hans C. Anderson. The multiscale coarse-grianing method. v. isothermal-isobaric ensemble. *J. Chem. Phys*, 132:164106, 2010.

[95] Nicholas J. H. Dunn and W. G. Noid. Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids. *J. Chem. Phys*, 143:24318, 2015.

[96] Stefano Falletta, Andrea Cepellotti, Chuin Wei Tan, Anders Johansson, Albert Musaelian, Cameron J. Owen, and Boris Kozinsky. Unified differentiable learning of the electric enthalpy and dielectric properties with exact physical constraints. *arXiv preprint arXiv:2403.17207*, 2024.

[97] Michael Gastegger, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of solvent effects on molecular spectra and reactions. *Chem. Sci.*, 12:11473–11483, 2021.

[98] Hwijae Son, Jin Woo Jang, Woo Jin Han, and Hyung Ju Hwang. Sobolev training for physics-informed neural networks. *Commun. Math. Sci.*, 21:1679–1705, 2023.

[99] Nikolaos N. Vlassis and WaiChing Sun. Sobolev training of thermodynamic-informed neural networks for interpretable elasto-plasticity models with level set hardening. *Computer Methods in Applied Mechanics and Engineering*, 377:113695, 2021.

[100] Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Neural Information Processing Systems*, 2017.

[101] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science, 2006.

[102] Timothy D. Loose, Patrick G. Sahrmann, Thomas S. Qu, and Gregory A. Voth. Coarse-graining with equivariant neural networks: A path toward accurate and data-efficient models. *J. Phys. Che. B*, 127:10564–10572, 2023.

[103] Giulio Imbalzano, Yongbin Zhuang, Venkat Kapil, Kevin Rossi, Edgar A. Engel, Federico Grasselli, and Michele Ceriotti. Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.*, 154:074102, 2021.

[104] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

[105] L. Kaufman and H. Bernstein. *Computer calculation of phase diagrams*. Academic Press Inc, 1970.

[106] Davud Rosenberger, Kipton Barros, Timothy C. Germann, and Nicholas Lubbers. Machine learning of consistent thermodynamic models using automatic differentiation. *Phys. Rev. E*, 105:045301, 2022.

[107] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dulak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment - a python library for working with atoms. *J. Condens. Matter Phys.*, 29, 2017.

[108] L.S Dodda, J. Z. Vilseck, J. Rives-Tirado, and W. L. Jorgensen. 1.14*cm1a-lbcc: Localized bond-charge corrected cm1a charges for condensed-phase simulations. *J. Phys. Chem. B*, 121:3864–3870, 2017.

[109] L. S. Dodda, I. Cabeza de Vaca, J. Rives-Tirado, and W. L. Jorgensen. Ligpargen web server: An automatic opls-aa parameter generator for organic ligands. *Nucleic Acids Research*, 45:W331–336, 2017.

[110] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.*, 117:1–19, 1995.

[111] J. A. Racker, Z. Wang, C. Lu, M. L. Laury, L. Lagardere, M. J. Schnieders, J-P Piquemal, P. Ren, and J. W. Ponder. Tinker 8: Software tools for molecular design. *J. Chem. Theory Comput.*, 14:5273–5289, 2018.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.