TECHNISCHE UNIVERSITÄT MÜNCHEN

PROFESSUR FÜR KONTINUUMSMECHANIK

Probabilistic Machine Learning Strategies for Coarse-Graining of Molecular Dynamics at Equilibrium

Markus Josef Johann SCHÖBERL

Vollständiger Abdruck der von der Fakultät für Maschinenwesen der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzende:	Prof. Dr. Julija ZAVADLAV
Prüfer der Dissertation:	1. Prof. Phaedon-Stelios KOUTSOURELAKIS, Ph.D.
	2. Prof. Nicholas ZABARAS, Ph.D.

Die Dissertation wurde am 13.02.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 24.07.2020 angenommen.

© Markus Schöberl

"What I cannot create, I do not understand."

Richard Feynman (1918 – 1988)

Abstract

Advances in biomolecular processes, materials science, and nanotechnology are hindered because the scales that are used in atomistic systems are not parallel. Resolving local oscillations in robust all-atom molecular dynamics simulations requires time steps in the order of femtoseconds (1.0×10^{-15} s), while relevant biochemical processes take place on timescales that exceed several milliseconds (1.0×10^{-3} s).

This discrepancy – of more than 12 orders of magnitude between the simulation time horizons and the required molecular dynamics time-step resolution (femtoseconds) – results in a prohibitive number of simulation steps. Even so, prevalent spatiotemporal limitations can be overcome by using simulation tools that overarch multiple scales.

In the context of equilibrium statistical mechanics, this project involved developing data-driven and variational coarse-graining approaches based on an atomistic scale. Ultimately, we offer a novel approach to mapping between scales. While existing methodologies rely on many-to-one, fine-to-coarse mappings (e.g., defined by summarizing atoms to macromolecules), we introduced a probabilistic coarse-to-fine map. This approach corresponds to a directed probabilistic graphical model, wherein coarse-grained variables are implicitly defined by the introduced probabilistic coarse-to-fine map. Hence, the coarse-grained variables, which are latent variables, serve as generators of fully atomistic representations.

Essentially, we reformulated the approaches to a likelihood-based maximization problem that is embedded in a consistent Bayesian framework. This approach allowed for the reconstruction of a fully atomistic scale, which enabled estimations of macroscopic observables that are governed by fine-scale interdependencies. Further, this allowed for the determination of posterior distributions of model parameters, which express uncertainties due to limited training data. The prevalent uncertainties were propagated to a predictive posterior distribution over relevant quantities. More broadly, the predictive distributions reflect the credibility of the coarse-grained model and quantify uncertainties in the available data. In the end, we were left to either focus exclusively on low amounts of training data (50–1000) or completely circumvent the production of training data by adopting variational approaches.

In addition to developing predictive machine-learned coarse-graining frameworks, we sought to discover physical insights on the absence of any system-dependent knowledge. One component of this work focuses on obtaining sparse, physically interpretable solutions for the interactions of coarse-grained variables. Another seeks to reveal parsimonious lower-dimensional representations through expediting the discovery of collective variables in the reference system.

Overall, the sparse learning methods provide robust and interpretable models that can be readily generalized for further unsupervised learning problems. The framework's capabilities are demonstrated through the coarse-graining of physically-relevant reference systems (i.e., the Ising model, SPC/E water, alanine dipeptide, and ALA-15).

Zusammenfassung

Der antiparallele Verlauf inhärenter Zeit- und Längenskalen atomistischer Systeme beeinträchtigt die Erforschung von biomolekularen Prozessen und Fortschritten in Materialwissenschaft und Nanotechnologie. Während biochemische Prozesse einen Zeithorizont von einigen Millisekunden (1.0×10^{-3} s) einnehmen, benötigen Molekulardynamik-Simulationen zur Auflösung lokaler Oszillationen einen Zeitschritt von Femtosekunden (1.0×10^{-15} s).

Eine Diskrepanz von mehr als zwölf Größenordnungen zwischen relevanten Simulationszeithorizonten und der erforderlichen Zeitschrittauflösung führt zu einer unerschwinglichen Anzahl von Simulationsschritten. Simulationstools, die mehrere Skalen überbrücken, können jedoch weit verbreitete raumzeitliche Einschränkungen effizient überwinden.

Diese Arbeit adressiert die Entwicklung datengetriebener sowie variationeller Multiskalenmodelle, basierend auf sich im thermodynamischen Gleichgewicht befindenden atomistischen Systemen. Die grundlegende Neuheit des entwickelten Ansatzes liegt in einer neuen Perspektive der Verknüpfung von involvierten Skalen. Existierende Methoden basieren auf einer Vorschrift, die ausgehend von einer atomistischen Beschreibung auf eine vergröberte, dimensionsreduzierte Beschreibung schließt (Fein-zu-grob-Ansatz). Dies geschieht beispielsweise durch das Zusammenfassen mehrerer Atome zu Makromolekülen. Wir verfolgen hingegen in dieser Arbeit eine entgegengesetzte Perspektive bezüglich einer mathematischen Verknüpfung involvierter Skalen. Wir führen ein probabilistisches Mapping ein, das ausgehend von einer vergröberten Darstellung (Makromoleküle) auf die ursprüngliche atomistische Beschreibung schließen lässt (Grob-zu-fein-Ansatz). Die entwickelte Methode kann als gerichtetes probabilistisches grafisches Modell interpretiert werden, wobei das eingeführte Grob-zu-fein-Mapping die grobkörnige Beschreibung impliziert. Koordinaten der grobkörnigen Beschreibung sind latente Variablen und dienen zur Generierung der feinskaligen atomistischen Beschreibung.

Die im Kontext atomistischer Systeme entwickelten Dimensionsreduktionsansätze haben eine informationstheoretische Untermauerung. Wir betrachten diese Ansätze aus einem wahrscheinlichkeitstheoretischen Blickwinkel und entwickeln ein konsistentes bayessches Rahmenwerk. Die entwickelte Methode ermöglicht eine probabilistische Rekonstruktion der atomistischen Beschreibung, ausgehend von der grobkörnigen Darstellung. Die atomistische Auflösung erlaubt wiederum eine Vorhersage makroskopischer Eigenschaften, die von gegenseitigen feinskaligen Abhängigkeiten bestimmt sind. Darüber hinaus gestattet ein bayesscher Blickwinkel

die Bestimmung von A-posteriori-Wahrscheinlichkeitsverteilungen der Modellparameter. Diese Verteilungen quantifizieren die Unsicherheit involvierter Modellparameter aufgrund begrenzter Trainingsdaten. Wir propagieren Parameterunsicherheiten in Bezug auf eine prädiktive Wahrscheinlichkeitsverteilung relevanter atomistischer Eigenschaften und berechnen Glaubwürdigkeitsintervalle, die den Erkenntniswert, basierend auf einer limitierten Datenmenge, quantifizieren. Wir entwickeln in dieser Arbeit robuste maschinelle Lernverfahren, die trotz geringer Datenmengen (50–1.000) einen zuverlässigen Erkenntnisgewinn ermöglichen.

Mittels entwickelter prädiktiver und maschinell gelernter Multiskalenmodelle decken wir physikalisch relevante Einblicke atomistischer Komplexe ohne systemspezifisches Vorwissen auf. Ein Teil dieser Arbeit forciert physikalisch interpretierbare makromolekulare Interaktionspotentiale, die mit möglichst wenigen Modellparametern prädiktiv sind. Darüber hinaus zeigen wir Methoden zur Identifikation physikalisch essentieller Koordinaten, die zur beschleunigten Exploration komplexer atomistischer Systeme beitragen können.

Entwickelte maschinelle Lernalgorithmen führen zu robusten und interpretierbaren Modellen, die im Allgemeinen für weitere Problemstellungen leicht verallgemeinert werden können. Wir demonstrieren die Leistungsfähigkeit des entwickelten Ansatzes anhand physikalisch relevanter Referenzsysteme (d. h. das Ising-Modell, SPC/E-Wassermodell, Alanin-Dipeptid und ALA-15).

Acknowledgements

I am grateful to have been part of an active and open-minded scientific community while working in predictive multiscale modeling and uncertainty quantification. Working with thoughtful and supportive advisors and esteemed colleagues in environments that were free of hierarchy enabled me to build my training and capacity in the field.

First and foremost, I am indebted to my advisers, Nicholas Zabaras and Phaedon-Stelios Koutsourelakis for their commitment, time, insights, and intellectual charity. It has been an honor and a great fortune to conduct research alongside you both, not to mention presenting our work at international conferences and publishing in international journals. Through your joint efforts, you helped me to grow throughout my training and struck the perfect balance between challenging me and empowering me to succeed on my own. Particularly in the early stages of my PhD, I benefitted from Phaedon-Stelios Koutsourelakis' constructive research guidance and careful supervision. Above and beyond his role as my adviser, I want to thank Nicholas Zabaras for mentoring me and enabling me to build relationships with a number of scientists and industry representatives.

I also want to recognize Phaedon-Stelios Koutsourelakis and Isabell Franck for bringing this interdisciplinary field to my attention while I was completing my master's thesis.

In terms of my formal training, I am grateful for my research placement at the Center for Informatics and Computational Science at the University of Notre Dame. In addition to being immersed in the latest research on probabilistic machine learning and leveraging physical models, I had the opportunity to exchange and cross-pollinate ideas with many visiting scientists who delivered thoughtful and timely presentations. Specifically, my discussions with Ilias Bilionis, Markos Katsoulakis, Jesper Kristensen, and Ben Leimkuhler enriched my views.

I would also like to take this opportunity to express my deepest thanks to my friends and colleagues affiliated with the Professorship for Continuum Mechanics – TUM, particularly Isabell Franck, Constantin Grigo, Maximilian Rixner, Sigrid Harnauer, Sebastian Kaltenbach, Luca Beradocco, Jonas Nitzler, and Atul Agrawal. Likewise, I am grateful to my colleagues and friends at the Center for Informatics and Computational Science at the University of Notre Dame, including Nicholas Geneva, Yinhao Zhu, Steven Atkinson, Shaoxing Mo, Souvik Chakraborty, Sina Malakpour Estalaki, Yingzhi Xia, Navid Shervanitabar, and Govinda Anantha Padmanabha. I have also benefitted tremendously from interactions and exchanges with my colleagues and friends at the Warwick Centre for Predictive Modelling (UK).

More broadly, I want to recognize the Institute for Advanced Study at the Technical University of Munich and the associated financial support I received from the European Commission. I am also thankful for the support I received from the FGZ-MW Team.

With gratitude, I acknowledge both the mentorship and financial support I received from the Hanns-Seidel-Foundation (funded by the German Federal Ministry of Education and Research) and am obliged to the heads of the scholarship system, Hans-Peter Niedermeier and Andreas Burtscheidt. I would be remiss if I did not mention the wonderful colleagues I met through the mind-broadening seminars and our invaluable discussions.

Finally, I thank my friends for their companionship, and, above all, I am grateful to my family for enabling me to advance along this scientific journey through their unconditional support.

Contents

Chapter 1

Introduction

In this thesis, we focus on the development of methodologies for overcoming the spatial- and temporal-scale limitations inherent in direct atomistic simulation techniques while still being able to reason about the fully resolved scale and thus, enable the estimation of properties that depend on the fine resolution. We propose coarse-graining (CG) approaches to learning a parsimonious lower-dimensional representation without assuming system-specific knowledge, which can still reveal physical insights. Further, we develop methodologies favoring sparse models that facilitate robust machine learning in the low-data regime. The approaches presented follow a fully Bayesian framework to produce probabilistic estimates that account for epistemic uncertainty.

This chapter outlines the motivation for covering an extended range of spatial and temporal scales compared to brute-force molecular dynamics (MD) simulations. In this work, the atomistic resolution depicts the scale of origin and we focus on identifying predictive CG descriptions. The meaning of the term "predictive" is specified in this introductory chapter along with an overview of various simulation techniques for scales relevant in modeling materials. We summarize existing methods and point out their advantages and limitations.

While writing this work, we have found several different approaches for extending the spatial and temporal scales based on the atomistic scale. We categorize the methodologies as (i) data-driven approaches, (ii) enhanced sampling approaches, and (iii) variational approaches, which we all introduce in this work. However, the focus is on the development of predictive data-driven CG methodologies that unveil physical insights with limited data. Two minor sections address collective variable discovery for enhanced sampling and strategies that fully embed the available physics without producing any atomistic data beforehand.

We provide an outline of the thesis at the end of this introductory chapter.

1.1 Background and context

The macroscopic properties of materials or biochemical systems are evoked by their microscopic behavior [1–3]. Microscopic insights are, thus, essential for understanding the overarching physicochemical processes in multiple interdigitating scientific

disciplines. Nowadays, atomistic simulations are well supported and routinely applied in materials science [4, 5], chemical physics, biochemistry, biophysics, and pharmacology [6–9].

The foundations of MD simulations were developed by Adler and Wainwright [10, 11] in the 1950s when they were studying the collisions of 32 hard spheres. Later, in the 1960s, Rahman [12] carried out an MD simulation of 864 argon atoms in the liquid phase to study structural correlations using one of the first high-performance computers, the IBM 704, achieving 40 kFLOPS. Continuing progress in computer architecture and processor technology has lead to compute farms with up to approximately 200 000 TFLOPS [13]. Despite the availability of petascale computing resources utilizing GPUs and extensively parallelized implementations, the discrepancy between the practically accessible spatial and temporal scales compared to those relevant for understanding biochemical processes remains significant.

1.1.1 Limitations of all-atom MD simulations

Robust all-atom MD simulations require time steps of the order of femtoseconds (1.0×10^{-15} s) for resolving local oscillations, i.e., bonded interactions [14, 15]. Unlike the time steps in MD simulations, interesting biochemical processes, e.g., conformational changes of proteins (relevant for understanding diseases like HIV [16]) and protein-folding processes [17–20], take place on timescales exceeding several milliseconds (1.0×10^{-3} s). The discrepancy of more than 12 orders of magnitude between relevant simulation periods (milliseconds) and the required MD time-step resolution (femtoseconds) results in a prohibitive number of simulation steps of more than one trillion [20].

However, simulating events that span milliseconds is key to understanding diseases due to protein misfolding [21], such as type-2 diabetes [22, 23], Alzheimer's [24, 25], and Parkinson's [26, 27]. Gaining an extensive microscopic understanding of the mechanisms causing such diseases could tremendously accelerate the development of effective medications and treatments [28]. The significance of reducing the computational cost for research into inherently large biological systems is emphasized in Figure 1.1, which compares the required wall clock time with the physical simulation time. The figure relies on findings in [29]. The antiparallel nature of scales is not only hampering the exploration of biomolecular processes but also progress in materials science and nanotechnology [30–32]. MD simulations are an effective way to conduct in silico experiments in the development of novel materials [33, 34], e.g., for investigating crack propagation in complex material compounds [35, 36]. Physical timescales that exceed the practically available computational resources are, likewise, prevalent in the simulation of alloy-hardening processes [37]. Thus, by overcoming the spatial limitations, atomistic and macroscopic simulations could become a better computational microscope [20, 38], enabling the rapid exploration of material properties under various mechanical and thermodynamic conditions.

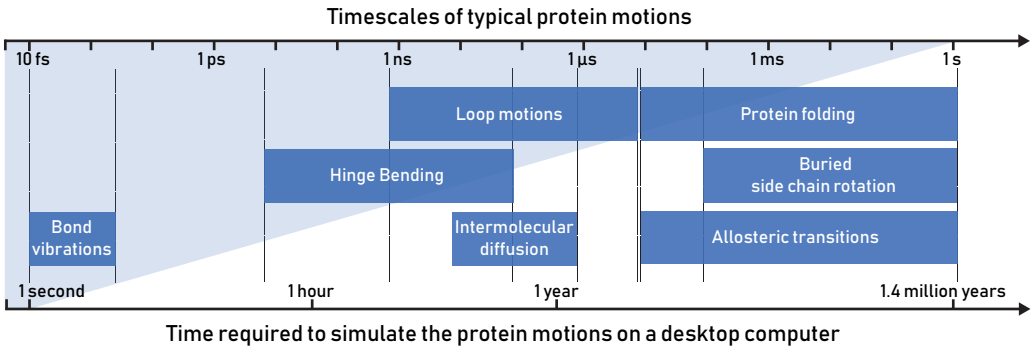


FIGURE 1.1: Approximate computation time (wall clock time) versus the actual physical simulation time indicated on the timeline for a typical protein–protein system solvated in explicit water. The shortest events, i.e., bonded interactions, require femtosecond time steps that are equivalent to milliseconds of physical simulation time. Figure inspired by [29].

Algorithmic advances and approximations, like Ewald summation [39] and thresholding the influence of interactions up to a certain cutoff radius [14], can reduce the computational cost per step from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$, while leaving the number of steps required unchanged. Here, n is the number of atoms considered in the system. The increasing availability of computational resources and recent advances in utilizing GPUs accompanied by extensively parallelized atomistic simulation algorithms [40–44] foster simulations by decomposing the simulation domain into local regions distributed to multiple processors. Computations are performed spatially in parallel [45], not parallel in time. The parallel in time integration of ordinary differential equations (ODEs) is an active area of research. For an overview, see [46, 47]. Parallel time-integration involves the reformulation as an optimization problem over several time steps. The main focus has been on expediting MD simulations in condensed matter physics.

Beyond the algorithmic progress, are hardware improvements for CPUs and GPUs mostly driven by increasing transistor densities. Increased densities are beneficial for improving performance for multiple parallel tasks, while the clock rate stagnates, where increasing clock rates would be useful for MD simulations [20, 48].

The computational challenges due to the broad range of scales involved in biochemical and solid-state systems have led to the development of advanced simulation techniques, hierarchically targeting different scales of interest. At the bottom of the hierarchy concerning different simulation approaches, depicted in Figure 1.2, strategies resolve quantum mechanical effects and electron structure to provide an accurate but computationally intensive description of interactions between atoms by solving the Schrödinger equation. The primary domain of ab initio methods are systems with a few atoms over a timescale of picoseconds [49].

Unlike ab initio methods, extended spatial and temporal scales are accessible with atomistic descriptions, such as MD simulations, which sought to solve ODEs

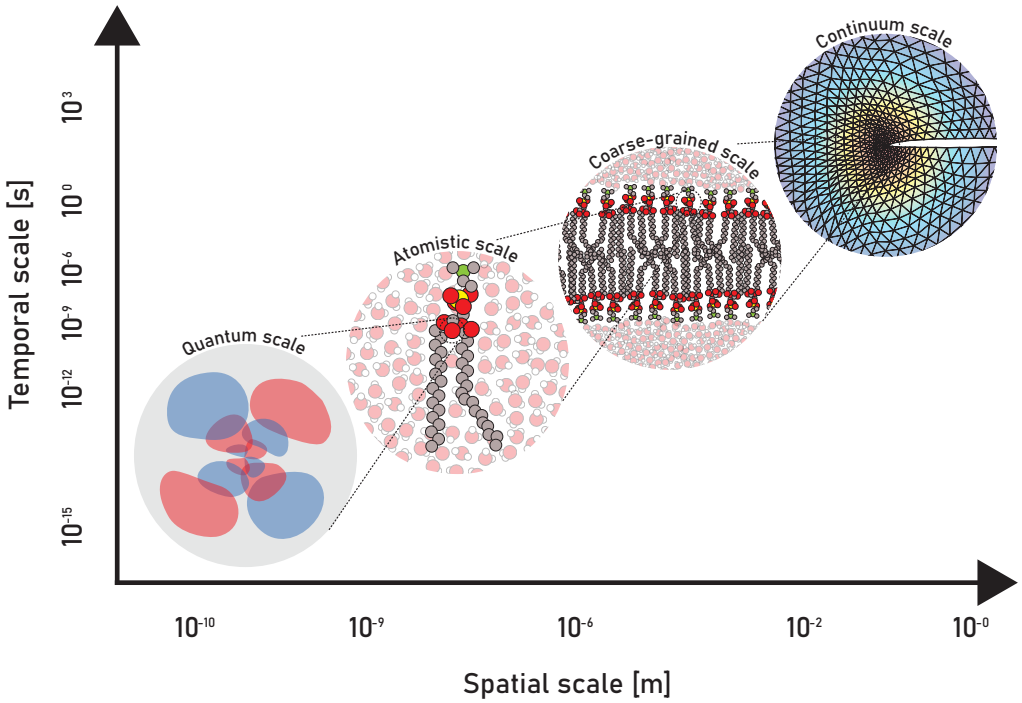


FIGURE 1.2: The figure shows practical simulation techniques for multiple temporal and spatial scales. However, there is a fuzzy transition between scales. Figure is inspired by [54, 55].

based on classical mechanics and empirical force fields. By ignoring the atomistic details, CG models summarize the essential collective particle motion [50, 51]. Macroscopic continuum approaches at the top of the hierarchy in Figure 1.2 capture scales up to hours and meters, depending on the spatial and temporal discretization and model complexity, in terms of higher-order test or shape functions [52, 53].

Before focusing on the reference simulation techniques and MD simulations used in this work, and their limitations, we provide a brief overview of different simulation resolutions with potential application domains in the multiscale paradigm.

1.1.2 Computational methods in materials modeling

This section provides a brief overview of simulation techniques for specific application domains with regards to spatial and temporal scales.

Density functional theory

Ab initio methods considering electron structure are not directly the focus of this work; however, they are useful tools for estimating the atomistic force fields required for MD simulations. We provide a brief introduction to one of the most popular approaches for obtaining approximate solutions, which is used instead of solving the Schrödinger equation [56].

Density functional theory (DFT) goes back to the work of Thomas [57] and Fermi [58] in the 1920s, who independently proposed an expression for the electron density in an external field. Hohenberg and Kohn [59, 60] proved that the ground-state energy of an atomistic system is fully defined by its ground-state electron density, which uniquely minimizes the energy functional. Refs. [61] and [62] provide a comprehensive overview of the development of DFT methods and their various characteristics.

The computational efficiency of the density functional approach is relatively moderate compared to other simulation methods that resolve quantum effects. DFT simulations achieve spatial and temporal scales of up to a few nanometers (1.0×10^{-9} m) and picoseconds (1.0×10^{-12} s) [49, 63, 64]. Although ab initio calculations are currently not appropriate for addressing scales relevant to biomolecular systems, they are crucial, along with experimental data, for estimating the parameters of a force field [65, 66]. The force fields considered in MD simulations do not actively account for the electron density but implicitly contain such information from DFT and provide a description for overarching biochemical processes [67, 68]. See [69] for an overview of various DFT approaches and their application in chemistry.

An active area of research is the development of ab initio MD approaches, which employ forces evaluated with DFT calculations instead of relying on empirical force fields that integrate out the electronic structure. An ab initio MD approach was first proposed by Car and Parinello [70–73] for simulations of dense metallic systems at scales inaccessible by pure DFT calculations [74, 75]. Ab initio MD approaches can resolve bond breaking and forming processes by resolving the electron structure but at the cost of restricted spatial and temporal scales.

Attempts to overcome scale limitations with quantum accuracy initiated an active research avenue on learning surrogate models for DFT calculations. Novel approaches, e.g., machine-learned Gaussian approximations [76–78] and neural networks [79] trained on transferable effective force fields based on DFT data, have facilitated MD simulations at almost DFT accuracy. Learned potentials are computationally efficient to evaluate compared to performing full DFT simulations, as done in the Car–Parinello MD approach. However, using these potentials is unavoidably accompanied by decreasing accuracy. Instead of simulations, other approaches use flexible machine-learned models based on large DFT data sets to predict molecular properties directly [80–84]. More recently, a novel research direction, empowered by the rise of machine learning, aims to predict molecular properties based solely on their chemical structure [85]. Though the model training relies on properties of reference structures, it provides explorative features for any molecular graph by learning the molecular structure. For this purpose, Kondor and Anderson developed covariant compositional networks [86], which contain information about atomistic interactions and capture the multiscale nature of molecular graphs using a hierarchy of subgraphs.

Molecular dynamics

Resolving the electronic structure becomes cumbersome for large systems [87]. For biochemical processes, detailed electron density information is, in many cases, not essential and can be implicitly incorporated into an atomistic description. Atomistic MD simulations compute the dynamic evolution based on classical mechanics of positions and velocities, in which their nuclei represent the system's atoms. The atoms interact via a force field or interaction potential, which usually decompose into simple parametrized functions. The different components model physically motivated multi-body contributions from bonded and non-bonded interactions [14, 88]. Although ab initio quantum mechanical calculations are not appropriate for addressing scales relevant to biochemical processes, they are crucial, along with experimental data, for estimating the parametrization of a force field [65, 66, 89]. Given a system's specifications, such as geometry, boundary conditions, and the force field, its equations of motion based on classical mechanics can be integrated stepwise with numerical methods [90]. Symplectic numerical schemes, as required in MD, must conserve the phase-space volume. Time reversibility and energy conservation are relevant criteria if time integration methods are to give trustworthy trajectories [14]. The standard choice is the Velocity-Verlet algorithm, which provides up to second order in time accuracy [91]. In general, analytic solutions are not available for the N-body problems addressed.

Without requiring any additional modifications of the integration scheme, the MD trajectories obtained, follow the microcanonical (NVE) ensemble. However, many in silico experiments require different thermodynamic conditions. Thermostats can augment the time integration such that the particle dynamics follow the desired ensemble. For example, the earliest and simplest approach adjusts the system temperature by rescaling the particle velocities (strong coupling) [92]. Other methodologies rely on a coupled random force and constant friction on the particles to satisfy the fluctuation–dissipation theorem [93]. The findings of Nosé and Hoover [94, 95] led to a deterministic formulation based on augmented equations of motion with an additional friction degree of freedom associated with a fictitious mass [96] for the canonical ensemble. These approaches can be transformed for an isothermal isobaric (NPT) ensemble.

The anticipated added value of atomistic MD simulations depends on the quality of the force fields utilized [97, 98], which have reached an acceptable level of accuracy for representing, e.g., protein-folding processes. We refer to [99] for a survey comparing the predictive capabilities of various force fields.

We elaborate more on MD simulations that depict the reference simulator used in this work in the methodology part.

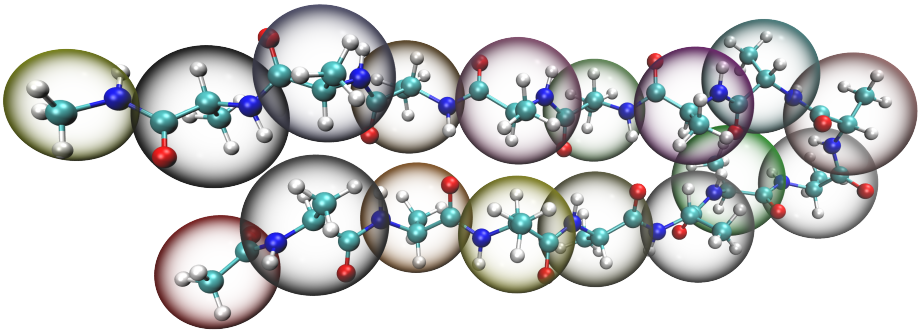


FIGURE 1.3: Typical mapping from atoms to CG variables. Multiple atoms form one CG bead, depicted as semi-transparent spheres. We render all atomistic representations in this work with VMD [112].

Coarse-graining methods

Ab initio methods are applied for simulation timescales spanning a few picoseconds, whereas MD simulations reach hundreds of nanoseconds. However, the simulation of biochemical processes from classical first-principles methods remains prohibitively costly for millisecond timescales, even when considering current and foreseeable advances in high-performance computing [50, 100]. CG approaches aim to overcome the scale limitations of first-principles methods by integrating out atomistic detail [101, 102], while still being able to represent essential system phenomena, like the diffusion of polymer melt [103, 104] and conformational changes of proteins [105–108], which occur over long timescales. CG representations are sought to contain essential atomistic features while omitting less relevant atomistic fluctuations. Thus, a CG mapping defines which fine-grained (FG) degrees of freedom are summarized and map to CG degrees of freedom, which are also termed CG variables [109]. These CG variables should capture the most salient physical features of the FG simulations. One possible strategy for obtaining a CG mapping is motivated by a molecule structure that summarizes several atoms and treats them as a single interaction site or CG particle in the CG description [110, 111], e.g., as shown in Figure 1.3.

As well as the CG mapping, a second crucial component for CG approaches is the potential energy surface, which is a function depending on the CG variables. The CG potential energy describes the interactions between CG variables or sites, and thus, the CG interaction forces. For a given CG mapping, there is a many-body potential of mean force (PMF) [113, 114], which yields an equivalent distribution of CG system states as the reference all-atom distribution represented in the CG coordinates. For relevant systems, the direct calculation of the PMF is computationally

impractical. Remedy provides the reformulation of identifying the PMF as an optimization problem [115]. Different objectives are motivated by bottom-up and top-down CG strategies. Bottom-up approaches rely on an FG atomistic reference simulation, which provides highly resolved information from classical first-principles or structural correlations derived from the reference simulation. In contrast, top-down approaches aim for reproducing thermodynamic observables obtained from in silico or in situ experiments. Section 1.2 discusses the advantages and disadvantages of the methods. Additionally, there is a concise review by Noid [116].

The reduced number of degrees of freedom in CG models (which lowers the computational effort for the calculation of forces, the time integration, and Monte Carlo sampling) is an essential reason for the efficiency gain compared to all-atom simulations. A further computational speed-up is due to the smoother potential energy surface prevalent in CG models compared to the highly resolved atomistic counterparts. High-frequency bonds, e.g., hydrogen bonds, are integrated out and modeled implicitly by a smoothed CG potential energy surface [117]. Smoother potential energy surfaces facilitate more substantial time steps (in MD simulations) or more significant trail movements (in Monte Carlo simulations). However, having fewer particles in a CG model results in reduced friction in the system, which changes the dynamics [118–120]. In this work, we are interested in systems in equilibrium and corresponding observables, where the enhanced diffusion due to the smoothed potential energy surface is beneficial and leads to an accelerated exploration of the phase space [100, 121].

Many CG methodologies require the learning of process-targeted CG models. Training these models relies upon, either long FG atomistic trajectories or reference thermodynamic or structural observables. After successfully training the CG model based on some objective, the model is supposed to estimate observables on extended time-horizons. Next to CG models, enhanced sampling approaches are accelerating the efficient exploration of scales. These are not in the focus of this work, but we develop among this work an essential component to enhanced sampling methods, which is the identification of collective variables [122].

Collective variables and enhanced sampling

Many research questions in biochemistry are related to conformational changes and protein folding of biological macromolecules [123]. However, biomolecules are known for a rugged free-energy surface with several local minima separated by substantial barriers [124]. Simulations of such molecules with high free-energy barriers tend to spend much computational effort on exploring one minimum of the free-energy surface without efficiently exploring other relevant minima [125, 126]. Deep basins hamper the exploration of the configurational space [127, 128], which could result in a biased trajectory that does not reflect the actual all-atom density [129–131]. Enhanced sampling methods sought to overcome this issue by enabling a more efficient exploration of the configurational space.

One category of enhanced sampling relies on a variety of combinations of tempering schemes, utilizing parallel replicas and reweighting. The post-processing step of reweighting to adjust sampling based on the correct target temperature is essential. Methods such as parallel tempering, replica-exchange MD [132, 133], multicanonical sampling [134], and enveloping distribution sampling [135] are only a few out of the many strategies.

Other enhanced sampling methods utilize collective variables for biasing the potential energy to evoke favorable circumstances for escaping the current conformation and its corresponding local minimum in the free-energy surface [124]. Several enhanced sampling methods use this concept, e.g., umbrella sampling [136–138], potential smoothing [139], conformational flooding [140], adaptive biasing force method [141], metadynamics [124, 142], and variationally enhanced sampling [143].

Since enhanced sampling is not the main topic of this thesis, refer to the review articles for more details [122, 131, 144].

1.2 Related work and motivation

This section reviews existing CG approaches and discusses their advantages and shortcomings.

1.2.1 Coarse-graining methods

Top-down CG approaches aim to reproduce thermodynamic reference observables and large-scale properties obtained from experimental measurements. Since there is no underlying atomistic model, low-resolution top-down CG models, in general, are not able to capture the essential high-resolution physics. The modeler's physical intuition often motivates the parametrization of a top-down CG potential, and parameters are optimized to match large-scale properties. In many cases, there is no unique set of parameters where different CG physics or CG interactions can lead to the same thermodynamic properties [145, 146]. However, top-down models successfully delivered CG models in [100, 147]. Since we are interested in CG models providing interpretable physical insight for a complex atomistic all-atom model, we focus on approaches with connections to an underlying FG model.

Bottom-up CG approaches count on a fully atomistic simulation with a reference trajectory, forces, and structural correlations (e.g., radial distribution functions). Instead of reasoning from a single thermodynamic observable, as is common in top-down approaches, bottom-up CG approaches use a structured procedure to identify low-resolution features from high-resolution reference simulations, consistent with statistical mechanics [116]. Note that for bottom-up CG approaches to capture the information contained in reference simulations, they require sufficiently long reference simulations. Enhanced sampling methods [122] are efficient for exploring the free-energy landscape extensively. Existing bottom-up approaches require the definition of a fine-to-coarse mapping for connecting the all-atom scale with the CG

representation and a (parametrized) interaction potential expressed in the dependency of CG coordinates. The many-to-one mapping needs to be defined a priori and is usually motivated by existing physical insights. Various methods are after that employed to learn effective low-resolution interactions.

The following discusses structural and variational bottom-up CG approaches and the role of the fine-to-coarse mapping.

Correlation function approaches

Some bottom-up CG methodologies focus on reproducing structural observables, such as pairwise distribution functions, higher-order correlations, and angular and dihedral distribution functions defined by three or four atoms. The reference structural distribution functions utilize a reference all-atom simulation mapped onto the coarse scale. Reference structural observables must be defined on the same coarse-scale for consistency, since there is, as yet, no reverse coarse-to-fine mapping in CG methodologies. Examples of structural bottom-up CG approaches are inverse Monte Carlo (IMC) [148] and (iterative) Boltzmann inversion (IBI) [114, 149].

IMC methods, also known as reverse Monte Carlo methods [148, 150, 151], rely on renormalization group Monte Carlo methods that were initially developed to simulate CG phase transitions of spin systems [152, 153]. The approach is iterative and estimates the linear response of all pairwise radial distribution functions based on a change in the CG pair potential for a given interaction type at a specified pair distance between CG sites. This leads to a susceptibility matrix of the pairwise distributions with respect to the pairwise interaction potential. The susceptibility equals the covariance of pairwise distribution functions multiplied by the inverse temperature. Coupling both expressions for the susceptibility matrix results in a linear system of equations that can be solved to obtain the CG potential energy. The susceptibility matrix is proportional to the covariance of pairwise distribution functions and, by definition, always positive definite. Thus, an IMC approach converges to a global optimum and accounts for cross-correlations between pairwise distributions when updating the interaction potential. The original approach employs a CG interaction potential represented as a piecewise constant function with a discretization of 0.5 Å, which is restrictive, especially when considering the detailed resolution of repulsive interactions. An extension providing more flexibility for the CG potential utilizes a linear combination of radial basis functions [154]. However, the basis functions have fixed support. Expressivity needs to be maintained by providing various kernels, which leads to an optimization problem of increased complexity. A formulation of renormalization group Monte Carlo [153] focused on CG representations of the Ising model [155, 156] with further extensions and amendments for identifying CG models for DNA and other large biomolecules [157, 158].

Boltzmann inversion (BI) was initially proposed in [149] and is, like IMC, a structural bottom-up CG approach. BI assumes that CG structural observables, such as two-body distribution functions, are uncorrelated, where ξ denotes a CG variable,

e.g., corresponding to the distance between CG particles. Given a two-body correlation function, [149] assumes that the observed distribution relies on the corresponding Boltzmann factor and that the CG interaction potential $U_c(\xi)$ is in the exponent: $p(\xi) \propto \exp\left(-\beta U_c(\xi)\right)$. The inversion for obtaining the CG potential is

$$U_c(\xi) = -\frac{1}{\beta} \log \frac{p(\xi)}{J(\xi)}, \tag{1.1}$$

where the Jacobian $J(\xi)$ accounts for representations not using Cartesian coordinates. BI focuses on pairwise interactions and CG variables that depend on only a single coordinate to enable direct inversion. An explicit limitation of this approach is the assumption of independent pairwise distribution functions between CG particles, which is invalid in systems of practical interest [159]. The assumption can be justified, however, for all types of (stiff) bonded interactions but provides limited capabilities in identifying non-bonded CG interactions [149, 160]. Iterative methods, like IMC approaches or the following iterative version of BI, provide better descriptions of correlations between different pairwise interactions by accounting for them implicitly in an iterative procedure that balances the contributions [161, 162].

Instead of directly using inversion to determine the CG potential, IBI [114] proposes an iterative updating procedure. The initial guess of the potential employs direct Boltzmann inversion, which is a reasonable starting point. Differences between the reference atomistic observable $p_{ref}(\xi)$ and the observable obtained from a CG simulation utilizing the current CG potential $U_c^i(\xi)$ at iteration i, $p_c^i(\xi)$, motivate the per iteration update:

$$\Delta U_c^i(\xi) = \frac{1}{\beta} \log \frac{p_c^i(\xi)}{p_{ref}(\xi)}. \tag{1.2}$$

Equation 1.2 adjusts the CG potential at ξ if $p_c^i(\xi) < p_{ref}(\xi)$ such that $U_(\xi)$ becomes more attractive at ξ and vice versa according to the update rule:

$$U_c^{(i+1)}(\xi) = U_c^{(i)}(r) + \alpha \Delta U_c^{(i)}(\xi). \tag{1.3}$$

The original formulation suggests changing the CG potential for one type of interaction (e.g., bonded interactions) while keeping it fixed for the others [114]. The optimization should start from the least correlated and thus, stiffest pairwise interactions, in the order: bonded, angular, dihedral, and non-bonded interactions. The iterative approach tries to account for cross-correlations, which, however, hampers the convergence [163].

IBI and IMC methods focus on reproducing certain structural observables for a specific thermodynamic state. More recent enhancements generalize CG models based on IBI by considering reference observables from different thermodynamic

states. However, they still focus on structural observables [164]. CG models obtained from IBI and IMC approaches tend to lack accuracy in matching thermodynamic properties. Thermodynamic constraints, introduced via Kirkwood–Buff integrals [165, 166], provide a potential avenue for addressing the mismatch in the thermodynamic properties as presented in [167, 168]. Adding thermodynamic constraints [169], however, is unavoidably accompanied by a lower accuracy in reproducing pairwise distribution functions [170]. Refer to [171] for a comprehensive review of IBI methods with thermodynamic correction terms. Structural CG approaches rely on pairwise distribution functions. Although including three-body correlations from the atomistic reference and providing three-body terms in the CG potential is feasible, however, results in poor convergence while adding a tedious additional computational cost [172]. Moreover suffers the accuracy of estimates concerning three-body correlations under prevalent noise.

An IMC approach requires the calculation of cross-correlations between different pairwise interactions, which improves the accuracy of the CG potentials obtained. However, computing correlations requires, in general, more extended CG simulations per IMC iteration to balance noise-afflicted estimates and to achieve reasonable convergence. By utilizing noisy estimates, using correlations results in IMC approaches having accelerated convergence compared to IBI approaches for pairwise interactions. Theoretically, both IMC and IBI frameworks can accommodate the identification of higher-order interaction terms in the CG potential energy. However, these frameworks are impractical for systems of interest [114, 148, 173, 174].

CG models obtained from the IMC, BI, and IBI approaches aim to reproduce pairwise structural observables, thus seek to reproduce an approximate pairwise PMF. Different CG models can lead to the same observables [116, 170, 175]. However, such CG models may not reveal or encode higher-order physicochemical insight.

Variational approaches

In approximating the many-body PMF, CG frameworks, such as the IBI and IMC, only use a fraction (expressed by structural properties) of the information contained in atomistic reference simulations. Objectives matching certain structural distribution functions can approximate only an effective PMF expressed by the distribution functions provided. These are mostly pairwise distributions. Even by adding flexibility through adding higher-order interactions, they can still access only a limited portion of the information contained in the atomistic reference trajectory, expressed by the selected pairwise correlation functions.

Instead of relying on a particular selection of structural correlation functions up to a specified order, variational CG approaches have access to the full information resolved by atomistic reference simulations in the form of the trajectory. In the limit, variational CG approaches allow the identification of the many-body PMF. As well as defining a mapping from fine to coarse coordinates, a characteristic functional

defining an optimization objective is required for such methods. This objective results in an optimization problem that is a rigorous approach for systematically learning the parametrization of CG potentials, not only for pairwise contributions but flexible parametric forms, including contributions up to the desired interaction order [116].

Multiscale coarse-graining (MS-CG) [109, 176, 177] and the relative entropy framework [178–181] are the most prominent representatives of variational CG methodologies.

Multiscale CG approaches These are an extension of the original force-matching approach (FM-CG) first introduced in [182]. The approaches endeavor matching forces of atomistic reference simulations projected onto the coarse-scale with the forces of the CG model evoked by the CG potential U_c [109, 176, 177, 183, 184]. The framework projects atomistic reference forces onto the CG scale for maintaining consistency [109, 184]. Differences between forces obtained from the CG model and reference forces are expressed by a mean squared error functional, which is subject to minimization concerning the CG potential U_c. In the variational approach, the objective functional has a global minimum if the CG potential U_c equals the many-body PMF (for a proof, see [177, 185]) or, in the vicinity of the minimum, if it is an approximation to the many-body PMF. CG forces, evoked by the CG potential U_c, are evaluated at atomistic reference configurations represented in terms of CG coordinates. Evaluating forces at coordinates governed by the reference trajectory implies that MS-CG approaches do not require simulations of the CG model for learning the involved parameters by minimizing the objective [186]. MS-CG approaches can reproduce the effective average forces in terms of CG coordinates. However, since there is no connection to the reference distribution functions, individual correlation functions, e.g., radial or angular distribution functions, are not accurately captured [116, 187]. [188] fixes this by reformulating the MS-CG approach by iteratively matching forces and gradients of derivatives of two-body correlation functions. Convergence is though not guaranteed and the methodology fails in reproducing three-body distribution functions. In most cases, the MS-CG approaches rely on a linear mapping from a fine scale to CG coordinates, which was extended by [189] to nonlinear mappings that facilitate the use of the collective variables [190] identified in the MS-CG framework.

Relative entropy CG approaches These approaches [178–181] belong to the group of variational CG methods that minimize a functional, which yields, in the limit of infinite basis functions representing the CG potential, the many-body PMF. Relative entropy CG relies on an information theoretic objective that seeks to minimize the overlap of two probability distributions. The relative entropy or Kullback–Leibler (KL) divergence is a special case of the Rényi α divergence [191–193]. The KL divergence is asymmetric and is greater than or equal to zero. The distributions involved

(the atomistic reference and the CG distribution) require expressions concerning atomistic coordinates. The reference Boltzmann density is naturally an expression in terms of the fine-scale coordinate, while the corresponding CG distribution evaluated on the fine-scale involves a normalization. The renormalization relates to the degeneracy of CG states, which is the mapping of multiple FG configurations to identical CG variables. The so-called mapping entropy term in the proposed objective links the degeneracy of multiple fine-scale configurations to one CG state [180]. This term is independent of the parametrization of the CG potential and expresses the unavoidable loss of information due to utilizing a lower resolution description of the atomistic ensemble caused by the fine-to-coarse mapping. Expressing the CG potential as a linear combination of basis functions and parameters, as shown in [180], yields an objective providing a global maximum, which is a compelling property and enables the application of a broad set of optimization algorithms, as investigated in [194]. Refer to Section 2.2.2 for more details on the relative entropy CG methodology.

Both approaches yield, in the limit of a complete set of basis functions for expressing the CG potential, the many-body PMF, or, in case of employing a limited set of basis functions, an optimal approximation to the many-body PMF. All references above for the MS-CG and relative entropy CG approaches include terms up to second order in the CG potential functional. However, as emphasized by [187, 188], relevant multi-body distribution functions [195], e.g., the angular distribution, are not resolved to sufficient accuracy when considering terms up to second order in the CG potential. In contrast, three-body interactions are responsible for phenomena in polymers [196, 197] and liquid crystal physics [198]. Variational objectives (MS-CG and relative entropy approaches) circumvent the estimation of reference correlation functions and CG higher-order correlations since they rely directly on the trajectory compared to associated statistics. The first approaches that extend the expressivity of the CG potential using three-body terms [115, 117, 173] have had significant success in improving CG water models and provide better approximations to angular structural distributions. Further developments extend the MS-CG approaches by considering fixed functional forms of three-body terms in the CG interaction potential. Applying a cutoff radius of 3–4 Å, balances the computational burden with model flexibility and accuracy. Additional advancements [185] introduce a set of basis functions for two- and three-body interactions, which softens the assumption of the functional form. Adaptive basis functions describing the CG potential have been introduced based on the MS-CG approach [199]. Refer to [200] for a more detailed discussion on variational CG approaches and explicitly MS-CG and relative entropy CG.

Introduced CG methods suffer when many metastable states with differing optimal collective or slow variables exist. For each conformation, a different set of CG variables and thus, a different fine-to-coarse mapping may be beneficial for obtaining simple CG interaction potentials. Ultra coarse-graining (UCG) [201] addresses

challenges arising due to multimodality by facilitating various CG mappings. The authors introduced a latent discrete random variable to select the conformation (e.g., folded or unfolded) and assign to each particular mode a predefined individual mapping and, likewise, a parametrized local CG interaction potential. The UCG approach is readily incorporated in variational CG frameworks, MS-CG, and relative entropy, which drive the parameter learning process. Beyond coping with multiple local CG mappings, UCG provides improved convergence in the associated optimization problem. [202] provides the numerical details of the UCG approach combined with the FM-CG variational objective for 1,2-dichloroethane, which has two distinct conformations governed by dihedral angles. Local CG potentials associated with corresponding modes have been considered as independent thus far. [203] extends the local, independent CG potentials by considering interaction terms shared over various conformations and weighted by conformation-dependent mixtures. Sharing the parametrization combines advantages of conformation-specific CG variables with having simple and structured CG interactions based on mode-dependent responsibilities. The mode-associated weights depict conditional posterior distributions or responsibilities (in the context of Gaussian mixture models) of interactions in a particular state. Nonetheless, UCG requires insights regarding existent conformations and individually specified fine-to-coarse mappings. Thus, the definition of CG variables in each basin of the free-energy landscape. However, as mentioned in [201], defining useful fine-to-coarse mappings in the absence of prerequisite system knowledge depicts a ubiquitous problem. The modeler's insight may not suffice, recognizing most characteristic slow coordinates for each conformation and introducing appropriate CG variables.

Further CG approaches are based on free-energy computations [204, 205]. However, their primary purpose is to escape from deep free-energy basins and are limited to few CG variables. Adaptive multiscale formulations [206–211], developed with mathematical rigor for equilibrium and non-equilibrium situations, focus on CG lattice systems. Extensions for time-dependent variational frameworks rely on force matching, relative entropy, and probabilistic objectives [212–218]. An attractive Bayesian formulation of CG dynamics was presented in [219]. Since the focus of this work is on equilibrium methods, we refer to the comprehensive review in [220] on non-equilibrium CG methodologies.

Thus far, all discussed CG methodologies require two modeling components:

(i) A definition of a fine-to-coarse mapping function, and thus the definition of the CG variables.

(ii) A parametrized formulation of the CG potential describing how CG variables interact.

The rise of deep learning models [221], backed by algorithmic advances that can efficiently utilize modern GPU platforms [222], has encouraged research into novel CG approaches that incorporate deep models. DeepPCG [223] is a CG approach

based on force matching in which the CG potential is parametrized via a deep neural network instead of via a physically motivated composition of basis functions, such as pairwise and three-body terms. Reference [224] poses the identification of a CG potential as a supervised learning problem. Given input CG particle positions, the methods predicts CG interaction forces based on a neural network [225]. Similar to DeepPCG, the overall objective compares implied CG forces to forces of a reference trajectory represented on the CG scale. However, [226] implements this with a physically motivated feed-forward neural network. CGnet provides a mapping from input Cartesian coordinates to associated CG forces, while Cartesian coordinates propagate through a first "featurization" layer of the neural network to a relative representation for further use. The relative representation utilizes pairwise distances and sines and cosines of angles defined by three adjacent CG particles. The relative representation removes partially nonlinearities of a purely Cartesian representation and alleviates the required complexity of the neural network. The layers after featurization employ a composition of artificial neural networks that propagate the input to the corresponding CG potential energies and their gradient, the CG interaction forces. However, a definition of the essential CG features in terms of a fine-to-coarse mapping is still required a priori. In addition to formulating the CG process as a supervised learning problem, as proposed in [224], reference [227] critically and extensively discusses the use of dimensionality reduction methods and unsupervised and supervised learning approaches in the context of insightful CG approaches.

Following implicit generative models [228, 229], [230] proposes an adversarial minimax objective in the context of CG. The approach minimizes a general f divergence [192] between the FG reference Boltzmann distribution projected on the CG scale and the CG counterpart while maximizing the objective concerning the parameter specifying the f divergence distance metric. Special cases relate to relative entropy CG approaches [231] and the Hellinger distance [232, 233].

Despite the rise of machine learning, the approaches mentioned above still require a decent amount of physical insight in defining a fine-to-coarse mapping and the transformation to a featurized representation. Even if neural network layers address this featurization, the employed features are pre-specified and not revealed after a learning process. Thus, physical insight is required a priori.

A similar unsupervised CG approach, introduced in [234], relies on an alternation of a variational autoencoder [235]. The employed loss function decomposes into a reconstruction loss term and a mean force regularization term multiplied by a weight factor. The weight factor depicts a hyperparameter. Advantageous of the proposed formulation is a discrete and parametrized mapping that assigns atoms to CG macromolecules. The mapping can reveal optimal CG beads and, thus, the approach has the potential for identifying appropriate CG resolutions. Reference [234] implements the discrete mapping which is efficiently realized by a categorical reparametrization with a Gumbel-softmax transformation [236]. However,

the involved weight parameter associated with the regularization term affects the optimal model parametrization, which requires further detailed study [237].

Interesting recent attempts utilize the reverse definition of the relative entropy supported by invertible neural networks [238, 239], which enables the learning of CG models without simulating an atomistic reference trajectory but instead by evaluating the atomistic interaction potential and corresponding forces for samples generated from the CG model [240, 241]. In their work, invertible neural networks enable inference without introducing further approximations, like constructing upper or lower bounds on the optimization objective. However, the samples generated from the CG model do not match the reference state probabilities, and thus, the FG reference Boltzmann distribution. The work corrects the state probabilities by employing a post-processing step utilizing reweighting importance sampling. Such physics-informed deep learning approaches are successfully applied in the context of partial differential equations [242, 243].

The approaches introduced thus far utilizing deep learning require a vast amount of training data to stabilize the training of overparametrized neural network representations. Graph-based CG methodologies can make efficient use of training data since each molecule's structure is reflected in the model parametrization [244–246].

[247] was the first study to compare CG interaction potential models under a fully Bayesian framework. The authors developed an automated algorithm for model selection and validation, utilizing the model evidence and the posterior model plausibility. The iterative algorithm was based on Ockham's razor [248] and refines the CG interaction potential by utilizing the most plausible models from the preceding iterative steps. This methodology involves computing the model evidence, and thus, the marginalization concerning the model parameters, which is computationally impractical for sophisticated and flexible interactions. Furthermore, the model validation step depends on the choice of observables. Reference [249] studies a Bayesian framework for model selection considering multiple mappings, which imply multiple CG resolutions for representing water. These fine-to-coarse mappings yield CG beads consisting of three to six water molecules. The CG beads encompass one, two, or three bonded CG particles per bead, while inter-particle interactions per CG bead consider combinations of bonded and electrostatic components. Non-bonded interactions between CG beads follow a Lennard-Jones potential. The likelihood, which [249] utilizes in the generative model, represents observables (density, dielectric constant, surface tension, isothermal compressibility, and shear viscosity) estimated from long FG simulations. However, this approach involves marginalizing of the model parameters, which requires a vast number of CG MD simulations associated with a tremendous computational cost per parameter set. The authors address the computational burden by incorporating a Gaussian process regression [250] that directly computes required observables based on the input model parameters and thus circumvents extensive CG MD simulations. However, training the Gaussian process and producing a few reference observables for the given model parameter

set, induces a moderate computational overhead.

Mapping

In the limit of employing a complete set of basis functions for the CG potentials, methodologies such as the MS-CG and relative entropy CG provide a general framework for identifying the many-body PMF. Even if it sounds compelling to have identified the many-body PMF, or an approximation that is close to it, this does not imply that the CG simulations provide the same information as a simulation of the fully resolved atomistic system. The loss of information is unavoidable due to the reduction in the number of degrees of freedom, which is explained well in the context of relative entropy CG approaches with mapping entropy. The prespecified non-parametrized mapping induces the information loss [251]. The mapping and thus, the CG variables define the resolution in the CG model and the level of detail needed to encompass the information provided by atomistic simulations. The identification or selection of CG variables is crucial to the expected predictability of observables and phenomena captured in the CG model.

[252] studied different mappings for polystyrene. The fine-to-coarse mappings yielded similar CG resolutions with different groups of atoms defining a single CG bead. A further systematic study on fine-to-coarse mappings was undertaken by [253]. These studies provide a rigorous way to understand the dependence of the accuracy of the CG model on the fine-to-coarse mapping. However, a modeler's chemical insight and experience are still required to construct the CG beads and thus, the definition of the mapping.

The essential-dynamics CG methodology [254] utilizes knowledge from performing a principal component analysis [255] for defining optimal CG sites. Other approaches rely on optimizing an elastic network [256, 257], such as for large biomolecules [258].

An automated approach for testing different mappings, from an atomistic representation to a CG model, was proposed by [259]. The authors utilized a mapping operator tree consisting of all symmetry-preserving mappings in a graph-based representation of a molecule. At the highest resolution, nodes, which represent atoms, are connected with edges, which depict atomistic bonds. Their approach relies on minimizing the difference in the summed node entropy [260] between paths in the mapping operator tree. The entropy utilized for comparison is the entropy of the velocities of a group of atoms and provides a measure of how homogeneous atoms move within a group. [259] develops automated procedures for finding optimal mappings in terms of the entropy. However, obtained fine-to-coarse mappings always rely on the center of mass of the group of atoms. As a natural extension of linear mappings, can nonlinear or multi-component dimensionality reduction methods, such as temporal-independent component analysis [261] or kernel principal component analysis [262], improve the information content per CG variable [190, 263].

[264] represented a molecule of interest as a molecular graph and also applied grouping methodologies based on the eigenvector centrality. Nodes rank by their contribution to the largest eigenvalue of the adjacency matrix. Potential CG beads are then synthesized from the corresponding descendants [265].

The approach proposed in [266] identifies, from a set of given CG mappings and corresponding CG interaction potentials, the one which is most promising for observed distributions compared to the atomistic reference via the Jensen–Shannon divergence [267, 268]. This approach, however, does not train mapping-specific CG potentials but instead utilizes the Martini force field[1] [270, 271] with predefined interactions between a library of CG beads. The interactions are specific to the type of CG macromolecules and depend on which species of atoms map to one CG site. Further descriptions can be found in [114, 272–274].

Adaptive concurrent multiscale modeling approaches [275] can couple resolutions from the atomistic viewpoint up to the continuum description of materials, which is relevant, e.g., in the context of predictive simulations of crack propagation [276, 277]. The adaptive resolution scheme (AdResS) proposed in [278–281] concurrently couples multiple resolutions, e.g., a full atomistic description with a CG viewpoint, by conserving linear momentum. Adaptive reslolutions are especially compelling when the atomistic dynamics of a solvent influence the solute (e.g., a peptide) in a spatially limited volume fraction alongside the solute but have less influence in the far-field. Predefined transition areas appropriately weight the CG forces and effective atomistic forces per molecule [282]. An extension of AdResS connects MD simulations with dissipative particle dynamics simulations [283], which enables the simulation of correct hydrodynamics.

1.2.2 Summary and challenges

All CG approaches, as mentioned earlier, rely either on structural correlations, thermodynamic observables, or directly on FG reference trajectories. While relative entropy CG employs reference Cartesian coordinates, use force-matching or MS-CG approaches reference forces obtained from fine-scale atomistic trajectories. Structural CG approaches utilize pairwise distance distributions (IBI) or, besides, correlations between pairwise radial distribution functions (IMC) from FG reference simulations. All data-driven CG approaches employ non-parametrized predefined mappings, which represent a functional relation from the input atomistic coordinates to the analogous CG variables. However, more than one set of atomistic coordinates can yield the same CG representation, e.g., when utilizing the centers of mass as CG variables for the atomistic description of water molecules [284]. The definition of many-to-one mappings is predominantly influenced by the modeler's physical intuition and experience [100]. Automated approaches for identifying the most promising fine-to-coarse mapping still require physical insight in creating a

[1] The Martini CG force field was obtained by using top-down CG approaches to match experimental properties [269].

comprehensive library of mappings. A systematic approach based on molecular graphs suffers from combinatorial possibilities forming CG beads and the computational burden associated. Furthermore is the superiority of particular fine-to-coarse mappings dependent on the observables sought to predict. All considerations, as mentioned earlier, naturally yield the question of how to assess the quality of CG models and rigorously compare different models [285, 286].

The CG methodologies reviewed earlier utilize all-atom data (trajectories and forces) projected onto the CG scale and reference observables computed from the CG representation of all-atom reference trajectories. Thus far, after applying the fine-to-coarse projection, the detailed information from classical first-principles simulations is not available to the CG model anymore and cannot be recovered due to lacking coarse-to-fine connections. Potential mappings need to be well-conceived while they may lead to observable-dependent capabilities. Integrating out degrees of freedom prevents CG models from reproducing observables explicitly depending on FG coordinates. E.g., employing a molecule's center of mass as a CG description prohibits the estimation of observables depending on structural correlations between atoms within this molecule. There is no structural correlation defined with regards to the CG representation in the previous case. Models that cannot reason about the FG variables only suffice for quantities of interest that are fully defined by a CG representation of the FG atomistic coordinates. In general, however, depend quantities of interest on FG coordinates, which leads to the problem of representability as discussed in [145, 284, 287].

Furthermore, matching quantities concerning reference observables expressed in terms of the CG scale is no guarantee that the associated CG model encodes FG physics. FG fluctuations, however, do influence the thermodynamic observables in particular and are essential for providing a fully predictive CG model that would allow for reasoning FG coordinates given a coarse representation [288, 289]. Attempting to establish a reverse coarse-to-fine mapping raises two questions: What CG resolution is sufficient? How many CG variables are needed to encode the all-atom behavior? The term "sufficient" for the quality of CG predictions requires a consistent definition enabling comparability between CG models [290]. A fine-to-coarse mapping is a many-to-one mapping and is thus not invertible, which is problematic when elaborating a consistent mapping back to the fully resolved atomistic scale. Beyond developing a parametrizable map, are mappings interesting, which could reason about FG coordinates given the coarse representation. Providing a connection back to the fine-scale supports encoding and revealing relevant physics rather than employing physical intuition for pre-specifying mappings. Additionally, given a CG variable, mapping back to FG coordinates enables the employment of the CG potential for biasing FG simulations to positively influence the exploration of the configurational space providing favorable circumstances for escaping from free-energy minima [291]. Besides the reconstruction mapping from coarse to fine, it is

relevant to identify assigned CG variables given fully atomistic coordinates for model interpretability, in agreement with [292]. In particular, we need to develop CG methodologies that illustrate and reveal the underlying physical processes and thus enhance our understanding of atomistic systems using CG representations rather than assuming physical intuition a priori for creating CG models.

The availability of training data directly affects the credibility of estimates of observables obtained from CG models [293, 294]. The uncertainty due to the limited availability of training data needs to be accounted for in design processes relying on CG models, which we also address with our proposed approaches [295].

Thus far, the CG potentials decompose into physically motivated parametrized interaction terms for pairwise bonded interactions or non-bonded interactions. Some approaches attempt to include three-body interactions, which increases the computational burden. It is challenging to choose the right interaction terms in the CG potential. Doing so requires a vast amount of prerequisite physical insight, which we instead seek to reveal with the CG methods proposed in this work. Providing a complete set of basis functions is one possibility. However, the strategy mentioned earlier works only in a big data regime, whereas our focus is on small data, as is applicable in MD simulations. Instead, we develop an automated process to select relevant basis functions given limited training data. Coping with limited data facilitates the identification of physically relevant features that are associated with the appropriate basis functions. Instead of identifying the most salient features from a large set of basis functions, we explore besides the opposite approach, where we start with a few basis functions and iteratively refine the CG potential by adding relevant features. Moreover, we explore how to learn flexible coarse-to-fine mappings, which can reveal physically relevant features instead of predefining mappings and, thus, the meaning of CG variables. Given the enhanced mapping flexibility, it suffices to model CG variables with simple distributions [235].

In general, we seek to develop fully Bayesian predictive CG methodologies to facilitate the reconstruction of the FG picture and enable thus the estimation of observables that depend on all-atom coordinates. All estimates are fully probabilistic and account for uncertainties due to the limited data. We are interested in achieving this without any prior physical insight. We can reveal the relevant slow coordinates of an FG simulation and the most relevant CG interactions using an automated process. We present methodologies that enable the robust training of variational autoencoders and neural networks, in general, while fully incorporating the available physics.

1.3 Impact and contributions

We develop a CG framework to address the shortcomings above by employing generative probabilistic graphical models [296, 297]. The CG framework introduced

coherently follows the Bayesian paradigm and rigorously addresses model selection and validation [298, 299]. Unlike most CG approaches, which explicitly define the coarse variables by a many-to-one fine-to-coarse mapping, we propose a probabilistic coarse-to-fine mapping that implicitly defines CG variables. Involved lower-dimensional CG variables are supposed to give rise, through the probabilistic mapping, to the full atomistic resolution. The proposed approach follows the concept of probabilistic generative models, which implies the generation of observed atomistic configuration through an underlying hidden CG structure. The generative process allows for obtaining atomistic configurations through producing CG states and employing the probabilistic coarse-to-fine mapping, which is a conditional probability distribution given the CG state. The framework developed enables the reconstruction of FG states and thus, the estimation of observables governed by complex interdependencies in the FG model. Beyond point estimates of observables allows the Bayesian formulation obtaining posterior distributions over the model parameters that account for the credibility of model parameters. We can propagate this credibility to estimates of observables and provide error bars or credible intervals for observables that reflect the uncertainty induced [300, 301]. The credibility depends on the amount of training data and of the unavoidable information loss.

Finding expressions for the CG interaction potential relies on having a rich set of basis functions, whereas we automatically synthesize the most relevant features using functional sparsity priors. The relevant features correspond to physical correlations or the smoothness of the CG interaction potential induced by the reference FG model. The identified features allow us to gain physical insights into the underlying all-atom model [302, 303]. Beyond revealing physics by identifying the most salient features, sparse models can cope with a limited amount of training data to provide a robust CG framework. Alternatively to selecting the most relevant features from a rich basis, we develop an approach that adaptively adds the most promising features, starting from a simple basis. Specifically, we address the addition of optimal radial basis functions where the kernels are optimized to maximize the anticipated benefit when adding them. The developed approach provides an efficient way to decide which type of basis functions are most relevant, whithout predefining a comprehensive library of feature functions. The proposed strategy can be adopted in the context of high dimensions, since providing a sufficiently rich set of basis functions is impractical in such context.

Instead of providing flexibility in the CG potential, we explore an approach for learning flexible coarse-to-fine maps, which can reveal physically relevant insights and slow collective variables. E.g., in the context of peptides, identified CG variables indicate the relevance of the dihedral angles. Taking the dihedral angles as a lower-dimensional description of peptides is known as an apt description which we discover without any prior physicochemical insight. The description obtained can be further employed in enhanced sampling methods, as we propose in the outline of this thesis.

Providing sufficient training data can be a challenging problem, as there may be multiple free-energy minima. The rugged free-energy surface prevents MD simulators from escaping basins and thus, exploring different conformations [304]. We propose a methodology that avoids simulating a FG system by using a reverse KL divergence objective. This yields a physics-embedding deep learning approach where a CG model can be optimized only by assessing the potential energy and forces of the FG system for samples generated from the CG model [242, 243, 305]. We successfully apply the developed methods to bimodal distributions, which are difficult to sample. The proposed approach is a basis for exploring the application of CG models to large biomolecules.

1.4 Outline

The remainder of this thesis is structured as follows. After specifying the research objectives and the contributions of the methodologies developed in this thesis, Chapter 2 introduces the notational conventions and mathematical basics of MD simulations. The chapter has a section on the consistency of CG models and the computation of observables with FG or CG simulations. We introduce Bayesian modeling and inference and further include a short introduction to the Monte Carlo methods applied in this work. One specific CG approach, the relative entropy methodology, is introduced as well since some of our developments build on this framework. Chapter 3 introduces a predictive CG framework that enables the reconstruction of the FG picture given CG variables and the quantification of uncertainties due to limited data and information loss, reflected in credible intervals for observables. Chapter 4 introduces a methodology for the adaptive addition of the most promising features in the CG potential by maximizing the anticipated benefit. Chapter 5 describes a methodology that employs deep neural networks to reveal the essential collective variables and to quantify the uncertainties. We present in this chapter further an approach for sparse neural networks and robust learning in the low-data regime. A CG framework that completely circumvents the need for reference simulations by fully incorporating the available physics is proposed in Chapter 6. Chapter 7 provides an overarching discussion, summarizes, and outlines the work presented in this thesis. The methods developed may be employed in novel research on enhanced sampling approaches.

Chapter 2

Methodologies

This chapter introduces the methodological basis of the present work. We summarize the atomistic simulation tools for estimating observables, which we employ for reference and predictive estimates (Section 2.1). We define fine-to-coarse mappings and introduce the relative entropy CG approach on which our work is based. Section 2.2 discusses the consistency of CG variables and observables based on CG variables compared to fine-scale coordinates. We introduce the probabilistic generative modeling approach and the motivation for uncertainty quantification and propagation in Sections 2.3 and 2.4. In Section 2.5, methodologies for inferring latent variables and model parameters are introduced. A core component of inference builds Monte Carlo (MC) estimates of integrals, as presented in Section 2.6. Section 2.7 introduces stochastic optimization, which is an important tool in the context of approximate inference and the presence of noisy gradient estimators.

2.1 Atomistic simulations

Atomistic simulations allow us to resolve detailed microscopic behavior for a variety of systems in solid-state physics and biochemistry, whereas molecular dynamics (MD) simulations rely on a time integration of the classical equations of motion. Alternative methods of simulating atomistic systems are MC sampling approaches [306], which are especially common in solid-state physics [156, 307, 308]. An MD simulation estimates quantities of interest as time averages while a MC method utilizes averages over the statistical ensemble, which are the same in the limit [309, 310]. In the following, we focus on MD simulations. For more on MC methods, see the general introduction in Section 2.6.

2.1.1 Molecular dynamics

MD simulations are based on solving the classical equations of motion of many-body systems. N_p atoms are modeled as particles based on the position of their nuclei. These particles move due to interatomic forces.

The classical equations of motion follow Newton's second law:

$$m_i \ddot{\mathbf{x}}_i = \mathbf{f}_i, \quad \forall i \in \{1, \dots, N_p\}, \tag{2.1}$$

where $\mathbf{x}_i$ is a vector of Cartesian coordinates, m_i the atomic mass, and $\mathbf{f}_i$ the force acting on particle i. Atomic positions and velocities are a function of the current time step t and thus, also the interatomic forces $\mathbf{f}_i(\mathbf{x}_1, \ldots, \mathbf{x}_{N_p})$ depending on the current particle positions. For Equation 2.1, there exists a unique solution for a given set of initial positions $\{\mathbf{x}_i(t=0)\}_{i=1}^{N_p}$ and velocities $\{\dot{\mathbf{x}}_i(t=0)\}_{i=1}^{N_p}$ at $t=0$. Analytic solutions of Equation 2.1 exist only in a very few cases, however, and not for practically relevant systems. Solving the equations of motion thus requires numerical iterative schemes yielding approximate solutions at discretized time increments with an increment Δt [311, 312].

To explain the requirements for numerical integration schemes in MD, it is beneficial to introduce the Hamiltonian representation of the equations of motion from Equation 2.1:

$$H(\mathbf{p},\mathbf{x}) \equiv H\left(\mathbf{p}_1, \ldots, \mathbf{p}_{N_p}, \mathbf{x}_1, \ldots, \mathbf{x}_{N_p}\right) = \sum_{i=1}^{N_p} \frac{\mathbf{p}_i^2}{2m_i} + U\left(\mathbf{x}_1, \ldots, \mathbf{x}_{N_p}\right), \tag{2.2}$$

where the momenta $\mathbf{p}_i = m_i \dot{\mathbf{x}}_i$ and the interaction potential $U(\mathbf{x}_1, \ldots, \mathbf{x}_{N_p})$. Given the interaction potential, the interatomic forces $\mathbf{f}_i$ are obtained with

$$\mathbf{f}_i = -\frac{\partial U}{\partial \mathbf{x}_i}. \tag{2.3}$$

Transforming the Hamiltonian in Equation 2.2 to give the positions and momenta:

$$\dot{\mathbf{x}}_i = \frac{\partial H}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} \tag{2.4}$$

$$\dot{\mathbf{p}}_i = -\frac{\partial H}{\partial \mathbf{x}_i} = -\frac{\partial U}{\partial \mathbf{x}_i} = \mathbf{f}_i\left(\mathbf{x}_1, \ldots, \mathbf{x}_{N_p}\right). \tag{2.5}$$

The equations of motion in Equation 2.1 can be derived by taking the time derivative of Equation 2.4 and replacing the expression for the momentum in Equation 2.5 [91, 98]. Two main properties of Equation 2.1 are worth mentioning:

1. The equations of motion are reversible in time.
2. They conserve the Hamiltonian (Equation 2.2), thus [14]:

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \sum_{i=1}^{N_p} \left[\frac{\partial H}{\partial \mathbf{x}_i}\dot{\mathbf{x}}_i + \frac{\partial H}{\partial \mathbf{p}_i}\dot{\mathbf{p}}_i\right] = \sum_{i=1}^{N} \left[\frac{\partial H}{\partial \mathbf{x}_i}\frac{\partial H}{\partial \mathbf{p}_i} - \frac{\partial H}{\partial \mathbf{p}_i}\frac{\partial H}{\partial \mathbf{x}_i}\right] = 0. \tag{2.6}$$

   Conserving the Hamiltonian is the same as conserving the total energy of the system, which is relevant when linking MD and statistical mechanics [97]. More details on statistical mechanics are provided in Section 2.1.2.

A well-studied time integration method for MD simulations is the Velocity-Verlet algorithm, which provides a reasonable balance between accuracy and efficiency. It preserves volumes in the phase space, thus it is symplectic. The relevant update

equations for obtaining positions and velocities at $t + \Delta t$ are

$$\mathbf{x}_i(t + \Delta t) = \mathbf{x}_i(t) + \Delta t \dot{\mathbf{x}}_i(t) + \frac{\Delta t^2}{m_i}\mathbf{f}_i(t) + \mathcal{O}\left(\Delta t^3\right), \tag{2.7}$$

$$\dot{\mathbf{x}}_i(t + \Delta t) = \dot{\mathbf{x}}_i(t) + \frac{\Delta t}{2m_i}\left(\mathbf{f}_i(t) + \mathbf{f}_i(t + \Delta t)\right) + \mathcal{O}\left(\Delta t^3\right). \tag{2.8}$$

Simulating the equations of motion provides a trajectory resembling the microcanonical ensemble with constant energy E for N_p particles and constant volume V. Thermostats [313, 314] are employed to simulate a trajectory preserving other statistical ensembles, for example, for the canonical ensemble that preserves the temperature T instead of the total energy E. The number of particles N_p and volume V are also preserved in canonical and microcanonical ensembles.

All reference trajectories produced in this work utilize the velocity-Verlet algorithm, which is implemented in different MD packages, for example, LAMMPS [315] and GROMACS [316–322], and OpenMM [44].

For more details on alternative time integration methods, incuding the advantages and disadvantages, refer to Refs. [323, 324]. For higher-order methods, see Ref. [325].

2.1.2 Equilibrium statistical mechanics

Statistical mechanics provide a link between equilibrium thermodynamic properties and microscopic details, i.e., the positions of atoms. Multiple microscopic configurations yield identical macroscopic properties [326], so for a large system, it is unnecessary to know explicitly the particle motions in detail.[1] Different ensembles, corresponding to different partition functions or normalization constants, can yield the same probability distributions of microscopic states [327].

We discuss here the probability distribution and corresponding partition function or normalization constant for the, in our case pertinent, canonical ensemble. More details on other ensembles can be found in Refs. [328, 329].

Atomistic states that are characterized by the generalized coordinates $\mathbf{x} \in \mathcal{M}_\mathrm{f} \subset \mathbb{R}^{n_\mathrm{f}}$, with $n_\mathrm{f} = \dim(\mathbf{x})$, follow the Boltzmann–Gibbs density for the canonical ensemble:

$$p_\mathrm{f}(\mathbf{x}; \beta) = \frac{1}{Z_\mathrm{f}(\beta)} e^{-\beta U_\mathrm{f}(\mathbf{x})}, \tag{2.9}$$

where the inverse temperature $\beta = 1/k_B T$ for the Boltzmann constant k_B. The temperature-dependent denominator in Equation 2.9, $Z_\mathrm{f}(\beta)$, normalizes the probability of states and is referred to as the *partition function* or normalization constant:

$$Z_\mathrm{f}(\beta) = \int_{\mathcal{M}_\mathrm{f}} e^{-\beta U_\mathrm{f}(\mathbf{x})}\, d\mathbf{x}. \tag{2.10}$$

[1] Unless one is interested in dynamics.

For atomistic systems, the integration involved in determining $Z_f(\beta)$ in Equation 2.10 depends on high-dimensional $\mathbf{x}$, making it impractical. MC methods can circumvent the estimation of $Z_f(\beta)$ and allow samples $\boldsymbol{x} \sim p_f(\mathbf{x};\beta)$ to be drawn from the ensemble distribution [330]. The *Helmholtz free energy*, computed by utilizing the the partition function Z_f,

$$F = -\beta^{-1} \log Z_f, \tag{2.11}$$

is a fundamental quantity in molecular systems, since it completely defines the internal energy or the entropy of the system [330, 331].

The distribution of the fine-scale coordinates $\mathbf{x}$ introduced in Equation 2.9 is also referred to as the *target* distribution, since we want to recover it using approximate distributions. Therefore, the notation $p_{\text{target}}(\mathbf{x};\beta)$ with subscript "target" refers to the fine-scale distribution $p_f(\mathbf{x};\beta)$.

2.1.3 Force fields

The force fields employed are traditionally split into bonded and non-bonded interactions. Non-bonded interactions, with contributions from one-body, pairwise, three-body, and even, higher-order interactions, contribute significantly to the overall computational cost per time step [332]. Considering non-bonded interactions up to a limited distance, or cut-off radius r_{cut}, balances the computational burden and accuracy of simulations, as critically discussed in Ref. [333].

We employ in this work well established force fields. For simulations of water, we use the SPC/E model [334–336] and for peptides, AMBER ff96 [337–339]. Clearly, a variety of force fields have been optimized for different fine-scale characteristics, as discussed for peptides in Ref. [340]. Investigating different fine-scale force fields is, however, out of scope for this work.

To demonstrate the different contributions and their physical motivation, we denote the general form of the AMBER force field as

$$\begin{aligned} U_{\text{AMBER}}(\mathbf{x}) = &\sum_{\text{bonds}} K_r \left(r(\mathbf{x}) - r_{\text{eq}}\right)^2 + \sum_{\text{angles}} K_\theta \left(\theta(\mathbf{x}) - \theta_{\text{eq}}\right)^2 \\ &+ \sum_{\text{dihedrals}} \frac{K_n}{2}[1 + \cos(n\phi(\mathbf{x}) - \gamma(\mathbf{x}))] \\ &+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}(\mathbf{x})^{12}} + \frac{q_{ij}}{R_{ij}(\mathbf{x})^6} + \frac{q_{ij}}{\epsilon R_{ij}(\mathbf{x})}\right] \\ &+ \sum_{\text{H-bonds}} \left[\frac{C_{ij}}{R_{ij}(\mathbf{x})^{12}} - \frac{D_{ij}}{R_{ij}(\mathbf{x})^{10}}\right] \end{aligned} \tag{2.12}$$

Above summations encompass dihedral bonds including four atoms, angular bonds involving three bonds and pairwise harmonics. All components in Equation 2.12 that depend on the atomic positions $\mathbf{x}$ are either distances between particles ($r(\mathbf{x})$ and $R_{ij}(\mathbf{x})$) or angles ($\boldsymbol{\theta}(\mathbf{x})$, $\phi(\mathbf{x})$, and $\gamma(\mathbf{x})$). Components with subscripts $(\cdot)_{\text{eq}}$ in Equation 2.12 refer to the equilibrium state and are treated as parameters.

$\{\epsilon, q_{ij}, A_{ij}, C_{ij}, D_{ij}\}$ are parameters for non-bonded interactions, whereas $\{K_r, K_\theta, K_n\}$ represent the stiffnesses of the bonds considered [332].

2.1.4 Calculating quantities of interest

Let us denote the function $a(\mathbf{x}) : \mathcal{M}_\mathrm{f} \to \mathbb{R}$. It corresponds to a physical observable (e.g., structural pairwise distances). Then, we can utilize the ensemble approach and compute macroscopic quantities of interest as phase averages:

$$\langle a \rangle_{p_\mathrm{f}} = \int_{\mathcal{M}_\mathrm{f}} a(\mathbf{x}) p_\mathrm{f}(\mathbf{x}; \beta)\, d\mathbf{x}. \tag{2.13}$$

where $\langle \cdot \rangle_p$ represents a phase average or expectation, which is equivalent to the notation $\mathbb{E}_p[\cdot]$. In the MD approach, quantities of interest are computed as time averages given a trajectory $\{\mathbf{x}(t) | t = 0, \ldots, \tau\}$:

$$\bar{a}_\tau = \frac{1}{\tau} \int_0^\tau a\left(\mathbf{x}(t)\right)\, dt. \tag{2.14}$$

In the limit, according to the ergodic hypothesis, quantities obtained by ensemble averages (Equation 2.13) are equivalent to quantities estimated as time averages (Equation 2.14):

$$\lim_{\tau \to \infty} \bar{a}_\tau = \langle a \rangle_{p_\mathrm{f}}. \tag{2.15}$$

For more details on the ergodic hypothesis, see Refs. [97, 327].

Clearly, the integral in equation (2.13) is analytically intractable and also numerically impractical even when applying Markov chain-based approximators, which we will introduce in Section 2.6. Estimating Equation 2.14 requires long and computationally expensive trajectories produced from MD simulations. In MC and MD approaches, the prohibitively large computational cost directly relates to the high-dimensional phase space $\mathbf{x}$ [117]. Instead of estimating quantities of interest $a(\mathbf{x})$ that depend on FG states $\mathbf{x}$, utilizing CG surrogate models reduces the amount of computation since they reduce the dimensionality of the phase space of the reference fine-scale system. Thus, simulating and estimating observables for a reduced CG description provides a speedup.

2.2 Coarse-graining

As well as an efficient means of estimating quantities of interest expressed in terms of CG variables $\mathbf{z}$, we develop a CG methodology for estimating expectations of any observable defined with regards to fine-scale coordinates $\mathbf{x}$. Existing approaches limit the estimation of observables depending on only CG variables $\mathbf{z}$. They do not allow us to reason about fine-scale coordinates $\mathbf{x}$ and thus, connected observables depending on $\mathbf{x}$, as emphasized in the introduction in Section 1.2. However, before

considering the novelties of the proposed CG approaches, we introduce the required notation.

We denote CG coordinates as $\mathbf{z}$, with $\mathbf{z} \in \mathcal{M}_c \subset \mathbb{R}^{n_c}$. The CG variables have a lower dimensionality compared to the fine-scale coordinates, and thus, $n_c \ll n_f$. The CG variables $\mathbf{z}$ interact with a corresponding CG potential $U_c(\mathbf{z})$, so we can introduce the CG density,

$$p_c(\mathbf{z}) = \frac{1}{Z_c} e^{-\beta U_c(\mathbf{z})}, \tag{2.16}$$

and the normalization constant,

$$Z_c = \int_{\mathcal{M}_c} e^{-\beta U_c(\mathbf{z})} d\mathbf{z}. \tag{2.17}$$

Thus far has, we have not introduced any connection between fine-scale coordinates $\mathbf{x}$ and coarse-scale variables $\mathbf{z}$. As mentioned in the introduction in Section 1.2, most CG approaches rely on an explicit restriction, i.e., a fine-to-coarse map $\mathcal{R} : \mathcal{M}_f \to \mathcal{M}_c$. CG variables are defined by the restriction $\mathbf{z} = \mathcal{R}(\mathbf{x})$. In general, this restriction maps multiple fine-scale states to identical CG variables. This implies a many-to-one map, which is not invertible for obtaining FG variables $\mathbf{x}$ given a CG state $\mathbf{z}$ [194].

2.2.1 Consistency in CG models

Given the restriction $\mathcal{R}$ and the reference ensemble probability distribution $p_f(\mathbf{x})$, the many-body PMF defines the optimal CG interaction potential. If we are interested in expectation values of observables that depend directly on the CG variables $\mathbf{z}$, then the fine-scale observable $a(\mathbf{x})$ can be replaced by an observable expressed in terms of CG variables $A(\mathcal{R}(\mathbf{x})) = A(\mathbf{z})$. The following equality holds:

$$a(\mathbf{x}) = A\left(\mathcal{R}(\mathbf{x})\right) = A(\mathbf{z}). \tag{2.18}$$

With this equality, we can express the estimation of observables in terms of CG variables as follows [145, 341, 342]:

$$\begin{aligned}
\mathbb{E}_{p_f}[a] &= \int_{\mathcal{M}_f} a(\mathbf{x}) p_f(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{M}_f} A\left(\mathcal{R}(\mathbf{x})\right) p_f(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{M}_f} \left(\int_{\mathcal{M}_c} A(\mathbf{z}) \delta\left(\mathbf{z} - \mathcal{R}(\mathbf{x})\right)\, d\mathbf{z} \right) p_f(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{M}_c} A(\mathbf{z}) \underbrace{\left(\int_{\mathcal{M}_f} \delta\left(\mathbf{z} - \mathcal{R}(\mathbf{x})\right) p_f(\mathbf{x})\, d\mathbf{x} \right)}_{= p_c^{\mathrm{opt}}(\mathbf{z})} d\mathbf{z} \\
&= \int_{\mathcal{M}_c} A(\mathbf{z}) p_c^{\mathrm{opt}}(\mathbf{z}) d\mathbf{z}.
\end{aligned} \tag{2.19}$$

Taking a closer look at the second last line in Equation 2.19, we identified the optimal CG density $p_c^{opt}(\mathbf{z})$, which can be obtained by marginalization with respect to $\mathbf{x}$:

$$p_c^{opt}(\mathbf{z}) = \int_{\mathcal{M}_f} \delta(\mathbf{z} - \mathcal{R}(\mathbf{x})) p_f(\mathbf{x})\, d\mathbf{x}. \tag{2.20}$$

Connecting Equation 2.20 with the expression that includes a CG potential U_c, Equation 2.16, we obtain the many-body PMF with

$$U_c^{opt}(\mathbf{z}) = -\frac{1}{\beta} \log \int_{\mathcal{M}_f} \delta(\mathbf{z} - \mathcal{R}(\mathbf{x})) p_f(\mathbf{x})\, d\mathbf{x}. \tag{2.21}$$

While CG approaches provide more or less accurate approximations of $U_c^{opt}(\mathbf{z})$, finding an exact estimate of $U_c^{opt}(\mathbf{z})$ is computationally challenging and usually an intractable task. Even if we assume that we could obtain $U_c^{opt}(\mathbf{z})$, simulating the CG potential would lead to exact quantities of interest only for observables fulfilling $a(\mathbf{x}) = A(\mathcal{R}(\mathbf{x})) = A(\mathbf{z})$. However, many observables depend explicitly on fine-scale coordinates, for example, correlations between atoms within CG macromolecules. In that case it is necessary to reconstruct the fully atomistic picture $\mathbf{x}$, given CG variables $\mathbf{z}$, which are sampled from $p_c(\mathbf{z})$.

Thus far, CG methodologies have not been able to provide fine-scale coordinates $\mathbf{x}$ given a CG variable $\mathbf{z}$. Nevertheless, when introducing a consistent pseudo-reconstruction mapping, it must be assumed that all $\mathbf{x}$ leading to the same $\mathbf{z}$ are equally probable and therefore, uniformly distributed in $\mathcal{M}_f$, given a CG variable $\mathbf{z}$. We can, thus, write the conditional distribution:

$$p_{\mathcal{R}}(\mathbf{x}|\mathbf{z}) = \frac{\delta\left(\mathbf{z} - \mathcal{R}(\mathbf{x})\right)}{Z_{\mathcal{R}}(\mathbf{z})}, \tag{2.22}$$

with the normalization, counting the realizations $\mathbf{x}$ leading to the same $\mathbf{z}$:

$$Z_{\mathcal{R}}(\mathbf{z}) = \int \delta(\mathbf{z} - \mathcal{R}(\mathbf{x})) d\mathbf{x}. \tag{2.23}$$

With the probability density function (PDF) for the CG variables $p_c(\mathbf{z})$ (Equation 2.16) and the conditional PDF needed for a consistent reconstruction in Equation 2.22, the all-atom PDF is [184]:

$$\begin{aligned} p_{\mathcal{R}}(\mathbf{x}) &= \int p_{\mathcal{R}}(\mathbf{x}|\mathbf{z}) p_c(\mathbf{z}) d\mathbf{z} \\ &= \int \frac{\delta(\mathbf{z} - \mathcal{R}(\mathbf{x}))}{Z_{\mathcal{R}}(\mathbf{z})} p_c(\mathbf{z}) d\mathbf{z} \\ &= \frac{p_c(\mathcal{R}(\mathbf{x}))}{Z_{\mathcal{R}}(\mathcal{R}(\mathbf{x}))}. \end{aligned} \tag{2.24}$$

2.2.2 Relative entropy CG approaches

With the CG density defined on the FG manifold $\mathcal{M}_\mathrm{f}$ in Equation 2.24, relative entropy CG approaches provide an information theoretic framework[2] for comparing the density $p_\mathcal{R}$, which is defined by the CG potential U_c, $\mathcal{R}$, and the reference ensemble density p_f. Relative entropy CG approaches utilize a metric that quantifies the deviation from a reference or target PDF p_f (also denoted as p_target) to the model PDF $p_\mathcal{R}$, the Kullback–Leibler (KL) divergence or relative entropy [345, 346]:

$$\begin{aligned} 0 &\leq D_\mathrm{KL}\left(p_\mathrm{f}(\mathbf{x}) \| p_\mathcal{R}(\mathbf{x})\right) \\ &= -\int p_\mathrm{f}(\mathbf{x}) \log \frac{p_\mathcal{R}(\mathbf{x})}{p_\mathrm{f}(\mathbf{x})}\, d\mathbf{x} \\ &= -\mathbb{E}_{p_\mathrm{f}(\mathbf{x})}\left[\log p_\mathrm{c}(\mathcal{R}(\mathbf{x})\right] + \mathbb{E}_{p_\mathrm{f}(\mathbf{x})}\left[\log Z_\mathcal{R}\left(\mathcal{R}(\mathbf{x})\right)\right] - \mathbb{H}\left(p_\mathrm{f}\right). \end{aligned} \tag{2.25}$$

The entropy follows with

$$\mathbb{H}\left(p_\mathrm{f}\right) = -\int p_\mathrm{f}(\mathbf{x}) \log p_\mathrm{f}(\mathbf{x})\, d\mathbf{x}. \tag{2.26}$$

Note that only the first term in the objective in Equation 2.25 can be optimized with respect to U_c since $\mathcal{R}$ is fixed. The second term is fixed once we prescribe a fine-to-coarse mapping, since it depends solely on $Z_\mathcal{R}$ and thus, $\mathcal{R}$. The expectation $\mathbb{E}_{p_\mathrm{f}(\mathbf{x})}[\log Z_\mathcal{R}(\mathcal{R}(\mathbf{x}))]$ is a fixed penalty that relates the fine-to-coarse mapping to the associated information loss due to the reduced coordinate representation [179, 200]. The original work [231] calls this term the *mapping entropy*.

2.3 Predictive modeling and uncertainty quantification

In the following, we review relevant[3] components of probability theory from a Bayesian perspective, which provides a consistent approach for plausible reasoning. Bayesian inference enables probabilistic conclusions to be drawn from limited data and the consistent use of new sequential evidence. A Bayesian approach allows us to incorporate speculative information or prior knowledge that is available before we analyze the raw data [347, 348]. In general, uncertain quantities, such as model parameters, are treated in a Bayesian approach as random variables, which enables the propagation and quantification of the associated uncertainty [349].

Using numerical values to express the degree of belief and including coherent rules to adjusting these quantities was done in Refs. [350, 351]. These laws are equivalent to the sum and product rules in provability theory. For a broader view and for discussions that differ from the Bayesian perspective, such as frequentist and maximum entropy approaches, refer to Refs. [352–355].

[2] An interesting connection between information theory and statistical mechanics was made in the middle of the last century [343, 344].

[3] Relevant to this work's developments and contributions in context of CG approaches.

2.3.1 Bayes' theorem

Bayes' theorem allows us to combine information from some prior notion of uncertain model parameters $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$, expressed as a *prior* distribution $p(\boldsymbol{\theta})$, with evidence obtained from observed data $\mathbf{x}^{\mathcal{D}_N} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. Bayes' theorem transforms the prior distribution into a posterior distribution by incorporating observed evidence. Observations have a conditional distribution $p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta})$, which indicates the *likelihood* of the observed data and thus, how probable or how well the data are explained for given parameters $\boldsymbol{\theta}$. Bayes' rule [347] combines a prior and a likelihood:

$$p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) = \frac{p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x}^{\mathcal{D}_N})}. \tag{2.27}$$

The *posterior* distribution $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ accounts for the credibility of a realization of $\boldsymbol{\theta}$ after the data $\mathbf{x}^{\mathcal{D}_N}$ have been observed and expresses the uncertainty in terms of a distribution over $\boldsymbol{\theta}$ [356]. The denominator in Equation 2.27 accounts for normalization. It yields a proper distribution function of the posterior:

$$p(\mathbf{x}^{\mathcal{D}_N}) = \int p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{2.28}$$

The measure $p(\mathbf{x}^{\mathcal{D}_N})$ in Equation 2.28 is also known as model evidence. It provides an estimate of the probability of observing the given data independent of the parameters $\boldsymbol{\theta}$. The evidence $p(\mathbf{x}^{\mathcal{D}_N})$ is required to obtain the posterior distribution in which we are interested in Bayesian inference. However, determining $p(\mathbf{x}^{\mathcal{D}_N})$ is a computationally intractable problem because it requires an integration in the high-dimensional parameter space of $\mathbb{R}^{d_\theta}$. Approximation methods [357, 358], such as MC methods (Section 2.6) and modern variational inference (Section 2.5), may be useful.

Including prior knowledge expressed by prior distributions raises the obvious question about which probability distribution function is appropriate. As stated in Ref. [359, 360], the subjective choice of prior distributions is a criticism of Bayesian inference. We recommend the Bayesian approach since the influence of the prior on the posterior distribution decays as further evidence, such as observed data, is included. Even if the suspected knowledge is not appropriate, the data will correct for it. Moreover, if we have some understanding of the system of interest expressed by prior distributions, parameter learning can cope with fewer data.

As in sharing common parameters for modeling the likelihood $p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta})$, the parameters $\boldsymbol{\theta}$ themselves may share a common hyperparameter γ that is specified by a prior distribution $p(\gamma)$. This provides a coherent approach for hierarchical extensions of Bayesian models. Assuming the conditional independence of θ_k given γ, a hierarchical extension is

$$p(\boldsymbol{\theta}) = \int p(\gamma) \prod_{k=1}^{d_\theta} p(\theta_k|\gamma)\, d\gamma. \tag{2.29}$$

Hierarchical Bayesian models can reduce the need to tune hyperparameters or they imply a structure [298, 361, 362]. Such an extension can be used for hierarchical functional priors. It supports machine learning of sparse models to unveil physically relevant features, which we are interested in in this work [363–366].

Finally, Bayesian inference provides a consistent framework for drawing probabilistic conclusions expressed as posterior probability distributions (i.e., $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$) when making predictive estimates of quantities of interest.

2.3.2 Bayesian learning and prediction

We are interested in making predictions after including priors and observations. For this purpose, we utilize the identified data-producing process $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ with

$$p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})\, d\boldsymbol{\theta}, \tag{2.30}$$

which is the *predictive* distribution. Again, solving the integral in Equation 2.30 requires enormous computational effort.

In addition to learning from data, Bayesian modeling can be generalized for model comparison, averaging, and selection tasks. Suppose we are interested in comparing K models $\{m_k\}_{k=1}^K$. Given speculation in the form of a prior distribution for a model $p(m_k)$, we seek to obtain the posterior distribution of the model m_k given the observations $\mathbf{x}^{\mathcal{D}_N}$:

$$p(m_k|\mathbf{x}^{\mathcal{D}_N}) = \frac{p(\mathbf{x}^{\mathcal{D}_N}|m_k)p(m_k)}{p(\mathbf{x}^{\mathcal{D}_N})}, \tag{2.31}$$

for the model evidence

$$p(\mathbf{x}^{\mathcal{D}_N}|m_k) = \int p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}, m_k)p(\boldsymbol{\theta}|m_k)\, d\boldsymbol{\theta}. \tag{2.32}$$

This can be interpreted as the probability of generating the observations $\mathbf{x}^{\mathcal{D}_N}$ for a model m_k where the model parameters are sampled from the prior distribution $p(\boldsymbol{\theta}|m_k)$ [367, 368].

Model averaging is a compelling way to combine individual models for predictive purposes. The predictive distributions of individual models (i.e., Equation 2.30) are then weighted according to their posterior probability $p(m_k|\mathbf{x}^{\mathcal{D}_N})$:

$$p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N}) = \sum_{k=1}^{K} p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N}, m_k)p(m_k|\mathbf{x}^{\mathcal{D}_N}). \tag{2.33}$$

Simpler approximations of Equation 2.33 utilize the most probable model with maximal $p(m_k|\mathbf{x}^{\mathcal{D}_N})$ [356, 369].

2.4 Probabilistic generative models

Probabilistic modeling supplemented by deep learning leverages a variety of research avenues [221, 297, 370]. In this work, we focus on developing methodologies for learning CG models while extracting physically relevant features in terms of the CG potential or on unveiling the slow coordinates of the system in the absence of any model-specific insight. We use an unsupervised learning approach [371] and rely, in particular, on probabilistic generative models [372, 373].

The quote from Richard Feynman, "What I cannot create, I do not understand," summarizes well the motivation behind generative models: to understand the hidden processes that give rise to atomistic observations $\mathbf{x}^{\mathcal{D}_N}$ obtained from the reference fine-scale Boltzmann distribution $p_{\mathrm{f}}(\mathbf{x})$. The motion of even a complex atomistic system with high-dimensional $\mathbf{x}$ is governed by a set of a few slow or collective variables (CVs) [374]. Thus, the CVs are potent candidates, serving as an efficient parsimonious lower-dimensional representation of $\mathbf{x}$. Although CVs and CG variables $\mathbf{z}$ are unobserved, they give rise to the observed atomistic coordinates $\mathbf{x}$, and thus, we are interested in learning the hidden features that lead to the observed fine-scale samples $\boldsymbol{x} \sim p_{\mathrm{f}}(\mathbf{x})$.

We introduce a joint probability distribution for the observed atomistic coordinates $\mathbf{x}$ and latent CG variables $\mathbf{z}$:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})\, p(\mathbf{z}), \tag{2.34}$$

where $\dim(\mathbf{z}) \ll \dim(\mathbf{x})$ [372]. Extending the probability space with latent variables next to the observed variables facilitates the expressiveness of the joint distribution. Furthermore the hierarchical structure, with lower-dimensional latent CG variables $\mathbf{z}$, promotes the unveiling of physically relevant features, as we show in the following chapters. The probabilistic model introduced in Equation 2.34 has two components:

(i) The density $p(\mathbf{z})$ is a generator for and describes the distribution of the CG variables $\mathbf{z}$.

(ii) The coarse-to-fine mapping is probabilistic and implicitly defines CG variables by the conditional probability distribution $p(\mathbf{x}|\mathbf{z})$.

Instead of defining a fine-to-coarse mapping operator, in this approach CG variables are regarded as latent generators that lead, through the probabilistic coarse-to-fine mapping $p(\mathbf{x}|\mathbf{z})$, to the observable fine-scale atomistic picture $\mathbf{x}$. Flexibility is provided through the two components of the model: (1) in the form of the coarse-to-fine mapping $p(\mathbf{x}|\mathbf{z})$ and (2) with regard to the CG description $p(\mathbf{z})$. For example, the coarse-to-fine mapping could have a simple linear form (e.g., Gaussian) or a flexible non-linear mapping (e.g., Gaussian, with mean and variance obtained from an expressive neural network). Additionally, we can think of multiple coarse-to-fine

mappings as each being responsible for a subset of the CG variables $\mathbf{z}$, which corresponds to a local mapping. Similar flexibility provides the coarse description $p(\mathbf{z})$, which is hierarchically extendable by introducing multiple layers corresponding to different levels of CG approaches [375].

FG configurations $\boldsymbol{x}^{(i)}$ are produced by a generative process:

(i) Obtain a CG realization $\boldsymbol{z}^{(i)}$ from the CG distribution: $\boldsymbol{z}^{(i)} \sim p(\mathbf{z})$.

(ii) Draw a sample $\boldsymbol{x}^{(i)}$ with $\boldsymbol{x}^{(i)} \sim p(\mathbf{x}|\boldsymbol{z}^{(i)})$.

The generative framework endows the CG framework with a truly predictive probability density in the sense of producing new fine-scale atomistic realizations, which, when $a(\mathbf{x}) \neq A(\mathbf{z})$, can be utilized to estimate the observables depending on the fine-scale resolution. The predictive distribution follows from a marginalization of Equation 2.34:

$$p(\mathbf{x}) = \int_{\mathcal{M}_\mathrm{c}} p(\mathbf{x},\mathbf{z})\, d\mathbf{z} = \int_{\mathcal{M}_\mathrm{c}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\, d\mathbf{z}. \tag{2.35}$$

We parametrize the densities with $\boldsymbol{\theta}_\mathrm{cf}$ and $\boldsymbol{\theta}_\mathrm{c}$ and thus, denote the parametrized probability distributions as $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}_\mathrm{cf})$ and $p(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c})$.[4] The parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\mathrm{cf}, \boldsymbol{\theta}_\mathrm{c}\}$ summarizes all parameters used in the generative model.

We are interested in comparing the previously introduced relative entropy CG procedure with the approach proposed in this work. Therefore, we start from an information theoretic viewpoint in optimizing the generative distribution $p(\mathbf{x}|\boldsymbol{\theta})$. The KL divergence[5] quantifies the difference between the target PDF $p_\mathrm{target}(\mathbf{x})$ (which is the fine-scale Boltzmann distribution $p_\mathrm{target}(\mathbf{x}) \equiv p_\mathrm{f}(\mathbf{x})$) and the generative model $p(\mathbf{x})$:

$$\begin{aligned} D_\mathrm{KL}\left(p_\mathrm{target}(\mathbf{x}) \| p(\mathbf{x}|\boldsymbol{\theta})\right) &= -\int_{\mathcal{M}_\mathrm{f}} p_\mathrm{target}(\mathbf{x}) \log \frac{p(\mathbf{x}|\boldsymbol{\theta})}{p_\mathrm{target}(\mathbf{x})}\, d\mathbf{x} \\ &= -\int_{\mathcal{M}_\mathrm{f}} p_\mathrm{target}(\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} \\ &\quad + \int_{\mathcal{M}_\mathrm{f}} p_\mathrm{target}(\mathbf{x}) \log p_\mathrm{target}(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{2.36}$$

This equation provides an objective function that is minimized in identifying the optimal parametrization $\boldsymbol{\theta}_\mathrm{KL} = \arg\min_{\boldsymbol{\theta}} D_\mathrm{KL}\left(p_\mathrm{target}(\mathbf{x}) \| p(\mathbf{x}|\boldsymbol{\theta})\right)$.

Minimizing the KL divergence in Equation 2.36 is the same as maximizing the component $\int_{\mathcal{M}_\mathrm{f}} p_\mathrm{target}(\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}$. One can show, by approximating the integral empirically with

$$\frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{x} - \boldsymbol{x}^{(i)}) \log p(\mathbf{x}|\boldsymbol{\theta})$$

[4] To meet the guidelines for a journal, the notation $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}_\mathrm{cf})$ and $p(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c})$ was modified to $p_{\boldsymbol{\theta}_\mathrm{cf}}(\mathbf{x}|\mathbf{z})$ and $p_{\boldsymbol{\theta}_\mathrm{c}}(\mathbf{z})$ and both forms will be used interchangeably in this thesis.

[5] The KL divergence is one of many metrics from the family of α divergences for quantifying the closeness between two PDFs. See Refs. [376–378] for a more in-depth consideration.

with samples $\boldsymbol{x}^{(i)} \sim p_{\text{target}}(\mathbf{x})$, that maximizing the aforementioned empirical estimator is the same as, up to the multiplicative constant $1/N$, maximizing the marginal log-likelihood of $p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta})$ with a given dataset $\mathbf{x}^{\mathcal{D}_N} = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$:

$$\begin{aligned}\log p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \int_{\mathcal{M}_c} p_{\boldsymbol{\theta}_{cf}}(\boldsymbol{x}^{(i)}|\mathbf{z}^{(i)}) p_{\boldsymbol{\theta}_c}(\mathbf{z}^{(i)})\, d\mathbf{z}^{(i)},\end{aligned} \tag{2.37}$$

since $p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}) = \prod_{i=1}^N p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$. Equation 2.37 involves the observed fine-scale atomistic description $\boldsymbol{x}^{(i)}$ and the corresponding random variable of the latent variable $\mathbf{z}^{(i)}$, which we interpret as a lower-dimensional pre-image of the observed atomistic configuration $\boldsymbol{x}^{(i)}$. The maximum likelihood (ML) estimate arises by maximizing $\log p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta})$ [379, 380]:

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}) \right\}. \tag{2.38}$$

Compared to regarding the optimization from an information theoretic viewpoint, as introduced in Equation 2.36, the log-likelihood in Equation 2.37 can be embedded in a Bayesian framework. Thus, prior information $p(\boldsymbol{\theta})$ can be included, which leads to the maximum a posteriori (MAP) estimate [381]:

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \right\}. \tag{2.39}$$

The posterior distribution of the model parametrization $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ is relevant for propagating uncertainties to observables. It is expressed by applying Bayes' rule:

$$\begin{aligned}p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) &= \frac{p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{x}^{\mathcal{D}_N})} \\ &= \frac{\prod_{i=1}^N \left(\int_{\mathcal{M}_c} p_{\boldsymbol{\theta}_{cf}}(\boldsymbol{x}^{(i)}|\mathbf{z}^{(i)}) p_{\boldsymbol{\theta}_c}(\mathbf{z}^{(i)})\, d\mathbf{z}^{(i)} \right) p(\boldsymbol{\theta})}{p(\mathbf{x}^{\mathcal{D}_N})}.\end{aligned} \tag{2.40}$$

The posterior $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ for interesting systems is intractable but approximation methods are discussed in Sections 2.6 and 2.5.4. The parameter credibility after having seen data $\mathbf{x}^{\mathcal{D}_N}$ is represented as a posterior density and is propagated to predictions with predictive posterior $p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N})$, which involves the marginalization of hidden

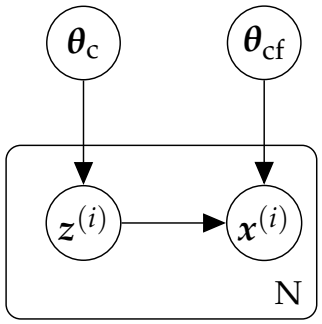


FIGURE 2.1: Representation as a probabilistic graphical model.

variables $\mathbf{z}$ and parameters $\boldsymbol{\theta}$:

$$\begin{aligned} p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N}) &= \int \underbrace{p(\mathbf{x}|\boldsymbol{\theta})}_{\text{Equation 2.35}} \; p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \, d\boldsymbol{\theta} \\ &= \int \left(\int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \, d\mathbf{z} \right) p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \, d\boldsymbol{\theta} \\ &= \int \left(\int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf}) p(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c}) \, d\mathbf{z} \right) p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \, d\boldsymbol{\theta} \\ &= \int \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf}) p(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c}) p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \, d\mathbf{z} \, d\boldsymbol{\theta}. \end{aligned} \tag{2.41}$$

This equation is concisely represented in Figure 2.1 as a directed graphical model. It enables the application of the inference models as graphical models, as discussed in Refs. [372, 382].

As well as generating atomistic representations $\boldsymbol{x}^{(i)}$, Equation 2.41 is employed for efficiently approximating expectation values of observables, as introduced in Equation 2.19, by replacing the reference fine-scale Boltzmann distribution $p_\mathrm{target}(\mathbf{x})$ with the predictive distribution $p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N})$:

$$\begin{aligned} \mathbb{E}_{p_\mathrm{target}(\mathbf{x})}[a(\mathbf{x})] &\approx \mathbb{E}_{p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N})}[a(\mathbf{x})] \\ &= \int a(\mathbf{x}) p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N}) \, d\mathbf{x} \\ &= \int a(\mathbf{x}) \left(\int p_\mathrm{cf}(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf}) p_\mathrm{c}(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c}) p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \, d\mathbf{z} \, d\boldsymbol{\theta} \right) d\mathbf{x} \\ &= \int \left(\int a(\mathbf{x}) p_\mathrm{cf}(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf}) p_\mathrm{c}(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c}) \, d\mathbf{z} \, d\mathbf{x} \right) p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \, d\boldsymbol{\theta} \\ &= \int \underbrace{\left(\int a(\mathbf{x}) p_\mathrm{cf}(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf}) p_\mathrm{c}(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c}) \, d\mathbf{z} \, d\mathbf{x} \right)}_{\hat{a}(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \, d\boldsymbol{\theta} \\ &= \int \hat{a}(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \, d\boldsymbol{\theta}. \end{aligned} \tag{2.42}$$

Clearly, the estimator $\mathbb{E}_{p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N})}[a(\mathbf{x})]$ is an approximation to $\mathbb{E}_{p_\mathrm{target}(\mathbf{x})}[a(\mathbf{x})]$, since we

use finite data and imply models for the distributions. However, the above expressions incorporate the aforementioned approximations in terms of the posterior density $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$, which are propagated through the observable estimator $\hat{a}(\boldsymbol{\theta})$. Given a set of parameters $\boldsymbol{\theta}^{(i)}$ sampled from $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$, the predictive estimator $\hat{a}(\boldsymbol{\theta}^{(i)})$ provides the corresponding observable. One can either compute the expected value of $\hat{a}(\boldsymbol{\theta})$, as indicated in the last line of Equation 2.42, or present the uncertainty as credible intervals around the MAP or ML estimate [383, 384]. The latter would imply replacing $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ with $\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP/ML}})$.

As we discussed in Section 2.3.1, a Bayesian approach has several advantages. Thus, we address in the following section Bayesian inference algorithms that facilitate the drawing of conclusions about the latent variables $\mathbf{z}$ and estimating the (approximate) posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$, which is utilized in making predictions and in uncertainty quantification.

2.5 Inference

In the following, we discuss methods facilitating inference in latent variable models. Without loss of generality, we focus on the model introduced in Section 2.4 but collective approaches are applicable to any latent variable models. Note that the seminal work is Ref. [357].

2.5.1 Point-based approximations: ML and MAP

Point-based ML and MAP estimates are the simplest way to approximate the posterior $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$. They avoid the computationally expensive integration needed to obtain the fully predictive posterior of Equation 2.41. In latent variable models, we usually have a set of observed data $\mathbf{x}^{\mathcal{D}_N} = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$ and a set of hidden variables $\mathbf{z}^{\mathcal{D}_N} = \{\boldsymbol{z}^{(i)}\}_{i=1}^N$, which gives rise to the observed dataset $\mathbf{x}^{\mathcal{D}_N}$ through the generative model. The observed and latent variables are probabilistically connected through the parameters $\boldsymbol{\theta}$. This leads to the marginal log-likelihood, Equation 2.37:

$$\log p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int_{\mathcal{M}_c} p(\boldsymbol{x}^{(i)}|\mathbf{z}^{(i)}, \boldsymbol{\theta}_{\text{cf}}) p(\mathbf{z}^{(i)}|\boldsymbol{\theta}_c)\, d\mathbf{z}^{(i)}.$$

The log-likelihood is a function of the model parameters $\boldsymbol{\theta}$ given the observed data $\mathbf{x}^{\mathcal{D}_N}$:

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N}) \equiv \log p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}). \tag{2.43}$$

For non-trivial densities, the hidden variables and thus, the induced probabilistic dependency between model parameters, latent variables, and observables makes direct optimization of the marginal (log-)likelihood in Equation 2.43 intractable. Every change to $\boldsymbol{\theta}$ would require solving the computationally prohibitive integral

$$\int_{\mathcal{M}_c} p(\boldsymbol{x}^{(i)}|\mathbf{z}^{(i)}, \boldsymbol{\theta}_{\text{cf}}) p(\mathbf{z}^{(i)}|\boldsymbol{\theta}_c)\, d\mathbf{z}^{(i)}$$

for every datum $\boldsymbol{x}^{(i)}$.

The maximization of $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$ can be simplified by introducing a set of auxiliary distributions $\{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})\}_{i=1}^N$, which allows us to apply Jensen's inequality [385]. Each element of the summation in the marginal log-likelihood can be lower bounded:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N}) &= \sum_{i=1}^{N} \log \left(\int p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)} | \boldsymbol{\theta}) \, d\mathbf{z}^{(i)} \right) \\
&= \sum_{i=1}^{N} \log \left(\int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \frac{p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)} | \boldsymbol{\theta})}{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})} \, d\mathbf{z}^{(i)} \right) \qquad (2.44) \\
&\geq \sum_{i=1}^{N} \underbrace{\left(\int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \log \frac{p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)} | \boldsymbol{\theta})}{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})} \, d\mathbf{z}^{(i)} \right)}_{\equiv \mathcal{F}\left(q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}), \boldsymbol{\theta}; \boldsymbol{x}^{(i)}\right)} \\
&= \sum_{i=1}^{N} \mathcal{F}\left(q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}), \boldsymbol{\theta}; \boldsymbol{x}^{(i)}\right) \\
&= \mathcal{F}\left(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N}\right), \qquad (2.45)
\end{aligned}$$

with $q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}) = \prod_{i=1}^N q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})$, which refers to the whole set of free distributions, $\{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})\}_{i=1}^N$. Thus far, we have not specified these free distributions and they will have an important role, especially for the physical interpretation of latent variables.

Expectation maximization

An iterative scheme alternating between an expectation step (E step) and a maximization step (M step) was introduced by Refs. [386, 387]. The expectation maximization (EM) procedure produces distributions $\{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})\}_{i=1}^N$ given a current set of model parameters $\boldsymbol{\theta}^{(i)}$ (E step) and then maximizes the lower bounds and thus, the marginal log-likelihood, for a given set of distributions $\{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})\}_{i=1}^N$ with respect to $\boldsymbol{\theta}$ (M step). Indicating with t the iteration of a combined EM step and assuming some initial values for the parameters $\boldsymbol{\theta}^{(t)}$ at $t = 0$, we can write:

- E step: Maximize the lower bound $\mathcal{F}(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}^{(t)}; \mathbf{x}^{\mathcal{D}_N})$ with respect to every distribution over the latent variables $q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})$ given the current $\boldsymbol{\theta}^{(t)}$ to obtain $q_{\boldsymbol{x}^{(i)}}^{(t+1)}(\mathbf{z}^{(i)})$:

$$q_{\boldsymbol{x}^{(i)}}^{(t+1)}(\mathbf{z}^{(i)}) \leftarrow \underset{q_{\boldsymbol{x}^{(i)}}}{\arg\max} \mathcal{F}\left(q_{\boldsymbol{x}^{(i)}}^{(t)}(\mathbf{z}), \boldsymbol{\theta}^{(t)}; \mathbf{x}^{\mathcal{D}_N}\right), \quad \forall i \in \{i, \ldots, N\}. \qquad (2.46)$$

- M step: Maximize the lower bound $\mathcal{F}(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}^{(t)}; \mathbf{x}^{\mathcal{D}_N})$ with respect to $\boldsymbol{\theta}$ given the posterior over the latent variables $q^{(t+1)}_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z})$:

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{F}\left(q^{(t+1)}_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N}\right) \tag{2.47}$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(\int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \log p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}|\boldsymbol{\theta}) \, d\mathbf{z}^{(i)} \right). \tag{2.48}$$

The optimal distributions $q^{\text{opt}}_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})$ are obtained when they equal the true conditional posterior distribution over the latent variables $\mathbf{z}^{(i)}$ given the observation $\boldsymbol{x}^{(i)}$:

$$\begin{aligned} q^{(t+1),\,\text{opt}}_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) &= p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)}), \qquad \forall i \in \{i, \ldots, N\} \\ &\propto p(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}, \boldsymbol{\theta}^{(t)}_{\text{cf}}) p(\mathbf{z}^{(i)}|\boldsymbol{\theta}^{(t)}_{\text{c}}). \end{aligned} \tag{2.49}$$

Using the expression $q^{(t+1),\,\text{opt}}_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})$ from Equation 2.49 and by replacing the corresponding terms in Equation 2.44, we can prove that the inequality becomes an equality (with a tightened lower bound), i.e.:

$$\mathcal{F}\left(q^{(t+1),\,\text{opt}}_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}^{(t)}; \mathbf{x}^{\mathcal{D}_N}\right) = \sum_{i=1}^{N} \log p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathbf{x}^{\mathcal{D}_N}). \tag{2.50}$$

The full proof for Equation 2.50 and further details are provided in Ref. [357]. Obtaining the full posterior distribution over latent variables is computationally prohibitive and the approximation of the actual posterior in Equation 2.49, with some statistics, suffices in general.

In general and also in the context of this work, the posterior distributions

$$\left\{ p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}) \right\}_{i=1}^{N}$$

are analytically intractable due to the interdependence of the latent variables, as discussed in Refs. [388–391]. To circumvent the intractable computation of the true posterior $p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})$, we follow two approaches, one of which includes Markov chain Monte Carlo (MCMC) methods and the other variational Bayesian approaches.

Expectation maximization with MCMC E step

The M step requires the maximization of the expression in Equation 2.48:

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(\int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \log p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}|\boldsymbol{\theta}) \, d\mathbf{z}^{(i)} \right).$$

Instead of attempting to solve the above integration directly, one can introduce an approximate version of it. An approximate estimator of the aforementioned integral with a finite set of m_t samples from the posterior over the hidden variables $\boldsymbol{z}^{(i,j)} \sim$

$p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)})$ takes the form

$$\int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \log p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}|\boldsymbol{\theta}) \, d\mathbf{z}^{(i)} \approx \frac{1}{m_t} \sum_{j=1}^{m_t} \log p(\boldsymbol{x}^{(i)}, z^{(i,j)}|\boldsymbol{\theta}). \tag{2.51}$$

Clearly, the quality of such approximate estimators depends on the number of (effective) samples m_t and the quality of the sample set in the sense of the induced correlation between subsequent samples since we rely on MCMC rather than independent and identically distributed (iid) MC samples [392–394]. Relatively small changes to $\boldsymbol{\theta}^{(t)}$ offer the potential to reuse the samples obtained from the predecessor of the current EM step. The population of given samples can be re-weighted to best approximate the posterior over latent variables of the current step t instead of $t-1$ by importance sampling [395]. There are even more advanced and adaptive sampling schemes, including annealing, such as adaptive sequential MC approaches [205, 396]. MC methods will be discussed in more detail in Section 2.6.

Variational expectation maximization

Instead of introducing an E step to approximate the true posterior over latent variables by sampling $z^{(i,j)} \sim p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)})$, we can use a variational approach that constrains the auxiliary distribution $q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z})$ and that stems from a particular parametrized family of distributions such that expectation values in the lower bound $\mathcal{F}(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$ can be simplified, for example, by utilizing sufficient statistics [357]. The variational E step does not attempt to obtain the exact posterior distributions $p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}^{(t)})$ but distributions from a constrained family that resemble the exact posterior distributions [357]. We reformulate the lower bound:

$$\begin{aligned}
&\mathcal{F}(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N}) \\
&\quad = \sum_{i=1}^{N} \int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \log \frac{p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}|\boldsymbol{\theta})}{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})} \, d\mathbf{z}^{(i)} \\
&\quad = \sum_{i=1}^{N} \int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \log \frac{p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}) p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})}{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})} \, d\mathbf{z}^{(i)} \\
&\quad = \sum_{i=1}^{N} \int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \log p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}) \, d\mathbf{z}^{(i)} + \int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \log \frac{p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})}{q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})} \, d\mathbf{z}^{(i)} \\
&\quad = \sum_{i=1}^{N} \log p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}) - D_{\mathrm{KL}}\left(q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \| p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})\right).
\end{aligned} \tag{2.52}$$

According to Ref. [357] and from Equation 2.52, maximizing the lower bound $\mathcal{F}(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$ with respect to all $q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})$ is equivalent to minimizing the KL divergence between the variationally approximate posterior $q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})$ and the exact posterior $p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})$. The KL divergence is greater than or equal to zero if and only if $q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})$. In the latter case, the lower bound $\mathcal{F}(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$ gets "tight" and equals the marginal log-likelihood $\log p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta})$. In the following M

step, the parameters are optimized such that $\mathcal{F}(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$ is maximized given the current variational posterior distributions $q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})$.

The minimization of $D_{\mathrm{KL}}(q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) \| p(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}))$ and thus, the maximization of the lower bound $\mathcal{F}(q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$ with regards to all $q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)})$ becomes a minimization or maximization of the respective quantities with respect to the parametrization of the variational posterior distributions. We can introduce a set of independent parameters $\boldsymbol{\phi}^{(i)} \in \mathbb{R}^{d_\phi}$ for each datum $\boldsymbol{x}^{(i)}$ and summarize all parameters defining the variational posterior distributions with $\boldsymbol{\Phi} = \{\boldsymbol{\phi}^{(i)}\}_{i=1}^{N}$. Then we write for the posterior belonging to a particular parametrized family [397, 398][6]:

$$q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}) = q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}|\boldsymbol{\phi}^{(i)}), \quad \forall i \in \{1, \ldots, N\}. \tag{2.53}$$

By utilizing the parametrized posterior in Equation 2.53, the E step translates to maximizing the lower bound $\mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$, which depends implicitly on $q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}|\boldsymbol{\Phi})$ through $\boldsymbol{\Phi}$ with respect to the variational parameters $\boldsymbol{\Phi}$:

- E step: Update the variational parameters with:

$$\boldsymbol{\Phi}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\Phi}} \mathcal{F}\left(\boldsymbol{\Phi}^{(t)}, \boldsymbol{\theta}^{(t)}; \mathbf{x}^{\mathcal{D}_N}\right). \tag{2.54}$$

  Thereafter, the updated posterior distributions are obtained by utilizing the updated set of parameters $\boldsymbol{\Phi}^{(t+1)}$: $q_{\mathbf{x}^{\mathcal{D}_N}}(\mathbf{z}|\boldsymbol{\Phi}^{(t+1)})$. The expectation values needed to compute the lower bound

$$\sum_{i=1}^{N} \int q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}|\boldsymbol{\phi}^{(t+1)}) p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}|\boldsymbol{\theta}) \, d\mathbf{z}^{(i)}$$

  are either analytically tractable or we can leverage efficient MCMC approximations of $\mathcal{F}(\boldsymbol{\Phi}^{(t+1)}, \boldsymbol{\theta}^{(t)}; \mathbf{x}^{\mathcal{D}_N})$ due to the choice of a particular family of distributions for the posterior, which we can easily sample from.

- M step: Update the model parameters $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{F}\left(\boldsymbol{\Phi}^{(t+1)}, \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N}\right). \tag{2.55}$$

We investigate in this work different parametrization strategies for the approximate posterior distributions over the latent variables $q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}|\boldsymbol{\phi}^{(i)})$, which are discussed in Section 2.5.5.

Variants of the EM algorithm (for example, with multiple M steps per E step or in which a particular subset of latent variables and approximate posterior distributions are updated in the E step followed by a full update in the M step) could improve the efficiency. For discussions on potent variants of the EM algorithm refer to Refs. [399, 400].

[6] Both forms of notation for the variational posterior, $q_{\boldsymbol{x}^{(i)}}(\mathbf{z}^{(i)}|\boldsymbol{\phi}^{(i)})$ and $q_{\boldsymbol{\phi}^{(i)}}(\mathbf{z}^{(i)})$, have the same meaning and are employed equivalently in this work.

2.5.2 Variational Bayesian inference

As partly introduced in Section 2.5.1 for an EM algorithm with a variational E step, variational Bayesian inference transforms cumbersome inference tasks into an optimization problem by defining a flexible family of distributions over latent variables and treats parameters also as hidden variables [382, 401, 402]. Variational methods can yield the exact posterior distribution over latent variables. However, such methods result in an approximate posterior because usually the family of distributions we are optimizing is restricted, for example, by limiting them to quadratic functions or distributions governed by a linear combination of a fixed set of basis functions, or by assuming independence between the posterior on latent variables and model parameter posterior distributions in full Bayesian inference.

Developments of versatile variational inference algorithms for distributions from the conjugate exponential family are documented in Refs. [403–405]. There is an increasing degree of automation of variational inference, resulting in black-box variational inference [406] and variational Bayesian autoencoders [235, 407] and relying on stochastic backpropagation and reparametrization [408] of the variational posterior.

Extensions, such as stochastic variational inference, address the efficiency of variational inference by utilizing subsamples of the dataset $\mathbf{x}^{\mathcal{D}_N}$ for inferring latent variables in combination with stochastic optimization methods [409]. The success of stochastic approaches is significantly driven by the capabilities of the stochastic optimization methods employed for coping with noisy, but unbiased, gradient estimates of the (evidence) lower bound $\mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$ [410–412].

As stated in Ref. [356], the only difference from the variational EM introduced in Section 2.5.1 is the set of latent variables, which, in variational Bayesian approaches, encompasses the model parameters $\boldsymbol{\theta}$. The corresponding marginal log-likelihood $\log p(\mathbf{x}^{\mathcal{D}_N})$ is

$$\log p(\mathbf{x}^{\mathcal{D}_N}) = \sum_{i=1}^{N} \log \int p(\boldsymbol{x}^{(i)}|\mathbf{z}^{(i)}, \boldsymbol{\theta}_{\mathrm{cf}}) p(\mathbf{z}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}) p(\boldsymbol{\theta})\, d\mathbf{z}^{(i)}\, d\boldsymbol{\theta}. \tag{2.56}$$

The prior factorizes over the model parametrization with $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_{\mathrm{c}}) p(\boldsymbol{\theta}_{\mathrm{cf}})$, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathrm{cf}}, \boldsymbol{\theta}_{\mathrm{c}}\}$. In the probabilistic model defined in Equation 2.56, Ref. [409] distinguishes between local and global latent variables using conditional dependencies. The latent variables $\mathbf{z}^{(i)}$ giving rise to the observed atomistic configurations $\boldsymbol{x}^{(i)}$ are local hidden variables, since they are conditionally independent of the latent variables $\mathbf{z}_{\setminus i}^{\mathcal{D}_N}$ and observations $\boldsymbol{x}_{\setminus i}^{\mathcal{D}_N}$.[7] Therefore, the complete joint likelihood is

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{i=1}^{N} p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}|\boldsymbol{\theta}), \tag{2.57}$$

[7] $(\cdot)_{\setminus i}^{\mathcal{D}_N}$ denotes the set of variables $\{(\cdot)^{(1)}, \ldots, (\cdot)^{(i-1)}, (\cdot)^{(i+1)}, \ldots, (\cdot)^{(N)}\}$.

which gives rise to the posterior distribution

$$p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{\int p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})\, d\mathbf{z}\, d\boldsymbol{\theta}}. \tag{2.58}$$

We started our discussion with the marginal log-likelihood $\log p(\mathbf{x}^{\mathcal{D}_N})$ but now with $\boldsymbol{\theta}$ being a random variable itself:

$$\begin{aligned} \log p(\mathbf{x}^{\mathcal{D}_N}) &= \log \int p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})\, d\mathbf{z}\, d\boldsymbol{\theta} \\ &= \log \int q(\mathbf{z}, \boldsymbol{\theta}) \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{q(\mathbf{z}, \boldsymbol{\theta})}\, d\mathbf{z}\, d\boldsymbol{\theta} \\ &\geq \int q(\mathbf{z}, \boldsymbol{\theta}) \log \frac{p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})}{q(\mathbf{z}, \boldsymbol{\theta})}\, d\mathbf{z}\, d\boldsymbol{\theta} \\ &= \mathcal{F}(q(\mathbf{z}, \boldsymbol{\theta}); \mathbf{x}^{\mathcal{D}_N}). \end{aligned} \tag{2.59}$$

A decomposition can be obtained in a similar fashion, as shown in Equation 2.52, with:

$$\log p(\mathbf{x}^{\mathcal{D}_N}) = \mathcal{F}(q(\mathbf{z}, \boldsymbol{\theta}); \mathbf{x}^{\mathcal{D}_N}) + D_{\mathrm{KL}}\left(q(\mathbf{z}^{\mathcal{D}_N}, \boldsymbol{\theta}) \| p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})\right). \tag{2.60}$$

The evidence lower bound $\mathcal{F}(q(\mathbf{z}, \boldsymbol{\theta}); \mathbf{x}^{\mathcal{D}_N})$ is maximized, which, thus, minimizes $D_{\mathrm{KL}}(q(\mathbf{z}^{\mathcal{D}_N}, \boldsymbol{\theta}) \| p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}))$. The minimization of the aforementioned KL divergence implies that the variational posterior $q(\mathbf{z}^{\mathcal{D}_N}, \boldsymbol{\theta})$ comes close in terms of the KL metric to the exact posterior $p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$. Based on mean-field theory, which originated in statistical physics [413, 414], each latent variable is treated independently and the corresponding approximate posterior distribution has its own parameter [409]:

$$q(\mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{\Phi}) = q(\boldsymbol{\theta}|\boldsymbol{\lambda}) \prod_{i=1}^{N} q(\mathbf{z}^{(i)}|\boldsymbol{\phi}^{(i)}), \tag{2.61}$$

with the free parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\Phi} = \left\{\boldsymbol{\phi}^{(i)}\right\}_{i=1}^{N}$.

All these methodologies rely on the computation of gradients of the lower bound $\nabla_{\{\boldsymbol{\Phi}, \boldsymbol{\theta}\}} \mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$, which we will address directly in the specific model setting in the publications based on this thesis [375, 415] and in Chapters 4 and 6 of this work.

This section and Section 2.5.1 are mostly based on the seminal works [357, 416], which we highly recommended for more insights on variational methods for Bayesian learning.

2.5.3 Prior specification

This work employs flexible and potentially overparametrized models. The training of overparametrized models can lack robustness since they can yield unbounded likelihoods and ill-posed optimization problems [417, 418]. The Bayesian response to regularization in deterministic settings is to use functional priors, which we will exploit for regularizing the log-likelihood [419, 420]. The prior distribution occurs

in an additive fashion in the log-likelihood, as first introduced in Equation 2.39:

$$\arg\max_{\boldsymbol{\theta}} \left\{ \log p(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \right\}.$$

The primary motivation for employing such functional priors is to reveal parsimonious physical insights. In addition to that, we employ functional sparsity-inducing priors for enhancing stability during training by reducing the number of parameters to only those needed for explaining the data $\mathbf{x}^{\mathcal{D}_N}$ [417].

Automatic relevance determination (ARD) [421] will be employed in this work, which depicts a hierarchical prior using two components:

$$p(\boldsymbol{\theta}|\boldsymbol{\tau}) \equiv \prod_k \mathcal{N}(\theta_k|0, \tau_k^{-1}), \quad \tau_k \sim \text{Gamma}(\tau_k|a_0, b_0). \tag{2.62}$$

This hierarchical prior implies that the prior on $\boldsymbol{\theta}$ factorizes with independent Gaussian distributions with zero mean and precision τ_k. In the hierarchical extension for the ARD prior, τ_k is modeled with a Gamma distribution, which is conjugate to the Gaussian distribution. The marginal prior $p(\theta_k)$ has a heavy-tailed Student's t-distribution. A Student's t-distribution favors θ_k close to zero and thus, induces sparsity.

To optimize the lower bound $\mathcal{F}$, which encompasses now the additive log-prior contribution, gradients with respect to $\boldsymbol{\theta}$ need to be estimated. This can be done by regarding τ_k as a latent variable and performing an inner-loop EM step [422]:

- E step: Estimate

$$\mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k] = \frac{a_0 + \frac{1}{2}}{b_0 + \frac{\theta_k^2}{2}}. \tag{2.63}$$

- M step: Maximize

$$\frac{\partial \log p(\boldsymbol{\theta})}{\partial \theta_k} = -\mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k]\,\theta_k. \tag{2.64}$$

Moreover, the second derivative of the log-prior with respect to $\boldsymbol{\theta}$ is

$$\frac{\partial^2 \log p(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} = \begin{cases} -\mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k], & \text{if } k = l, \\ 0, & \text{otherwise.} \end{cases} \tag{2.65}$$

The hierarchical prior contains the hyperparameters a_0 and b_0, which need to be specified. Based on Ref. [421], in ARD, $a_0 = b_0 = 1.0 \times 10^{-5}$.

2.5.4 Approximate Bayesian inference using Laplace's approximation

Compared to the approaches introduced in Section 2.5.2, the Laplace approximation is an efficient way to obtain information about the distribution of $\boldsymbol{\theta}$ around $\boldsymbol{\theta}^{\text{MAP}}$, which are the maximum a posteriori estimates of the parameters $\boldsymbol{\theta}$. Determining the

full posterior distribution of the model parameters $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ comes with the computational burden of determining normalization constants. Therefore, we now estimate the approximate posterior over the model parameters $\boldsymbol{\theta}$ using Laplace's method [423], which was rediscovered as an efficient approximation method for quantifying uncertainties of neural network weights [424].

The Laplace approach allows us to approximate the exact posterior distribution with a Gaussian centered in the vicinity of the ML or MAP optimum. The Gaussian posterior is defined with the mean at $\mu_\mathrm{L} = \boldsymbol{\theta}^\mathrm{MAP}$ and a covariance matrix defined as the negative inverse of the Hessian of the log-posterior in the mode $\boldsymbol{\theta}^\mathrm{MAP}$. We assume in this work a covariance matrix with diagonal structure: $\boldsymbol{S}_\mathrm{L} = \mathrm{diag}(\sigma_L^2)$. Thus, we can write for the approximate posterior

$$p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N}) \approx \mathcal{N}\left(\boldsymbol{\mu}_\mathrm{L}, \boldsymbol{S}_\mathrm{L} = \mathrm{diag}(\sigma_\mathrm{L}^2)\right), \tag{2.66}$$

where

$$\boldsymbol{\mu}_\mathrm{L} = \boldsymbol{\theta}_\mathrm{MAP}, \tag{2.67}$$

and the diagonal entries of $\boldsymbol{S}_\mathrm{L}^{-1}$ are

$$\sigma_{\mathrm{L},k}^{-2} = -\frac{\partial^2 \mathcal{F}(\boldsymbol{\phi}, \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})}{\partial \theta_k^2}\bigg|_{\boldsymbol{\theta}_\mathrm{MAP}, \boldsymbol{\phi}_\mathrm{MAP}} + \mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k]. \tag{2.68}$$

The final term in Equation 2.68, $\mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k]$, stems from the prior employed in Equation 2.65. The mean and the variance in Equations 2.67 and 2.68 are used after the last iteration upon convergence of the lower bound $\mathcal{F}(\boldsymbol{\phi}, \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N})$ in the optimization.

2.5.5 Approaches for model parametrization

The introduced inference tools are general and widely applicable in developing predictive CG methodologies. Thus far, we have not discussed the specifics of the implementations and gradient computation, which are specific to the model. Since we explore a variety of approaches with different parametrization strategies for involved distributions, we discuss the particular objectives in the corresponding chapters and related papers, which present the novel contributions. However, the following is an overview of the approaches we use to model a predictive distribution $p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N})$.

The proposed CG method has two components: (1) the hidden generator of CG variables $p(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c})$ and (2) the generative mapping, which provides predictions of observable atomistic configurations, $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf})$.

The following strategies ensure that the models are flexible and can cope with small data to reveal physical insights:

(i) Provide rich and flexible descriptions from the beginning and automatically select those parameters required for describing reference atomistic trajectories

well. Parameters that are less significant are automatically turned off during model learning. This avenue is followed in Chapters 3 and 5.

We realize this in Chapter 3 by employing a rich set of basis functions describing $p(\mathbf{z}|\boldsymbol{\theta}_c) \propto e^{-\beta U(\mathbf{z};\boldsymbol{\theta}_c)}$, with the CG potential expressed as a linear combination of basis functions $\boldsymbol{\phi}(\mathbf{z})$:

$$U(\mathbf{z};\boldsymbol{\theta}_c) = \boldsymbol{\theta}_c^T \boldsymbol{\phi}(\mathbf{z}), \tag{2.69}$$

and automatically set components in $\boldsymbol{\theta}_c$ that are not required to zero. The proposed methodology does not require any specific prerequisite modeling knowledge. As we propose a fully Bayesian framework, we consistently search for sparsity implied by functional priors using methods such as ARD [421]. For a more detailed discussion of different sparsity-favoring priors, we refer to Refs. [420, 425]. Sparsity priors naturally induce sparse solutions without actively fixing any parameters to zero, which would be impractical. The prior competes with the available data and allows us to switch previously redundant parameters on again if they turn out to be required for explaining the data $\mathbf{x}^{\mathcal{D}_N}$.

In addition to providing a rich set of basis functions, in Chapter 5 we explore the possibilities of employing deep neural networks [426, 427] to reveal sparse features. This can be done by equipping $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}_{cf})$ with a flexible neural network description, for example, a Gaussian with mean and variance output from a flexible deep neural network:

$$p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}_{cf}) = \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_{\boldsymbol{\theta}^{\mu}_{cf}}(\mathbf{z}), \boldsymbol{S}_{\boldsymbol{\theta}^{S}_{cf}}), \tag{2.70}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\theta}^{\mu}_{cf}}(\mathbf{z}) = f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}^{\mu}_{cf}}(\mathbf{z}) \tag{2.71}$$

is a non-linear mapping $\mathbf{z} \mapsto f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}^{\mu}_{cf}}(\mathbf{z})$ ($f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}^{\mu}_{cf}} : \mathbb{R}^{n_c} \mapsto \mathbb{R}^{n_f}$) parametrized by an expressive multilayer perceptron [428–430].

(ii) The opposite approach to implementing sparse methods, while following the same overall goal of developing interpretable learning strategies for CG models, is as follows. Start with a simple model for the distributions, learn the parametrization given the current model structure until convergence. The convergence criterion is the lower bound $\mathcal{F}$, which provides a valuable indication [431, 432]. Subsequently, add complexity and optimize the new model with enhanced flexibility given the data $\mathbf{x}^{\mathcal{D}_N}$ until convergence. This can be realized by providing a model for $p(\mathbf{z}|\boldsymbol{\theta}_c) \propto e^{-\beta U(\mathbf{z};\boldsymbol{\theta}_c)}$ with the CG potential expressed

as a linear combination of basis functions, as in Equation 2.69, where the vectors are

$$\boldsymbol{\theta}_{\mathrm{c}} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_L \end{pmatrix} \quad \text{and} \quad \boldsymbol{\phi}(\mathbf{z}) = \begin{pmatrix} \phi_1(\mathbf{z};\boldsymbol{\lambda}_1) \\ \vdots \\ \phi_L(\mathbf{z};\boldsymbol{\lambda}_L) \end{pmatrix}. \tag{2.72}$$

This linear combination allows us iteratively to add basis functions $\phi_i(\mathbf{z};\boldsymbol{\lambda}_i)$ from a family of basis functions (e.g., radial basis functions), which is parametrized with $\boldsymbol{\lambda}_i$. This even enables us to add the optimal basis function $\phi_i(\mathbf{z};\boldsymbol{\lambda}_i^{\mathrm{opt}})$ by maximizing an information theoretic criterion we develop with respect to $\boldsymbol{\lambda}_i$. After identifying the optimal $\boldsymbol{\lambda}_i^{\mathrm{opt}}$, it is fixed, and we proceed with the optimization of the associated parameter θ_i and previous $\boldsymbol{\theta}_{\mathrm{c}}$ (associated with previously added features) such that the lower bound increases. We present this strategy in Chapter 4.

The sparse solutions, as we will show, are beneficial for dealing with only a few data points N compared to the dimension of $\mathbf{x}$. Expediting the available evidence, based on limited data, naturally leads to descriptions that reveal relevant physics, which is the sparsest solution itself.

Beyond the discussion above, we provide adaptive models once in the mapping $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}_{\mathrm{cf}})$ and once in the coarse description $p(\mathbf{z};\boldsymbol{\theta}_{\mathrm{c}})$. In the latter case, we seek to identify sparse features correlated with physical insights in terms of the interactions of the CG potential expressed by $U(\mathbf{z};\boldsymbol{\theta}_{\mathrm{c}})$ given a probabilistic simple (but still parameter-dependent) coarse-to-fine mapping $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}_{\mathrm{cf}})$. We discuss this strategy in Chapters 3 and 4. In the opposite case, with an adaptive and flexible $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}_{\mathrm{cf}})$ and simple $p(\mathbf{z};\boldsymbol{\theta}_{\mathrm{c}})$, we seek to reveal insights in terms of expressive mappings that reveal slow coordinates or CVs of the reference system. The identified CVs provide a simple description of the complex observed trajectories $\mathbf{x}^{\mathcal{D}_N}$ (simple, since CG variables are generated from a simple $p(\mathbf{z};\boldsymbol{\theta}_{\mathrm{c}})$; thus, they cannot encode the complex structure in $\mathbf{z}$). We present the identification of slow coordinates or CVs that depict a parsimonious description of complex physics in Chapters 5 and 6.

2.5.6 Outline

Finally, we suggest the consideration of recent progress on invertible generative distributions $p(\mathbf{x}|\mathbf{x}^{\mathcal{D}_N})$, which depict an appealing framework for facilitating variational inference by circumventing the intractable marginalization over the latent variables:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\, d\mathbf{z}.$$

Normalizing flow based approaches also belong to the family of generative models and rely on a change of variables for transforming simple distributions into rich expressive ones by the aforementioned mappings. Reference [433] reviews of normalizing flow based approaches.

Normalizing flow based generative models rely on simple distributions over latent variables $\mathbf{z}$, whereas complexity is induced by a series of bijective mappings that add expressivity in the generative process $p_x(\mathbf{x})$ [434]. Latent variables $\boldsymbol{z}^{(i)} \sim p_z(\mathbf{z})$ are mapped to the corresponding observable variables $\boldsymbol{x}^{(i)}$ with a bijective function $f : \boldsymbol{x} \to \boldsymbol{z}$ and its inverse function $g = f^{-1} : \boldsymbol{z} \to \boldsymbol{x}$ [435]. The mapping function must be bijective and thus, invertible, which implies $\dim(\mathbf{z}) = \dim(\mathbf{x})$. With the change of variables, a simple distribution $p(\mathbf{z})$ is transformed into a complex distribution $p_x(\mathbf{x})$, which gives rise to the observed data $\mathbf{x}^{\mathcal{D}_N}$:

$$p_x(\mathbf{x}) = p_z(f(\mathbf{x})) \left| \det \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} \right) \right|. \tag{2.73}$$

Here

$$\left| \det \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} \right) \right|$$

is the determinant of a square matrix $\boldsymbol{A}$ and $\partial f(\mathbf{x})/\partial \mathbf{x}^T$ is the Jacobian of f and $\mathbf{x}$. In the above discussions on expectation maximization (Section 2.5.1) and variational inference (Section 2.5.2), the intractable marginalization over the latent variables $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\, d\mathbf{z}$ caused most of the computational problems, which normalizing flows circumvent by employing Equation 2.73.

However, the challenging aspect in normalizing flows is constructing expressive functions, for example, a composition of neural network layers: $f(\mathbf{x}) = f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{x})$, which offer the computation of reasonably tractable determinants of the Jacobian [435, 436]. RealNVP [238, 437] and GLOW [239] are attractive and promising approaches for constructing such efficient bijective mappings. In particular, normalizing flows provide in the context of generative models an explicit definition of likelihood, which allows us to use existing inference algorithms for learning the parameters [435].

Further approaches for implicit likelihood models are generative adversarial networks (GANs), which have a two-player min-max objective [438]. This approach has had promising results in image and video segmentation and even generation, and is also used in physics [439]. There are many variants of GANs. For reviews, see Refs. [440, 441].

2.6 Monte Carlo methods

A wide range of problems in different scientific disciplines requires the computation of integrals of the form:

$$I = \int g(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x}, \tag{2.74}$$

where $\mathbf{x}$ is a random variable distributed according to $\pi(\mathbf{x})$ and $\mathbf{x}$ is high dimensional. In our work, such integrals occur when computing the expected values of

observables (Section 2.1.4) or when estimating expectations in inference tasks (Section 2.5) where $\pi(\mathbf{x})$ is the posterior distribution over hidden variables. MC methods overcome two major obstacles:

1. Observables $a(\mathbf{x})$ are complex functions and a vast amount of computational effort is required to evaluate their expected values. We seek to obtain unbiased, but eventually noisy, approximations to such expectations [295].

2. In many cases, for example, in variational inference, the computation of the normalization (or the model evidence) for obtaining posterior distributions is prohibitive. MC methods allow sampling from unnormalized distributions.

Moreover, an optimization problem can be reformulated as a MC problem for some temperature T and the function to be optimized denoted as $h(\mathbf{x})$:

$$\pi_T(\mathbf{x}) \propto e^{-h(x)/T}. \tag{2.75}$$

Then, with sufficiently small T, most states $\boldsymbol{x}^{(i)}$ are sampled in the vicinity of the minimum of $h(\mathbf{x})$ [442].

The theory beind MC methods suggests an approximation of the the integral in Equation 2.74 by a finite sum with a limited set of independent samples $\boldsymbol{x}^{(l)} \sim \pi(\mathbf{x})$:

$$\hat{I}_L = \frac{1}{L} \sum_{l=1}^{L} g(\boldsymbol{x}^{(l)}). \tag{2.76}$$

The law of large numbers states that the approximate estimator $\hat{I}_L$ in this equation converges to the exact value of the integral I:

$$\lim_{L \to \infty} \hat{I}_L = I, \tag{2.77}$$

in the limit of large numbers of independent samples representing $\pi(\mathbf{x})$. Very notable is the convergence rate of the approximate estimator $\hat{I}_L$ [442]:

$$\sqrt{L}(\hat{I}_L - I) \to \mathcal{N}(0, \sigma^2), \tag{2.78}$$

where

$$\sigma^2 = \text{var}\left[\hat{I}\right] = \frac{1}{L}\mathbb{E}_{\pi(\mathbf{x})}\left[(g - \mathbb{E}\left[g\right])^2\right],$$

which is independent of the dimension of $\mathbf{x}$ [356]. Thus, using a very few samples can result in a sufficient estimate of the expected value provided they are of "good" quality, which means independent. However, most sampling schemes provide correlated samples so that the effective sample size (ESS) translates the apparent number of samples into the number of independent ones [443]. From the perspective of signal processing, a sample autocorrelation provides insights about the quality of the samples utilized [444] for the MCMC methods that are introduced later.

2.6.1 Importance sampling

Assume that we are interested in computing $\mu = \mathbb{E}_{\pi(\mathbf{x})}[g] = \int_{\mathcal{X}} g(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x}$ with the PDF $\pi(\mathbf{x})$ defined on $\mathcal{X} \subset \mathbb{R}^{d_x}$. Further, suppose we can evaluate $\pi(\mathbf{x})$ at any $\mathbf{x}$ but we are not able to sample from $\pi(\mathbf{x})$. A vanilla MC approach could uniformly discretize $\mathcal{X}$ and then compute $g(\mathbf{x})\pi(\mathbf{x})$ at the grid points to yield the estimator:

$$\mu \approx \sum_{l=1}^{L} g(\boldsymbol{x}^l)\pi(\boldsymbol{x}^l). \tag{2.79}$$

The terms in the above summation increase exponentially as the dimension d_x increases. Assuming the PDF $\pi(\mathbf{x})$ has the probability mass distributed in narrow regions of $\mathcal{X}$, most evaluations of $\pi(\mathbf{x})$ will yield zero, whereas interesting regions with high probability mass, thus large $\pi(\mathbf{x})$, could be missed. The idea, initially proposed in Ref. [445], is that, to reduce the computational effort, the focus of sampling should lie in regions of importance [442]. Areas of importance are those that yield large absolute values of the product $g(\mathbf{x})\pi(\mathbf{x})$ and thus, make a significant contribution to the sum in Equation 2.79.

Thus, we introduce a distribution $q(\mathbf{x})$ that is easy to sample and obtain L samples $\boldsymbol{x}^{(l)} \sim q(\mathbf{x})$. An approximate estimator the integral of interest is

$$\begin{aligned}
\mathbb{E}_{\pi(\mathbf{x})}[g] &= \int g(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} \\
&= \int g(\mathbf{x})\frac{\pi(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})\, d\mathbf{x} \\
&\approx \frac{1}{L}\sum_{l=1}^{L}\frac{\pi(\boldsymbol{x}^l)}{q(\boldsymbol{x}^l)}g(\boldsymbol{x}^{(l)}).
\end{aligned} \tag{2.80}$$

The term

$$r_l = \frac{\pi(\boldsymbol{x}^l)}{q(\boldsymbol{x}^l)}$$

in Equation 2.80 relates to the corrections for importance weights when sampling from $q(\mathbf{x})$ instead of $\pi(\mathbf{x})$ while keeping all L samples in the sum of Equation 2.80. The above discussion implies that the distributions are normalized. However, often we are not able to compute the normalization constant. We denote with $\tilde{\pi}(\mathbf{x})$ the unnormalized part, which is normalized with the constant Z_π. Therefore, $\pi(\mathbf{x}) = \tilde{\pi}(\mathbf{x})/Z_\pi$ and likewise $q(\mathbf{x}) = \tilde{q}(\mathbf{x})/Z_q$. Then the estimator becomes:

$$\begin{aligned}
\mathbb{E}_{\pi(\mathbf{x})}[g] &= \int g(\mathbf{x})\pi(\mathbf{x})\, d\mathbf{x} \\
&= \frac{Z_q}{Z_\pi}\int g(\mathbf{x}) \underbrace{\frac{\tilde{\pi}(\mathbf{x})}{\tilde{q}(\mathbf{x})}}_{\equiv \tilde{r}_l} q(\mathbf{x})\, d\mathbf{x} \\
&\approx \frac{Z_q}{Z_\pi}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l g(\boldsymbol{x}^{(l)}).
\end{aligned} \tag{2.81}$$

Thus far we have not defined how to obtain an estimate of the ratio of normalizing constants Z_q/Z_π in Equation 2.81. This can be done by utilizing the same samples $\{\boldsymbol{x}^{(l)}\}_{l=1}^{L}$ with:

$$\frac{Z_\pi}{Z_q} = \frac{1}{Z_q}\int \tilde{\pi}(\mathbf{x})\, d\mathbf{x} = \int \frac{\tilde{\pi}(\mathbf{x})}{\tilde{q}(\mathbf{z})} \underbrace{\frac{\tilde{q}(\mathbf{z})}{Z_q}}_{=q(\mathbf{x})} d\mathbf{x} \approx \frac{1}{L}\sum_{l=1}^{L} \tilde{r}_l. \tag{2.82}$$

By defining the normalized importance weights,

$$w_l = \frac{\tilde{r}_l}{\sum_{k=1}^{L} \tilde{r}_k}, \tag{2.83}$$

a short version of the approximate integral estimator in importance sampling is

$$\mathbb{E}_{\pi(\mathbf{x})}[g] \approx \sum_{l=1}^{L} w_l g(\boldsymbol{x}^{(l)}), \quad \text{with } \boldsymbol{x}^{(l)} \sim q(\mathbf{x}). \tag{2.84}$$

Weighting samples is a relevant prerequisite for more advanced sequential Monte Carlo (SMC) schemes, which is why importance sampling has been introduced here. However, the direct applicability of importance sampling is limited by several drawbacks:

- The success of importance sampling depends crucially on the closeness of $q(\mathbf{x})$ to the target distribution $\pi(\mathbf{x})$. If $\pi(\mathbf{x})$ is very concentrated, most weights r_l will be zero and the sum in Equation 2.82 is determined only by a small subset of samples. For such cases, it is interesting for diagnostic purposes to estimate the ESS:
$$\mathrm{ESS} = \frac{1}{\sum_{l=1}^{L}\left(w_l^2\right)}, \tag{2.85}$$
which indicates the amount of samples the approximation of $\mathbb{E}_{\pi(\mathbf{x})}[g]$ relies on [446].
- Most problematic is when none of the samples $\boldsymbol{x}^{(l)}$ falls into regions with large $r_l g(\boldsymbol{x}^{(l)})$. The variance in r_l is still small in such cases and the ESS is close to N_L, which makes a reliable diagnosis challenging.

Reference [447] contains an extended chapter on importance sampling techniques from a physical point of view plus a discussion of extensions to vanilla importance sampling.

2.6.2 Markov Chain Monte Carlo

Section 2.6.1 introduces importance sampling, which enables the approximation of expected values. However, importance sampling gives poor results in case of high dimensional $\mathbf{x}$. This section discusses MCMC approaches, which provide a general framework for sampling from a variety of densities.

The word "chain" describes well the difference from MC methods. Here we define a distribution for sampling proposals that is conditionalized on the current state $\boldsymbol{x}^{(t)}$, $q(\mathbf{x}|\boldsymbol{x}^{(t)})$. The subsequent samples $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots$ are form a Markov chain [356]. Since we need to sample from the proposal density $q(\mathbf{x}|\boldsymbol{x}^{(t)})$, the chain should be simple. The overall goal is to generate samples from $\pi(\mathbf{x}) = \tilde{\pi}(\mathbf{x})/Z_\pi$ where we can evaluate $\tilde{\pi}(\mathbf{x})$ for any $\mathbf{x}$. The algorithm proceeds by drawing a potential candidate $\mathbf{x}'$ from the proposal distribution $q(\mathbf{x}|\boldsymbol{x}^{(t)})$[8] and accept or reject the candidate based on a criterion that depends on the particular implementation of the MCMC algorithm.

The early Metropolis algorithm [306] used symmetric proposal distributions with $q(\mathbf{x}'|\mathbf{x}^*) = q(\mathbf{x}^*|\mathbf{x}')$. The proposal $\mathbf{x}'$ is accepted for the new state $\mathbf{x}^{(t+1)}$ with probability

$$\mathcal{A}(\mathbf{x}', \mathbf{x}^{(t)}) = \min\left\{1, \frac{\tilde{\pi}(\mathbf{x}')}{\tilde{\pi}(\mathbf{x}^{(t)})}\right\}. \tag{2.86}$$

Proposals that yield $\tilde{\pi}(\mathbf{x}') > \tilde{\pi}(\mathbf{x}^{(t)})$ are certainly kept. If the proposal gets accepted, we set $\mathbf{x}^{(t+1)} = \mathbf{x}'$ and $t \leftarrow t+1$. If $\mathcal{A}$ suggests the rejection of the proposal $\mathbf{x}'$, we set $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$ and $t \leftarrow t+1$. The latter implies that we store copies of $\mathbf{x}^{(t)}$ if the proposal is rejected. This can be done also by weighting the sample to make storage more efficient. The proposal distribution q is, in both cases, conditionalized on $\mathbf{x}^{(t+1)}$. For $t \to \infty$, the distribution of samples $\mathbf{x}^{(t)}$ tend to $\pi(\mathbf{x})$ [356]. The set of samples subsequently obtained, however, is not independent and exhibits correlation between the samples. To circumvent this, one can discard most samples within the sequence and keep only every mth sample.

For a Markov chain to converge, the distribution of interest $\pi(\mathbf{x})$ must be invariant under the transition function $T(\mathbf{x}, \mathbf{x}^{(t)})$. For homogeneous Markov chains, with the same $T(\mathbf{x}, \mathbf{x}^{(t)})$ for each step t, the distribution is invariant if

$$\pi(\mathbf{x}^{(t)}) = \int_{\mathbf{x}'} T(\mathbf{x}^{(t)}, \mathbf{x}')\pi(\mathbf{x}')\, d\mathbf{x}'. \tag{2.87}$$

A restrictive criterion ensuring convergence of the Markov chain is called *detailed balance*. It is a sufficient but not necessary condition:

$$\pi(\mathbf{x})T(\mathbf{x}, \mathbf{x}') = \pi(\mathbf{x}')T(\mathbf{x}', \mathbf{x}) \tag{2.88}$$

This results in a reversible Markov chain. Independent of the initial distribution, as $t \to \infty$, the Markov chain must converge to the target distribution $\pi(\mathbf{x})$. Ergodicity means that the equilibrium distribution occurs for $t \to \infty$. For a more detailed discussion of convergence criteria, read Chapter 5.3 in Ref. [442] and Chapter 11.2.1 in Ref. [356].

[8] The literature often introduces a general transition function $T(\mathbf{x}, \boldsymbol{x}^{(t)})$. A probability transition function takes only non-negative values and $\int T(\mathbf{x}, \boldsymbol{x}^{(t)})\, d\mathbf{x} = 1$.

2.6.3 Metropolis–Hastings algorithm

Unlike Section 2.6.2, in which the proposal distribution was symmetric, the Metropolis–Hastings MCMC method [448] is a generalization for asymmetric proposal distributions. Given a current state $\mathbf{x}^{(t)}$, a proposal move $\mathbf{x}'$ is drawn from $q(\mathbf{x}|\mathbf{x}^{(t)})$. For the Metropolis–Hastings update criterion, we must additionally incorporate the reverse transition probability $q(\mathbf{x}^{(t)}|\mathbf{x}')$:

$$\mathcal{A}(\mathbf{x}', \mathbf{x}^{(t)}) = \min\left\{1, \frac{\tilde{\pi}(\mathbf{x}')q(\mathbf{x}^{(t)}|\mathbf{x}')}{\tilde{\pi}(\mathbf{x}^{(t)})q(\mathbf{x}'|\mathbf{x}^{(t)})}\right\}. \tag{2.89}$$

For a symmetric proposal distribution $q(\mathbf{x}^{(t)}|\mathbf{x}') = q(\mathbf{x}'|\mathbf{x}^{(t)})$, the original Metropolis acceptance probability can be recovered (Equation 2.86).

A further interesting MCMC sampling approach is Gibbs sampling, which provides samples of a subset of the complete random variable $\mathbf{x}$ at a time, for example, $\tilde{\mathbf{x}}$, of the distribution conditioned on all other values $\pi(\tilde{\mathbf{x}}|\mathbf{x}_{\setminus\tilde{\mathbf{x}}})$. The sampling steps are repeated either for a single component of the random variable $\mathbf{x}$ or a subset $\tilde{\mathbf{x}}$. The updated variables can change and can even be randomly selected from the random variable vector $\mathbf{x}$. The Gibbs sampling scheme [449] is a special case of the Metropolis–Hastings algorithm.

2.6.4 Metropolis-adjusted Langevin algorithm

The Metropolis-adjusted Langevin algorithm (MALA) [450, 451] relies on the Metropolis–Hastings acceptance probability for a proposed state $\mathbf{x}'$, which was introduced in Section 2.6.3. MALA differs in the way we obtain a proposal $\mathbf{x}'$. It incorporates gradient information, which increases efficiency in exploring the phase space. It moves efficiently to regions of interest, that is, those with a high probability mass. Therefore, with a given ratio of accepted and rejected samples, the mixing of samples is improved compared to random walk Metropolis–Hastings. Random walk proposals are governed by overdamped Langevin dynamics [452] that incorporate gradient information for the target distribution $\pi(\mathbf{x})$. The Langevin stochastic differential equation (SDE) is

$$d\mathbf{x}_t = \frac{\sigma^2}{2}\nabla_{\mathbf{x}} \log \pi(\mathbf{x}_t)\, dt + \sigma d\boldsymbol{W}_t, \tag{2.90}$$

where $\boldsymbol{W}_t$ corresponds to Brownian motion or a Wiener process. The discretized version of the SDE follows from $dt \approx \Delta t$, $d\mathbf{x}_t = \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}$, and $d\boldsymbol{W}_t \approx \sqrt{\Delta t} z_t, z_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$:

$$x^{(t+1)} = x^{(t)} + \Delta t \frac{\sigma^2}{2}\nabla_{\mathbf{x}} \log \pi(\mathbf{x}) + \sigma\sqrt{\Delta t} z_t, \tag{2.91}$$

which leads to the proposal distribution:

$$q(\mathbf{x}'^{(t+1)}|\mathbf{x}^{(t)}) = \mathcal{N}\left(\mathbf{x}^{(t)} + \frac{\sigma^2}{2}\Delta t \nabla_{\mathbf{x}} \log \pi(\mathbf{x}^{(t)}), \sigma^2 \Delta t\right). \tag{2.92}$$

The proposal for the random walk $\mathbf{x}^{(t+1)}$ in MALA, $\mathbf{x}'^{(t+1)}$ is accepted with probability

$$\mathcal{A}(\mathbf{x}'^{(t+1)}, \mathbf{x}^{(t)}) = \min\left\{1, \frac{\pi(\mathbf{x}'^{(t+1)})q(\mathbf{x}^{(t)}|\mathbf{x}'^{(t+1)})}{\pi(\mathbf{x}^{(t)})q(\mathbf{x}'^{(t+1)}|\mathbf{x}^{(t)})}\right\}. \tag{2.93}$$

If the proposal is accepted, then we set $\mathbf{x}^{(t+1)} = \mathbf{x}'^{(t+1)}$ otherwise $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$. The parameter σ needs to be specified and adjusted according to the desired acceptance/rejection ratio for proposal samples.

For further MCMC approaches relying on the Metropolis–Hastings criterion, see the review in Ref. [453].

For distributions with a very dense probability mass separated into multiple modes with high free-energy barriers, the aforementioned MCMC methods may suffer because they could sample only a single mode with limited trail steps. In theory, any MC method can explore the full phase space in an infinite number of steps, which, in practice, is infeasible.

Advances in sampling methods led to hybrid MC methods, which translate sampling from probability distributions into simulating a dynamic Hamiltonian system by introducing fictitious momentum variables conjugate to the position $\mathbf{x}$ [454–457]. Interesting approaches for lowering the high free-energy barriers provide tempering schemes that start initially at high temperatures, which allows the samples to mix [458–460]. We highly recommend Ref. [461], as it provides an overview of Langevin and Hamiltonian MC methods with powerful extensions.

2.6.5 Adaptive sequential Monte Carlo methods

In this work, we are interested in sampling complex high-dimensional random variables with multiple distinct modes. The SMC approach is an efficient method for sampling of multimodal distributions since it efficiently samples a sequence of (unnormalized) distributions exhibiting smooth transitions. The sequence of distributions could, for example, relate to a tempering scheme, which facilitates covering multiple modes. There is also a sequence of slowly chaining distributions within the stochastic optimization processes in variational inference. In variational inference, the parametrization of a distribution $\boldsymbol{\theta}_k$ changes relatively slow in each step and the absolute value of $(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k)/\boldsymbol{\theta}_k$ is small. Each step in the SMC method is embarrassingly parallelizable with respect to the particles involved [462–464].

A sequence of distributions can be generated by a tempering scheme. Tempering implies starting with a high temperature (low inverse temperature β_0) and sequentially lowering the temperature to the target temperature (or target inverse temperature $\beta_K = 1$) [465–468]. The sequence is, thus, defined by the inverse temperatures $0 \leq \beta_0 \leq \cdots \leq \beta_k \leq \cdots \leq \beta_K = 1$. The target distribution is $\pi(\mathbf{x}) = \tilde{\pi}(\mathbf{x})/Z$. Then

we can write the sequence of distributions in the following form:

$$\pi_k(\mathbf{x}|\beta_k) = \frac{1}{Z_k} e^{\beta_k \overbrace{\log \tilde{\pi}(\mathbf{x})}^{\equiv -V(\mathbf{x})}}, \quad \forall k \in \{1, \ldots, K\}. \tag{2.94}$$

In the contexts in which we apply adaptive SMC, it is associated with small changes to the governing parameters $\boldsymbol{\theta}_k$ of a distribution:

$$\pi_k(\mathbf{x}|\boldsymbol{\theta}_k) = \frac{1}{Z(\beta, \boldsymbol{\theta}_k)} e^{-\beta V(\mathbf{x};\boldsymbol{\theta}_k)}. \tag{2.95}$$

We have a set of N random samples (also interpreted as particles or replicas) at step k with parameters $\boldsymbol{\theta}_k$ denoted as $\{\mathbf{x}_k^{(i)}\}_{i=1}^N$, which approximate the distribution $\pi_k(\mathbf{x}|\boldsymbol{\theta}_k)$. In adaptive SMC, this set of particles $\{\mathbf{x}_k^{(i)}\}_{i=1}^N$ is updated based on a sequence of importance sampling, resampling, and rejuvenation steps [469]. Each of the particles $\mathbf{x}_k^{(i)}$ contributes to the particle-based approximation of $\pi_k(\mathbf{x}|\boldsymbol{\theta}_k)$ according to its normalized weight $w_k^{(i)}$:

$$\pi_k(\mathbf{x}|\boldsymbol{\theta}_k) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\mathbf{x} - \mathbf{x}_k^{(i)}). \tag{2.96}$$

With particle-based SMC approaches, we can approximate expectations of a function $g(\mathbf{x})$ as,

$$\mathbb{E}_{\pi_k(\mathbf{x}|\boldsymbol{\theta}_k)}[g(\mathbf{x})] \approx \sum_{i=1}^{N} w_k^{(i)} g(\mathbf{x}_k^{(i)}). \tag{2.97}$$

The adaptive component in an adaptive SMC scheme comes from building an adaptive path between $\pi_k(\mathbf{x}|\boldsymbol{\theta}_k)$ and $\pi_{k+1}(\mathbf{x}|\boldsymbol{\theta}_{k+1})$, which bridges the parameter increment. This step is relevant since the quality of sampling depends on the proximity of the two distributions and a smooth transition between $\pi_k(\mathbf{x}|\boldsymbol{\theta}_k)$ and $\pi_{k+1}(\mathbf{x}|\boldsymbol{\theta}_{k+1})$. Adaptive SMC, therefore, constructs the intermediate auxiliary distributions:

$$\begin{aligned} \pi_k^{\gamma}(\mathbf{x}|\boldsymbol{\theta}_k^{\gamma}) &\propto \tilde{\pi}_k^{\gamma}\left(\mathbf{x}|(1-\gamma)\boldsymbol{\theta}_k + \gamma\boldsymbol{\theta}_{k+1}\right), \qquad \gamma \in [0,1] \\ &= e^{-\beta V(\mathbf{x};\boldsymbol{\theta}_k^{\gamma})}, \end{aligned} \tag{2.98}$$

where $\boldsymbol{\theta}_k^{\gamma} = (1-\gamma)\boldsymbol{\theta}_k + \gamma\boldsymbol{\theta}_{k+1}$. The control parameter γ ensures that there is a smooth transition, which makes importance sampling applicable. In Ref. [205], the intermediate steps are automatically adjusted to balance the efficiency and the overall accuracy of the proposed scheme. The number of intermediate steps for changing γ is clearly problem dependent. We utilize in total S steps for approaching $\gamma = 1$ with $0 = \gamma_1 < \gamma_2 < \cdots < \gamma_S = 1$ and thus, ultimately $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k^{\gamma_S=1}$. The ESS controls the adjustment of the intermediate step size (as introduced in Equation 2.85):

$$\mathrm{ESS}_k^{\gamma_s} = \frac{1}{\sum_{i=1}^{N} \left(w_{k,\gamma_s}^{(i)}\right)^2}.$$

The extremes of the ESS involve cases where the whole distribution is approximated by only one sample or where the ESS equals N because the samples are uniformly distributed in terms of the weights $w_k^{(i)}$. When updating γ_s to γ_{s+1}, the ESS should not change significantly since the affiliated distributions are similar. If changes in the ESS for γ_{s+1} compared to γ_s are noticeable, the corresponding increment from γ_s to γ_{s+1} should be reduced. The threshold in Ref. [205] was adopted with $\text{ESS}_{\gamma_{s+1}} \geq \zeta \, \text{ESS}_{\gamma_s}$, where $\zeta = 0.6$. If the population $\{\mathbf{x}_k^{(i)}\}_{i=1}^N$ is far from representing the distribution $\pi_k(\mathbf{x})$, which is indicated by the ESS dropping below $\text{ESS}_{\text{min}} = N/2$ [205], the existing samples are resampled by sampling a multinomial distribution with the corresponding weights $w_k^{(i)}$. After a potential resampling step, the new population rejuvenates by applying a MALA proposal distribution with $q(\mathbf{x}_{k,s+1}^{(i)}|\mathbf{x}_{k,s}^{(i)})$. Algorithm 1 lists the steps.

Algorithm 1: Adaptive SMC algorithm

Input: $s = 1$, $\gamma_1 = 0$, samples $\{\mathbf{x}_{k,s}^{(i)}\}_{i=1}^N$ approximating $\pi_k^{\gamma_1}(\mathbf{x}|\boldsymbol{\theta}_k^{\gamma_1})$ as defined in Equation 2.98 and such that Equation 2.96 holds.

Output: New population $\{\mathbf{x}_{k,S}^{(i)}\}_{i=1}^N$ that approximates $\pi_{k+1}(\mathbf{x}|\boldsymbol{\theta}_{k+1})$.

while $\gamma_s < 1$ **do**

$s \leftarrow s + 1$.

Reweighting and importance sampling:

Estimate updated population weights with

$$w_{k,s}^{(i)}(\gamma_s) = w_{k,s-1}^{(i)} \frac{\pi_k^{\gamma_s}(\mathbf{x}_{s-1}^{(i)}|\boldsymbol{\theta}_k^{\gamma_s})}{\pi_k^{\gamma_{s-1}}(\mathbf{x}_{s-1}^{(i)}|\boldsymbol{\theta}_k^{\gamma_{s-1}})}$$

depending on the current adapted step γ_s.

Adjust $\gamma_s \in (\gamma_{s-1}, 1]$ such that $\text{ESS}_s = \zeta \, \text{ESS}_{s-1}$.

Resampling:

if $\text{ESS}_s \leq \text{ESS}_{min}$ **then**

Resample utilizing a multinomial distribution with the current weights $w_{k,s}^{(i)}(\gamma_s)$.

Rejuvenation:

Given the population, move according to an MCMC proposal and acceptance step (in our case, we utilize MALA).

Utilize a MCMC proposal distribution $q(\mathbf{x}_{k,s}^{(i)}|\mathbf{x}_{k,s-1}^{(i)})$ to ensure the invariance of $\pi_k^{\gamma_s}$ under the Markov chain approach.

Update:

$$\left\{\mathbf{x}_{k,s}^{(i)}\right\}_{i=1}^N \leftarrow \left\{\mathbf{x}_{k,s-1}^{(i)}\right\}_{i=1}^N.$$

2.7 Stochastic optimization

All inference tasks in this work translate to optimization problems in terms of the model parameters $\boldsymbol{\theta}$ and the parameters $\boldsymbol{\phi}$ governing the approximate posterior distributions over the latent variables [470, 471]. The gradients of the objective $\mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{\theta}; \boldsymbol{x}^{(i)})$[9] involve expectations of the complete log-likelihood $\log p(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}|\boldsymbol{\theta})$. We consider, for instance, the gradient with respect to the model parameters $\boldsymbol{\theta}$:

$$\begin{aligned} \mathcal{G}^{(i)}(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{\theta}; \mathbf{x}^i) \\ &= \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\mathbf{x}^{(i)}}(\mathbf{z}^{(i)}|\boldsymbol{\phi}^{(i)})} \left[\log p(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}|\boldsymbol{\theta}) \right]. \end{aligned} \tag{2.99}$$

The integral with respect to $\mathbf{z}^{(i)}$ can be approximated by utilizing the MC methods in Section 2.6 with $\hat{\mathcal{G}}^{(i)}(\boldsymbol{\theta}) \approx \mathcal{G}^{(i)}(\boldsymbol{\theta})$ and the corresponding approximate estimator:

$$\hat{\mathcal{G}}^{(i)}(\boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x}^{(i)}, \boldsymbol{z}_m^{(i)}|\boldsymbol{\theta}), \tag{2.100}$$

where samples are drawn with $\boldsymbol{z}_m^{(i)} \sim q_{\mathbf{x}^{(i)}}(\mathbf{z}^{(i)}|\boldsymbol{\phi}^{(i)})$. Clearly the above approximate estimator is unavoidably affected by noise. Stochastic optimization algorithms provide a remedy and enable robust optimization, even for noisy gradient estimates. We focus on two schemes utilizing first-order gradients: (1) one of the earliest schemes for stochastic optimization, the Robbins–Monro algorithm [410], and (2) the more recent ADAM scheme [472].

Given $\boldsymbol{\theta}^t$, the general per iteration update rule for obtaining the parameters $\boldsymbol{\theta}^{t+1}$ when maximizing an objective is

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \eta_t \sum_{i=1}^{N} \hat{\mathcal{G}}(\boldsymbol{\theta}). \tag{2.101}$$

2.7.1 Robins–Monro stochastic optimization

The Robins–Monro algorithm is guaranteed to converge to the extremum under the following conditions [412]:

$$\sum_{t=1}^{\infty} \eta_t = +\infty \text{ and } \sum_{t=1}^{\infty} \eta_t^2 < \infty. \tag{2.102}$$

The step size of the gradient ascent step follows to

$$\eta_t = \frac{\alpha}{(A+t)^{\rho}}, \tag{2.103}$$

where $\rho \in (0.5, 1]$ and A and ρ are problem-dependent parameters.

[9] We consider here an objective based on a single datum that contributes as a summand to the overall objective. This is specified in the corresponding chapters.

2.7.2 ADAM stochastic optimization

ADAM is an adaptive optimization scheme that requires little user interaction in handling a large amount of data and many parameters [472]. The procedure synthesizes the advantages of two earlier developments, AdaGrad [473] and RMSProp [474], which utilize step-size annealing. The procedure is summarized in Algorithm 2, where $\odot$ denotes the element-wise product of two vectors.

Algorithm 2: ADAM [472]. Standard settings for the free parameters are as follows: $\alpha = 1 \times 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$.

Input: Step size α, exponential decay rates β_1, $\beta_2 \in [0, 1)$, stochastic gradient $\hat{\mathcal{G}}(\boldsymbol{\theta})$, initial parameters $\boldsymbol{\theta}_0$, ϵ.

Output: Optimized parameters $\boldsymbol{\theta}_t$.

Initialization:

$\mathbf{m}_0 \leftarrow 0$, $\mathbf{v}_0 \leftarrow 0$ (first- and second-moment vectors), $t \leftarrow 0$ (step).

while $\boldsymbol{\theta}_t$ *not converged* **do**

$t \leftarrow t + 1$.

Compute the gradient: $\hat{\mathcal{G}}_t(\boldsymbol{\theta}_{t-1})$.

Update biased first- and second-moment estimates:

$\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1)\hat{\mathcal{G}}_t$.

$\mathbf{v}_t \leftarrow \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2)\hat{\mathcal{G}}_t \odot \hat{\mathcal{G}}_t$.

Estimate bias-corrected moment estimates:

$$\hat{\mathbf{m}}_t \leftarrow \frac{\mathbf{m}_t}{(1 - \beta_1^t)}.$$

$$\hat{\mathbf{v}}_t \leftarrow \frac{\mathbf{v}_t}{(1 - \beta_2^t)}.$$

Update parameters:

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} + \alpha \frac{\hat{\mathbf{m}}_t}{(\sqrt{\hat{\mathbf{v}}_t}) + \epsilon}.$$

For a detailed overview of the development of stochastic optimization algorithms with a focus on deep learning, we refer to Ref. [475].

Chapter 3

Predictive coarse-graining

This chapter has been published in

M. Schöberl, N. Zabaras, P.-S. Koutsourelakis.
"Predictive coarse-graining".
In: *Elsevier Journal of Computational Physics* 333 (2017), pp. 49-77.

The following provides a summary of the scientific achievements of the above work and describes the individual contributions before closing this section.

3.1 Motivation and summary

The coarse-graining (CG) framework presented in the context of equilibrium statistical mechanics in [375] disrupts the perspective of existing CG approaches. Thus far, CG approaches have built on a many-to-one, fine-to-coarse mapping, e.g., defined by lumping multiple atoms into an effective interaction site or pseudo-molecule. In such approaches, CG variables represent the center of mass of the pseudo-molecules, and the fine-to-coarse mapping is fixed once it has been introduced. These approaches are discussed in more detail in Section 1.2.

The novelty of the data-driven predictive CG approach lies in the implicit definition of the CG variables by introducing a probabilistic and parametrizable coarse-to-fine mapping that corresponds to a directed probabilistic graphical model [476]. In this graphical model, the CG variables serve as latent generators that yield, through the probabilistic coarse-to-fine mapping, the full atomistic fine-scale representations. Here, we point out the novelties and advantages of the developed methodology based on a related information theoretic CG approach [231]. Ref. [375] rigorously builds on an information theoretic perspective and generalizes the objective towards a versatile Bayesian framework. A consistent Bayesian approach enables the quantification of epistemic uncertainties [477], which unavoidably occur in CG processes [251, 478]. We derive an efficient approach for obtaining a posterior distribution of the model parameters in the CG context. The posterior distribution expresses the aforementioned uncertainties and facilitates truly predictive estimates of macroscopic observables. First, as we enable the reconstruction of a fully atomistic picture, we can learn and predict fine-scale interdependencies and thus provide estimates

for macroscopic observables with dependence on the fine-scale coordinates. Second, using the derived predictive posterior distribution, we demonstrate the computation of credible intervals of macroscopic observables. These intervals reflect the model's predictive confidence based on limited data and the aforementioned resolution-dependent uncertainties. The Bayesian framework developed here is seamlessly hierarchically extendable, supporting model complexity and model selection tasks by functional priors; the latter promotes for the discovery of sparse solutions to reveal prominent model features. Identification of features provides physical insights that facilitate knowledge extraction given a limited amount of data. We develop a computationally efficient and embarrassingly parallelizable Monte Carlo expectation maximization scheme, addressing inference and learning of latent CG variables and model parameters.

We comprehensively assess the proposed methodology with two relevant numerical illustrations. As a demonstration of the algorithmic and conceptual enhancements, we use the identification of CG models for a lattice spin system (Ising model [479]) and an SPC/E water[1] model [334]. We employ relatively simple coarse-to-fine mappings $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\mathrm{cf}})$ and ensure the overall model flexibility of the predictive distribution by an expressive CG description $p(\mathbf{z}; \boldsymbol{\theta}_{\mathrm{c}})$.

The CG Ising model encompasses a probabilistic mapping $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\mathrm{cf}})$ (Bernoulli distribution), while the CG variables $\mathbf{z}$ interact via an expressive Boltzmann density defined by a flexible CG interaction potential $U_{\mathrm{c}}(\mathbf{z}; \boldsymbol{\theta}_{\mathrm{c}})$. The interaction potential comprises a rich set of basis functions representing the order of considered interactions and the considered CG interaction length. The CG water model involves a coarse-to-fine mapping that describes the center of mass of the fine-grained atoms probabilistically. A rich set of sine and cosine basis functions with different wavelengths compose the CG interaction potential of CG variables. The wave lengths correspond to the volatility of the two-body CG interaction. Besides two-body interations, we include a three-body term sought to capture the tetrahedral structure of water molecules [481]. The observables represent structural properties, i.e., the pairwise and angular distribution functions.

Both numerical examples provide insight into the influence of the available training data on the credibility of the model's prediction compared to a fine-scale reference estimate. In the context of the CG Ising model, we further investigate the influence of the CG level on the model's confidence. We show that, in both cases, the use of functional priors (automatic relevance determination [421]) yields sparse and physically interpretable solutions.

We reproduce the original publication with permission of ELSEVIER in Appendix B.

[1] SPC/E water serves as the solvent in most biochemical molecular dynamic simulations, where 80% of the computational time is spent on simulating water [480].

3.2 Declaration of the author's individual contribution

Conceptual drafts and strategical developments of the proposed predictive CG methodology, published in [375], were performed in collaboration with my supervisors, N. Zabaras and P.-S. Koutsourelakis. Detailed model specifications, all implementations, numerical illustrations, and conclusions were produced by me. I provide a detailed list of my contributions below.

- Research on existing CG methods with critical review.
- Research on and detailed mathematical derivation of the suggested generative probabilistic model.
- Research on inference schemes with the adoption of posterior approximation methods and sampling approaches for the development of an expectation maximization mechanism.
- Research on and mathematical derivation of expressive distributions from an exponential family based on the linear combination of basis functions.
- Suggestion of fine-scale reference systems with relevant observables to demonstrate the methodological advantages.
- Implementation of a related information theoretic CG approach, the relative entropy method [231], for comparison.
- Simulation of the reference Ising model with Monte Carlo approaches and its implementation.
- Setting up an SPC/E water reference model and carrying out molecular dynamics simulations with LAMMPS including data preparation for model learning.
- Definition of model distributions for the Ising and SPC/E water model.
- Complete C++ implementation of the proposed CG approach and parallelization with OpenMPI on an HPC cluster.
- Designing and performing numerical experiments with a focus on the influence of the amount of available training data, the level of CG, and sparseness.
- Preparation of the manuscript for submission with all necessary computation, graphics, and visualizations.

Chapter 4

Adaptive sequential model refinement for Bayesian coarse-graining

4.1 Introduction

We are interested in overcoming spatial and temporal limitations prevalent in brute-force molecular dynamics (MD) simulations by using coarse-grained (CG) atomistic systems. Scale limitations arise due to the unparalleled nature of scales of interest. For example, protein-folding simulations can take milliseconds [48] and employ a time step of femtoseconds for resolving bonded interactions [97]. CG methodologies often rely on physical intuition and chemical expertise, which is not always available for interesting complex atomistic systems [116].

Rather than assuming physical pre-knowledge, we aim to reveal physicochemical insights by learning a CG model in which the lower-dimensional characteristic CG variables serve as a generator and give rise to a fully atomistic representation. Thus, by dimensionality reduction, parsimonious physically relevant features are made accessible by inferring the latent CG variables from observed data. The general framework for this probabilistic and predictive CG approach was introduced in Chapter 3.

The following is a recap of Chapter 3 and Ref. [375]. In Ref. [375], we proposed a predictive CG framework with two components:

(i) A generator for the latent CG variables $\mathbf{z}$, $q(\mathbf{z}|\boldsymbol{\theta}_{\mathrm{c}})$, which involves parameters $\boldsymbol{\theta}_{\mathrm{c}}$.

(ii) A probabilistic coarse-to-fine mapping $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\mathrm{cf}})$, with the parameters $\boldsymbol{\theta}_{\mathrm{cf}}$.

The CG variables are implicitly defined by $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\mathrm{cf}})$ and we train the model by minimizing the KL divergence from the target Boltzmann density $p_{\mathrm{target}}(\mathbf{x})$ to the predictive density of the CG model, which we obtain by marginalizing $\mathbf{z}$:

$$q(\mathbf{x}|\boldsymbol{\theta}) = \int q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\mathrm{cf}}) q(\mathbf{z}|\boldsymbol{\theta}_{\mathrm{c}}) \, d\mathbf{z}. \tag{4.1}$$

The objective is then

$$\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}\left(p_{\mathrm{target}}(\mathbf{x}) \| q(\mathbf{x}|\boldsymbol{\theta})\right), \tag{4.2}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathrm{cf}}, \boldsymbol{\theta}_{\mathrm{c}}\}$. As we elaborated in Chapter 3, minimizing Equation 4.2 can be reformulated as maximizing the marginal likelihood or marginal log-likelihood:

$$\log q(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log q(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}), \tag{4.3}$$

which allows us to incorporate the CG approach into a Bayesian framework.

In Chapter 3, we utilized a simple density for modeling the coarse-to-fine map $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\mathrm{cf}})$ (e.g., Bernoulli or Gaussian), compensated for by an expressive $q(\mathbf{z}|\boldsymbol{\theta}_{\mathrm{c}})$, where

$$q(\mathbf{z}|\boldsymbol{\theta}_{\mathrm{c}}) \propto e^{-\beta U_{\mathrm{c}}(\mathbf{z};\boldsymbol{\theta}_{\mathrm{c}})}. \tag{4.4}$$

The potential $U_{\mathrm{c}}(\mathbf{z};\boldsymbol{\theta}_{\mathrm{c}})$, which expresses the interactions of CG variables $\mathbf{z}$, is linear with respect to $\boldsymbol{\theta}_{\mathrm{c}}$:

$$U_{\mathrm{c}}(\mathbf{z};\boldsymbol{\theta}_{\mathrm{c}}) = \boldsymbol{\theta}_{\mathrm{c}}^{T} \boldsymbol{\phi}(\mathbf{z}),$$

and

$$\boldsymbol{\theta}_{\mathrm{c}} = \begin{pmatrix} \theta_1^{\mathrm{c}} \\ \vdots \\ \theta_L^{\mathrm{c}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\phi}(\mathbf{z}) = \begin{pmatrix} \phi_1(\mathbf{z};\boldsymbol{\lambda}_1) \\ \vdots \\ \phi_L(\mathbf{z};\boldsymbol{\lambda}_L) \end{pmatrix}.$$

Note that in Ref. [375], $\boldsymbol{\lambda}_l$, the parameters of the basis, are assumed to be fixed.

Previous work follows the approach of using a rich set of basis functions $\boldsymbol{\phi}(\mathbf{z})$ (e.g., polynomials, sines, cosines, or kernels) and then, during the learning task, identifying those that are relevant for explaining the atomistic reference data. For this task, we employed a sparsity-favoring prior [420, 421], explicitly the ARD prior, which pushes the parameters θ_k of unnecessary basis functions $\phi_k(\mathbf{z})$ to zero. The ARD approach is dynamic in the sense that it can reactivate previously unnecessary features $\phi_k(\mathbf{z})$, if they explain $p_{\mathrm{target}}(\mathbf{x})$.

Model sparsity can be favored as explained above by searching for the most relevant features $\phi_k(\mathbf{z})$ in a rich set of basis functions $\boldsymbol{\phi}(\mathbf{z})$. A second way to obtain an expressive but sparse description is the following. Start initially with only a limited number of basis functions, which may have significant support in $\mathbf{z}$, and iteratively add new features as learning proceeds. Adding a feature results in a refined and more expressive model for $U_{\mathrm{c}}(\mathbf{z};\boldsymbol{\theta})$. However, it raises the following two questions:

(i) When, in terms of training iterations, is a good moment for refining $U_{\mathrm{c}}(\mathbf{z};\boldsymbol{\theta})$ by adding a new basis function?

(ii) Which basis function should be added?

We provide in this work a consistent information theoretic approach for answering both questions simultaneously.

A second novelty compared to previous work [375] is the way we express the probabilistic coarse-to-fine mapping. The approach in this chapter circumvents the need for physical understanding by introducing a Gaussian coarse-to-fine map $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\text{cf}})$, involving a linear model for expressing its mean value depending on the latent CG variable [482–484]. We demonstrate the proposed methodology for alanine dipeptide (ALA-2). Further, we investigate the influence of the amount of training data on the probabilistic estimates of observables and provide credible intervals that quantify the epistemic uncertainty induced due to the limited training data.

The rest of this chapter is structured as follows. Section 4.2 elaborates on methodological advances, where we review the basic concepts of the predictive CG approach. We specify strategies for parametrizing the densities and discuss a variational Bayesian approach for obtaining posterior distributions on the parametrization of the coarse-to-fine map. Section 4.3 demonstrates the proposed adaptive Bayesian CG framework for the ALA-2 peptide. This chapter ends with a conclusion and an outlook in Section 4.4.

4.2 Methodology

The general notation is the same as in Chapter 2. We are interested in identifying an efficient and predictive CG model that seeks to approximate the target Boltzmann distribution:

$$p_{\text{target}}(\mathbf{x}; \beta) = \frac{1}{Z(\beta)} e^{-\beta U_{\text{f}}(\mathbf{x})}. \tag{4.5}$$

We consider systems in equilibrium at constant temperatures T, and we omit the temperature dependency of $p_{\text{target}}(\mathbf{x})$ in the following. The inverse temperature is $\beta = 1/k_B T$, where k_B denotes the Boltzmann constant. The dimension of the atomistic reference is $n_{\text{f}} = \dim(\mathbf{x})$. We seek $\dim(\mathbf{z}) \ll \dim(\mathbf{x})$.

4.2.1 Bayesian CG approach

We enhance the general predictive CG approach introduced earlier in Ref. [375] with:

$$q(\mathbf{x}|\boldsymbol{\theta}) = \int q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\text{cf}}) q(\mathbf{z}|\boldsymbol{\theta}_{\text{c}}) \, d\mathbf{z}, \tag{4.6}$$

in this chapter with regards to both components, the probabilistic coarse-to-fine mapping $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\text{cf}})$ and the description of the CG variables $\mathbf{z}$, $q(\mathbf{z}|\boldsymbol{\theta}_{\text{c}})$. We introduce for the probabilistic mapping a Gaussian distribution whose mean $\boldsymbol{\mu}_{\text{cf}}$ depends linearly on $\mathbf{z}$: $\boldsymbol{\mu}_{\text{cf}}(\mathbf{z}) = \boldsymbol{\mu} + \mathbf{W}\mathbf{z}$ [483, 485]. Thus, we can write

$$q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\text{cf}}) = \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{W}\mathbf{z}, \boldsymbol{S}_{\text{cf}}\right), \tag{4.7}$$

where $\boldsymbol{S}_{\text{cf}} = \text{diag}\left(\sigma_1^2, \ldots, \sigma_{n_{\text{f}}}^2\right)$ is a diagonal covariance matrix with $n_{\text{f}} = \dim(\mathbf{x})$. The parameters of the probabilistic mapping $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\text{cf}})$ are summarized in $\boldsymbol{\theta}_{\text{cf}} =$

$\{\boldsymbol{\mu}, \sigma_1^2, \ldots, \sigma_{n_\mathrm{f}}^2\}$. We consider the matrix $\mathbf{W}$ as latent variables (like $\mathbf{z}$) in the context of variational Bayesian inference [357]. The prior distribution on $\mathbf{W}$ is a hierarchical conjugate Gaussian-Gamma prior with the hyperparameters of the Gamma distribution set in accordance with the ARD values ($a_0 = b_0 = 1 \times 10^{-5}$ [421]), as discussed in Section 2.5.3. The prior on $\mathbf{W}$, which spans the dimension $n_\mathrm{f} \times n_\mathrm{c}$, factorizes row-wise:

$$q(\mathbf{W}) = \prod_{j=1}^{n_\mathrm{f}} q(\boldsymbol{w}^{(j)}), \tag{4.8}$$

where the parameters $\boldsymbol{w}^{(j)}$ are associated with row j in $\mathbf{W}$.

The distribution $q(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c})$, specified in the following,

$$q(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c}) \propto e^{-\beta U_\mathrm{c}(\mathbf{z};\boldsymbol{\theta}_\mathrm{c}) + \Pi_{[0,1]^{n_\mathrm{c}}}(\mathbf{z})}. \tag{4.9}$$

generates the CG variables. Interactions between CG variables $\mathbf{z}$ are expressed with the potential $U_\mathrm{c}(\mathbf{z};\boldsymbol{\theta}_\mathrm{c})$, which is linear concerning $\boldsymbol{\theta}_\mathrm{c}$:

$$U_\mathrm{c}(\mathbf{z};\boldsymbol{\theta}_\mathrm{c}) = \boldsymbol{\theta}_\mathrm{c}^T \boldsymbol{\phi}(\mathbf{z}). \tag{4.10}$$

The vectors $\boldsymbol{\theta}_\mathrm{c}$ and $\boldsymbol{\phi}$ decompose to

$$\boldsymbol{\theta}_\mathrm{c} = \begin{pmatrix} \theta_1^\mathrm{c} \\ \vdots \\ \theta_L^\mathrm{c} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\phi}(\mathbf{z};\boldsymbol{\Lambda}) = \begin{pmatrix} \phi_1(\mathbf{z};\boldsymbol{\lambda}_1) \\ \vdots \\ \phi_L(\mathbf{z};\boldsymbol{\lambda}_L) \end{pmatrix}. \tag{4.11}$$

We employ in this work radial basis functions, explicitly Gaussian kernels:

$$\phi_l(\mathbf{z};\boldsymbol{\lambda}_l) = \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_{\phi_l})^T \boldsymbol{S}_{\phi_l}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{\phi_l})\right). \tag{4.12}$$

This kernel is centered at $\boldsymbol{\mu}_{\phi_l}$ and has a diagonal covariance matrix $\boldsymbol{S}_{\phi_l} = \mathrm{diag}(\sigma_{\phi_l,1}^2, \ldots, \sigma_{\phi_l,n_\mathrm{c}}^2)$. Considering the aforementioned parametrization, a single kernel $\phi_l(\mathbf{z};\boldsymbol{\lambda}_l)$ is fully defined with the set of parameters $\boldsymbol{\lambda}_l = \{\boldsymbol{\mu}_{\phi_l}, \boldsymbol{S}_{\phi_l}\}$. We summarize all parameters specifying the kernels in $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_L\}$.

The auxiliary potential $\Pi_{[0,1]^{n_\mathrm{c}}}(\mathbf{z})$ in Equation 4.9 is constant within $[0,1]^{n_\mathrm{c}}$ and has a quadratic slope beyond. This additive term in the potential naturally limits the $\mathbf{z}$ domain, which is required since we use unnormalized kernels. The auxiliary potential restricts neither the generality nor the model expressivity but rather ensures that the latent representation $\mathbf{z}^{(i)}$ of its corresponding $\mathbf{x}^{(i)}$ span the same areas to allow comparability between different training runs. We employ the following expression for that purpose:

$$\Pi_{[0,1]^{n_\mathrm{c}}}(\mathbf{z}) = \begin{cases} 1, & \text{if } \mathbf{z} \in [0,1]^{n_\mathrm{c}} \\ 100\, \mathbf{z}^T\mathbf{z}, & \text{otherwise.} \end{cases} \tag{4.13}$$

4.2.2 Inference

In the following, we discuss the parameter-learning tasks of the generative model, thus $\boldsymbol{\theta}_{\text{cf}}$ and $\boldsymbol{\theta}_{\text{c}}$. The section mostly relies on components of variational inference that were introduced and discussed in Sections 2.3, 2.4, and 2.5.2; thus, we only briefly mention here the most relevant components in terms of the specific model setting used in this chapter.

To obtain point estimates of the model parameters and for inferring the the latent variables, we denote the marginal log-likelihood as $\log q(\mathbf{x}^{\mathcal{D}_N}|\boldsymbol{\theta})$. We consider atomistic data with $\mathbf{x}^{\mathcal{D}_N} = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$ and write for a single datum $\boldsymbol{x}^{(i)}$:

$$\begin{aligned}
\mathcal{L}^{(i)}(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}) &= \log \int q(\mathbf{x}|\mathbf{z}, \mathbf{W}, \boldsymbol{\theta}_{\text{cf}}) q(\mathbf{z}|\boldsymbol{\theta}_{\text{c}})\, d\mathbf{z}\, q(\mathbf{W})\, d\mathbf{W} \\
&= \log \int r(\mathbf{z}|\mathbf{x}) r(\mathbf{W}) \frac{q(\mathbf{x}|\mathbf{z}, \mathbf{W}, \boldsymbol{\theta}_{\text{cf}}) q(\mathbf{z}|\boldsymbol{\theta}_{\text{c}}) q(\mathbf{W})}{r(\mathbf{z}|\mathbf{x}) r(\mathbf{W})}\, d\mathbf{z}\, d\mathbf{W} \\
&\geq \int r(\mathbf{z}|\mathbf{x}) r(\mathbf{W}) \log \frac{q(\mathbf{x}|\mathbf{z}, \mathbf{W}, \boldsymbol{\theta}_{\text{cf}}) q(\mathbf{z}|\boldsymbol{\theta}_{\text{c}}) q(\mathbf{W})}{r(\mathbf{z}|\mathbf{x}) r(\mathbf{W})}\, d\mathbf{z}\, d\mathbf{W} \\
&= \mathcal{F}^{(i)}(r(\mathbf{W}) r(\mathbf{z}|\mathbf{x}), \boldsymbol{\theta}; \boldsymbol{x}^{(i)}).
\end{aligned} \tag{4.14}$$

The second line of Equation 4.14 implies that

$$r(\mathbf{z}, \mathbf{W}|\boldsymbol{x}^{(i)}) \approx r(\mathbf{W}) r(\mathbf{z}|\boldsymbol{x}^{(i)}).$$

The lower bound on the log-likelihood composes the contributions per datum $\boldsymbol{x}^{(i)}$:

$$\mathcal{F}(r(\mathbf{W}), r(\mathbf{z}^{(1)}|\mathbf{x}^{(1)}), \ldots, r(\mathbf{z}^{(N)}|\mathbf{x}^{(N)}); \mathbf{x}^{\mathcal{D}_N}) = \sum_{i=1}^{N} \mathcal{F}^{(i)}(r(\mathbf{W}) r(\mathbf{z}|\mathbf{x}), \boldsymbol{\theta}; \boldsymbol{x}^{(i)}). \tag{4.15}$$

Refs. [357, 391, 405] propose a variational Bayesian expectation-maximization (VB-EM) scheme and show that maximizing Equation 4.15 with respect to $r(\mathbf{W})$ and $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ tightens the lower bound, which equals the marginal log-likelihood if

$$r(\mathbf{W}) r(\mathbf{z}|\mathbf{x}) = \frac{q(\mathbf{x}|\mathbf{z}, \mathbf{W}) q(\mathbf{z}) q(\mathbf{W})}{q(\mathbf{x})}. \tag{4.16}$$

We employ a VB-EM scheme where the VBE step optimizes the lower bound,

$$\mathcal{F}\left(r(\mathbf{W}), r(\mathbf{z}^{(1)}|\mathbf{x}^{(1)}), \ldots, r(\mathbf{z}^{(N)}|\mathbf{x}^{(N)}), \boldsymbol{\theta}; \mathbf{x}^{\mathcal{D}_N}\right),$$

with respect to the auxiliary distributions $r(\cdot)$ and evaluates the expectations in Equation 4.14. The maximization of $\mathcal{F}$ in terms of $r(\cdot)$ is reformulated as an optimization with respect to the parametrization of $r(\mathbf{z}|\mathbf{x})$. $\boldsymbol{\phi}$ summarizes parameters of the latter $r(\mathbf{z}|\mathbf{x})$. The set of distributions approximating the posterior over the latent variables are $\{r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})\}_{i=1}^N$, which we model as Gaussian densities with,

$$r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)}) = \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\phi}}^{(i)}, \boldsymbol{S}_{\boldsymbol{\phi}}^{(i)}\right). \tag{4.17}$$

The vector $\boldsymbol{\phi}^{(i)} = (\boldsymbol{\mu}_{\boldsymbol{\phi}}^{(i)}, \boldsymbol{S}_{\boldsymbol{\phi}}^{(i)})$ summarizes the parameters where

$$\boldsymbol{S}_{\boldsymbol{\phi}}^{(i)} = \mathrm{diag}(\sigma_{\boldsymbol{\phi}^1}^2, \ldots, \sigma_{\boldsymbol{\phi}^1}^2).$$

This chapter employs non-amortized inference, which implies the direct learning of $\boldsymbol{\mu}_{\boldsymbol{\phi}}^{(i)}$ and $\boldsymbol{S}_{\boldsymbol{\phi}}^{(i)}$ as parameters and not as parametrized functions that give rise to $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(i)})$ and $\boldsymbol{S}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(i)})$. The latter, amortized inference, provides a differentiable encoder mapping from any input atomistic configuration $\mathbf{x}$ and assigns a value to the associated latent CG variable $\mathbf{z}$. This is compelling in the context of collective variable discovery and enhanced sampling methods due to the differentiability of $\mathbf{z}$ with respect to $\mathbf{x}$. The gradient can be used to guide the exploration of the configuration space [486].[1]

The following addresses the computation of the approximate posterior $r(\mathbf{W}|\mathbf{x}^{\mathcal{D}_N})$, which is modeled as

$$r(\mathbf{W}|\mathbf{x}^{\mathcal{D}_N}) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{W}}^{(k)}, \boldsymbol{S}_{\mathbf{W}}\right) \tag{4.18}$$

with

$$\boldsymbol{\mu}_{\mathbf{W}}^{(k)} = \boldsymbol{S}_{\mathrm{cf}}^{-1} \sum_{i=1}^{N} \left\langle \mathbf{x}^{(i)} \overline{\mathbf{z}}^{(i)\mathrm{T}} \right\rangle_{r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})} \boldsymbol{S}_{\mathbf{W}} \tag{4.19}$$

$$\boldsymbol{S}_{\mathbf{W}} = \left(\mathrm{diag}\langle \boldsymbol{\tau} \rangle + \sum_{i=1}^{N} \boldsymbol{S}_{\boldsymbol{\phi}^{(i)}}^{-1} \left\langle \mathbf{z}^{(i)} \mathbf{z}^{(i)T} \right\rangle_{r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})} \right)^{-1} \tag{4.20}$$

where

$$\overline{\mathbf{z}}^{(i)\mathrm{T}} = \boldsymbol{\Sigma}_{\overline{\mathbf{z}}} \mathbf{W}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{cf}}^{-1} \mathbf{x}^{(i)} \tag{4.21}$$

$$\boldsymbol{\Sigma}_{\overline{\mathbf{z}}} = \left(\left\langle \mathbf{W}^T \boldsymbol{S}_{\mathrm{cf}}^{-1} \mathbf{W} \right\rangle + \boldsymbol{I} \right)^{-1}. \tag{4.22}$$

The vector $\boldsymbol{\tau}$ contains the precision entries of the Gaussian distribution involved in the ARD prior based on Section 2.5.3 and the equations defining the approximate posterior distributions mostly rely on Refs. [490–494].

The VBE step requires the computation of gradients:

$$\nabla_{\boldsymbol{\phi}^{(i)}} \mathcal{F}^{(i)} \left(r(\mathbf{W}), r(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}^{(i)}), \boldsymbol{\theta}; \boldsymbol{x}^{(i)} \right).$$

Based on Equation 4.14, this involves, when regarding one datum $\boldsymbol{x}^{(i)}$, the following expression[2]:

$$\nabla_{\boldsymbol{\phi}} \mathcal{F} \left(r(\mathbf{W}), r(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}), \boldsymbol{\theta}; \boldsymbol{x}^{(i)} \right) = \int r(\mathbf{W}) \, \nabla_{\boldsymbol{\phi}} \left(\int r(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) \left(\log q(\mathbf{x}|\mathbf{z}, \mathbf{W}, \boldsymbol{\theta}_{\mathrm{cf}}) + \log q(\mathbf{z}|\boldsymbol{\theta}_{\mathrm{c}}) - \log r(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) \right) d\mathbf{z} \right) d\mathbf{W}. \tag{4.23}$$

[1] We consider an amortized inference approach [406, 407, 487–489] in Chapter 5.

[2] We omit here the superscript notation with $(\cdot)^{(i)}$ for improving readability.

The expectation under $r(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})$ in above equation is intractable and we utilize MC approximations. However, the MC estimate of the gradient suffers from noise, as discussed in Ref. [406]. A reparametrized version of Equation 4.23 provides remedy and reduces the induced noise in the estimator [406].

For thus purpose, we introduce an auxiliary random variable $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, which gives rise to [408],

$$\mathbf{z}(\boldsymbol{\epsilon};\boldsymbol{\phi}) = \boldsymbol{\mu}_{\boldsymbol{\phi}} + \boldsymbol{\sigma}_{\boldsymbol{\phi}} \odot \boldsymbol{\epsilon}. \tag{4.24}$$

By using the change of variables in Equation 4.24, the inner integral of Equation 4.23 becomes,

$$\begin{aligned}
&\nabla_{\boldsymbol{\phi}} \left\langle \log q(\mathbf{x},\mathbf{z}|\mathbf{W},\boldsymbol{\theta}) - \log r(\mathbf{z}|\mathbf{x},\boldsymbol{\phi}) \right\rangle_{r(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})} \\
&\qquad = \nabla_{\boldsymbol{\phi}} \left\langle \log q(\mathbf{x},\mathbf{z}|\mathbf{W},\boldsymbol{\theta}) - \log r(\mathbf{z}|\mathbf{x},\boldsymbol{\phi}) \right\rangle_{p(\boldsymbol{\epsilon})} \\
&\qquad = \left\langle \frac{\partial \log q(\mathbf{x},\mathbf{z}(\boldsymbol{\epsilon};\boldsymbol{\phi})|\mathbf{W},\boldsymbol{\theta})}{\partial \mathbf{z}} \frac{\partial \mathbf{z}(\boldsymbol{\epsilon};\boldsymbol{\phi})}{\boldsymbol{\phi}} \right\rangle_{p(\boldsymbol{\epsilon})} \\
&\qquad\quad - \left\langle \frac{\log r(\mathbf{z}(\boldsymbol{\epsilon};\boldsymbol{\phi})|\mathbf{x},\boldsymbol{\phi})}{\partial \mathbf{z}} \frac{\partial \mathbf{z}(\boldsymbol{\epsilon};\boldsymbol{\phi})}{\boldsymbol{\phi}} \right\rangle_{p(\boldsymbol{\epsilon})}.
\end{aligned} \tag{4.25}$$

Note that the above expression still needs to be assessed as an expectation under $r(\mathbf{W})$.

The VBM step, as discussed in more detail in Section 2.5.2, optimizes the lower bound $\mathcal{F}$ with respect to parameters $\boldsymbol{\theta}$. These include $\boldsymbol{\theta}_{\text{cf}}$, the mean and covariance of $q(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}_{\text{cf}})$, and $\boldsymbol{\theta}_{\text{c}}$ associated with the interaction potential of the CG variables $\mathbf{z}$, $U_{\text{c}}(\mathbf{z};\boldsymbol{\theta}_{\text{c}})$.

4.2.3 Exponential family densities: Uniqueness of solution

Our previous work on predictive CG (Appendix B or Section 2.4 in [375]) demonstrates the uniqueness of the solution when employing throughout densities belonging to the exponential family, as in Equation 4.26. A similar discussion is valid for this chapter, which we show below in an modified version of the original discussion in Ref. [375].

The distributions for $q(\mathbf{z}|\boldsymbol{\theta}_{\text{c}})$ and $q(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}_{\text{cf}})$ can be expressed in form the exponential family:

$$q(\mathbf{z}|\boldsymbol{\theta}_{\text{c}}) = \exp\{\boldsymbol{\theta}_{\text{c}}^T \boldsymbol{\phi}(\mathbf{z}) - A(\boldsymbol{\theta}_{\text{c}})\}, \tag{4.26}$$

$$q(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}_{\text{cf}}) = \exp\{\boldsymbol{\theta}_{\text{cf}}^T \boldsymbol{\psi}(\mathbf{x},\mathbf{z}) - B(\mathbf{z},\boldsymbol{\theta}_{\text{cf}})\}. \tag{4.27}$$

In the above equations, $A(\boldsymbol{\theta}_\mathrm{c})$ and $B(\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf})$ are normalization constants or partition functions, which are defined as

$$A(\boldsymbol{\theta}_\mathrm{c}) = \log \int e^{\boldsymbol{\theta}_\mathrm{c}^T \boldsymbol{\phi}(\mathbf{z})} d\mathbf{z}, \tag{4.28}$$

$$B(\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf}) = \log \int e^{\boldsymbol{\theta}_\mathrm{cf}^T \boldsymbol{\psi}(\mathbf{x},\mathbf{z})} d\mathbf{x}. \tag{4.29}$$

The derivatives of the partition function with respect to the associated parametrization are

$$\begin{aligned} \frac{\partial A(\boldsymbol{\theta}_\mathrm{c})}{\partial \theta_{\mathrm{c},k}} &= \langle \phi_k(\mathbf{z}) \rangle_{q(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c})}, \\ \frac{\partial^2 A(\boldsymbol{\theta}_\mathrm{c})}{\partial \theta_{\mathrm{c},k} \partial \theta_{\mathrm{c},l}} &= \mathrm{Cov}_{q(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c})}[\phi_k(\mathbf{z}), \phi_l(\mathbf{z})], \end{aligned} \tag{4.30}$$

and

$$\begin{aligned} \frac{\partial B(\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf})}{\partial \theta_{\mathrm{cf},k}} &= \langle \psi_k(\mathbf{x}, \mathbf{z}) \rangle_{q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf})}, \\ \frac{\partial^2 B(\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf})}{\partial \theta_{\mathrm{cf},k} \partial \theta_{\mathrm{cf},l}} &= \mathrm{Cov}_{q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_\mathrm{cf})} \left[\psi_k(\mathbf{x}, \mathbf{z}), \psi_l(\mathbf{x}, \mathbf{z})\right]. \end{aligned} \tag{4.31}$$

We denote with $\langle \cdot \rangle_p$ in the above expressions the expectation under the density p and $\mathrm{Cov}_p[\cdot, \cdot]$ is the covariance of the arguments with respect to p.

Thus, we can write for the gradients of the objective $\mathcal{F}$:

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial \theta_{\mathrm{c},k}} &= \sum_{i=1}^{N} \left(\left\langle \phi_k(\mathbf{z}^{(i)}) \right\rangle_{r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})} - \langle \phi_k(\mathbf{z}) \rangle_{q(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c})} \right), \\ \frac{\partial \mathcal{F}}{\partial \theta_{\mathrm{cf},k}} &= \sum_{i=1}^{N} \left(\left\langle \psi_k(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \right\rangle_{r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})} - \left\langle \psi_k(\mathbf{x}, \mathbf{z}^{(i)}) \right\rangle_{q(\mathbf{x}|\mathbf{z}^{(i)}, \boldsymbol{\theta}_\mathrm{cf}) r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})} \right). \end{aligned} \tag{4.32}$$

Fundamental to the proposed adaptive model refinement approach are gradients $\nabla \boldsymbol{\theta}_\mathrm{c} \mathcal{F}$, which we specify at this point based on the employed model parametrization for $U_\mathrm{c}(\mathbf{x}; \boldsymbol{\theta}_\mathrm{c})$:

$$\nabla_{\boldsymbol{\theta}_\mathrm{c}} \mathcal{F} = \beta \langle \boldsymbol{\phi}(\mathbf{z}) \rangle_{q(\mathbf{z}|\boldsymbol{\theta})} - \beta \langle \boldsymbol{\phi}(\mathbf{z}) \rangle_{r(\mathbf{z}|\mathbf{x}^{(i)}, \boldsymbol{\theta}_\mathrm{cf})}. \tag{4.33}$$

Given the gradient, we can obtain the Hessian with the general form for exponential family distributions:

$$\begin{aligned} \frac{\partial^2 \mathcal{F}}{\partial \theta_{\mathrm{c},k} \theta_{\mathrm{c},l}} &= -N \, \mathrm{Cov}_{q(\mathbf{z}|\boldsymbol{\theta}_\mathrm{c})} \left[\phi_k(\mathbf{z}), \phi_l(\mathbf{z})\right], \\ \frac{\partial^2 \mathcal{F}}{\partial \theta_{\mathrm{c},k} \theta_{\mathrm{cf},l}} &= 0, \\ \frac{\partial^2 \mathcal{F}}{\partial \theta_{\mathrm{cf},k} \theta_{\mathrm{cf},l}} &= -\sum_{i=1}^{N} \mathrm{Cov}_{q(\mathbf{x}|\mathbf{z}^{(i)}, \boldsymbol{\theta}_\mathrm{cf}) r(\mathbf{z}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})} \left[\psi_k(\mathbf{x}, \mathbf{z}), \psi_l(\mathbf{x}, \mathbf{z})\right]. \end{aligned} \tag{4.34}$$

Composing the above second derivatives into a Hessian matrix yields a block-diagonal structure with a linear combination of parameters $(\boldsymbol{\theta}_\mathrm{c}, \boldsymbol{\theta}_\mathrm{cf})$ and features $(\boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\psi}(\mathbf{x}))$. Since a block-diagonal Hessian is always negative definite, we conclude that

the objective $\mathcal{F}$ is concave and thus, has a unique maximum.

Note that the approximation of the Hessian in Equation 4.34 by MC estimators implies a noisy estimate. In addition to the Hessian, all gradient computations rely on MC estimators, which are likewise afflicted by noise. Developments in optimization schemes [495–499], with focus on CG approaches [194], lead, nevertheless, to convergent optimization procedures. This work relies on ADAM stochastic optimization with parameters as presented in Section 2.7 and originally in Ref. [472].

4.2.4 Adaptive sequential model refinement

In Section 4.2.2, we introduced a VB-EM algorithm that alternates between an E step and an M step in inferring hidden variables and optimizing the model parameters $\boldsymbol{\theta}$ with the overall goal of maximizing the objective $\mathcal{F}$. Convergence in maximizing the parameters is reached when $\mathcal{F}$ provides a plateau and does not improve further. Corresponding to the latter case, the gradients stochastically fluctuate around zero. This means that a sufficiently flexible model can be improved only: (1) by using new insights, (2) by employing better approximations in the inference, or (3) by reducing noise in the gradient estimators. However, if the model is too restrictive compared to the data that it is supposed to explain, it can converge at higher levels $\mathcal{F}$ by enhancing the flexibility of the model. In the proposed CG framework, we pursue this strategy by sequentially adding expressivity into $U_\mathrm{c}(\mathbf{x})$. Upon convergence of the objective $\mathcal{F}$ given the current model, we add a new feature that enhances the flexibility. Thereafter, learning the augmented model parameters continues, including that associated with the added feature.

The alternating process between adding features and optimizing the new model continues as long as enriching the model leads to convergence at higher levels of the objective $\mathcal{F}$. We operate on the refinement of $U_\mathrm{c}(\mathbf{x};\boldsymbol{\theta}_\mathrm{c})$. However, this does not limit the applicability of the process for other model components, such as $q(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}_\mathrm{cf})$. The most crucial question in the context of model refinement is how to identify features that maximize the anticipated added value of the model.

Recall that we express the CG potential $U_\mathrm{c}(\mathbf{z};\boldsymbol{\theta}_\mathrm{c})$ by a linear combination of basis functions, which we often refer to as features, and the associated parameters $\boldsymbol{\theta}_\mathrm{c}$, as originally given by Equation 4.10, as

$$U_\mathrm{c}(\mathbf{z};\boldsymbol{\theta}_\mathrm{c}) = \boldsymbol{\theta}_\mathrm{c}^T \boldsymbol{\phi}(\mathbf{z}),$$

where, as introduced in Equation 4.10,

$$\boldsymbol{\theta}_\mathrm{c} = \begin{pmatrix} \theta_1^\mathrm{c} \\ \vdots \\ \theta_L^\mathrm{c} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\phi}(\mathbf{z};\boldsymbol{\Lambda}) = \begin{pmatrix} \phi_1(\mathbf{z};\boldsymbol{\lambda}_1) \\ \vdots \\ \phi_L(\mathbf{z};\boldsymbol{\lambda}_L) \end{pmatrix}.$$

The family of feature functions in this work, i.e., the radial basis functions, and more explicitly Gaussian kernels, does not limit the generality of the proposed approach for other basis functions. The Gaussian kernels (Equation 4.12) are

$$\phi_l(\mathbf{z};\boldsymbol{\lambda}_l) = \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_{\phi_l})^T \mathbf{S}_{\phi_l}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{\phi_l})\right),$$

where $\mathbf{S}_{\phi_l} = \mathrm{diag}(\sigma^2_{\phi_l,1},\ldots,\sigma^2_{\phi_l,n_c})$ is the covariance matrix, which has a diagonal structure, and $\boldsymbol{\mu}_{\phi_l}$ defines the center of the kernels. For this parametrization, a single feature $\phi_l(\mathbf{z};\boldsymbol{\lambda}_l)$ is then fully defined by the set of parameters $\boldsymbol{\lambda}_l = \{\boldsymbol{\mu}_{\phi_l}, \mathbf{S}_{\phi_l}\}$. We summarize all the parameters specifying kernels as $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_L\}$.

We denote the current set of converged parameters with $\boldsymbol{\theta}_c^{\mathrm{conv}}$ and the associated CG potential as $U_c^{\mathrm{conv}}(\mathbf{z};\boldsymbol{\theta}_c^{\mathrm{conv}})$ based on the set of features $\boldsymbol{\phi}(\mathbf{z};\boldsymbol{\Lambda})$. The current model encompasses L features. Thus, $\dim(\boldsymbol{\theta}_c^{\mathrm{conv}}) = L$ and likewise $\dim(\boldsymbol{\phi}) = L$. We seek to refine $U_c^{\mathrm{conv}}(\mathbf{z})$ by adding a new feature $\phi_{L+1}(\mathbf{z};\boldsymbol{\lambda}_{L+1})$ and write

$$U_c(\mathbf{z},\boldsymbol{\theta}_c^{\mathrm{conv}},\theta_{L+1}) = U_c^{\mathrm{conv}}(\mathbf{z};\boldsymbol{\theta}_c^{\mathrm{conv}}) + \theta_{L+1}\,\phi_{L+1}(\mathbf{z};\boldsymbol{\lambda}_{L+1}), \tag{4.35}$$

with the initial value for $\theta_{L+1} = 0$, which implies that the potential before adding the new feature equals the potential after adding a new basis function, $U_c(\mathbf{z},\boldsymbol{\theta}_c^{\mathrm{conv}},\theta_{L+1}) = U_c(\mathbf{z},\boldsymbol{\theta}_c^{\mathrm{conv}})$. The overall optimization is driven by minimizing the KL divergence, which can be optimized by maximizing the lower bound on the marginal log-likelihood:

$$\begin{aligned}
D_{\mathrm{KL}}\left(p_{\mathrm{target}}(\mathbf{x})\|q(\mathbf{x}|\boldsymbol{\theta})\right) &= -\left\langle \log q(\mathbf{x}|\boldsymbol{\theta})\right\rangle_{p_{\mathrm{target}}(\mathbf{x})} + \left\langle \log p_{\mathrm{target}}(\mathbf{x})\right\rangle_{p_{\mathrm{target}}(\mathbf{x})} \\
&\approx -\frac{1}{N}\sum_{i=1}^{N}\log q(\mathbf{x}|\boldsymbol{\theta}) + \left\langle \log p_{\mathrm{target}}(\mathbf{x})\right\rangle_{p_{\mathrm{target}}(\mathbf{x})} \\
&\leq -\frac{1}{N}\sum_{i=1}^{N}\left(\left\langle \log q(\mathbf{x}^{(i)}|\mathbf{z},\boldsymbol{\theta}_{\mathrm{cf}})\right\rangle_{r(\mathbf{z}|\mathbf{x}^{(i)})} + \left\langle \log q(\mathbf{z}|\boldsymbol{\theta}_c)\right\rangle_{r(\mathbf{z}|\mathbf{x}^{(i)})}\right) \\
&\quad + \left\langle \log p_{\mathrm{target}}(\mathbf{x})\right\rangle_{p_{\mathrm{target}}(\mathbf{x})} \\
&= -\frac{1}{N}\sum_{i=1}^{N}\mathcal{F}(r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}),\boldsymbol{\theta};\boldsymbol{x}^{(i)}) + \left\langle \log p_{\mathrm{target}}(\mathbf{x})\right\rangle_{p_{\mathrm{target}}(\mathbf{x})},
\end{aligned} \tag{4.36}$$

where the lower bound on the log-likelihood or upper bound on the KL divergence $\mathcal{F}(r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}),\boldsymbol{\theta};\boldsymbol{x}^{(i)})$ in Equation 4.36 depends on the (approximate) posterior $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$. The latter does not have an explicit dependence on $U_c(\mathbf{z};\boldsymbol{\theta}_c)$. However, changing $U_c(\mathbf{z};\boldsymbol{\theta}_c)$ implies different optima for $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$.

We propose to add the feature $\phi_{L+1}(\mathbf{z};\boldsymbol{\lambda}_{L+1})$ defined by $\boldsymbol{\lambda}_{L+1}$, which maximizes the absolute value of the derivative of $\mathcal{F}$ with respect to the new parameter θ_{L+1} at $\theta_{L+1} = 0$. The new feature is then supposed to maximize the anticipated gain when augmenting the potential $U_c(\mathbf{z},\boldsymbol{\theta}_c^{\mathrm{conv}})$ by the feature $\phi_{L+1}(\mathbf{z};\boldsymbol{\lambda}_{L+1}^{\mathrm{opt}})$.

This involves the derivative of $\mathcal{F}$ with respect to θ_{L+1}:

$$\frac{d\mathcal{F}}{d\theta_{L+1}} = \sum_{i=1}^{N} \left\langle \frac{d \log q(\mathbf{z}|\boldsymbol{\theta}_c, \theta_{L+1})}{d\theta_{L+1}} \right\rangle_{r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi})}, \tag{4.37}$$

and the expression $\log q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{conv}}, \theta_{L+1})$ in terms of the old features and the feature to be added $\phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}^{\text{opt}})$:

$$\log q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{conv}}) = -\beta U_c(\mathbf{z}; \boldsymbol{\theta}_c^{\text{conv}}) - \beta\theta_{L+1}\phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}^{\text{opt}}) - Z(\boldsymbol{\theta}_c^{\text{conv}}, \theta_{L+1}), \tag{4.38}$$

where we consider $\boldsymbol{\theta}_c^{\text{conv}}$ as fixed while identifying the optimal feature $\phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}^{\text{opt}})$. We obtain the relevant argument of the expectation in Equation 4.37 with,

$$\begin{aligned} \left.\frac{d \log q(\mathbf{x}|\boldsymbol{\theta}_c^{\text{conv}})}{d\theta_{L+1}}\right|_{\theta_{L+1}=0} &= -\beta\phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}) - \left.\frac{d \log Z(\theta_{L+1})}{d\theta_{L+1}}\right|_{\theta_{L+1}=0} \\ &= -\beta\phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}) + \beta \left\langle \phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}) \right\rangle_{q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{comv}})}. \end{aligned} \tag{4.39}$$

Substituting Equation 4.39 into Equation 4.37 finally leads to

$$\frac{d\mathcal{F}}{d\theta_{L+1}} = \beta \sum_{i=1}^{N} \left(\left\langle \phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}) \right\rangle_{r(\mathbf{z}|\mathbf{x}^{(i)}, \boldsymbol{\phi})} - \left\langle \phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}) \right\rangle_{q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{conv}})} \right). \tag{4.40}$$

As mentioned earlier, the absolute value or squared value of the gradient of $\mathcal{F}$ with respect to θ_{L+1} at $\theta_{L+1} = 0$ is objective to maximization with respect to the parameters $\boldsymbol{\lambda}_{L+1}$ of the new feature $\phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1})$. We employ the squared value of Equation 4.40 which is proportional to:

$$H(\boldsymbol{\lambda}_{L+1}) = \left(\sum_{i=1}^{N} \left(\left\langle \phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}) \right\rangle_{r(\mathbf{z}|\mathbf{x}^{(i)}, \boldsymbol{\phi})} - \left\langle \phi_{L+1}(\mathbf{z}; \boldsymbol{\lambda}_{L+1}) \right\rangle_{q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{conv}})} \right) \right)^2. \tag{4.41}$$

Maximizing $H(\boldsymbol{\lambda}_{L+1})$ in Equation 4.41 with respect to $\boldsymbol{\lambda}_{L+1}$ implies searching for the basis ϕ_{L+1} that yields the largest discrepancy between its aggregated expectation under $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi})$ and $q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{conv}})$. By adding the new feature ϕ_{L+1}, the new latent generator $q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{conv}}, \theta_{L+1})$ should be enriched such that it provides a new density $q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{conv}}, \theta_{L+1})$, which is more promising for capturing the aggregated posterior $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi})$ over the latent CG variables. If the aggregated approximate posterior and the generative distribution over the latent variables $\mathbf{z}$ match exactly, then $H(\boldsymbol{\lambda}_{L+1}) = 0$. Thus, maximizing $H(\boldsymbol{\lambda}_{L+1})$ gives rise to the basis that is most poorly represented in $q(\mathbf{z}|\boldsymbol{\theta}_c^{\text{conv}})$ compared to the aggregated posterior distributions $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi})$ for all $\boldsymbol{x}^{(i)} \in \mathbf{x}^{\mathcal{D}_N}$.

After having obtained the optimal $\boldsymbol{\lambda}_{L+1}$, there are two potential avenues for optimizing $\boldsymbol{\theta}_c$:

(i) Fix all $\boldsymbol{\theta}_c$ and continue optimizing only the one parameter associated with the new feature ϕ_{L+1}, θ_{L+1}.

(ii) Optimize the whole set $\boldsymbol{\theta}_c$, including previous parameters and the new θ_{L+1} associated with the added feature.

Either way, once we added a new feature, specified by the optimal parameter $\boldsymbol{\lambda}_{L+1}$, it remains fixed in the continuing learning process.

Before ending this section, we note some potential pitfalls of the proposed sequential adaptive refinement scheme that require some consideration. The objective $H(\boldsymbol{\lambda}_{L+1})$ can be maximized by adjusting a multiplicative constant in ϕ_{L+1}, which is misleading. Therefore, we suggest employing only features ϕ that are insensitive to scalar multiplication, for example, by employing normalized kernels. Alternatively, and this is the approach we follow, restrict the overall domain of possible $\mathbf{z}$ values. We implement this by naturally limiting the domain to $\mathbf{z} \in [0, 1]^{n_c}$. The latter neither restricts the expressivity of the obtained CG model nor limits the proposed method.

Moreover, the proposed approach for sequentially adding features ϕ by maximizing the anticipated benefit expressed in H does not guarantee that the KL-divergence in Equation 4.36 will decay. This may have several reasons. First, we rely on suboptimal posterior distributions over the latent variables, $r(\mathbf{z}|\mathbf{x})$, which could differ from the exact $q(\mathbf{z}|\mathbf{x})$. Second, the optimization of H in Equation 4.41 is based on $\theta_{L+1} = 0$, which means that the KL divergence can become smaller, remain constant, or increase if $\theta_{L+1} \neq 0$.

Refer to Algorithm 3 for a summary of the proposed adaptive model refinement procedure.

Algorithm 3: Adaptive sequential model refinement

Input: Initial $U_c(\mathbf{x};\boldsymbol{\theta}) = \theta_1\phi_1(\mathbf{z};\boldsymbol{\lambda}_1)$, with $\phi_1(\mathbf{z};\boldsymbol{\lambda}_1)$ having large support in the $\mathbf{z}$ domain.[a] Provide a family of features $\phi(\mathbf{z};\boldsymbol{\lambda})$ parametrized by $\boldsymbol{\lambda}$.

Output: Refined CG potential $U_c(\mathbf{x};\boldsymbol{\theta})$.

$L \leftarrow 1$

while $D_{KL}\left(p_{target}(\mathbf{x})\|q(\mathbf{x}|\boldsymbol{\theta})\right)$ *continues decaying by adding new features* $\phi_{L+1}(\mathbf{z};\boldsymbol{\lambda}_{L+1})$ **do**

Model training:

Optimize the current model, as discussed in Section 4.2.2. Upon convergence of $\mathcal{F}$, obtain:

$\boldsymbol{\theta}_c^{\text{conv}}$ $(\dim(\boldsymbol{\theta}) = L)$.

Add a new feature by maximizing $H(\boldsymbol{\lambda})$, the (anticipated) benefit:

$\boldsymbol{\lambda}_{L+1} = \arg\max_{\boldsymbol{\lambda}} H(\boldsymbol{\lambda})$ (Equation 4.41).[b]

Augment the parameter vector:

$\boldsymbol{\theta}_c \leftarrow (\boldsymbol{\theta}_c^{\text{conv}}, \theta_{L+1})$, where $\theta_{L+1} = 0$.

Augment the parameter vector of the features:

$\boldsymbol{\Lambda} \leftarrow (\boldsymbol{\Lambda}_{1:L}, \boldsymbol{\lambda}_{L+1})$.

Note that after obtaining $\boldsymbol{\lambda}_{L+1}$, the parameters corresponding to the feature remain fixed. Only the corresponding parameter vector $\boldsymbol{\theta}_c$ is optimized thereafter.[c]

Augment the feature vector:

$\boldsymbol{\phi}(\mathbf{z};\boldsymbol{\Lambda}) \leftarrow (\boldsymbol{\phi}_{1:L}, \boldsymbol{\phi}_{L+1}(\mathbf{z};\boldsymbol{\lambda}_{L+1}))$.

Update the current step:

$L \leftarrow L + 1$.

[a] We initialize the mean of the added Gaussian kernels, $\boldsymbol{\mu}_L^{\phi}$, based on sampling from a uniform distribution in $[0,1]^{n_c}$ and use $\sigma_{\phi,1}^2 = 4 \cdot \mathbf{1}$, where $\mathbf{1}$ denotes a vector with all entries being one.

[b] We employ a Broyden–Fletcher–Goldfarb–Shanno optimization approach for maximizing $H(\boldsymbol{\lambda})$.

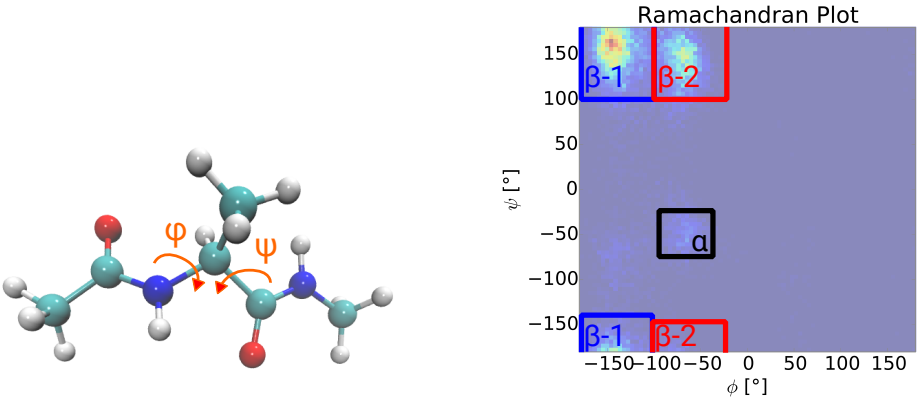
[c] There are different possibilities for optimizing $\boldsymbol{\theta}_c$, e.g., consider only the new θ_{L+1} in the optimizer or the whole $\boldsymbol{\theta}_c$.

4.3 Numerical illustration: ALA-2

This section assesses the performance of the proposed data-driven CG method for alanine dipeptide (ALA-2). We investigate, next to the adaptive model refinement, aspects of the influence of the amount of available training data N on predictive capabilities.

The ALA-2 peptide exhibits three distinct conformational modes[3], which are characterized by their corresponding dihedral angles (ϕ, ψ) (defined in Figure 4.1(a)).

[3] We observe three distinct modes but there are definitely more intermediate conformations [500].



(a) Dihedral angels for ALA-2.

(b) Characteristic conformations according [500].

FIGURE 4.1: Dihedral angles (left) and (ϕ, ψ) statistics of a reference simulation with characteristic modes (right).

The characteristic conformations are α, β-1, and β-2 depicted in the Ramachandran plot [501] in Figure 4.1(b). The considered ALA-2 peptide is a system with 22 atoms of the elements hydrogen, oxygen, nitrogen, and carbon. By ignoring rigid-body motions, one can fully describe the system with 60 degrees of freedom. We employ a Cartesian coordinate representation in $\mathbf{x}$ and we include the whole set of atomistic coordinates, thus $\dim(\mathbf{x}) = 66$. The coordinate ordering and the associated meaning as used in the vector $\mathbf{x}$ are explained in Appendix E.3.

The initial values for the mean of the basis functions, $\boldsymbol{\mu}_{\phi_l}$, are drawn from a uniform distribution in $[0,1]^{n_c}$ and we set the covariance matrix to $\boldsymbol{S}_{\phi_l} = 4 \cdot \boldsymbol{I}$. The latter resembles, given the domain $\mathbf{z} \in [0,1]^{n_c}$, a kernel that provides through $q(\mathbf{x}|\boldsymbol{\theta})$ a relatively wide support. We employ in the following experiments the model setting introduced in Section 4.2.1 and search sequentially for the optimal feature to be added based on Gaussian kernels.

We consecutively define all necessary numerical details for obtaining the presented results. The expressions in Section 4.2 involve expectations with respect to $r(\mathbf{z}|\mathbf{x})$, $r(\mathbf{W})$, and $q(\mathbf{z}|\boldsymbol{\theta})$. We approximate these with the MC methods [306, 502] that were introduced in Section 2.6.2. In contrast, the distributions $r(\mathbf{z}|\mathbf{x})$ and $r(\mathbf{W})$ are simple and have a tractable cumulative distribution function. The generator of CG variables, the distribution $q(\mathbf{z}|\boldsymbol{\theta}_c)$, is initially simple while we seek to incorporate complexity by refining $U_c(\mathbf{z})$. A refined $U_c(\mathbf{z})$ potentially leads to distinct modes so that employing random walk MCMC methods could hamper the sufficient exploration of $q(\mathbf{z}|\boldsymbol{\theta}_c)$ given a limited number of samples J. The whole procedure crucially depends upon having unbiased samples of $q(\mathbf{z}|\boldsymbol{\theta}_c)$. Adaptive sequential Monte Carlo [205] (SMC) leverages in this work the efficient exploration of multimodal potential energy surfaces evoked by adaptive model refinement. Adaptive SMC was introduced in Section 2.6.5 and summarized in Algorithm 1. We employ this parallelizable particle-based approach for obtaining $J = 4000$ samples from

$q(\mathbf{z}|\boldsymbol{\theta}_c)$ for estimating expectations $\langle\cdot\rangle_{q(\mathbf{z}|\boldsymbol{\theta}_c)}$. Given a datum $\boldsymbol{x}^{(i)}$, we approximate the expectations under $r(\mathbf{z}|\mathbf{x}^{(i)})$ with $M = 20$ samples. This is a relatively low number compared to J. However, note that convergence is also guaranteed even if we would employ only $M = 1$ and that we need draw M samples from the N posterior distributions $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$. $M = 20$, as used in the experiments, is a compromise between accuracy and efficiency. Further, as we see later, the variances in the VBE step for the approximate posterior distributions $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ are small compared to the domain of $\mathbf{z}$. Thus small M suffices.

The approximate posterior $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})$ is obtained by optimizing its parameters $\boldsymbol{\phi}^{(i)}$ during the VBE step (Section 4.2.2) by employing black box variational inference [406], as discussed in the methodology section. We optimize the mean and diagonal entries of the covariance matrix associated with $r(\mathbf{z}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})$ by employing the Robins–Monro method [410] with the parameter setting given in Section 2.7.1. We maximally perform 100 iterations per E step in the optimization of $\boldsymbol{\phi}^{(i)}$ and employ as initial values the converged $\boldsymbol{\phi}^{(i)}$ from the previous E step. Usually, convergence is reached after less than 15 Robins–Monro updates. Since we use non-amortized inference, the posterior distributions are independent given the datum $\boldsymbol{x}^{(i)}$. We make use of this independence by embarrassingly parallelizing the E step. The M step, which optimizes $\mathcal{F}$ with respect to $\boldsymbol{\theta}$, employs ADAM [472] as a stochastic optimization algorithm with the published standard parameter settings.

4.3.1 Adaptive feature learning

In the first part of the numerical illustration, we demonstrate the adaptive refinement procedure, as summarized in Algorithm 3. We emphasize that the results presented were obtained by optimizing the whole set of parameters in $\boldsymbol{\theta}_c$. Strategies where one optimizes only the θ_{L+1} associated with the new feature $\phi(\mathbf{z}; \boldsymbol{\lambda}_{L+1})$ while keeping all others fixed upon convergence in earlier stages are feasible. However, continuing with the complete $\boldsymbol{\theta}_c$ and thus, also those parameters associated with features added from earlier stages leads to superior overall performance. Note that, once the anticipated optimal feature is found by identifying $\boldsymbol{\lambda}_{L+1}$, the corresponding $\boldsymbol{\lambda}$ parameters remain fixed. For visual representability, we set $\dim(\mathbf{z}) = 2$. We employ for training a data set $\mathbf{x}^{\mathcal{D}_N}$ with $N = 526$ reference atomistic snapshots. The initial $U_c(\mathbf{z})$, including one basis function and associated $\theta_1 = 0$, is a uniform distribution on the domain of $\mathbf{z}$. We depict the initial $U_c(\mathbf{z})$ in Figure 4.2. Once θ_1 converged, we begin the adaptive refinement of $U_c(\mathbf{z})$ and alternate between optimizing $\boldsymbol{\theta}$ and refining $U_c(\mathbf{x})$ by adding new features by identifying $\boldsymbol{\lambda}$ of $\phi(\mathbf{z}; \boldsymbol{\lambda})$ maximizing the anticipated benefit, as discussed in Section 4.2.4. Figure 4.3 demonstrates the process of adding new features and the corresponding CG potential energy surface $U_c(\mathbf{x})$ (left column) at early training stages, i.e., after four features have been added. The right column in Figure 4.3 depicts the negative aggregate and average

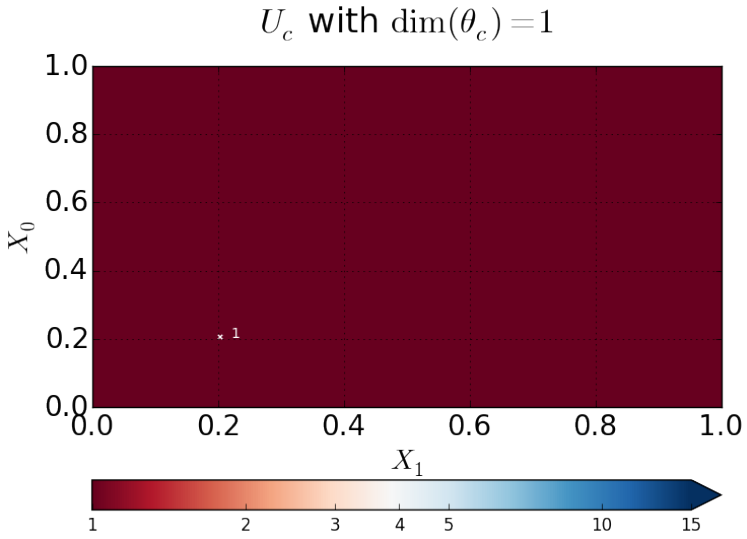


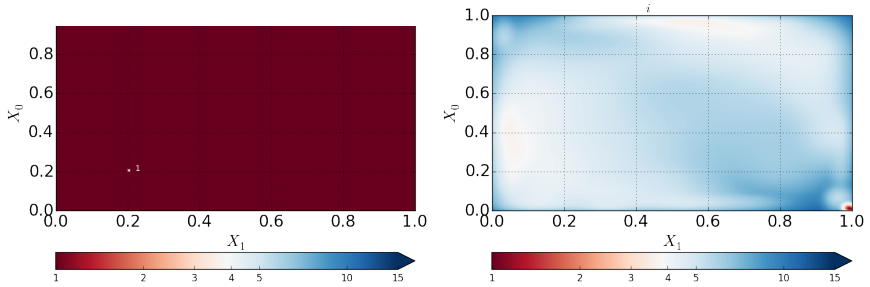
FIGURE 4.2: Initial potential $U_c(\mathbf{z};\boldsymbol{\theta})$ assessed on the relevant $\mathbf{z}$ domain, where $\boldsymbol{\theta}_1 = 0$, implying a uniform $q_c(\mathbf{z}|\boldsymbol{\theta}_c)$. The color bar below the figure indicates the values of $U_c(\mathbf{z};\boldsymbol{\theta})$. Note that we set the minimum of $U_c(\mathbf{z};\boldsymbol{\theta})$ for comparability with other steps in the adaptive refinement to one.

log-posterior, which we estimate as

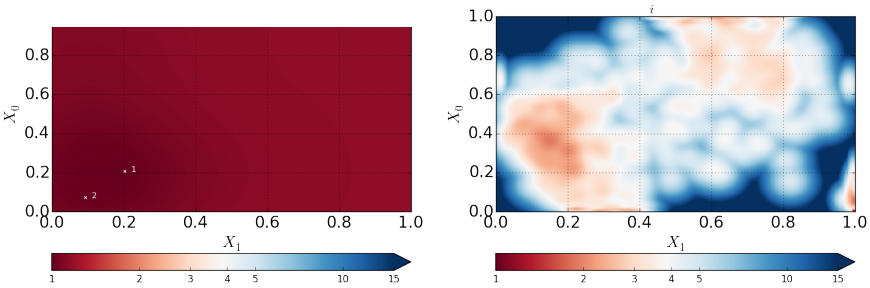
$$-\frac{1}{N}\sum_{i=1}^{N}\log r(\mathbf{z}|\mathbf{x}^{(i)}). \tag{4.42}$$

The latter expression is evaluated on a grid over the domain $\mathbf{z} \in [0,1]^{n_c}$. The refinement of $U_c(\mathbf{x})$ initially lags behind the already very characteristic posterior landscape visualized by Equation 4.42. Note that we wish to have the CG potential $U_c(\mathbf{z})$ close to the average log-posterior introduced in Equation 4.42. Adding features is continued in Figure 4.4. The anticipated benefit, expressed by $H(\boldsymbol{\lambda}_{L+1})$, of adding new features $\phi(\mathbf{z};\boldsymbol{\lambda}_{L+1})$ clearly drops with an increasing number of features, as depicted in Figure 4.5. This behavior makes sense and as we discussed earlier, one can employ $H(\boldsymbol{\lambda}_{L+1})$ as the stopping criterion for model refinement. With comparably low values of $H(\boldsymbol{\lambda}_{L+1})$, adding further features does not improve the model in terms of explaining the available data better. In this case, we could endeavor reducing noise in the gradient estimators by increasing the number of samples. Another avenue would involve improving the posterior inference, for example, by considering a Gaussian with full rank covariance or a flexible extension provided by amortized inference [407, 489].

In case of convergence of all model parameters and when the expected benefit of adding new features is fluctuating with low values compared to H at earlier stages, we are interested to see if any structure has been revealed in the latent space. The mean values of the approximate posterior distributions $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)},\boldsymbol{\phi}^{(i)})$ form clusters in the latent space according to the originating conformation (α, β-1, or β-2) of the atomistic coordinates $\boldsymbol{x}^{(i)}$. This reflects that the latent encoding is associated with

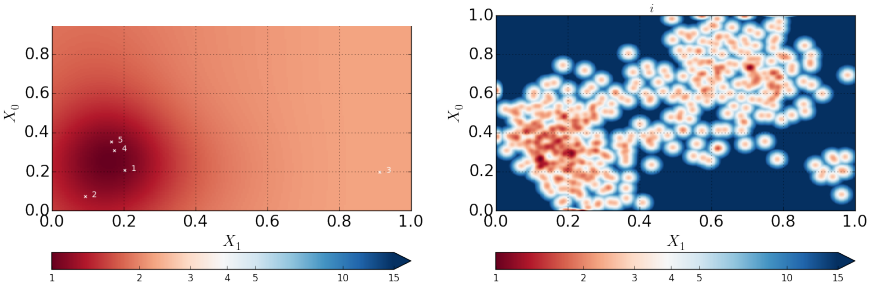


(a) $U_c(\mathbf{z};\boldsymbol{\theta})$ with $\dim(\boldsymbol{\theta}_c) = 1$.



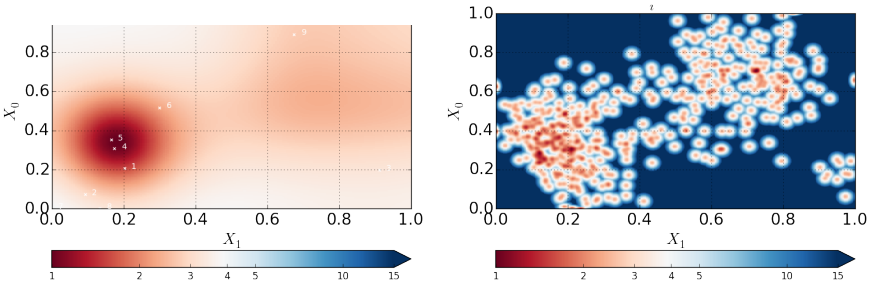
(c) $U_c(\mathbf{z};\boldsymbol{\theta})$ with $\dim(\boldsymbol{\theta}_c) = 2$.

(d) $-\frac{1}{N}\sum_{i=1}^{N} \log r(\mathbf{z}|\mathbf{x}^{(i)})$



(e) $U_c(\mathbf{z};\boldsymbol{\theta})$ with $\dim(\boldsymbol{\theta}_c) = 5$.

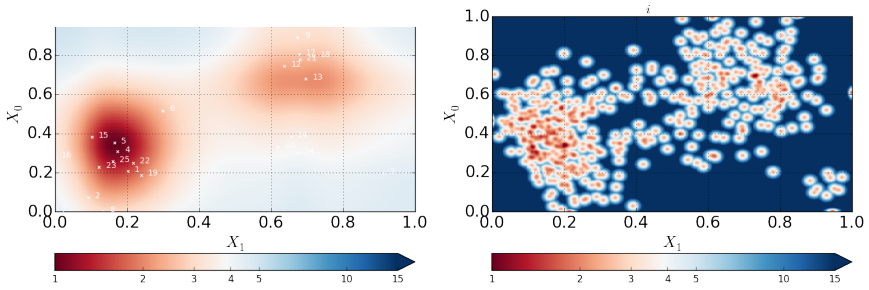
(f) $-\frac{1}{N}\sum_{i=1}^{N} \log r(\mathbf{z}|\mathbf{x}^{(i)})$



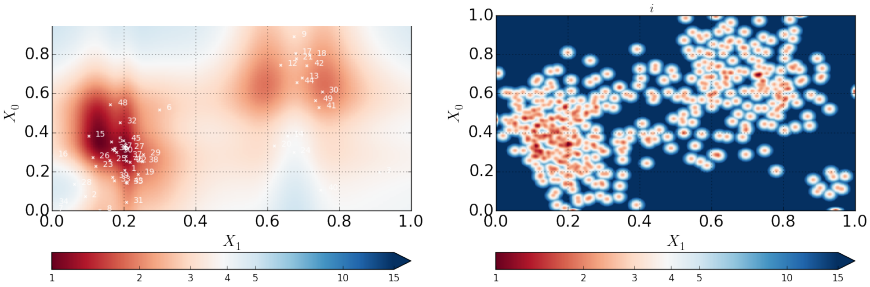
(g) $U_c(\mathbf{z};\boldsymbol{\theta})$ with $\dim(\boldsymbol{\theta}_c) = 10$.

(h) $-\frac{1}{N}\sum_{i=1}^{N} \log r(\mathbf{z}|\mathbf{x}^{(i)})$

FIGURE 4.3: Sequential adaptive refinement of $U_c(\mathbf{z};\boldsymbol{\theta})$. The left column depicts the potential energy $U_c(\mathbf{z};\boldsymbol{\theta})$ assessed in $\mathbf{z} \in [0,1]^{n_c}$. The potential $U_c(\mathbf{z};\boldsymbol{\theta})$ is composed of the indicated number of features. The right column represents the averaged aggregated log-posterior $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})$ assessed in $\mathbf{z} \in [0,1]^{n_c}$. The potential $U_c(\mathbf{z};\boldsymbol{\theta})$ should approach the averaged aggregated log-posterior $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})$. The first few alternating steps indicate that the posterior converges faster compared to how fast the model refines.

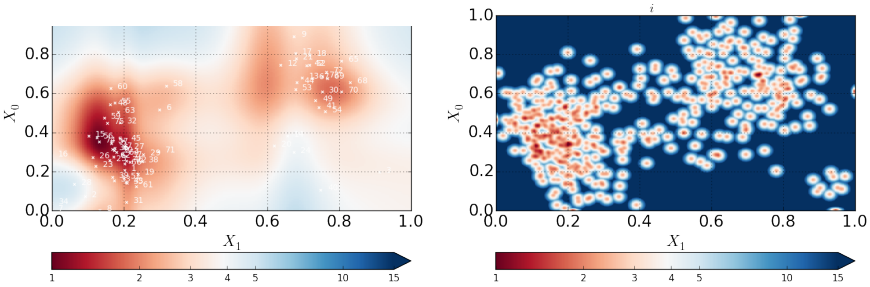


(a) $U_c(\mathbf{z};\boldsymbol{\theta})$ with $\dim(\boldsymbol{\theta}_c) = 25$.



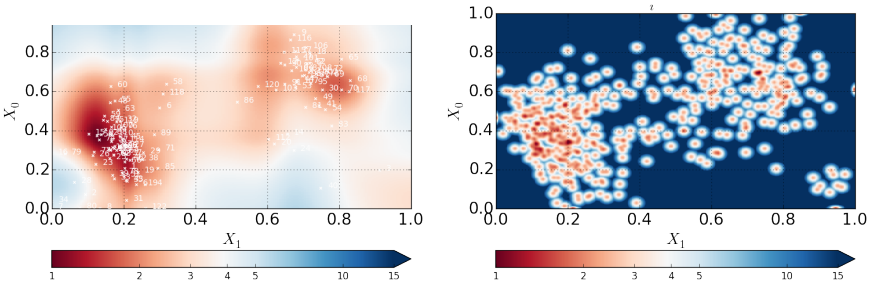
(c) $U_c(\mathbf{z};\boldsymbol{\theta})$ with $\dim(\boldsymbol{\theta}_c) = 50$.

(d) $-\frac{1}{N}\sum_{i=1}^{N} \log r(\mathbf{z}|\mathbf{x}^{(i)})$



(e) $U_c(\mathbf{z};\boldsymbol{\theta})$ with $\dim(\boldsymbol{\theta}_c) = 75$.

(f) $-\frac{1}{N}\sum_{i=1}^{N} \log r(\mathbf{z}|\mathbf{x}^{(i)})$



(g) $U_c(\mathbf{z};\boldsymbol{\theta})$ with $\dim(\boldsymbol{\theta}_c) = 122$.

(h) $-\frac{1}{N}\sum_{i=1}^{N} \log r(\mathbf{z}|\mathbf{x}^{(i)})$

FIGURE 4.4: Sequential adaptive refinement of $U_c(\mathbf{z};\boldsymbol{\theta})$. The left column depicts the potential energy $U_c(\mathbf{z};\boldsymbol{\theta})$ assessed in $\mathbf{z} \in [0,1]^{n_c}$. The potential $U_c(\mathbf{z};\boldsymbol{\theta})$ is composed of the indicated number of features. The right column represents the averaged aggregated log-posterior $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})$ assessed in $\mathbf{z} \in [0,1]^{n_c}$. The potential $U_c(\mathbf{z};\boldsymbol{\theta})$ should approach the averaged aggregated log-posterior $r(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\phi}^{(i)})$. The potential $U_c(\mathbf{z})$ is further refined and resembles more and more the averaged aggregated log-posterior.

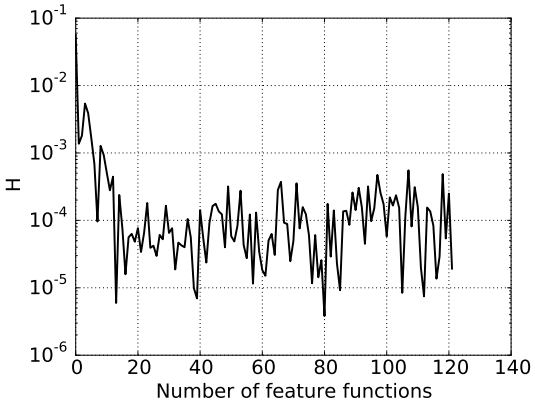


FIGURE 4.5: $H(\boldsymbol{\lambda}_{L+1})$ of Equation 4.41 for subsequently added features (horizontal axis) upon convergence of $\boldsymbol{\lambda}$, thus the maximum of H in terms of $\boldsymbol{\lambda}$.

the dihedrals (ϕ, ψ), since the latter distinguish the different conformations (Figure 4.6).

4.3.2 Predictive observable estimation

The overall goal of the proposed adaptive CG approach is to obtain a predictive model that is an efficient estimator of observables and to quantify epistemic uncertainties due to the limited amount of training data. Relevant observables for the ALA-2 peptide are statistics on the radius of gyration and the root-mean-square deviation from a reference α-helix configuration. Expressions for these observables are given in Appendix C.1.

We compare the maximum a posteriori (MAP) estimates with reference estimates based on 10 000 samples of a reference trajectory in Figs. 4.7 and 4.8. In addition to the depicted MAP prediction in these figures, we provide the 1–99% credible interval reflecting the epistemic uncertainty induced by relying on a limited amount of training data. By employing more evidence, expressed by larger data sets, the predictive model gains confidence and so the credible intervals shrink. The credible intervals are obtained as introduced in Equation 2.42 and Appendix A.1. In the present case, we propagate uncertainties expressed in the posterior of $\mathbf{W}$ based on variational Bayesian inference (Section 2.5.2) and the Laplace approximation (Section 2.5.4) to obtain an approximate posterior distribution on $\boldsymbol{\theta}_c$. All credible intervals shown are based on 4000 samples drawn from the joint approximate posterior $q(\mathbf{W}, \boldsymbol{\theta}_c | \mathbf{x}^{\mathcal{D}_N})$. For each sample $(\mathbf{W}^{(i)}, \boldsymbol{\theta}_c^{(i)})$, we use J samples from $q(\mathbf{x}|\mathbf{W}^{(i)}, \boldsymbol{\theta}_c^{(i)}, \boldsymbol{\theta}_{cf}^{\mathrm{MAP}})$ to estimate the observables associated with $(\mathbf{W}^{(i)}, \boldsymbol{\theta}_c^{(i)})$.

Predictions of the statistics as a function of the characteristic dihedral angles (ϕ, ψ) are shown in comparison to a reference estimate in Figure 4.9.

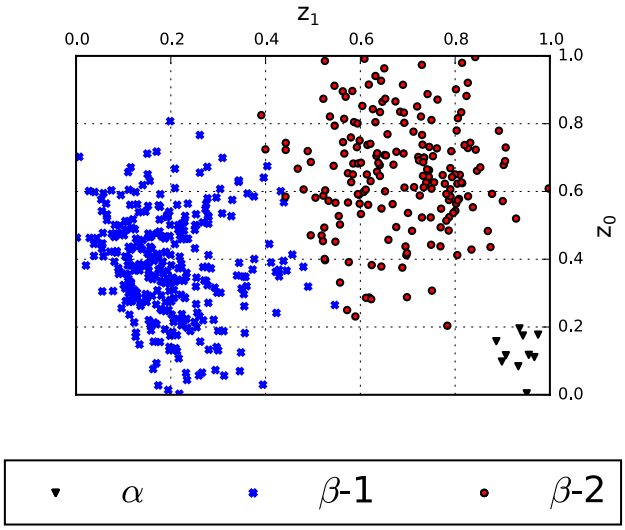
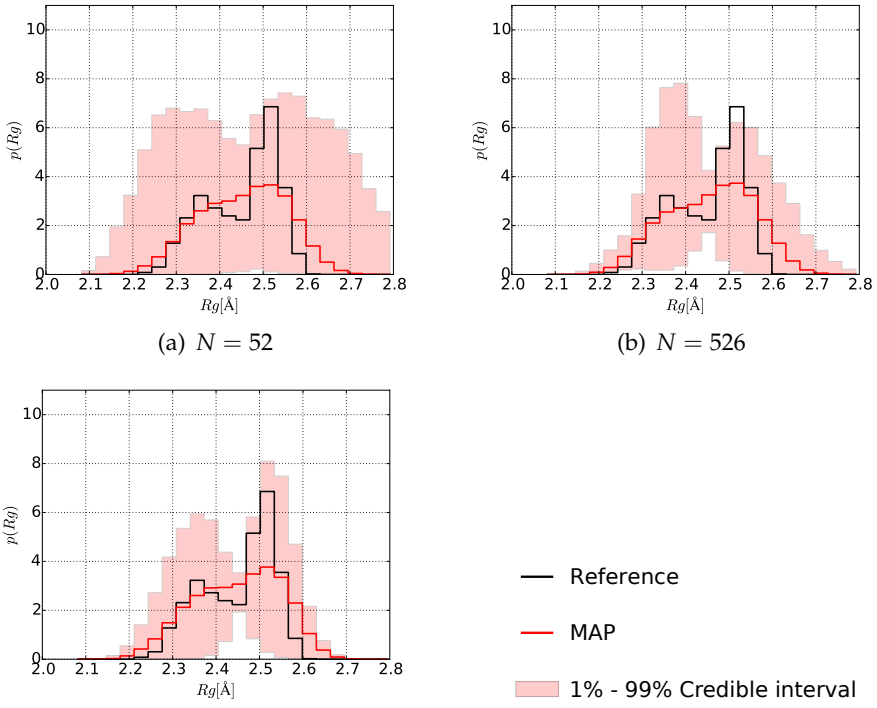


FIGURE 4.6: Mean values of the approximate posterior distributions $r(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\phi}^{(i)})$ given reference data $\boldsymbol{x}^{(i)}$. The latent representation of $\boldsymbol{x}^{(i)}$ employs markers according to their reference conformation, i.e., α, β-1, or β-2. Similar $\boldsymbol{x}^{(i)}$, in terms of being from the same conformational mode, map to a similar latent embedding and form clusters.

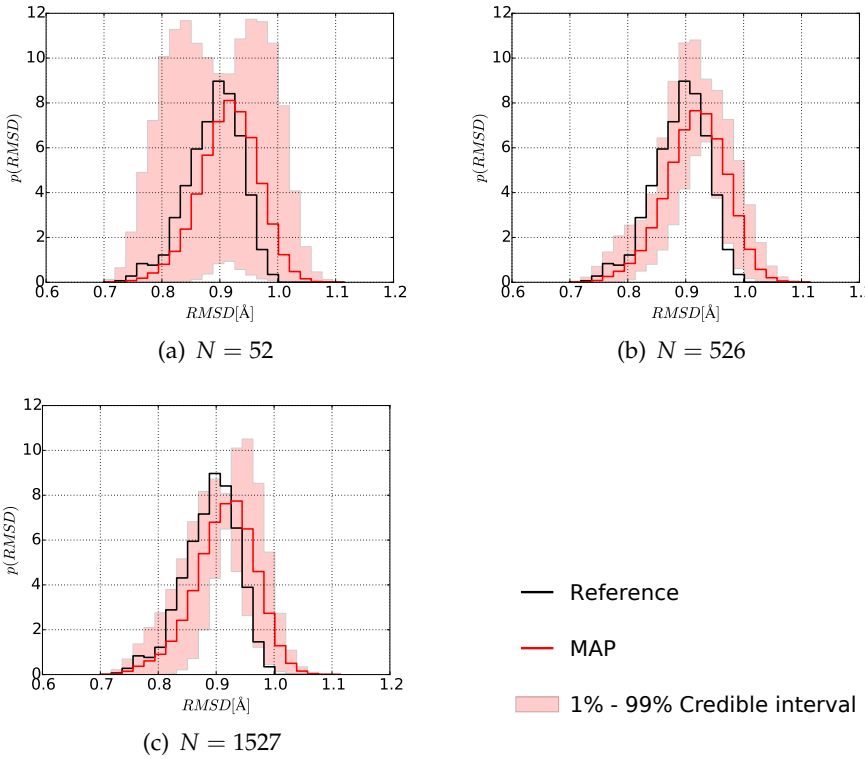


(a) $N = 52$

(b) $N = 526$

(c) $N = 1527$

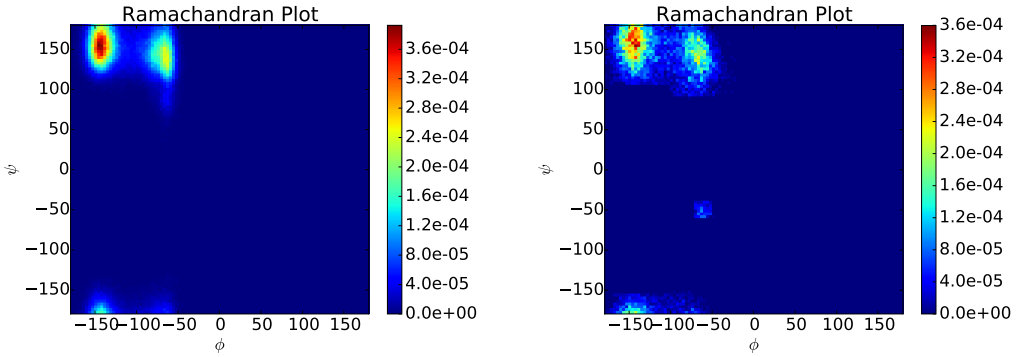
FIGURE 4.7: Predictive estimates of the radius of gyration for $\dim(\mathbf{z}) = 2$ for different sizes of data set N. The observable is defined in Appendix C.1. The MAP estimates (red lines) are compared to the reference estimates (black lines), which are based on a reference atomistic simulation with 10 000 samples. The shaded region around a MAP estimate depicts the 1–99% credible interval, which indicates the epistemic uncertainty due to the limited amount of training data. The credible intervals depicted were estimated using 4000 posterior samples of $q(\mathbf{W}, \boldsymbol{\theta}_c|\mathbf{x}^{D_N})$ based on Appendix A.1.



(a) $N = 52$ (b) $N = 526$

(c) $N = 1527$

FIGURE 4.8: Predictive estimates of the root-mean-square deviation for $\dim(\mathbf{z}) = 2$ for different size of data set N. The observable is defined in Appendix C.1. The MAP estimates (red lines) are compared to the reference estimates (black lines), which are based on a reference atomistic simulation with 10 000 samples. The shaded region around a MAP estimate depicts the 1–99% credible interval, which indicates the epistemic uncertainty due to the limited amount of training data. The credible intervals depicted were estimated using 4000 posterior samples of $q(\mathbf{W}, \boldsymbol{\theta}_c|\mathbf{x}^{\mathcal{D}_N})$ based on Appendix A.1.



(a) MAP prediction based on $N = 526$ and $\dim(\mathbf{z}) = 2$.

(b) Reference estimate based on 10 000 samples.

FIGURE 4.9: Predicted Ramachandran plot (left) compared to the reference estimate (right).

4.4 Summary and outlook

We have presented a Bayesian CG approach that both handles model selection tasks and provides a predictive CG framework.

In previous work, presented in Chapter 3, we identified the most relevant features in a rich set of basis functions. For applications differing from CG, in which we seek to obtain a lower-dimensional description, finding a rich set of basis functions for high-dimensional settings, e.g., $\dim(\mathbf{z}) = 100$, can be prohibitive [503]. Therefore, the proposed avenue is also beneficial in such settings since we initially learn a simple model. We drive the parameters to convergence and then add a wisely identified basis function that adds flexibility and improves the model's predictive capabilities. However, importantly, how do we know which new basis function to add? We develop an automated approach that searches among features $\phi(\mathbf{z};\boldsymbol{\lambda})$, parametrized by $\boldsymbol{\lambda}$, to add the one feature that provides the most considerable anticipated benefit. We developed an objective that reformulates the search among $\phi(\mathbf{z};\boldsymbol{\lambda})$ as an optimization problem concerning $\boldsymbol{\lambda}$. As an objective, we employ the gradient of the lower bound with respect to θ, associated with the feature to be added at $\theta = 0$. This gradient provides a metric for expressing the deviation between $q(\mathbf{z})$ and the aggregated and averaged posterior $r(\mathbf{z}|\mathbf{x})$. Maximizing the absolute value or squared value of this gradient pushes the new basis to regions where the discrepancy between $q(\mathbf{z})$ and the aggregated posterior $r(\mathbf{z}|\mathbf{x})$ is the largest. Thus we add the feature associated with the largest anticipated benefit.

In addition to this component, we proposed a Bayesian CG framework, and we developed a flexible coarse-to-fine mapping by employing a Gaussian encompassing a mean obtained by evaluating a linear model. The latter facilitates learning a coarse-to-fine mapping in the absence of any prior physical insight, which may not be available. We propose treating the matrix $\mathbf{W}$ by transforming $\mathbf{z}$ to $\mathbf{x}$ as a latent variable, and we make use of variational Bayesian inference to obtain the corresponding posterior distribution. Predictive observable estimates were augmented by credible intervals that reflect the uncertainty due to the limited amount of training data.

Next, our strategy is to extend the model hierarchically using a mixture of CG generators:

$$q(\mathbf{z}) = \prod_{i=1}^{I} \left(q(\mathbf{x}|\mathbf{z}_i) q_i(\mathbf{z}_i|\gamma)\right) q(\gamma)\, d\mathbf{z}\, d\gamma. \tag{4.43}$$

The proposed approach may enable local CG models in the context of each conformation. This formulation yield a posterior over γ, $q(\gamma|\mathbf{x}^{(i)})$, which is the responsibility for the corresponding mode for a given datum $\mathbf{x}^{(i)}$ and thus, gain additional physical insight.

Further, we are interested in amortized inference schemes, which provide an efficient probabilistic mapping from any $\mathbf{x}$ directly to $\mathbf{z}$. The flexible probabilistic mapping could support the discovery of latent collective variables (CVs), which can be

employed to drive the exploration of the configurational space. We suggest assuming a simple distribution of the latent variables accompanied by a complex probabilistic mapping. The latter enables the learning of complex transformations and thus, compress all relevant physics into an elementary distribution over the latent variable $\mathbf{z}$. The latter strategy has similarities with an Auto-Encoding Variational Bayes [407], which we explore in Chapter 5.

Chapter 5

Predictive collective variable discovery with deep Bayesian models

This chapter has been published in

M. Schöberl, N. Zabaras, P.-S. Koutsourelakis.
"Predictive collective variable discovery with deep Bayesian models".
In: *AIP The Journal of Chemical Physics* 150 (2019), 024109.

The following provides a summary of the scientific achievements of the above work and describes the individual contributions before closing this section.

5.1 Motivation and summary

The simulation of atomistic systems containing $M = 1 \times 10^5$ atoms for a time horizon of only 1×10^{-4} s, resolved by time steps of $\Delta t = 1 \times 10^{-15}$ s, requires a wall clock time of 1 year [504, 505]. An additional obstacle in atomistic systems results from the free-energy barriers between favorable atomistic configurations, which tremendously hamper the exploration of the full configurational space [506]. Accelerated exploration can be achieved by emplying enhanced sampling methods [507–514]. These rely on a lower-dimensional representation of the fully resolved atomistic description, which encodes system characteristics and changes in atomistic conformations (i.e., moving from one free-energy basin to another) [515]. Collective variables (CVs) depict such quantitative and lower-dimensional representations of relevant atomistic processes [507] and can be employed for guiding enhanced sampling methods [516]. However, the expected acceleration of enhanced sampling crucially depends on the quality of the obtained CVs. Inappropriate CVs result in enhanced sampling having even lower efficiency than brute force MD [517].

In [415], we developed a data-driven Bayesian framework leveraging the identification of CVs, providing physical insights without assuming any prerequisite

knowledge of the reference system. This methodology is suitable for use with limited atomistic reference data and thus copes well with noisy gradients. We employ $N = 50$[1] reference samples, whereas existing dimensionality reduction methods in the context of CV discovery are based on determining eigenvectors of transition matrices [508, 518–528], requiring densely sampled reference trajectories and comprehensive datasets encompassing more than 10 000 samples [529–531]. The developed methodology yields CVs that are differentiable with respect to fine-scale atomistic coordinates. Differentiability is important for seamlessly incorporating CVs into enhanced sampling schemes, e.g., for biasing the fine-scale potential $U_{\mathrm{f}}(\mathbf{x})$ and computing corresponding biasing forces. This work builds on findings of the publication summarized in Section 3.1 and considers CVs as latent generators yielding, through a probabilistic coarse-to-fine mapping, the full atomistic trajectory. This coarse-to-fine mapping can be interpreted as a decoder decrypting the latent CV to give the corresponding observed atomistic configurations. By training the generative model based on limited data, we seek to approximate the underlying complex reference density $p_{\mathrm{target}}(\mathbf{x})$. The complement to the generative decoder is the posterior distribution over the latent variables, given the observations. We call this the component encoder, as it translates atomistic observations to a parsimonious CV representation. This encoder is of paramount importance to revealing latent CVs with physical notion.

Beyond providing physical insights by identifying CVs, the framework enables the computation of probabilistic estimates of observables augmented by credible intervals. This is achieved though the predictive distribution. Credible intervals, as first introduced in the context of coarse-graining in Section 3.1, express the model's predictive confidence based on limited data. Identifying CVs and providing a predictive distribution are seamlessly addressed by deep Bayesian models. We utilize recent findings in deep learning research with an Auto-Encoding Variational Bayes modification [235, 407, 408] favoring sparse solutions.

We employ expressive deep neural networks for approximating the posterior distribution over latent variables ($q(\mathbf{z}|\mathbf{x})$) and likewise for the probabilistic mapping ($p(\mathbf{x}|\mathbf{z})$), while $p(\mathbf{z})$ remains simple, implying a simple representation of CVs. Expressivity and complexity are pushed to the probabilistic mappings[2] in combination with amortized inference [487, 488, 532]. Amortized inference enables drawing conclusions about latent CVs given any atomistic configuration input and thus facilitates the learning of a general encoder. An overview of the encoder and decoder is given in Figure 5.1.

We demonstrate the functionality of the proposed Bayesian CV discovery with deep learning by producing a parsimonious representation of alanine-dipeptide

[1] $N = 50$ is comparably low with respect to the effective dimensionality of the reference fine-scale system with $\dim(\mathbf{x}) = 60$.

[2] This is in contrast to the approach described in Chapter 3, where we explore an expressive distribution of latent variables $p(\mathbf{z})$ and a simple mapping $p(\mathbf{x}|\mathbf{z})$. The different parametrization strategies are emphasized in more detail in the closing discussion in Chapter 7

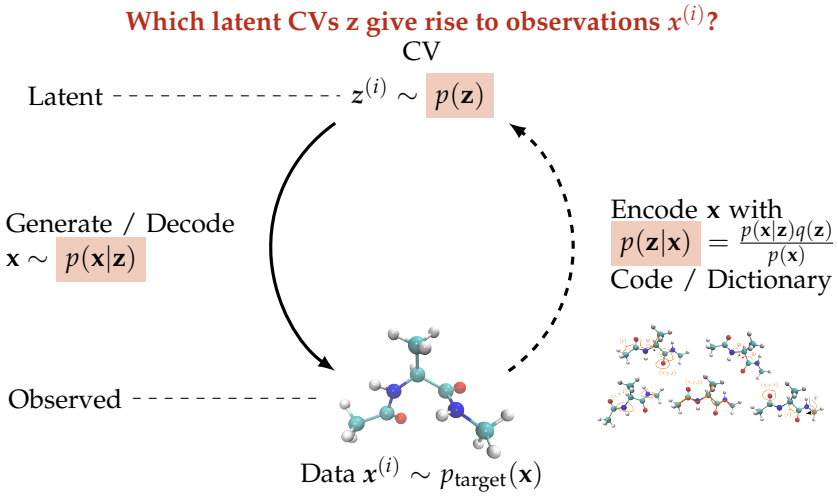
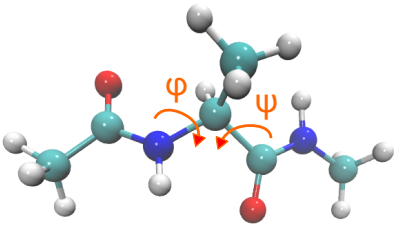


FIGURE 5.1: Conceptual sketch of predictive collective variable discovery with deep Bayesian models. For learning, a Bayesian deep model with reference data $x^{(i)}$ approximating the target Boltzmann distribution $p_{\text{target}}(\mathbf{x})$ is employed. Observed coordinates are encoded with the posterior $p(\mathbf{z}|\mathbf{x})$. The generative part allows, through ancestral sampling, the generation of new configurations based on a latent CV $z^{(i)} \sim p(\mathbf{z})$ and a decoder $p(\mathbf{x}|\mathbf{z})$. The encoder and decoder learn simultaneously via amortized variational Bayesian inference. Identifying CVs is thus reformulated to an stochastic optimization problem.

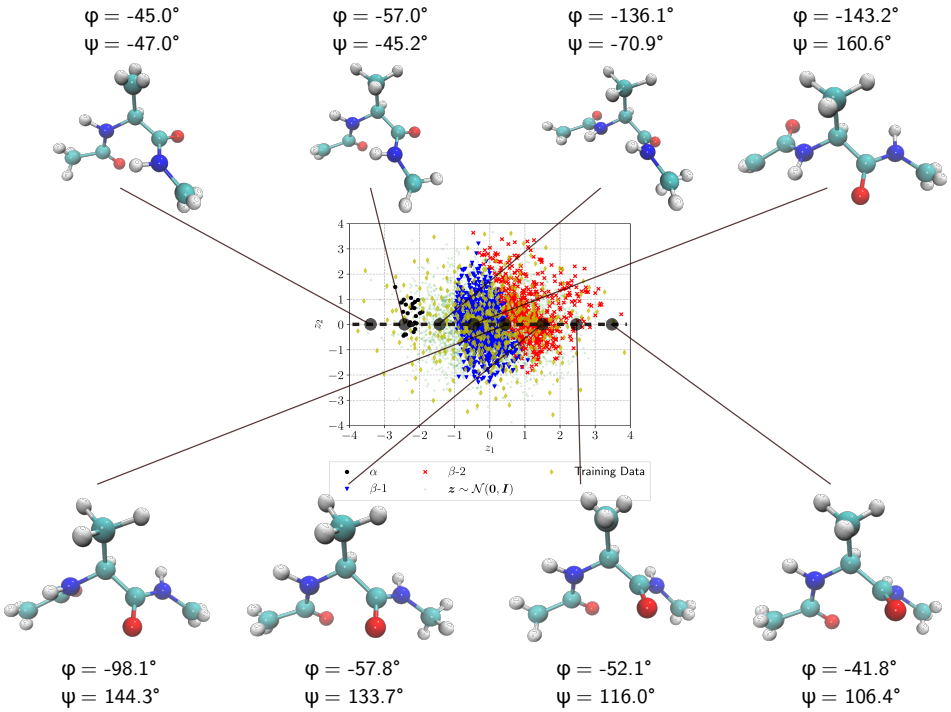
(ALA-2) based on as few as 50 training data points. We show (Figure 5.2(b)) that the identified CVs relate to the physically most compact description of ALA-2, that is, the dihedral angles[3]. Similar behavior has been demonstrated with a higher-dimensional alanine peptide consisting of 15 residues: ALA-15. In both examples, we provide probabilistic estimates of observables augmented by credible intervals.

We reproduce the original publication with permission of AIP Publishing in Appendix D.

[3]The dihedral angles of ALA-2 peptide (ϕ, ψ) are indicated in Figure 5.2(a).



(a)



(b)

FIGURE 5.2: Definition of dihedral angles and *predictions* given latent CVs (a) ALA-2 peptide with indicated dihedral angles. (b) *Predicted* configurations $\boldsymbol{x}^{(i)}$ with annotated dihedral angle values at $\boldsymbol{z}^{(i)}$ values are indicated by black filled dots. Moving along the z_1 axis, for the given CVs we obtain atomistic coordinates belonging to different conformations of ALA-2.

5.2 Declaration of the author's individual contribution

Conceptual ideas were generated and research and development of the presented CV discovery framework based on deep variational inference were performed individually. All reference simulations, the implementation of the developed framework, and construction of suitable neural networks were carried out on my own. All visualizations and illustrations with accompanying coding were produced without further assistance. My supervisors N. Zabaras and P.-S. Koutsourelakis provided valuable comments supporting the claimed statements in the numerical illustration section. The submitted manuscript has also been revised in collaboration with these supervisors.

Chapter 6

Embedded-physics machine learning for coarse-graining and collective variable discovery without data

Boltzmann densities, which are ensemble representations of equilibrium atomistic systems, are usually explored by molecular dynamics (MD) [11] or Monte Carlo-based (MC) techniques [502]. These versatile and general simulation techniques asymptotically guarantee unbiased estimates of observables [328]. However, these simulation techniques become computationally impractical in cases where the atomistic interaction potential exhibits several distinct minima or wells. Such complex potentials imply multimodal Boltzmann densities. Escaping such a well is rare and requires overcoming high free-energy barriers, resulting in impractically long simulation times or biased trajectories [506].

Key to exploring such multimodal Boltzmann densities is the recognition of appropriate slow coordinates or collective variables (CVs) that exhibit sensitivity in transition regions between modes. This requires tremendous physicochemical insight, which is not available per se. CVs, which provide an effective lower-dimensional description of high-dimensional atomistic systems, are key to accelerating the exploration of multimodal densities by biasing the dynamics to escape deep free-energy wells [516].

Identifying expressive CVs governing major conformational changes in the absence of physical insight requires data-driven strategies. However, in many cases, the identification of CVs requires a dense sample of the target Boltzmann distribution, as well as unbiased simulation trajectories. This creates a contradiction, as it is computationally impractical to obtain unbiased trajectories in the presence of multiple modes [190], restricting the potential use-cases of such approaches. Few nonlinear dimensionality reduction methods coping with low data provide CVs that are not differentiable with respect to their atomistic counterparts [529–531]; however, this is required for biasing the dynamics [508, 518–528]. Deep learning approaches

providing flexible and efficiently differentiable functions have also influenced research on the efficient exploration of multimodal Boltzmann distributions. However, these build on previously acquired reference data and do not account directly for the interaction potential that actually drives the MD or MC simulation.

This work provides a novel and fundamentally different perspective on data-driven deep learning approaches. Instead of relying on two separate process, acquiring data and then employing statistical learning of a model, we synthesize and embed physics, i.e., the Boltzmann density, with a machine learning objective. The advocated learning methodology proposes (atomistic) configurations to which the model is attracted to learn from the potential energy and associated interatomic forces. The proposed machine learning algorithm does not require any simulation of the Boltzmann density but only queries the physical model, i.e., the potential and forces, to gain relevant information by evaluating rather than simulating the Boltzmann density. The proposed learning algorithm includes a versatile nonlinear dimensionality-reduction routine, which simultaneously discovers relevant CVs while learning the Boltzmann density. We demonstrate the procedure using a double well potential and the alanine dipeptide.

The present work differs clearly from recent developments on Boltzmann generators [240] that rely on invertible neural networks such as RealNVP [437] and NICE [238]. As it employs invertible neural networks, the dimensionality of the latent generator must equal the dimensionality of the atomistic configuration, which detains a consistent dimensionality reduction. Generated atomistic realizations of the employed model in [240] do not reflect the statistics of the reference Boltzmann distribution and serve instead as an input to a subsequent re-weighting importance sampling step. However, importance sampling is difficult to monitor if the variance in the importance weights is low, implying a large effective sample size, when none of the proposed realizations yield relatively high probabilities as evaluated by the target density [442]. Furthermore, a good guess of CVs is provided in Boltzmann generators, which depict physical insights that may not be available. By contrast, the proposed approach (similar double well example) reveals the effective CVs *and* also provides a generator to produce samples that yield the correct statistics of the target.

In the following Section 6.1, we develop the proposed learning approach based on KL divergence minimization and derive a tractable upper bound based on hierarchical variational models [533]. We discuss the required gradient computation and provide a physically interpretable underpinning of the components involved. After introducing a general model parametrization, we provide an adaptive tempering scheme facilitating a robust machine learning procedure at the end of Section 6.1. The proposed physics-embedding learning procedure for revealing CVs and obtaining a coarse-grained (CG) model is numerically validated in Section 6.2 with a double well potential and the alanine dipeptide. We close this paper in Section 6.3, summarizing the main findings of this work and outlining simple but effective

further extensions and research directions. These include the generalization of the obtained predictive distribution for predictive purposes at any temperature.

6.1 Methodology

After introducing the notation in Section 6.1.1, we describe the general proposed framework in Section 6.1.2. A tractable optimization objective is provided in Section 6.1.3. We compare the proposed approach with data-driven objectives in Section 6.1.4. Relevant model specifications and gradient computations for training are discussed in Section 6.1.5, and we close with some notes on the actual training procedure in Section 6.1.6.

6.1.1 Equilibrium statistical mechanics

In equilibrium statistical mechanics, we seek to estimate ensemble averages of observables $a(\mathbf{x})$ with respect to the Boltzmann density,

$$\langle a \rangle_{p_{\text{target}}(\mathbf{x};\beta)} = \int_{\mathcal{M}_{\text{f}}} a(\mathbf{x}) p_{\text{target}}(\mathbf{x};\beta)\, d\mathbf{x}. \tag{6.1}$$

We denote the Boltzmann distribution, for which we aim to learn an efficient approximation, by $p_{\text{target}}(\mathbf{x};\beta)$:

$$\begin{aligned} p_{\text{target}}(\mathbf{x}) &= \frac{1}{Z(\beta)} \underbrace{e^{-\beta U(\mathbf{x})}}_{\pi(\mathbf{x};\beta)} \\ &= \frac{\pi(\mathbf{x};\beta)}{Z(\beta)}. \end{aligned} \tag{6.2}$$

In Equation 6.2, $Z(\beta) = \int e^{-\beta U(\mathbf{x})} d\mathbf{x}$ is the partition function or normalization constant, and $\beta = \frac{1}{k_B T}$ is the reciprocal or inverse temperature with the Boltzmann constant k_B and the temperature T. The interatomic potential $U(\mathbf{x})$ depends on generalized atomistic coordinates denoted by $\mathbf{x} \in \mathcal{M}_{\text{f}} \subset \mathbb{R}^{n_{\text{f}}}$, with $n_{\text{f}} = \dim(\mathbf{x})$. In equilibrium statistical mechanics, we are usually interested in phase averages at distinct constant temperatures; however, we will also demonstrate how to utilize the temperature to introduce an auxiliary sequence of target distributions to facilitate learning the actual target distribution. The auxiliary sequence stabilizes the parameter learning inspired by annealing [534, 535] and adaptive sequential MC [205].

6.1.2 Coarse-graining through probabilistic generative models

Data-driven coarse-graining methodologies are based on a limited set of N realizations obtained from the target density $p_{\text{target}}(\mathbf{x})$. The realizations $\boldsymbol{x}^{(i)}$ are produced by drawing samples from $p_{\text{target}}(\mathbf{x})$: $\boldsymbol{x}^{(i)} \sim p_{\text{target}}(\mathbf{x})$ with Markov Chain MC (MCMC) methods [306, 450] and/or, especially in the context of biochemical

atomistic systems, by MD simulations [14, 88]. Both methodologies yield a dataset $\mathbf{x}^{\mathcal{D}_N} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$, which approximates the target distribution with

$$\begin{aligned} p_{\text{target}}(\mathbf{x}) &\approx \tilde{p}(\mathbf{x}) \\ &\propto \prod_{i=1}^{N} \delta(\boldsymbol{x}^{(i)} - \mathbf{x}). \end{aligned} \tag{6.3}$$

The above approximation, given independent and identically distributed samples, may sufficiently resemble simple systems. However, atomistic many-body systems exhibit higher-order and long-range interactions [48, 97] involving multiple free energy modes separated by high barriers [106, 131]. Therefore, the collection of sufficient data becomes an insurmountable task: a protein folding process may take microseconds versus a time discretization of femtoseconds [536]. Given limited computational power, the relevant conformations and transitions are not guaranteed to be reflected by the reference simulation [537].

The quality of data-driven learning approaches depends strongly on the quality of the available set of reference data $\mathbf{x}^{\mathcal{D}_N}$. If, e.g., in the case of peptides, certain conformations are missed, it is an almost insurmountable challenge to obtain a data-driven model exploring such missed configurations [124, 538]. Enhanced sampling methods [508, 518–528] can support the exploration of the configuration space, while the efficiency crucially depends on the quality of utilized CVs [263, 517].

Instead of relying on reference data, which may be a distorted representation of $p_{\text{target}}(\mathbf{x})$, or gradually exploring the configuration space by enhanced sampling, we present a variational approach that learns the target distribution $p_{\text{target}}(\mathbf{x})$ by querying the unnormalized distribution $\pi(\mathbf{x})$ or the corresponding potential energy $U(\mathbf{x})$ (see Equation 6.2).

We are first interested in identifying latent CVs $\mathbf{z}$ ($\mathbf{z} \in \mathcal{M}_c \subset \mathbb{R}^{n_c}$) depending on the fully atomistic picture $\mathbf{x}$, which encode physically relevant characteristics (e.g., coordinates along transition regions between conformations) and provide insight into the unknown atomistic system we seek to explore. Second, we seek to identify a CG model expressed in terms of the latent CVs $\mathbf{z}$ that is predictive but nevertheless facilitates reasoning about all-atom coordinates $\mathbf{x}$ [375]. The obtained CG model is expected to serve as an approximation to $p_{\text{target}}(\mathbf{x})$ to enable the efficient computation of the expectations of observables (Equation 6.1) and most importantly to capture relevant configurations in the free energy landscape that were inaccessible by brute-force MD or MCMC approaches [539]. Fulfilling the latter requirement also implies capturing statistics of $p_{\text{target}}(\mathbf{x})$.

The CVs $\mathbf{z}$ serve as latent generators of the higher-dimensional generalized coordinates $\mathbf{x}$, where we seek $\dim(\mathbf{z}) \ll \dim(\mathbf{x})$. This generative process is expressed with two components,

(i) the conditional density $q_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{x}|\mathbf{z})$, parametrized by $\boldsymbol{\theta}_{\text{cf}}$,

(ii) and the density over the latent CVs $q_{\boldsymbol{\theta}_c}(\mathbf{z})$.

Combining both densities gives the following joint:

$$q_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = q_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{x}|\mathbf{z}) q_{\boldsymbol{\theta}_{\text{c}}}(\mathbf{z}). \tag{6.4}$$

Assuming we have obtained the optimal parameters $\boldsymbol{\theta}^{\text{opt}}$ after a training process based on an objective, which we will discuss later in this section, we can utilize the model for predictive purposes. This can be done by ancestral sampling [442], i.e., first draw $\boldsymbol{z}^{(i)} \sim q_{\boldsymbol{\theta}^{\text{opt}}}(\mathbf{z})$ and second $\boldsymbol{x}^{(i)} \sim q_{\boldsymbol{\theta}^{\text{opt}}}(\mathbf{x}|\boldsymbol{z}^{(i)})$.

For obtaining optimal parameters $\boldsymbol{\theta}$, many methods rely on minimizing a distance from the target distribution $p_{\text{target}}(\mathbf{x})$ to the marginal distribution $q_{\boldsymbol{\theta}}(\mathbf{x})$, which is given by:

$$q_{\boldsymbol{\theta}}(\mathbf{x}) = \int q_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \, d\mathbf{z} = \int q_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{x}|\mathbf{z}) q_{\boldsymbol{\theta}_{\text{c}}}(\mathbf{z}) \, d\mathbf{z}. \tag{6.5}$$

A commonly employed metric expressing the deviation between two densities is the Kullback–Leibler (KL) divergence, which belongs to the family of α-divergences [376–378]:

$$\begin{aligned} D_{\text{KL}}\left(p_{\text{target}}(\mathbf{x}) \| q_{\boldsymbol{\theta}}(\mathbf{x})\right) &= -\int p_{\text{target}}(\mathbf{x}) \log \frac{q_{\boldsymbol{\theta}}(\mathbf{x})}{p_{\text{target}}(\mathbf{x})} \, d\mathbf{x} \\ &= -\left\langle \log q_{\boldsymbol{\theta}}(\mathbf{x}) \right\rangle_{p_{\text{target}}(\mathbf{x})} + \underbrace{\left\langle \log p_{\text{target}}(\mathbf{x}) \right\rangle_{p_{\text{target}}(\mathbf{x})}}_{-\mathbb{H}(p_{\text{target}})}. \end{aligned} \tag{6.6}$$

Minimizing Equation 6.6 with respect to $\boldsymbol{\theta}$ leads to $q_{\boldsymbol{\theta}}(\mathbf{x})$ being closer to $p_{\text{target}}(\mathbf{x})$. However, in practice, the expectations in Equation 6.6 are intractable:

(i) the marginal $q_{\boldsymbol{\theta}}(\mathbf{x})$ requires the integration with respect to $\mathbf{z}$ which is intractable itself and

(ii) the involved expectation with respect to $p_{\text{target}}(\mathbf{x})$, $\langle \cdot \rangle_{p_{\text{target}}(\mathbf{x})}$ is analytically intractable since the normalization constant of $p_{\text{target}}(\mathbf{x})$ is unavailable (which would require solving an integral with respect to $\mathbf{x}$).

Considering the above challenges, the latter could be addressed by approximating $p_{\text{target}}(\mathbf{x})$ with data or samples $\mathbf{x}^{\mathcal{D}_N}$ and thus approximating the corresponding expectations with MC estimators. However, as we deal with complex multimodal Boltzmann densities $p_{\text{target}}(\mathbf{x})$, the data generating process (MCMC or MD) may miss relevant modes. By employing a biased set of samples or data not approximating $p_{\text{target}}(\mathbf{x})$ [295], we learn a biased estimator not approximating $p_{\text{target}}(\mathbf{x})$. The generation of the training dataset is thus decoupled from the learning process.

To circumvent the data-generating process and thus sampling from $p_{\text{target}}(\mathbf{x})$, we propose employing the other extreme of the family of α-divergences (as compared

to Equation 6.6), the *reverse* KL divergence:

$$\begin{aligned} D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\mathbf{x})\|p_{\mathrm{target}}(\mathbf{x})\right) &= -\int q_{\boldsymbol{\theta}}(\mathbf{x}) \log \frac{p_{\mathrm{target}}(\mathbf{x})}{q_{\boldsymbol{\theta}}(\mathbf{x})}\, d\mathbf{x} \\ &= \underbrace{\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x})}\left[\log q_{\boldsymbol{\theta}}(\mathbf{x})\right]}_{-\mathbb{H}(q(\mathbf{x}))} - \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x})}\left[\log p_{\mathrm{target}}(\mathbf{x})\right] \\ &= -\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x})}\left[\log p_{\mathrm{target}}(\mathbf{x})\right] - \mathbb{H}(q(\mathbf{x})). \end{aligned} \tag{6.7}$$

Minimizing Equation 6.7 with respect to $\boldsymbol{\theta}$ requires maximizing the log-likelihood $\log p_{\mathrm{target}}(\mathbf{x})$ assessed under $q_{\boldsymbol{\theta}}(\mathbf{x})$ (first component in Equation 6.7), and the maximization of the entropy of $q_{\boldsymbol{\theta}}(\mathbf{x})$, $\mathbb{H}(q(\mathbf{x}))$ (second component in Equation 6.7). Minimizing the reverse KL divergence balances the two terms, as maximizing only the log-likelihood $\log p_{\mathrm{target}}(\mathbf{x})$ assessed under $q_{\boldsymbol{\theta}}(\mathbf{x})$ would result in a degenerate case where $q_{\boldsymbol{\theta}}(\mathbf{x})$ would become a Dirac-delta placed at the (global) maximum of $p_{\mathrm{target}}(\mathbf{x})$ obtained at the (global) minimum of $U(\mathbf{x})$. The second component implies a regularization favoring a parametrization $\boldsymbol{\theta}$ such that the entropy of $q_{\boldsymbol{\theta}}(\mathbf{x})$ is maximized.

6.1.3 Inference and learning

In what follows, we use the negative of the KL divergence in Equation 6.7 to be maximized, which we denote with $\mathcal{L}$ for the sake of comparability with other learning approaches [407, 436]. At the end of this section, we compare the presented methodology with data-driven approaches relying on the forward KL divergence [406, 407, 409] and especially those addressing coarse-graining problems [415, 540–542].

The objective to be maxmized is

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x})}\left[\log p_{\mathrm{target}}(\mathbf{x}) - \log q_{\boldsymbol{\theta}}(\mathbf{x})\right], \tag{6.8}$$

where we can draw samples from $q_{\boldsymbol{\theta}}(\mathbf{x})$ as we can wisely select tractable hierarchical components composing to $q_{\boldsymbol{\theta}}(\mathbf{x})$. The optimization of the first component in $\mathcal{L}(\boldsymbol{\theta})$ relating to the log-likelihood is tractable as the normalization of $p_{\mathrm{target}}(\mathbf{x})$ does not depend on the parameters $\boldsymbol{\theta}$ and thus being able to evaluate $\pi(\mathbf{x})$ or $U(\mathbf{x})$ suffices. However, the entropy term is not tractable ad-hoc as it involves the marginal $q_{\boldsymbol{\theta}}(\mathbf{x}) = \int q_{\boldsymbol{\theta}_{\mathrm{cf}}}(\mathbf{x}|\mathbf{z}) q_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{z})\, d\mathbf{z}$, posing in most cases an intractable or least cumbersome task.

Therefore, we seek to construct a tractable lower bound on $\mathbb{H}(\mathbf{x})$ as presented in [533] by introducing an auxiliary density $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ parametrized by $\boldsymbol{\phi}$ and write:

$$\begin{aligned} -\mathbb{E}_{q(\mathbf{x})}\left[\log q(\mathbf{x})\right] &= -\mathbb{E}_{q(\mathbf{x})}\left[\log q(\mathbf{x}) + \underbrace{D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{x})\|q(\mathbf{z}|\mathbf{x})\right)}_{=0}\right] \\ &\geq -\mathbb{E}_{q(\mathbf{x})}\left[\log q(\mathbf{x}) + D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{x})\|r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right)\right] \\ &= -\mathbb{E}_{q(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log q(\mathbf{x}) + \log q(\mathbf{z}|\mathbf{x}) - \log r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]\right]. \end{aligned} \tag{6.9}$$

Adding $D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{x})\|q(\mathbf{z}|\mathbf{x})\right)$ in the first line of Equation 6.9 has no influence as the term is equal to zero. It involves the posterior distribution over the latent variables $\mathbf{z}$, $q(\mathbf{z}|\mathbf{x}) = \frac{q(\mathbf{x},\mathbf{z})}{q(\mathbf{x})}$, which is intractable. By utilizing an auxiliary distribution $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, the equality becomes an inequality as a consequence of $D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{x})\|r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right) \geq 0$ for $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ deviating from $q(\mathbf{z}|\mathbf{x})$. Replacing the exact log-posterior $\log q(\mathbf{z}|\mathbf{x})$ by

$$\log q(\mathbf{z}|\mathbf{x}) = \log q(\mathbf{z}) + \log q(\mathbf{x}|\mathbf{z}) - \log q(\mathbf{x}), \tag{6.10}$$

it follows that

$$\begin{aligned} -\mathbb{E}_{q(\mathbf{x})}\left[\log q(\mathbf{x})\right] \geq & -\mathbb{E}_{q(\mathbf{x})}\Big[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\big[\log q(\mathbf{x}) + \log q(\mathbf{z}) + \log q(\mathbf{x}|\mathbf{z}) \\ & - \log q(\mathbf{x}) - \log r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\big]\Big] \\ = & -\mathbb{E}_{q(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log q(\mathbf{z}) + \log q(\mathbf{x}|\mathbf{z}) - \log r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]\right]. \end{aligned} \tag{6.11}$$

Rewriting the expectation $\mathbb{E}_{q(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\cdot\right]\right]$ as $\mathbb{E}_{q(\mathbf{x},\mathbf{z})}\left[\cdot\right]$, Equation 6.11 depicts a tractable lower bound on the entropy term. Maximizing the lower bound in Equation 6.11 with respect to $\boldsymbol{\phi}$ minimizes $D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{x})\|r(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})\right)$ and thus tightens the bound on the entropy term. As mentioned earlier, the optimum[1] is obtained when we identify the exact posterior of the latent CVs $q(\mathbf{z}|\mathbf{x})$ with $r(\mathbf{z}|\mathbf{x};\boldsymbol{\phi}^{\mathrm{opt}}) = q(\mathbf{z}|\mathbf{x})$, thus $D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{x})\|r(\mathbf{z}|\mathbf{x};\boldsymbol{\phi}^{\mathrm{opt}})\right) = 0$. Utilizing the obtained bound in Equation 6.11, the objective $\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta})$ from Equation 6.8 becomes:

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}\left[\log p_{\mathrm{target}}(\mathbf{x}) - \log q_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{z}) - \log q_{\boldsymbol{\theta}_{\mathrm{cf}}}(\mathbf{x}|\mathbf{z}) + \log r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]. \tag{6.12}$$

The following shows the connection between the obtained objective and the KL divergence defined between the joint $q(\mathbf{x}|\mathbf{z})q(\mathbf{z})$ and $p_{\mathrm{target}}(\mathbf{x})r(\mathbf{z}|\mathbf{x})$ acting on the extended probability space:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}\left[\log p_{\mathrm{target}}(\mathbf{x}) - \log q(\mathbf{z}) - \log q(\mathbf{x}|\mathbf{z}) + \log r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \\ &= \mathbb{E}_{q(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}\left[\log \frac{p_{\mathrm{target}}(\mathbf{x})r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{q_{\boldsymbol{\theta}_{\mathrm{cf}}}(\mathbf{x}|\mathbf{z})q_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{z})}\right] \qquad (6.13) \\ &= -D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}_{\mathrm{cf}}}(\mathbf{x}|\mathbf{z})q_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{z})\|p_{\mathrm{target}}(\mathbf{x})r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right). \qquad (6.14) \end{aligned}$$

[1] The optimum with respect to $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and thus $\boldsymbol{\phi}$ that tightens the lower bound.

Based on Equation 6.13 and Equation 6.14, we show how the objective separates into two KL divergence terms:

$$\begin{aligned} D_{\mathrm{KL}}\left(q(\mathbf{x}|\mathbf{z})q(\mathbf{z})\|p_{\mathrm{target}}(\mathbf{x})r(\mathbf{z}|\mathbf{x})\right) &= -\mathbb{E}_{q(\mathbf{z})}\left[\mathbb{E}_{q(\mathbf{x}|\mathbf{z})}\left[\log\frac{p_{\mathrm{target}}(\mathbf{x})r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{x})q(\mathbf{x})}\right]\right] \\ &= -\mathbb{E}_{q(\mathbf{x})}\left[\log\frac{p_{\mathrm{target}}(\mathbf{x})}{q(\mathbf{x})}\right] - \mathbb{E}_{q(\mathbf{z})q(\mathbf{x}|\mathbf{z})}\left[\log\frac{r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{x})}\right] \\ &= D_{\mathrm{KL}}\left(q(\mathbf{x})\|p_{\mathrm{target}}(\mathbf{x})\right) + D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{x})\|r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right) \\ &\geq D_{\mathrm{KL}}\left(q(\mathbf{x})\|p_{\mathrm{target}}(\mathbf{x})\right) \end{aligned} \tag{6.15}$$

As mentioned earlier, the lower bound on $\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta})$ or upper bound on $D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}_{\mathrm{cf}}}(\mathbf{x}|\mathbf{z})q_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{z})\|p_{\mathrm{target}}(\mathbf{x})r(\mathbf{z}|\mathbf{x})\right)$ becomes tight when $r(\mathbf{z}|\mathbf{x};\boldsymbol{\phi}^{\mathrm{opt}}) = q(\mathbf{z}|\mathbf{x})$, which is Equation 6.15. Suboptimal $\boldsymbol{\phi}$ imply bounds on the objective owing to the positivity of $D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{x})\|r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right) \geq 0$.

The advantage of the proposed method for identifying CVs and learning a predictive coarse-graining model becomes clearer when we directly utilize the reference potential energy $U(\mathbf{x})$ (which we assume to be available in this paper). The objective $\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta})$, which is the negative KL divergence defined by the joint distributions, is subject to maximization with respect to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta}) &= -D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{z})q_{\boldsymbol{\theta}_{\mathrm{cf}}}(\mathbf{x}|\mathbf{z})\|p_{\mathrm{target}}(\mathbf{x})r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right) \\ &= \left\langle\log p_{\mathrm{target}}(\mathbf{x})\right\rangle_{q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})} + \left\langle\log\frac{r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{q_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{z})q_{\boldsymbol{\theta}_{\mathrm{cf}}}(\mathbf{x}|\mathbf{z})}\right\rangle_{q_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x})} \\ &= -\beta\left\langle U(\mathbf{x})\right\rangle_{q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})} + \left\langle\log\frac{r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{q_{\boldsymbol{\theta}_{\mathrm{c}}}(\mathbf{z})q_{\boldsymbol{\theta}_{\mathrm{cf}}}(\mathbf{x}|\mathbf{z})}\right\rangle_{q_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{x})} \end{aligned} \tag{6.16}$$

Maximizing Equation 6.16 solely involves expectations with respect to the generative model, from which it is easy to draw samples from. Explicitly there are no expectations with respect to the target density $p_{\mathrm{target}}(\mathbf{x})$, which would require an approximation with data. Instead of data, the target density $p_{\mathrm{target}}(\mathbf{x})$ contributes to the learning of the parameters $(\boldsymbol{\phi},\boldsymbol{\theta})$ through the interatomic potential energy $U(\mathbf{x})$ assessed for samples of the generative model $q(\mathbf{x}|\mathbf{z})q(\mathbf{z})$. Note that the normalization constant of $p_{\mathrm{target}}(\mathbf{x})$ is independent of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ and has been omitted in Equation 6.16. We are aware that the method requires a potential energy function $U(\mathbf{x})$, which can be assessed at $\mathbf{x}$. This is always the case for systems where we can set up MD or MCMC simulations, although we do circumvent the need to simulate a trajectory or draw reference samples by directly incorporating the available physics expressed by the potential energy.

6.1.4 Reverse or forward KL divergence?

In the following, we point out commonalities and differences between the proposed approach relying on the reverse KL divergence as introduced in Equation 6.7 and the

forward KL divergence (Equation 6.6). The latter has been successfully employed for the development of coarse-graining methodologies [540–542] and with a focus on CV discovery in combination with predictive coarse-graining in [415].

The data-driven objective is based on minimizing the following KL divergence:

$$D_{\mathrm{KL}}\left(p_{\mathrm{target}}(\mathbf{x})\|q_{\boldsymbol{\theta}}(\mathbf{x})\right). \tag{6.17}$$

Reformulating the minimization of Equation 6.17 to a maximization problem, the lower bound based on the summation over terms corresponding to each datum $\boldsymbol{x}^{(i)}$ of a set of $\mathbf{x}^{\mathcal{D}_N} = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{N}$ is written as:

$$\begin{aligned}\mathcal{L}^{\mathrm{forward}}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{\mathcal{D}_N}) &= \sum_{i=1}^{N}\mathbb{E}_{r_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)})}\left[-\log r_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)})+\log q_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)},\mathbf{z}^{(i)})\right]\\ &= -\sum_{i=1}^{N}D_{\mathrm{KL}}\left(r_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)})\|q_{\boldsymbol{\theta}}(\mathbf{z}^{(i)})\right)\\ &\quad+\sum_{i=1}^{N}\mathbb{E}_{r_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)})}\left[\log q_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\mathbf{z}^{(i)})\right].\end{aligned} \tag{6.18}$$

The objective above depicts the lower bound on the marginal log-likelihood and has been constructed in the context of data-driven variational inference [357, 406, 407]. The first component in Equation 6.18 implies minimizing $D_{\mathrm{KL}}\left(r_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)})\|q_{\boldsymbol{\theta}}(\mathbf{z}^{(i)})\right)$ in an aggregation of all considered $\boldsymbol{x}^{(i)}$. Hence, in aggregation the pre-images of $\boldsymbol{x}^{(i)}$, expressed by the approximate posterior, should resemble the generative component $q_{\boldsymbol{\theta}}(\mathbf{z})$, whereas the latter term in Equation 6.18 accounts for the reconstruction loss of encoded pre-images $\mathbf{z}^{(i)}$ (encoded through $r_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\boldsymbol{x}^{(i)})$) to its origin $\boldsymbol{x}^{(i)}$).

Minimizing the reverse KL divergence as introduced in Equation 6.7 with

$$D_{\mathrm{KL}}\left(q_{\boldsymbol{\theta}}(\mathbf{x})\|p_{\mathrm{target}}(\mathbf{x})\right)$$

implies a tractable maximization with respect to $(\boldsymbol{\phi},\boldsymbol{\theta})$ of the following objective based on [533]:

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta}) = \underbrace{-\beta\left\langle U(\mathbf{x})\right\rangle_{q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}}_{*} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}\left[\log r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]}_{\dagger} + \underbrace{\mathbb{H}(q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}))}_{\ddagger}. \tag{6.19}$$

We comment on the meaning of the indicated terms in Equation 6.19; however, note that the optimization always needs to be regarded in the composition of all terms.

$*$) Maximizing $\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta})$ seeks to minimize $\beta\left\langle U(\mathbf{x})\right\rangle_{q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}$, which corresponds to the average potential energy of the system evaluated under the generative model $q_{\boldsymbol{\theta}}(\mathbf{x})$.

$\dagger$) Maximize the expected log-probability that a given fine-scale realization $\boldsymbol{x}^{(i)}$ (with the corresponding latent pre-image $\boldsymbol{z}^{(i)}$) drawn from the joint of the generative model, $q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})$, can be reconstructed by $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ based on $\boldsymbol{x}^{(i)}$.

‡) Maximize the entropy of the generative model $\mathbb{H}(q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}))$.

Note that all aforementioned contributions must be seen in the composition, and maximizing $\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta})$ with respect to $(\boldsymbol{\phi},\boldsymbol{\theta})$ maximizes the balance of all. Most important is that the involved objective in the reverse KL divergence does not encompass any expectations with respect to $p_{\text{target}}(\mathbf{x})$, which need to be approximated by data as is the case in $\mathcal{L}^{\text{forward}}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{\mathcal{D}_N})$.

We discuss in the next section the particulars of the optimization with respect to $\boldsymbol{\theta},\boldsymbol{\phi}$, and also specify the form of the densities involved, i.e., $q_{\boldsymbol{\theta}}$ and $r_{\boldsymbol{\phi}}$.

6.1.5 Model specification and gradient derivation

In the sequel we introduce a general approach for parametrizing distributions $q_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})$ and $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and provide an approach for optimizing parameters with variance-reduction methods, enabling accelerated convergence.

Model specification

We base the model specification on previous work in the context of data-driven CVs discovery [415]. The model involves two components, ($q(\mathbf{x}|\mathbf{z})$ and $q(\mathbf{z})$), with respect to the generative path and the encoder $r(\mathbf{z}|\mathbf{x})$ in the recognition path.

As we seek to obtain a set of lower-dimensional coordinates representing characteristic and slow coordinates of the system, we aim to provide complexity in the mapping and thus the encoder and decoder components $r(\mathbf{z}|\mathbf{x})$ and $q(\mathbf{x}|\mathbf{z})$, respectively, and simple descriptions of the CVs through $q(\mathbf{z})$. Pushing complexity to the involved mappings and assuming simple correlations in $q(\mathbf{z})$ yields CVs capturing the most relevant features of the atomistic system compressed in low dimensions [543, 544].

The distribution $q_{\boldsymbol{\theta}_c}(\mathbf{z})$, which the obtained CVs are supposed to follow and which we desire to be simple, is represented as a standard Gaussian with unit diagonal variance:

$$q_{\boldsymbol{\theta}_c}(\mathbf{z}) = q(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\boldsymbol{I}). \tag{6.20}$$

The simplicity induced by Equation 6.20 is balanced by employing a flexible mapping given latent CVs $\mathbf{z}$ to fine-scale atomistic coordinates $\mathbf{x}$ (probabilistic decoder) with

$$q_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{z}),\boldsymbol{S}_{\boldsymbol{\theta}_{\text{cf}}}\right), \tag{6.21}$$

where the nonlinear mapping

$$\boldsymbol{\mu}_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{z}) = f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{z}), \tag{6.22}$$

with $\mathbf{z} \mapsto f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{z})$ ($f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}_{\text{cf}}} : \mathbb{R}^{n_c} \mapsto \mathbb{R}^{n_f}$) is expressed by a flexible (multilayer) neural network [428–430]. The Gaussian in Equation 6.21 with the flexible mean $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{z})$ is then fully defined by considering a diagonal covariance matrix with $\boldsymbol{S}_{\boldsymbol{\theta}_{\text{cf}}} = \text{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}_{\text{cf}}})$

[417]. We omit the subscripts of $\boldsymbol{\theta}$, as the latent generator $q(\mathbf{z})$ does not depend on parameters. Thus, we write $\boldsymbol{\theta} = \boldsymbol{\theta}_{\theta_{\text{cf}}}$. We treat the entries $\sigma^2_{\boldsymbol{\theta},j}$ directly as parameters without dependence on latent CVs $\mathbf{z}$. Maintaining $\sigma^2_{\boldsymbol{\theta},j} > 0$ is ensured by optimizing $\log \sigma^2_{\boldsymbol{\theta},j}$ instead.

In a similar fashion, compared to the model of $q_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{x}|\mathbf{z})$, we express the encoder that approximates the actual posterior distribution $p(\mathbf{z}|\mathbf{x})$ as follows:

$$r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \boldsymbol{S}_{\boldsymbol{\phi}}(\mathbf{x})\right), \tag{6.23}$$

with the diagonal covariance matrix $\boldsymbol{S}_{\boldsymbol{\phi}}(\mathbf{x}) = \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})\right)$. Likewise, $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$ and $\log \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})$ are obtained from encoding neural networks $f^{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$ and $f^{\sigma}_{\boldsymbol{\phi}}(\mathbf{x})$, respectively:

$$\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) = f^{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) \quad \text{and} \quad \log \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}) = f^{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}). \tag{6.24}$$

The actual but intractable posterior $q(\mathbf{z}|\mathbf{x})$ will differ from a multivariate normal distribution, for which we compensate by providing a flexible mean in $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. Structural correlations revealed by a full rank covariance matrix represent an interesting avenue to be explored [545]; however, this is not part of this paper. The employed models resemble those developed earlier in the context of CV discovery. Therefore, we refer to the discussion in [415] justifying the use of the neural networks.

We utilize the following general structure for the decoding neural network $f^{\mu}_{\boldsymbol{\theta}_{\text{cf}}}(\mathbf{z})$:

$$f^{\mu,K_q}_{\boldsymbol{\theta}}(\mathbf{z}) = \left(l^{(K_q+1)}_{\boldsymbol{\theta}_{\text{cf}}} \circ \tilde{a}^{(K_q)} \circ l^{(K_q)}_{\boldsymbol{\theta}} \circ \cdots \circ \tilde{a}^{(1)} \circ l^{(1)}_{\boldsymbol{\theta}}\right)(\mathbf{z}). \tag{6.25}$$

with K_q hidden layers. In a similar manner, we define the encoding networks for $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$ and $\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})$ of $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$:

$$f^{K_r}_{\boldsymbol{\phi}}(\mathbf{x}) = \left(a^{(K_r)} \circ l^{(K_r)}_{\boldsymbol{\phi}} \circ \cdots \circ a^{(1)} \circ l^{(1)}_{\boldsymbol{\phi}}\right)(\mathbf{x}), \tag{6.26}$$

which leads to the mean and diagonal terms of the covariance matrix with

$$f^{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) = l^{(K_r+1)}_{\boldsymbol{\phi}}\left(f^{K_r}_{\boldsymbol{\phi}}(\mathbf{x})\right) \quad \text{and} \quad f^{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}) = l^{(K_r+2)}_{\boldsymbol{\phi}}\left(f^{K_r}_{\boldsymbol{\phi}}(\mathbf{x})\right). \tag{6.27}$$

The linear layers used in the above expressions are denoted as $l^{(i)}$, e.g., mapping a variable $\boldsymbol{y}$ to the output with $l^{(i)}(\boldsymbol{y}) = \boldsymbol{W}^{(i)}\boldsymbol{y} + \boldsymbol{b}^{(i)}$. The nonlinearities in $f^{(\cdot)}_{(\cdot)}$ are implied by activation $a(\cdot)$. Encoding and decoding functions are indicated by the superscripts $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively. Activation functions belonging to the encoder are $a^{(i)}$, and those involved in decoding $\mathbf{z}$ are $\tilde{a}^{(i)}$. The size of $\boldsymbol{W}^{(i)}$ is specified by the input dimension, which could be the output of a precedent layer $l^{(i-1)}(\boldsymbol{y})$, and the output dimension, which we specify with $d_{l^{(i)}}$. This leads to a matrix $\boldsymbol{W}^{(i)}$ of dimension $d_{l^{(i)}} \times d_{l^{(i-1)}}$. The corresponding parametrization details with depth K and activations of the networks are specified with the corresponding numerical illustrations in Section 6.2.

Gradient computation and reparametrization

This section is devoted to deriving relevant gradients of the objective $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta})$ in Equation 6.12, which involve the fine-scale potential energy $U(\mathbf{x})$. We show a noise-reducing gradient estimator by utilizing reparametrization [408, 546].

The focus is on the first component in Equation 6.12, which depends only on the parameters $\boldsymbol{\theta}$. We write for the corresponding derivative:

$$\begin{aligned} -\beta \frac{\partial}{\partial \boldsymbol{\theta}} \langle U(\mathbf{x}) \rangle_{q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})} &= -\beta \frac{\partial}{\partial \boldsymbol{\theta}} \int \int q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) q(\mathbf{z}) U(\mathbf{x}) \, d\mathbf{x} \, d\mathbf{z} \\ &= -\beta \int q(\mathbf{z}) \underbrace{\frac{\partial}{\partial \boldsymbol{\theta}} \left(\int q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) U(\mathbf{x}) \, d\mathbf{x} \right)}_{\nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}[U(\mathbf{x})]} d\mathbf{z}. \end{aligned} \tag{6.28}$$

In the last line of the above equation, we note the expression $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})} [U(\mathbf{x})]$; this is for the case of using approximate MC estimators, highly affected by noise, as discussed in [406]. This would hamper the optimization even when employing stochastic techniques. The variance of the approximate estimator of $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})} [U(\mathbf{x})]$ can be reduced by the reparametrization of $q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. This is done by introducing an auxiliary random variable $\boldsymbol{\epsilon}$, which gives rise to $\mathbf{x}$ by a differentiable transformation:

$$\mathbf{x} = g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}; \mathbf{z}) \text{ with } \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}). \tag{6.29}$$

With the mapping, $g_{\boldsymbol{\theta}} : \boldsymbol{\epsilon} \to \mathbf{z}$, the following holds by change of variables:

$$q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = p\left(g_{\boldsymbol{\theta}}^{-1}(\mathbf{x}; \mathbf{z})\right) \left| \frac{\partial g_{\boldsymbol{\theta}}^{-1}(\mathbf{x}; \mathbf{z})}{\partial \mathbf{x}} \right|, \tag{6.30}$$

where the inverse function of $g_{\boldsymbol{\theta}}$, $g_{\boldsymbol{\theta}}^{-1} : \mathbf{x} \to \boldsymbol{\epsilon}$ leads to $\boldsymbol{\epsilon} = g_{\boldsymbol{\phi}}^{-1}(\mathbf{x}; \mathbf{z})$. Different possibilities of auxiliary distributions and invertible transformations are discussed in more detail in [547]. With the introduced transformation, we can rewrite the derivative with:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})} [U(\mathbf{x})] &= \mathbb{E}_{p(\boldsymbol{\epsilon})} \left[\nabla_{\boldsymbol{\theta}} U \left(g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}; \mathbf{z}) \right) \right] \\ &= \mathbb{E}_{p(\boldsymbol{\epsilon})} \left[\frac{\partial U \left(g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}; \mathbf{z}) \right)}{\partial \mathbf{x}} \frac{\partial g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}; \mathbf{z}) \right)}{\partial \boldsymbol{\theta}} \right]. \end{aligned} \tag{6.31}$$

The auxiliary random variables $\boldsymbol{\epsilon}$ follow a Gaussian with $\boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The corresponding transformation for representing the random variables $\mathbf{x}$ is:

$$\mathbf{x} = g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}; \mathbf{z}) = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}) + \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{z}) \odot \boldsymbol{\epsilon}. \tag{6.32}$$

Replacing the expression $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}\left[U(\mathbf{x})\right]$ in Equation 6.28 with Equation 6.31 leads to:

$$
\begin{aligned}
-\beta \frac{\partial}{\partial \boldsymbol{\theta}} \left\langle U(\mathbf{x}) \right\rangle_{q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})} &= -\beta \left\langle \mathbb{E}_{p(\boldsymbol{\epsilon})} \left[\frac{\partial U\left(g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon};\mathbf{z})\right)}{\partial \mathbf{x}} \frac{\partial g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon};\mathbf{z}))}{\partial \boldsymbol{\theta}} \right] \right\rangle_{q(\mathbf{z})} \\
&= -\beta \Bigg\langle \underbrace{\frac{\partial U\left(g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon};\mathbf{z})\right)}{\partial \mathbf{x}}}_{=-\mathbf{F}(\mathbf{x})} \frac{\partial g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon};\mathbf{z}))}{\partial \boldsymbol{\theta}} \Bigg\rangle_{p(\boldsymbol{\epsilon})q(\mathbf{z})} \quad .
\end{aligned} \tag{6.33}
$$

First, the physically less interesting part in Equation 6.33 is the contribution $\frac{\partial g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon};\mathbf{z}))}{\partial \boldsymbol{\theta}}$, which can be estimated by employing efficient backpropagation and automatic differentiation algorithms for neural networks [428, 548]. However, the more physically relevant component, the gradient of the atomistic potential, $\frac{\partial U(g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon};\mathbf{z}))}{\partial \mathbf{x}}$, is involved in Equation 6.33. The gradient of the potential $U(\mathbf{x})$ with respect to $\mathbf{x}$ equals the negative interatomic force $\mathbf{F}(\mathbf{x})$, evaluated at $\mathbf{x}$, where $\mathbf{x} = g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon};\mathbf{z})$. This latter term incorporates physics into the gradient computation of $\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta})$ in the form of interatomic forces. This is the source from which physics are embedded into our proposed model and drives the optimization of $\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta})$ by querying the force field at samples of $q_{\boldsymbol{\theta}}(\mathbf{x})$. Notably, the forces are incorporated at atomistic positions $\mathbf{x}$, which are determined by sampling as follows.

(i) Draw a sample from the generative distributions: $\boldsymbol{z}^{(i)} \sim q(\mathbf{z})$ which is simple to sample from.

(ii) Then obtain a sample from the auxiliary distribution: $\boldsymbol{\epsilon}^{(j)} \sim p(\boldsymbol{\epsilon})$.

(iii) Determine the corresponding atomistic representation of $(\boldsymbol{z}^{(i)}, \boldsymbol{\epsilon}^{(j)})$ with: $\mathbf{x}^{(i,j)} = g_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}^{(j)}; \boldsymbol{z}^{(i)}) = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)}) + \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)}) \odot \boldsymbol{\epsilon}^{(j)}$.

This means we evaluate the force $\mathbf{F}$ at samples $\mathbf{x}^{(i,j)}$; no reference data are required in this process.

The force evaluation at atomistic coordinates $\mathbf{x}$ is the heart of common MD software such as LAMMPS [315], GROMACS [316–322], and OpenMM [44]. The MD simulators are highly sophisticated in terms of efficiency and allow us to employ this optimized force evaluation function in our development.

In this work, we develop a PyTorch module that incorporates OpenMM [44] in the backward pass, which enables efficient optimization by querying the forces computed by OpenMM at input positions governed by $q_{\boldsymbol{\theta}}(\mathbf{x})$. We are continuously developing the software on GPU platforms, and it will be made available[2].

6.1.6 Training

Training the model parameters $(\boldsymbol{\phi},\boldsymbol{\theta})$ requires some attention as variational models tend to be mode-focusing [356, 489]. If parameters update too rapidly, in terms of

[2]Software available upon publication on `https://github.com/ms.../...`.

configurations of $p_{\text{target}}(\mathbf{x})$ that have been explored by $q_{\boldsymbol{\theta}}(\mathbf{x})$ thus far, relevant conformations could be missed. However, compared with the data-driven approach, the proposed variational coarse-graining methodology offers strategies ensuring that relevant conformations are captured and incorporates querying of the potential $U(\mathbf{x})$ into the learning procedure. In data-driven schemes, once the data is obtained, there is no control on exploring unobserved conformations [190, 544]. Remedy, next to employing stochastic optimization with adaptive step size control [412], provide tempering approaches [458, 549]. These start at high initial temperatures or low inverse temperatures, with, e.g., $0 \leq \beta_1$, whereas $\beta = 0$ resembles a uniform target distribution. A sequence of K temperatures and related inverse temperatures $0 \leq \beta_1 \cdots \leq \beta_k \leq \ldots \beta_K$ yields a sequence of target distributions with [462–464]

$$p_{\text{target}}(\mathbf{x};\beta_k) = \frac{1}{Z(\beta_k)} e^{-\beta_k U(\mathbf{x})}, \quad \forall\, k \in \{1,\ldots,K\}, \tag{6.34}$$

while β_K equals the target simulation temperature β_{target}.

Instead of directly minimizing $D_{\text{KL}}\left(q(\mathbf{x}) \| p_{\text{target}}(\mathbf{x};\beta_{\text{target}})\right)$, we minimize subsequent $D_{\text{KL}}\left(q(\mathbf{x}) \| p_{\text{target}}(\mathbf{x};\beta_{\mathbf{k}})\right)$ while we obtain optimal $(\boldsymbol{\phi}_k, \boldsymbol{\theta}_k)$, which we use as initial parameters for minimizing $D_{\text{KL}}\left(q(\mathbf{x}) \| p_{\text{target}}(\mathbf{x};\beta_{\mathbf{k+1}})\right)$. However, the size of the increment between two subsequent temperature steps $\Delta\beta_k = \beta_{k+1} - \beta_k$ is a difficult choice.

Therefore, we develop an adaptive scheme for gradually increasing β_k, which adjusts the proposed $\Delta\beta_k$ such that the relative difference in subsequent KL divergences estimated at β_k and β_{k+1} does not exceed a threshold $c_{\max}$. We define the relative increase of the KL divergence between β_k and β_{k+1} with:

$$\frac{D_{\text{KL}}\left(q(\mathbf{x}) \| p_{\text{target}}(\mathbf{x};\beta_{k+1})\right) - D_{\text{KL}}\left(q(\mathbf{x}) \| p_{\text{target}}(\mathbf{x};\beta_k)\right)}{D_{\text{KL}}\left(q(\mathbf{x}) \| p_{\text{target}}(\mathbf{x};\beta_k)\right)}. \tag{6.35}$$

By employing the derived upper bound on the KL divergence, which is defined in Equation 6.16, we can rewrite Equation 6.35 as

$$c_k = \frac{\log(Z(\beta_{k+1})) - \log(Z(\beta_k)) + (\beta_{k+1} - \beta_k)\left\langle U(\mathbf{x})\right\rangle_{q(\mathbf{x},\mathbf{z})}}{\log Z(\beta_k) + \beta_k \left\langle U(\mathbf{x})\right\rangle_{q(\mathbf{x},\mathbf{z})} - \left\langle \log r(\mathbf{z}|\mathbf{x})\right\rangle_{q(\mathbf{x},\mathbf{z})} - \mathbb{H}(q(\mathbf{x},\mathbf{z}))}. \tag{6.36}$$

Besides the (log-)difference of the normalization constants, $\log(Z(\beta_{i+1})) - \log(Z(\beta_i))$, and $\log(Z(\beta_k)$, all remaining components in Equation 6.36 are accessible through MC estimators. The supporting material in Appendix E.2 includes an approximation of $\log(Z(\beta_{k+1})) - \log(Z(\beta_k)$ and $\log(Z(\beta_k)$. The procedure for updating the

temperature is summarized in Algorithm 4.

Algorithm 4: Tempering scheme for updating β_k. We set $\Delta\beta_{\max} = 1.0 \times 10^{-3}$ and $c_{\max} = 1.0$.

Input: Converged model with its parameter $(\boldsymbol{\phi}_k, \boldsymbol{\theta}_k)$ at current inverse temperature β_k; $c_{\max}$, maximal relative increase in D_{KL}; $\Delta\beta_{\max}$, the temperature increment; Current step k.

Output: β_{k+1}

Initialize: $s := 0$, $f_s := 1.0$.

while $c_k^s > c_{max}$ **do**

  Propose new inverse temperature β_{k+1}^s**:**

  $\beta_{k+1}^s = \beta_k + f_s \Delta\beta_{\max}$

  Estimate rel. increase c_k^s **with proposed** β_{k+1}^s**:**

  See computation in Equation 6.36.

  Update f_s **for proposing a new maximal increase in** β**:**

  $f_{s+1} = 0.6 f_s$

  Update step:

  $s = s + 1$.

Set: $\beta_{k+1} = \beta_{k+1}^s$

Update: $k = k + 1$

Continue optimization with: $\log p_{\mathrm{target}}(\mathbf{x}; \beta_k) \propto e^{\beta_k U(\mathbf{x})}$

6.2 Numerical illustrations

The following section demonstrates the developed methodology based on a double well potential in Section 6.2.1 and an alanine dipeptide in Section 6.2.2.

6.2.1 Double well

This section shows the capabilities of the proposed method in the context of a two-dimensional double well potential energy function $U(\mathbf{x})$ ($\dim(\mathbf{x}) = 2$) that exhibits two distinct modes distinguishable in the x_1 direction. One of the modes is favorably explored owing to its lower potential energy. The potential is quadratic in the x_2 direction, as depicted in Figure 6.1:

$$U(\mathbf{x}) = \frac{1}{4} x_1^4 - 3 \cdot x_1^2 + x_1 + \frac{1}{2} x_2^2. \tag{6.37}$$

The double well potential in Equation 6.37 and the implied target distribution $p_{\mathrm{target}}(\mathbf{x}; \beta = 1) \propto e^{-\beta U(\mathbf{x})}$ result in a distribution that is challenging to explore with purely random walk MCMC and without performing extensive fine-tuning of the proposal step. A test MCMC estimator, which was as fair as possible, did not discover the second mode for $x_1 > 0$ after 1×10^5 steps. The natural CV of the potential

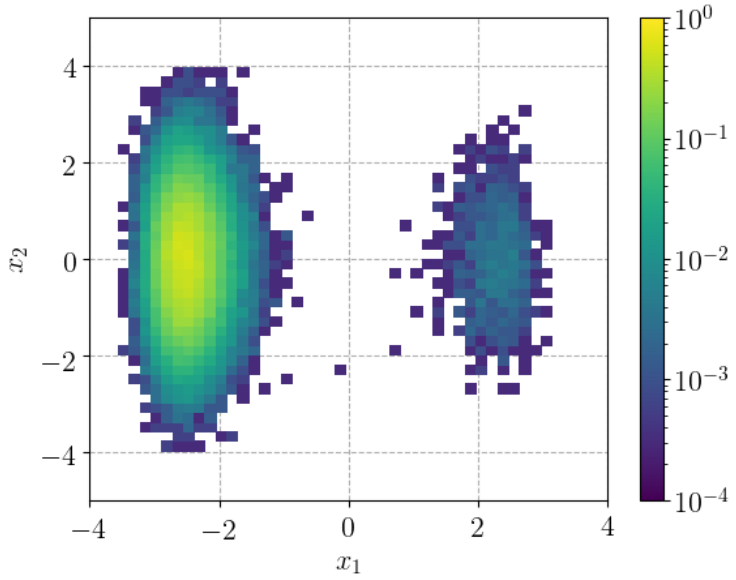


FIGURE 6.1: Reference potential energy $U(\mathbf{x})$. The color relates to the value of $U(\mathbf{x})$ quantified by the *logarithmic* color bar on the right. Most MCMC random walk approaches will discover only one of the depicted potential energy basins.

$U(\mathbf{x})$ and thus of $p_{\text{target}}(\mathbf{x}) \propto e^{-U(\mathbf{x})}$ is the x_1 coordinate. The x_1 direction distinguishes the two modes that $p_{\text{target}}(\mathbf{x})$ exhibits. We expect our algorithm to reveal CVs z "equal" to x_1 or having high correlation with x_1. We put "equal" in quotes as we work in a probabilistic framework. The dimensionality of $\dim(\mathbf{z})$ is 1.

The functional form and parameters have been taken from [240] to ensure comparability. However, note that we seek to identify simultaneously the lower-dimensional characteristics revealing the relevant physics, encoded in CVs, and obtain a generative CG model for predictive purposes. In [240], the focus was on the generative component. The CVs utilized for learning are *selected* rather than revealed from the physics. The latent CVs $\mathbf{z}$ have the same dimensionality as $\mathbf{x}$ owing to the use of *invertible* neural networks that require $\dim(\mathbf{z}) = \dim(\mathbf{x})$ [437].

We employ the model as introduced in Section 6.1.5 and define the unspecified options such as the number of layers, layer dimensions, and activation functions used in the encoder and decoder as in Tables 6.1 and 6.2, respectively. To train the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$, we employ a tempering scheme as introduced in Section 6.1.6 and specified in Algorithm 4 with initial $\beta_0 = 1 \times 10^{-10}$, while the target is defined with $\beta_K = 1$. For all numerical illustrations, we employ ADAM stochastic optimization [472] with $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon_{\text{ADAM}} = 1.0 \times 10^{-8}$. The expectations with respect to $q(\mathbf{x}, \mathbf{z})$ are computed based on $J = 1000$ samples.

We will assess the trained model with respect to its predictive power and the obtained CVs in the following.

Linear layer	Input dimension	Output dimension	Activation layer	Activation function
$l_{\boldsymbol{\phi}}^{(1)}$	$\dim(\mathbf{x}) = 2$	d_1	$a^{(1)}$	SeLu[3]
$l_{\boldsymbol{\phi}}^{(2)}$	d_1	d_2	$a^{(2)}$	SeLu
$l_{\boldsymbol{\phi}}^{(3)}$	d_2	d_3	$a^{(3)}$	TanH
$l_{\boldsymbol{\phi}}^{(4)}$	d_3	$\dim(\mathbf{z})$	None	-
$l_{\boldsymbol{\phi}}^{(5)}$	d_3	$\dim(\mathbf{z})$	None	-

TABLE 6.1: Network specification of the encoding neural network with $d_{\{1,2,3\}} = 100$.

Linear layer	Input dimension	Output dimension	Activation layer	Activation function
$l_{\boldsymbol{\theta}}^{(1)}$	$\dim(\mathbf{z}) = 1$	d_3	$\tilde{a}^{(1)}$	Tanh
$l_{\boldsymbol{\theta}}^{(2)}$	d_3	d_2	$\tilde{a}^{(2)}$	Tanh
$l_{\boldsymbol{\theta}}^{(3)}$	d_1	$\dim(\mathbf{x})$	None	-

TABLE 6.2: Network specification of the decoding neural network with $d_{\{1,2,3\}}$ as defined in Table 6.1.

Predictive CG model

Figures 6.2 and 6.3 show intermediate results obtained while training the model. The left columns depict a two-dimensional (2D) histogram containing the target histogram based on a long reference simulation obtained by employing the Metropolis-adjusted Langevin algorithm [451] at $\beta = 1$. Next to the histogram of $p_{\text{target}}(\mathbf{x}; \beta = 1)$, we provide 2D histograms of intermediate predictions at β_k, as indicated in the sub-caption. The predictive histograms are obtained by drawing J samples from the predictive distribution $q_{\boldsymbol{\theta}}(\mathbf{x})$. The latter is very simple and computationally efficient owing to the use of ancestral sampling [442] of the generative model, as explained in the Section 6.1.2. The right columns of Figures 6.2 and 6.3 provide the reference potential energy $U(x_1, x_2 = 0)$, the intermediate target potential $\beta_k U(x_1, x_2 = 0)$, and the predicted potential $U_k^{\text{pred}}(x_1, x_2 = 0)$ after convergence of $(\boldsymbol{\phi}, \boldsymbol{\theta})$ at temperature β_k. For the intermediate steps, we estimate $U_k^{\text{pred}}(x_1, x_2 = 0)$ as follows:

$$U_k^{\text{pred}}(x_1, x_2 = 0) \propto -\frac{1}{\beta_k} \log q_{\boldsymbol{\theta}}(x_1, x_2 = 0). \tag{6.38}$$

We note that the evaluation of $\log q_{\boldsymbol{\theta}}(x_1, x_2 = 0)$ requires approximation of the integral $\log q_{\boldsymbol{\theta}}(x_1, x_2 = 0) = \int q(\mathbf{x}|\mathbf{z}) q(\mathbf{z})\, d\mathbf{z}$, which induces noise. The aforementioned integral has been approximated by $N = 5000$ samples drawn from $q(\mathbf{z})$.

Figure 6.4(a) shows the overall convergence of the model, expressed in the form of the reverse KL divergence (Equation 6.7) and the forward KL divergence (Equation 6.6); the latter, which relies on the data, is only used for illustrative purposes. Data for evaluating $D_{\text{KL}}\left(p_{\text{target}}(\mathbf{x}) \| q_{\boldsymbol{\theta}}(\mathbf{x})\right)$ were not used in the training process. We

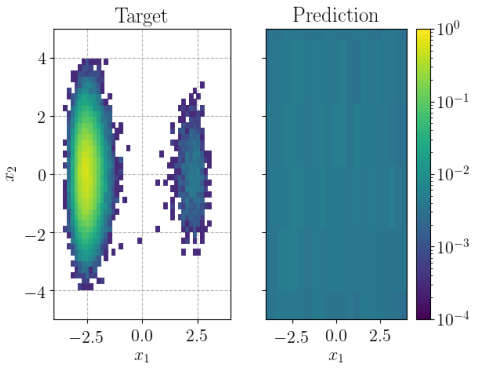
compare reference statistics (again based on data which were not used during training) with statistics estimated based on the efficient predictive distribution $q_{\boldsymbol{\theta}}(\mathbf{x})$ in Figure 6.4(b)

Predictive collective variables

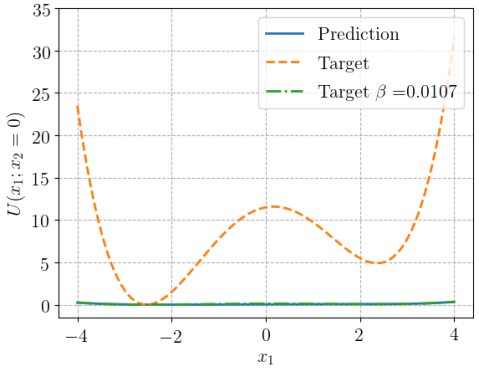
The proposed approach provides an efficient CG model that can be employed for predictive purposes, as described in the previous section. We claim that in addition to obtaining a CG model, we can provide relevant insights by identifying CVs of the system. In the double well example, one would expect the CV to be the x_1 coordinate that separates the two modes, where conformational changes are implied by moving along x_1.

To visualize the assigned CVs given samples $\boldsymbol{x}^{(i)} \sim q_{\boldsymbol{\theta}}(\mathbf{x})$, we plot samples as dots in Figure 6.5, while the color of the $\boldsymbol{x}^{(i)}$ is assigned based on the corresponding value of the CV. We note that for every $\boldsymbol{x}^{(i)}$ there exists a whole distribution of CVs $r_{\boldsymbol{\phi}}(z|\boldsymbol{x}^{(i)})$, as we work in a probabilistic framework. The assigned color in Figure 6.5 is based on the mean of $r_{\boldsymbol{\phi}}(z|\boldsymbol{x}^{(i)})$, which is obtained by evaluating $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(i)})$.

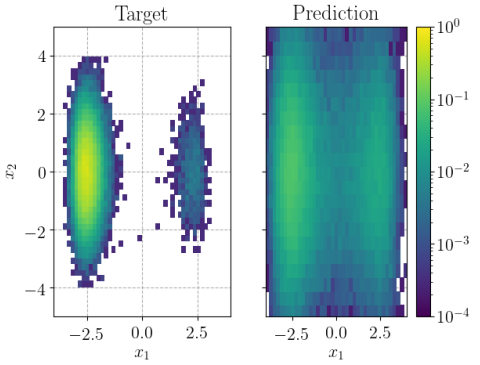
The (color) gradient of z with respect to $\mathbf{x}$ is almost exactly parallel to the x_1-direction, which implies that the revealed CV z is (probabilistically) parallel to the x_1 axis and thus meets our expectations. The proposed approach reveals the relevant, slow, CV x_1 solely by evaluating $U(\mathbf{x})$ under $q_{\boldsymbol{\theta}}(\mathbf{x})$.



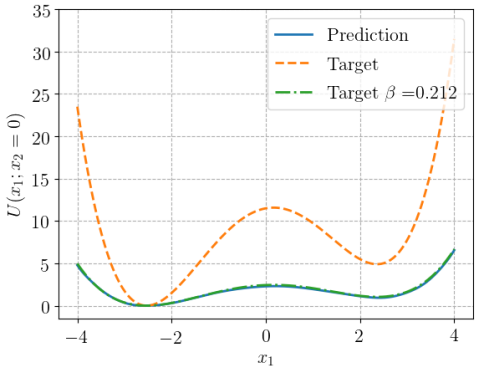
(a) Histogram of $q_{\theta}(\mathbf{x})$ at $\beta \approx 0$ and of $p_{\text{target}}(\mathbf{x})$.



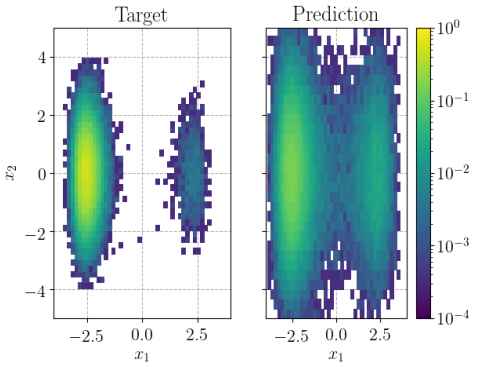
(b) $U(x_1, x_2 = 0)$ at $\beta \approx 0$.



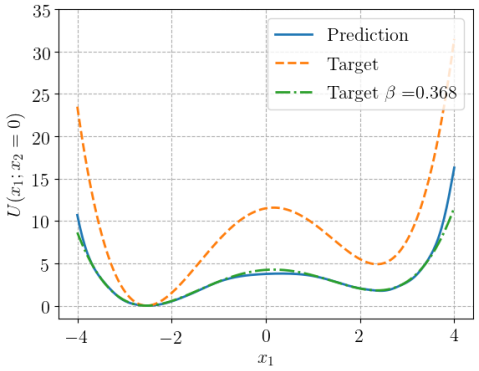
(c) Histogram of $q_{\theta}(\mathbf{x})$ at $\beta \approx 0.2$ and of $p_{\text{target}}(\mathbf{x})$.



(d) $U(x_1, x_2 = 0)$ at $\beta \approx 0.2$.

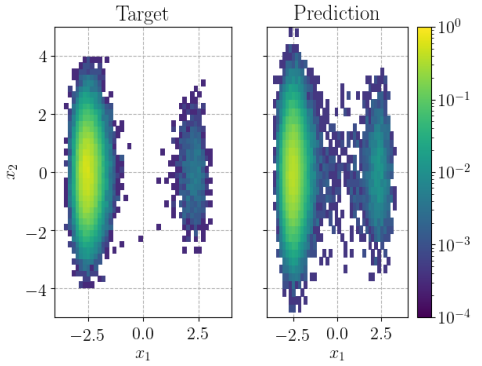


(e) Histogram of $q_{\theta}(\mathbf{x})$ at $\beta \approx 0.36$ and of $p_{\text{target}}(\mathbf{x})$.

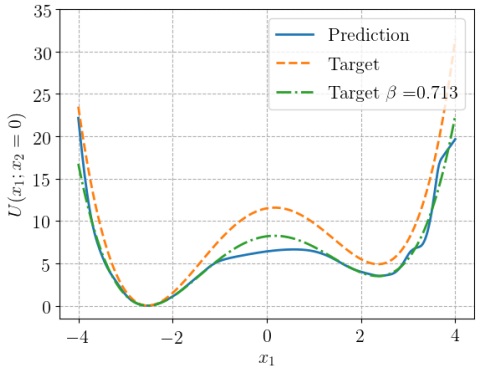


(f) $U(x_1, x_2 = 0)$ at $\beta \approx 0.36$.

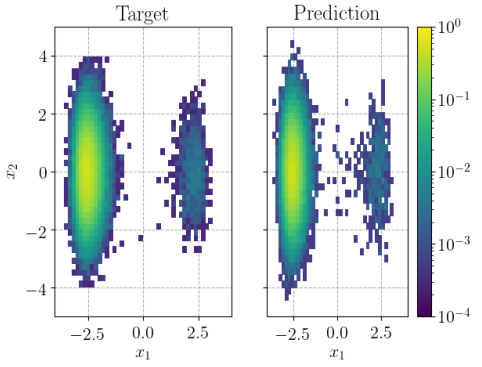
FIGURE 6.2: The left column shows histograms of the target $p_{\text{target}}(\mathbf{x})$ (at $\beta = 1$) and predictions based on $q_{\phi}(\mathbf{x})$ at the indicated temperature β in the subcaptions. The right column shows a 1D slice through the potential energy $U(\mathbf{x})$ at $x_2 = 0$, emphasizing the two distinct modes. The figures include the reference potential for the indicated temperature β_k with $\beta_k U(\mathbf{x})$ and an estimation of $U_k^{\text{pred}}(\mathbf{x})$ based on $q_{\theta}(\mathbf{x})$.



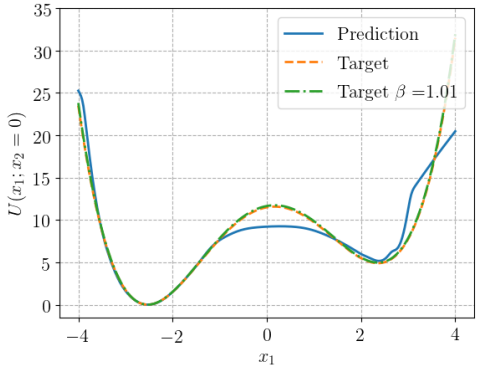
(a) Histogram of $q_{\boldsymbol{\theta}}(\mathbf{x})$ at $\beta \approx 0.7$ and of $p_{\text{target}}(\mathbf{x})$.



(b) $U(x_1, x_2 = 0)$ at $\beta \approx 0.7$.

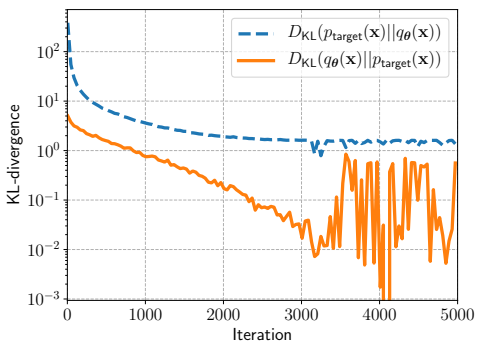


(c) Histogram of $q_{\boldsymbol{\theta}}(\mathbf{x})$ at $\beta \approx 1$ and of $p_{\text{target}}(\mathbf{x})$.

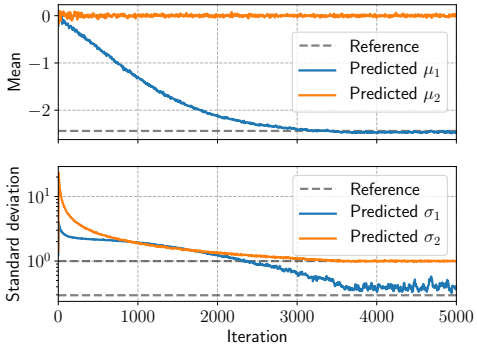


(d) $U(x_1, x_2 = 0)$ at $\beta \approx 1$.

FIGURE 6.3: The left column shows histograms of the target $p_{\text{target}}(\mathbf{x})$ (at $\beta = 1$) and predictions based on $q_{\boldsymbol{\phi}}(\mathbf{x})$ at the indicated temperature β in the subcaptions. The right column shows a 1D slice through the potential energy $U(\mathbf{x})$ at $x_2 = 0$, emphasizing the two distinct modes. The figures include the reference potential for the indicated temperature β_k with $\beta_k U(\mathbf{x})$ and an estimation of $U_k^{\text{pred}}(\mathbf{x})$ based on $q_{\boldsymbol{\theta}}(\mathbf{x})$.

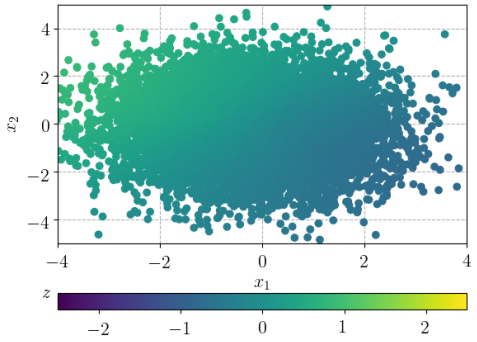


(a) Upper and lower bounds of the training objective.

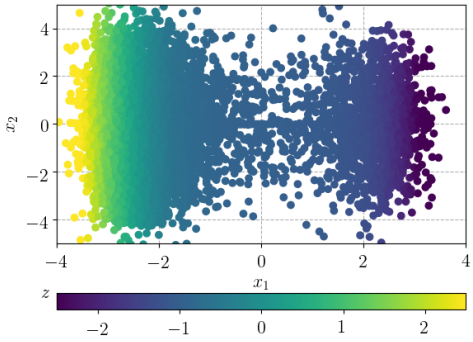


(b) Predicted mean and standard deviations compared with reference data-based estimates. The subscripts of μ and σ indicate the corresponding x_1 and x_2 directions.

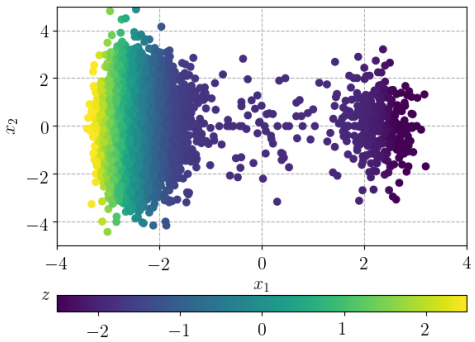
FIGURE 6.4: Convergence of the KL divergences (left) and predicted statistics compared with reference estimates (right).



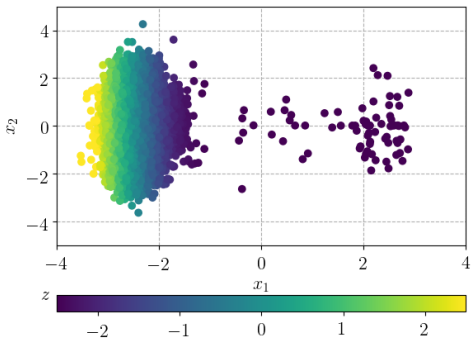
(a) $\beta \approx 0$



(b) $\beta \approx 0.36$

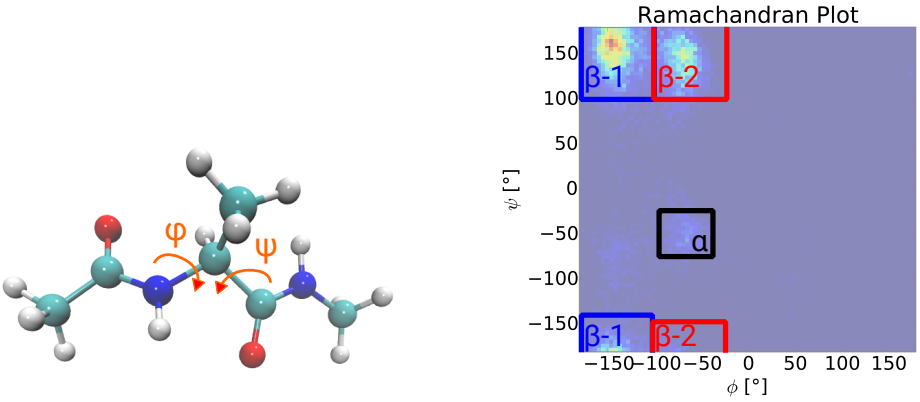


(c) $\beta \approx 0.7$



(d) $\beta \approx 1$

FIGURE 6.5: Samples $\boldsymbol{x}^{(i)} \sim q_{\boldsymbol{\theta}}(\mathbf{x})$ at the indicated temperature β are depicted as dots, whereas the assigned color of $\boldsymbol{x}^{(i)}$ corresponds to its latent CV obtained by the mean of $r_{\boldsymbol{\phi}}(\mathbf{z}|\boldsymbol{x}^{(i)})$. The color bar below the images shows the color corresponding to the assigned value of the CV $\mathbf{z}$ given $\mathbf{x}$. The figure is based on $N = 1 \times 10^4$ samples $\boldsymbol{x}^{(i)}$.



(a) Dihedral angels for ALA-2.

(b) Characteristic conformations according [500].

FIGURE 6.6: Dihedral angles (left) and (ϕ, ψ) statistics of a reference simulation with characteristic modes (right).

6.2.2 ALA-2

After demonstrating the functionality of the proposed scheme for a double well potential energy, we are interested in addressing an atomistic system. The following is devoted to the CV discovery of alanine dipeptide (ALA-2) in the context of an implicit solvent. Characteristic coordinates of the ALA-2 peptide include the dihedral angles (ϕ, ψ), as shown in Figure 6.6(a). Distinct combinations of the dihedral angles characterize three distinguishable (α, β-1, β-2) conformations, as provided in the Ramachandran plot [501] in Figure 6.6(b) [500]. The peptide consisting of 22 atoms can be described by 60 effective degrees of freedom (rigid body motion removed); however, we store the complete Cartesian coordinate vector $\mathbf{x}$ with $\dim(\mathbf{x}) = 66$, where six degrees of freedom are fixed. The exact representation of ALA-2 in $\mathbf{x}$ with coordinate bookkeeping is given in the Appendix E.3.

Reference model setting

Applying the proposed methodology does not require the production of any reference atomistic trajectories. However, we are interested in comparing our obtained predictions from the generative CG model to reference observables estimated by a reference MD simulation. We refer to Appendix E.4 for all necessary details obtaining the MD trajectory. Nevertheless, for the sake of evaluating forces, we need to specify system properties such as the force field, which in this case is AMBER ff96 [337–339]. We employ an implicit water model based on the generalized Born approach [551, 552], which serves as a solvent. The temperature of interest is $T = 330\,\mathrm{K}$.

Linear layer	Input dimension	Output dimension	Activation layer	Activation function
$l_{\boldsymbol{\phi}}^{(1)}$	$\dim(\mathbf{x}) = 66$	d_1	$a^{(1)}$	SeLu[4]
$l_{\boldsymbol{\phi}}^{(2)}$	d_1	d_2	$a^{(2)}$	SeLu
$l_{\boldsymbol{\phi}}^{(3)}$	d_2	d_3	$a^{(3)}$	Log Sigmoid [5]
$l_{\boldsymbol{\phi}}^{(4)}$	d_3	$\dim(\mathbf{z})$	None	-
$l_{\boldsymbol{\phi}}^{(5)}$	d_3	$\dim(\mathbf{z})$	None	-

TABLE 6.3: Network specification of the encoding neural network with $d_{\{1,2,3\}} = 170$.

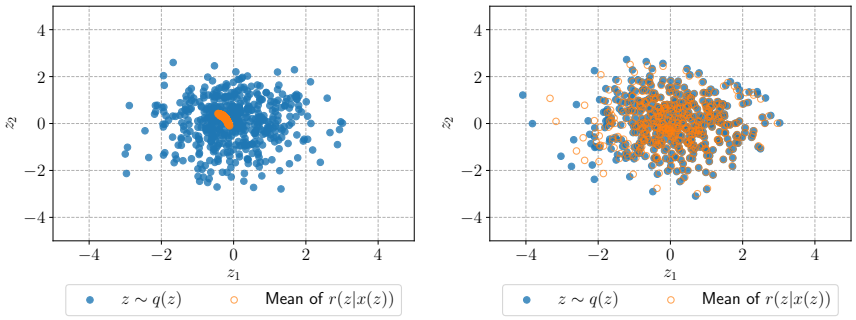
Linear layer	Input dimension	Output dimension	Activation layer	Activation function
$l_{\boldsymbol{\theta}}^{(1)}$	$\dim(\mathbf{z}) = 2$	d_3	$\tilde{a}^{(1)}$	Tanh
$l_{\boldsymbol{\theta}}^{(2)}$	d_3	d_2	$\tilde{a}^{(2)}$	Tanh
$l_{\boldsymbol{\theta}}^{(3)}$	d_2	d_1	$\tilde{a}^{(3)}$	Tanh
$l_{\boldsymbol{\theta}}^{(4)}$	d_1	$\dim(\mathbf{x})$	None	-

TABLE 6.4: Network specification of the decoding neural network with $d_{\{1,2,3\}} = 120$.

Model specification

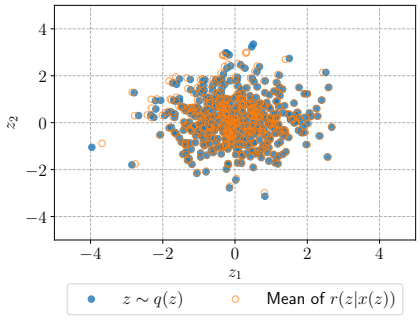
The general model structure introduced earlier in Section 6.1.5 is also employed in the context of the ALA-2 setting. We mostly rely on findings in [415], where an identical system was explored on the basis of data-driven forward KL divergence minimization. All required details for the model are specified in Tables 6.3 and 6.4 for the encoder ($r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$) and decoder ($q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$), respectively. Similar to the previous example in Section 6.2.1, we employ a tempering scheme as introduced in Section 6.1.6 and specified in Algorithm 4 with initial $\beta_0 = 1 \times 10^{-14} \cdot \beta_K$, while the target temperature is defined by $\beta_K = \frac{1}{k_B T}$ and $T = 330\,\mathrm{K}$. The inverse temperature β_0 occurs as a multiplicative factor, multiplying the potential energy $U(\mathbf{x})$. For gradient estimation, the interatomic force $\mathbf{F}(\mathbf{x})$ is multiplied by β_k. In the variational approach presented in this work, we evaluate the force field under samples from $q_{\boldsymbol{\theta}}(\mathbf{x})$. However, when $q_{\boldsymbol{\theta}}(\mathbf{x})$ has not yet learned, samples $\boldsymbol{x}^{(i)}$ will potentially yield high-energy states associated with large forces. According to experimental results, the magnitude of $\mathbf{F}(\mathbf{x})$ in early training stages reaches $\pm 1 \times 10^{-18}$. Therefore, the initial inverse temperature is chosen such that $\beta_0 \mathbf{F}(\mathbf{x})$ evaluated under $q_{\boldsymbol{\theta}}(\mathbf{x})$ yields values of $\pm 1 \times 10^{1}$. This implies that the embedded physics, expressed by interatomic forces $\mathbf{F}(\mathbf{x})$, are weak in the early training stages and are emphasized as the learning process proceeds with increasing β_k. For further details, refer to Appendix E.6.

The stochastic optimization algorithm is ADAM [472] with $\alpha = 0.001, \beta_1 =$



(a) Iteration 700.

(b) Iteration 2200.



(c) Iteration 7200.

FIGURE 6.7: Samples from $q_{\boldsymbol{\theta}}(\mathbf{z})$ (blue, filled) are decoded with $q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and encoded with $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (orange, no facecolor). We consider the means of aforementioned distributions for performing the decoding and encoding processes. Re-encoding the decoded $\boldsymbol{z}^{(i)}$ matches its origin.

$0.9, \beta_2 = 0.999, \epsilon_{\text{ADAM}} = 1.0 \times 10^{-8}$. We employ $J = 2000$ samples for computing expectations with respect to $q_{\boldsymbol{\theta}}(\mathbf{x})$ throughout the training process. Initially, in the early training stages, using fewer samples does not influence the training. The number of samples should be increased once the model has been refined and comes closer to $p_{\text{target}}(\mathbf{x})$.

Collective variables

When training the model with its encoder and decoder components $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, it is important that these consistently map a generated sample $\boldsymbol{z}^{(i)} \sim q(\mathbf{z})$ to $\boldsymbol{x}^{(i)}$ through the decoder $q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, and from the decoded atomistic configuration $\boldsymbol{x}^{(i)}$ back to its origin, the value of the CV $\boldsymbol{z}^{(i)}$ it has been generated from. The projection from $\boldsymbol{x}^{(i)}$ to the according CV is enabled through the encoder $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. After some initial iterations optimizing $(\boldsymbol{\phi}, \boldsymbol{\theta})$, the encoder and decoder work consistently as depicted in Figure 6.7.

In Figure 6.8 we utilize the identified encoder $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, which assigns CVs to an input atomistic configuration, for encoding a reference test dataset. This dataset has not been used for training and is used here solely for visualization purposes. The test data (generated according to Appendix E.4) contains atomistic configurations

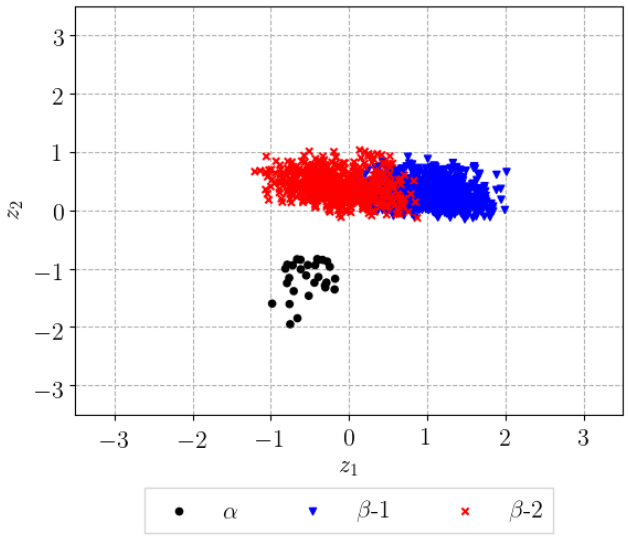


FIGURE 6.8: Representation of $\mathbf{z}$-coordinates of test data assigned by the mean of $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{z})$, which we learn by minimizing the reverse KL divergence without reference data. Characteristic conformations of ALA-2 are indicated in: α black, β-1 blue, and β-2 red color. Without any prior physical information and in the complete absence of any data, the encoder identifies physically relevant coordinates, which are related to ϕ, ψ values.

from multiple characteristic modes based on their dihedral angle values (ϕ, ψ) as shown in Figure 6.6(b). Given a datum from the test dataset $\boldsymbol{x}^{(i)}$, we can assign the corresponding value of its CV by employing the mean $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(i)})$ of the approximate posterior over the latent variables $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. The assigned CV depicts the pre-image of the atomistic configuration in the reduced CV space. It is important to note in Figure 6.8 that atomistic configurations belonging to characteristic conformations $(\alpha, \beta\text{-}1, \beta\text{-}2)$ are identified by $r_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and form clusters in the CV space. We note that the conformations β-1, β-2 interleave with each other in regions around $z_1 = 0$. An explanation for this overlap is the similarity of (ϕ, ψ) combinations in the Ramachandran plot in Figure 6.6(b). Separate from the β configurations is the cluster associated with α configurations in the CV space. The latter differ significantly with respect to the (ϕ, ψ) pairs from the β conformations.

The implied similarity of, e.g., β conformations in the CV space is in accordance with the expectations on dimensionality reduction methods. Similar atomistic—or, in general, observed—coordinates should map to similar regions in their latent lower-dimensional embedding, as emphasized in [516]. This is achieved in, e.g., multidimensional scaling [553] or isomap [518]. The presented dimensionality reduction relies solely on evaluating the force field $\mathbf{F}(\mathbf{x})$ at generated samples from $q_{\boldsymbol{\theta}}(\mathbf{x})$, without using any data, and is differentiable with respect to $\mathbf{x}$.

The hidden and lower-dimensional physically characteristic generative process is emphasized further in Figure 6.9. We illustrate *predicted* atomistic configurations $\mathbf{x}$ given the marked (circle) values of the CVs $\mathbf{z}$. The change of characteristic (ϕ, ψ)

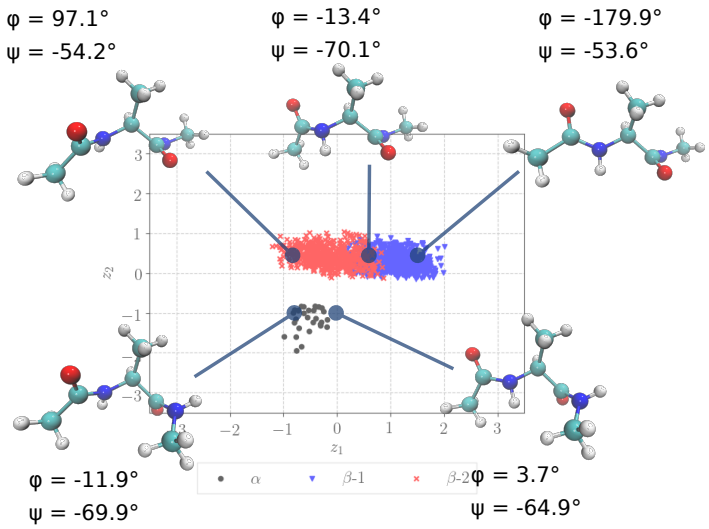


FIGURE 6.9: *Predicted* configurations $\mathbf{x}$ (including dihedral angle values) with $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})$ of $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. As one moves along the z_1 axis, we obtain for the given CVs atomistic configurations $\mathbf{x}$ reflecting the conformations α, β-1, and β-2. All rendered atomistic representations in this work were created by VMD [112].

dihedrals can be observed by moving from the red (β-2) to the blue region (β-1) in the CV space and observing the configurational change in the predicted atomistic configurations, given the indicated CVs. The depicted atomistic configurations are obtained using the input CV $\mathbf{z}$ and the mean of $q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, which is expressed as a neural network with $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})$. The probabilistic decoder $q_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is a distribution, implying that given one value of the CV, several atomistic realizations can be produced. For illustrative reasons we represent the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})$.

To obtain a better understanding of the meaning of identified CVs in terms of the dihedral angles (ϕ, ψ), we visualize them by mapping given values of $\mathbf{z}$ to atomistic configurations and compute the (ϕ, ψ) values assigned to the corresponding $\mathbf{z}$, as shown in Figure 6.10. Again, in the probabilistic model, we draw multiple atomistic realizations $\mathbf{x}$ given one CG representation $\mathbf{z}$. The realization for a given $\mathbf{z}$ fluctuates in terms of bonded vibrations rather than any change in the dihedrals (ϕ, ψ). We observe a strong correlation between (ϕ, ψ) and the CVs $\mathbf{z}$.

In addition to the visual assessment given in Figure 6.9, we show quantitatively that the structural properties of atomistic configurations generated through $q_{\boldsymbol{\theta}}(\mathbf{x})$ truly capture those of a reference trajectory at $T = 330\,\mathrm{K}$, as shown in Figure 6.11. Figure 6.11 provides histograms over bonding distances over all bonded atoms in the system. Reference statistics of bond lengths are compared with those based on generated samples of the predictive distribution $q_{\boldsymbol{\theta}}(\mathbf{x})$. Figure 6.12 provides estimated observables based on the predictive model and a reference trajectory. The observables are computed as explained in Appendix E.5.

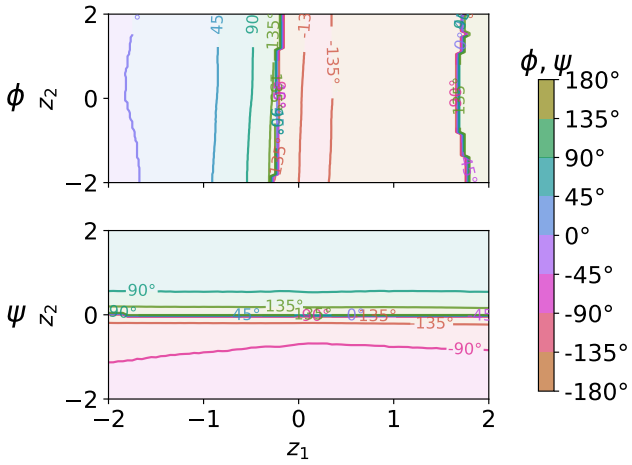


FIGURE 6.10: Predicted dihedrals (ϕ, ψ) for the latent CVs. The depicted (ϕ, ψ) values were obtained from atomistic configurations given a CV value $\mathbf{z}$ through the mean of $q_{\theta}(\mathbf{x}|\mathbf{z})$, $\boldsymbol{\mu}_{\theta}(\mathbf{z})$.

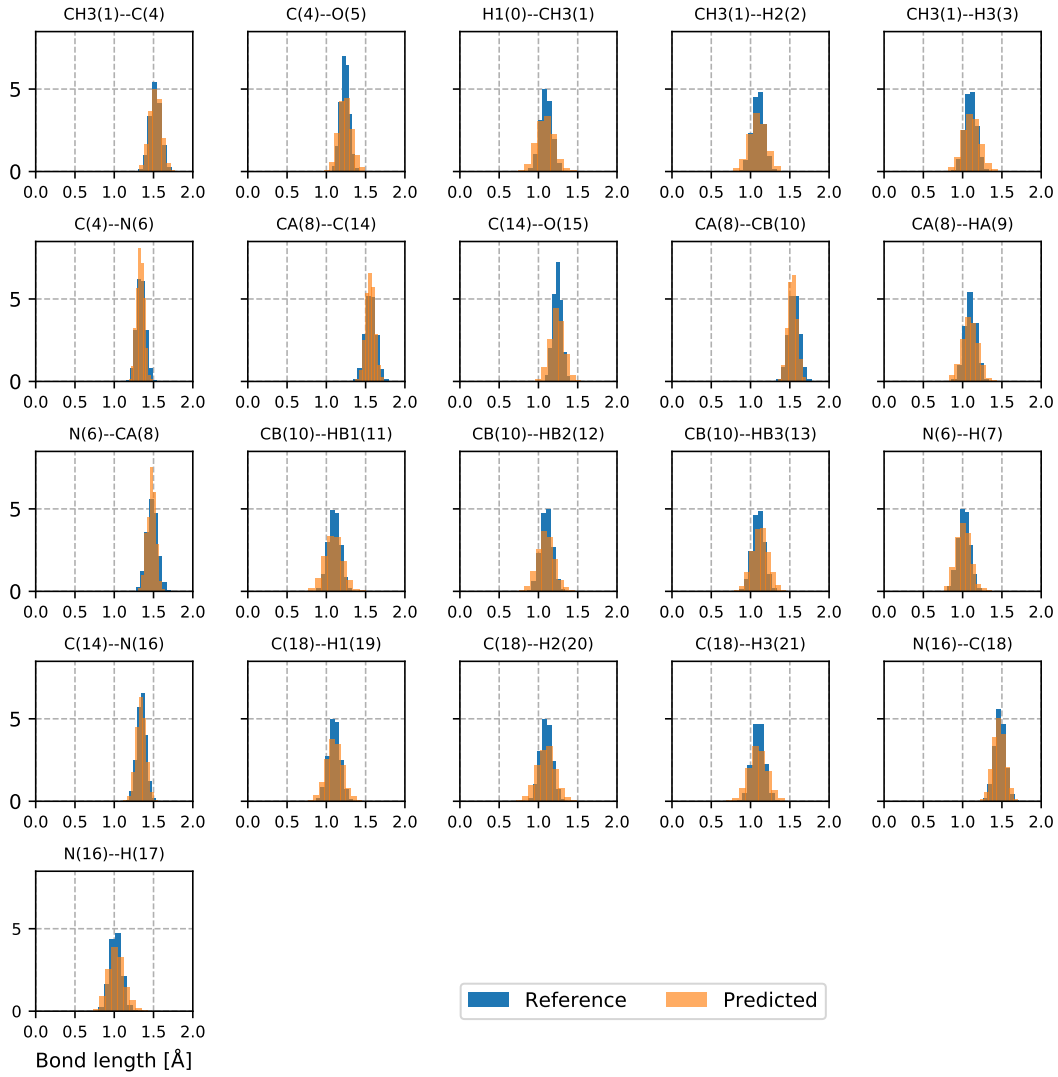
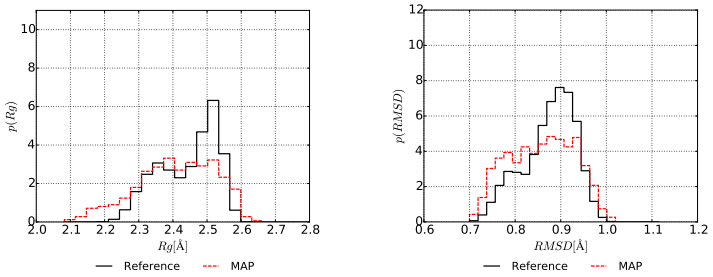


FIGURE 6.11: Bonding distance statistics. In ALA-2, bonded atoms of a reference simulation (blue) compared with histograms of the bond lengths of the *predicted* atomistic ensemble based on $q_{\boldsymbol{\theta}}(\mathbf{x})$ (semi-transparent in the foreground in orange). The titles of the subplots indicate the relevant atom names, and the corresponding atom id of the structure file of ALA-2 as provided in Appendix E.3 is shown in brackets. The physics, in the form of bonding distances, is well maintained in the generated realizations. Predictive estimates are obtained by employing $J = 2000$ samples of $q_{\boldsymbol{\theta}}(\mathbf{x})$, and the reference is based on $N = 4000$ MD snapshots.



(a) (b)

FIGURE 6.12: Predicted observables compared with reference estimates. Radius of gyration (left) and root-mean-squared deviation (right). Predictive estimates are obtained by employing $J = 2000$ samples of $q_{\boldsymbol{\theta}}(\mathbf{x})$ and the reference by $N = 4000$ MD snapshots. Observables are estimated according to Appendix E.5.

6.3 Summary and outlook

We have presented a variational approach for revealing CVs of atomistic systems that additionally yield a predictive CG model. We circumvent the need for reference data, which is supposed to provide an approximation of the target distribution $p_{\text{target}}(\mathbf{x})$. The simulation of complex biochemical systems and thus the obtained data may suffer from bias owing to insufficiently exploring all relevant conformations. Conformations are separated by high free-energy barriers, which hamper efficient exploration with brute-force MD [190]. The presented variational coarse-graining and CV discovery approach is guided by evaluating interatomic forces under the predictive distribution $q_{\boldsymbol{\theta}}(\mathbf{x})$, where sampling is computationally efficient. By embedding the atomistic force components, $q_{\boldsymbol{\theta}}(\mathbf{x})$ learns from the target distribution $p_{\text{target}}(\mathbf{x})$. We derived an upper bound on the reverse KL divergence, in which all terms are tractable, and discuss the physical underpinning of the components involved. The derived upper bound is subject to minimization with respect to all model parameters. We provide a variance-reducing gradient estimator based on reparametrization. Whereas variational approaches are known for being mode focusing, remedy provides the introduced consistent tempering scheme, alleviating the simultaneous learning of modes. We demonstrate the proposed algorithmic advances with a double well potential energy and the ALA-2 peptide. Characteristic CVs have been identified by the introduced optimization objective.

The following steps will be pursued in continuation of this work. Atomistic forces and thus gradients span many orders of magnitude at initial iterations. This could lead to numerical instabilities. Thus, we are interested in synthesizing the advantages of the forward and reverse KL divergences in the context of atomistic systems. We propose an adaptive learning scheme that may rely in its early training stages on a few data points. These are not required to reflect the whole phase space, but it is important to have a basis for learning, e.g., the structure of the atomistic system with its approximate bond lengths. This eases the problem of evaluating $\mathbf{F}(\mathbf{x})$ for the un-physical realizations that may be predicted by $q_{\boldsymbol{\theta}}(\mathbf{x})$ in an early training phase:

$$\mathcal{E} = \gamma D_{\text{KL}}\left(q_{\boldsymbol{\theta}}(\mathbf{x}) \| p_{\text{target}}(\mathbf{x})\right) + (1-\gamma) D_{\text{KL}}\left(p_{\text{target}}(\mathbf{x}) \| q_{\boldsymbol{\theta}}(\mathbf{x})\right). \tag{6.39}$$

With $\gamma \in [0,1]$ weighting the overall contribution from the reverse and forward KL divergences, we use an adaptive weight, $\gamma(k)$, which implies dependence on the current iteration k. With the proceeding learning process (increasing k), γ could increase up to $\gamma = 1$, so that we fully rely on the variational approach and thus the associated physics expressed by the potential $U(\mathbf{x})$ and forces $\mathbf{F}(\mathbf{x})$. Minimizing the above objective Equation 6.39 with respect to model distributions synthesizes the findings of this work and those of [415].

We furthermore propose the employment of the obtained $q_{\boldsymbol{\theta}}(\mathbf{x})$ for predictive purposes of systems at different temperatures. This can be achieved by obtaining the implicitly learned predictive potential expressed in terms of fine-scale coordinates at

β_{target} based on $q_{\boldsymbol{\theta}}(\mathbf{x})$:

$$U_{\boldsymbol{\theta}}^{\text{pred}}(\mathbf{x}) = -\frac{1}{\beta_{\text{target}}} \log q_{\boldsymbol{\theta}}(\mathbf{x}) + \text{const.} \tag{6.40}$$

Assuming we are interested in simulating the same system at β_{new} where $\beta_{\text{new}} \neq \beta_{\text{target}}$, we can readily provide a generalized predictive distribution for any β_{new},

$$\tilde{q}_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{-\beta_{\text{new}} U_{\boldsymbol{\theta}}^{\text{pred}}(\mathbf{x})}, \tag{6.41}$$

by employing the predictive potential $U_{\boldsymbol{\theta}}^{\text{pred}}(\mathbf{x})$ defined in Equation 6.40.

Finally, we emphasize the most relevant findings of this work. We have reformulated the identification of CVs as an optimization problem, which additionally provides a predictive CG model. CVs are revealed in the absence of any prior physical knowledge or insight, and thus in the absence of any system-dependent assumptions. Instead of relying on reference data, we employ the minimization of the reverse KL divergence and develop an inference scheme in the context of atomistic systems. Thus, the optimization is solely guided by the *evaluation* of the potential $U(\mathbf{x})$ and/or forces $\mathbf{F}(\mathbf{x})$ at samples of the predictive distributions $q_{\boldsymbol{\theta}}(\mathbf{x})$. We have also developed an adaptive tempering scheme based on findings of [205].

Chapter 7

Discussion, conclusions, and outlook

In the sequel we point out commonalities and differences among the publications and chapters of this thesis. We focus on the methodological advancements and the evoked advantages in the context of coarse-graining atomistic simulations. For suggested future work with advances and extensions of the presented work, refer to the outlook provided in each chapter or publication.

7.1 Discussion and conclusions

We have presented a novel coarse-graining approach with a focus on atomistic systems in equilibrium. Compared with most existing coarse-graining schemes, which rely on a many-to-one mapping from fine-to-coarse, we have developed strategies that follow the opposite path by proposing a probabilistic coarse-to-fine mapping. The strategy developed in this work results in a directed probabilistic model, in which coarse-grained variables serve as latent generators yielding fully atomistic coordinates through a probabilistic coarse-to-fine mapping. The latent coarse-grained variables thus depict a pre-image of the all-atom representation, where the essential physical features are expressed by the lower-dimensional[1] coarse-grained variables. The proposed probabilistic graphical model readily enables the quantification of the uncertainties that unavoidably occur during coarse-graining processes. We demonstrate the origins of the proposed approach by drawing similarities with a related coarse-graining framework that relies on the minimization of an information theoretic objective, the KL divergence from the target Boltzmann distribution to the distribution of the coarse-grained model. We generalize the relative entropy method to produce a truly predictive framework and reformulate the minimization of the KL divergence to a Bayesian likelihood-based approach. The Bayesian framework enables the consistent incorporation of functional prior information. It provides a predictive distribution that allows the probabilistic reconstruction of the microscopic all-atom description and thus the estimation of macroscopic observables with interdependencies in the fine-scale microscopic representation. Beyond obtaining

[1] Lower-dimensional in comparison with the observed atomistic coordinates.

point estimates of observables, the Bayesian framework enables a predictive posterior distribution over any quantity of interest. Posterior distributions over model parameters are propagated to a predictive posterior expressing the model's confidence. The model's confidence is reflected by error bars or credible intervals around the maximum a posteriori estimate of the corresponding observable. We show the dependency of the obtained credible intervals on the availability of training data.

Model selection is critical in all machine learning methods in any context. Here, we aim to provide a flexible model to capture all relevant features encompassed in the training data. The downside of flexible models is the resulting large set of unknown parameters, which leads to increased computational effort and susceptibility to overfitting. The probabilistic graphical model we follow encompasses two main components:

(i) a probabilistic coarse-to-fine mapping $q(\mathbf{x}|\mathbf{z})$;

(ii) the description of the latent coarse-grained variables $q(\mathbf{z})$.

This provides additional freedom with regards to the model selection. In Chapters 3 and 4, we follow the idea of having a simple but parametrized probabilistic coarse-to-fine mapping while providing flexibility in the coarse-grained description induced by a flexible coarse-grained interaction potential $U_c(\mathbf{z})$. Pushing complexity to the coarse-grained description enables us to reveal physically relevant features in the coarse-grained potential $U_c(\mathbf{z})$, such as the relevant interaction order of the coarse-grained variables or the interaction length of coarse-grained variables relevant to generating the reference atomistic data.

We explore two strategies for providing an expressive parametrization of the coarse-grained interaction potential $U_c(\mathbf{z})$:

(i) we provide initially a flexible $U_c(\mathbf{z})$ expressed by a rich set of basis functions and search for those features required for explaining atomistic reference data and deactivate all others (see Chapter 3);

(ii) we initially define a simple coarse-grained interaction potential that sequentially enriches complexity upon demand by adding optimal basis functions to maximize the anticipated benefit (see Chapter 4).

We will explain the commonalities and differences of the two approaches for providing an expressive coarse-grained interaction potential.

The first approach involves providing a rich set of basis functions associated with physically meaningful interactions, such as the interaction length, interaction order between coarse-grained variables, or different wavelengths when considering a basis of trigonometric functions. In the latter case, the wavelength relates to the frequency of changes of the interaction potential in dependencies of pairwise distances. Considering the combinatorial possibilities of the aforementioned basis functions, we obtain a flexible and expressive functional form of $U_c(\mathbf{z})$. The proposed Bayesian framework allows us to incorporate prior models that support the discovery

of sparse solutions and thus to reveal dominant features associated with the basis functions. Unnecessary features are automatically turned off if they are not relevant for describing the atomistic data. For this purpose, we employ a hierarchical prior along the lines of automatic relevance determination (ARD) [421]. This successfully reveals physically relevant features associated with the most prominent interaction lengths, interaction orders, or wavelengths when employing a set of trigonometric basis functions. The algorithm is adaptive: as training proceeds and previously unrequired features become relevant to the learning process, these can be re-activated again.

Starting with a simple coarse-grained interaction potential $U_c(\mathbf{z})$ alleviates parameter training, especially in the early stages of the learning process, as only a few model parameters need to be optimized. The model is refined sub-sequentially as training proceeds and more expressive models are required. We demonstrate this in Chapter 4 by developing an objective that expresses the anticipated gain mathematically by adding a basis function from a parametrized family of features. This is based on the squared value of the gradient of the lower bound with respect to the parameter associated with the feature to be added. We postulate that the added basis function should be placed in regions where the coarse-grained distribution $q(\mathbf{z})$ differs most from the averaged and aggregated posterior distributions, which depict the latent pre-image of the training data. The objective for identifying the most promising basis function is expressed by maximizing the squared gradient of the lower bound with respect to the parametrization of the feature functions. The proposed algorithm thus searches the feature functions and employs the one that best approaches the averaged and aggregated posterior distributions by the distribution of the coarse-grained variables $q(\mathbf{z})$ by enriching $U_c(\mathbf{z})$. We suggest adding features once the previous set of parameters has converged in terms of the lower bound on the log-likelihood. We stop adding new features when the anticipated gain drops below a certain level. In Chapter 4, we introduce a coarse-to-fine mapping that does not require a priori physical insight, in contrast to the procedure in 3. However, both methodologies are adaptive and reveal coarse-grained potentials $U_c(\mathbf{z})$ with a physical underpinning, providing insight into the physics of the fine-scale simulation.

Chapters 3 and 4 develop efficient methods, combining advances in Markov chain Monte Carlo (MC) methods [205, 554] and non-amortized variational inference [382]. The posterior distribution is obtained in both cases based on the Laplace approximation centered at the MAP parameter estimate with a covariance relying on the negative inverse of the Hessian matrix. We also provide a variational Bayesian posterior approximation of the parameters defining the coarse-to-fine mapping in Chapter 4. The approximate posterior distributions obtained through the Laplace and variational Bayes approximations reflect the uncertainties in the model parameters due to limited data. The associated implementations are embarrassingly parallelizable with regards to the expectation step.

The publications discussed earlier and presented in Chapters 3 and 4 both rely

on rather simple probabilistic coarse-to-fine mappings, balanced by incorporating rich and adaptive descriptions of the coarse-grained potential. Pushing complexity towards the coarse-grained interactions provides insight into the most prominent interactions of the coarse-grained representation. Approaches for enhanced sampling methods rely on slow coordinates of the system, called collective variables, to efficiently guide the exploration of the fine-scale configurations and overcome high free-energy barriers. This makes it necessary to identify collective variables that encode the most relevant motions in the all-atom description into a low-dimensional description. The latter implies that the coarse-grained variables are distributed according to a simple distribution, which should be compensated by a flexible mapping. Pushing flexibility towards the probabilistic coarse-to-fine mapping allows us to learn complex mapping functions that lead to a simple distribution of coarse-grained variables in the latent space associated with the most physically relevant coordinates, given the low dimension of the latent collective variables. A flexible coarse-to-fine mapping is presented in Chapter 5. We develop a methodology that reformulates the identification of collective variables by Bayesian inference based on unsupervised learning of a generative model. The focus is thus on the posterior distribution, which encodes observed atomistic configurations to latent collective variables. One can interpret the encoding component as a dictionary translating atomistic configurations to latent collective variables. Its counterpart represents a decoder transforming the values of latent variables to fully atomistic representations. In contrast to previous work presented in Chapters 3 and 4, we address this using variational Bayesian autoencoders and amortized inference (the aforementioned chapters relied on non-amortized inference). In the identification of collective variables, amortized inference has the advantage that it can learn a function for assigning latent collective variables to any input atomistic configuration. The black box variational inference approach followed in Chapters 3 and 4 treats the mean and variance (assuming Gaussian approximate posteriors) as parameters and learns an associated posterior distribution for every single datum $\mathbf{x}^{(i)}$. Amortized inference, which we employ in Chapters 5 and 6, learns functions that give rise to the mean and variance (assuming Gaussian approximate posteriors), given any input atomistic configuration $\mathbf{x}^{(i)}$. These expressive functions have been modeled with deep neural networks. As discussed earlier, overfitting can be a problem for flexible models. As in Chapters 3 and 4, we provide an adaptive learning algorithm favoring sparse solutions of the employed neural networks. Beyond obtaining sparse solutions, the ARD prior enables robust machine learning of neural network parameters ($\dim(\boldsymbol{\theta}) \approx 13\,000$) with only, e.g., 52 data points from the target distribution when $\dim(\mathbf{x}) = 60$. The latter approach identifies neural networks that cut 60% of the neurons. We have tested the proposed approach with alanine dipeptide (ALA-2) and ALA-15, revealing collective variables based on a small dataset ($N = 52$) that are highly correlated with the dihedral angles of the corresponding peptides. The dihedral angles are known to be a physically parsimonious representation of the peptide. However, we

reveal these collective variables in the absence of any physical insight or assumptions, instead reformulating the identification of collective variables as a Bayesian inference problem. Beyond the identification of collective variables, the proposed approach is predictive, as shown in Chapters 3 and 4. We provide an efficient means for the quantification of parameter uncertainties in neural networks induced by limited amounts of training data. We provide an approximate posterior for the network parameters, capturing the aforementioned uncertainty. The parameter uncertainties of the neural networks are propagated, and we provide a predictive distribution over the observables. MAP estimates are augmented by credible intervals.

All developments presented in Chapters 3, 4, and 5 rely on the minimization of the KL divergence from the target Boltzmann density to the predictive distribution of the coarse-grained model. This implies the use of data for MC approximations of the involved expectations with respect to the target density. This raises the question of whether one can assume a sufficient quantity of data representing the configuration space in all biochemically relevant systems. Instead of simulating the target Boltzmann density, e.g., with MC or molecular dynamics approaches for obtaining reference data, we start at an earlier stage, proposing an approach that circumvents the need to obtain reference data of complex atomistic systems (see Chapter 6). Instead of relying on the KL divergence from the target Boltzmann distribution to the predictive coarse-grained model, we flip the order and pursue the minimization of the reverse KL divergence from the predictive distribution of the coarse-grained model to the target. This actually implies that no expectations occur, with respect to the target distribution, that depend on model parameters. This naturally leads to an objective that does not require any reference data. We develop an approach among the lines of hierarchical variational inference [533], which provides a predictive coarse-grained model and also yields collective variables based on the reverse KL divergence. In this approach we employ, as discussed in Chapter 5, an encoding and decoding structure, where the only difference is the optimization objective. We discuss commonalities and differences between the forward (data-driven) KL divergence and the proposed reverse (variational) KL divergence objective. Direct optimization of the reverse KL divergence is in general intractable. We present an approach for obtaining a tractable upper bound on the reverse KL divergence based on [533] and provide noise reducing estimators for learning the predictive coarse-grained model. We obtain a variational amortized inference scheme, which, as in Chapter 5, enables the assignment of collective variables for any atomistic input coordinate $\mathbf{x}$. We also show that the machine learning of the model fully embeds the available physics. The derived gradient formulations involve interatomic forces assessed for samples of the predictive coarse-grained distribution $q(\mathbf{x})$. Furthermore, the objective balances maximizing the entropy of the predictive distribution $q(\mathbf{x})$ and minimizing the average potential energy of the target distribution $U_{\mathrm{f}}(\mathbf{x})$ assessed under $q(\mathbf{x})$. After developing the theoretical framework, we assess the methodology using a 'toy' double well example and, at the end of Chapter 6, the alanine dipeptide.

The major development of this work is a predictive coarse-graining framework that is predictive in terms of generating all-atom representations while consistently accounting for uncertainties induced by limited training data. These uncertainties are propagated to posterior predictive distributions expressed in credible intervals around MAP estimates of observables. We develop adaptive models that provide sparse solutions associated with the coarse-grained interaction potential. These methods follow different strategies for providing an expressive coarse-grained interaction potential by automatically deactivating unnecessary features (Chapter 3) or by refining the model describing the coarse-grained potential (Chapter 4), thereby revealing physically interpretable interaction features. Whereas Chapters 3 and 4 provide a flexible description of the coarse-grained potential, we present the opposite strategy in Chapters 5 and 6, with a simple description of the coarse-grained variables and a flexible coarse-to-fine mapping. This approach seeks to identify slow or collective variables of the system, which can be readily employed by enhanced sampling methods as the employed mapping is differentiable with respect to $\mathbf{x}$. We explore data-driven methodologies relying on limited data in Chapters 3, 4, and 5 and develop a variational approach that is based solely on assessing the fine-scale interaction potential and does not require any reference data. We use this approach for learning a predictive coarse-grained model that also yields physically relevant collective variables, as presented in Chapter 6.

7.2 Outlook

An outlook of the subsequent steps for building upon the presented methodologies is provided at the end of each chapter or publication. The following describes some potential future directions for incorporating the proposed approaches as building blocks of enhanced sampling methods.

The potential acceleration of the exploration of the configuration space with enhanced sampling methods crucially depends on the identified collective variables. A second important component for enhanced sampling methods is the identification of schemes that utilize collective variables for the construction of biasing potentials to lift deep free-energy wells.

The construction of a biasing potential to help escape sufficiently explored free-energy basins often relies on many cross-validation steps to identify the optimal parameters [291, 511, 555]. We advocate employing the lower-dimensional pre-image of the currently explored region of the potential energy surface. For this purpose, we use the posterior (encoder) $r(\mathbf{z}|\mathbf{x}^{(i)}, \boldsymbol{\phi}(\mathbf{x}^{(i)}))$ to construct a local biasing potential around the currently explored configuration $\mathbf{x}^{(i)}$. The biasing potential is constructed as follows:

$$U^{\mathbf{x}^{(i)}}_{\text{bias}}(\mathbf{x}) \propto -\log \int q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) r(\mathbf{z}|\mathbf{x}^{(i)}, \boldsymbol{\phi}) \, d\mathbf{z}. \tag{7.1}$$

The simulation of the reference potential requires the gradient of $U(\mathbf{x})$ with respect to $\mathbf{x}$, which represents the interatomic forces acting in the system. The proposed expression in Equation 7.1 can be readily incorporated, and gradient computation with respect to $\mathbf{x}$ is feasible when relying on backpropagation [548] in cases where the distributions are based on neural networks. Adding biasing forces can direct and accelerate the exploration of the configurational space.

Next, we propose an approach that combines active learning of a biasing potential with enhancing the exploration of the configurational space, and which is also predictive of observable estimations. For this purpose, we introduce a distribution encompassing the reference fine-scale interaction potential $U_{\mathrm{f}}(\mathbf{x})$ and a predictive distribution, e.g., $q(\mathbf{x}) = \int q(\mathbf{x}|\mathbf{z})q(\mathbf{z})\, d\mathbf{z}$:

$$p_{\mathrm{bias}}(\mathbf{x}) = \frac{1}{Z_p} e^{-\beta U_{\mathrm{f}}(\mathbf{x}) - \log q(\mathbf{x})}. \tag{7.2}$$

The distribution $p_{\mathrm{bias}}(\mathbf{x})$ becomes uniform when $q(\mathbf{x}) = p_{\mathrm{target}}(\mathbf{x})$. A potential optimization objective thus describes the minimization of the KL divergence from a uniform distribution to $p_{\mathrm{bias}}(\mathbf{x})$ with respect to q. Alternatively, one could employ a sequence of objectives guided by an auxiliary distribution $p_n(\mathbf{x})$, which could be close to a reference configuration $\mathbf{x}_{\mathrm{ref}}$ defined by a Gaussian $p_n(\mathbf{x}) = \mathcal{N}(\mathbf{x}_{\mathrm{ref}}, \sigma_n^2 \boldsymbol{I})$:

$$\min_q D_{\mathrm{KL}}\left(p_n(\mathbf{x}) \| p_{\mathrm{bias}}(\mathbf{x})\right). \tag{7.3}$$

Initially, we propose starting with small values for σ_n^2, which could be increased as learning proceeds (see Appendix F). As described in Chapter 6, flipping the order and employing the reverse version of Equation F.3 would circumvent the need to simulate the biased fine-scale system.

Appendix A

Methodology

A.1 Estimating credible intervals

The following provides an algorithm for estimating credible intervals. Bayesian inference algorithms, as introduced in Section 2.5, yield approximations of the posterior distributions $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ (Equation 2.66). Given this posterior distribution we can propagate uncertainties towards observables as specified in Algorithm 5.

Algorithm 5: Estimation of credible intervals

Input: Posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$, observable $a(\mathbf{x})$.

Output: Credible interval.

1 **for** *all* $i = 1, \ldots, I$ **do**

  Obtain a posterior sample:

2 $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ (Equation 2.66).

  Calculate the predictive estimate $\hat{a}(\boldsymbol{\theta}^{(i)})$ shown in Equation 2.42:

3

$$\hat{a}(\boldsymbol{\theta}^{(i)}) = \left(\int a(\mathbf{x}) p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\text{cf}}^{(i)}) p(\mathbf{z}|\boldsymbol{\theta}_{\text{c}}^{(i)}) \, d\mathbf{z} \, d\mathbf{x} \right). \tag{A.1}$$

  We approximate involved integrals with Monte Carlo methods.

4 **Compute desired intervals by employing the obtained samples $\hat{a}(\boldsymbol{\theta}^{(1\ldots I)})$.**

Please note that even in cases with symmetric posterior distributions over the model parameters $p(\boldsymbol{\theta}|\mathbf{x}^{\mathcal{D}_N})$ (Equation 2.66), the obtained credible intervals may not exhibit symmetry around the observable associated with $\boldsymbol{\theta}_{\text{MAP}}$, $\hat{a}(\boldsymbol{\theta}_{\text{MAP}})$.

Appendix B

Predictive coarse-graining

Reproduced from

M. Schöberl, N. Zabaras, P.-S. Koutsourelakis.
"Predictive coarse-graining".
In: *Elsevier Journal of Computational Physics* 333 (2017), pp. 49-77.

with the permission of ELSEVIER.

Journal of Computational Physics 333 (2017) 49–77

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Predictive coarse-graining

Markus Schöberl [a], Nicholas Zabaras [b,c], Phaedon-Stelios Koutsourelakis [a,*]

[a] *Continuum Mechanics Group, Technical University of Munich, Boltzmannstraße 15, 85748 Garching, Germany*
[b] *Institute for Advanced Study, Technical University of Munich, Lichtenbergstraße 2a, 85748 Garching, Germany*
[c] *Department of Aerospace and Mechanical Engineering, University of Notre Dame, 365 Fitzpatrick Hall, Notre Dame, IN 46556, USA*

ARTICLE INFO

Article history:
Received 26 May 2016
Received in revised form 28 September 2016
Accepted 14 October 2016
Available online 21 December 2016

Keywords:
Coarse-graining
Generative models
Bayesian
Uncertainty quantification
SPC/E water
Lattice systems

ABSTRACT

We propose a data-driven, coarse-graining formulation in the context of equilibrium statistical mechanics. In contrast to existing techniques which are based on a fine-to-coarse map, we adopt the opposite strategy by prescribing a *probabilistic coarse-to-fine* map. This corresponds to a directed probabilistic model where the coarse variables play the role of latent generators of the fine scale (all-atom) data. From an information-theoretic perspective, the framework proposed provides an improvement upon the relative entropy method [1] and is capable of quantifying the uncertainty due to the information loss that unavoidably takes place during the coarse-graining process. Furthermore, it can be readily extended to a fully Bayesian model where various sources of uncertainties are reflected in the posterior of the model parameters. The latter can be used to produce not only point estimates of fine-scale reconstructions or macroscopic observables, but more importantly, predictive posterior distributions on these quantities. Predictive posterior distributions reflect the confidence of the model as a function of the amount of data and the level of coarse-graining. The issues of model complexity and model selection are seamlessly addressed by employing a hierarchical prior that favors the discovery of sparse solutions, revealing the most prominent features in the coarse-grained model. A flexible and parallelizable Monte Carlo – Expectation–Maximization (MC-EM) scheme is proposed for carrying out inference and learning tasks. A comparative assessment of the proposed methodology is presented for a lattice spin system and the SPC/E water model.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Molecular dynamics simulations [2] are nowadays commonplace in physics, chemistry, and engineering and represent one of the most reliable tools in the analysis of complex processes and the design of new materials [3–5]. Direct simulations are hampered by the gigantic number of degrees of freedom, complex, potentially long-range and high-order interactions, and as a result, are limited to small spatio-temporal scales with current and foreseeable computational resources.

An approach towards making complex simulations practicable over extended time and space scales is coarse-graining (CG) [6]. Coarse-graining methods attempt to summarize the atomistic detail in much fewer degrees of freedom which in turn lead to shorter simulation times, with potentially larger time-steps and enable the analysis of systems that occupy

* Corresponding author.
E-mail addresses: m.schoeberl@tum.de (M. Schöberl), nzabaras@gmail.com (N. Zabaras), p.s.koutsourelakis@tum.de (P.-S. Koutsourelakis).
URLs: http://www.zabaras.com (N. Zabaras), http://www.contmech.mw.tum.de (P.-S. Koutsourelakis).

http://dx.doi.org/10.1016/j.jcp.2016.10.073
0021-9991/© 2016 Elsevier Inc. All rights reserved.

larger spatial domains. Furthermore, from a reductionist's point of view, they can provide insight into the fundamental components or processes associated with the macroscopic behavior and properties of molecular ensembles.

A systematic strategy towards coarse-graining is offered in the context of free-energy computation methods [7,8]. Nevertheless, their primary goal is to escape deep, free-energy wells and are generally limited to a relatively small number of CG variables. A mathematically rigorous approach to coarse-graining lattice systems and a rich set of multi-level, adaptive algorithms for equilibrium and nonequilibrium settings, has been developed in [9–14]. Inversion-based methods such as the Direct or Iterative Boltzmann Inversion [15,16] and Inverse Monte Carlo [17], represent a popular strategy where the parameters of the CG model are adjusted to reproduce macroscopic observables [18]. Molecular Renormalization Group CG [19] is founded upon the ideas first presented in [20] and is based on matching correlators, obtained from atomistic and coarse-grained simulations, for observables that explicitly enter the coarse-grained Hamiltonian. Data-driven, variational CG methods such as Multiscale CG [21,22], Relative Entropy [1], Ultra GG [23], offer a rigorous way of learning CG models by approximating the Potential of Mean Force (PMF) [24] with respect to the CG variables on the basis of appropriate functionals.

It is obvious that unless there are known redundancies in the all-atom or fine-grained (FG) description, any coarse-graining scheme will result in information loss [25,26]. A manifestation of this can be seen if one attempts to reconstruct the microscopic, FG configurations from the CG states [27,28]. Discrepancies will appear not only because the CG statistics are not captured correctly, but because the CG variables do not encode all the details needed to reproduce the FG picture. Despite this, predictions generated by existing CG schemes are always in the form of *point estimates* that do not reflect any of the predictive uncertainty which the aforementioned information loss induces. It is also reasonable to expect that this information loss increases the larger the difference between the dimension of fine and coarse descriptions becomes. Nevertheless given two competing CG descriptions of the same dimension, it is unlikely that both will capture the FG picture equally well. The discovery of a good set of CG variables (analogous to finding good reaction coordinates or collective variables in free energy computations [29]) is, on one hand, a function of the macroscopic quantities of interest but more importantly of the complex structure of inter-dependencies in the FG model.

The starting point of all CG schemes is the prescription of the coarse variables through a many-to-one, *fine-to-coarse* map. Such maps are dictated by the analysis objectives but also by physical insight on which FG features might be important [30]. For example several atoms/molecules can be lumped into a single, effective, pseudo-molecule with coordinates defined by considering the center of mass. A central component of the present work is the implicit definition of the CG variables through a *coarse-to-fine* map. This is achieved by a *probabilistic generative model* that treats the CG degrees of freedom as latent variables and explicitly quantifies the uncertainty in the reconstruction of the FG states from the CG description. The model is complemented with a distribution for the CG variables. Both densities are parametrized and the optimal values are determined on the basis of an information-theoretic objective (e.g. minimizing a Kullback–Leibler divergence as in [1]) which is shown to be a special case of a more general, Bayesian framework. The latter offers a critical advantage over existing techniques as it enables the prediction of macroscopic observables not only in the form of point estimates, but by providing whole distributions. These reflect the uncertainty due the aforementioned information loss as well as the fact that finite amounts of training data were used.

The emphasis on this amplified predictive ability of the proposed framework is the reason behind the title chosen for the present paper *predictive coarse-graining* (PCG). The Bayesian framework advocated offers a superior setting for model selection. We make use of hierarchical prior models that promote the discovery of a sparse set of features in the aforementioned model components. This enables the search to be carried out over a very large set of feature functions for the CG potential which naturally amplifies the expressivity of the model [30]. We note that a Bayesian framework towards uncertainty quantification for force field parameters in molecular dynamics was introduced in [31,32]. Other Bayesian formulations of coarse-graining problems using macroscopic observables were presented in [33,34] where also the issues of model calibration and validation were discussed.

The structure of the rest of the paper is as follows. Section 2 presents the basic model components, compares them with other CG schemes (primarily the relative entropy method), provides details on the exponential family of distributions employed for which uniqueness of solution can be proven and discusses in detail algorithmic and computational aspects. Numerical evidence of the capabilities of the proposed framework is provided in Section 3 where coarse-graining efforts for an Ising lattice system as well as for the SPC/E water model are documented. In all numerical examples, we report results on the *predictive uncertainty* as a function of the level of coarse graining, and the amount of data available. Finally, Section 4, summarizes the main contributions and discusses natural extensions of the proposed framework.

2. Methodology

This section introduces the notational conventions adopted and presents the proposed modeling and computational frameworks. We frequently draw comparisons with the relative entropy method introduced in [1] and further expanded and studied in [35,36] in order to shed light on the aspects related to information loss and to emphasize the need for quantifying the resulting uncertainty in the predictions.

2.1. Equilibrium statistical mechanics

We consider molecular ensembles in equilibrium described by an n_f-dimensional vector denoted by $\boldsymbol{x} \in \mathcal{M}_\mathrm{f} \subset \mathbb{R}^{n_\mathrm{f}}$. This generally consists of the coordinates of the atoms which follow the Boltzmann–Gibbs density[1]:

$$p_\mathrm{f}(\boldsymbol{x}|\beta) = \frac{\exp\{-\beta U_\mathrm{f}(\boldsymbol{x})\}}{Z_\mathrm{f}(\beta)}, \tag{1}$$

where $U_\mathrm{f}(\boldsymbol{x})$ is the all-atom (fine-grained) potential, $\beta = \frac{1}{k_b T}$ where k_b is the Boltzmann constant and T is the temperature, and $Z_\mathrm{f}(\beta)$ is the normalization constant (partition function) given by:

$$Z_\mathrm{f}(\beta) = \int_{\mathcal{M}_\mathrm{f}} \exp\{-\beta U_\mathrm{f}(\boldsymbol{x})\} d\boldsymbol{x}. \tag{2}$$

In the following, we assume that the temperature T (or equivalently β) is constant as it is commonly done in coarse-graining literature, even though it is generally of interest to derive coarse-grained descriptions that are suitable for all (or at least a wide range) of temperatures [30]. In this setting and in order to simplify the notation, we drop the temperature dependence.

If $a(\boldsymbol{x}) : \mathcal{M}_\mathrm{f} \to \mathbb{R}$ denotes an observable (e.g. magnetization in Ising models), then the corresponding macroscopic properties can be computed as an expectation with respect to $p_\mathrm{f}(\boldsymbol{x})$ as follows:

$$\mathbb{E}_{p_\mathrm{f}(\boldsymbol{x})}[a(\boldsymbol{x})] = \int_{\mathcal{M}_\mathrm{f}} a(\boldsymbol{x}) p_\mathrm{f}(\boldsymbol{x}) d\boldsymbol{x}. \tag{3}$$

Such expectations are (approximately) computed using long and cumbersome simulations as explained in the introduction e.g. by a long MCMC run [37]. Our goal is two-fold. Firstly, to construct a coarse-grained description of the system that would be easier and faster to simulate, and secondly to use this in order to predict expectations of any observable as in Eq. (3). A distinguishing aspect of the proposed PCG framework is that we also compute quantitative metrics of the predictive uncertainty in those estimates. At a third level, one would also want the coarse-grained description to provide a decomposition of the original, all-atom ensemble into physically interpretable terms and interactions. We defer such a discussion on how the proposed model can achieve this goal for the conclusions.

We denote by $\boldsymbol{X}$ the coarse-grained variables and assume that they take values in $\mathcal{M}_\mathrm{c} \subset \mathbb{R}^{n_\mathrm{c}}$. It is obviously desirable that $n_\mathrm{c} \ll n_\mathrm{f}$. Let also $U_\mathrm{c}(\boldsymbol{X})$ denote the potential associated with $\boldsymbol{X}$ and $p_\mathrm{c}(\boldsymbol{X})$ the corresponding density:

$$p_\mathrm{c}(\boldsymbol{X}) = \frac{\exp\{-\beta U_\mathrm{c}(\boldsymbol{X})\}}{Z_\mathrm{c}}, \tag{4}$$

with the normalization constant,

$$Z_\mathrm{c} = \int_{\mathcal{M}_\mathrm{c}} \exp\{-\beta U_\mathrm{c}(\boldsymbol{X})\} d\boldsymbol{X}. \tag{5}$$

In existing coarse-graining formulations, the coarse variables $\boldsymbol{X}$ are defined using a restriction, fine-to-coarse map $\mathcal{R} : \mathcal{M}_\mathrm{f} \to \mathcal{M}_\mathrm{c}$ i.e. $\boldsymbol{X} = \mathcal{R}(\boldsymbol{x})$. As this is generally a many-to-one map, it is not invertible [36]. If the observables of interest actually depend on $\boldsymbol{X}$ i.e. if $a(\boldsymbol{x}) = A(\mathcal{R}(\boldsymbol{x})) = A(\boldsymbol{X})$, then one can readily show that it suffices that $p_\mathrm{c}(\boldsymbol{X})$ is equal to the marginal of $\boldsymbol{X}$ with respect to $p_\mathrm{f}(\boldsymbol{x})$, or equivalently that $U_\mathrm{c}(\boldsymbol{X}) = U_\mathrm{c}^\mathrm{opt}(\boldsymbol{X})$ where:

$$U_\mathrm{c}^\mathrm{opt}(\boldsymbol{X}) = -\beta^{-1} \log \int \delta(\boldsymbol{X} - \mathcal{R}(\boldsymbol{x}))\ p_\mathrm{f}(\boldsymbol{x}) d\boldsymbol{x}. \tag{6}$$

That is the coarse-scale potential $U_\mathrm{c}(\boldsymbol{x})$ coincides with the potential of mean-force of $\boldsymbol{X}$. This is a consequence of the following equalities:

$$\begin{aligned}\mathbb{E}_{p_\mathrm{f}}[a] &= \int_{\mathcal{M}_\mathrm{f}} a(\boldsymbol{x})\ p_\mathrm{f}(\boldsymbol{x})\ d\boldsymbol{x} \\ &= \int_{\mathcal{M}_\mathrm{f}} A(\mathcal{R}(\boldsymbol{x}))\ p_\mathrm{f}(\boldsymbol{x})\ d\boldsymbol{x}\end{aligned}$$

[1] In the following, we assume all probability measures are absolutely continuous with the Lebesgue measure and therefore work exclusively with the corresponding probability density functions.

$$
\begin{aligned}
&= \int_{\mathcal{M}_\mathrm{f}} \left(\int_{\mathcal{M}_\mathrm{c}} A(\boldsymbol{X}) \delta(\boldsymbol{X} - \mathcal{R}(\boldsymbol{x}))\, d\boldsymbol{X} \right) p_\mathrm{f}(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{\mathcal{M}_\mathrm{c}} A(\boldsymbol{X}) \left(\int_{\mathcal{M}_\mathrm{f}} \delta(\boldsymbol{X} - \mathcal{R}(\boldsymbol{x}))\, p_\mathrm{f}(\boldsymbol{x})\, d\boldsymbol{x} \right) d\boldsymbol{X} \\
&= \int_{\mathcal{M}_\mathrm{c}} A(\boldsymbol{X})\, p_c(\boldsymbol{X})\, d\boldsymbol{X}.
\end{aligned}
$$

Nevertheless, even if one is able to compute or approximate sufficiently well $U_\mathrm{c}^{\mathrm{opt}}(\boldsymbol{X})$, there is no guarantee that expectations of other observables that do not solely depend on $\boldsymbol{X}$ can be accurately computed. Consistent reconstructions of the all-atom configurations $\boldsymbol{x}$, given $\boldsymbol{X}$ samples from $p_\mathrm{c}(\boldsymbol{X})$, can be obtained from the conditional:

$$
p_{\mathcal{R}}(\boldsymbol{x}|\boldsymbol{X}) = \frac{\delta(\boldsymbol{X} - \mathcal{R}(\boldsymbol{x}))}{Z_{\mathcal{R}}(\boldsymbol{X})}, \tag{7}
$$

i.e. the uniform density on the manifold in $\mathcal{M}_\mathrm{f}$ implied by the map $\mathcal{R}$,[2] where:

$$
Z_{\mathcal{R}}(\boldsymbol{X}) = \int \delta(\boldsymbol{X} - \mathcal{R}(\boldsymbol{x}))\, d\boldsymbol{x}. \tag{8}
$$

Given a coarse-grained potential U_c (not necessarily the optimal as in Eq. (6)) and the density $p_\mathrm{c}(\boldsymbol{X})$ in Eq. (4), the corresponding reconstruction density of the all-atom description consistent with the map $p_{\mathcal{R}}(\boldsymbol{x}|\boldsymbol{X})$ (Eq. (7)) is given by:

$$
\begin{aligned}
p_{\mathcal{R}}(\boldsymbol{x}) &= \int p_{\mathcal{R}}(\boldsymbol{x}|\boldsymbol{X})\, p_\mathrm{c}(\boldsymbol{X})\, d\boldsymbol{X} \\
&= \int \frac{\delta(\boldsymbol{X} - \mathcal{R}(\boldsymbol{x}))}{Z_{\mathcal{R}}(\boldsymbol{X})}\, p_\mathrm{c}(\boldsymbol{X})\, d\boldsymbol{X} \\
&= \frac{p_\mathrm{c}(\mathcal{R}(\boldsymbol{x}))}{Z_{\mathcal{R}}(\mathcal{R}(\boldsymbol{x}))}.
\end{aligned} \tag{9}
$$

We note that in the context of the relative entropy method [1], which like ours, is data-driven and has an information-theoretic underpinning, the goal is to identify the U_c (within a certain class) that brings $p_{\mathcal{R}}(\boldsymbol{x})$ (Eq. (9)) as close as possible to the reference, FG density $p_\mathrm{f}(\boldsymbol{x})$ (Eq. (1)). For that purpose the Kullback–Leibler (KL) divergence [39] $\mathrm{KL}(p_\mathrm{f}(\boldsymbol{x})||p_{\mathcal{R}}(\boldsymbol{x}))$ is employed as the objective which, based on Eq. (9), is given by:

$$
\begin{aligned}
0 \leq \mathrm{KL}(p_\mathrm{f}(\boldsymbol{x})||p_{\mathcal{R}}(\boldsymbol{x})) &= -\int p_\mathrm{f}(\boldsymbol{x}) \log \frac{p_{\mathcal{R}}(\boldsymbol{x})}{p_\mathrm{f}(\boldsymbol{x})}\, d\boldsymbol{x} \\
&= -\mathbb{E}_{p_\mathrm{f}(\boldsymbol{x})}[\log\, p_\mathrm{c}(\mathcal{R}(\boldsymbol{x})] + \mathbb{E}_{p_\mathrm{f}(\boldsymbol{x})}[\log Z_{\mathcal{R}}(\mathcal{R}(\boldsymbol{x}))] - H(p_\mathrm{f}),
\end{aligned} \tag{10}
$$

where $H(p_\mathrm{f})$ is the entropy of $p_\mathrm{f}(\boldsymbol{x})$, which is independent of U_c and can be ignored in the minimization. As it has been identified in several investigations [35,36,38], while the first term can be reduced by adjusting U_c (it can be shown that the minimum is attained when $U_\mathrm{c}(\boldsymbol{X}) = U_\mathrm{c}^{\mathrm{opt}}(\boldsymbol{X})$), the second term is fixed once the restriction map $\mathcal{R}$ that defines the coarse-grained variables has been selected. It represents a constant penalty reflecting the information loss that takes place due to the coarse-grained (and generally lower-dimensional) description adopted. Our goal is to reduce this component of information loss.

2.2. Probabilistic generative model

We propose a *probabilistic, generative model* [40] in which the coarse description is treated as a latent (hidden) state. In particular, we define a *joint* density $\bar{p}(\boldsymbol{X}, \boldsymbol{x})$ for $\boldsymbol{X}$ and $\boldsymbol{x}$ as follows:

$$
\bar{p}(\boldsymbol{X}, \boldsymbol{x}) = p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X})\, p_\mathrm{c}(\boldsymbol{X}). \tag{11}
$$

This consists of two components i.e.:

(i) a density $p_\mathrm{c}(\boldsymbol{X})$ describing the statistics of the coarse-grained description $\boldsymbol{X}$,
(ii) a **probabilistic, coarse-to-fine mapping** implied by the conditional density $p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X})$.

[2] In [38] this is further generalized by introducing an additional, weighting density.

We discuss the form and parametrization of the aforementioned densities in the sequel. We emphasize at this stage the different definition of the coarse-grained variables as latent generators that give rise to the observables through the probabilistic *lifting* operator implied by p_{cf} [9], in contrast to the restriction operators employed in other schemes explained previously. Such mappings can take various forms (e.g. local or global, linear or nonlinear) and can be extended to many hierarchical levels, as it will be shown. Understanding the meaning of the latent variables can only be done through the prism of this generative mapping. According to this, each FG configuration $\boldsymbol{x}^{(i)}$ is generated as follows:

- Draw a CG configuration $\boldsymbol{X}^{(i)}$ from $p_{\mathrm{c}}(\boldsymbol{X})$.
- Draw $\boldsymbol{x}^{(i)}$ from $p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}^{(i)})$.

As we will show, an advantage of the proposed framework is that it readily provides a (predictive) probability density for the observables of interest. The marginal density of the FG description $\boldsymbol{x}$ is given from Eq. (11) by integrating out $\boldsymbol{X}$:

$$\bar{p}_{\mathrm{f}}(\boldsymbol{x}) = \int_{\mathcal{M}_{\mathrm{c}}} p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X})\; p_{\mathrm{c}}(\boldsymbol{X}) d\boldsymbol{X}. \tag{12}$$

Suppose the aforementioned component densities are parametrized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathrm{c}}, \boldsymbol{\theta}_{\mathrm{cf}})$ i.e. $p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}})$ and $p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})$, and we attempt to minimize the KL-divergence between the reference density $p_{\mathrm{f}}(\boldsymbol{x})$ and the marginal $\bar{p}_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{\theta})$ implied by the generative model proposed :

$$\begin{aligned} \mathrm{KL}(p_{\mathrm{f}}(\boldsymbol{x})||\bar{p}_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{\theta})) &= -\int_{\mathcal{M}_{\mathrm{f}}} p_{\mathrm{f}}(\boldsymbol{x}) \log \frac{\bar{p}_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{\theta})}{p_{\mathrm{f}}(\boldsymbol{x})} d\boldsymbol{x} \\ &= -\int p_{\mathrm{f}}(\boldsymbol{x}) \log \bar{p}_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{\theta})\; d\boldsymbol{x} + \int p_{\mathrm{f}}(\boldsymbol{x}) \log p_{\mathrm{f}}(\boldsymbol{x}) d\boldsymbol{x}. \end{aligned} \tag{13}$$

This is equivalent to *maximizing* $\int p_{\mathrm{f}}(\boldsymbol{x}) \log \bar{p}_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{\theta})\; d\boldsymbol{x}$ which, given samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ from $p_{\mathrm{f}}(\boldsymbol{x})$, is approximated by the *log-likelihood* of $\bar{p}_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{\theta})$[3]:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log \bar{p}_{\mathrm{f}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \left(\int p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}) d\boldsymbol{X}^{(i)} \right). \end{aligned} \tag{14}$$

We note in the expression above that we associate a latent, coarse configuration $\boldsymbol{X}^{(i)}$ to each sample $\boldsymbol{x}^{(i)}$ which is effectively its pre-image. More importantly, the objective in the aforementioned expression accounts for both the density of the coarse-grained description as well as the reconstruction (lifting) of the all-atom configuration from the (latent) coarse-grained one. Maximizing $\mathcal{L}(\boldsymbol{\theta})$ naturally leads to the Maximum Likelihood estimate $\boldsymbol{\theta}_{\mathrm{MLE}}$.

Furthermore the interpretation of the objective as the log-likelihood makes the progression into Bayesian formulations much more straightforward. If for example we define a prior density $p(\boldsymbol{\theta})$ then maximizing:

$$\arg\max_{\boldsymbol{\theta}} \; \{\mathcal{L}(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\}, \tag{15}$$

is equivalent to obtaining a Maximum a Posteriori (MAP) estimate $\boldsymbol{\theta}_{\mathrm{MAP}}$ [41]. The next step from point estimates for the model parameters is of course obtaining the full posterior $p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})$ using Bayes formula as:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)}) &\propto p(\boldsymbol{x}^{(1:N)}|\boldsymbol{\theta})\; p(\boldsymbol{\theta}) \\ &\propto e^{\mathcal{L}(\boldsymbol{\theta})}\; p(\boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{N} \left(\int p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}) d\boldsymbol{X}^{(i)} \right) p(\boldsymbol{\theta}). \end{aligned} \tag{16}$$

The aforementioned relationship can be concretely represented in the form of a directed graphical model as depicted in Fig. 1.

We discuss a strategy for approximating this posterior in the next subsections. It is more important to emphasize at this stage that given this posterior, we can produce not just point estimates of the expectation of any observable $a(\boldsymbol{x})$, but also

[3] This result can be obtained (up to $1/N$) by substituting $p_{\mathrm{f}}(\boldsymbol{x})$ in Eq. (13) by the empirical measure $\frac{1}{N}\sum_{i=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$. The likelihood of N samples drawn from $p_{\mathrm{f}}(\boldsymbol{x})$ is trivially $\prod_{i=1}^{N} \bar{p}_{\mathrm{f}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$.

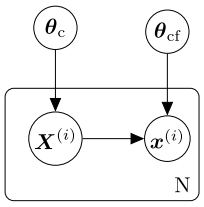


Fig. 1. Probabilistic graphical model representation.

compute its predictive posterior. For that purpose we make use of the predictive posterior $p(\boldsymbol{x}|\boldsymbol{x}^{(1:N)})$ of our model which is determined by marginalizing the latent variables $\boldsymbol{X}$ and the model parameters $\boldsymbol{\theta}$:

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{x}^{(1:N)}) &= \int p(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})\, d\boldsymbol{X}d\boldsymbol{\theta} \\ &= \int p(\boldsymbol{x}, \boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{x}^{(1:N)})\; p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})\, d\boldsymbol{X}d\boldsymbol{\theta}. \end{aligned} \tag{17}$$

By replacing the joint density with the proposed generative model in Eq. (11), the predictive posterior $p(\boldsymbol{x}|\boldsymbol{x}^{(1:N)})$ becomes:

$$p(\boldsymbol{x}|\boldsymbol{x}^{(1:N)}) = \int p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}})\; p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})\, d\boldsymbol{X}d\boldsymbol{\theta}. \tag{18}$$

The latter can be used in place of the FG distribution $p_{\mathrm{f}}(\boldsymbol{x})$ in Eq. (3), to obtain approximations to the expectation of any observable $a(\boldsymbol{x})$ as follows:

$$\begin{aligned} \mathbb{E}_{p_{\mathrm{f}}(\boldsymbol{x})}[a(\boldsymbol{x})] &\approx \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{x}^{(1:N)})}[a(\boldsymbol{x})] \\ &= \int a(\boldsymbol{x})\; p(\boldsymbol{x}|\boldsymbol{x}^{(1:N)})\, d\boldsymbol{x} \\ &= \int a(\boldsymbol{x}) \left(\int p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}})\; p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})\, d\boldsymbol{X}\, d\boldsymbol{\theta} \right) d\boldsymbol{x} \\ &= \int \underbrace{\left(\int a(\boldsymbol{x})\; p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}})\, d\boldsymbol{X}\, d\boldsymbol{x} \right)}_{\hat{a}(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})\, d\boldsymbol{\theta} \\ &= \int \hat{a}(\boldsymbol{\theta})\; p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})\, d\boldsymbol{\theta}. \end{aligned} \tag{19}$$

The approximation in the first line reflects the quality of the model as well as the uncertainty arising from the finite data $\boldsymbol{x}^{(1:N)}$ that were used to calibrate it. This derivation suggests that $\hat{a}(\boldsymbol{\theta})$ represents the predictive estimate of the expectation of $a(\boldsymbol{x})$ for a given value $\boldsymbol{\theta}$ of the model's parameters. Averaging over the posterior of the latter provides the expected (a posteriori) value of this quantity. More importantly though by propagating the (posterior) uncertainty of $\boldsymbol{\theta}$ through $\hat{a}(\boldsymbol{\theta})$, one can readily obtain the predictive distribution of the observable. In the numerical examples we frequently plot such posterior statistics, usually in the form of credible intervals (see also A.1). Point estimates can be easily recovered if the analyst wishes to do so by employing for example the MAP (or MLE) estimate $\boldsymbol{\theta}_{\mathrm{MAP}}$ in the aforementioned equation i.e. if $p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)}) \equiv \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})$.

2.3. Inference and learning (point estimates)

This section is concerned with the computational aspects of training the proposed model. We pay particular attention to distributions in the exponential family for which the concavity of the maximum-likelihood problem can be analytically shown. Furthermore, we discuss strategies for parallelizing these tasks and improving the computational efficiency. We finally discuss particular prior specifications that are suitable for sparse feature recovery and model selection.

We begin our discussion with a strategy for obtaining point estimates for the model parameters $\boldsymbol{\theta}$ by maximizing the log-likelihood (or the log-posterior) as given in Eq. (14) (or Eq. (15)). The difficulty in the optimization problem stems from the intractability of the log-likelihood due to the integration with respect to the latent variables $\boldsymbol{X}^{(i)}$ (except for trivial cases for $p_{\mathrm{c}}, p_{\mathrm{cf}}$). To address this we employ an *Expectation–Maximization* (EM) scheme [42,43] where MCMC is used to approximate the E-step (MCEM) [44] and stochastic approximations to handle the Monte Carlo noise in the gradient estimates of the M-Step [45,46]. The EM algorithm allows the maximization of the log-likelihood by circumventing the need for repeated evaluations of the aforementioned intractable integrals and normalization constants. To motivate the derivation, we note that for an arbitrary set of densities $q_i(\boldsymbol{X}^{(i)})$ we can construct lower bounds, denoted by $\mathcal{F}^{(i)}(q_i(\boldsymbol{X}^{(i)}),\ \boldsymbol{\theta})$, *for each term* in the sum that makes up the log-likelihood as follows:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log\left(\int p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}) d\boldsymbol{X}^{(i)}\right) \\
&= \sum_{i=1}^{N} \log\left(\int \frac{p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}})}{q_i(\boldsymbol{X}^{(i)})} q_i(\boldsymbol{X}^{(i)})\, d\boldsymbol{X}^{(i)}\right) \\
&\geq \sum_{i=1}^{N} \underbrace{\left(\int q_i(\boldsymbol{X}^{(i)}) \log \frac{p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}})}{q_i(\boldsymbol{X}^{(i)})}\, d\boldsymbol{X}^{(i)}\right)}_{:=\mathcal{F}^{(i)}(q_i(\boldsymbol{X}^{(i)}),\,\boldsymbol{\theta})} \\
&= \sum_{i=1}^{N} \mathcal{F}^{(i)}(q_i(\boldsymbol{X}^{(i)}),\,\boldsymbol{\theta}) \\
&= \mathcal{F}(\boldsymbol{q}(\boldsymbol{X}),\,\boldsymbol{\theta}),
\end{aligned}
\tag{20}
$$

where $\boldsymbol{q}(\boldsymbol{X}) = \prod_{i=1}^{N} q_i(\boldsymbol{X}^{(i)})$, and the result in the third step is a consequence of Jensen's inequality. We note that the optimal $q_i^{\mathrm{opt}}(\boldsymbol{X}^{(i)})$ for each of the aforementioned terms is:

$$
q_i^{\mathrm{opt}}(\boldsymbol{X}^{(i)}) = q_i(\boldsymbol{X}^{(i)}|\boldsymbol{x}^{(i)},\boldsymbol{\theta}) \propto p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}),
\tag{21}
$$

i.e. the conditional posterior of the latent variables $\boldsymbol{X}^{(i)}$ given $\boldsymbol{x}^{(i)}$ and $\boldsymbol{\theta}$. This is optimal in the sense that the inequality becomes an equality [41] i.e.:

$$
\mathcal{F}^{(i)}(q_i^{\mathrm{opt}}(\boldsymbol{X}^{(i)}),\,\boldsymbol{\theta}) = \log\left(\int p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{\mathrm{cf}})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}) d\boldsymbol{X}^{(i)}\right).
\tag{22}
$$

All other q_i's lead to suboptimal schemes that fall under the category of Variational Bayesian Expectation–Maximization (VB-EM, [47]). More importantly, the aforementioned derivation suggests an iterative algorithm where one alternates (until convergence) between the following two steps, i.e. at each iteration t:

E-step: Given the current estimate of $\boldsymbol{\theta} \equiv \boldsymbol{\theta}^{(t)}$, evaluate:

$$
\mathcal{F}(\boldsymbol{q}^{\mathrm{opt},\,t}(\boldsymbol{X}),\,\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{N} \mathcal{F}^{(i)}(q_i^{\mathrm{opt},\,t}(\boldsymbol{X}^{(i)}),\,\boldsymbol{\theta}^{(t)}),
\tag{23}
$$

where $q_i^{\mathrm{opt},\,t}$ is given in Eq. (21) for $\boldsymbol{\theta} \equiv \boldsymbol{\theta}^{(t)}$.

M-step: Given the current $q_i^{\mathrm{opt},\,t}(\boldsymbol{X}^{(i)})$, find:

$$
\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathcal{F}^{(i)}(q_i^{\mathrm{opt},\,t}(\boldsymbol{X}^{(i)}),\,\boldsymbol{\theta}^{(t)}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(\int q_i^{\mathrm{opt},\,t}(\boldsymbol{X}^{(i)}) \log\left(p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{\mathrm{cf}}^{(t)})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}^{(t)})\right) d\boldsymbol{X}^{(i)}\right).
\end{aligned}
\tag{24}
$$

We discuss in detail each of the two steps.

- The E-step of the algorithm requires computing expectations with respect to the intractable distributions in Eq. (21). As it can be seen in Eq. (24) only the terms in $\mathcal{F}^{(i)}$ that depends on $\boldsymbol{\theta}$ needs to be computed which we approximate by a Monte Carlo estimator:

$$
\begin{aligned}
&\int q_i^{\mathrm{opt},\,t}(\boldsymbol{X}^{(i)}) \log\left(p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{\mathrm{cf}}^{(t)})\; p_{\mathrm{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}^{(t)})\right) d\boldsymbol{X}^{(i)} \approx \\
&\qquad \approx \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\log p_{\mathrm{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}_j^{(i)},\boldsymbol{\theta}_{\mathrm{cf}}^{(t)})\; p_{\mathrm{c}}(\boldsymbol{X}_j^{(i)}|\boldsymbol{\theta}_{\mathrm{c}}^{(t)})\right).
\end{aligned}
\tag{25}
$$

The m_t samples used at each iteration t are drawn using MCMC from $q_i^{\mathrm{opt},\,t}(\boldsymbol{X}^{(i)})$. Compared to i.i.d. Monte Carlo samples, the use of MCMC introduces theoretical complications with regards to the stability and the error in the approximation [48,49]. A recent treatment of the convergence conditions for such schemes is contained in [50]. The obvious

error source arises from the bias in the MCMC samples which are *approximately* distributed according to the target density. In addition the samples generated are correlated. Such errors can be subdued by increasing the sample size m_t. Heuristically speaking, at the first few iterations t, even a crude estimate of the objective generally suffices to drive the parameter $\boldsymbol{\theta}$-updates toward the region of interest. As the EM iterations proceed, the number of samples should increase in order to zoom-in at the optimum and minimize the oscillatory behavior due to the noise in the estimates. Several strategies have been proposed to optimize m_t or even devise an automatic schedule by making use of error estimates [51–54]. In this work, we used a constant sample size i.e. $m_t = m, \forall t$ that we report in the numerical examples. We found through several cross-validation runs that this had no noticeable effect to the optima identified. We note finally that other Monte Carlo schemes can be utilized. One would expect that Importance Sampling [55], where previously generated samples are re-weighted and re-used, could be quite effective particularly when $\boldsymbol{\theta}^{(t)}$ do not change much and the corresponding $q_i^{\text{opt},\,t}$ are quite similar. A more potent alternative is offered by Sequential Monte Carlo schemes (SMC) [8,56] which combine the benefits of MCMC and Importance Sampling.

- The maximization of the lower bound with respect to $\boldsymbol{\theta}$ is not analytically tractable even when a Monte Carlo approximation of the objective, as discussed previously, is used. For that purpose, we make use of a gradient ascent scheme that employs the partial derivatives of $\mathcal{F}$:

$$\begin{aligned}\mathcal{G}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\,\mathcal{F} &= \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \mathcal{F}^{(i)} \quad \left(= \sum_{i=1}^{N} \mathcal{G}^{(i)}(\boldsymbol{\theta})\right) \\ &= \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \left(\int q_i^{\text{opt},\,t}(\boldsymbol{X}^{(i)}) \log\left(p_{\text{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{\text{cf}}^{(t)})\; p_{\text{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\text{c}}^{(t)})\right) d\boldsymbol{X}^{(i)} \right), \end{aligned} \tag{26}$$

where at each iteration t, each term $\mathcal{G}^{(i)}(\boldsymbol{\theta})$ is approximated by a Monte Carlo estimate (see discussion before) as:

$$\begin{aligned}\mathcal{G}^{(i)}(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \int q_i^{\text{opt},\,t}(\boldsymbol{X}^{(i)}) \log\left(p_{\text{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{\text{cf}}^{(t)})\; p_{\text{c}}(\boldsymbol{X}^{(i)}|\boldsymbol{\theta}_{\text{c}}^{(t)})\right) d\boldsymbol{X}^{(i)} \\ &\approx \frac{1}{m_t} \sum_{j=1}^{m_t} \nabla_{\boldsymbol{\theta}} \log\left(p_{\text{cf}}(\boldsymbol{x}^{(i)}|\boldsymbol{X}_j^{(i)}, \boldsymbol{\theta}_{\text{cf}}^{(t)})\; p_{\text{c}}(\boldsymbol{X}_j^{(i)}|\boldsymbol{\theta}_{\text{c}}^{(t)})\right) \\ &= \hat{\mathcal{G}}_t^{(i)}. \end{aligned} \tag{27}$$

The latter are used to update $\boldsymbol{\theta}$ as follows[4]:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} + \eta_t \sum_{i=1}^{N} \hat{\mathcal{G}}_t^{(i)}. \tag{28}$$

The step sizes η_t are defined in the context of the Robbins–Monro scheme [45] which is designed to handle the unavoidable Monte Carlo noise in the gradient estimates. They should satisfy the following conditions [57]:

$$\sum_{t=1}^{\infty} \eta_t = +\infty, \text{ and } \sum_{t=1}^{\infty} \eta_t^2 < \infty. \tag{29}$$

In this work, we employ [36]:

$$\eta_t = \frac{\alpha}{(A+t)^{\rho}}, \tag{30}$$

with $\rho \in (0.5, 1]$. The choice for the values α, ρ, and A is problem dependent and is explicitly given in Sections 3.1 and 3.2 for the Ising and water problems, respectively.

- We note finally that the gradient needed for the $\boldsymbol{\theta}$-updates, involves the sum of N independent terms, one for each datum (i.e. FG configuration) available. Apart from the obvious opportunity for parallelization that this offers, it also suggests that fine-scale data can be successively added. Hence the optimization can be initiated with a small number of data points N and the changes in the optimal $\boldsymbol{\theta}$ identified can be monitored as more fine-scale data are generated/added to ensure that convergence is achieved with the smallest such effort. Another strategy for reducing the computational effort is to perform the E-step i.e. sample from $q_i^{\text{opt},\,t}$ only for a subset of the data $i = 1, \ldots, N$ at a time. While this has the potential of reducing the overall number of MCMC steps needed, convergence is still guaranteed [43].

[4] As discussed in the seminal work of Neal and Hinton [43], more than one updates of $\boldsymbol{\theta}$ per EM iteration can be performed.

2.4. Exponential family densities – uniqueness of solution

In order to provide some insight to the log-likelihood maximization, we consider the case of model densities belong to the exponential family [41,58]. As it will be shown in the numerical illustrations, this represents a very large set of flexible densities where by appropriate selection of the feature functions $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ in the equations below one can capture interactions of various order (e.g. 2nd, 3rd) [36,38]. Such densities have the form:

$$p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}}) = \exp\{\boldsymbol{\theta}_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{X}) - A(\boldsymbol{\theta}_{\mathrm{c}})\}, \tag{31}$$

and:

$$p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}}) = \exp\{\boldsymbol{\theta}_{\mathrm{cf}}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{X}) - B(\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})\}, \tag{32}$$

where $A(\boldsymbol{\theta}_{\mathrm{c}})$ and $B(\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})$ are the log-partition functions given by:

$$\begin{aligned} A(\boldsymbol{\theta}_{\mathrm{c}}) &= \log \int e^{\boldsymbol{\theta}_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{X})} d\boldsymbol{X}, \\ B(\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}}) &= \log \int e^{\boldsymbol{\theta}_{\mathrm{cf}}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{X})} d\boldsymbol{x}. \end{aligned} \tag{33}$$

One can readily show that:

$$\begin{aligned} \frac{\partial A(\boldsymbol{\theta}_{\mathrm{c}})}{\partial \theta_{\mathrm{c},k}} &= < \phi_k(\boldsymbol{X}) >_{p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}})}, \\ \frac{\partial^2 A(\boldsymbol{\theta}_{\mathrm{c}})}{\partial \theta_{\mathrm{c},k} \partial \theta_{\mathrm{c},l}} &= \mathrm{Cov}_{p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}})}[\phi_k(\boldsymbol{X}), \phi_l(\boldsymbol{X})], \end{aligned} \tag{34}$$

and:

$$\begin{aligned} \frac{\partial B(\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})}{\partial \theta_{\mathrm{cf},k}} &= < \psi_k(\boldsymbol{x}, \boldsymbol{X}) >_{p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})}, \\ \frac{\partial^2 B(\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})}{\partial \theta_{\mathrm{cf},k} \partial \theta_{\mathrm{cf},l}} &= \mathrm{Cov}_{p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}})}[\psi_k(\boldsymbol{x}, \boldsymbol{X}), \psi_l(\boldsymbol{x}, \boldsymbol{X})], \end{aligned} \tag{35}$$

where $< \cdot >_p$ denotes the expectation with respect to the density p and $\mathrm{Cov}_p[\cdot, \cdot]$ the covariance of the arguments with respect to p. Hence, for p_{c} and p_{cf} as above, the gradient of the objective $\mathcal{F}$ in Eq. (24) is given by[5]:

$$\frac{\partial \mathcal{F}}{\partial \theta_{\mathrm{c},k}} = \sum_{i=1}^{N} \left(< \phi_k(\boldsymbol{X}^{(i)}) >_{q_i(\boldsymbol{X}^{(i)})} - < \phi_k(\boldsymbol{X}) >_{p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}})} \right),$$

and

$$\frac{\partial \mathcal{F}}{\partial \theta_{\mathrm{cf},k}} = \sum_{i=1}^{N} \left(< \psi_k(\boldsymbol{x}^{(i)}, \boldsymbol{X}^{(i)}) >_{q_i(\boldsymbol{X}^{(i)})} - < \psi_k(\boldsymbol{x}, \boldsymbol{X}^{(i)}) >_{p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{\mathrm{cf}}) q_i(\boldsymbol{X}^{(i)})} \right). \tag{36}$$

Furthermore, the Hessian is:

$$\begin{aligned} \frac{\partial^2 \mathcal{F}}{\partial \theta_{\mathrm{c},k} \theta_{\mathrm{c},l}} &= -N \, \mathrm{Cov}_{p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}})}[\phi_k(\boldsymbol{X}), \phi_l(\boldsymbol{X})], \\ \frac{\partial^2 \mathcal{F}}{\partial \theta_{\mathrm{c},k} \theta_{\mathrm{cf},l}} &= 0, \\ \frac{\partial^2 \mathcal{F}}{\partial \theta_{\mathrm{cf},k} \theta_{\mathrm{cf},l}} &= -\sum_{i=1}^{N} \mathrm{Cov}_{p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}^{(i)}, \boldsymbol{\theta}_{\mathrm{cf}}) q_i(\boldsymbol{X}^{(i)})}[\psi_k(\boldsymbol{x}, \boldsymbol{X}), \psi_l(\boldsymbol{x}, \boldsymbol{X})]. \end{aligned} \tag{37}$$

The block-diagonal Hessian is negative definite (at least when linearly independent feature functions are employed) which ensures that the objective is concave and has a unique maximum (whether arbitrary q_i are employed or q_i^{opt} as in Eq. (21)). We note also that Monte Carlo estimates of the Hessian can also be obtained and used in the $\boldsymbol{\theta}$-updates. These however tend to be more noisy than the gradients and special treatment is needed unless one is willing to generate large numbers of MCMC samples [36]. Finally, there is a wealth of stochastic approximation schemes that have been proposed and exhibit accelerated convergence [59–62].

[5] We compare gradients of PCG with the relative entropy method in A.2.

2.5. Prior specification

The incorporation of priors for $\boldsymbol{\theta}$ does not pose any computational difficulties as their contribution is additive (see Eq. (15)) to the log-likelihood and its partial derivatives. While priors for $\boldsymbol{\theta}_{\text{cf}}$, i.e. the parameters in the coarse-to-fine map, are unavoidably problem-dependent due to their special physical meaning, a more general strategy can be adopted for the $\boldsymbol{\theta}_{\text{c}}$, i.e. the parameters associated with the density of the coarse-grained variables $\boldsymbol{X}$. For exponential family distributions as in Eq. (31), each $\theta_{\text{c},k}$ is associated with a feature function $\phi_k(\boldsymbol{X})$. As it will become apparent in the numerical examples, each of these feature functions encapsulates low- or high-order dependencies (or components thereof) between $\boldsymbol{X}$. It is obviously impossible to know a priori which of the $\phi(\boldsymbol{X})$ are relevant for a particular problem and how these depend on the dimension of $\boldsymbol{X}$ or the coarse-to-fine probabilistic map p_{cf}. This underpins an important *model selection* issue that has been of concern in several coarse-graining studies [30,33,34,38]. One strategy to address this is to initiate the search with a small number of features $\phi(\boldsymbol{X})$ and progressively add more. These can be selected from a pool of candidates by employing appropriate criteria. In [8,63] for example, the feature function that causes the largest (expected) decrease (or increase) in the KL-divergence (or the log-likelihood) that we seek to minimize (or maximize), is added at each step. In this work, we adopt a different approach whereby *all* available $\phi(\boldsymbol{X})$ contained in the vocabulary of feature functions, are simultaneously considered. Consequently this leads to a vector of unknowns $\boldsymbol{\theta}_{\text{c}}$ of very large dimension which not only impedes computations but can potentially lead to multiple local maxima, if the Hessian in Eq. (37) becomes semi-negative definite i.e. if linear dependencies between the selected $\phi(\boldsymbol{X})$ are present. More importantly though (at least when the number of data points N is small), it can obstruct the identification of the most salient features of the coarse-grained model which provide valuable physical insight [30].

To address this, we propose the use of sparsity-enforcing priors that are capable of identifying solutions in which only a (small) subset of $\boldsymbol{\theta}_{\text{c}}$ are non-zero and therefore only the corresponding $\phi(\boldsymbol{X})$ are active [64,65]. A lot of the prior models that have been proposed along these lines can be readily cast in the context of *hierarchical Bayesian models* where *hyper-parameters* are introduced in the prior. In this work, we adopt the Automatic Relevance Determination (ARD, [66]) model which consists of the following:

$$p(\boldsymbol{\theta}_{\text{c}}|\boldsymbol{\tau}) \equiv \prod_k \mathcal{N}(\theta_{\text{c},k}|0, \tau_k^{-1}), \quad \tau_k \sim Gamma(\tau_k|a_0, b_0). \tag{38}$$

This implies that each $\theta_{\text{c},k}$ is modeled (a priori) with an independent, zero-mean, Gaussian, with a precision hyper-parameter τ_k which is in turn modeled (independently) with a (conjugate) Gamma density. We note that when $\tau_k \to \infty$, then $\theta_{\text{c},k} \to 0$. The resulting prior for $\theta_{\text{c},k}$ arising by marginalizing the hyper-parameter is a heavy-tailed, Student's t-distribution. For the purposes of learning of $\boldsymbol{\theta}_{\text{c}}$ and in order to compute derivatives of the log-prior, we retain the τ_k's and treat them as latent variables in an inner-loop EM scheme [67] (see derivation in A.3) which consists of:

- E-step: evaluate:

$$\langle \tau_k \rangle_{p(\tau_k|\theta_{\text{c},k})} = \frac{a_0 + \frac{1}{2}}{b_0 + \frac{\theta_{\text{c},k}^2}{2}}. \tag{39}$$

- M-step: evaluate:

$$\frac{\partial \log p(\boldsymbol{\theta}_{\text{c}})}{\partial \theta_{\text{c},k}} = -\langle \tau_k \rangle_{p(\tau_k|\theta_{\text{c},k})} \theta_{\text{c},k}. \tag{40}$$

We note also that the second derivative of the log-prior with respect to $\boldsymbol{\theta}_{\text{c}}$ can be similarly obtained as:

$$\frac{\partial^2 \log p(\boldsymbol{\theta}_{\text{c}})}{\partial \theta_{\text{c},k} \partial \theta_{\text{c},l}} = \begin{cases} -\langle \tau_k \rangle_{p(\tau_k|\theta_{\text{c},k})}, & \text{if } k = l \\ 0, & \text{otherwise.} \end{cases} \tag{41}$$

2.6. Approximate Bayesian inference – Laplace's approximation

The discussion thus far has been limited to point estimates for $\boldsymbol{\theta}$. A fully Bayesian treatment would pose significant computational challenges. These stem from the intractability of the log-partition function $A(\boldsymbol{\theta}_{\text{c}})$ of p_{c} in the exponential family of models (see Eq. (31)). Sampling or approximating the full posterior of $\boldsymbol{\theta}_{\text{c}}$ would require repeated evaluations of this and potentially its derivatives, a difficulty which is only amplified when $\dim(\boldsymbol{\theta}_{\text{c}}) \gg 1$. For that reason, we adopt an approximation based on the Laplace's method [68]. According to this, the target posterior $p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})$ is modeled with a Gaussian (Fig. 2) with mean equal to the MAP estimate $\boldsymbol{\theta}_{\text{MAP}}$ and a covariance $\boldsymbol{S}$ equal to the inverse of the negative Hessian of the log-posterior at $\boldsymbol{\theta}_{\text{MAP}}$ (see Eqs. (37) and (41)). These two quantities are readily obtained at the last iteration (upon convergence) of the MC-EM scheme described previously. Hence:

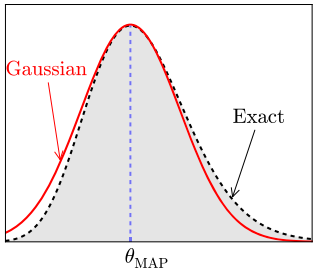


Fig. 2. Schematic illustration of the Laplace's approximation.

$$\boldsymbol{S}^{-1} = \begin{bmatrix} \boldsymbol{S}_{cc} & 0 \\ 0 & \boldsymbol{S}_{ff} \end{bmatrix}, \tag{42}$$

where the block-matrices above are given by:

$$\begin{aligned} \boldsymbol{S}_{cc} &= N\,\mathrm{Cov}_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)}[\boldsymbol{\phi}(\boldsymbol{X}), \phi_l(\boldsymbol{X})] + \mathrm{diag}(\langle\tau_k\rangle_{p(\tau_k|\theta_{c,k})}) \\ \boldsymbol{S}_{ff} &= \sum_{i=1}^{N} \mathrm{Cov}_{p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}^{(i)},\boldsymbol{\theta}_{\mathrm{cf}})q_i(\boldsymbol{X}^{(i)})}[\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{X})]. \end{aligned} \tag{43}$$

Laplace's approximation can also be interpreted as a second-order Taylor series expansion of the log-posterior at $\boldsymbol{\theta}_{\mathrm{MAP}}$. Some remarks:

- For $\theta_{c,k}$ that are effectively turned off when using the ARD prior (i.e. $\theta_{c,k,\mathrm{MAP}} = 0$), $\langle\tau_k\rangle_{p(\tau_k|\theta_{c,k})} \to \infty$ and thus dominate the corresponding terms in $\boldsymbol{S}^{-1}$. As a result, the (approximate) posterior covariance of these $\theta_{c,k}$ approaches 0.
- We note that when the number of data points $N \to \infty$, the corresponding terms in $\boldsymbol{S}^{-1}$ increase and as a result the (approximate) posterior covariance goes to 0, as one would expect.

Algorithm 1 summarizes the basic steps of the scheme advocated.

Algorithm 1 Proposed MC-EM scheme.

1: Initialize $\boldsymbol{\theta}^0 = \{\boldsymbol{\theta}_c{}^0, \boldsymbol{\theta}_{\mathrm{cf}}{}^0\}$.
2: Select parameters $\{a, \rho, A\}$ for the Robbins–Monro optimization algorithm (Eq. (30)).
3: Step $t = 0$
4: **while** (*not* converged) **do**
5: MC-E-step:
6: **for** all $i = 1, \ldots, N$ **do**
7: Generate MCMC samples from the (conditional) posterior distribution $q_i(\boldsymbol{X}^{(i)})$ in Eq. (21)
8: **end for**
9: M-step:
10: Construct Monte Carlo gradient estimators $\hat{\mathcal{G}}^{(i)}$ (Eq. (27)) augmented by the prior gradient (Eq. (40)).
11: Update the parameters $\boldsymbol{\theta}$ based on Eq. (28))
12: $t \leftarrow t + 1$
13: **end while**
14: Compute Hessian of the log-posterior Eq. (15) at $\boldsymbol{\theta}_{\mathrm{MAP}}$ (Eqs. (37), (41)) to construct Laplace's approximation of the posterior $p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})$ (Eq. (42)).

3. Numerical illustrations

We illustrate the proposed PCG framework in two examples. We particularize the definition of coarse-grained variables $\boldsymbol{X}$ which unavoidably differs from problem to problem. We emphasize through several illustrations the ability of the proposed method to produce predictive estimates of various macroscopic observables as well as quantify the predictive uncertainty as a function of the amount of training data N used and the level of coarse-graining i.e. the ratio of the amount of fine to coarse variables. We also provide comparisons with the results obtained by employing the relative entropy method. Finally, we demonstrate how the ARD prior advocated can lead to the discovery of sparse solutions revealing the most prominent feature functions in the coarse potential and possibly the most significant types of interactions that this should contain. Whenever such a hierarchical prior (ARD) is employed (Eq. (38)) for the parameters $\boldsymbol{\theta}_c$ in the coarse potential, the following values were used for the hyperparameters: $a_0 = b_0 = 10^{-5}$.

3.1. Ising model

The Ising model serves as abstraction of various physical problems, e.g. for modeling electromagnetism or lattice gas systems [69,70]. It has been the subject of detailed studies and several strategies for coarse-graining in equilibrium [9, 11–13,35,36] and nonequilibrium [9] settings.

We consider a periodic, one-dimensional lattice consisting of $n_\mathrm{f} = 64$ sites. Each site i is associated with a binary variable $x_i, i = 1, \dots, n_\mathrm{f}$ which takes values ± 1. The n_f-dimensional vector $\boldsymbol{x} = \{x_i\}_{i=1}^{n_\mathrm{f}}$ follows $p_\mathrm{f}(\boldsymbol{x}) \propto \exp\{-\beta U_\mathrm{f}(\boldsymbol{x})\}$ with the fine-scale potential given by:

$$U_\mathrm{f}(\boldsymbol{x}) = -\frac{1}{2}\sum_{k=1}^{L_\mathrm{f}} J_k \left(\sum_{|i-j|=k} x_i x_j \right) - \mu \sum_{i=1}^{n_\mathrm{f}} x_i. \tag{44}$$

The expression $|i - j| = k$ implies a summation over all lattice sites i, j that are k-sites apart (periodic boundary conditions are assumed). The parameter L_f expresses the maximal interaction length. Following [9,28,71], we use a decaying interaction strength J_k with,

$$J_k = \frac{K}{k^a}, \tag{45}$$

and the normalization,

$$K = \frac{J_0}{L_\mathrm{f}^{a-1} \sum_{k=1}^{L_\mathrm{f}} k^{-a}}. \tag{46}$$

Finally, the parameter μ denotes the external field.

The values $A = 25$, $\alpha = 0.15$, and $\rho = 0.75$ were used for the Robbins–Monro updates (Eq. (28)) based on suggestions given in [36]. We used $m = 170$ samples for the MCMC estimates of the gradients in Eqs. (25) and (27).

3.1.1. Observables

As pointed out previously, the framework proposed readily allows for reconstructions of the whole fine-scale description and therefore probabilistic predictions can be computed for any observable. For comparative purposes, we focus on two such quantities. The first one is the magnetization $m(\mu)$ and its dependence on the external field parameter μ. This is associated with the following observable:

$$a^{(m)}(\boldsymbol{x}) = \frac{1}{n_\mathrm{f}} \sum_i x_i, \tag{47}$$

i.e. $m(\mu) = \mathbb{E}_{p_\mathrm{f}(\boldsymbol{x})}[a^{(m)}(\boldsymbol{x})]$. The second quantity is the correlation $R(k)$ at various separation distances k which captures second-order statistical information of the fine-scale configurations. The corresponding observable is:

$$a^{(R)}(\boldsymbol{x}; k) = \frac{1}{n_\mathrm{f}} \sum_{|i-j|=k} x_i x_j, \tag{48}$$

i.e. $R(k) = \mathbb{E}_{p_\mathrm{f}(\boldsymbol{x})}[a^R(\boldsymbol{x}; k)]$.

3.1.2. Coarse-variables $\boldsymbol{X}$ and coarse-to-fine map

While the framework proposed offers great flexibility in the definition of the coarse variables $\boldsymbol{X}$, in this work we make perhaps the most intuitive choice by assuming that $\boldsymbol{X}$ are (also) binary and have a *local* dependence on $\boldsymbol{x}$. This offers a direct appraisal on the level of coarse-graining as well as a natural, visual interpretation of the coarse variables and their role.

In particular, we assume that each coarse variable $X_I, I = 1, \dots, n_\mathrm{c}$ is associated with a one-dimensional lattice that is a coarser version of the fine-scale one, i.e. with $n_\mathrm{c} < n_\mathrm{f}$ sites (Fig. 3). We can construct such descriptions by regularly coarsening by a factor of 2 such that $n_\mathrm{c} = n_\mathrm{f}/2^d$, with $d = 1, \dots, D$. We assume that each X_I (parent) is associated with $S = \frac{n_\mathrm{f}}{n_\mathrm{c}}$ fine-scale variables (children) denoted by $x_{(I-1)S+s} = x_{s,I}$ (where $s = 1, \dots, S$, Fig. 3). We define a coarse-to-fine map of the form:

$$\begin{aligned} p_\mathrm{cf}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_\mathrm{cf}) &= \prod_{I=1}^{n_\mathrm{c}} \prod_{s=1}^{S} p(x_{s,I}|X_I, \boldsymbol{\theta}_\mathrm{cf}) \\ &= \prod_{I=1}^{n_\mathrm{c}} \prod_{s=1}^{S} p_0^{\frac{1+x_{s,I}X_I}{2}} (1-p_0)^{\frac{1-x_{s,I}X_I}{2}} \\ &= p_0^{\sum_{I=1}^{n_\mathrm{c}} \sum_{s=1}^{S} \frac{1+x_{s,I}X_I}{2}} (1-p_0)^{\sum_{I=1}^{n_\mathrm{c}} \sum_{s=1}^{S} \frac{1-x_{s,I}X_I}{2}}. \end{aligned} \tag{49}$$

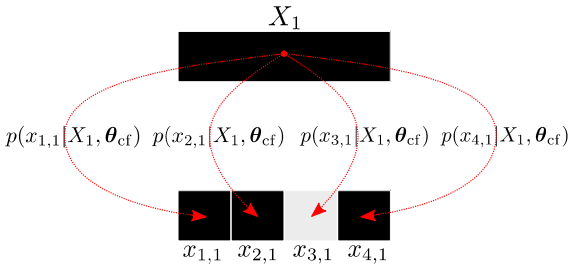


Fig. 3. Probabilistic coarse-to-fine map $p_{\rm cf}(\boldsymbol{x}|\boldsymbol{X},\boldsymbol{\theta}_{\rm cf})$. The coarse-variable X_1 is e.g. associated with $x_{1\ldots4,1}$ fine-scale variables through the probabilistic coarse-to-fine map $p_{\rm cf}$ (Eq. (49)). Each $x_{s,1}$ is *conditionally* independent from the other.

The expression above implies that each $x_{s,I}$ is *conditionally* independent and follows a Bernoulli distribution with probability p_0 of being of the same value as its parent X_I, and probability $(1-p_0)$ of having the opposite spin. We emphasize that this does not imply that $x_{s,I}$ are also independent. In fact they will be correlated as a result of the dependencies between the coarse variables $\boldsymbol{X}$ induced by the coarse model $p_{\rm c}$ which is discussed in the next subsection. The density $p_{\rm cf}$ above belongs to the exponential family (Section 2.4) and is controlled by a single parameter, $p_0 \in [0,1]$. Given the symmetry of the model, we restrict $p_0 \in [0.5,1]$. To ensure that it stays within this interval during the MC-EM updates (Algorithm 1), we operate instead on $\theta_{\rm cf} \in \mathbb{R}$ defined as follows:

$$p_0 = \frac{1}{2}\left(1 + \frac{1}{1+e^{-\theta_{\rm cf}}}\right). \tag{50}$$

The derivatives needed for the updates of the EM-scheme in Eq. (27) and Eq. (37) are:

$$\begin{aligned}
\frac{\partial \log p_{\rm cf}}{\partial \theta_{\rm cf}} &= \frac{\partial \log p_{\rm cf}}{\partial p_0}\frac{\partial p_0}{\partial \theta_{\rm cf}},\\
\frac{\partial^2 \log p_{\rm cf}}{\partial \theta_{\rm cf}^2} &= \frac{\partial^2 \log p_{\rm cf}}{\partial p_0^2}\left(\frac{\partial p_0}{\partial \theta_{\rm cf}}\right)^2 + \frac{\partial \log p_{\rm cf}}{\partial p_0}\frac{\partial^2 p_0}{\partial \theta_{\rm cf}^2},
\end{aligned} \tag{51}$$

where:

$$\begin{aligned}
\frac{\partial \log p_{\rm cf}}{\partial p_0} &= \frac{\psi(\boldsymbol{x},\boldsymbol{X})}{p_0} - \frac{1-\psi(\boldsymbol{x},\boldsymbol{X})}{1-p_0},\\
\frac{\partial^2 \log p_{\rm cf}}{\partial p_0^2} &= -\frac{\psi(\boldsymbol{x},\boldsymbol{X})}{p_0^2} - \frac{1-\psi(\boldsymbol{x},\boldsymbol{X})}{(1-p_0)^2},
\end{aligned} \tag{52}$$

and $\psi(\boldsymbol{x},\boldsymbol{X}) = \sum_{I=1}^{n_{\rm c}}\sum_{s=1}^{S}\frac{1+x_{s,I}X_I}{2}$.

3.1.3. Coarse model

The coarse potential $U_{\rm c}(\boldsymbol{X};\boldsymbol{\theta}_{\rm c})$ employed includes first-, second- and third-order interactions with various interaction lengths. In particular, we prescribe:

$$U_{\rm c}(\boldsymbol{X};\boldsymbol{\theta}_{\rm c}) = -\frac{1}{2}\left\{\theta_{\rm c}^{(1)}\sum_i X_i + \sum_i X_i \sum_k^{L_{\rm c}^{(2)}} \theta_{{\rm c},k}^{(2)} X_{i\pm k} + \sum_i X_i \sum_{\substack{k=1\\l=1}}^{L_{\rm c}^{(3)}} \theta_{{\rm c},kl}^{(3)} X_{i\pm k} X_{i\pm k\pm l}\right\} - \mu\sum_i X_i. \tag{53}$$

The parameters $L_{\rm c}^{(2)}$ and $L_{\rm c}^{(3)}$ denote the maximal second- and third- order interactions, respectively. With superscripts (1), (2), (3) we distinguish between the coarse potential parameters $\boldsymbol{\theta}_{\rm c}$ that are associated with the first, two-body and third-body interactions, respectively. These parameters determine also the number of $\boldsymbol{\theta}_{\rm c}$ which is equal to $1 + L_{\rm c}^{(2)} + (L_{\rm c}^{(3)})^2$.

In order to compare the proposed method with the relative entropy method, as briefly summarized in Section 2.1, a deterministic fine-to-coarse mapping $\mathcal{R}(\boldsymbol{x})$ is needed. We note that in [35,36] such efforts have been made by "coarse-graining" the interactions rather than the degrees of freedom i.e. $\boldsymbol{x} \equiv \boldsymbol{X}$. In order to truly assess the performance in cases where the coarse variables are of lower dimension and of the same type as in this study (i.e. binary), we prescribe the following map:

$$X_I = \begin{cases} +1, & \frac{1}{S}\sum_s^S x_{s,I} \geq 0\\ -1, & \frac{1}{S}\sum_s^S x_{s,I} < 0. \end{cases} \tag{54}$$

This implies a "majority rule" where the label of the parent X_I is determined by the majority of the children. The same model as in Eq. (53) was used for the coarse potential. In order to reconstruct the fine configurations $\boldsymbol{x}$ and estimate the observables of interest from the coarse description $\boldsymbol{X}$, a consistent sampling was performed from the conditional in Eq. (7) for the $\mathcal{R}$ above.

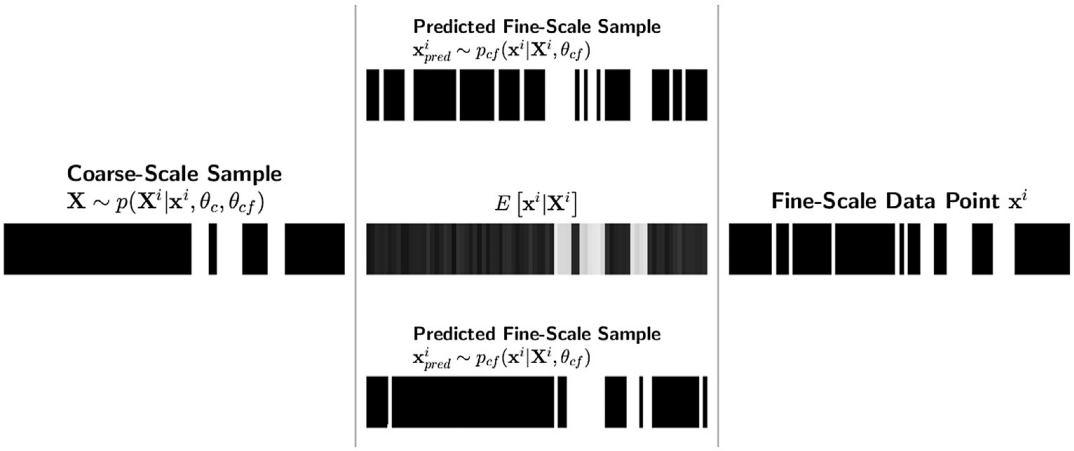


Fig. 4. For the FG datum $\boldsymbol{x}^{(i)}$ (right), the image on the left shows a sample from the posterior of the CG $\boldsymbol{X}^{(i)}$ (upon convergence of the Algorithm 1) i.e. one of the possible *pre-images* of $\boldsymbol{x}^{(i)}$. The three images in the center illustrate the predictions/reconstructions of the fine-scale: the top and bottom are samples drawn from the p_{cf} and the center is the expected FG configuration according to p_{cf}.

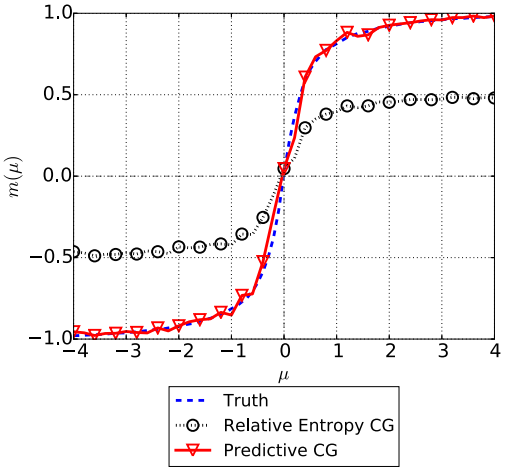
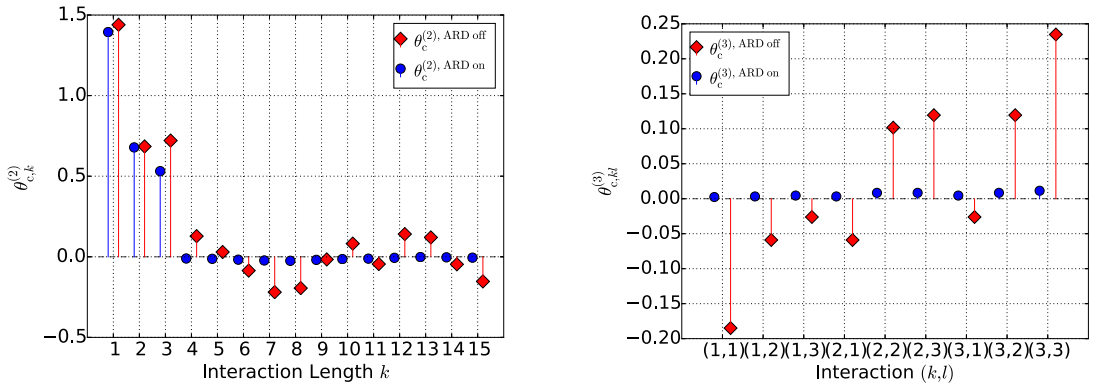


Fig. 5. Comparison of the reference magnetization (computed with the FG configuration) with posterior mean of predictive CG and relative entropy CG. $N = 20$, $\frac{n_{\text{f}}}{n_{\text{c}}} = 2$, $L_{\text{c}}^{(2)} = 15$, $L_{\text{c}}^{(3)} = 3$.

3.1.4. Results

The ensuing results are based on the following values for the fine-scale potential: $J_0 = 1.5$, $a = 0.8$, $L_{\text{f}} = 8$, $\beta = 0.3$, $n_{\text{f}} = 64$. We generated data from the fine scale model for each of 41 values of the external field μ, equidistantly distributed within $[-4, 4]$. A different CG model is trained for every μ value considered. One could also envision introducing a dependence of the CG model's components on μ which would allow a single model to be inferred and to be used for making predictions even for values of μ not contained in the data. Fig. 4 provides some insight on the role of the CG variables, their posterior and their ability to represent/reconstruct the FG configuration. Fig. 5 compares point-estimates of the predicted magnetization as obtained with the proposed method (red) and the relative entropy method (for fine-to-coarse mapping as given in Eq. (54)). While one can claim that better results can be obtained with a different set of CG variables (Eq. (54)), the point in this comparison is to demonstrate the information loss that takes place which can lead to poor predictions when not quantified. Given the same amount of training data N, the information loss in the relative entropy method is driven by the not adjusted map in the consistent density of the fine-scale variables $p_{\mathcal{T}\mathcal{R}}(\boldsymbol{x})$ denoted in Eq. (9) compared to PCG. While in PCG the probabilistic map $p_{\text{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\text{cf}})$ (Eq. (49)) is parametrized and optimized within the parametric family of p_{cf}. We note further that the relative entropy method can lead to good approximations of the Potential of Mean Force, and as a result, accurate estimates (as shown earlier) of expectations of observables that depend solely on $\boldsymbol{X}$. We could therefore select $\boldsymbol{X}$ in such a way that the magnetization is only a function of $\boldsymbol{X}$ in which case the result of the relative entropy method would probably be good. If however another expectation was sought (that does not depend on the current $\boldsymbol{X}$) a new set of $\boldsymbol{X}$ would need to be defined and a new CG model would need to be retrained.

When $\frac{n_{\text{f}}}{n_{\text{c}}} = 2$, $L_{\text{c}}^{(2)} = 15$, $L_{\text{c}}^{(3)} = 3$, the total number of unknowns parameters $\boldsymbol{\theta}_{\text{c}}$ in the potential U_{c} is $1 + L_{\text{c}}^{(2)} + (L_{\text{c}}^{(3)})^2 = 25$. This is not a particularly large number, but we demonstrate nevertheless the effect of the sparsity enforcing prior in Fig. 6 when $N = 20$ data points are used. In the absence of the ARD prior (Eq. (38)), all $\boldsymbol{\theta}_{\text{c}}$ are non-zero and



(a) Parametrization of two-body interactions $\theta_{c,k}^{(2)}$

(b) Parametrization of three-body interactions $\theta_{c,kl}^{(3)}$

Fig. 6. Parametrization of two- and three-body interactions at $\mu = 0.0$, obtained with and without ARD prior. A sparse solution is obtained with active ARD prior at the same predictive accuracy (see Fig. 7). $N = 20$, $\frac{n_f}{n_c} = 2$, $L_c^{(2)} = 15$, $L_c^{(3)} = 3$.

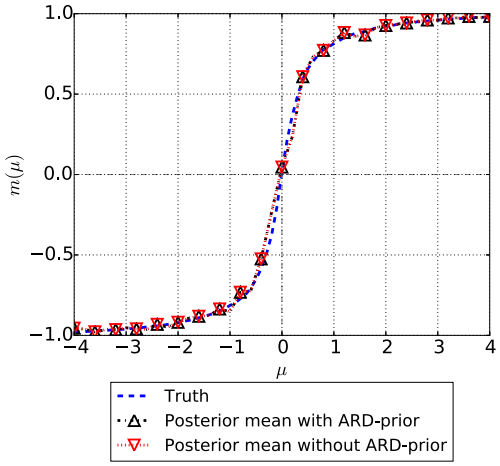


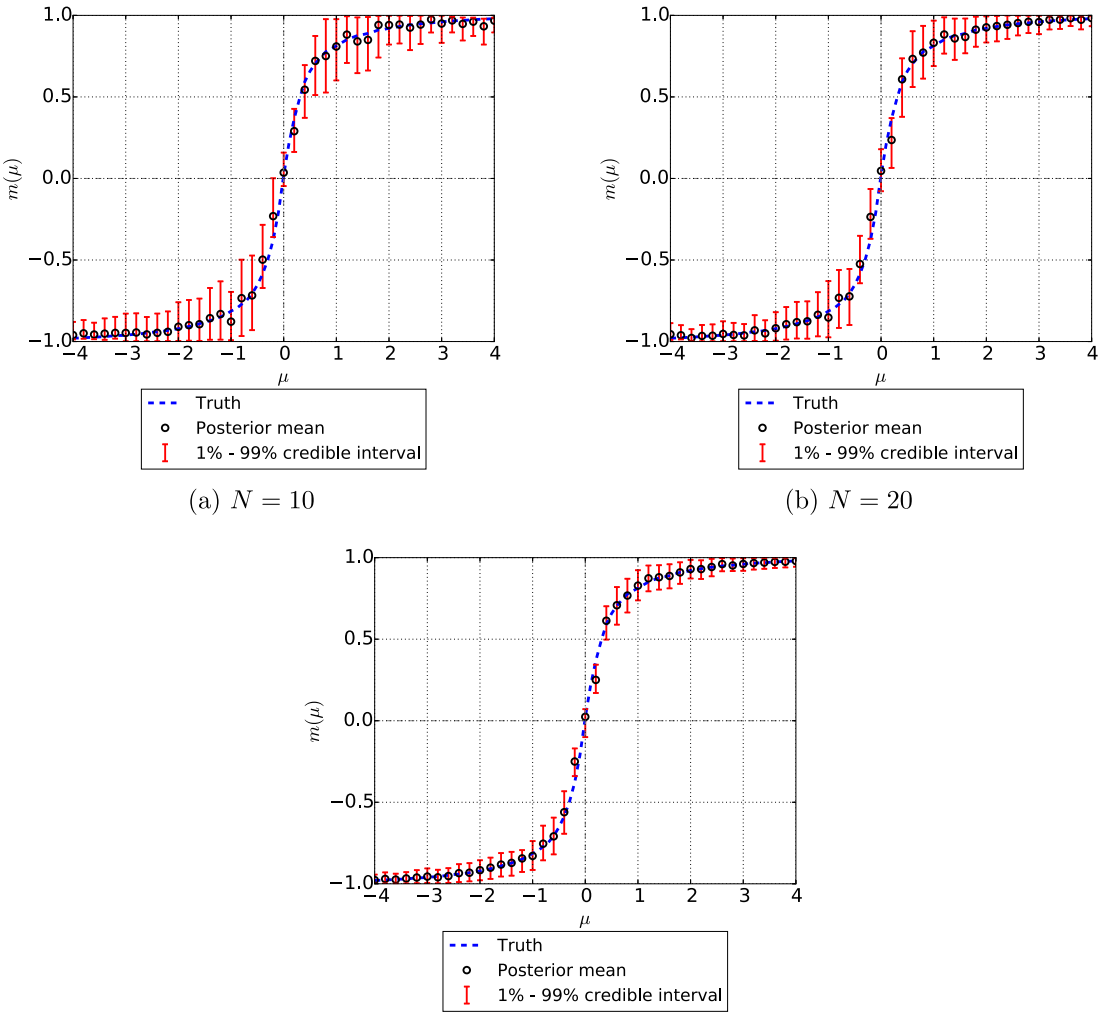
Fig. 7. Comparison of predicted magnetization with and without ARD prior. $N = 20$, $\frac{n_f}{n_c} = 2$, $L_c^{(2)} = 15$, $L_c^{(3)} = 3$.

the corresponding feature functions are all active (Eq. (53)). On the contrary, when the ARD prior is employed, the learning scheme identifies only 3 non-zero $\boldsymbol{\theta}_c$. Interestingly these are associated with two-body interactions up to separation 3 whereas all other terms corresponding to two- and three-body interactions are found to be unnecessary, despite having equal predictive accuracy as shown in Fig. 7 where point estimates of the magnetization are plotted (with and without the ARD prior).

Fig. 8 depicts the effect of adding more training data N in the predictive posterior estimates for the magnetization at various μ values. One observes that as N increases, not only the posterior mean estimates approach the reference solution, but more importantly, the posterior credible intervals shrink around it reflecting the fact that the model becomes more confident. Credible intervals are obtained by sampling the (approximate) posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})$ (Eq. (16)) and determining the observable for each sample $\boldsymbol{\theta}^{(i)}$ with the predictive estimator $\hat{a}(\boldsymbol{\theta}^{(i)})$ (Eq. (19)). We use the predictive samples $\hat{a}(\boldsymbol{\theta}^{(i)})$ to determine desired quantiles (see A.1 for more details). The same observations can be made when attempting to predict second-order statistics of the fine-scale i.e. the correlation at various separations k (Fig. 9).

The decreasing variance for increasing N can also be observed in the model parameters e.g. the coarse-to-fine mapping parameter p_0 (Eq. (49)), the (approximate) posterior of which is shown in Fig. 10.

Finally in Figs. 11 and 12, the predictive ability of the model is compared for different levels of coarse-graining. In the formulation adopted, this is quantified by the ratio between the dimension of fine $\boldsymbol{x}$ and coarse $\boldsymbol{X}$ descriptions i.e. $\frac{n_f}{n_c}$. We consider two cases i.e. $\frac{n_f}{n_c} = 2, 8$. As one would expect, the posterior mean estimates are superior when $\frac{n_f}{n_c} = 2$ but also the predictive posterior uncertainty increases as the coarse-graining becomes more pronounced. This is easily understood by the fact that the fewer CG variables used, the higher the information loss becomes. It is important to note though that even when $\frac{n_f}{n_c} = 8$, the predictive posterior's credible intervals always include the reference solution.



(a) $N = 10$ (b) $N = 20$

(c) $N = 50$

Fig. 8. Comparison of the reference magnetization (computed with the FG configuration) with posterior mean and credible intervals corresponding to 1% and 99% posterior quantiles. $N = 20$, $\frac{n_f}{n_c} = 2$, $L_c^{(2)} = 15$, $L_c^{(3)} = 3$.

3.2. Coarse-graining SPC/E water

The second example addresses the coarse-graining of a water model which is described at the atomistic scale by oxygen and hydrogen atoms. Water has been the focus of several studies in coarse-graining as it plays the role of the solvent in various biological and chemical systems and as a result it can take up to 80% of the total simulation time [30]. Furthermore there exist several well-documented properties which can serve as a measure of comparison. In this study, we employ the Simple Point Charge/Extended (SPC/E) water model introduced in [72,73] for the FG (all-atom) description. In the context of the relative entropy method, coarse-graining of the SPC/E water is addressed in [36,74–76]. In particular, we consider a system of $M = 100$ water molecules at a temperature of $T = 300\,\mathrm{K}$, and a pressure of $p = 1.0\,\mathrm{bar}$. The equilibrium box length is $l_{\mathrm{box}} = 14.56\,\text{Å}$ and a time step of $\Delta t = 2.0\,\mathrm{fs}$ is used. Periodic boundary conditions are applied in every dimension while ensuring the NVT ensemble by the Nosé–Hoover thermostat [77,78]. The $\boldsymbol{x}$ vector contains the coordinates of the 100 oxygen and 200 hydrogen atoms i.e. $\dim(\boldsymbol{x}) = 900$. The fine-scale potential $U_{\mathrm{f}}(\boldsymbol{x})$ under the SPC/E model consists of a Lennard–Jones (LJ) potential for non-bonded interactions and a Coulomb potential for long-range interactions. Parameters for the LJ potential,

$$U_{\mathrm{f}}^{\mathrm{LJ}}(\boldsymbol{x}) = \frac{1}{2}\sum_{j\neq k} 4\epsilon\left(\left(\frac{\sigma}{R_{ij}(\boldsymbol{x})}\right)^{12} - \left(\frac{\sigma}{R_{ij}(\boldsymbol{x})}\right)^{6}\right), \tag{55}$$

are $\sigma = 3.166\,\text{Å}$ and $\epsilon = 0.650\,\frac{\mathrm{kJ}}{\mathrm{mol}}$, with the distance between particle i and j denoted as R_{ij}.

The electric load of Hydrogen (H) and Oxygen (O) atoms are given by $q_{\mathrm{O}} = -0.8476\,e$, $q_{\mathrm{H}} = +0.4238\,e$ where e represents the elementary charge. The SPC/E model assumes the bonded interaction to be rigid with a bonding angle defined between the two H-atoms and the central O-atom as $\omega_{\mathrm{HOH}} = 109.47°$. The bond-length used in this study is $l_{\mathrm{OH}} = 1.0\,\text{Å}$. The

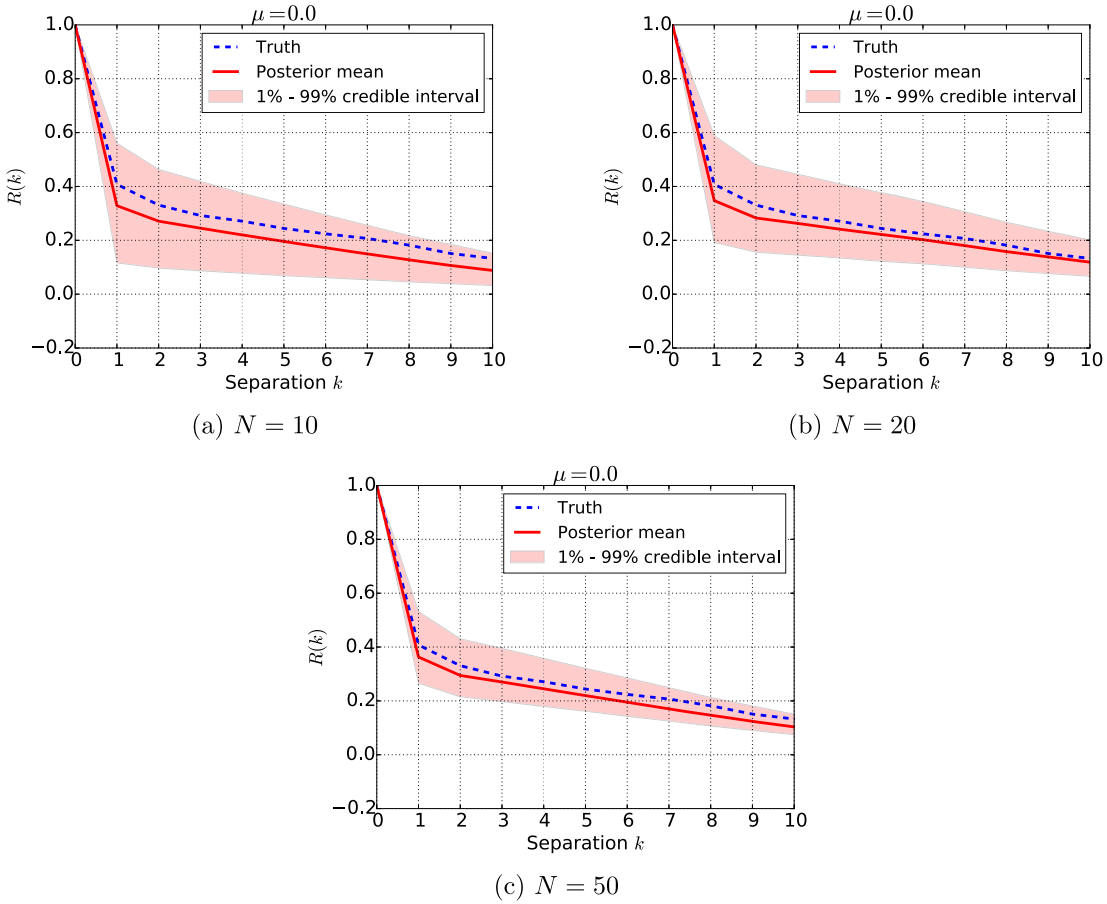


Fig. 9. Comparison of the reference correlation (computed with the FG configuration) with posterior mean and credible intervals corresponding to 1% and 99% posterior quantiles. $N = 20$, $\frac{n_f}{n_c} = 2$, $L_c^{(2)} = 15$, $L_c^{(3)} = 3$.

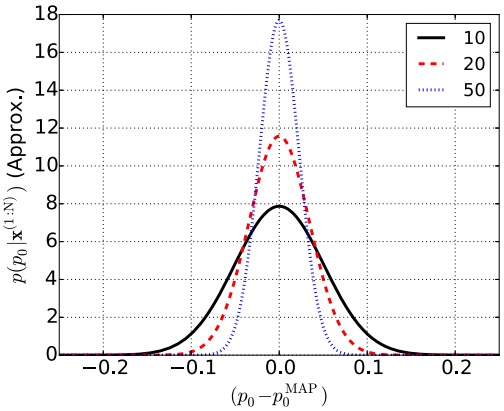


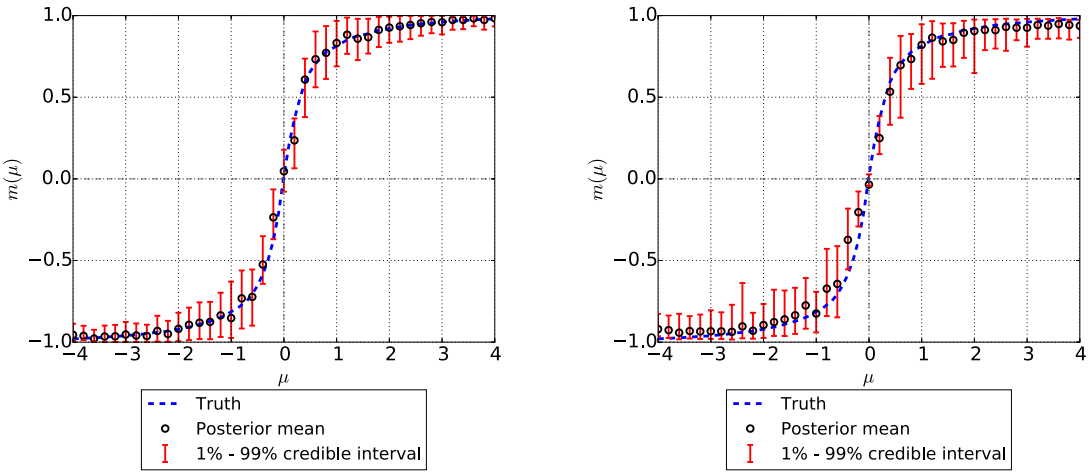
Fig. 10. Posterior $p(p_0|\boldsymbol{x}^{(1:N)})$ at $\mu = 0.0$ for $N = 10, 20, 50$. $\frac{n_f}{n_c} = 2$, $L_c^{(2)} = 15$, $L_c^{(3)} = 3$.

equilibration for the NVT ensemble was performed as in [36,75]. For both fine- and coarse-scale simulations the molecular dynamics software package LAMMPS [79] was used. Further details are contained in B.1.

The values $A = 9$, $\alpha = 0.05$, and $\rho = 0.60$ were used for the Robbins–Monro updates (Eq. (28)) based on suggestions given in [36]. We used $m = 160$ samples for the MCMC estimates of the gradients in Eqs. (25) and (27).

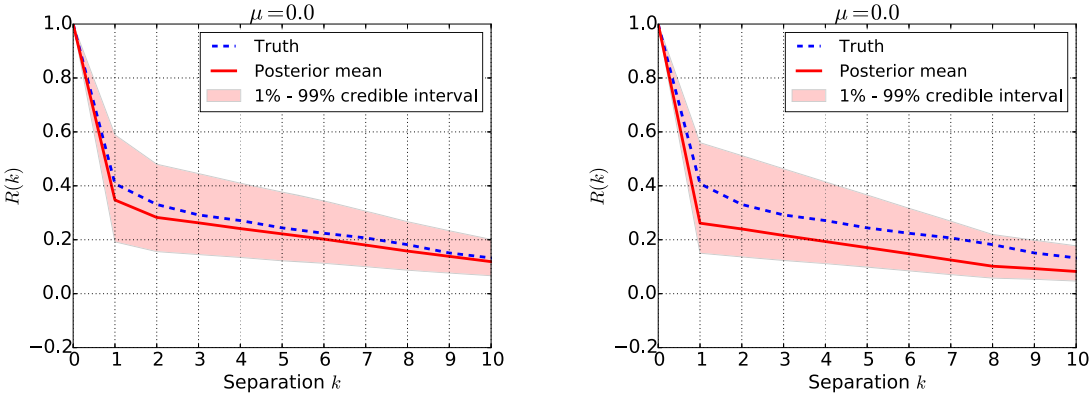
3.2.1. Observables

The first macroscopic observable of interest is the Radial Distribution Function (RDF) $g(r)$ which represents a characteristic and well-studied property in water models. Several computational and experimental results related to the RDF are described in [80]. As a pair correlation function, $g(r)$ depends on the statistics of the distances r_{jk} between each pair of molecules j, k. To compute these distances, we employ the coordinates of the center of mass of each water molecule $\hat{\boldsymbol{x}}_j$:



(a) $\frac{n_f}{n_c} = 2$ (b) $\frac{n_f}{n_c} = 8$

Fig. 11. Magnetization for different level of coarse graining, i.e. ratio of the amount fine/coarse variables $\frac{n_f}{n_c} = 2$ ($L_c^{(2)} = 15$, $L_c^{(3)} = 3$) and $\frac{n_f}{n_c} = 8$ ($L_c^{(2)} = 3$, $L_c^{(3)} = 1$). Both models were trained with the same data $N = 20$.



(a) $\frac{n_f}{n_c} = 2$ (b) $\frac{n_f}{n_c} = 8$

Fig. 12. Correlation for different level of coarse graining, i.e. ratio of the amount fine/coarse variables $\frac{n_f}{n_c} = 2$ ($L_c^{(2)} = 15$, $L_c^{(3)} = 3$) and $\frac{n_f}{n_c} = 8$ ($L_c^{(2)} = 3$, $L_c^{(3)} = 1$). Both models were trained with the same data $N = 20$.

$$\hat{\boldsymbol{x}}_j = \frac{\boldsymbol{x}_{\mathrm{O},j} m_{\mathrm{O}} + \boldsymbol{x}_{\mathrm{H},j_1} m_{\mathrm{H}} + \boldsymbol{x}_{\mathrm{H},j_2} m_{\mathrm{H}}}{m_{\mathrm{O}} + 2m_{\mathrm{H}}}, \tag{56}$$

where $\boldsymbol{x}_{\mathrm{O},j}$ are the coordinates of the oxygen atom of molecule j, $\boldsymbol{x}_{\mathrm{H},j_1}, \boldsymbol{x}_{\mathrm{H},j_2}$ are the coordinates of the two hydrogen atoms of the same molecule, and $m_{\mathrm{O}}, m_{\mathrm{H}}$ are the masses of oxygen and hydrogen atoms, respectively (see B.1). If $r_{jk} = |\hat{\boldsymbol{x}}_j - \hat{\boldsymbol{x}}_k|$, then the corresponding observable of interest is [81]:

$$a^{\mathrm{RDF}}(\boldsymbol{x}) = \frac{V}{M^2} \sum_{j}^{M} \sum_{j \neq k}^{M} \delta(r - r_{jk}), \tag{57}$$

where V denotes the volume of the simulation box (14.56^3 Å^3) and $M = 100$ the number of molecules in the system. Additional details can be found in B.2.

The second property of interest involves the tetrahedral structure of water. Neighboring water molecules temporarily build such tetrahedral clusters due to the hydrogen bonds. Several measures of tetrahedrality have been proposed which relate to the deviation from the perfect tetrahedral structure $\omega_0 = 109.471°$ [74,82]. In this work, we employ the angular distribution function which considers the eight closest neighbors $n_c = 8$ for a given molecule j. It is defined as follows:

$$a^{\mathrm{tetra}}(\boldsymbol{x}; \omega) = \frac{1}{M n_\omega} \sum_{j=1}^{M} \sum_{k=1}^{n_c} \sum_{l \neq j}^{n_c - 1} \delta(\omega - \omega_{jkl}), \tag{58}$$

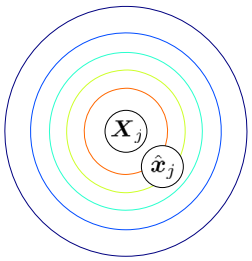


Fig. 13. Probabilistic mapping $p_{\text{cf}}(\hat{\boldsymbol{x}}_j|\boldsymbol{X}_j, \boldsymbol{\theta}_{\text{cf}})$, with mean $\boldsymbol{X}_j$ and predicted fine-scale variable $\hat{\boldsymbol{x}}_i$. The contours depict the isotropic Gaussian distribution of Eq. (59) with mean $\boldsymbol{X}_j$ and variance σ^2.

with ω_{jkl} the angle between molecules j, k, l, with the central molecule j (as computed using the centers of mass $\hat{\boldsymbol{x}}$ in Eq. (56)) and $n_\omega = \binom{n_c}{3} = 56$. The product (Mn_ω) normalizes a^{tetra} with respect to the considered angular triplets.

We note that since the observables of interest depend only on the centers of mass $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{x})$, it suffices to use a coarse-to-fine map that relates the coarse variables $\boldsymbol{X}$ directly with $\hat{\boldsymbol{x}}$ (Eq. (19)).

3.2.2. Coarse-variables $\boldsymbol{X}$ and coarse-to-fine map

Since the observables of interest depend on the centers of mass $\hat{\boldsymbol{x}}$ (Eq. (56)), the coarse-to-fine probabilistic map assumes the form $p_{\text{cf}}(\hat{\boldsymbol{x}}|\boldsymbol{X})$. As frequently done in CG studies of water, each molecule j is represented by a CG variable $\boldsymbol{X}_j \in \mathbb{R}^3$. We then prescribe a p_{cf} of the following form:

$$p_{\text{cf}}(\hat{\boldsymbol{x}}|\boldsymbol{X}, \boldsymbol{\theta}_{\text{cf}}) = \prod_{j=1}^{M} \mathcal{N}(\hat{\boldsymbol{x}}_j|\boldsymbol{X}_j,\ \sigma^2 \boldsymbol{I}), \tag{59}$$

where $\boldsymbol{I}$ is the 3×3 identity matrix. This suggests that each $\boldsymbol{X}_j$, $j = 1, \ldots, M$ determines the center of mass $\hat{\boldsymbol{x}}_j$ up to an isotropic Gaussian with mean $\boldsymbol{X}_j$ and variance σ^2 (see Fig. 13). The latter quantifies the uncertainty in the prediction of the fine-scale (up to centers of mass) from the CG description. Large values of σ^2 imply that $\boldsymbol{X}$ provides an imprecise reconstruction of $\hat{\boldsymbol{x}}$ and vice versa. Hence there is only one parameter in the coarse-to-fine map i.e. $\sigma^2 \geq 0$. In order to ensure non-negativity during updates we operate instead on $\boldsymbol{\theta}_{\text{cf}} = -\log \sigma^2$ which leads to the following derivatives needed in Eqs. (27) and (37):

$$\begin{aligned} \frac{\partial \log p_{\text{cf}}}{\partial \theta_{\text{cf}}} &= \frac{3M}{2} - \frac{1}{2\sigma^2} \sum_{j=1}^{M} |\hat{\boldsymbol{x}}_j - \boldsymbol{X}_j|^2, \\ \frac{\partial^2 \log p_{\text{cf}}}{\partial \theta_{\text{cf}}^2} &= -\frac{1}{2\sigma^2} \sum_{j=1}^{M} |\hat{\boldsymbol{x}}_j - \boldsymbol{X}_j|^2. \end{aligned} \tag{60}$$

Naturally, more complex descriptions involving an anisotropic covariance or a mixture of Gaussians could be used.

3.2.3. Coarse model

The coarse potential $U_{\text{c}}(\boldsymbol{X}; \boldsymbol{\theta}_{\text{c}})$ employed consists of two- and three-body interactions. It assumes the form:

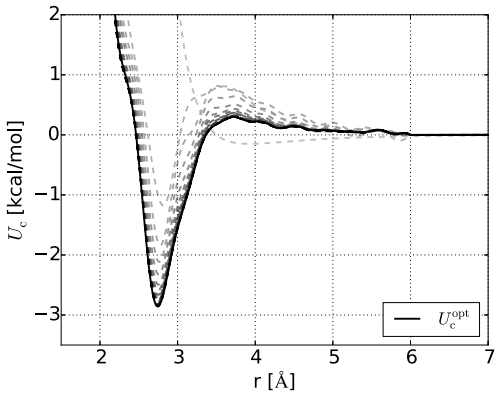
$$U_{\text{c}}(\boldsymbol{X}; \boldsymbol{\theta}_{\text{c}}) = \underbrace{U^{\text{SW}}(\boldsymbol{X})}_{\text{fixed}} + \tilde{U}(\boldsymbol{X}; \boldsymbol{\theta}_{\text{c}}), \tag{61}$$

where $U^{\text{SW}}(\boldsymbol{X})$ is a fixed term described below and $\tilde{U}(\boldsymbol{X}; \boldsymbol{\theta}_{\text{c}})$ represents the "correction" that is learned from the data using the framework advocated. In particular, the fixed term $U^{\text{SW}}(\boldsymbol{X})$ is given by (a variation of) the Stillinger–Weber (SW) potential proposed in [83] and discussed in B.3. The remaining part $\tilde{U}(\boldsymbol{X}; \boldsymbol{\theta}_{\text{c}})$ consists only of two-body interaction terms i.e.

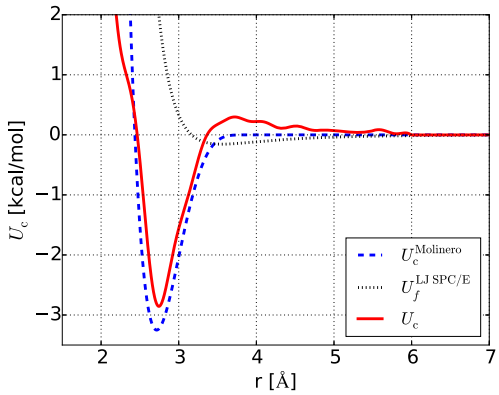
$$\tilde{U}(\boldsymbol{X}; \boldsymbol{\theta}_{\text{c}}) = \frac{1}{2} \sum_{j \neq k} u^{(2)}(R_{jk}; \boldsymbol{\theta}_{\text{c}}), \tag{62}$$

where $R_{jk} = |\boldsymbol{X}_j - \boldsymbol{X}_k|$ and the pairwise potential $u^{(2)}(R; \boldsymbol{\theta}_{\text{c}})$ is parametrized as follows:

$$u^{(2)}(R; \boldsymbol{\theta}_{\text{c}}) = u^{\text{LJ}}(R; \boldsymbol{\theta}_{\text{c}}^{\text{LJ}}) + \sum_{k=1}^{K} \theta_{\text{c},k}^{\text{cor}} \phi_k(R), \quad R > 0. \tag{63}$$



(a) Evolution of $u^{(2)}(R;\boldsymbol{\theta}_c)$ in 63 at various iterations of the Algorithm 1. Darker lines correspond to more proceeded iteration steps in the optimization scheme. The solid line shows the converged solution $u^{(2)}(R;\boldsymbol{\theta}_{c,MAP})$.



(b) Comparison of $u^{(2)}(R;\boldsymbol{\theta}_c)$ identified with the proposed method (red) with the two-body potential computed using the relative entropy method in [81] (dashed blue) and the LJ part of the fine-scale SPC/E model $U_f^{LJ\ SPC/E}$.

Fig. 14. Coarse-graining SPC/E water using $N = 20$ training data. Computed two-body, coarse-scale potential $u^{(2)}(R;\boldsymbol{\theta}_c)$ and comparisons.

In the equation above, $u^{LJ}(R;\boldsymbol{\theta}_c^{LJ})$ is a Lennard–Jones potential and the feature functions $\boldsymbol{\phi} = \{\phi_k(R)\}_{k=1}^K$ are a combination of sines and cosines truncated in the interval $I_c = [R_{min} = 2.0\,\text{Å},\ R_{max} = 6.0\,\text{Å}]$. The bounds R_{min}, R_{max} define an effective window where the LJ potential is corrected to capture the associated CG interactions. In particular:

$$\phi_k(R) = \begin{cases} 1_{I_c}(R)\ \sin 2\pi\nu_k R, & k = \text{odd}, \\ 1_{I_c}(R)\ \cos 2\pi\nu_k R, & k = \text{even}, \end{cases} \tag{64}$$

where $1_{I_c}(R)$ is the indicator function of the interval I_c. The wave-numbers ν_k offer a Fourier-like decomposition of the second-order potential and were defined as follows:

$$\nu_{2k'} = \nu_{2k'+1} = 1 + \frac{19}{K/2}k', \quad k' = 0, 2, \ldots, K/2 - 1, \tag{65}$$

i.e. at a uniform grid in $[1, 20]$. By increasing the total number K of these terms, one can potentially learn finer fluctuations of this potential. Naturally one would want to use as many feature functions as possible in order to ensure greater flexibility of the model, which gives rise to the need for sparsity-enforcing priors for $\theta_{c,k}^{cor}$ as discussed previously. In this study, $K = 100$ was used.

The superimposed LJ potential ensures that $\lim_{R\to 0} u^{(2)}(R;\boldsymbol{\theta}_c) = \infty$ and is of the form:

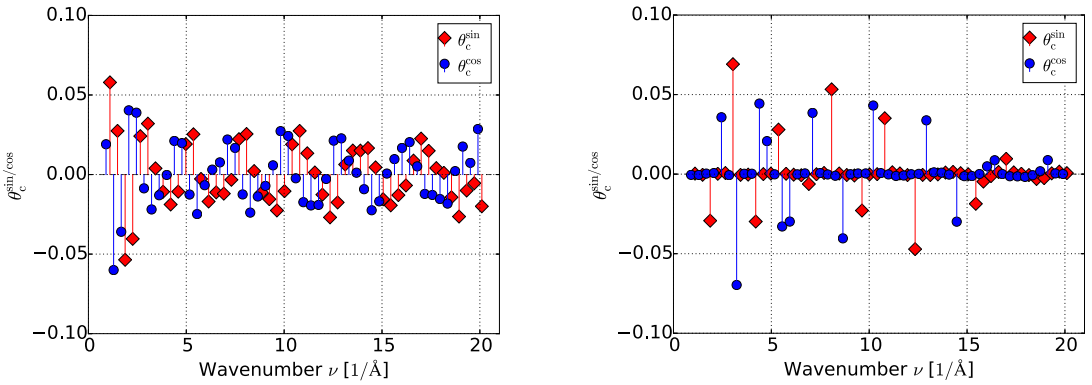
$$u^{LJ}(R;\boldsymbol{\theta}_c^{LJ}) = 4\epsilon\left(\left(\frac{\sigma_{LJ}}{R}\right)^{12} - \left(\frac{\sigma_{LJ}}{R}\right)^{6}\right), \tag{66}$$

where $\boldsymbol{\theta}_c^{LJ} = (\sigma_{LJ}, \epsilon)$. The total number of parameters associated with the two-body term was $K + 2 = 102$ and consists of $\boldsymbol{\theta}_c = (\boldsymbol{\theta}_c^{LJ}, \boldsymbol{\theta}_c^{cor})$. The ARD prior is employed only for $\boldsymbol{\theta}_c^{cor}$ and an (improper) uniform prior is employed for the rest $\boldsymbol{\theta}_c^{LJ}$. We note that due to the LJ part, the corresponding distribution p_c is not in the exponential family anymore (Section 2.4) and the possibility of multiple local maxima cannot be excluded.

3.2.4. Results

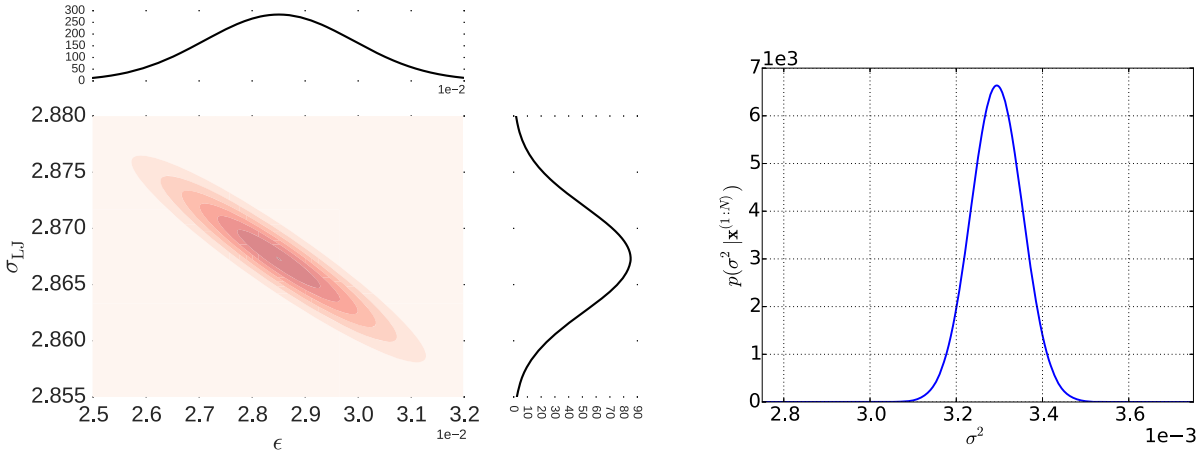
We first run the proposed algorithm for $N = 20$ fine-scale (all-atom) realizations. Fig. 14a depicts the evolution of the inferred coarse-scale potential $u^{(2)}(R;\boldsymbol{\theta}_c)$ (Eq. (63)) at various iterations of the EM-scheme. We initialize with $\boldsymbol{\theta}_c^{cor} = \mathbf{0}$ and $\boldsymbol{\theta}_c^{LJ} = (\epsilon = 0.15\,\frac{\text{kcal}}{\text{mol}}, \sigma_{LJ} = 3.5\,\text{Å})$. After 194 iterations, the converged result $u^{(2)}(R;\boldsymbol{\theta}_{c,MAP})$ is depicted with a solid black line. In Fig. 14b, we compare this converged result (red) with the two-body potential computed in [81] (dashed blue) using the relative entropy method and the LJ part (black) of the fine-scale SPC/E model. The former two exhibit similarities but also differences which stem from the different structure of these two models. These differences persist even if more training data N are used.

Fig. 15 depicts the effect of the ARD prior on $\boldsymbol{\theta}_c^{cor}$. One observes in Fig. 15a that if no such prior is used (instead a uniform was employed) almost all $\boldsymbol{\theta}_c^{cor}$ are non-zero and as a result almost all the corresponding feature functions $\phi_k(R)$ in Eq. (63) are active and the model is unable to distinguish their relative importance (unless N becomes very large). In contrast, the inclusion of the ARD prior in Fig. 15b leads to a sparse solution in which most $\phi_k(R)$ are deactivated (roughly



(a) Without ARD prior (b) With ARD prior

Fig. 15. $\theta^{\text{cor}}_{\text{c,MAP}}$ without and with the ARD prior with respect to the wavenumber ν_k (Eq. (64)). Superscripts sin (red) and cos (blue) indicate whether the corresponding $\theta^{\text{cor}}_{\text{c},k}$ (Eq. (63)) is associated with a sine or cosine feature function respectively.



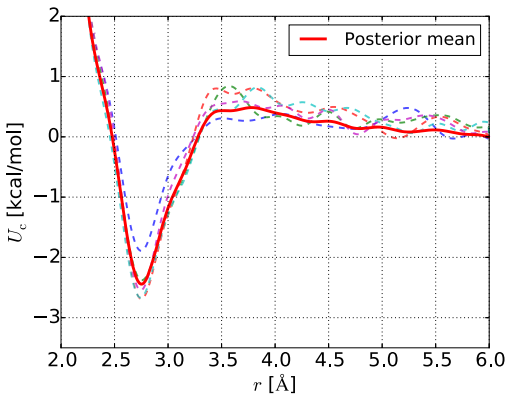
(a) Joint posterior $p(\sigma_{\text{LJ}}, \epsilon|\boldsymbol{x}^{(1:N)})$. (b) Posterior $p(\sigma^2|\boldsymbol{x}^{(1:N)})$.

Fig. 16. Posterior of $\boldsymbol{\theta}^{\text{LJ}}_{\text{c}} = (\sigma_{\text{LJ}}, \epsilon)$ in Eq. (66) and σ^2 in Eq. (59) for $N = 20$.

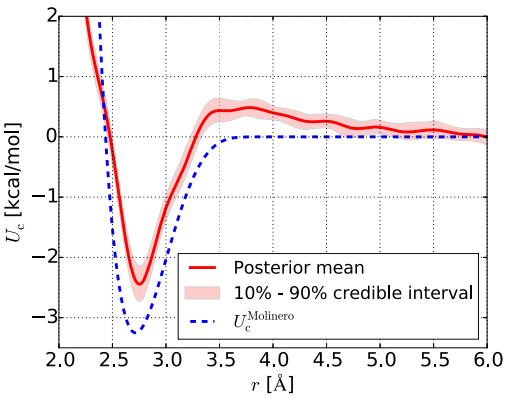
80 out of 100 in this case). It can be clearly seen as well that feature functions (sines/cosines) with high wave-numbers (small wave-lengths) are largely unnecessary for the description of the coarse potential. Although not demonstrated in this run, we envision that this modeling feature will eventually allow us to identify not only the most important terms in each potential term but also the most suitable order of interactions in the coarse potential. Fig. 16 depicts the (approximate) posterior obtained for $\boldsymbol{\theta}^{\text{LJ}}_{\text{c}} = (\sigma_{\text{LJ}}, \epsilon)$ (Eq. (66)) and σ^2 (Eq. (59)) for $N = 20$.

Fig. 17 provides information with regards to the (approximate) posterior of $\boldsymbol{\theta}_{\text{c}}$, computed using the Laplace's approximation proposed, as reflected in the $u^{(2)}(R; \boldsymbol{\theta}_{\text{c}})$. In particular in Fig. 17a, we plot sample realizations of $u^{(2)}(R; \boldsymbol{\theta}_{\text{c}})$ corresponding to different samples of $\boldsymbol{\theta}_{\text{c}}$ from the (approximate) Gaussian posterior (Section 2.6). We note that all realizations suggest the same location for the minimum of the potential. Variability is observed in the depth of this well as well as in its shape to the right of the minimum. Fig. 17b depicts the posterior mean of $u^{(2)}(R; \boldsymbol{\theta}_{\text{c}})$ as well as credible intervals at 10% and 90% posterior quantiles which reflect the inferential uncertainties discussed.

We finally report results illustrating the predictive capability of the model in terms of the macroscopic observables of interest i.e. the RDF and the angular distribution function discussed previously. To that end, we consider three data settings with $N = 10, 20$ and 100 fine-scale (all-atom) training data. While the MAP estimates do not exhibit prominent differences, the advantage of the method proposed is the predictive posterior that is furnished (Eq. (19)) and quantifies the uncertainty in the predictions that the coarse-grained model produces. Figs. 18 and 19 depict the posterior means and credible intervals corresponding to 10% and 90% posterior quantiles for the RDF $g(r)$ (i.e. the expected value of the observable in Eq. (57)) and the angular distribution function $p(\omega)$ (i.e. the expected value of the observable in Eq. (58)). In all cases, the posterior means are very close to the reference values obtained by simulating the all-atom SPC/E model. It is interesting to point out that when only $N = 10$ data were used, the posterior mean overestimates the first peak in the RDF (Fig. 18a). Nevertheless the true solution is contained within the credible intervals computed. As one would expect, the breath of the credible intervals decreases as more training data N is introduced, reflecting the reduction in the predictive uncertainty of the model. Details for the computation of these credible intervals can be found in A.1.

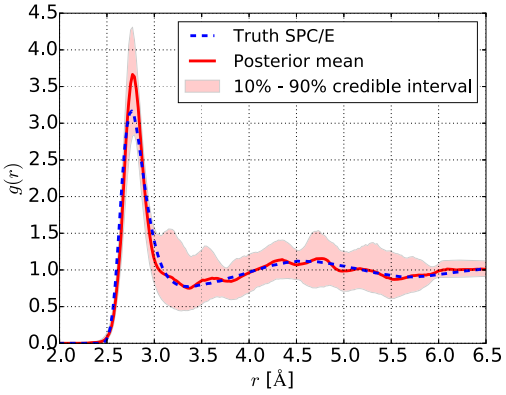


(a) Realizations of $u^{(2)}(R;\boldsymbol{\theta}_c)$ for random samples of $\boldsymbol{\theta}_c$ drawn from the approximate posterior.

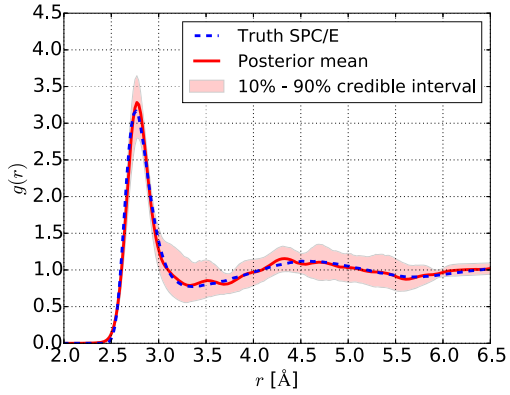


(b) Posterior mean and credible intervals for $u^{(2)}(R;\boldsymbol{\theta}_c)$. We compare this with with the two-body potential computed in [81] (dashed blue) using the relative entropy method.

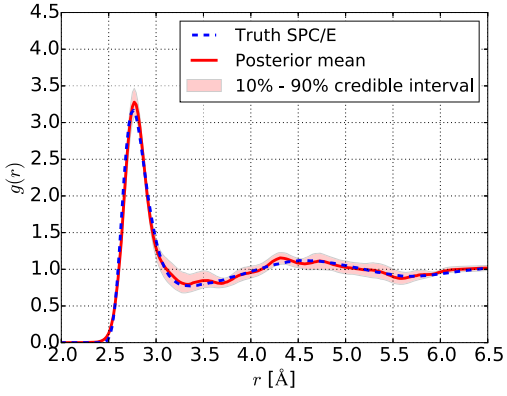
Fig. 17. Posterior of $u^{(2)}(R;\boldsymbol{\theta}_c)$ for $N = 20$.



(a) $N = 10$



(b) $N = 20$

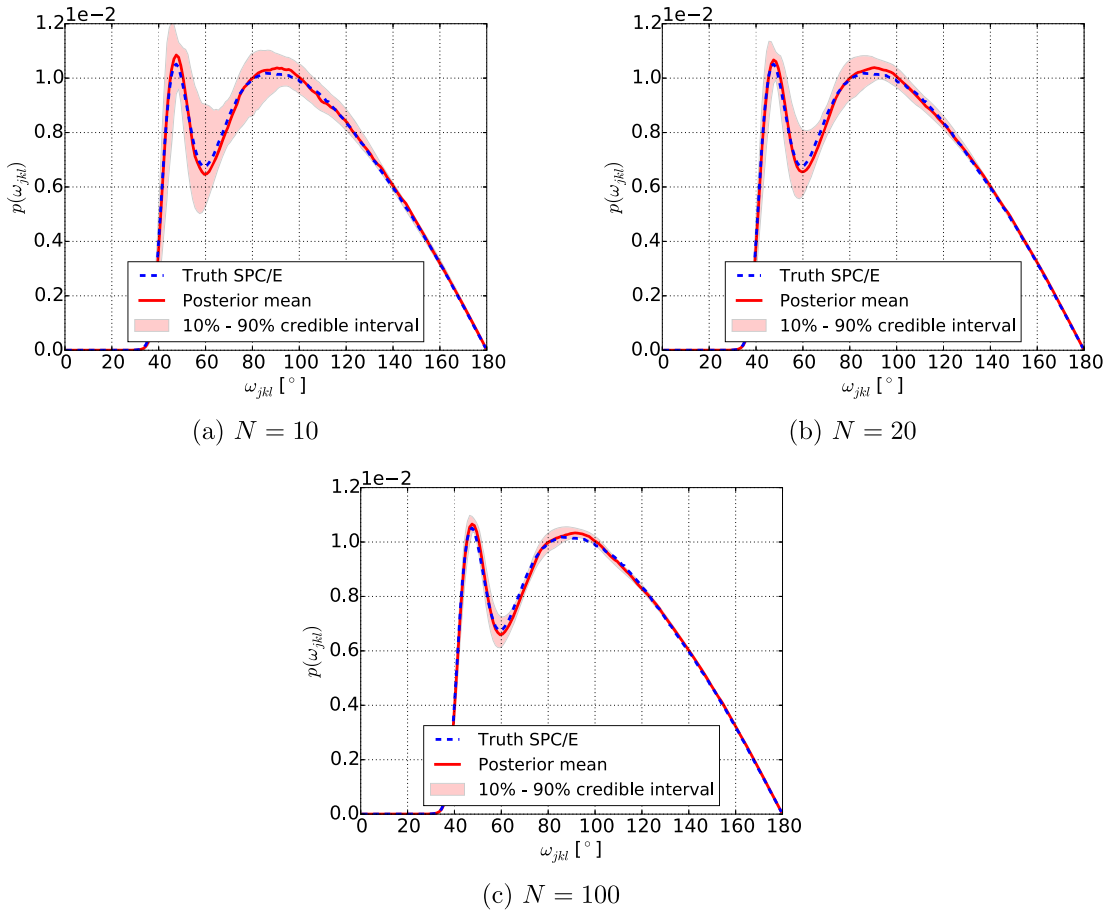


(c) $N = 100$

Fig. 18. Comparison of the reference RDF $g(r)$ (computed with all-atom simulations using the SPC/E model) with posterior mean and credible intervals corresponding to 10% and 90% posterior quantiles.

4. Conclusions

We presented a novel, data-driven coarse-graining scheme of atomistic ensembles in equilibrium. In contrast to existing techniques which are based on a restriction, fine-to-coarse map, we adopt the opposite strategy by prescribing a *probabilistic coarse-to-fine* map. This corresponds to a directed probabilistic model where the coarse variables play the role of latent generators of the fine scale (all-atom) data. Such a model can readily quantify the uncertainty due to the information loss



(a) $N = 10$ (b) $N = 20$

(c) $N = 100$

Fig. 19. Comparison of the reference ADF $p(\omega)$ (computed with all-atom simulations using the SPC/E model) with posterior mean and credible intervals corresponding to 10% and 90% posterior quantiles.

that unavoidably occurs during the CG process. We showed that from an information-theoretic perspective, the framework proposed broadens the relative entropy method. Furthermore, it can be readily extended to a fully Bayesian model where various sources of uncertainties are reflected in the posterior of the model parameters. The latter can be used to produce not only point estimates of fine-scale reconstructions or macroscopic observables, but more importantly, predictive posterior distributions on these quantities. We show how these can quantify the confidence of the model as a function of the amount of data and the level of coarse-graining, i.e. the contrast in the dimension between fine and coarse descriptions.

A critical issue in all CG methods pertains to the form of the coarse model or coarse potential. On one hand, it is desirable to introduce not only as many feature functions as possible but also to capture interactions of the highest-order possible. On the other hand, such an intricate representation leads to a large number of unknown parameters, augmented computational cost and an increased possibility of overfitting. Such challenges can be readily addressed within the Bayesian framework adopted by the incorporation of appropriate prior models that promote the discovery of sparse solutions and are capable of revealing the most dominant features in the coarse potential. We demonstrated how such a hierarchical prior model, namely the ARD, is capable of distinguishing the most prominent feature functions.

The computational engine of the proposed framework is based on an MC-EM scheme that alternates between expectations with respect to the posterior of the latent variables and maximization with respect to the model parameters. This leads to MAP estimates of the model parameters which serve as the basis for the Laplace's model that approximates their posterior. We note that this represents a very basic approximation that we intend to extend by exploiting advanced MCMC schemes [84] and/or variational inference schemes [85]. From a practical point of view, we note that the algorithm proposed is embarrassingly parallelizable with regards to the expectation step (which is also the most expensive) and incremental variants can be readily adopted leading to improvements in computational efficiency.

The generative definition of the CG variables through a probabilistic coarse-to-fine map allows for great flexibility in the type and number of CG variables used. For example in [23], the FG configuration space is partitioned and within each of these subdomains a different set of CG variables and CG models is learned. This is a reasonable strategy not only because a *globally*-good set of CG variables is difficult to find, but also because the local CG variables can be lower-dimensional as they need only to work on a limited subdomain. In the context of the directed, probabilistic model advocated, the same effect can be readily achieved by using a mixture model [86]. Consider for example augmenting the set of (latent) CG variables with a discrete-valued variable, S which can take values between 1 and L (which is the number of partitions). The (latent)

variable S characterizes a finite number of discrete states of the system. Depending on the value S takes, the number and type of CG variables $\boldsymbol{X}$ can change by affecting the two distributions making up the mode, i.e:

$$p_{\mathrm{c}}(\boldsymbol{X}, S = s|\boldsymbol{\theta}_{\mathrm{c}}) = p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}}^{s})p_{\mathrm{c}}(S = s), \tag{67}$$

where each $p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}}^{s})$ can be of the same or different form (e.g. exponential family) but with different parametrizations $\boldsymbol{\theta}_{\mathrm{c}}^{s}, s = 1, \dots L$. Similarly for the coarse-to-fine map, we can define:

$$p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, S = s, \boldsymbol{\theta}_{\mathrm{cf}}) = p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}}^{s}), \tag{68}$$

where again the parametrization can depend or not on S, $\boldsymbol{\theta}_{\mathrm{cf}}^{s}, s = 1, \dots L$. Infinite mixture models [87–89] based on Dirichlet process priors could provide a rigorous strategy on determining the number L of such hidden states needed to describe the atomistic ensemble. We note finally that, in nonequilibrium settings, by appropriate modeling of the time dependence of S one would recover Hidden Markov Models (HMM, [46]) which have been employed in coarse-graining frameworks [90,91].

Another potentially powerful extension, involves the use of deep, hierarchical models. Deep learning tools have revolutionized various machine learning tasks [92] by stacking multiple layers of simple representations. In the context of coarse-graining, such a scheme could be materialized by augmenting the set of CG variables as $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots \boldsymbol{X}_L$ and the CG model as:

$$p_{\mathrm{c}}(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots \boldsymbol{X}_L) = p_{\mathrm{c},1}(\boldsymbol{X}_1|\boldsymbol{X}_2, \boldsymbol{\theta}_{\mathrm{c}}^{1}) \dots p_{\mathrm{c},L-1}(\boldsymbol{X}_{L-1}|\boldsymbol{X}_L, \boldsymbol{\theta}_{\mathrm{c}}^{L-1})p_{\mathrm{c},L}(\boldsymbol{X}_L|\boldsymbol{\theta}_{\mathrm{c}}^{L}). \tag{69}$$

If $\dim(\boldsymbol{X}_1) > \dim(\boldsymbol{X}_2) > \dots > \dim(\boldsymbol{X}_L)$, then such a structure could provide a hierarchical decomposition of the CG picture, starting from a highly coarse description and gradually reaching the more detailed abstraction $\boldsymbol{X}_1$. The coarse-to-fine map could be controlled by $\boldsymbol{X}_1$ as $p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}_1, \boldsymbol{\theta}_{\mathrm{c}})$.

Acknowledgements

We acknowledge the support by the Hans Fisher Senior Fellowship of Nicholas Zabaras of the Technical University of Munich–Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement No. 291763. Nicholas Zabaras also acknowledges support from the Computer Science and Mathematics Division of ORNL under the DARPA EQUiPS program.

Appendix A. Methodology

A.1. Estimating credible intervals

This note summarizes necessary steps for estimating credible intervals. The Bayesian inference algorithms described in Sections 2.3 and 2.6, lead to (Gaussian) approximations of the posterior $p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})$ (Eq. (16)). The credible intervals shown in Figs. 11, 12, 18, and 19 are constructed from Monte Carlo samples $\hat{a}(\boldsymbol{\theta}^{(i)})$ of the observables of interest. These are generated on the basis of Eq. (19) as follows:

Algorithm 2 Estimating credible intervals.

1: **for** all $i = 1, \dots, I$ **do**
2: Obtain a posterior sample: $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})$ (Eq. (16)).
3: Calculate the predictive estimate $\hat{a}(\boldsymbol{\theta}^{(i)})$ shown in Eq. (19):

$$\hat{a}(\boldsymbol{\theta}^{(i)}) = \int a(\boldsymbol{x})\ p_{\mathrm{cf}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{\theta}_{\mathrm{cf}}^{(i)})\ p_{\mathrm{c}}(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{c}}^{(i)})\ d\boldsymbol{X}\ d\boldsymbol{x}. \tag{A.1}$$

The integrations involved are performed with Monte Carlo sampling. We note that this requires simulating only the CG model as the mapping implied by p_{cf} is straightforward.
4: **end for**
5: Compute desired quantiles with the given samples $\hat{a}(\boldsymbol{\theta}^{(1\dots I)})$.

We note that the estimated quantiles of the corresponding predictive posterior are not necessarily symmetric around its MAP estimate $\hat{a}(\boldsymbol{\theta}_{\mathrm{MAP}})$, even in the case of a symmetric posterior of the model's parameters $p(\boldsymbol{\theta}|\boldsymbol{x}^{(1:N)})$ (Eq. (16)).

A.2. Comparison of gradients between relative entropy method and PCG

This section compares the gradients with respect to the parameters of the coarse potential $\boldsymbol{\theta}_{\mathrm{c}}$, between the proposed scheme and the relative entropy method. These are used for fitting the model parameters $\boldsymbol{\theta}_{\mathrm{c}}$. In our case, the gradient is given by:

$$\frac{\partial \mathcal{F}}{\partial \theta_{c,k}} = \sum_{i=1}^{N} \left(< \phi_k(\boldsymbol{X}^{(i)}) >_{q_i(\boldsymbol{X}^{(i)})} - < \phi_k(\boldsymbol{X}) >_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)} \right), \tag{A.2}$$

whereas for the relative entropy method (when the objective $\mathcal{F}_{\mathrm{KL}}$ is given as in Eq. (10)):

$$\begin{aligned} \frac{\partial \mathcal{F}_{\mathrm{KL}}}{\partial \theta_{c,k}} &= \left(< \phi_k(\mathcal{R}(\boldsymbol{x})) >_{p_f(\boldsymbol{x})} - < \phi_k(\boldsymbol{X}) >_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)} \right) \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \left(\phi_k(\mathcal{R}(\boldsymbol{x}^{(i)})) - < \phi_k(\boldsymbol{X}) >_{p_c(\boldsymbol{X}|\boldsymbol{\theta}_c)} \right). \end{aligned} \tag{A.3}$$

In the latter case, the expectations with respect to $p_f(\boldsymbol{x})$ are estimated using the fine-scale data $\boldsymbol{x}^{(i)}$ whereas in the former these involve averaging over the *posterior* of the CG variables $\boldsymbol{X}$. This emphasizes the role of the CG variables play in our model as latent (hidden) generators of the fine-scale.

A.3. ARD prior

We adopt the Automatic Relevance Determination (ARD, [66]) which is formulated in the context of hierarchical Bayesian models. The prior on the parameters $\boldsymbol{\theta}_c$ is modeled as independent Gaussian for each $\theta_{c,k}$ with zero mean and precision hyper-parameter τ_k:

$$p(\boldsymbol{\theta}_c|\boldsymbol{\tau}) \equiv \prod_k \underbrace{\mathcal{N}(\theta_{c,k}|0, \tau_k^{-1})}_{p(\theta_{c,k}|\tau_k)}. \tag{A.4}$$

The precision (hyper-)parameters τ_k follow a Gamma distribution,

$$\tau_k \sim Gamma(\tau_k|a_0, b_0). \tag{A.5}$$

Anytime derivatives of the log-prior are needed, an inner-loop Expectation–Maximization scheme can be employed which is based on the same ideas presented previously. In particular, for any set of densities $q_k(\tau_k)$ we can obtain a lower bound on the log-prior as follows:

$$\begin{aligned} \log p(\boldsymbol{\theta}_c) &= \log \left(\prod_k \int p(\theta_{c,k}|\tau_k)\ p(\tau_k|a_0, b_0)\, d\tau_k \right) \\ &= \sum_k \log \int q_k(\tau_k) \frac{p(\theta_{c,k}|\tau_k)\ p(\tau_k|a_0, b_0)}{q(\tau_k)}\, d\tau_k \\ &\geq \sum_k \int q_k(\tau_k) \log \frac{p(\theta_{c,k}|\tau_k)\ p(\tau_k|a_0, b_0)}{q_k(\tau_k)}\, d\tau_k \qquad \text{(Jensen's inequality)} \end{aligned} \tag{A.6}$$

The optimal q_k i.e. the posteriors $p(\tau_k|\theta_{c,k})$ (for which the lower bound becomes tight) can be analytically computed and are Gamma densities with parameters $a_k = a_0 + \frac{1}{2},\ b_k = b_0 + \frac{\theta_{c,k}^2}{2}$ [67], where the current values of $\theta_{c,k}$'s are used. This leads to the extremely simple iterations of the following form [67]:

- E-step: evaluate:

$$\langle \tau_k \rangle_{p(\tau_k|\theta_{c,k})} = \frac{a_k}{b_k} = \frac{a_0 + \frac{1}{2}}{b_0 + \frac{\theta_{c,k}^2}{2}}. \tag{A.7}$$

- M-step: evaluate:

$$\begin{aligned} \frac{\partial \log p(\boldsymbol{\theta}_c)}{\partial \theta_{c,k}} &= \frac{\partial}{\partial \theta_{c,k}} \int q_k(\tau_k) \log p(\theta_{c,k}|\tau_k)\, d\tau_k \\ &= -\int q_k(\tau_k) \tau_k\, d\tau_k\, \theta_{c,k} \\ &= -\langle \tau_k \rangle_{p(\tau_k|\theta_{c,k})}\, \theta_{c,k}. \end{aligned} \tag{A.8}$$

Appendix B. Numerical examples

B.1. SPC/E model, parameters and simulation details

The following SPC/E parameters as given in [36,75] are used for producing the fine-scale data.

- LJ-potential: $\sigma = 3.166\,\text{Å}$, $\epsilon = 0.650\,\frac{\text{kJ}}{\text{mol}}$.
- Electrostatic load: $q_\text{H} = +0.4238\,e$, $q_\text{O} = -0.8476\,e$.
- Structural properties of rigid water model: bond-length $l_\text{OH} = 1.0\,\text{Å}$ and bond-angle $\theta_\text{HOH} = 109.47°$.
- Masses: $m_\text{O} = 15.994\,\frac{\text{g}}{\text{mol}}$ and $m_\text{H} = 1.00794\,\frac{\text{g}}{\text{mol}}$.

B.1.1. Simulation steps

In this work, we consider a system of $N_w = 100$ water molecules at a temperature $T = 300\,\text{K}$. The following steps for obtaining training data are performed:

1. NPT simulation with $p = 1\,\text{bar}$ and a timestep of $\Delta t = 2.0\,\text{fs}$. Simulate the system for $t = 100\,\text{ns}$.
2. Use the last $t = 80\,\text{ns}$ for calculating the equilibrium box size. We found $l_\text{box} = 14.5459665\,\text{Å}$.
3. Fix the box length to the one obtained from the previous step. Simulate the system in NVT ensemble for $t = 45\,\text{ns}$ with a timestep of $\Delta t = 2.0\,\text{fs}$. Use the last $t = 40\,\text{ns}$ and write the trajectory every 200 steps.

B.2. Radial distribution function

The radial distribution function $g(r)$ is defined by,

$$g(r) = \left\langle \frac{V}{N^2} a^\text{RDF}(r) \right\rangle.$$

The discrete version follows with the number of bins n_bin and a bin size Δr:

$$g(r_1) = \frac{1}{N n_\text{bin}} \frac{\langle a^\text{RDF}(r_1) \rangle}{\rho_\text{ideal}},$$

with,

$$\rho_\text{ideal} = N/V,$$

$$a^\text{RDF}(r_1) = \frac{n(r_1)}{\Delta V} = \frac{\sum_{ij} \int_{r_1}^{r_1+\Delta r} \delta(r_{ij} - r) dr}{\frac{4}{3}\pi((r_1 + \Delta r)^3 - r_1^3)}.$$

B.3. Stillinger–Weber (SW) potential

The Stillinger–Weber (SW) potential originally proposed in [83] and extended in [81], contains both two- and three-body interactions. In this work, we make use only of the latter three-body contribution:

$$U^\text{SW}(\boldsymbol{X}) = \sum_j \sum_{k \neq j} \sum_{l > k} \phi_3^\text{SW}(r_{jk}, R_{jl}, \omega_{jkl}), \tag{B.1}$$

where the three-body term $\phi_3^\text{SW}(r_{jk}, r_{jl}, \omega_{jkl})$ is given by:

$$\phi_3^\text{SW}(r_{jk}, r_{jl}, \omega_{jkl}) = \lambda\epsilon \left[\cos\omega_{jkl} - \cos\omega_0\right]^2 \exp\left(\frac{\gamma\sigma}{r_{jk} - a_3\sigma_\text{SW}}\right) \exp\left(\frac{\gamma\sigma}{r_{jl} - a_3\sigma_\text{SW}}\right), \tag{B.2}$$

with r_{jk} being the pairwise distances between molecules j and k and ω_{jkl} is the angle between molecules j, k, l. The following values for the parameters were used [81]: $\lambda = 0.762$, $\epsilon = 83.5737$, $\cos\omega_0 = -0.487217$, $\gamma = 0.291321$, $a_3 = 0.586097$, $\sigma_\text{SW} = 6.4144$.

References

[1] M.S. Shell, The relative entropy is fundamental to multiscale and inverse thermodynamic problems, J. Chem. Phys. 129 (14) (2008) 144108.
[2] B.J. Alder, T.E. Wainwright, Studies in molecular dynamics. I. General method, J. Chem. Phys. 31 (2) (1959) 459–466, http://dx.doi.org/10.1063/1.1730376, http://scitation.aip.org/content/aip/journal/jcp/31/2/10.1063/1.1730376.
[3] M. Karplus, J.A. McCammon, Molecular dynamics simulations of biomolecules, Nat. Struct. Biol. 9 (9) (2002) 646–652, http://dx.doi.org/10.1038/nsb0902-646.
[4] M.J. Buehler (Ed.), Atomistic Modeling of Materials Failure, Springer US, Boston, MA, 2008, http://link.springer.com/10.1007/978-0-387-76426-9.

[5] C. Peter, K. Kremer, Multiscale simulation of soft matter systems – from the atomistic to the coarse-grained level and back, Soft Matter 5 (2009) 4357–4366, http://dx.doi.org/10.1039/B912027K.
[6] G. Voth.
[7] T. Lelièvre, M. Rousset, G. Stoltz, Free Energy Computations: A Mathematical Perspective, Imperial College Press, London, Hackensack (N.J.), Singapore, 2010, http://opac.inria.fr/record=b1131369.
[8] I. Bilionis, P.S. Koutsourelakis, Free energy computations by minimization of Kullback–Leibler divergence: an efficient adaptive biasing potential method for sparse representations, J. Comput. Phys. 231 (9) (2012) 3849–3870, http://dx.doi.org/10.1016/j.jcp.2012.01.033, wOS:000302501500020.
[9] M. Katsoulakis, A. Majda, D. Vlachos, Coarse-grained stochastic processes and Monte Carlo simulations in lattice systems, J. Comput. Phys. 186 (1) (2003) 250–278, http://dx.doi.org/10.1016/S0021-9991(03)00051-2.
[10] A. Chatterjee, D.G. Vlachos, M.A. Katsoulakis, Spatially adaptive lattice coarse-grained Monte Carlo simulations for diffusion of interacting molecules, J. Chem. Phys. 121 (22) (2004) 11420–11431, http://scitation.aip.org/content/aip/journal/jcp/121/22/10.1063/1.1811601.
[11] M.A. Katsoulakis, P. Plecháč, A. Sopasakis, Error analysis of coarse-graining for stochastic lattice dynamics, SIAM J. Numer. Anal. 44 (6) (2006) 2270–2296, http://epubs.siam.org/doi/abs/10.1137/050637339.
[12] M.A. Katsoulakis, P. Plecháč, L. Rey-Bellet, Numerical and statistical methods for the coarse-graining of many-particle stochastic systems, J. Sci. Comput. 37 (1) (2008) 43–71, http://dx.doi.org/10.1007/s10915-008-9216-6, http://www.springerlink.com/index/10.1007/s10915-008-9216-6.
[13] E. Kalligiannaki, M.A. Katsoulakis, P. Plecháč, D.G. Vlachos, Multilevel coarse graining and nano-pattern discovery in many particle stochastic systems, J. Comput. Phys. 231 (6) (2012) 2599–2620, http://dx.doi.org/10.1007/s10915-008-9216-6, http://www.sciencedirect.com/science/article/pii/S0021999111007212.
[14] M.A. Katsoulakis, P. Plecháč, Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems, J. Chem. Phys. 139 (7) (2013) 074115, http://dx.doi.org/10.1063/1.4818534, http://scitation.aip.org/content/aip/journal/jcp/139/7/10.1063/1.4818534.
[15] W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, O. Hahn, Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates, Acta Polym. 49 (2–3) (1998) 61–74.
[16] D. Reith, M. Pütz, F. Müller-Plathe, Deriving effective mesoscale potentials from atomistic simulations, J. Comput. Chem. 24 (13) (2003) 1624–1636, http://dx.doi.org/10.1002/jcc.10307.
[17] A.P. Lyubartsev, A. Laaksonen, Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach, Phys. Rev. E 52 (1995) 3730–3737.
[18] J.F. Rudzinski, W.G. Noid, A generalized-Yvon–Born–Green method for coarse-grained modeling, Eur. Phys. J. Spec. Top. 224 (12) (2015) 2193–2216, http://dx.doi.org/10.1140/epjst/e2015-02408-9, http://link.springer.com/article/10.1140/epjst/e2015-02408-9.
[19] A. Savelyev, G.A. Papoian, Molecular renormalization group coarse-graining of polymer chains: application to double-stranded {DNA}, Biophys. J. 96 (10) (2009) 4044–4052, http://dx.doi.org/10.1016/j.bpj.2009.02.067, http://www.sciencedirect.com/science/article/pii/S0006349509006729.
[20] R.H. Swendsen, Monte Carlo renormalization group, Phys. Rev. Lett. 42 (14) (1979) 859.
[21] S. Izvekov, G.A. Voth, Multiscale coarse graining of liquid-state systems, J. Chem. Phys. 123 (13) (2005) 134105.
[22] W.G. Noid, J. Chu, G.S. Ayton, G.A. Voth, Multiscale coarse-graining and structural correlations: connections to liquid-state theory, J. Phys. Chem. B 111 (16) (2007) 4116–4127.
[23] J.F. Dama, A.V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A.R. Dinner, G.A. Voth, The theory of ultra-coarse-graining. 1. General principles, J. Chem. Theory Comput. 9 (5) (2013) 2466–2480, http://dx.doi.org/10.1021/ct4000444, pMID: 26583735.
[24] A.R. Leach, Molecular Modelling, Prentice Hall, 2001, http://www.ebook.de/de/product/3246977/andrew_r_leach_molecular_modelling.html.
[25] M.A. Katsoulakis, J. Trashorras, Information loss in coarse-graining of stochastic particle dynamics, J. Stat. Phys. 122 (1) (2006) 115–135, http://link.springer.com/article/10.1007/s10955-005-8063-1.
[26] T.T. Foley, M.S. Shell, W.G. Noid, The impact of resolution upon entropy and information in coarse-grained models, J. Chem. Phys. 143 (24) (2015) 243104, http://dx.doi.org/10.1063/1.4929836, http://scitation.aip.org/content/aip/journal/jcp/143/24/10.1063/1.4929836.
[27] M.A. Katsoulakis, P. Plecháč, A. Sopasakis, Error analysis of coarse-graining for stochastic lattice dynamics, SIAM J. Numer. Anal. 44 (6) (2006) 2270–2296, http://dx.doi.org/10.1137/050637339.
[28] J. Trashorras, D. Tsagkarogiannis, From mesoscale back to microscale: reconstruction schemes for coarse-grained stochastic lattice systems, SIAM J. Numer. Anal. 48 (5) (2010) 1647–1677, http://dx.doi.org/10.1137/080722382, https://hal.archives-ouvertes.fr/hal-00275802.
[29] M.A. Rohrdanz, W. Zheng, C. Clementi, Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions, Annu. Rev. Phys. Chem. 64 (2013) 295–316, http://www.annualreviews.org/doi/abs/10.1146/annurev-physchem-040412-110006.
[30] W.G. Noid, Perspective: Coarse-grained models for biomolecular systems, J. Chem. Phys. 139 (9) (2013), http://dx.doi.org/10.1063/1.4818908, http://scitation.aip.org/content/aip/journal/jcp/139/9/10.1063/1.4818908.
[31] P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework, J. Chem. Phys. 137 (14) (2012), http://dx.doi.org/10.1063/1.4757266, http://scitation.aip.org/content/aip/journal/jcp/137/14/10.1063/1.4757266.
[32] P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Data driven, predictive molecular dynamics for nanoscale flow simulations under uncertainty, J. Phys. Chem. B 117 (47) (2013) 14808–14816.
[33] K. Farrell, J.T. Oden, Calibration and validation of coarse-grained models of atomic systems: application to semiconductor manufacturing, Comput. Mech. 54 (1) (2014) 3–19, http://dx.doi.org/10.1007/s00466-014-1028-y.
[34] K. Farrell, J.T. Oden, D. Faghihi, A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems, J. Comput. Phys. 295 (2015) 189–208, http://dx.doi.org/10.1016/j.jcp.2015.03.071, http://www.sciencedirect.com/science/article/pii/S0021999115002430.
[35] A. Chaimovich, M.S. Shell, Coarse-graining errors and numerical optimization using a relative entropy framework, J. Chem. Phys. 134 (9) (2011) 094112, http://dx.doi.org/10.1063/1.3557038, http://www.ncbi.nlm.nih.gov/pubmed/21384955.
[36] I. Bilionis, N. Zabaras, A stochastic optimization approach to coarse-graining using a relative-entropy framework, J. Chem. Phys. 138 (4) (2013) 044313, http://dx.doi.org/10.1063/1.4789308, http://www.ncbi.nlm.nih.gov/pubmed/23387590.
[37] E. Cances, F. Legoll, G. Stoltz, Theoretical and numerical comparison of some sampling methods for molecular dynamics, edpsciences.org, http://www.edpsciences.org/articles/m2an/abs/2007/02/m2an0588/m2an0588.html.
[38] J.F. Rudzinski, W.G. Noid, Coarse-graining entropy, forces, and structures, J. Chem. Phys. 135 (2011) 214101, http://dx.doi.org/10.1063/1.3663709, http://scitation.aip.org/content/aip/journal/jcp/135/21/10.1063/1.3663709.
[39] T. Cover, J. Thomas, Elements of Information Theory, John Wiley & Sons, 1991.
[40] C. Bishop, Latent variable models, in: M.I. Jordan (Ed.), Learning in Graphical Models, MIT Press, 1999, pp. 371–403.
[41] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
[42] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. R. Stat. Soc. B 39 (1) (1977) 1–38.
[43] R. Neal, G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in: Learning in Graphical Models, Kluwer Academic Publishers, 1998, pp. 355–368.
[44] G.C.G. Wei, M.A. Tanner, A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, J. Am. Stat. Assoc. 85 (411) (1990) 699–704, http://dx.doi.org/10.1080/01621459.1990.10474930, http://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474930.

[45] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. 22 (3) (1951) 400–407.
[46] O. Cappe, E. Moulines, T. Ryden, Inference in Hidden Markov Models, Springer-Verlag, 2005.
[47] M.J. Beal, Variational Algorithms for Approximate Bayesian Inference, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003, http://www.cse.buffalo.edu/faculty/mbeal/thesis/index.html.
[48] L. Younes, On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates, Stoch. Stoch. Rep. 65 (3–4) (1999) 177–228, http://dx.doi.org/10.1080/17442509908834179, http://www.tandfonline.com/doi/abs/10.1080/17442509908834179.
[49] C. Andrieu, E. Moulines, P. Priouret, Stability of stochastic approximation under verifiable conditions, SIAM J. Control Optim. 44 (1) (2005) 283–312, http://dx.doi.org/10.1137/S0363012902417267, http://epubs.siam.org/doi/abs/10.1137/S0363012902417267.
[50] G. Fort, E. Moulines, A. Schreck, M. Vihola, Convergence of Markovian stochastic approximation with discontinuous dynamics, SIAM J. Control Optim. (2016) 866–893, http://dx.doi.org/10.1137/140962723.
[51] J.G. Booth, J.P. Hobert, Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, J. R. Stat. Soc., Ser. B, Stat. Methodol. 61 (1) (1999) 265–285, http://www.jstor.org/stable/2680750.
[52] R.A. Levine, G. Casella, Implementations of the Monte Carlo EM algorithm, J. Comput. Graph. Stat. 10 (3) (2001) 422–439, http://www.jstor.org/stable/1391097.
[53] G. Fort, E. Moulines, Convergence of the Monte Carlo expectation maximization for curved exponential families, Ann. Stat. 31 (4) (2003) 1220–1259, http://dx.doi.org/10.1214/aos/1059655912, http://projecteuclid.org/euclid.aos/1059655912.
[54] R.A. Levine, J. Fan, An automated (Markov chain) Monte Carlo EM algorithm, J. Stat. Comput. Simul. 74 (5) (2004) 349–360, http://dx.doi.org/10.1080/0094965031000147704.
[55] J.S. Liu, Monte Carlo Strategies in Scientific Computing, Springer Publishing Company, 2008, Incorporated.
[56] P. Del Moral, Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications, Springer, New York, 2004.
[57] J.C. Spall, Introduction to Stochastic Search and Optimization, 1st edition, John Wiley & Sons, Inc., New York, NY, USA, 2003.
[58] S. Mohamed, K.A. Heller, Z. Ghahramani, Bayesian exponential family PCA, in: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (Eds.), Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8–11, 2008, in: Advances in Neural Information Processing Systems, vol. 21, Curran Associates, Inc., 2008, pp. 1089–1096, http://papers.nips.cc/paper/3532-bayesian-exponential-family-pca.
[59] P. Moritz, R. Nishihara, M.I. Jordan, A linearly-convergent stochastic L-BFGS, algorithm.
[60] R.H. Byrd, G.M. Chin, W. Neveitt, J. Nocedal, On the use of stochastic Hessian information in optimization methods for machine learning, SIAM J. Optim. 21 (3) (2011) 977–995.
[61] C. Chen, D. Carlson, Z. Gan, C. Li, L. Carin, Bridging the gap between stochastic gradient MCMC and stochastic optimization.
[62] H. Kushner, G.G. Yin, Stochastic Approximation and Recursive Algorithms and Applications, vol. 35, Springer Science & Business Media, 2003.
[63] S. Della Pietra, V. Della Pietra, J. Lafferty, Inducing features of random fields, IEEE Trans. Pattern Anal. Mach. Intell. 19 (4) (1997) 380–393.
[64] M.A.T. Figueiredo, Adaptive sparseness for supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1150–1159.
[65] M. West, Bayesian factor regression models in the “large p, small n” paradigm, in: J. Bernardo, M. Bayarri, J. Berger, A.P. Dawid, D. Heckerman, A. Smith, M. West (Eds.), Bayesian Stat., vol. 7, 2003.
[66] D.J.C. MacKay, R.M. Neal, Automatic Relevance Determination for Neural Networks, Tech. rep., University of Cambridge, 1994.
[67] C.M. Bishop, M.E. Tipping, Variational relevance vector machines, in: UAI, 2000, pp. 46–53.
[68] D.J.C. MacKay, Information Theory, Inference & Learning Algorithms, Cambridge University Press, New York, NY, USA, 2002.
[69] J.V. Selinger, Introduction to the Theory of Soft Matter, Springer-Verlag GmbH, 2015, http://www.ebook.de/de/product/24265794/jonathan_v_selinger_introduction_to_the_theory_of_soft_matter.html.
[70] N. Ashcroft, N. Mermin, Solid State Physics, Saunders College, Philadelphia, 1976.
[71] S. Are, M.A. Katsoulakis, P. Plecháč, L.R. Bellet, Multibody interactions in coarse-graining schemes for extended systems, SIAM J. Sci. Comput. 31 (2) (2008) 987–1015, https://dx.doi.org/10.1137/080713276.
[72] H.J.C. Berendsen, J.R. Grigera, T.P. Straatsma, The missing term in effective pair potentials, J. Phys. Chem. 91 (24) (1987) 6269–6271, http://dx.doi.org/10.1021/j100308a038.
[73] P.G. Kusalik, I.M. Svishchev, The spatial structure in liquid water, Science 265 (5176) (1994) 1219–1221, http://dx.doi.org/10.1126/science.265.5176.1219, http://science.sciencemag.org/content/265/5176/1219.
[74] A. Chaimovich, M.S. Shell, Tetrahedrality and structural order for hydrophobic interactions in a coarse-grained water model, Phys. Rev. E 89 (2014) 022140, http://dx.doi.org/10.1103/PhysRevE.89.022140, http://link.aps.org/doi/10.1103/PhysRevE.89.022140.
[75] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, D. Andrienko, Versatile object-oriented toolkit for coarse-graining applications, J. Chem. Theory Comput. 5 (12) (2009) 3211–3223.
[76] R. Erban, Coupling all-atom molecular dynamics simulations of ions in water with Brownian dynamics, Proc. R. Soc. Lond. A, Math. Phys. Eng. Sci. 472 (2186) (2016), http://dx.doi.org/10.1098/rspa.2015.0556, http://rspa.royalsocietypublishing.org/content/472/2186/20150556.
[77] W.G. Hoover, Canonical dynamics: equilibrium phase-space distributions, Phys. Rev. A 31 (1985) 1695–1697, http://dx.doi.org/10.1103/PhysRevA.31.1695, http://link.aps.org/doi/10.1103/PhysRevA.31.1695.
[78] S. Nosé, A unified formulation of the constant temperature molecular dynamics methods, J. Chem. Phys. 81 (1) (1984) 511–519, http://dx.doi.org/10.1063/1.447334, http://scitation.aip.org/content/aip/journal/jcp/81/1/10.1063/1.447334.
[79] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, J. Comput. Phys. 117 (1) (1995) 1–19, http://dx.doi.org/10.1006/jcph.1995.1039, http://www.sciencedirect.com/science/article/pii/S002199918571039X.
[80] G.N. Clark, C.D. Cappa, J.D. Smith, R.J. Saykally, T. Head-Gordon, The structure of ambient water, Mol. Phys. 108 (11) (2010) 1415–1433.
[81] J. Lu, Y. Qiu, R. Baron, V. Molinero, Coarse-graining of TIP4P/2005, TIP4P-Ew, SPC/E, and TIP3P to monatomic anisotropic water models using relative entropy minimization, J. Chem. Theory Comput. 10 (9) (2014) 4104–4120, http://dx.doi.org/10.1021/ct500487h, pMID: 26588552.
[82] H. Wang, C. Junghans, K. Kremer, Comparative atomistic and coarse-grained study of water: what do we lose by coarse-graining?, Eur. Phys. J. E 28 (2) (2009) 221–229, http://dx.doi.org/10.1140/epje/i2008-10413-5.
[83] F.H. Stillinger, T.A. Weber, Computer simulation of local order in condensed phases of silicon, Phys. Rev. B 31 (1985) 5262–5271, http://dx.doi.org/10.1103/PhysRevB.31.5262, http://link.aps.org/doi/10.1103/PhysRevB.31.5262.
[84] F. Liang, A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants, J. Stat. Comput. Simul. 80 (9) (2010) 1007–1022, http://dx.doi.org/10.1080/00949650902882162.
[85] M. Wainwright, M. Jordan, Graphical models, exponential families, and variational inference, in: Foundations and Trends in Machine Learning, vol. 1, 2008, pp. 1–305.
[86] C.M. Bishop, M. Svenskn, Bayesian hierarchical mixtures of experts, in: Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, UAI'03, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003, pp. 57–64, http://dl.acm.org/citation.cfm?id=2100584.2100591.
[87] C. Antoniak, Mixtures of Dirichlet processes with applications to nonparametric Bayesian problems, Ann. Stat. 2 (1974) 1152–1174.
[88] C.E. Rasmussen, The infinite Gaussian mixture model, in: Advances in Neural Information Processing Systems 12, NIPS Conference, Denver, Colorado, USA, November 29–December 4, 1999, 1999, pp. 554–560, http://papers.nips.cc/paper/1745-the-infinite-gaussian-mixture-model.

[89] P. Chen, N. Zabaras, I. Bilionis, Uncertainty propagation using infinite mixture of Gaussian processes and variational Bayesian inference, J. Comput. Phys. 284 (2015) 291–333, http://dx.doi.org/10.1016/j.jcp.2014.12.028, http://www.sciencedirect.com/science/article/pii/S0021999114008456.
[90] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, C. Schütte, Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models, J. Comput. Chem. 28 (2007) 2453–2464.
[91] I. Horenko, F. Noe, C. Hartmann, C. Schütte, Data-based parameter estimation of generalized multidimensional Langevin processes, Phys. Rev. E 78 (2007) 016706.
[92] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, http://dx.doi.org/10.1038/nature14539, http://www.nature.com/nature/journal/v521/n7553/full/nature14539.html.

Appendix C

Bayesian coarse-graining and adaptive sequential model refinement

C.1 Observable estimation for ALA-2

We are interested in estimating observables based on predictive models, in contrast to those obtained through reference MD simulations. In general, observables are evaluated as ensemble (MC) or phase (MD) averages, $\int a(\mathbf{x}) p_{\text{target}}(\mathbf{x})\, d\mathbf{x}$, by making use of $q_{\boldsymbol{\theta}}(\mathbf{x})$ and samples drawn by ancestral sampling.

C.1.1 Radius of gyration

We illustrate the radius of gyration (Rg) [68, 556], given as:

$$a_{\text{Rg}}(\mathbf{x}) = \sqrt{\frac{\sum_p m_p \|\mathbf{x}_p - \mathbf{x}_{\text{COM}}\|^2}{\sum_p m_p}} \tag{C.1}$$

The sum in Equation (C.1) considers all system atoms $p = 1, \ldots, P$, with the atom mass m_p and Cartesian coordinate $\mathbf{x}_p$ of each atom. The center of mass of the peptide is denoted by $\mathbf{x}_{\text{COM}}$. A histogram of $a_{\text{Rg}}(\mathbf{x})$ reflects the statistics of the peptide's average size, which characterize its various conformations [68].

C.1.2 Root-mean-square deviation

The root-mean-square deviation (RMSD) from a reference atomistic configuration provides relevant structural information in the context of biochemistry. It often refers to a α helical configuration which we denote with $\mathbf{x}_{\alpha\text{-ref}}$. Histograms on the RMSD reveal the frequency of deviations within a certain range. We calculate the RMSD with

$$a_{\text{RMSD}}(\mathbf{x}) = \sqrt{\frac{1}{P} \sum_{p=1}^{P} \left(\mathbf{x}_p - \mathbf{x}_{\alpha\text{-ref},p}\right)^2}. \tag{C.2}$$

Appendix D

Predictive collective variable discovery with deep Bayesian models

Reproduced from

M. Schöberl, N. Zabaras, P.-S. Koutsourelakis.
"Predictive collective variable discovery with deep Bayesian models".
In: *AIP The Journal of Chemical Physics* 150 (2019), 024109,

with the permission of AIP Publishing.

The Journal of Chemical Physics

Predictive collective variable discovery with deep Bayesian models

Cite as: J. Chem. Phys. **150**, 024109 (2019); https://doi.org/10.1063/1.5058063
Submitted: 16 September 2018 . Accepted: 23 December 2018 . Published Online: 14 January 2019

Markus Schöberl, Nicholas Zabaras, and Phaedon-Stelios Koutsourelakis

View Online Export Citation CrossMark

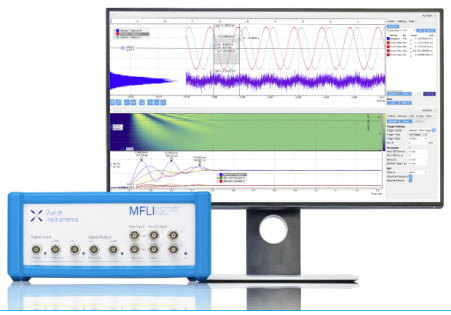
ARTICLES YOU MAY BE INTERESTED IN

Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)
The Journal of Chemical Physics **149**, 072301 (2018); https://doi.org/10.1063/1.5025487

Perspective: Identification of collective variables and metastable states of protein dynamics
The Journal of Chemical Physics **149**, 150901 (2018); https://doi.org/10.1063/1.5049637

Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics
The Journal of Chemical Physics **148**, 241703 (2018); https://doi.org/10.1063/1.5011399

Challenge us.
What are your needs for periodic signal detection?
Zurich Instruments
Watch



© 2019 Author(s).

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Predictive collective variable discovery with deep Bayesian models

Cite as: J. Chem. Phys. 150, 0241 09 (201 9); doi: 1 0.1 063/1 .5058063
Submitted: 1 6 September 201 8 • Accepted: 23 December 201 8 •
Published Online: 1 4 January 201 9

View Online Export Citation CrossMark

Markus Schöberl,[1,2,a] Nicholas Zabaras,[1,b] and Phaedon-Stelios Koutsourelakis[2,c]

AFFILIATIONS

[1] Center for Informatics and Computational Science, University of Notre Dame, 311 Cushing Hall, Notre Dame, Indiana 46556, USA
[2] Continuum Mechanics Group, Technical University of Munich, Boltzmannstraße 15, 85748 Garching, Germany

[a] Electronic mail: mschoeberl@gmail.com
[b] Electronic mail: nzabaras@gmail.com. URL: https://cics.nd.edu.
[c] Electronic mail: p.s.koutsourelakis@tum.de. URL: http://www.contmech.mw.tum.de.

ABSTRACT

Extending spatio-temporal scale limitations of models for complex atomistic systems considered in biochemistry and materials science necessitates the development of enhanced sampling methods. The potential acceleration in exploring the configurational space by enhanced sampling methods depends on the choice of collective variables (CVs). In this work, we formulate the discovery of CVs as a Bayesian inference problem and consider the CVs as hidden generators of the full-atomistic trajectory. The ability to generate samples of the fine-scale atomistic configurations using limited training data allows us to compute estimates of observables as well as our probabilistic confidence on them. The methodology is based on emerging methodological advances in machine learning and variational inference. The discovered CVs are related to physicochemical properties which are essential for understanding mechanisms especially in unexplored complex systems. We provide a quantitative assessment of the CVs in terms of their predictive ability for alanine dipeptide (ALA-2) and ALA-15 peptide.

Published under license by AIP Publishing. https://doi.org/1 0.1 063/1 .5058063

I. INTRODUCTION

Molecular dynamics (MD) simulations, in combination with prevalent algorithmic enhancements and tremendous progress in computational resources, have contributed to new insights into mechanisms and processes present in physics, chemistry, biology, and engineering. However, their applicability in systems of practical relevance poses insurmountable computational difficulties.[1,2] For example, the simulation of $M = 10^5$ atoms over a time horizon of a mere $T \approx 10^{-4}$ s with a time step of $\Delta t = 10^{-15}$ s implies a computational time of one year.[3] A rugged free-energy surface and configurations separated by high free-energy barriers lead to unobserved conformations even in very long simulations.

Enhanced sampling methods[4] provide a framework for accelerating the exploration of the configurational space.[5–11] These methods rely on the existence of a lower-dimensional representation of the atomistic details. Lower-dimensional system variables (reaction coordinates), capture the characteristics of the system, allow us to understand relevant processes and conformational changes,[12] and can enable guided and enhanced MD simulations. *Reaction coordinates* provide quantitative understanding of macromolecular motion, whereas *order parameters* are of qualitative nature, as discussed in Ref. 13. In the following, we use the term *collective variables* (CVs), combining the quantitative and qualitative properties of reaction coordinates and order parameters, respectively. References 4 and 13 review the challenges in the exploration of the free-energy landscape and the identification of "good" collective variables.

Adding an appropriate biasing potential or force, based on CVs, results in an accelerated exploration of the configurational space.[13] Such algorithms might employ a constant

Published under license by AIP Publishing

bias term (e.g., umbrella sampling,[14] hyperdynamics,[15] accelerated MD,[16] etc.) or a time-dependent one (e.g., local elevation,[17] conformational flooding,[18] metadynamics,[3,19,20] adaptive biasing force,[21,22] etc.). The crucial ingredient for almost all of the aforementioned algorithms is the *right choice* of the *collective variables*. The potential benefit and justification of enhanced sampling algorithms strongly depend on the quality of the collective variables, as comprehensively elaborated in Refs. 23–25. Physical intuition, experience gathered from previous simulation as well as quantitative methods for dimensionality reduction [e.g., by utilizing principal component analysis[26] (PCA)], potentially supports the choice of reasonable collective variables. For complex materials-design problems and large-scale biochemical processes, complexity exceeds our intuition and the question of "good" collective variables remains unanswered. Enhanced sampling methods employing inappropriate collective variables can be outperformed by brute force MD simulations.[27] Thus, the identification of collective variables or reaction coordinates poses an important and difficult problem.

A systematic, robust, and general approach is needed for the discovery of lower-dimensional representations. Recent developments in dimensionality reduction methods provide a systematic strategy for discovering CVs.[13] For completeness, we give a brief overview of significant tools addressing CV discovery and dimensionality reduction in the context of molecular systems. An early study[28] found a steep decay in the eigenvalues of peptide trajectories indicating the existence of a low-dimensional representation that is capable of capturing essential physics. This study is based on PCA[26,29] which identifies a linear coordinate transformation for best capturing the variance. However, the linear coordinate transformations employed merely describe local fluctuations in the context of peptide trajectories. Multidimensional scaling (MDS)[30,31] identifies a lower-dimensional embedding such that pairwise distances [e.g., root-mean-square deviation (RMSD)] between atomistic configurations are best preserved. A sketch map[32] focuses on preserving "middle" ranged RMSD between trajectory pairs. Middle ranged RMSD pairs are the most relevant for observing pertinent behavior of the system.[32] An isometric feature map or ISOMAP[33] follows a similar idea of preserving geodesic distances. The aforementioned methods require dense sampling and encounter problems if the training data are non-uniformly distributed.[34–36] Furthermore, we note that these methods involve a mapping from the atomistic configurations to the CVs, whereas predictive tasks require a generative mapping from the CVs to the atomistic configuration.

Another group of non-linear dimensionality reduction methods follows the idea of approximating the eigenfunctions of the backward Fokker-Plank operator[37] by identifying eigenvalues and eigenvectors of transition kernels. The employed kernels resemble transition probabilities between configurations that we aim to preserve. For example, the diffusion map[38–40] retains the diffusion distance by the identified coordinates for dynamic[41] and stochastic systems.[42] A variation of diffusion maps exploits locally scaled diffusion maps (LSDMaps)[34] which calculate the transition probabilities between two configurations, utilizing the RMSD instead of an Euclidean distance. An additional local scale parameter, indicating the distance around a specific configuration presumably, could be well approximated by a low-dimensional hyperplane tangent. LSDMap is applied in Ref. 43 and enhances the exploration of the configurational space, as shown in Ref. 44. More recent approaches to collective variable discovery work under a common variational approach for conformation (VAC) dynamics[45] and employ a combination of basis functions for defining the eigenfunctions to the backward Fokker-Planck operator. One approach under VAC was developed in the context of metadynamics[19] combining ideas from time-lagged independent component analysis and well-tempered metadynamics.[46] Further developments have focused on alternate distance metrics, relying either on a kinetic distance which measures how slowly configurations interconvert[47] or on the commute distance[48] which provides an extension (arising by integration) of the former.

Several methods rely on the estimation of the eigenvectors of transition matrices which is an expensive task in terms of computational cost. The need for "large" training datasets (e.g., 10 000 datapoints are required for robustness of the results[13]) limits the applicability of these methods to less complex systems. We refer to Ref. 49 for a critical review and comparison of the various methodologies mentioned before.

In this work, we propose a data-driven reformulation of the identification of CVs under the paradigm of probabilistic (Bayesian) inference. The methodology implies a generative model, considering CVs as lower-dimensional (latent) generators[50] of the full atomistic trajectory. The focus furthermore is on problems where limited atomistic training data are available that prohibit the accurate calculation of statistics for quantities of interest. Our approach is to compute an approximation of the underlying probabilistic distribution of the data. We then use this approximate distribution in a generative manner to perform accurate Monte Carlo estimation of the quantities of interest. To account for the limited information provided by small size training datasets, epistemic uncertainties on quantities of interest are also computed within the Bayesian paradigm.

In the context of coarse-graining atomistic systems, latent variable models have been introduced in Refs. 51 and 52. We optimize a flexible non-linear mapping between CVs and atomistic coordinates which implicitly specifies the meaning of the CVs. The identified CVs provide physical/chemical insight into the characteristics of the considered system. In the proposed model, the posterior distribution of the CVs for a given atomistic datapoint is computed. This posterior provides a pre-image of the atomistic representation in the lower-dimensional latent space. We utilize recent developments in machine learning and deep Bayesian modeling (auto-encoding variational Bayes[53,54]). While typically deep learning

Published under license by AIP Publishing

models rely on huge amounts of data, we demonstrate the robustness of the proposed methodology considering only small and highly variable datasets (e.g., 50 datapoints compared to 10 000 as required in the aforementioned methods). The proposed strategy requires significantly less data as compared to MDS,[30,31] ISOMAP,[33] and diffusion map[38,39,41] and simultaneously enables the quantification of uncertainties arising from limited data. We also discuss how additional datapoints can be readily incorporated by efficiently updating the previously trained model.

Apart from the possibility of utilizing the discovered CVs for dimensionality reduction and enhanced sampling, we exploit them for *predictive* purposes, i.e., for generating new atomistic configurations and estimating macroscopic observables. One could draw similarities between the identification of CVs and the problem of identifying a good coarse-grained representation.[51,55–66] In addition, rather than solely obtaining point estimates of observables, the Bayesian framework adopted provides whole distributions which capture the epistemic uncertainty. This uncertainty propagates in the form of error bars around the predicted observables.

Several recent publications focus on similar problems.[67–69] The present work clearly differs from that of Ref. 69 where the data are provided in a pre-processed form of sine and cosine backbone dihedral angles, i.e., not as the full-atom configurations. The approach in Ref. 68 utilizes a pre-reduced representation of heavy atom positions as training data. While this is valid, it necessitates physical insight which might be not available for unexplored complex chemical compounds. In contrast, we rely on training data represented as Cartesian coordinates comprising all atoms of the considered system. We do not consider any physically or chemically motivated transformation nor do we perform any pre-processing of the dataset. Instead, we reveal, given the dimensionality of the CVs, important characteristics (i.e., dihedral angles and heavy atom positions) or less relevant fluctuations (noise) from the full atomistic picture. This work is also distinguished by following throughout a formalism based on Bayesian learning. Instead of adopting or designing optimization objectives or loss functions, we consistently work within a Bayesian framework where the objective naturally arises. Furthermore, this readily allows us to make use of sparsity-inducing priors which reveal parsimonious features. The work of Ref. 8 is based on auto-associative artificial neural networks (autoencoders) which allow the encoding and reconstruction of atomistic configurations given an input datum. The work of Ref. 8 relies on reduced Cartesian coordinates in the form of backbone atoms which induces information loss. In addition, the focus of Ref. 8 was on CV discovery and enhanced sampling, whereas we focus on CV discovery and obtaining a predictive model accounting for epistemic uncertainty.

The structure of the rest of the paper is as follows. Section II presents the basic model components, the use of Variational Autoencoders (VAEs[53]) in the CV discovery, and provides details on the learning algorithms employed. Numerical evidence of the capabilities of the proposed framework is provided in Sec. III. We identify CVs for alanine dipeptide and show the correlation between the discovered CVs and the dihedral angles. We furthermore assess the predictive quality of the discovered CVs and estimate observables augmented by credible intervals. We show the dependence of credible intervals on the amount of training data. We also present the results of a similar analysis for a more complex and higher-dimensional molecule, i.e., the ALA-15 peptide. Finally, Sec. IV summarizes the key findings of this paper and provides a brief discussion on potential extensions.

II. METHODS

After introducing the main notational convention in the context of equilibrium statistical mechanics, this section is devoted to the key concepts of generative latent variable models and variational inference[70] with emphasis on the identification of collective variables in atomistic systems.

A. Equilibrium statistical mechanics

We denote the coordinates of atoms of a molecular ensemble as $\boldsymbol{x} \in \mathcal{M}_{\mathrm{f}} \subset \mathbb{R}^{n_{\mathrm{f}}}$, with $n_{\mathrm{f}} = \dim(\boldsymbol{x})$. The coordinates $\boldsymbol{x}$ follow the Boltzmann-Gibbs density,

$$p_{\mathrm{target}}(\boldsymbol{x}) = \frac{1}{Z(\beta)} e^{-\beta U(\boldsymbol{x})}, \tag{1}$$

with the interatomic potential $U(\boldsymbol{x})$, $\beta = \frac{1}{k_b T}$, where k_b is the Boltzmann constant and T is the temperature. The normalization constant is given as $Z(\beta) = \int_{\mathcal{M}_{\mathrm{f}}} \exp\{-\beta U(\boldsymbol{x})\}\, d\boldsymbol{x}$. MD simulations[71] or Monte-Carlo-based methods[72] allow us to obtain samples from the distribution defined in Eq. (1). In the following, we assume that a dataset, $\mathbf{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$, has been collected, where $\boldsymbol{x}^{(i)} \sim p_{\mathrm{target}}(\boldsymbol{x})$. N denotes the amount of datapoints considered. The dataset $\mathbf{X}$ will be used for training the generative model to be introduced in the sequel. The underlying assumption in this work is that the size of the available training dataset $\mathbf{X}$ is small and not sufficient to compute directly the statistics of observables. Our focus is thus on deriving an approximation to the distribution in Eq. (1) from which, in a computationally inexpensive manner, one can sample sufficient realizations of $\boldsymbol{x}$ to allow probabilistic estimates of observables.

As elaborated in Ref. 13, the collection of a dataset $\mathbf{X}$ that sufficiently captures the configurational space constitutes a difficult problem of its own. Hampered by free-energy barriers, a MD simulation is not guaranteed to visit all conformations of an atomistic system within a finite simulation time. The discovery of CVs can facilitate the development of enhanced sampling methods[3,19,23] to address the efficient exploration of the configurational space.

This study considers systems in equilibrium for a given constant temperature T and consequently constant β.

Published under license by AIP Publishing

Optimally, the CVs discovered should be suitable for a range of temperatures.[25]

B. Probabilistic generative models

Deep learning[73] integrated with probabilistic modeling[74] has impacted many research areas.[75] In this paper, we emphasize a subset of these models referred to as *probabilistic generative models*.[50,76]

The objective is to identify CVs associated with relevant configurational changes of the system of interest. We consider CVs as hidden (low-dimensional) generators, giving rise to the observed atomistic configurations $\boldsymbol{x}$.[77] Extending the variable space of atomistic coordinates $\boldsymbol{x}$ by latent CVs denoted as $\boldsymbol{z} \in \mathcal{M}_{\mathrm{CV}} \subset \mathbb{R}^{n_{\mathrm{CV}}}$, with $n_{\mathrm{CV}} = \dim(\boldsymbol{z})$ and $\dim(\boldsymbol{z}) \ll \dim(\boldsymbol{x})$, allows us to define a joint distribution over the observed data $\boldsymbol{x}$ and latent CVs[50,78] $p(\boldsymbol{x}, \boldsymbol{z})$. The joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$ is written as

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})\, p(\boldsymbol{z}). \tag{2}$$

In Eq. (2), $p(\boldsymbol{z})$ prescribes the distribution of the CVs and $p(\boldsymbol{x}|\boldsymbol{z})$ represents the conditional probability of the full atomistic coordinates $\boldsymbol{x}$ given their latent representation $\boldsymbol{z}$. The probabilistic connection between the latent CVs $\boldsymbol{z}$ and the atomistic representation $\boldsymbol{x}$ implicitly defines the meaning of the CVs.

Marginalizing the joint representation of Eq. (2) with respect to the CVs leads to $p(\boldsymbol{x})$,

$$p(\boldsymbol{x}) = \int_{\mathcal{M}_{\mathrm{CV}}} p(\boldsymbol{x}, \boldsymbol{z})\, d\boldsymbol{z} = \int_{\mathcal{M}_{\mathrm{CV}}} p(\boldsymbol{x}|\boldsymbol{z})\, p(\boldsymbol{z})\, d\boldsymbol{z}. \tag{3}$$

Equation (3) provides a generative model for the atomistic configurations $\boldsymbol{x}$ and will be utilized as an efficient estimator for observables of the atomistic system. Standard autoencoders in the context of CV discovery[8] do not yield a probabilistic, predictive model which is the focus of this work. With appropriate selection of $p(\boldsymbol{z})$ and $p(\boldsymbol{x}|\boldsymbol{z})$, the resulting predictive distribution $p(\boldsymbol{x})$ should resemble the atomistic reference $p_{\mathrm{target}}(\boldsymbol{x})$ in Eq. (1). In order to quantify the closeness of the approximating distribution $p(\boldsymbol{x})$ and the actual distribution $p_{\mathrm{target}}(\boldsymbol{x})$, a distance measure is employed. The KL-divergence is one possibility out of the family of α-divergences[79–81,137] measuring the similarity between $p_{\mathrm{target}}(\boldsymbol{x})$ and $p(\boldsymbol{x})$. The non-negative valued KL-divergence is zero if and only if the two distributions coincide, which leads to the minimization objective with respect to $p(\boldsymbol{x})$ of the following
form:

$$\begin{aligned} D_{\mathrm{KL}}(p_{\mathrm{target}}(\boldsymbol{x})||p(\boldsymbol{x})) &= -\int_{\mathcal{M}_{\mathrm{f}}} p_{\mathrm{target}}(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p_{\mathrm{target}}(\boldsymbol{x})}\, d\boldsymbol{x} \\ &= -\int_{\mathcal{M}_{\mathrm{f}}} p_{\mathrm{target}}(\boldsymbol{x}) \log\, p(\boldsymbol{x})\, d\boldsymbol{x} \\ &\quad + \int_{\mathcal{M}_{\mathrm{f}}} p_{\mathrm{target}}(\boldsymbol{x}) \log\, p_{\mathrm{target}}(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned} \tag{4}$$

We introduce a parametrization $\boldsymbol{\theta}$ of the approximating distribution as $p(\boldsymbol{x}|\boldsymbol{\theta}) = \int_{\mathcal{M}_{\mathrm{CV}}} p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})\, d\boldsymbol{z}$. Instead of minimizing the KL-divergence with respect to $p(\boldsymbol{x})$, one can optimize the objective with respect to the parameters $\boldsymbol{\theta}$. We note that the minimization of Eq. (4) is equivalent to maximizing the expression $\int_{\mathcal{M}_{\mathrm{f}}} p_{\mathrm{target}}(\boldsymbol{x}) \log\, p(\boldsymbol{x})\, d\boldsymbol{x}$. If we consider a data-driven approach, where $p_{\mathrm{target}}(\boldsymbol{x})$ is approximated by a finite-sized dataset $\mathbf{X}$, we can write the problem as the maximization of the marginal log-likelihood $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(\mathrm{N})})$,

$$\begin{aligned} \log p(\mathbf{X}|\boldsymbol{\theta}) &= \sum_{\mathrm{i}=1}^{\mathrm{N}} \log p(\boldsymbol{x}^{(\mathrm{i})}|\boldsymbol{\theta}) \\ &= \sum_{\mathrm{i}=1}^{\mathrm{N}} \log\left(\int_{\mathcal{M}_{\mathrm{CV}}} p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(\mathrm{i})}|\boldsymbol{z}^{(\mathrm{i})})\, p_{\boldsymbol{\theta}}(\boldsymbol{z}^{(\mathrm{i})})\, d\boldsymbol{z}^{(\mathrm{i})} \right). \end{aligned} \tag{5}$$

Maximizing Eq. (5) with respect to the model parameters $\boldsymbol{\theta}$ results in the maximum likelihood estimate (MLE) $\boldsymbol{\theta}_{\mathrm{MLE}}$. By introducing a prior $p(\boldsymbol{\theta})$ on the parameters, one can augment this optimization problem to compute the Maximum a Posteriori (MAP) estimate[82–84] of $\boldsymbol{\theta}$ as follows:

$$\arg\max_{\boldsymbol{\theta}} \{\log p(\mathbf{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\}. \tag{6}$$

The full posterior of the model parameters $\boldsymbol{\theta}$ could also be obtained by applying Bayes' rule,

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{X})}. \tag{7}$$

Quantifying uncertainties in $\boldsymbol{\theta}$ enables us to capture the epistemic uncertainty introduced from the limited training data. The discovery of CVs through Bayesian inference is elaborated in the sequel.

C. Inference and learning

This section focuses on the details of inference and parameter learning for the generative model introduced in Eq. (3). Both tasks are facilitated by approximate variational inference[85] and stochastic backpropagation[54,86,87] which we discuss below.

Direct optimization of the marginal likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ requires the evaluation of $p(\boldsymbol{x}|\boldsymbol{\theta}) = \int_{\mathcal{M}_{\mathrm{CV}}} p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})\, d\boldsymbol{z}$ which constitutes an intractable integration over $\mathcal{M}_{\mathrm{CV}}$. The posterior over the latent CVs, $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) / p(\boldsymbol{x}|\boldsymbol{\theta})$, is also computationally intractable. Therefore, direct application of expectation-maximization[88,89] is not feasible. To this end, we reformulate the marginal log-likelihood for the dataset $\mathbf{X} = \{\boldsymbol{x}^{(\mathrm{i})}\}_{\mathrm{i}=1}^{\mathrm{N}}$ by introducing auxiliary densities $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(\mathrm{i})}|\boldsymbol{x}^{(\mathrm{i})})$ parametrized by $\boldsymbol{\phi}$. The meaning of $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(\mathrm{i})}|\boldsymbol{x}^{(\mathrm{i})})$ will be specified later in the text. The marginal log-likelihood is as follows:

Published under license by AIP Publishing

$$\begin{aligned}\log\, p(\mathbf{X}|\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \int_{\mathcal{M}_{\mathrm{CV}}} p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}) p_{\boldsymbol{\theta}}(\mathbf{z}^{(i)})\, d\mathbf{z}^{(i)} \\ &= \sum_{i=1}^{N} \log \int_{\mathcal{M}_{\mathrm{CV}}} q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}) \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}) p_{\boldsymbol{\theta}}(\mathbf{z}^{(i)})}{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}\, d\mathbf{z}^{(i)} \\ &\geq \sum_{i=1}^{N} \underbrace{\int_{\mathcal{M}_{\mathrm{CV}}} q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}) p_{\boldsymbol{\theta}}(\mathbf{z}^{(i)})}{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}\, d\mathbf{z}^{(i)}}_{\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{(i)})}, \end{aligned} \tag{8}$$

where in the last step we have made use of Jensen's inequality. Note that for each datapoint $\mathbf{x}^{(i)}$, one latent CV $\mathbf{z}^{(i)}$ is assigned. The lower-bound of the marginal log-likelihood is

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{X}) = \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{(i)}) \tag{9}$$

and implicitly depends on $\boldsymbol{\phi}$ through the parametrization of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. For each datapoint $\mathbf{x}^{(i)}$ and from the definition of $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{(i)})$, one can rewrite the marginal log-likelihood $\log p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$ as

$$\begin{aligned}\log p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) &= \mathrm{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})\right) \\ &\quad + \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{(i)}) \geq \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{(i)}). \end{aligned} \tag{10}$$

Since the KL-divergence is always non-negative, the inequalities in Eqs. (8) and (10) become equalities if and only if $q_{\boldsymbol{\theta}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}) = p_{\boldsymbol{\theta}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ as in this case $\mathrm{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})\right) = 0$. Thus $q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ can be thought of as an approximation of the true posterior over the latent variables $\mathbf{z}$. If the lower-bound gets tight, $q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ equals the exact posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})$.

Equation (8) can also be written as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{X}) &= \sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}[-\log q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\mathbf{z}^{(i)})] \\ &= -\sum_{i=1}^{N} \mathrm{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z}^{(i)})\right) \\ &\quad + \sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})]. \end{aligned} \tag{11}$$

It is clear from Eq. (11) that the lower-bound balances the optimization of the following two objectives:[53]

1. Minimizing $\sum_{i=1}^{N} \mathrm{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})\right)$ regularizes the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ such that, *on average* over all datapoints $\mathbf{x}^{(i)}$, it resembles $p_{\boldsymbol{\theta}}(\mathbf{z})$. We expect highly probable atomistic configurations $\mathbf{x}^{(i)}$ to be encoded to CVs $\mathbf{z}^{(i)}$ located in regions with high probability mass in $p_{\boldsymbol{\theta}}(\mathbf{z})$. The approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ over the latent CVs $\mathbf{z}$ accounts for this and supports findings presented in Ref. 68.
2. $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})]$ is the negative expected reconstruction error employing the encoded pre-image of the atomistic configuration $\mathbf{x}^{(i)}$ in the latent CV space. For example, assuming $p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})$ to be a Gaussian with mean $\boldsymbol{\mu}(\mathbf{z}^{(i)})$ and variance $\boldsymbol{\sigma}^2$, one can rewrite $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})]$ as

$$\begin{aligned}&\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})] \\ &\quad = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}\left[-\frac{1}{2}\frac{\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}(\mathbf{z}^{(i)})\right)^2}{\boldsymbol{\sigma}^2}\right] + \text{const.} \\ &\quad \propto -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})}\left[\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}(\mathbf{z}^{(i)})\right)^2\right] \\ &\quad = -\int_{\mathcal{M}_{\mathrm{CV}}} q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}) \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}(\mathbf{z}^{(i)})\right)^2 d\mathbf{z}^{(i)}. \end{aligned} \tag{12}$$

The second line of Eq. (12) is the negative expected error of reconstructing the atomistic configuration $\mathbf{x}^{(i)}$ through the decoder $p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})$. The expectation [see the last line in Eq. (12)] is evaluated with respect to $q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ and therefore with respect to all CVs $\mathbf{z}^{(i)}$ probabilistically assigned to $\mathbf{x}^{(i)}$.

The approximate posterior $q_{\boldsymbol{\phi}}$ of the latent variables $\mathbf{z}$ serves as a recognition model and is called the encoder.[53] Atomistic configurations $\mathbf{x}$ can be mapped via $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to their lower-dimensional representation $\mathbf{z}$ in the CV space. Hence, each $\mathbf{z}$ could be interpreted as a (latent) *encoding* of an $\mathbf{x}$. Its counterpart, the decoder $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, probabilistically maps CVs $\mathbf{z}$ to atomistic configurations $\mathbf{x}$. As will be demonstrated in the sequel, $\mathbf{z}$ sampled from $p_{\boldsymbol{\theta}}(\mathbf{z})$ will be used to reconstruct atomistic configurations via $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. The corresponding graphical model is presented in Fig. 1. Note that we do not require any physicochemical meaning assigned to the latent CVs that are identified implicitly during the training process.

The (approximate) inference task of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ has been reformulated as an optimization problem with respect to the parameters $\boldsymbol{\phi}$. These will be updated in combination with the parameters $\boldsymbol{\theta}$ as described in the following. At this point, we emphasize that the lower-bound $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{(i)})$ on the marginal log-likelihood (unobserved CVs are marginalized out) of Eq. (11) has been used as a negative "loss" function in non-Bayesian applications of autoencoders in the context of atomistic simulations as in Refs. 67 and 69.

In order to carry out the optimization $\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta};\mathbf{X})$ with respect to $\{\boldsymbol{\phi},\boldsymbol{\theta}\}$, first-order derivatives are needed of terms involving expectations with respect to $q_{\boldsymbol{\phi}}$ as it can be seen in Eq. (11). Consider in general a function $f(\mathbf{z})$ and the corresponding expectation $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})]$. Its gradient with respect to $\boldsymbol{\phi}$ can be expressed as

$$\nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[f(\mathbf{z}) \nabla_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right], \tag{13}$$

Published under license by AIP Publishing

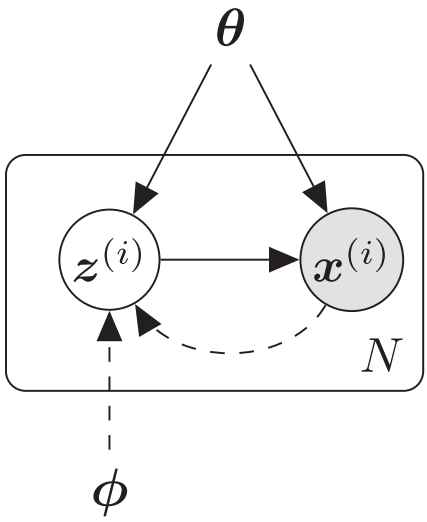


FIG. 1. Probabilistic graphical model representation following Ref. 53 with the latent CV representation $\boldsymbol{z}^{(i)}$ of each configuration $\boldsymbol{x}^{(i)}$ obtained by the approximate variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$ using the parametrization $\boldsymbol{\phi}$. The variational approximation is indicated with dashed edges and the generative model $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ with solid edges. $\boldsymbol{\theta}$ is the parametrization of the generative model.

and the expectation $\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\cdot]$ on the right hand-side can be approximated via a Monte-Carlo (MC) estimate using samples of $\boldsymbol{z}$ drawn from $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$. It is, however, known[86] that the variance of such estimators can be very high which adversely affects the optimization process. The high variance of the estimator in Eq. (13) can be addressed with the so-called reparametrization trick.[53,54] It is based on expressing $\boldsymbol{z}$ by auxiliary random variables $\boldsymbol{\epsilon}$ and a differentiable transformation $g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \boldsymbol{x})$ as

$$\boldsymbol{z} = g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \boldsymbol{x}) \text{ with } \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}). \tag{14}$$

Using the mapping, $g_{\boldsymbol{\phi}}: \boldsymbol{\epsilon} \to \boldsymbol{z}$, we can write the following for the densities $p(\boldsymbol{\epsilon})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = p\left(g_{\boldsymbol{\phi}}^{-1}(\boldsymbol{z}; \boldsymbol{x})\right)\left|\frac{\partial g_{\boldsymbol{\phi}}^{-1}(\boldsymbol{z}; \boldsymbol{x})}{\partial \boldsymbol{z}}\right|. \tag{15}$$

In Eq. (15), $g_{\boldsymbol{\phi}}^{-1}: \boldsymbol{z} \to \boldsymbol{\epsilon}$ denotes the inverse function of $g_{\boldsymbol{\phi}}$ which gives rise to $\boldsymbol{\epsilon} = g_{\boldsymbol{\phi}}^{-1}(\boldsymbol{z}; \boldsymbol{x})$. Several such transformations have been documented for typical densities (e.g., Gaussians).[90] The change of variables leads to the following expression for the gradient:

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[f(\boldsymbol{z})] &= \mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\phi}} f(g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \boldsymbol{x}))] \\ &= \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\frac{\partial f(g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \boldsymbol{x}))}{\partial \boldsymbol{z}}\frac{\partial g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \boldsymbol{x})}{\partial \boldsymbol{\phi}}\right],\end{aligned} \tag{16}$$

which can in turn be calculated by Monte Carlo using samples of $\boldsymbol{\epsilon}$ drawn from $p(\boldsymbol{\epsilon})$. Based on this, we define the following modified estimator for the lower-bound:[53]

$$\tilde{\mathcal{L}}(\boldsymbol{\phi}, \boldsymbol{\theta}; \boldsymbol{x}^{(i)}) = -\mathrm{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})||p_{\boldsymbol{\theta}}(\boldsymbol{z})\right) + \frac{1}{\mathrm{L}}\sum_{l=1}^{\mathrm{L}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,l)}),$$

with

$$\boldsymbol{z}^{(i,l)} = g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(l)}; \boldsymbol{x}^{(i)}) \text{ and } \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon}). \tag{17}$$

Note that for the particular forms of $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ selected in Sec. III A 2, $\mathrm{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})||p_{\boldsymbol{\theta}}(\boldsymbol{z})\right)$ becomes an analytically tractable expression. In order to increase the computational efficiency, we work with a sub-sampled minibatch $\mathbf{X}^{\mathrm{M}}$ comprising M datapoints from $\mathbf{X}$, with M < N. This leads to $\lfloor \mathrm{N}/\mathrm{M} \rfloor$ minibatches, each uniformly sampled from $\mathbf{X}$. The corresponding estimator of the lower-bound on the marginal log-likelihood is then given as

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \mathbf{X}) \simeq \tilde{\mathcal{L}}^{\mathrm{M}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^{\mathrm{M}}) = \frac{\mathrm{N}}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)}), \tag{18}$$

with $\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)})$ computed in Eq. (17). The factor N/M in Eq. (18) rescales $\sum_{i=1}^{\mathrm{M}}\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)})$ such that the lower-bound $\tilde{\mathcal{L}}^{\mathrm{M}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^{\mathrm{M}})$ computed by M < N datapoints approximates the actual lower-bound $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \mathbf{X})$ computed with N datapoints.[53] However, note that using a subset of the datapoints unavoidably increases the variance in the stochastic gradient estimator Eq. (17). Strategies compensating this increase are presented in Refs. 91 and 92, and a rigorous study of optimization techniques with enhancements in the context of coarse-graining is given in Ref. 62. The overall inference procedure is summarized in Algorithm 1.

We finally note that new data can be readily incorporated by augmenting accordingly the objective and initializing the algorithm with the optimal parameter values found up to that point. In fact this strategy was adopted in the results presented in Sec. III and led to significant efficiency gains. One can envision running an all-atom simulation which sequentially generates new training data that are automatically and quickly ingested by the proposed coarse-grained model which is in turn used to produce predictive estimates, as will be described in the sequel. In contrast, other dimensionality reduction methods based on the solution of an eigenvalue problem are required to solve a new system for the whole dataset when new data are presented.

D. Predicting atomistic configurations—Leveraging the exact likelihood

After training the model as described in Sec. II C, we are interested in obtaining the predictive distribution $p(\boldsymbol{x}|\boldsymbol{\theta}) = \int_{\mathcal{M}_{\mathrm{CV}}} p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{z})\, d\boldsymbol{z}$ [see Eq. (3)] which poses a demanding computational task. One approach for predicting configurations $\boldsymbol{x}$ distributed according to $p(\boldsymbol{x}|\boldsymbol{\theta})$ is ancestral sampling. First, one can generate a sample $\boldsymbol{z}^{l}$ from $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ and second sample $\boldsymbol{x}^{(k,l)} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}^{l})$. The variance of such estimators significantly increases with increasing $\dim(\boldsymbol{z})$. Ancestral

Published under license by AIP Publishing

Algorithm 1. Stochastic variational inference algorithm.

$\{\boldsymbol{\theta}, \boldsymbol{\phi}\} \leftarrow$ Initialize parameters.

repeat

$\mathbf{X}^{\mathrm{M}} \leftarrow$ Random minibatch of M datapoints drawn from dataset $\mathbf{X}$.

$\boldsymbol{\epsilon} \leftarrow$ Random sample(s) from noise distribution $p(\boldsymbol{\epsilon})$.

$\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{\phi},\boldsymbol{\theta}} \tilde{\mathcal{L}}^{\mathrm{M}}(\boldsymbol{\phi}, \boldsymbol{\theta}; \mathbf{X}^{\mathrm{M}})$ Calculate gradients with the estimator in Eq. (18).

$\{\boldsymbol{\phi}, \boldsymbol{\theta}\} \leftarrow$ Update parameters with gradient $\boldsymbol{g}$ (e.g., employing ADAM[93]).

until Convergence of $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$.

return $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$

sampling does not account for training the model by employing an *approximate* posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ instead of the actual posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ of the CVs $\mathbf{z}$. The *Metropolis-within-Gibbs* sampling scheme[94] accounts for grounding the optimization of the objective in Eq. (11) on a variational approximation. This approach builds upon findings in Ref. 54 and proposes that generated samples $\bar{\mathbf{x}}$ follow a Markov chain $(\mathbf{z}_t, \bar{\mathbf{x}}_t)$ for steps $t \geq 1$. Reference 94 proposes employing the following Metropolis[95,96] update criterion ρ_t reflecting a ratio of importance ratios:

$$\rho_t = \frac{\frac{p_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_{t-1}|\tilde{\mathbf{z}}_t)\, p_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}_t)}{p_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_{t-1}|\mathbf{z}_{t-1})\, p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1})}}{\frac{q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_t|\bar{\mathbf{x}}_{t-1})}{q_{\boldsymbol{\phi}}(\mathbf{z}_{t-1}|\bar{\mathbf{x}}_{t-1})}}. \tag{19}$$

Equation (19) provides the needed correction when using the approximate latent variable posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. When the CV's exact posterior is identified, i.e., when $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = 0$, all proposals $\mathbf{z}_t$ in Algorithm 2 are accepted with $\rho_t = 1$.

E. Prior specification

The recent work of Ref. 94 discusses the pitfalls of overly expressive, deep, latent variable models which can yield infinite likelihoods and ill-posed optimization problems.[97] We address these issues by regularizing the log-likelihood with functional priors.[98,99] The prior contribution is added as an additional component in the log-likelihood, as indicated in Eq. (6). In addition to enhanced stability during training,[94] sparsity inducing priors alleviate the overparameterized nature of complex neural networks.

We adopt the Automatic Relevance Determination (ARD[100]) model which consists of the following distributions:

$$p(\boldsymbol{\theta}|\boldsymbol{\tau}) \equiv \prod_k \mathcal{N}(\theta_k|0, \tau_k^{-1}), \quad \tau_k \sim \mathrm{Gamma}(\tau_k|a_0, b_0). \tag{20}$$

Equation (20) implies modeling each θ_k with an independent Gaussian distribution. The Gaussian distribution has zero-mean and an independent precision hyper-parameter τ_k, modeled with a (conjugate) Gamma density. The resulting prior $p(\theta_k)$ follows (by marginalizing the hyper-parameter τ_k) a heavy-tailed Student's t-distribution. This distribution favors *a priori* sparse solutions with θ_k close to zero. In order to compute derivatives of the log-prior, required for learning the parameters $\boldsymbol{\theta}$, we treat the τ_k's as latent variables in an inner-loop expectation-maximization scheme[101] which consists of the following steps:

Algorithm 2. Metropolis-within-Gibbs sampler.[94]

Input Trained model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z})$ and approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. Total steps T.

Initialize $(\mathbf{z}_0, \bar{\mathbf{x}}_0)$.

for t = 1 **to** T **do**

$\tilde{\mathbf{z}}_t \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\bar{\mathbf{x}}_{t-1})$ Draw proposal $\tilde{\mathbf{z}}_t$ from the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\bar{\mathbf{x}}_{t-1})$.

$\rho_t = \frac{p_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_{t-1}|\tilde{\mathbf{z}}_t)}{p_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_{t-1}|\mathbf{z}_{t-1})} \frac{p_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}_t)}{p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1})} \frac{q_{\boldsymbol{\phi}}(\mathbf{z}_{t-1}|\bar{\mathbf{x}}_{t-1})}{q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_t|\bar{\mathbf{x}}_{t-1})}$ Estimate the Metropolis acceptance ratio, correcting for the use of the approximate posterior distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$.

$\mathbf{z}_t = \begin{cases} \tilde{\mathbf{z}}_t & \text{with probability } \rho_t \\ \mathbf{z}_{t-1} & \text{with probability } 1 - \rho_t. \end{cases}$

$\bar{\mathbf{x}}_t \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_t)$

end for

return $\bar{\mathbf{x}}_{1:\mathrm{T}}$.

Published under license by AIP Publishing

Algorithm 3. Predictive collective variable discovery.

Input Dataset $\mathbf{X}$ with N samples $\boldsymbol{x}^{(i)} \sim p_{\text{target}}(\boldsymbol{x})$.

1: $\{\boldsymbol{\theta}, \boldsymbol{\phi}\} \leftarrow$ Specify the generative model $p_{\boldsymbol{\theta}}(\boldsymbol{z})$, $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ in Eq. (3) and the approximate posterior of the latent CVs $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ introduced in Eq. (8) with the corresponding parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively.

2: $\{\boldsymbol{\theta}_{\text{MAP}}, \boldsymbol{\phi}_{\text{MAP}}\} \leftarrow$ Maximize the lower-bound in Eq. (8) with stochastic variational inference, see Algorithm 1, and obtain the MAP estimates of the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

3: $p(\boldsymbol{\theta}|\mathbf{X}) \leftarrow$ Perform approximate Bayesian inference for obtaining the posterior distribution of the parameters of the generative model $\boldsymbol{\theta}$. See Sec. II F.

4: Predict the atomistic trajectory with Algorithm 2 for samples from the approximate posterior of the generative model parameters $\boldsymbol{\theta}^j \sim p(\boldsymbol{\theta}|\mathbf{X})$.

5: Estimate credible intervals of observables. This step is summarized in Algorithm 4.

Return Probabilistic estimates of observables accounting for epistemic uncertainty.

- E-step—evaluate:

$$\langle \tau_k \rangle_{p(\tau_k|\theta_k)} = \frac{a_0 + \frac{1}{2}}{b_0 + \frac{\theta_k^2}{2}}. \tag{21}$$

- M-step—evaluate:

$$\frac{\partial \log p(\boldsymbol{\theta})}{\partial \theta_k} = -\mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k]\theta_k. \tag{22}$$

The second derivative of the log-prior with respect to $\boldsymbol{\theta}$ is obtained as

$$\frac{\partial^2 \log p(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} = \begin{cases} -\mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k], & \text{if } k = l, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

The ARD choice of the hyper-parameters is $a_0 = b_0 = 1.0 \times 10^{-5}$. In similar settings, e.g., coarse-graining of atomistic systems, the ARD prior identified the most salient features,[51] whereas in this context it improves stability and turns off unnecessary parameters for describing the training data.

F. Approximate Bayesian inference for model parameters—Laplace's approximation

This subsection addresses the calculation of an approximate posterior of the model parameters $\boldsymbol{\theta}$. Thus far, we have considered point estimates of the model parameters $\boldsymbol{\theta}$ (either MLE or MAP). A fully Bayesian treatment, however, requires the evaluation of the normalization constant of the exact posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ of the model parameters $\boldsymbol{\theta}$, which is computationally impractical. We advocate an approximation to the posterior of $\boldsymbol{\theta}$ that is based on Laplace's method.[77] The latter has been rediscovered as an efficient approximation for weight uncertainties in the context of neural networks in Ref. 102.

In Laplace's approach, the exact posterior is approximated with a normal distribution with mean $\boldsymbol{\theta}_{\text{MAP}}$ and covariance, the inverse of the negative Hessian of the log-posterior at $\boldsymbol{\theta}_{\text{MAP}}$. Here, we assume a Gaussian with diagonal covariance matrix $\mathbf{S}_{\text{L}} = \text{diag}(\boldsymbol{\sigma}_{\text{L}}^2)$ as follows:

$$p(\boldsymbol{\theta}|\mathbf{X}) \approx \mathcal{N}\big(\boldsymbol{\mu}_{\text{L}}, \mathbf{S}_{\text{L}} = \text{diag}(\boldsymbol{\sigma}_{\text{L}}^2)\big), \tag{24}$$

with

$$\boldsymbol{\mu}_{\text{L}} = \boldsymbol{\theta}_{\text{MAP}}, \tag{25}$$

and the diagonal entries of $\mathbf{S}_{\text{L}}^{-1}$,

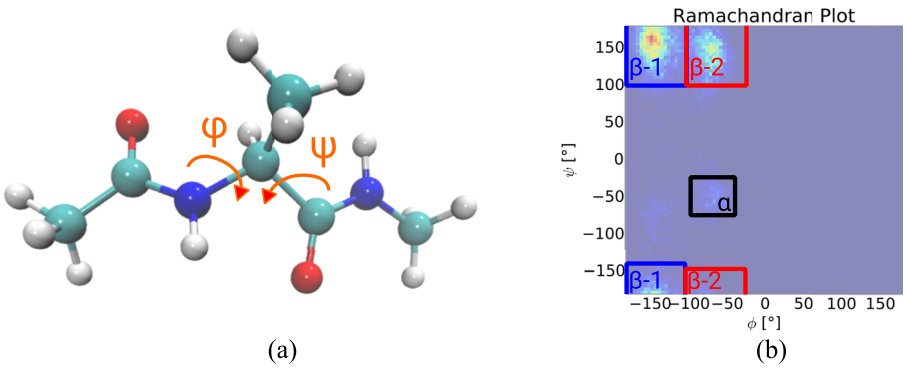
$$\sigma_{\text{L},k}^{-2} = -\frac{\partial^2 \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \mathbf{X})}{\partial \theta_k^2}\bigg|_{\boldsymbol{\theta}_{\text{MAP}}, \boldsymbol{\phi}_{\text{MAP}}} + \mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k], \tag{26}$$

where the term $\mathbb{E}_{p(\tau_k|\theta_k)}[\tau_k]$ arises from the prior via Eq. (23). The quantities in Eqs. (25) and (26) are obtained at the last iteration (upon convergence) of the auto-encoding variational Bayes algorithm. We summarize the procedure in Algorithm 3.

III. NUMERICAL ILLUSTRATIONS

Section III is devoted to the application of the proposed procedure for identifying collective variables of alanine dipeptide (ALA-2[103,104]) as well as of a longer peptide, i.e., ALA-15. We discuss the performance and robustness of the proposed methodology in the presence of a small amount of training data and emphasize the predictive capabilities of the model by the Ramachandran plot[105] and the radius of gyration. The predictions are augmented by error bars capturing epistemic uncertainty. The source code and data needed to reproduce all results presented in Secs. III A and III B are available at https://github.com/cics-nd/predictive-cvs.

Published under license by AIP Publishing



(a) (b)

FIG. 2. Definition of the dihedral angles and the labelling of characteristic modes as utilized in this paper. (a) ALA-2 peptide with indicated dihedral angles. (b) Characteristic conformations and their labelling as used in the sequel.

A. ALA-2

1. Simulation of ALA-2

Alanine dipeptide consists of 22 atoms leading to dim(**x**) = 66 in a Cartesian representation comprising the coordinates of *all* atoms which we will use later on as the model input. The actual degrees of freedom (DOF) are 60 after removing rigid-body motion. It is well-known that ALA-2 exhibits distinct conformations which are categorized depending on the dihedral angles (ϕ, ψ) [as indicated in Fig. 2(a)] of the atomistic configuration. We label the three characteristic modes as α, β-1, and β-2 in accordance with Ref. 106 [see Fig. 2(b)].

The procedure for generating the training data for ALA-2 is similar to that in Ref. 107. The atoms of the alanine dipeptide interact via the AMBER ff96[108–110] force field, and we employ an implicit water model based on the generalized Born/solvent accessible surface area model.[111,112] However, we note that an explicit water model would better represent an experimental environment. We employ an Andersen thermostat, and the simulations were carried out at constant temperature T = 330 K using Gromacs.[113–119] The time step is taken as Δt = 1 fs with an equilibration phase of 50 ns. The training dataset consisted of snapshots taken every 10 ps after the equilibration phase. Rigid-body motions have been removed from the dataset.

For demonstrating the encoding into the latent CV space of atomistic configurations not contained in the training dataset, we used a test dataset selected so that the dihedral angles (ϕ, ψ) had values belonging to all three modes, i.e., α, β-1, and β-2 [defined in Fig. 2(b)].

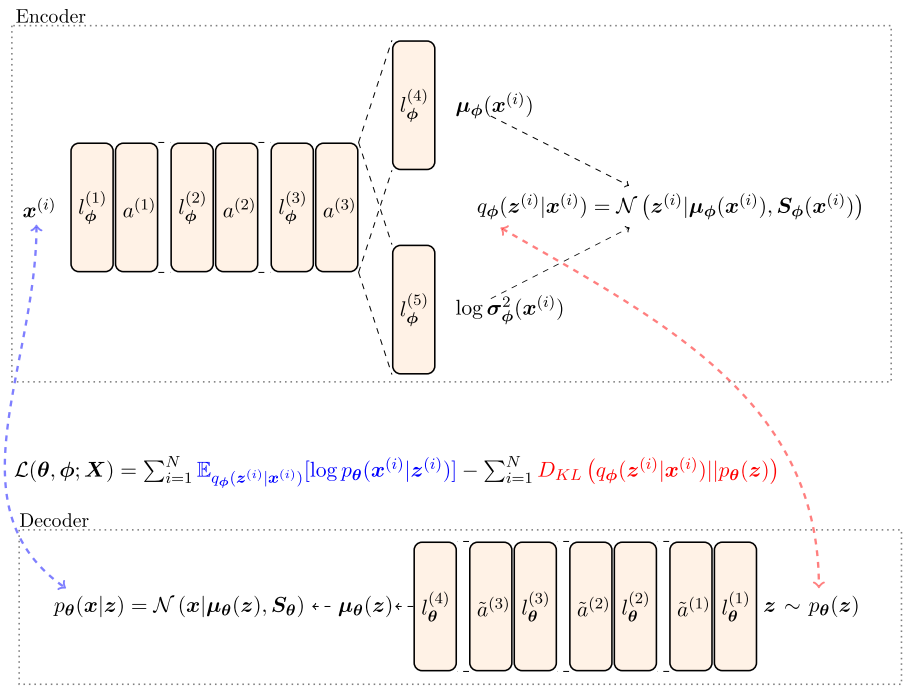


FIG. 3. Schematic of the AEVB depicting the employed network architecture. Fully connected linear layers are denoted with $l^{(i)}$ and non-linear activation functions with $a^{(i)}$. The indices $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ indicate encoding and decoding networks, respectively. The maximization of the lower-bound on the marginal log-likelihood $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{X})$ in Eq. (11) simultaneously optimizes the parametrization of the encoder and decoder. The first term in $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{X})$ accounts for the reconstruction of the training data $\boldsymbol{x}^{(i)}$ with $\boldsymbol{z}^{(i)}$ distributed according to $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$. The second term, in aggregation of all data $\boldsymbol{x}^{(i)}$, ensures that $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$ is close to $p(\boldsymbol{z})$.

Published under license by AIP Publishing

TABLE I. Network specification of the encoding neural network with $d_1 = 50$, $d_2 = 100$, and $d_3 = 100$.

Linear layer	Input dimension	Output dimension	Activation layer	Activation function
$l^{(1)}_{\boldsymbol{\phi}}$	$\dim(\boldsymbol{x})$	d_1	$a^{(1)}$	SeLu[a]
$l^{(2)}_{\boldsymbol{\phi}}$	d_1	d_2	$a^{(2)}$	SeLu
$l^{(3)}_{\boldsymbol{\phi}}$	d_2	d_3	$a^{(3)}$	Log sigmoid[b]
$l^{(4)}_{\boldsymbol{\phi}}$	d_3	$\dim(\boldsymbol{z})$	None	...
$l^{(5)}_{\boldsymbol{\phi}}$	d_3	$\dim(\boldsymbol{z})$	None	...

[a]SeLu: $a(x) = \begin{cases} \alpha(e^x - 1) & \text{if } x < 0, \\ x & \text{otherwise.} \end{cases}$ See Ref. 126 for further details.

[b]Log sigmoid: $a(x) = \log \frac{1}{1+e^{-x}}$.

2. Model specification

The model requires the specification of three components. Two components are needed to describe the generative model $p(\boldsymbol{x}|\boldsymbol{\theta})$: the probabilistic mapping $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ and the distribution of the CVs $p_{\boldsymbol{\theta}}(\boldsymbol{z})$. The third component is the approximate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ of the latent CVs, as shown in Eq. (8).

Following Ref. 53, the distribution of the CVs is taken to be a standard Gaussian,

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}) = p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I}). \tag{27}$$

The simplicity in the distribution in Eq. (27) is compensated by a flexible mapping from $\boldsymbol{z}$ to the atomistic coordinates $\boldsymbol{x}$. This probabilistic mapping (decoder) is given by a parametrized Gaussian as follows:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{z}), \boldsymbol{S}_{\boldsymbol{\theta}}), \tag{28}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{z}) = f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\boldsymbol{z}) \tag{29}$$

is a non-linear mapping $\boldsymbol{z} \mapsto f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\boldsymbol{z})$ ($f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}} : \mathbb{R}^{n_{\mathrm{CV}}} \mapsto \mathbb{R}^{n_f}$) parametrized by an expressive multilayer perceptron.[120–122]

We consider a diagonal covariance matrix, i.e., $\boldsymbol{S}_{\boldsymbol{\theta}} = \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}})$,[94] where its entries $\sigma^2_{\boldsymbol{\theta},j}$ are treated as model parameters and do not depend on the latent CVs $\boldsymbol{z}$. In order to ensure the non-negativity of $\sigma^2_{\boldsymbol{\theta},j} > 0$ while performing unconstrained optimization, we operate instead on $\log \sigma^2_{\boldsymbol{\theta},j}$.

The approximate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$ of the latent variables (encoder, approximating $p_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$) introduced in Eq. (8) is modeled by a Gaussian with flexible mean and variance represented by a neural network. For each pair of $\boldsymbol{x}^{(i)}$, $\boldsymbol{z}^{(i)}$ [for notational simplicity, we drop the index (i)],

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}), \boldsymbol{S}_{\boldsymbol{\phi}}(\boldsymbol{x})), \tag{30}$$

where the covariance matrix is assumed to be diagonal, i.e., $\boldsymbol{S}_{\boldsymbol{\phi}}(\boldsymbol{x}) = \mathrm{diag}\big(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\boldsymbol{x})\big)$. Furthermore $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x})$ and $\log \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\boldsymbol{x})$ are taken as the outputs of the encoding neural networks $f^{\boldsymbol{\mu}}_{\boldsymbol{\phi}}(\boldsymbol{x})$ and $f^{\boldsymbol{\sigma}}_{\boldsymbol{\phi}}(\boldsymbol{x})$, respectively,

$$\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}) = f^{\boldsymbol{\mu}}_{\boldsymbol{\phi}}(\boldsymbol{x}) \quad \text{and} \quad \log \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\boldsymbol{x}) = f^{\boldsymbol{\sigma}}_{\boldsymbol{\phi}}(\boldsymbol{x}). \tag{31}$$

We provide further details later in this section along with the structure of the employed networks. In our model, we assume a diagonal Gaussian approximation for $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$.

We are aware that the actual, but intractable, posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ could differ from a diagonal Gaussian and even from a multivariate normal distribution. However, the low variance $\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}$ observed in test cases justifies the assumption of a diagonal Gaussian in this context. An enriched model for the approximate posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ over the CVs could rely on, e.g., normalizing flows.[123] Recent developments on autoregressive flows[124] overcome the practical restriction of normalizing flows to low-dimensional latent spaces. This discussion equally holds for the assumption of a Gaussian with a diagonal covariance matrix for the generative distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. In the latter case, the diagonal entries of the covariance matrix

TABLE II. Network specification of the decoding neural network with $d_{\{1,2,3\}}$ as defined in Table I.

Linear layer	Input dimension	Output dimension	Activation layer	Activation function
$l^{(1)}_{\boldsymbol{\theta}}$	$\dim(\boldsymbol{z})$	d_3	$\tilde{a}^{(1)}$	Tanh
$l^{(2)}_{\boldsymbol{\theta}}$	d_3	d_2	$\tilde{a}^{(2)}$	Tanh
$l^{(3)}_{\boldsymbol{\theta}}$	d_2	d_1	$\tilde{a}^{(3)}$	Tanh
$l^{(4)}_{\boldsymbol{\theta}}$	d_1	$\dim(\boldsymbol{x})$	None	...

Published under license by AIP Publishing

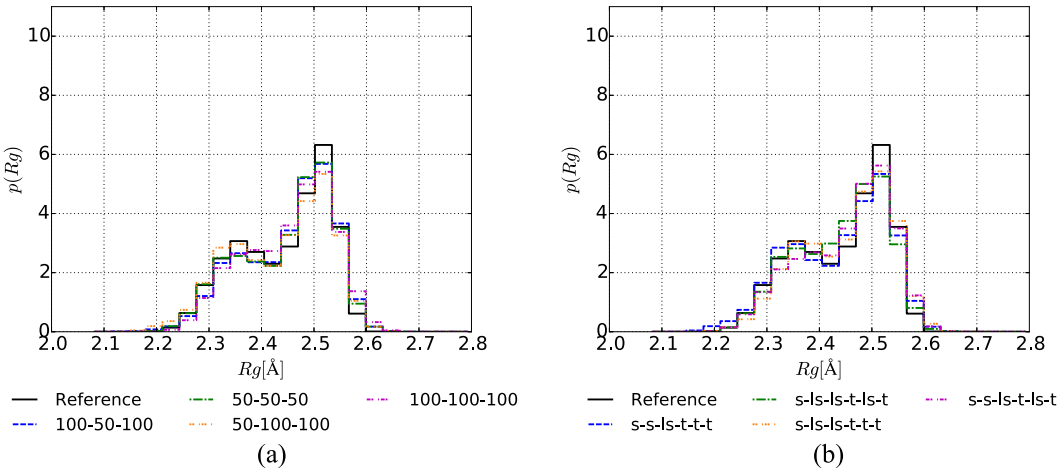


FIG. 4. Prediction of the radius of gyration with differing networks, in terms of (a) the dimensionality of the layers and (b) regarding the type of activation functions used. Changes in the network specification lead to similar predictions. This model has been trained with a dataset of size $N = 500$. (a) Varying dimensionality of the layers $l^{(i)}_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}}$. The labels represent the dimensionality of the layers in the format $d_1 - d_2 - d_3$ as specified in Tables I and II. We use the activation functions as denoted in Tables I and II. (b) Testing different activation functions for $a^{(i)}$. The labels specify the utilized activation functions in the following manner: $a^{(1)} - a^{(2)} - a^{(3)} - \tilde{a}^{(1)} - \tilde{a}^{(2)} - \tilde{a}^{(3)}$. We use the abbreviations: t: Tanh, s: SeLu, and ls: Log sigmoid.

$\mathbf{S}_{\boldsymbol{\theta}} = \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}})$ were modeled as parameters independent of $\mathbf{z}$. Using either $\mathbf{S}_{\boldsymbol{\theta}} = \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}})$ or introducing a dependency on the latent CVs, $\mathbf{S}_{\boldsymbol{\theta}}(\mathbf{z}) = \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\mathbf{z}))$ does not influence the predictive quality in terms of observables and predicted atomistic configurations. This statement is particularly valid when an expressive model for the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})$ in $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ (as in this work) is considered. It would be of interest employing more complex noise models for $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ which could be achieved by a Cholesky parametrization.[125] This might reveal structure correlations while reducing the need for higher complexity in $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})$.

As noted in Eq. (17), we employ the reparametrization trick by writing each random variable $\mathbf{z}^{(i,l)} \sim q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$ as

$$\mathbf{z}^{(i,l)} = g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(l)}; \mathbf{x}^{(i)}) = \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}^{(i)}) + \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}^{(i)}) \odot \boldsymbol{\epsilon}^{(l)} \tag{32}$$

and

$$\boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{33}$$

where $\odot$ denotes the element-wise vector product.

We utilize the following structure for the decoding neural network $f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{z})$:

$$f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{z}) = \left(l^{(4)}_{\boldsymbol{\theta}} \circ \tilde{a}^{(3)} \circ l^{(3)}_{\boldsymbol{\theta}} \circ \tilde{a}^{(2)} \circ l^{(2)}_{\boldsymbol{\theta}} \circ \tilde{a}^{(1)} \circ l^{(1)}_{\boldsymbol{\theta}}\right)(\mathbf{z}). \tag{34}$$

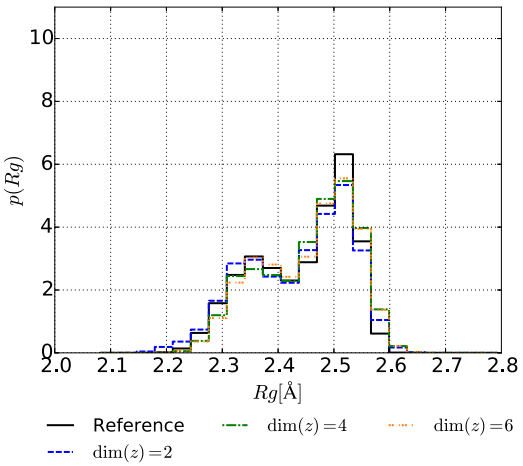


FIG. 5. Predicted radius of gyration for models utilizing different dim($\mathbf{z}$). The predictions are based on a model as specified in Tables I and II with $N = 500$.

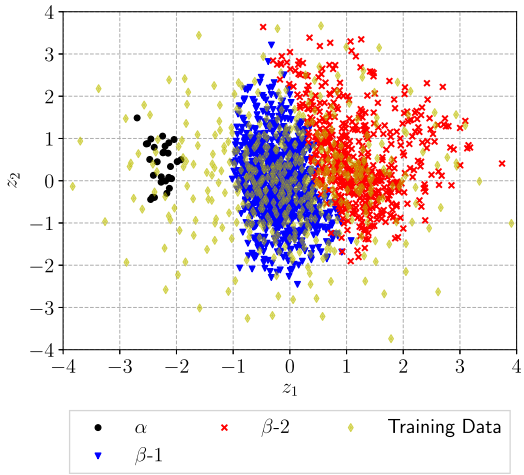


FIG. 6. Representation of the $\mathbf{z}$-coordinates of the training data $\boldsymbol{X}$ with $N = 500$ in the CV space (yellow diamonds). Using the trained model and the mean of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{z})$, we computed the $\mathbf{z}$-coordinates of 1527 test samples corresponding to different conformations of the alanine dipeptide to α (black), β-1 (blue), and β-2 (red). Without any prior physical information, the encoder yields three distinct clusters in the CV space.

Published under license by AIP Publishing

The encoding networks for obtaining $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$ and $\boldsymbol{\sigma}^2{}_{\boldsymbol{\phi}}(\mathbf{x})$ of the approximate posterior $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ over the latent CVs share the structure,

$$f_{\boldsymbol{\phi}}(\mathbf{x}) = \left(a^{(3)} \circ l^{(3)}_{\boldsymbol{\phi}} \circ a^{(2)} \circ l^{(2)}_{\boldsymbol{\phi}} \circ a^{(1)} \circ l^{(1)}_{\boldsymbol{\phi}}\right)(\mathbf{x}), \tag{35}$$

which gives rise to $f^{\boldsymbol{\mu}}_{\boldsymbol{\phi}}(\mathbf{x})$ and $f^{\boldsymbol{\sigma}}_{\boldsymbol{\phi}}(\mathbf{x})$ with

$$f^{\boldsymbol{\mu}}_{\boldsymbol{\phi}}(\mathbf{x}) = l^{(4)}_{\boldsymbol{\phi}}(f_{\boldsymbol{\phi}}(\mathbf{x})) \quad \text{and} \quad f^{\boldsymbol{\sigma}}_{\boldsymbol{\phi}}(\mathbf{x}) = l^{(5)}_{\boldsymbol{\phi}}(f_{\boldsymbol{\phi}}(\mathbf{x})). \tag{36}$$

In Eqs. (34)–(36), we consider linear layers $l^{(i)}$ of a variable $\mathbf{y}$ with $l^{(i)}(\mathbf{y}) = \mathbf{W}^{(i)}\mathbf{y} + \mathbf{b}^{(i)}$ and non-linear activation functions denoted with $a(\cdot)$. The indices $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ of the linear layers $l^{(i)}$ reflect correspondence to either the encoding or

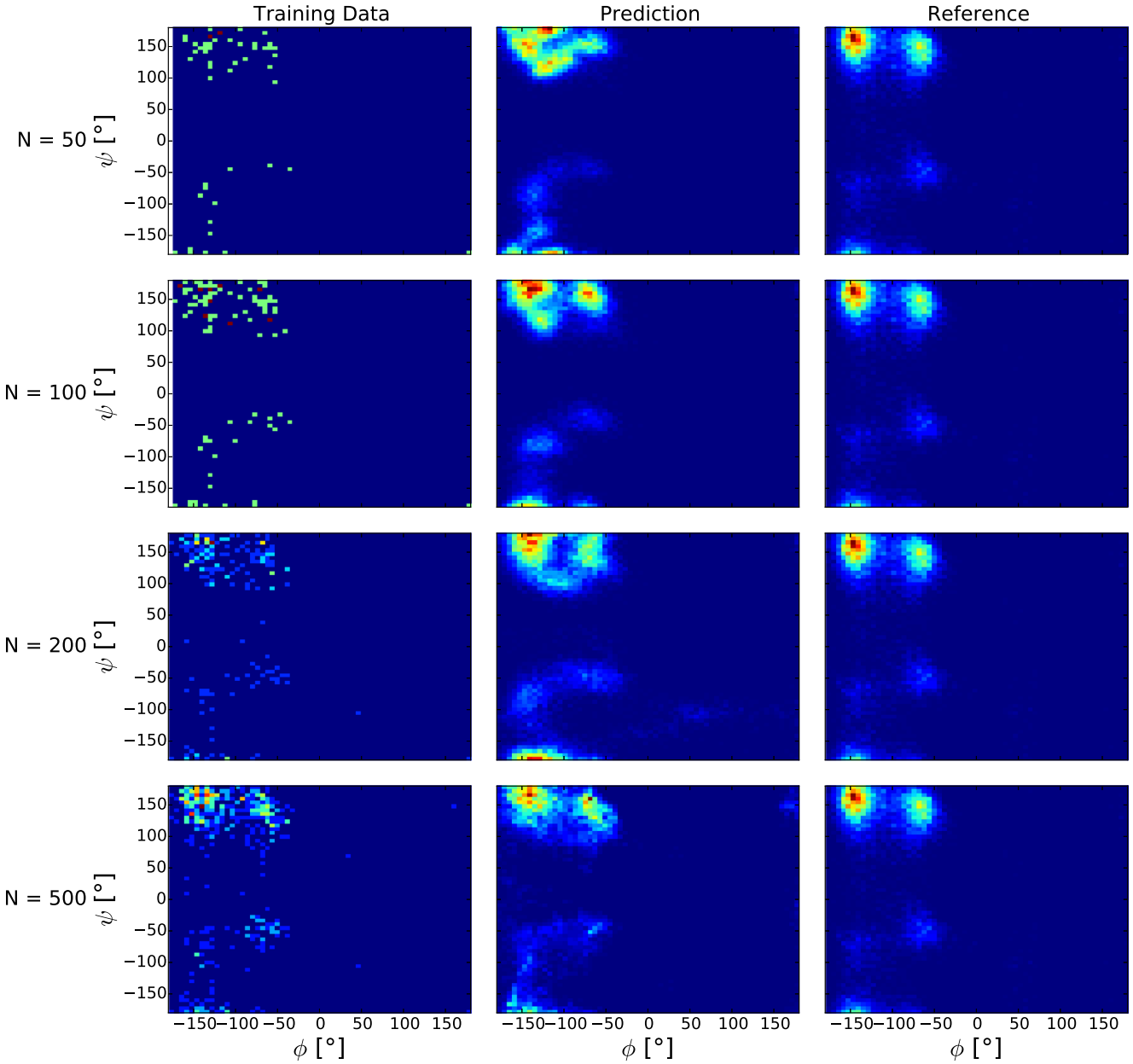


FIG. 7. Ramachandran plots estimated with the training data $\boldsymbol{X}$ (left column), using predictions of the trained model (middle column), and the reference (right column, estimated with N = 10 000). Each row refers to different sizes N of training datasets (the figure on the right column is repeated to allow easy comparison with the results on the first two columns). The represented predictions are obtained by applying Algorithm 2 with T = 10 000 samples. The generative nature of the model allows more accurate estimates than when using the training data alone. In addition, the Bayesian approach allows for predictions with their associated uncertainties as discussed subsequently.

Published under license by AIP Publishing

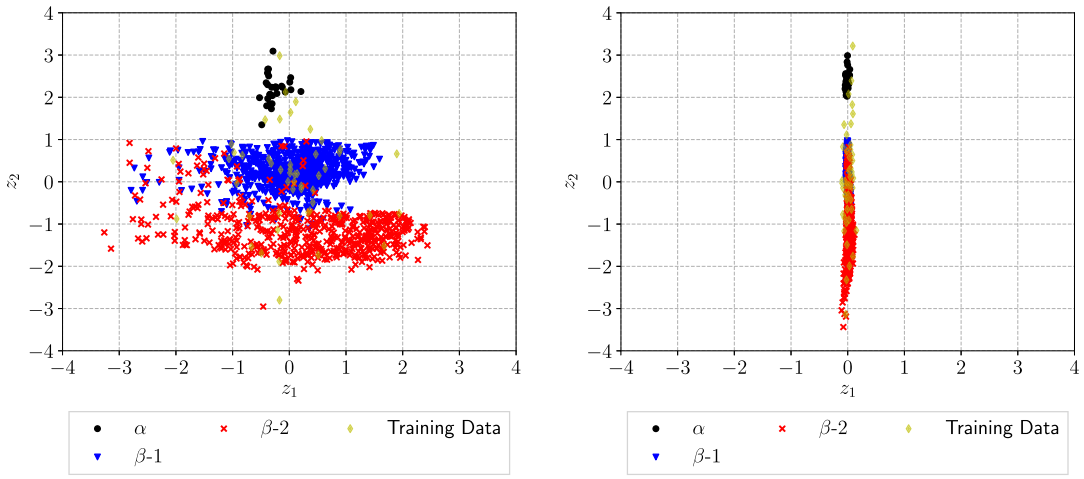


FIG. 8. Representation of the $\boldsymbol{z}$-coordinates of the training data $\boldsymbol{X}$ with $N = 50$ in the CV space (yellow diamonds). Using the trained model and the mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{z})$, we computed the $\boldsymbol{z}$-coordinates of 1527 test samples corresponding to different conformations of the alanine dipeptide to α (black), β-1 (blue), and β-2 (red). In the case of limited training data, the ARD prior facilitates the identification of physically meaningful CVs (a) compared to the representation (b) obtained without the ARD prior. Note that the changed positioning of the conformations in the CV space compared to Fig. 6 is due to symmetries in $p_{\boldsymbol{\theta}}(\boldsymbol{z})$.

decoding network, respectively. $\boldsymbol{\phi}$ comprises all parameters of the encoding networks $f^{\boldsymbol{\mu}}_{\boldsymbol{\phi}}(\boldsymbol{x})$ and $f^{\boldsymbol{\sigma}}_{\boldsymbol{\phi}}(\boldsymbol{x})$, $\boldsymbol{\theta}$ all parameters of the decoding network $f_{\boldsymbol{\theta}}(\boldsymbol{z})$ including the parameters $\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}$ discussed in Eq. (28). We differentiate the encoding and decoding activation functions by denoting them as $a^{(i)}$ and $\tilde{a}^{(i)}$, respectively. All layers considered were fully connected. The general architecture of the neural networks employed and how these affect the objective $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X})$ are depicted in Fig. 3.

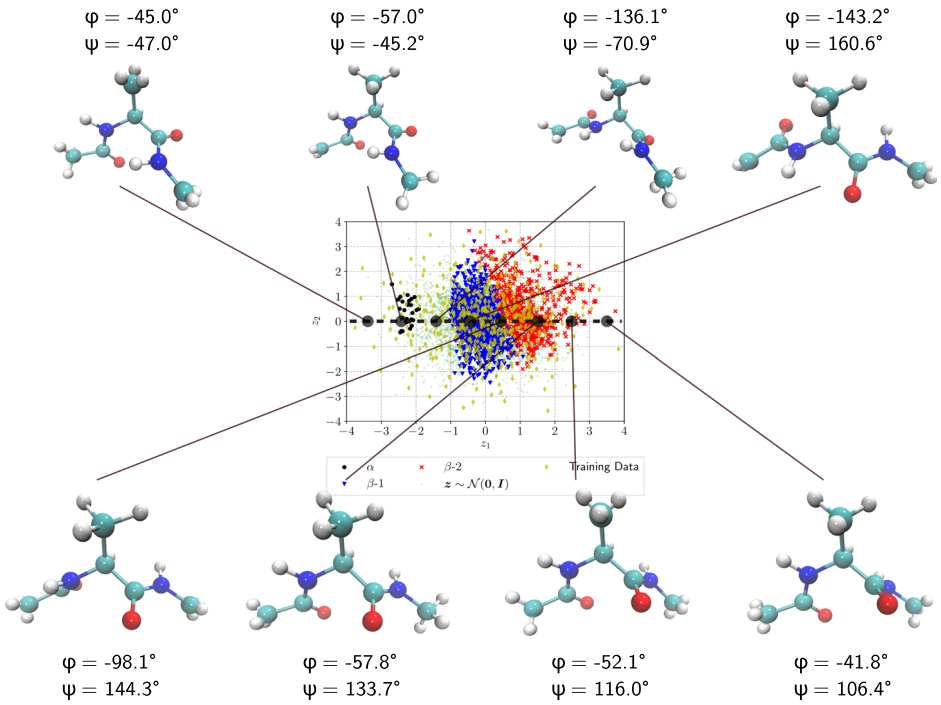


FIG. 9. *Predicted* configurations $\boldsymbol{x}$ (including dihedral angle values) for $\{\boldsymbol{z}|z_1 = \{-3.5, -2.5, \ldots, 3.5\}, z_2 = 0\}$ with $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{z})$ of $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. As one moves along the z_1 axis, we obtain for the given CVs atomistic configurations $\boldsymbol{x}$ reflecting the conformations α, β-1, and β-2. All rendered atomistic representations in this work are created by VMD.[128]

Published under license by AIP Publishing

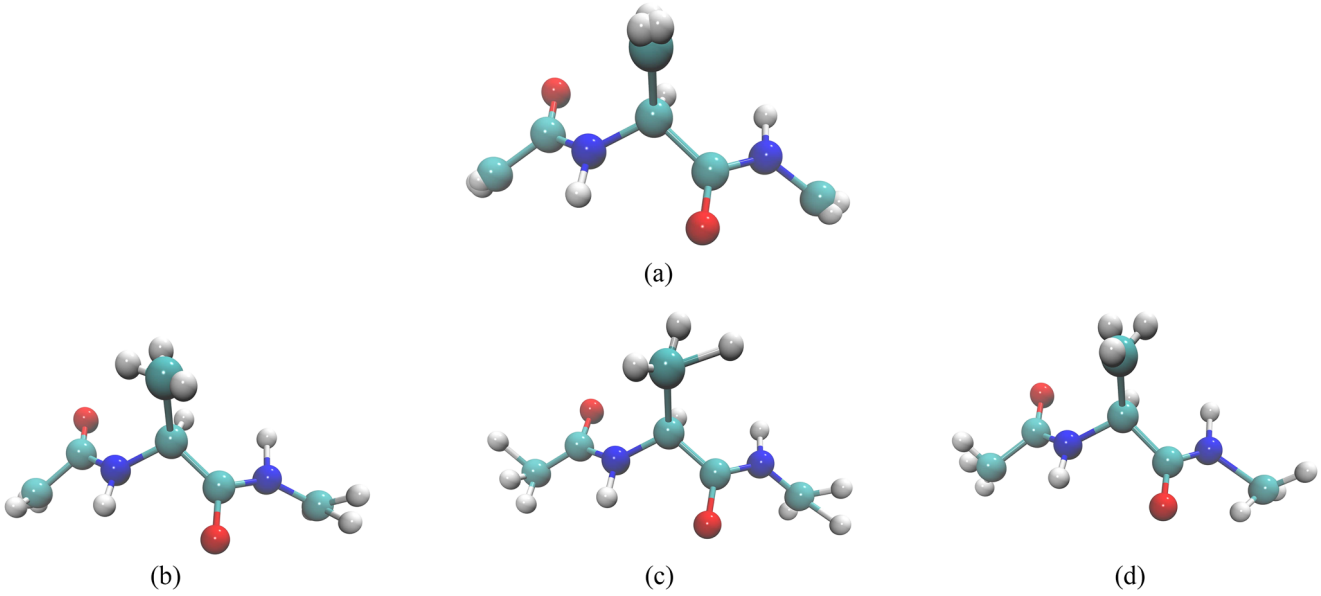


FIG. 10. Visualization of the mean prediction (a) for a sample $\mathbf{z}^0 \sim p(\mathbf{z})$, obtained from the decoding network $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}^0) = f_{\boldsymbol{\theta}}(\mathbf{z}^0)$, and realizations [(b)–(d)] $\mathbf{x}^{i,0} \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}^0)$. Less relevant positions of the outer hydrogen atoms are captured by the noise $\boldsymbol{\sigma}_{\boldsymbol{\theta}}$ of the model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}^0) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{z}^0_{\boldsymbol{\theta}}), \mathbf{S}_{\boldsymbol{\theta}} = \text{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}))$. (a) Mean prediction $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}^0)$ for a sample $\mathbf{z}^0 \sim p(\mathbf{z})$. (b) Realization $\mathbf{x}^{0,0} \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}^0), \mathbf{S}_{\boldsymbol{\theta}} = \text{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}))$. (c) Realization $\mathbf{x}^{1,0} \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}^0), \mathbf{S}_{\boldsymbol{\theta}} = \text{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}))$. (d) Realization $\mathbf{x}^{2,0} \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}^0), \mathbf{S}_{\boldsymbol{\theta}} = \text{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}))$.

The optimization of the objective is carried out by a stochastic gradient ascent algorithm. In our case, we employ ADAM[93] with the parameters chosen as $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon_{\text{ADAM}} = 1.00 \times 10^{-8}$. Gradients of the lower-bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X})$ with respect to the model parametrization $\{\boldsymbol{\phi}, \boldsymbol{\theta}\}$ are estimated by the backpropagation procedure.[120] The required gradients for optimizing the parameters $\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}$ can be computed analytically. For an entry $\sigma^2_{\text{j},\boldsymbol{\theta}}$, we can write the following:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(\text{i})})}{\partial \log \sigma^2_{\text{j},\boldsymbol{\theta}}} &= \frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(\text{i})}|\mathbf{z})}{\partial \log \sigma^2_{\text{j},\boldsymbol{\theta}}} \\
&= \frac{\partial}{\partial \log \sigma^2_{\text{j},\boldsymbol{\theta}}} \left(-\frac{1}{2} \sum_{\text{j}=1}^{\dim(\mathbf{x}^{(\text{i})})} \frac{\left(x^{(\text{i})}_{\text{j}} - \mu_{\text{j},\boldsymbol{\theta}}(\mathbf{z})\right)^2}{\sigma^2_{\text{j},\boldsymbol{\theta}}} \right) \\
&= \frac{\partial}{\partial \log \sigma^2_{\text{j},\boldsymbol{\theta}}} \left(-\frac{1}{2} \sum_{\text{j}=1}^{\dim(\mathbf{x}^{(\text{i})})} \frac{\left(x^{(\text{i})}_{\text{j}} - \mu_{\text{j},\boldsymbol{\theta}}(\mathbf{z})\right)^2}{\exp\left(\log\left(\sigma^2_{\text{j},\boldsymbol{\theta}}\right)\right)} \right) \\
&= \frac{1}{2} \frac{\left(x^{(\text{i})}_{\text{j}} - \mu_{\text{j},\boldsymbol{\theta}}(\mathbf{z})\right)^2}{\sigma^2_{\text{j},\boldsymbol{\theta}}}. \qquad (37)
\end{aligned}$$

Studying different combinations of activation functions and layers for the encoding network $f^{\boldsymbol{\mu},\boldsymbol{\sigma}}_{\boldsymbol{\phi}}(\mathbf{x})$ and the decoding network $f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{z})$ led to the network architecture depicted in Tables I and II, respectively. This network provided a repeatedly stable optimization during training. Variations of the given network architecture resulted in similar predictive capabilities, as shown in Fig. 4. Stability is not limited to symmetric encoding and decoding activation functions. An automated approach for selecting or learning the best architecture is an active research area.[127] Increasing the dimension of $\mathbf{z}$ did not improve the predictive capabilities, as shown in Fig. 5. This implies that CVs with $\dim(\mathbf{z}) = 2$ suffice to capture the physics encapsulated in the ALA-2 dataset with $\dim(\mathbf{x}) = 66$ or 60 DOF.

3. Results

In the following illustrations, we trained the model by varying the number of snapshots N. We utilized a sub-sampled

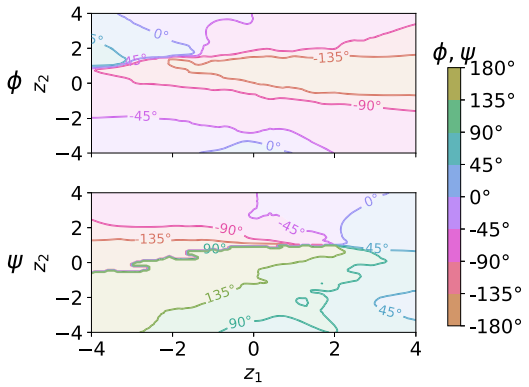


FIG. 11. Predicted dihedral angles (ϕ, ψ) given the latent variables $\mathbf{z} \in [-4, 4]^2$.

Published under license by AIP Publishing

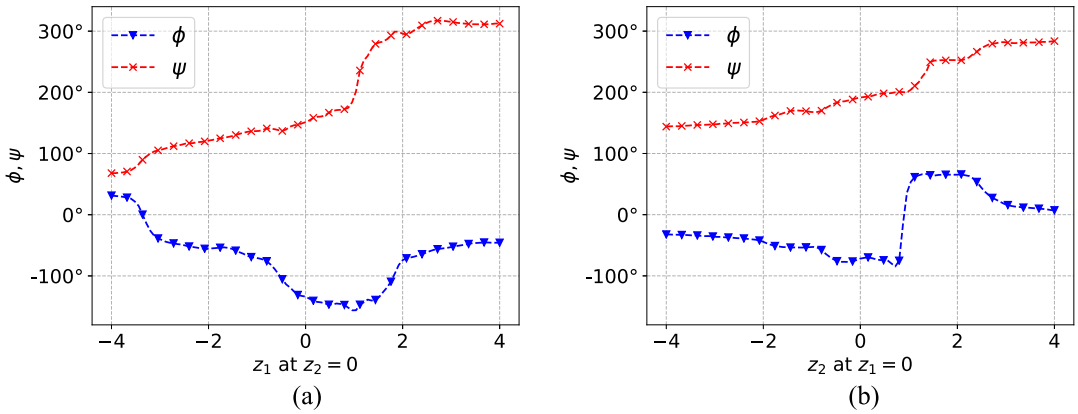


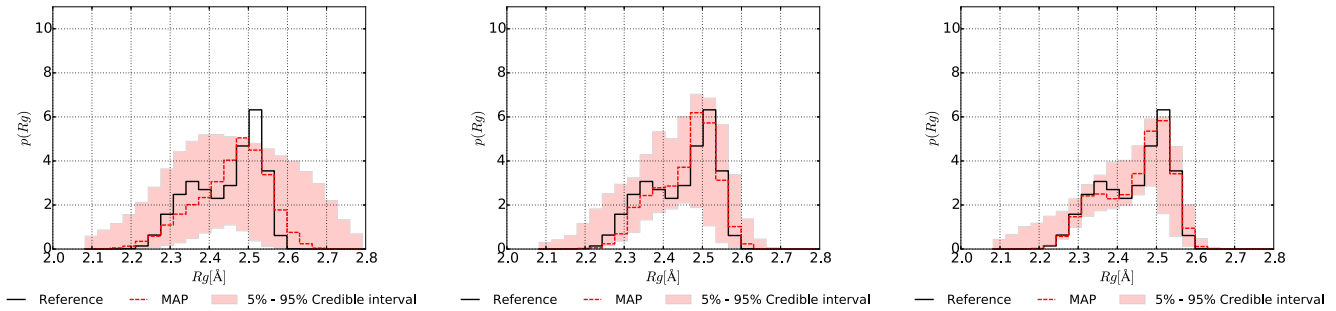
FIG. 12. Predicted dihedral angles (ϕ, ψ) given the latent variables (a) $\{z_1, z_2 | z_1 \in [-4, 4], z_2 = 0\}$ and (b) $\{z_1, z_2 | z_1 = 0, z_2 \in [-4, 4]\}$.

batch of size M = 64 from the dataset of size N. In cases where N < 64, we set M = N. The hyper parameters of the ARD prior in Eq. (20) are set to $a_0 = b_0 = 1.0 \times 10^{-5}$. Other values for a_0, b_0 in the range of $[1.0 \times 10^{-8}, 1.0 \times 10^{-4}]$ were also employed without a significant effect on the obtained sparsity patterns or the predictive accuracy of the model.

Figure 6 depicts the $\boldsymbol{z}$-coordinates of N = 500 training data as well as those of 1527 test data which have been classified into the three modes based on the values of the dihedral angles [see Fig. 2(b)]. In order to obtain the $\boldsymbol{z}$-coordinates of the test data, we made use of the mean $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(i)})$ of the inferred approximate posterior $q_{\boldsymbol{\phi}}$ as obtained after training. The resulting picture essentially provides the pre-images of the atomistic configurations in the CV space. Interestingly, similar atomistic configurations, i.e., belonging to one of the three modes, α, β-1, β-2, are recognized by $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and mapped to clusters in the identified CV space. β-1 configurations are encoded by $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ to regions with high probability mass in $p_{\boldsymbol{\theta}}(\boldsymbol{z})$, i.e., CVs $\boldsymbol{z}$ close to the center of $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ = N($\boldsymbol{0}$, $\boldsymbol{I}$) are assigned. This is in accordance with the reference Boltzmann distribution $p(\boldsymbol{x})$, where β-1 is the most probable conformation.

Various dimensionality reduction methods are designed in order to keep similar $\boldsymbol{x}$ close in their embedding on the lower-dimensional CV manifold, e.g., multidimensional scaling[30] or ISOMAP.[33] In the presented scheme, the generative model learns that mapping similar $\boldsymbol{x}$ to similar $\boldsymbol{z}$ leads to an expressive (in terms of the marginal likelihood) lower-dimensional representation. This similarity is revealed by inferring the approximate latent variable posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$. Therefore, the desired similarity mentioned in Ref. 13 between configurations in the atomistic representation $\boldsymbol{x}$ and via $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in the assigned CVs $\boldsymbol{z}$ is achieved.

In contrast to several other dimensionality reduction techniques (e.g., isomap[33] and diffusion maps[38–41]), which as mentioned in the Introduction require large amounts of training data, e.g., N > 10 000,[13,49] the proposed method can perform well in the small data regime, e.g., for N = 50 as shown



(a) $N = 50$. (b) $N = 200$. (c) $N = 500$.

FIG. 13. Predicted radius of gyration with dim($\boldsymbol{z}$) = 2 for various sizes N of the training dataset. The MAP estimate indicated in red is compared to the reference (black) solution. The latter is estimated by N = 10 000. The shaded area represents the 5%–95% credible interval, reflecting the induced epistemic uncertainty from the limited amount of training data.

Published under license by AIP Publishing

Algorithm 4. Estimating credible intervals.

Input J the number of samples to be drawn, optimal values of $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{MAP}}$ and $\boldsymbol{\phi} = \boldsymbol{\phi}_{\mathrm{MAP}}$.

Compute Laplace's approximation $\mathcal{N}(\boldsymbol{\mu}_{\mathrm{L}}, \mathbf{S}_{\mathrm{L}} = \mathrm{diag}(\boldsymbol{\sigma}^2_{\mathrm{L}}))$ to the posterior $p(\boldsymbol{\theta}|\mathbf{X})$ [Eq. (24)].

for $j = 1$ **to** J **do**

Draw a posterior sample: $\boldsymbol{\theta}^j \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{L}}, \mathbf{S}_{\mathrm{L}} = \mathrm{diag}(\boldsymbol{\sigma}^2_{\mathrm{L}}))$.

Obtain a predictive trajectory $\overline{\boldsymbol{x}}^j_{1:\mathrm{T}}$, given the parametrization $\boldsymbol{\theta}^j$ utilizing Algorithm 2.

Estimate the observable $\hat{a}(\boldsymbol{\theta}^j) = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} a(\overline{\boldsymbol{x}}^j_{\mathrm{t}})$, given the trajectory $\overline{\boldsymbol{x}}^j_{1:\mathrm{T}}$.

end for

Estimate the desired quantiles with $\hat{a}(\boldsymbol{\theta}^{1:J})$.

in Fig. 7. The latter depicts the Ramachandran plot in terms of the dihedral angles based on various amounts of training data N and compares it with the one predicted by the trained model on the same N as well as with the reference (obtained with N = 10 000). We note that the trained model yields Ramachandran plots that more closely resemble the reference as compared to the ones computed by the training data alone. The encoder, trained with N = 200, is capable of generating atomistic configurations leading to (ϕ, ψ) tuples which are not included in the training data.

The ARD prior in Eq. (20) drives 58% of the parameters $\boldsymbol{\theta}$ to zero (as a threshold, we consider a parameter to be inactive when its value drops below 1.0×10^{-4}). In contrast, all network parameters $\boldsymbol{\theta}$ remain active while optimizing the objective without the ARD prior. Apart from the qualitative advantage, the sparsity-inducing prior provides a strong regularization in the presence of limited data and yields superior predictive estimates. In addition to obtaining sparse solutions, the ARD prior facilitates the identification of physically meaningful latent representations for limited data (e.g., N = 50), as shown in Fig. 8. Without the ARD prior, the data are encoded in a rather small region of the latent space.

In Fig. 9, we attempt to provide insight into the physical meaning of the CVs $\boldsymbol{z}$ identified. In particular, we plot the atomistic configurations $\boldsymbol{x}$ corresponding to various values of the first CV z_1 while keeping $z_2 = 0$. The conformational transition in predicted atomistic configurations can be clearly recognized in the peptides of Fig. 9. We note that we start on the left ($z_1 < 0$) with α configurations, then move towards β-1 (starting at $z_1 \approx -1$), and finally obtain β-2 configurations. For illustration purposes, the predictions in Fig. 9 are based solely on the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})$ of the probabilistic decoder $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \mathbf{S}_{\boldsymbol{\theta}} = \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}))$. We note that for each value of the CVs $\mathbf{z}$ several atomistic realizations $\mathbf{x}$ can be drawn from $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, as depicted in Fig. 10. This figure reveals the characteristic and relevant movement of the backbone that is captured by the predictive mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}) = f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{z})$. Fluctuations of less relevant outer hydrogen atoms [see Figs. 10(b)–10(d)] are recognized as noise of the decoder $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}\big(\boldsymbol{\mu}(\mathbf{z}), \mathbf{S}_{\boldsymbol{\theta}} = \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}})\big)$ denoted in Eq. (28). We also note that the corresponding entries of $\boldsymbol{\sigma}_{\boldsymbol{\theta}}$ responsible for the outer hydrogen atoms are five times larger compared to the remaining atoms. The proposed model can therefore in an unsupervised fashion identify the central role of the backbone coordinates whereas this physical insight is pre-assumed in Refs. 8 and 68.

In order to gain further insight into the relation between the dihedral angles ϕ, ψ and the discovered CVs $\mathbf{z}$, we plot in Figs. 11 and 12 the corresponding maps for various combinations of $\mathbf{z}$-values. While it is clear that the map is not always bijective, the figures reveal the strong correlation between the two sets of variables. It should also be noted that in contrast to the dihedral angles, the $\mathbf{z}$ value for a given atomistic configuration $\mathbf{x}$ is not unique but rather there is a whole distribution as implied by $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. For the aforementioned plots, we computed the $\mathbf{z}$ from the mean of this density, i.e., $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x})$.

The trained model can also be employed in computing predictive estimates of observables $\int a(\mathbf{x})\ p_{target}(\mathbf{x})\ d\mathbf{x}$ by

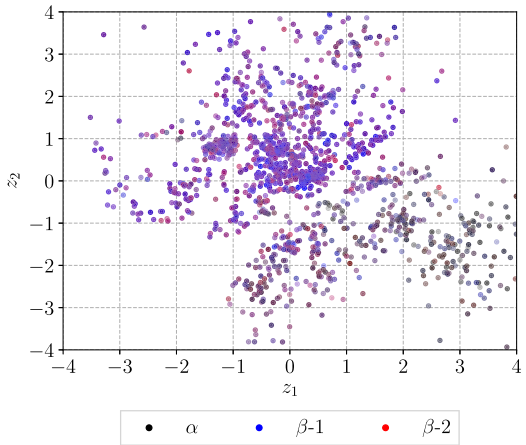


FIG. 14. Representation of the training data $\boldsymbol{X}$ with $N = 3000$ in the encoded collective variable space. The inferred approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\boldsymbol{x})$ of the latent CVs separates residues mostly belonging to the β conformations (mixture of red and blue) and peptide configurations containing largely residues in the α configuration (black). Here, the mean $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x})$ of the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}), \mathbf{S}_{\boldsymbol{\phi}} = \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\boldsymbol{x})))$ is depicted.

Published under license by AIP Publishing

making use of $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and samples drawn from it, as described in Sec. II D. We illustrate this by computing the radius of gyration (Rg)[107,129] given as

$$a_{\mathrm{Rg}}(\boldsymbol{x}) = \sqrt{\frac{\sum_p m_p ||\boldsymbol{x}_p - \boldsymbol{x}_{\mathrm{COM}}||^2}{\sum_p m_p}}. \tag{38}$$

The sum in Eq. (38) considers all atoms $p = 1, \ldots, P$ of the peptide, where m_p and $\boldsymbol{x}_p$ denote the mass and the coordinates of each atom, respectively. $\boldsymbol{x}_{\mathrm{COM}}$ denotes the center of mass of the peptide. The histogram of $a_{\mathrm{Rg}}(\boldsymbol{x})$ reflects the distribution of the size of the peptide and is correlated with the various conformations.[129]

In the estimates that we depict in Fig. 13, we have also employed the posterior approximation of the model parameters $\boldsymbol{\theta}$ obtained as described in Sec. II F in order to compute credible intervals for the observable. These credible intervals are estimated as described in Algorithm 4 utilizing $J = 3000$ samples. We observe that the model's predictive confidence increases with the size of the training data. This is reflected in shrinking credible intervals in Fig. 13 for increasing N.

In summary for ALA-2, we note that the proposed methodology for identifying CVs (Fig. 6) and predicting observables (Figs. 7 and 13) works well with small size datasets, e.g., N = {50, 200, 500}.

B. ALA-15

1. Simulation of ALA-15 and model specification

The following example considers a larger alanine peptide with 15 residues, ALA-15 which consists of 162 atoms giving rise to $\dim(\boldsymbol{x}) = 486$ with 480 DOF. The reference dataset $\mathbf{X}$ has been obtained in a similar manner as specified in Sec. III A 1 with the only difference being that we utilize a replica-exchange molecular dynamics[130] algorithm with 21 temperature replicas distributed according to $T_i = T_0 e^{\kappa \cdot i}$ ($T_0 = 270$ K, and $\kappa = 0.04$). This leads to an analogous simulation setting as employed in Ref. 107. The datasets are obtained as mentioned in the previous example. We consider here N = {300, 3000, 5000}. Using the same model specifications as in Sec. III A 2, we present next a summary of the obtained results.

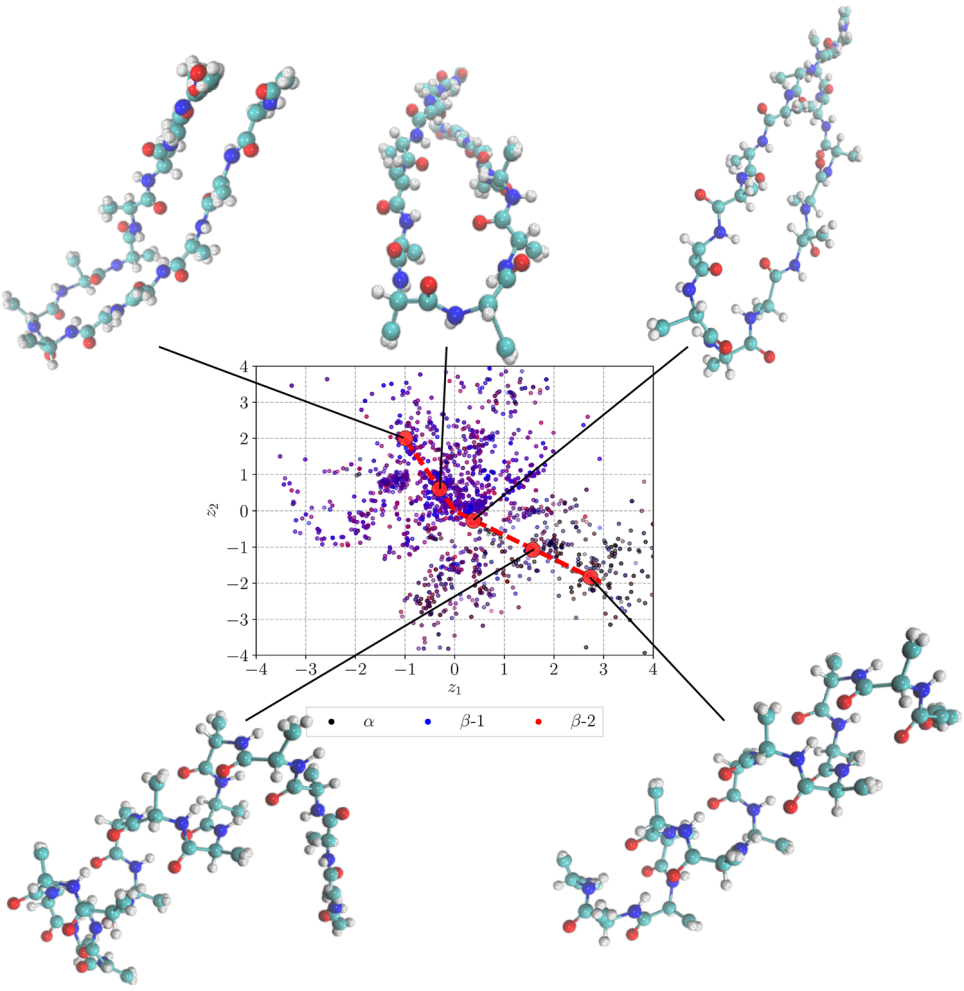


FIG. 15. *Predicted* configurations $\boldsymbol{x}$ for decoding CVs indicated as red points on the dashed line in the plot. Depicted configurations have been produced by evaluating the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})$ of $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\mathbf{z})$. Moving along the path, we obtain atomistic configurations $\boldsymbol{x}$ partially consisting of the conformations α, β-1, and β-2 in the ALA-15 peptide resulting in peptide secondary structures such as β-sheet (top left), β-hairpin (top middle and right), and α-helix (bottom row).

Published under license by AIP Publishing

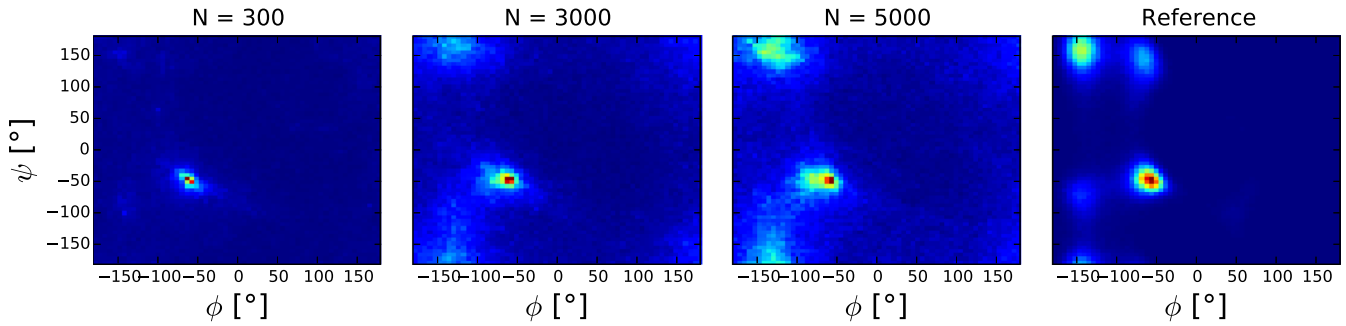
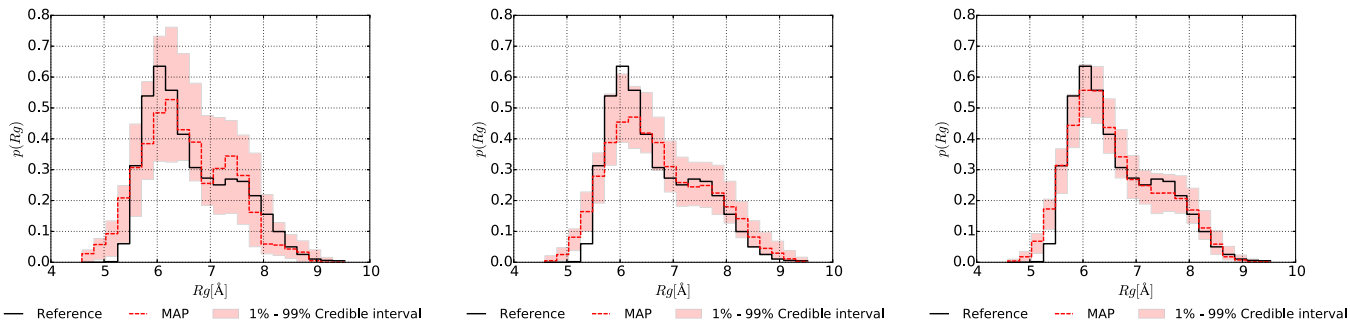


FIG. 16. Predicted Ramachandran plots with dim(**z**) = 2 for various sizes N of the training dataset (first three plots from the left). Depicted predictions are MAP estimates based on T = 10 000 samples. The plot on the right is the reference MD prediction with N = 10 000 configurations.

2. Results

For visualization purposes of the latent CV space, we assumed dim(**z**) = 2 in the following, even though the presence of 15 residues each requiring a pair of dihedral angles (ϕ, ψ) would potentially suggest a higher-dimensional representation. However, when considering test cases with dim(**z**) = {15, 30}, no significant differences were observed in the predictive capabilities. This is in agreement with Ref. 131 where it is argued based on density functional theory calculations that not all dihedral angles are equally relevant. The (ϕ, ψ) pairs within a peptide chain show high correlation. Multilayer neural networks provide the capability of transforming independent CVs (as considered in this study) to correlated ones by passing them through the subsequent network layers. This explains the reasonable predictive quality of the model using independent and low-dimensional CVs with dim(**z**) = 2. Considering more expressive $p_{\boldsymbol{\theta}}(\mathbf{z})$ than the standard Gaussian employed, could have accounted (in part) for such correlations. In this example, by employing the ARD prior, only 43% of the decoder parameters $\boldsymbol{\theta}$ remained effective.

Figure 14 depicts the posterior means of the N = 3000 training data in the CV space **z**. Given that a peptide configuration contains residues from different conformations labelled here as α, β-1, and β-2 and residues in intermediate (ϕ, ψ) states, we applied the following rule for labelling/coloring each datapoint. The assigned color in Fig. 14 is a mixture between the RGB colors: black (for α), blue (for β-1), and red (for β-2). The mixture weights of the assigned color are proportional to the number of residues belonging to the α (black), β-1 (blue), and β-2 (red) conformations, normalized by the total amount of residues which can be clearly assigned to α, β-1, and β-2. Additionally, we visualize the amount of intermediate (ϕ, ψ) states of the residues by the opacity of the scatter points. The opacity reflects the amount of residues which are clearly assigned to the α, β-1, and β-2 conformations compared to the total amount of residues in the peptide. For example, if all residues of a peptide configuration correspond to a specific mode, the opacity is taken as 100%. If all residues are in non-classified intermediate states, the opacity is set to the minimal value which is here taken as 20%.



(a) $N = 300$. (b) $N = 3000$. (c) $N = 5000$.

FIG. 17. Predicted radius of gyration with dim(**z**) = 2 for various sizes N of the training dataset. The MAP estimate indicated in red is compared to the reference (black) solution. The latter is estimated by N = 10 000. The shaded area represents the 1%–99% credible interval, reflecting the induced epistemic uncertainty from the limited amount of training data.

Published under license by AIP Publishing

We note that peptide configurations, in which the majority of residues belong to β-1 (blue) or in the β-2 conformation (red), are clearly separated in the CV space from datapoints with residues predominantly in the α conformation (black). Nevertheless, we observe that the encoder has difficulties separating blue (β-1) and red (β-2) datapoints. We remark though that the related secondary structures[132] resulting from the assembly of residues in β-1 and β-2, such as the β-sheet and β-hairpin, share a similar atomistic representation $\boldsymbol{x}$ which explains the similarity in the CV space.

When one moves in the CV space $\boldsymbol{z}$ along the path indicated by a red dashed line in Fig. 15 and reconstructs the corresponding $\boldsymbol{x}$ using the mean function of the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, we obtain atomistic configurations of the ALA-15 partially consisting of the conformations α, β-1, and β-2 which correspond to the aforementioned secondary structures, i.e., β-sheet (top left), β-hairpin (top middle and right), and α-helix (bottom row).

The ambiguity between β-1 and β-2 states is also reflected in the predicted Ramachandran plot in Fig. 16. Nevertheless properties, independent of the explicit separation of configurations predominantly consisting of residues in β-1 and β-2 states, are predicted accurately by the framework. This is demonstrated with the computed radius of gyration in Fig. 17. The MAP estimate is complemented by the credible intervals which reflect the epistemic uncertainty and are able to envelop the reference profile. As in the previous example, the breadth of the credible intervals shrinks with increasing training data N.

IV. CONCLUSIONS

We presented an unsupervised learning scheme for discovering CVs of atomistic systems. We defined the CVs as latent generators of the atomistic configurations and formulated their identification as a Bayesian inference task. Inference of the posterior distribution of the latent CVs given the fine-scale atomistic training data identifies a probabilistic mapping from the space of atomistic configurations to the latent space. This posterior distribution resembles a dictionary translating atomistic configurations to the lower-dimensional CV space which is inferred during the training procedure. Compared to other dimensionality reduction methods, the proposed scheme is capable of performing well with comparably heterogeneous and small datasets.

We presented the capabilities of the model for the test case of an ALA-2 peptide (see Sec. III). When the dimensionality of the CVs dim($\boldsymbol{z}$) was set to 2, the model discovered variables that correlate strongly with the widely known dihedral angles (ϕ, ψ). Other dimensionality reduction methods[26,30,31,33,38,39,41] rely on an objective keeping small distances between configurations in the atomistic space also small in the latent space. Rather than enforcing this requirement directly, the proposed framework identifies a lower-dimensional representation that clusters configurations in the CV space which show similarities in the atomistic space. The Bayesian formulation presented allows for a rigorous quantification of the unavoidable uncertainties and their propagation in the predicted observables. The ARD prior chosen was shown to lead to on average 45% less parameters compared to the optimization without it.

We presented an approach for approximating the intractable posterior of the decoding model parameters [Eq. (24)] and provided an algorithm (Algorithm 4) for estimating credible intervals. The uncertainty propagated to the observables captures the parameter uncertainty of the decoding neural network $f^{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\boldsymbol{z})$.

In addition to discovering CVs, the generative model employed is able to predict atomistic configurations by sampling the CV space with $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ and mapping the CVs probabilistically via $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ to full atomistic configurations. We showed that the predictive mapping $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ recognizes essential backbone behavior of the peptide while it models fluctuations of the outer hydrogen atoms with the noise of $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ (see Fig. 10). We use the model for predicting observables and quantifying the uncertainty arising from limited training data.

We emphasize that the whole work was based on data represented by Cartesian coordinates $\boldsymbol{x}$ of all the atoms of the ALA-2 (dim($\boldsymbol{x}$) = 66, and 60 DOF adjusted by removing rigid-body motion) and ALA-15 (dim($\boldsymbol{x}$) = 486, and 480 DOF adjusted by removing rigid-body motion) peptides. Considering a pre-processed dataset, e.g., by considering solely coordinates of the backbone atoms, heavy atom positions, or a representation by dihedral angles, assumes the availability of tremendous physical insight. The aim of this work was to reveal CVs with physicochemical meaning and the prediction of observables of complex systems without using any domain-specific physical notion.

Besides the framework proposed, generative adversarial networks (GANs)[133] and their Bayesian reformulation in Ref. 134 open an additional promising avenue in the context of CV discovery and enhanced sampling of atomistic systems. GANs are accompanied by a two player (generator and discriminator) min-max objective which poses known difficulties in training the model. The training of GANs is not as robust as the VAE employed here, and Bayesian formulations are not well studied. In addition, one needs to address the mode collapse issue (see Ref. 135) which is critical for atomistic systems.

Future work involves the use of the CVs discovered in the context of enhanced sampling techniques that can lead to an accelerated exploration of the configurational space. In addition to identifying good CVs, a crucial step for enhanced sampling methods is the biasing potential for lifting deep free-energy wells. In contrast to the ideas, e.g., presented in Refs. 8, 9, and 136, we would advocate a formulation where the biasing potential is based on the lower-dimensional pre-image of the currently visited free-energy surface. To this end, we envision using the posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ to construct a locally optimal biasing potential defined in the CV space which gets updated on the fly as the simulations explore the

Published under license by AIP Publishing

configuration space. The biasing potential can be transformed by the probabilistic mapping of the generative model $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ to the atomistic description as follows:

$$U^{\boldsymbol{x}^{(i)}}_{\text{bias}}(\boldsymbol{x}) \propto -\log \int_{\mathcal{M}_{\text{CV}}} p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})\, d\boldsymbol{z}. \tag{39}$$

Equation (39) is differentiable with respect to atomistic coordinates. Subtracting it from the atomistic potential could accelerate the simulation by "filling-in" the deep free-energy wells.

ACKNOWLEDGMENTS

The authors acknowledge support from the Defense Advanced Research Projects Agency (DARPA) under the Physics of Artificial Intelligence (PAI) program (Contract No. HR00111890034). M.S. gratefully acknowledges the non-material and financial support of the Hanns-Seidel-Foundation, Germany, funded by the German Federal Ministry of Education and Research. M.S. also acknowledges the support of NVIDIA Corporation.

REFERENCES

[1] J. R. Perilla, B. C. Goh, C. K. Cassidy, B. Liu, R. C. Bernardi, T. Rudack, H. Yu, Z. Wu, and K. Schulten, Curr. Opin. Struct. Biol. **31**, 64 (2015).
[2] P. Koutsourelakis, N. Zabaras, and M. Girolami, J. Comput. Phys. **321**, 1252 (2016).
[3] A. Barducci, M. Bonomi, and M. Parrinello, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 826 (2011).
[4] F. Pietrucci and W. Andreoni, Phys. Rev. Lett. **107**, 085504 (2011).
[5] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, J. Chem. Phys. **134**, 135103 (2011).
[6] W. Zheng, M. A. Rohrdanz, and C. Clementi, J. Phys. Chem. B **117**, 12769 (2013).
[7] O. Valsson and M. Parrinello, Phys. Rev. Lett. **113**, 090601 (2014).
[8] W. Chen and A. L. Ferguson, J. Comput. Chem. **39**, 2079 (2018).
[9] P.-Y. Chen and M. E. Tuckerman, J. Chem. Phys. **148**, 024106 (2018).
[10] A. Mitsutake, Y. Mori, and Y. Okamoto, "Enhanced sampling algorithms," in *Biomolecular Simulations: Methods and Protocols*, edited by L. Monticelli and E. Salonen (Humana Press, Totowa, NJ, 2013), pp. 153–195.
[11] C. Bierig and A. Chernov, J. Comput. Phys. **314**, 661 (2016).
[12] *Physico-Chemical and Computational Approaches to Drug Discovery*, RSC *Drug Discovery*, edited by J. Luque and X. Barril (The Royal Society of Chemistry, 2012), pp. FP001–418.
[13] M. A. Rohrdanz, W. Zheng, and C. Clementi, Annu. Rev. Phys. Chem. **64**, 295 (2013).
[14] G. Torrie and J. Valleau, J. Comput. Phys. **23**, 187 (1977).
[15] A. F. Voter, J. Chem. Phys. **106**, 4665 (1997).
[16] D. Hamelberg, J. Mongan, and J. A. McCammon, J. Chem. Phys. **120**, 11919 (2004).
[17] T. Huber, A. E. Torda, and W. F. van Gunsteren, J. Comput.-Aided Mol. Des. **8**, 695 (1994).
[18] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).
[19] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[20] A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 020603 (2008).
[21] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, J. Chem. Phys. **128**, 144120 (2008).
[22] J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein, J. Chem. Theory Comput. **6**, 35 (2010).
[23] F. Pietrucci, Rev. Phys. **2**, 32 (2017).
[24] A. C. Pan, T. M. Weinreich, Y. Shan, D. P. Scarpazza, and D. E. Shaw, J. Chem. Theory Comput. **10**, 2860 (2014).
[25] C. D. Fu, L. F. L. Oliveira, and J. Pfaendtner, J. Chem. Theory Comput. **13**, 968 (2017).
[26] H. Hotelling, J. Educ. Psychol. **24**, 498 (1933).
[27] R. T. McGibbon, B. E. Husic, and V. S. Pande, J. Chem. Phys. **146**, 044109 (2017).
[28] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Struct., Funct., Bioinf. **17**, 412 (1993).
[29] K. Pearson, Philos. Mag. **2**, 559 (1901).
[30] J. M. Troyer and F. E. Cohen, Proteins: Struct., Funct., Bioinf. **23**, 97 (1995).
[31] W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis* (Springer Berlin Heidelberg, 2007).
[32] M. Ceriotti, G. A. Tribello, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **108**, 13023 (2011).
[33] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, Science **290**, 2319 (2000).
[34] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).
[35] M. Balasubramanian and E. L. Schwartz, Science **295**, 7 (2002).
[36] D. L. Donoho and C. Grimes, Proc. Natl. Acad. Sci. U. S. A. **100**, 5591 (2003).
[37] H. Risken and T. Frank, *The Fokker-Planck Equation: Methods of Solution and Applications*, Springer Series in Synergetics (Springer, 1996).
[38] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U. S. A. **102**, 7426 (2005).
[39] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U. S. A. **102**, 7432 (2005).
[40] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Chem. Phys. Lett. **509**, 1 (2011).
[41] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, Appl. Comput. Harmonic Anal. **21**, 113 (2006), special Issue: Diffusion Maps and Wavelets.
[42] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Multiscale Model. Simul. **7**, 842 (2008).
[43] M. A. Rohrdanz, W. Zheng, B. Lambeth, J. Vreede, and C. Clementi, PLoS Comput. Biol. **10**, e1003797 (2014).
[44] W. Zheng, A. V. Vargiu, M. A. Rohrdanz, P. Carloni, and C. Clementi, J. Chem. Phys. **139**, 145102 (2013).
[45] F. Noé and F. Nüske, Multiscale Model. Simul. **11**, 635 (2013).
[46] J. McCarty and M. Parrinello, J. Chem. Phys. **147**, 204109 (2017).
[47] F. Noé and C. Clementi, J. Chem. Theory Comput. **11**, 5002 (2015).
[48] F. Noé, R. Banisch, and C. Clementi, J. Chem. Theory Comput. **12**, 5620 (2016).
[49] M. Duan, J. Fan, M. Li, L. Han, and S. Huo, J. Chem. Theory Comput. **9**, 2490 (2013).
[50] *Learning in Graphical Models*, edited by M. I. Jordan (MIT Press, Cambridge, MA, USA, 1999).
[51] M. Schöberl, N. Zabaras, and P.-S. Koutsourelakis, J. Comput. Phys. **333**, 49 (2017).
[52] L. Felsberger and P. Koutsourelakis, "Communications in computational physics" (to be published); e-print arXiv:1802.03824.
[53] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," e-print arXiv:1312.6114 (2013).
[54] D. J. Rezende, S. Mohamed, and D. Wierstra, in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21–26 June 2014* (PMLR, 2014), pp. 1278–1286, http://proceedings.mlr.press/v32/rezende14.html.
[55] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, Chem. Rev. **116**, 7898 (2016).
[56] W. G. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, J. Phys. Chem. B **111**, 4116 (2007).

Published under license by AIP Publishing

[57]M. S. Shell, J. Chem. Phys. **129**, 144108 (2008).
[58]C. Peter and K. Kremer, Soft Matter **5**, 4357 (2009).
[59]J. Trashorras and D. Tsagkarogiannis, SIAM J. Numer. Anal. **48**, 1647 (2010).
[60]E. Kalligiannaki, M. A. Katsoulakis, P. Plecháč, and D. G. Vlachos, J. Comput. Phys. **231**, 2599 (2012).
[61] V. Harmandaris, E. Kalligiannaki, M. Katsoulakis, and P. Plecháč, J. Comput. Phys. **314**, 355 (2016).
[62]I. Bilionis and N. Zabaras, J. Chem. Phys. **138**, 044313 (2013).
[63]J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth, J. Chem. Theory Comput. **9**, 2466 (2013).
[64]W. G. Noid, J. Chem. Phys. **139**, 090901 (2013).
[65]T. T. Foley, M. S. Shell, and W. G. Noid, J. Chem. Phys. **143**, 243104 (2015).
[66]M. Langenberg, N. E. Jackson, J. J. de Pablo, and M. Müller, J. Chem. Phys. **148**, 094112 (2018).
[67]C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, and V. S. Pande, J. Chem. Theory Comput. **14**, 1887 (2017).
[68]C. Wehmeyer and F. Noé, J. Chem. Phys. **148**, 241703 (2018).
[69]M. M. Sultan, H. K. Wayment-Steele, and V. S. Pande, J. Chem. Theory Comput. **14**, 1887 (2018).
[70]M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
[71] B. J. Alder and T. E. Wainwright, J. Chem. Phys. **31**, 459 (1959).
[72]D. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, New York, NY, USA, 2005).
[73]Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).
[74]Z. Ghahramani, Nature **521**, 452 (2015).
[75]W. von der Linden, V. Dose, and U. von Toussaint, *Bayesian Probability Theory: Applications in the Physical Sciences* (Cambridge University Press, 2014), p. 649.
[76]A. Y. Ng and M. I. Jordan, in *Advances in Neural Information Processing Systems* 14, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, 2002), pp. 841–848.
[77]D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003).
[78]C. Bishop, in *Learning in Graphical Models* (MIT Press, 1999), p. 371403.
[79]A. Cichocki and S.-i. Amari, Entropy **12**, 1532 (2010).
[80]S.-H. Cha, Int. J. Math. Mod. Meth. Appl. Sci. **1**, 300 (2007), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.154.8446.
[81] Inference on the generalized α-divergence is addressed in Ref. 137.
[82]D. J. C. MacKay, Neural Comput. **4**, 448 (1992).
[83]A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, Bayesian Data Analysis, 3rd Edition, Chapman and Hall/CRC 2013, ISBN 9781439840955.
[84]E. T. Jaynes, Math. Intell. **27**, 83 (2005).
[85]M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, J. Mach. Learn. Res. **14**, 1303 (2013).
[86]R. Ranganath, S. Gerrish, and D. Blei, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* [PMLR 33, 814–822 (2014)], http://proceedings.mlr.press/v33/ranganath14.html.
[87]J. W. Paisley, D. M. Blei, and M. I. Jordan, in *International Conference on Machine Learning* (Omnipress, 2012).
[88]A. P. Dempster, N. M. Laird, and D. B. Rubin, J. R. Stat. Soc., Ser. B: Methodol. **39**, 1 (1977).
[89]R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models* (MIT Press, Cambridge, MA, USA, 1999), pp. 355–368.
[90]T. M. Ruiz, J. R. Francisco, and D. Blei, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2016), pp. 460–468.
[91] P. Zhao and T. Zhang, "Accelerating minibatch stochastic gradient descent using stratified sampling," e-print arXiv:1405.3080 (2014).
[92]L. Bottou, F. Curtis, and J. Nocedal, SIAM Rev. **60**, 223 (2018).
[93]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," e-print arXiv:1412.6980 (2014).
[94]P.-A. Mattei and J. Frellsen, Adv. Neural Info. Proc. Sys. **31**, 3859 (2018), https://papers.nips.cc/paper/7642-leveraging-the-exact-likelihood-of-deep-latent-variable-models.
[95]N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[96]W. K. Hastings, Biometrika **57**, 97 (1970).
[97]L. L. Cam, Int. Stat. Rev. **58**, 153 (1990).
[98]M. West, *Bayesian Statistics* (Oxford University Press, 2003), pp. 723–732.
[99]M. A. Figueiredo and S. Member, IEEE Trans. Pattern Anal. Mach. Intell. **25**, 1150 (2003).
[100]D. J. C. MacKay and R. M. Neal, "Automatic relevance determination for neural networks," Technical Report, University of Cambridge, 1994 (unpublished).
[101] M. E. Tipping, J. Mach. Learn. Res. **1**, 211 (2001).
[102]H. Ritter, A. Botev, and D. Barber, in *International Conference on Learning Representations* (ICLR, 2018), https://iclr.cc/Conferences/2018/Schedule?showEvent=224.
[103]P. E. Smith, J. Chem. Phys. **111**, 5568 (1999).
[104]J. Hermans, Proc. Natl. Acad. Sci. U. S. A. **108**, 3095 (2011).
[105]G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, J. Mol. Biol. **7**, 95 (1963).
[106]R. Vargas, J. Garza, B. P. Hay, and D. A. Dixon, J. Phys. Chem. A **106**, 3213 (2002).
[107]S. P. Carmichael and M. S. Shell, J. Phys. Chem. B **116**, 8383 (2012).
[108]E. J. Sorin and V. S. Pande, Biophys. J. **88**, 2472 (2005).
[109]A. J. DePaul, E. J. Thompson, S. S. Patel, K. Haldeman, and E. J. Sorin, Nucleic Acids Res. **38**, 4856 (2010).
[110]M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, New York, NY, USA, 1989).
[111]A. Onufriev, D. Bashford, and D. A. Case, Proteins: Struct., Funct., Bioinf. **55**, 383 (2004).
[112]W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, J. Am. Chem. Soc. **112**, 6127 (1990).
[113]H. Berendsen, D. van der Spoel, and R. van Drunen, Comput. Phys. Commun. **91**, 43 (1995).
[114]E. Lindahl, B. Hess, and D. van der Spoel, Mol. Model. Annu. **7**, 306 (2001).
[115]D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 1701 (2005).
[116]B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[117]S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, Bioinformatics **29**, 845 (2013).
[118]S. Páll, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl, in *Solving Software Challenges for Exascale*, edited by S. Markidis and E. Laure (Springer International Publishing, Cham, 2015), pp. 3–27.
[119]M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1-2**, 19 (2015).
[120]D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation* (MIT Press, Cambridge, MA, USA, 1986), pp. 318–362.
[121] C. Van Der Malsburg, in *Brain Theory*, edited by G. Palm and A. Aertsen (Springer Berlin Heidelberg, Berlin, Heidelberg, 1986), pp. 245–248.
[122]S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. (Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998).
[123]D. J. Rezende and S. Mohamed, in *Proceedings of the 32nd International Conference on Machine Learning* [PMLR **37**, 1530–1538 (2015)], http://proceedings.mlr.press/v37/rezende15.html.
[124]D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, Adv. Neural Info. Proc. Syst. **29**, 4743 (2016), http://papers.nips.cc/paper/6581-improved-variational-inference-with-inverse-autoregressive-flow.

Published under license by AIP Publishing

[125]J. C. Pinheiro and D. M. Bates, Stat. Comput. **6**, 289 (1996).
[126]G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, in *Advances in Neural Information Processing Systems* 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), pp. 971–980.
[127]P. Ramachandran, B. Zoph, and Q. V. Le"Searching for activation functions" (2017), arXiv:1710.05941.
[128]W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[129]A. M. Fluitt and J. J. de Pablo, Biophys. J. **109**, 1009 (2015).
[130]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[131]A. Marini and R. Y. Dong, Phys. Rev. E **83**, 041712 (2011).
[132]Y. Zhou, A. Kloczkowski, E. Faraggi, and Y. Yang, "Prediction of protein secondary structure," in *Methods in Molecular Biology* (Springer, New York, 2016).
[133]I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Adv. Neural Info. Proc. Syst. **27**, 2672 (2014), https://papers.nips.cc/paper/5423-generative-adversarial-nets.
[134]Y. Saatchi and A. G. Wilson, Adv. Neural Info. Proc. Syst. **30**, 3622 (2017), https://papers.nips.cc/paper/6953-bayesian-gan.
[135]T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, in *Advances in Neural Information Processing Systems* 29, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 2234–2242.
[136]R. Galvelis and Y. Sugita, J. Chem. Theory Comput. **13**, 2489 (2017).
[137]J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernandez-Lobato, and R. Turner, in *Proceedings of the 33rd International Conference on Machine Learning*, PMLR, Vol. 48, edited by M. F. Balcan and K. Q. Weinberger (PMLR, New York, New York, USA, 2016), pp. 1511–1520, http://proceedings.mlr.press/v48/hernandez-lobatob16.html.

Published under license by AIP Publishing

Appendix E

Embedded-physics machine learning for coarse-graining and collective variable discovery without data

E.1 Relation with Expectation-Propagation

This section emphasizes the relationship of hierarchical variational models with expectation-propagation (EP) [557].

The following is not directly relevant to optimization of the objective Equation (6.7), but it shows the existence of an upper bound of the entropy term $-\mathbb{E}_{q(\mathbf{x})}\left[\log q(\mathbf{x})\right]$. Similar to Equation (6.9) one denotes,

$$\begin{aligned}
-\mathbb{E}_{q(\mathbf{x})}\left[\log q(\mathbf{x})\right] &= -\mathbb{E}_{q(\mathbf{x})}\left[\log q(\mathbf{x}) - D_{KL}\left(q(\mathbf{z}|\mathbf{x})||q(\mathbf{z}|\mathbf{x})\right)\right] \\
&\leq \mathbb{E}_{q(\mathbf{x})}\left[-\log q(\mathbf{x}) + D_{KL}\left(r(\mathbf{z}|\mathbf{x})||q(\mathbf{z}|\mathbf{x})\right)\right] \\
&= \mathbb{E}_{q(\mathbf{x})}\left[\mathbb{E}_{r(\mathbf{z}|\mathbf{x})}\left[-\log q(\mathbf{x}) - \log q(\mathbf{z}|\mathbf{x}) + \log r(\mathbf{z}|\mathbf{x})\right]\right] \\
&= \mathbb{E}_{q(\mathbf{x})}\left[\mathbb{E}_{r(\mathbf{z}|\mathbf{x})}\left[-\log q(\mathbf{x}) - \log \frac{q(\mathbf{x}|\mathbf{z})q(\mathbf{z})}{q(\mathbf{x})} + \log r(\mathbf{z}|\mathbf{x})\right]\right] \\
&= \mathbb{E}_{q(\mathbf{x})}\left[\mathbb{E}_{r(\mathbf{z}|\mathbf{x})}\left[-\log q(\mathbf{x}|\mathbf{z}) - \log q(\mathbf{z}) + \log r(\mathbf{z}|\mathbf{x})\right]\right]. \qquad \text{(E.1)}
\end{aligned}$$

The bound in Equation (E.1) is tractable if sampling from $q(\mathbf{x})$ and $r(\mathbf{z}|\mathbf{x})$ is feasible. Both bounds (Eqs. E.1 and 6.9) show similarities to the derivation of EP [557] and variational Bayesian inference [416]. However, note that the lower bound in Equation (6.11) is connected to the objective in EP, although EP only minimizes $D_{\mathrm{KL}}(q\|r)$ locally. The bound derived with $q(\mathbf{x})$ results in a tighter bound compared with variational autoencoders with $q(\mathbf{x}|\mathbf{z})$, as $\mathbb{H}[q(\mathbf{x})] \geq \mathbb{H}[q(\mathbf{x}|\mathbf{z})]$ (for details, see [558]).

E.2 Estimating the relative increase of the KL divergence

The relative increase of the KL divergence induced by decreasing the temperature is denoted as in Equation (6.36), with

$$c_k = \frac{\log(Z(\beta_{k+1})) - \log(Z(\beta_k)) + (\beta_{k+1} - \beta_k)\,\langle U(\mathbf{x})\rangle_{q(\mathbf{x},\mathbf{z})}}{\log Z(\beta_k) + \beta_k\,\langle U(\mathbf{x})\rangle_{q(\mathbf{x},\mathbf{z})} - \langle \log r(\mathbf{z}|\mathbf{x})\rangle_{q(\mathbf{x},\mathbf{z})} - \mathbb{H}(q(\mathbf{x},\mathbf{z}))}.$$

The following addresses the estimation of $\log(Z(\beta_{k+1})) - \log(Z(\beta_k))$ with $\Delta\beta_k = \beta_{k+1} - \beta_k$:

$$\begin{aligned} Z(\beta_k + \Delta\beta_k) &= \int e^{-(\beta_k+\Delta\beta_k)U(\mathbf{x})} d\mathbf{x} \\ &= \int \frac{e^{-(\beta_k+\Delta\beta_k)U(\mathbf{x})}}{\frac{e^{-\beta_k U(\mathbf{x})}}{Z(\beta_k)}} \frac{e^{-\beta_k U(\mathbf{x})}}{Z(\beta_k)}\, d\mathbf{x} \\ &= Z(\beta_k) \int e^{-\Delta\beta_k U(\mathbf{x})} p_{\text{target}}(\mathbf{x};\beta_k)\, d\mathbf{x} \\ &= Z(\beta_k) \int e^{-\Delta\beta_k U(\mathbf{x})} \frac{e^{-\beta_k U(\mathbf{x})}\; r(\mathbf{z}|\mathbf{x})}{q(\mathbf{x}|\mathbf{z})\; q(\mathbf{z})} q(\mathbf{x},\mathbf{z})\, d\mathbf{x}\, d\mathbf{z}. \end{aligned} \tag{E.2}$$

We are interested in $\log(Z(\beta_{i+k})) - \log(Z(\beta_k))$. Therefore, we write:

$$\begin{aligned} \log(Z(\beta_{i+1})) - \log(Z(\beta_i)) &= \log \int e^{-\Delta\beta U(\mathbf{x})} \underbrace{\frac{e^{-\beta_i U(\mathbf{x})}\; r(\mathbf{z}|\mathbf{x})}{q(\mathbf{x}|\mathbf{z})\; q(\mathbf{z})}}_{w} q(\mathbf{x},\mathbf{z})\, d\mathbf{x}\, d\mathbf{z} \\ &\approx \log \sum_{i=1}^{N} e^{-\Delta\beta U(\boldsymbol{x}^{(i)})} W^{(i)}. \end{aligned} \tag{E.3}$$

Equation (E.3) depicts a noisy Monte Carlo estimator for $\log(Z(\beta_{k+1})) - \log(Z(\beta_k))$ based on importance sampling [559] with the following normalized weights:

$$W^{(i)} = \frac{w^{(i)}}{\sum w^{(i)}} \quad \text{with} \quad w^{(i)} \propto \frac{e^{-\beta_i U(\mathbf{x})}\; r(\mathbf{z}|\mathbf{x})}{q(\mathbf{x}|\mathbf{z})\; q(\mathbf{z})}. \tag{E.4}$$

As $e^{-\beta_i U(\mathbf{x})} r(\mathbf{z}|\mathbf{x})$ may be small for samples $(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \sim q(\mathbf{x},\mathbf{z})$, we use instead $\log \bar{w}^{(i)}$ with $\log \bar{w}^{(i)} = \log w^{(i)} - a$ and $a = \max\{\log w^{(i)}\}$ to avoid numerical issues.

Whereas above we showed an approximate estimator for $\log(Z(\beta_{k+1})) - \log(Z(\beta_k))$, the following addresses $\log(Z(\beta_k))$. To estimate the relative increase in the KL divergence, one requires the normalization constant as mentioned in Equation (6.36). Multistage sampling [560] provides a way to approximate $Z(\beta_i)$, given all previous $Z(\beta_k)$ with $k < i$ and $\beta_i > \beta_{i-1}$:

$$\frac{Z(\beta_i)}{Z(0)} = \frac{Z(\beta_1)}{Z(\beta_0)} \cdot \frac{Z(\beta_2)}{Z(\beta_1)} \cdots \frac{Z(\beta_{i-1})}{Z(\beta_{i-2})}. \tag{E.5}$$

The ratios $\frac{Z(\beta_{k-1})}{Z(\beta_{k-2})}$ are given by Equation (E.3). The remaining component to be estimated is $Z(0)$, as we utilize the expression from Equation (E.3) to estimate the ratios of normalization factors. To avoid learning parametrizations $\boldsymbol{\theta}$ yielding almost uniform $q(\mathbf{x})$ on an infinite domain, which occurs in the limit when $\beta = 0$, we define the following auxiliary potential to restrict the domain:

$$U_{aux}(\mathbf{x}) = \begin{cases} U(\mathbf{x}), & \text{if } \mathbf{x} \in [-b, b]^{\dim(\mathbf{x})} \\ -\frac{u}{\beta}\mathbf{x}, & \mathbf{x} < -b \\ \frac{u}{\beta}\mathbf{x}, & \mathbf{x} > b, \end{cases} \tag{E.6}$$

with $u = 10 \times 10^2$. The above extension does not influence the potential energy $U(\mathbf{x})$ at relevant temperatures.

The initial $Z(\beta_0)$ is computed with importance sampling. This is done only once upon convergence of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ for β_0:

$$\begin{aligned} Z(\beta_0) &= \int e^{-\beta_0 U(\mathbf{x})}\, d\mathbf{x} \\ &= \int e^{-\beta_0 U(\mathbf{x})}\, r(\mathbf{z}|\mathbf{x})\, d\mathbf{x}\, d\mathbf{z} \\ &= \int \underbrace{\frac{e^{-\beta_0 U(\mathbf{x})}\, r(\mathbf{z}|\mathbf{x})}{q(\mathbf{x}, \mathbf{z})}}_{w} q(\mathbf{x}, \mathbf{z})\, d\mathbf{x}\, d\mathbf{z}. \end{aligned} \tag{E.7}$$

With samples $(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) \sim q(\mathbf{x}, \mathbf{z})$, we obtain the following unnormalized weights:

$$w^{(i)} = \frac{e^{-\beta_0 U(\boldsymbol{x}^{(i)})}\, r(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})}{q(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)})}, \tag{E.8}$$

or $\log w^{(i)} = -\beta_0 U(\boldsymbol{x}^{(i)}) + \log r(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}) - \log q(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)})$. Then,

$$\log Z(\beta_0) = -\log N + \log \sum_{i=1}^{N} e^{\log w^{(i)} - c} + c, \tag{E.9}$$

with $c = \max(\log w^{(i)})$.

E.3 ALA-2 coordinate representation

We show the structure of the ALA-2 petpide in Figure E.1. The numbers in the circles, which depict the involved atoms of ALA-2, correspond to the order in which we assemble block-wise the Cartesian coordinates (x_i, y_i, z_i) of atom i to

$$\mathbf{x} = (x_1, y_1, z_1, x_2, y_2, \ldots, x_{22}, y_{22}, z_{22})^T,$$

where i is the atom number as depicted in Figure E.1. For removing rigid-body motion, we fix the Cartesian coordinates (x_6, y_6, z_6) of atom 6, (x_9, y_9) of atom 9, and

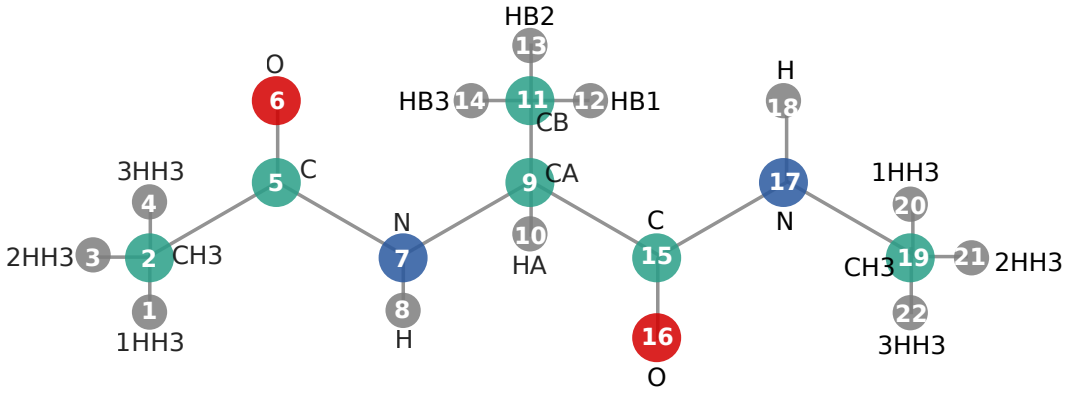


FIGURE E.1: ALA-2 structure with numbered atoms as used for decomposing $\mathbf{x}$.

(y_{15}) of atom 15. The employed PDB structure file is available online at https://github.com/cics-nd/predictive-cvs/blob/master/data_peptide/ala-2/ala2_adopted.pdb.

E.4 Simulation of ALA-2

The procedure for generating a reference trajectory for computing reference observables of ALA-2 is identical to that employed in [415], whereas the data generation approach relies on [556]. The utilized interaction force field is AMBER ff96 [337–339], resolved by an implicit water model based on the generalized Born model [551, 552]. Incorporating an explicit water model would, obviously, provide trajectories that would yield observables closer to the *experimental* reference. An Andersen thermostat is used to maintain fluctuations around the desired temperature $T = 330\,\mathrm{K}$. All reference simulations are carried out using Gromacs [316–322]. The time step is $\Delta t = 1\,\mathrm{fs}$, with a preceding equilibration phase of 50 ns. Thereafter, a trajectory snapshot is taken every 10 ps. Rigid-body motions have been removed from the Cartesian coordinates.

E.5 Observable estimation for ALA-2

We are interested in estimating observables based on predictive models, in contrast to those obtained through reference MD simulations. In general, observables are evaluated as ensemble (MC) or phase (MD) averages, $\int a(\mathbf{x}) p_{\text{target}}(\mathbf{x})\, d\mathbf{x}$, by making use of $q_{\boldsymbol{\theta}}(\mathbf{x})$ and samples drawn by ancestral sampling. We illustrate the radius of gyration (Rg) [68, 556], given as:

$$a_{\text{Rg}}(\mathbf{x}) = \sqrt{\frac{\sum_p m_p \|\mathbf{x}_p - \mathbf{x}_{\text{COM}}\|^2}{\sum_p m_p}}. \tag{E.10}$$

The sum in Equation (E.10) considers all system atoms $p = 1, \ldots, P$, with the atom mass m_p and Cartesian coordinate $\mathbf{x}_p$ of each atom. The center of mass of the peptide is denoted by $\mathbf{x}_{\text{COM}}$. A histogram of $a_{\text{Rg}}(\mathbf{x})$ reflects the statistics of the peptide's average size, which characterize its various conformations [68].

E.6 Gradient normalization

During optimization of the objective in the context of atomistic systems, we encounter significant forces, $\mathbf{F}(\mathbf{x})$. These differ in magnitude owing to sampling atomistic realizations, which induce, e.g., relatively small distances between bonded atoms. This leads to extreme force components. Gradient normalization [561, 562] circumvents disruption of the current set of learned parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ via a single component attached with extreme magnitudes, owing to, e.g., short bonded distances. Once training proceeds, and predicted atomistic realizations are closer to reasonable ones, the gradient normalization becomes redundant, affecting only gradients in extreme settings where the absolute values of $\mathbf{F}(\mathbf{x}) \geq 1 \times 10^{15}$. After an initial learning phase, such extreme magnitudes do not occur, and thus the normalization does not affect or distort the physics induced by evaluating the force field $\mathbf{F}(\mathbf{x})$.

Given a batch of I samples $\left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{I}$ obtained from $q_{\boldsymbol{\theta}}(\mathbf{x})$, we estimate the gradient of the objective, $\mathbf{g}^i(\boldsymbol{x}^{(i)})$ and calculate its ℓ-2 norm:

$$l^i = \|\mathbf{g}^i\|_2. \tag{E.11}$$

The average gradient norm is $\bar{l} = 1/I \sum_{i=1}^{I} l^i$, and we allow a maximal gradient norm based on the mean with $l_{\text{max}} = \kappa \bar{l}$, $\kappa = 3.0$. κ was determined by an empirical study. Those gradients with $l^i > l_{\text{max}}$ are normalized such that

$$\boldsymbol{g}_n^i = \frac{l_{\text{max}}}{l_i} \mathbf{g}^i. \tag{E.12}$$

As mentioned earlier, realistic atomistic systems at relevant temperatures are not exposed to ℓ-2 norms of gradients differing more as twice as compared with the gradient with the lowest ℓ-2 norm. Thus, the gradient normalization is inactive when learning realistic atomistic configurations.

Appendix F

On-the-fly coarse-graining

We are interested in simulating the following distribution, which exhibits a complex inter-atomic potential $U_{\mathrm{f}}(\mathbf{x})$ describing the interactions between fine-scale degrees of freedom $\mathbf{x}$:

$$p_{\mathrm{target}}(\mathbf{x}) = \frac{e^{-\beta U_{\mathrm{f}}(\mathbf{x})}}{Z(\beta)} = \frac{\pi(\mathbf{x})}{Z(\beta)}, \tag{F.1}$$

where β is the inverse temperature and $Z(\beta)$ is the partition function.

F.1 Methodology

The following describes an approach that combines actively learning a biasing potential, enhancing the exploration of the configurational space, and making predictions for observables. For this purpose, we introduce a distribution that includes the reference fine-scale interaction potential $U_{\mathrm{f}}(\mathbf{x})$ and a predictive distribution, e.g., $q(\mathbf{x}) = \int q(\mathbf{x}|\mathbf{z})q(\mathbf{z})\, d\mathbf{z}$:

$$p_{\mathrm{bias}}(\mathbf{x}) = \frac{1}{Z_p} e^{-\beta U_{\mathrm{f}}(\mathbf{x}) - \log q(\mathbf{x})}. \tag{F.2}$$

The distribution $p_{\mathrm{bias}}(\mathbf{x})$ becomes uniform when $q(\mathbf{x}) = p_{\mathrm{target}}(\mathbf{x})$. Therefore, it minimizes the KL-divergence from a uniform distribution to $p_{\mathrm{bias}}(\mathbf{x})$ with respect to q, a valid objective for learning a predictive distribution, while enhancing the exploration of the configurational space. One could furthermore employ a sequence of objectives guided by an auxiliary distribution $p_n(\mathbf{x})$, which could be close to a reference configuration $\mathbf{x}_{\mathrm{ref}}$ defined by a Gaussian $p_n(\mathbf{x}) = \mathcal{N}(\mathbf{x}_{\mathrm{ref}}, \sigma_n^2 \boldsymbol{I})$:

$$\min_q D_{\mathrm{KL},n}\left(p_n(\mathbf{x}) \| p_{\mathrm{bias}}(\mathbf{x})\right). \tag{F.3}$$

For the predictive component $q(\mathbf{x})$, we employ the usual probabilistic model as developed in this work, with

$$q(\mathbf{x}|\boldsymbol{\theta}) = \int q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\mathrm{cf}}) q(\mathbf{z}|\boldsymbol{\theta}_{\mathrm{c}})\, d\mathbf{z} \tag{F.4}$$

parametrized by $\boldsymbol{\theta}$, where $\mathbf{z}$ denotes the latent lower-dimensional CG variables that gives rise to the observations $\mathbf{x}$.

We introduce the objectives $H_n(\boldsymbol{\theta}) = D_{\mathrm{KL},n}\left(p_n(\mathbf{x}) \| p_{\mathrm{bias}}(\mathbf{x})\right)$ as follows:

$$\begin{aligned} H_n(\boldsymbol{\theta}) &= D_{\mathrm{KL},n}\left(p_n(\mathbf{x}) \| p_{\mathrm{bias}}(\mathbf{x})\right), \\ \min_{\boldsymbol{\theta}} H_n(\boldsymbol{\theta}) &= \left\langle \log q(\mathbf{x}|\boldsymbol{\theta}) \right\rangle_{p_n(\mathbf{x})} + \log \mathrm{Z}(\boldsymbol{\theta}). \end{aligned} \tag{F.5}$$

We employ gradient-based stochastic optimization to minimize the objective $H_n(\boldsymbol{\theta})$ (or a sequence of it). The gradient with respect to $\boldsymbol{\theta}$ is as follows:

$$\frac{\partial H_n}{\partial \boldsymbol{\theta}} = \left\langle \left\langle \frac{\partial \log q(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle_{q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_{\mathrm{cf}})} \right\rangle_{p_n(\mathbf{x})} - \left\langle \left\langle \frac{\partial \log q(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle_{q(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_{\mathrm{cf}})} \right\rangle_{p_{\mathrm{bias}}(\mathbf{x})}. \tag{F.6}$$

F.2 Numerical illustration

The following demonstrates the basic capabilities of the proposed approach using a target distribution that is a mixture of two Gaussians, defined as:

$$p_{\mathrm{target}}(\mathbf{x}) = a_1 \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2) + a_2 \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2). \tag{F.7}$$

We employ the following setting: $a_1 = a_2 = 0.5, \boldsymbol{\mu}_1 = (-0.5,\ 0.5)^T, \boldsymbol{\mu}_2 = (0.5,\ -0.5)^T$ and $\boldsymbol{\sigma}_1^2 = (1.0 \times 10^{-3},\ 5.0 \times 10^{-4})^T$, $\boldsymbol{\sigma}_2^2 = (3.0 \times 10^{-3},\ 1.0 \times 10^{-3})^T$. Random walk Markov chain Monte Carlo approaches usually become trapped in one of the depicted modes for the previously defined reference distribution. The resulting distribution is depicted with $-\log p_{\mathrm{target}}(\mathbf{x})$ in Figure F.1.

For $q(\mathbf{x}|\boldsymbol{\theta})$, we employ the following model, inspired by a mixture of factor analyzers [490]:

$$q(\mathbf{x}|\boldsymbol{\theta}) = \int q(\mathbf{x}|\mathbf{z}, \boldsymbol{\gamma}, \boldsymbol{\theta}_{\mathrm{cf}}) p(\mathbf{z}|\boldsymbol{\theta}_{\mathrm{c}}) q(\boldsymbol{\gamma})\, d\mathbf{z}\, d\boldsymbol{\gamma}. \tag{F.8}$$

The distributions are specified with:

- J probabilistic coarse-to-fine mapping distributions,

$$q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_{\mathrm{cf}}^j, \gamma_j = 1) = \mathcal{N}\left(\boldsymbol{\mu}_j + \mathbf{W}_j \mathbf{z}, \boldsymbol{S}_j\right), \tag{F.9}$$

  with $\boldsymbol{\theta}_{\mathrm{cf}}^j = \{\boldsymbol{\mu}_j, \mathbf{W}_j, \boldsymbol{S}_j\}$ where $\boldsymbol{S}_j = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{n_{\mathrm{f}}}^2)$.

- The distribution of the CG variables is independent of $\boldsymbol{\gamma}$,

$$q(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{F.10}$$

- The distribution governing the J mixture components $\boldsymbol{\gamma}$, where we employ a 1-of-J representation with $\gamma_j \in \{0, 1\}$ and $\sum_{j=1}^J = 1$, is:

$$q(\boldsymbol{\gamma}) = \prod_{j=1}^{J} \alpha_j^{\gamma_j}, \tag{F.11}$$

  with the mixing coefficient α_j of the j-th component.

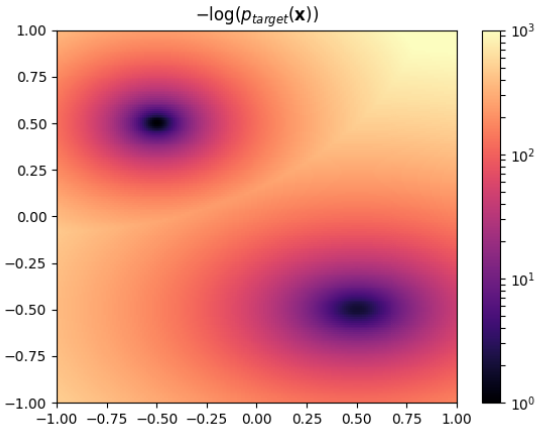
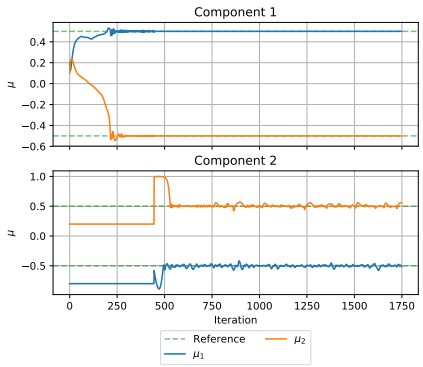


FIGURE F.1: Potential energy surface of the target distribution $-\log p_{\text{target}}(\mathbf{x})$ (log-scale).

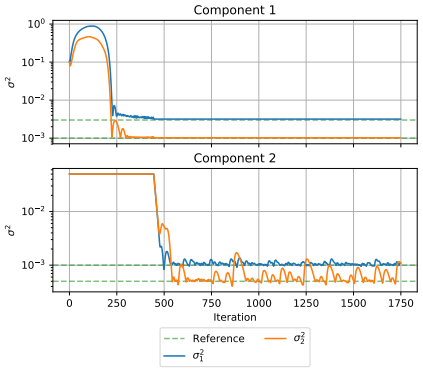
The above model specifications yield the predictive distribution:

$$q(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^{J} \alpha_j q(\mathbf{x}|\gamma_j = 1) = \sum_{j=1}^{J} \alpha_j \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{W}_j \mathbf{W}_j^T + \boldsymbol{S}_j). \tag{F.12}$$

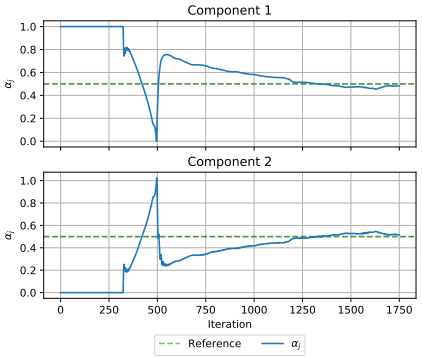
We choose $\dim(\mathbf{z}) = 2$ and start the training with $\sigma_0 = 0.02$, which increases linearly. We optimize $H_n(\boldsymbol{\theta})$ initially with one active component in the mixture model and activate the second component upon convergence of H_n. The training process is visualized in Figure F.2 and we show resulting potential energy surface of p_{bias} in Figure F.3.



(a) Mean values of the components $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}^j_{\text{cf}}, \gamma_j)$.

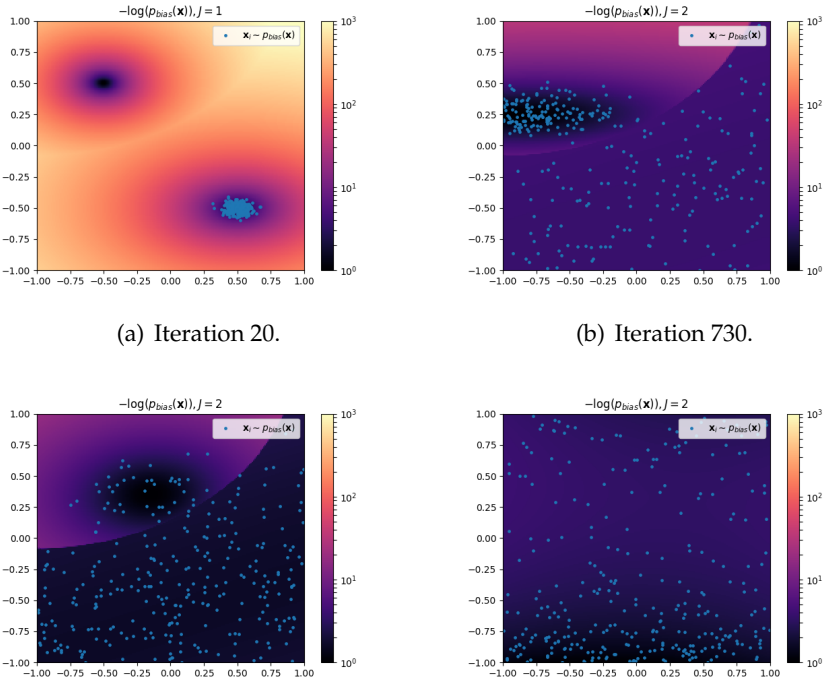


(b) Variance values of the components $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}^j_{\text{cf}}, \gamma_j)$.



(c) Mixture weights α_j.

FIGURE F.2: Properties of the model components $q(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}^j_{\text{cf}}, \gamma_j)$ and corresponding mixture weights α_j during training, compared with the target (indicated by dashed lines).



(a) Iteration 20.

(b) Iteration 730.

(c) Iteration 1500.

(d) Iteration 1700.

FIGURE F.3: Implied potential energy by p_{bias} during the training process. As objective we seek to obtain $p_{\text{bias}}(\mathbf{x})$ such that $p_{\text{bias}}(\mathbf{x})$ is close to a uniform distribution, then $q(\mathbf{x}|\boldsymbol{\theta}) = p_{\text{target}}(\mathbf{x})$.

Bibliography

[1] K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern, Nature **450**, 913 (2007).

[2] W. E, *Principles of multiscale modeling* (Cambridge University Press, Aug. 2011).

[3] J. Shen, J. Liu, H. Li, Y. Gao, X. Li, Y. Wu, and L. Zhang, Phys. Chem. Chem. Phys. **17**, 7196 (2015).

[4] U. Landman, in Computer simulation studies in condensed matter physics, edited by D. P. Landau, K. K. Mon, and H.-B. Schüttler (1988), pp. 108–123.

[5] M. O. Steinhauser and S. Hiermaier, International journal of molecular sciences **10**, PMC2801990[pmcid], 5135 (2009).

[6] D. Vlachakis, E. Bencurova, N. Papangelopoulos, and S. Kossida, in , Vol. 94, edited by R. Donev, Advances in Protein Chemistry and Structural Biology (Academic Press, 2014), pp. 269–313.

[7] P. W. Hildebrand, A. S. Rose, and J. K. Tiemann, Trends in Biochemical Sciences (2019).

[8] D. A. Antunes, D. Devaurs, and L. E. Kavraki, Expert Opinion on Drug Discovery **10**, PMID: 26414598, 1301 (2015).

[9] X. Liu, D. Shi, S. Zhou, H. Liu, H. Liu, and X. Yao, Expert Opinion on Drug Discovery **13**, PMID: 29139324, 23 (2018).

[10] B. J. Alder and T. E. Wainwright, The Journal of Chemical Physics **27**, 1208 (1957).

[11] B. J. Alder and T. E. Wainwright, The Journal of Chemical Physics **31**, 459 (1959).

[12] A. Rahman, Phys. Rev. **136**, A405 (1964).

[13] J. Dongarra and P. Luszczek, "Top500", in *Encyclopedia of parallel computing*, edited by D. Padua (Springer US, Boston, MA, 2011), pp. 2055–2057.

[14] D. Frenkel and B. Smit, eds., *Understanding molecular simulation: from algorithms to applications*, 1st (Academic Press, Inc., Orlando, FL, USA, 1996).

[15] S. Kim, Physics Procedia **53**, 26th Annual CSP Workshop on "Recent Developments in Computer Simulation Studies in Condensed Matter Physics", CSP 2013, 60 (2014).

[16] S. Zeng, G. Zhou, J. Guo, F. Zhou, and J. Chen, Scientific Reports **6**, Article, 24906 EP (2016).

[17] W. Y. Yang and M. Gruebele, Nature **423**, 193 (2003).

[18] J.-C. Horng, V. Moroz, and D. P. Raleigh, Journal of Molecular Biology **326**, 1261 (2003).

[19] M. B. Prigozhin and M. Gruebele, Physical chemistry chemical physics : PCCP **15**, PMC3632410[pmcid], 3372 (2013).

[20] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw, Annual Review of Biophysics **41**, PMID: 22577825, 429 (2012).

[21] C. M. Dobson, Nature **426**, 884 (2003).

[22] A. Mukherjee, D. Morales-Scheihing, P. C. Butler, and C. Soto, Trends in molecular medicine **21**, S1471-4914(15)00087-8[PII], 439 (2015).

[23] S. Costes, Current Opinion in Pharmacology **43**, 104 (2018).

[24] P. T. Lansbury and H. A. Lashuel, Nature **443**, 774 (2006).

[25] I. A. Tayubi, A. Firoz, and A. Malik, "Protein misfolding and amyloid formation in alzheimer's disease", in *Proteostasis and chaperone surveillance*, edited by L. R. Singh, T. A. Dar, and P. Ahmad (Springer India, New Delhi, 2015), pp. 119–135.

[26] C. M. Dobson, Philosophical transactions of the Royal Society of London. Series B, Biological sciences **356**, PMC1088418[pmcid], 133 (2001).

[27] D. J. Selkoe, Nature Cell Biology **6**, 1054 (2004).

[28] F. U. Hartl, Annual Review of Biochemistry **86**, PMID: 28441058, 21 (2017).

[29] M. C. Zwier and L. T. Chong, Current opinion in pharmacology **10**, 745 (2010).

[30] E. A. Carter, Science **321**, 800 (2008).

[31] M. Lundstrom, P. Cummings, and M. Alam, "Investigative tools: theory, modeling, and simulation", in *Nanotechnology research directions for societal needs in 2020*, Vol. 1 (Springer, 2011), pp. 29–69.

[32] X. L. Hu, S. Piccinin, A. Laio, and S. Fabris, ACS Nano **6**, PMID: 23145574, 10497 (2012).

[33] H.-P. Hsu and K. Kremer, The Journal of Chemical Physics **144**, 154907 (2016).

[34] P. Kordt, J. J. M. van der Holst, M. Al Helwi, W. Kowalsky, F. May, A. Badinski, C. Lennartz, and D. Andrienko, Advanced Functional Materials **25**, 1955 (2015).

[35] Y. Hou, L. Wang, D. Wang, X. Qu, and J. Wu, Applied Sciences **7**, 770 (2017).

[36] J. R. Kermode, T. Albaret, D. J. Sherman, N. Bernstein, P. Gumbsch, M. C. Payne, G. Csányi, and A. D. Vita, Nature **455**, 1224 (2008).

[37] W.-M. Choi, Y. H. Jo, S. S. Sohn, S. Lee, and B.-J. Lee, npj Computational Materials **4**, 1 (2018).

[38] E. H. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten, Structure **17**, 1295 (2009).

[39] P. P. Ewald, Annalen der Physik **369**, 253 (1921).

[40] W. Liu, B. Schmidt, G. Voss, and W. Müller-Wittig, in High performance computing – hipc 2007, edited by S. Aluru, M. Parashar, R. Badrinath, and V. K. Prasanna (2007), pp. 185–196.

[41] J. A. Anderson, C. D. Lorenz, and A. Travesset, Journal of Computational Physics **227**, 5342 (2008).

[42] M. S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. Legrand, A. L. Beberg, D. L. Ensign, C. M. Bruns, and V. S. Pande, Journal of Computational Chemistry **30**, 864 (2009).

[43] M. J. Harvey, G. Giupponi, and G. D. Fabritiis, Journal of Chemical Theory and Computation **5**, PMID: 26609855, 1632 (2009).

[44] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, PLoS computational biology **13**, PCOMPBIOL-D-16-01696[PII], e1005659 (2017).

[45] D. Janežič, U. Borštnik, and M. Praprotnik, "Parallel approaches in molecular dynamics simulations", in *Parallel computing: numerics, applications, and trends*, edited by R. Trobec, M. Vajteršic, and P. Zinterhof (Springer London, London, 2009), pp. 281–305.

[46] M. Gander and S. Vandewalle, SIAM Journal on Scientific Computing **29**, 556 (2007).

[47] E. J. Bylaska, J. Q. Weare, and J. H. Weare, The Journal of Chemical Physics **139**, 074114 (2013).

[48] D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. B. Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young, in Sc '14: proceedings of the international conference for high performance computing, networking, storage and analysis (Nov. 2014), pp. 41–53.

[49] H. Lin and D. G. Truhlar, Theoretical Chemistry Accounts **117**, 185 (2006).

[50] M. G. Saunders and G. A. Voth, Annual Review of Biophysics **42**, PMID: 23451897, 73 (2013).

[51] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, Chemical Reviews **116**, PMID: 27333362, 7898 (2016).

[52] T. Belytschko, W. Liu, B. Moran, and K. Elkhodary, *Nonlinear finite elements for continua and structures*, No Longer used (Wiley, 2013).

[53] R. Zhang, L. Wen, J. Xiao, and D. Qian, Comput. Mech. **63**, 455 (2019).

[54] A. Chakrabarty, (2007).

[55] K. Kreis, *Advanced adaptive resolution methods for molecular simulation* (Univ., Mainz, 2018).

[56] K. Burke, The Journal of Chemical Physics **136**, 150901 (2012).

[57] L. H. Thomas, Mathematical Proceedings of the Cambridge Philosophical Society **23**, 542 (1927).

[58] E. Fermi, Zeitschrift für Physik **48**, 73 (1928).

[59] W. Kohn and L. J. Sham, Physical Review **140**, 1133 (1965).

[60] W. Kohn, Rev. Mod. Phys. **71**, 1253 (1999).

[61] K. Burke and L. O. Wagner, International Journal of Quantum Chemistry **113**, 96 (2013).

[62] R. O. Jones, Rev. Mod. Phys. **87**, 897 (2015).

[63] F. Gygi and G. Galli, Materials Today **8**, 26 (2005).

[64] L. E. Ratcliff, S. Mohr, G. Huhs, T. Deutsch, M. Masella, and L. Genovese, Wiley Interdisciplinary Reviews: Computational Molecular Science **7**, e1290 (2017).

[65] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, Journal of Computational Chemistry **26**, 1668 (2005).

[66] V. Lindahl, A. Villa, and B. Hess, PLOS Computational Biology **13**, 1 (2017).

[67] G. Collier, N. A. Vellore, J. A. Yancey, S. J. Stuart, and R. A. Latour, Biointerphases **7**, 24 (2012).

[68] A. M. Fluitt and J. J. de Pablo, Biophysical Journal **109**, 1009 (2015).

[69] T. van Mourik, M. Bühl, and M.-P. Gaigeot, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences **372**, rsta.2012.0488[PII], 20120488 (2014).

[70] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).

[71] P. Carloni, U. Rothlisberger, and M. Parrinello, Accounts of Chemical Research **35**, PMID: 12069631, 455 (2002).

[72] D. Marx and J. Hutter, *Ab initio molecular dynamics: basic theory and advanced methods* (Cambridge University Press, 2009).

[73] M. Oka, H. Kamisaka, T. Fukumura, and T. Hasegawa, Phys. Chem. Chem. Phys. **17**, 29057 (2015).

[74] W. Kohn and L. J. Sham, Phys. Rev. (2) **140**, A1133 (1965).

[75] S. M. Rassoulinejad-Mousavi and Y. Zhang, Scientific Reports **8**, 2424 (2018).

[76] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[77] V. L. Deringer and G. Csányi, Phys. Rev. B **95**, 094203 (2017).

[78] S.-H. Li and L. Wang, (2018).

[79] G. P. P. Pun, R. Batra, R. Ramprasad, and Y. Mishin, Nature Communications **10**, 2339 (2019).

[80] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, Journal of Chemical Theory and Computation **9**, PMID: 26584096, 3404 (2013).

[81] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).

[82] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nature Communications **8**, 13890 (2017).

[83] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, Nature Communications **8**, 872 (2017).

[84] M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Mueller, and K. Burke, *Density functionals with quantum chemical accuracy: from machine learning to molecular dynamics*, May 2019.

[85] T. S. Hy, S. Trivedi, H. Pan, B. M. Anderson, and R. Kondor, The Journal of Chemical Physics **148**, 241745 (2018).

[86] R. Kondor, H. T. Son, H. Pan, B. M. Anderson, and S. Trivedi, CoRR **abs/1801.02144** (2018).

[87] J. A. McCammon and S. C. Harvey, *Dynamics of proteins and nucleic acids* (Cambridge University Press, 1987).

[88] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids: second edition* (Oxford University Press, Oxford, 2017), p. 640.

[89] P. E. M. Lopes, O. Guvench, and A. D. MacKerell, "Current status of protein force fields for molecular dynamics simulations", in *Molecular modeling of proteins*, edited by A. Kukol (Springer New York, New York, NY, 2015), pp. 47–71.

[90] B. J. Leimkuhler, S. Reich, and R. D. Skeel, "Integration methods for molecular dynamics", in *Mathematical approaches to biomolecular structure and dynamics*, edited by J. P. Mesirov, K. Schulten, and D. W. Sumners (Springer New York, New York, NY, 1996), pp. 161–185.

[91] A. R. Leach, *Molecular Modelling: Principles and Applications*, 2001.

[92] L. Woodcock, Chemical Physics Letters **10**, 257 (1971).

[93] G. S. Grest and K. Kremer, Phys. Rev. A **33**, 3628 (1986).

[94] W. G. Hoover and B. L. Holian, Physics Letters A **211**, 253 (1996).

[95] S. Nosé, The Journal of Chemical Physics **81**, 511 (1984).

[96] G. J. Martyna, M. L. Klein, and M. Tuckerman, The Journal of Chemical Physics **97**, 2635 (1992).

[97] M. E. Tuckerman and G. J. Martyna, The Journal of Physical Chemistry B **104**, 159 (2000).

[98] M. Tuckerman, *Statistical mechanics: theory and molecular simulation* (Oxford university press, 2010).

[99] F. Martín-García, E. Papaleo, P. Gomez-Puertas, W. Boomsma, and K. Lindorff-Larsen, PloS one **10**, PONE-D-14-40928[PII], e0121114 (2015).

[100] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, and S. J. Marrink, Wiley Interdisciplinary Reviews: Computational Molecular Science **4**, 225 (2014).

[101] M. Levitt and A. Warshel, Nature **253**, 694 (1975).

[102] P. J. Bond, J. Holyoake, A. Ivetac, S. Khalid, and M. S. Sansom, Journal of Structural Biology **157**, Advances in Molecular Dynamics Simulations, 593 (2007).

[103] R. C. Picu and A. Rakshit, The Journal of Chemical Physics **127**, 144909 (2007).

[104] A. Karatrantos, R. J. Composto, K. I. Winey, M. Kröger, and N. Clarke, Polymers **11**, PMC6571671[pmcid], 876 (2019).

[105] J. A. Mc Cammon and M. Karplus, Nature **268**, 765 (1977).

[106] M. KARPLUS and J. A. MCCAMMON, Nature **277**, 578 (1979).

[107] R. L. Melvin, R. C. Godwin, J. Xiao, W. G. Thompson, K. S. Berenhaut, and F. R. Salsbury, Journal of Chemical Theory and Computation **12**, PMID: 27802394, 6130 (2016).

[108] L. Orellana, Frontiers in Molecular Biosciences **6**, 117 (2019).

[109] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, The Journal of Chemical Physics **128**, 244114 (2008).

[110] G. A. Voth, *Coarse-graining of condensed phase and biomolecular systems*, 1st ed. (CRC Press, 2008).

[111] J. T. Padding and W. J. Briels, Journal of Physics: Condensed Matter **23**, 233101 (2011).

[112] W. Humphrey, A. Dalke, and K. Schulten, Journal of Molecular Graphics **14**, 33 (1996).

[113] J. G. Kirkwood, The Journal of Chemical Physics **3**, 300 (1935).

[114] D. Reith, M. Pütz, and F. Müller-Plathe, Journal of Computational Chemistry **24**, 1624 (2003).

[115] S. T. John and G. Csányi, The Journal of Physical Chemistry B **121**, PMID: 29117675, 10934 (2017).

[116] W. G. Noid, "Systematic methods for structurally consistent coarse-grained models", in *Biomolecular simulations: methods and protocols*, edited by L. Monticelli and E. Salonen (Humana Press, Totowa, NJ, 2013), pp. 487–531.

[117] C. Scherer and D. Andrienko, Phys. Chem. Chem. Phys. **20**, 22387 (2018).

[118] S. Izvekov and G. A. Voth, The Journal of Chemical Physics **125**, 151101 (2006).

[119] D. Fritz, K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt, and K. Kremer, Phys. Chem. Chem. Phys. **13**, 10412 (2011).

[120] S. Markutsya and M. H. Lamm, The Journal of Chemical Physics **141**, 174107 (2014).

[121] J. F. Rudzinski, Computation **7**, 42 (2019).

[122] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, The Journal of Chemical Physics **151**, 070902 (2019).

[123] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annual Review of Physical Chemistry **48**, PMID: 9348663, 545 (1997).

[124] A. Laio and M. Parrinello, Proceedings of the National Academy of Sciences **99**, 12562 (2002).

[125] Y. M. Rhee and V. S. Pande, Biophysical Journal **84**, 775 (2003).

[126] C. Bergonzo, N. M. Henriksen, D. R. Roe, J. M. Swails, A. E. Roitberg, and T. E. 3. Cheatham, Journal of chemical theory and computation **10**, PMC3893832[pmcid], 492 (2014).

[127] G. M. Torrie and J. P. Valleau, Journal of Computational Physics **23**, 187 (1977).

[128] C. Liao and J. Zhou, The Journal of Physical Chemistry B **118**, PMID: 24815540, 5843 (2014).

[129] C. Schütte, "Conformational dynamics: modelling, theory, algorithm, and application to biomolecules", PhD thesis (Freie Universität Berlin, Konrad-Zuse-Zentrum für Informationstechnik Berlin, July 1999).

[130] W. E and E. Vanden-Eijnden, Annual Review of Physical Chemistry **61**, PMID: 18999998, 391 (2010).

[131] R. C. Bernardi, M. C. Melo, and K. Schulten, Biochimica et Biophysica Acta (BBA) - General Subjects **1850**, Recent developments of molecular dynamics, 872 (2015).

[132] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).

[133] Y. Sugita and Y. Okamoto, Chemical Physics Letters **314**, 141 (1999).

[134] N. Nakajima, H. Nakamura, and A. Kidera, The Journal of Physical Chemistry B **101**, 817 (1997).

[135] C. D. Christ and W. F. van Gunsteren, The Journal of Chemical Physics **126**, 184110 (2007).

[136] G. Torrie and J. Valleau, Journal of Computational Physics **23**, 187 (1977).

[137] P. Maragakis, A. van der Vaart, and M. Karplus, The Journal of Physical Chemistry B **113**, PMID: 19284746, 4664 (2009).

[138] H. S. Hansen and P. H. Hünenberger, Journal of Computational Chemistry **31**, 1 (2010).

[139] L. Piela, J. Kostrowicki, and H. A. Scheraga, The Journal of Physical Chemistry **93**, 3339 (1989).

[140] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).

[141] E. Darve and A. Pohorille, The Journal of Chemical Physics **115**, 9169 (2001).

[142] O. Valsson, P. Tiwary, and M. Parrinello, Annual Review of Physical Chemistry **67**, PMID: 26980304, 159 (2016).

[143] O. Valsson and M. Parrinello, Phys. Rev. Lett. **113**, 090601 (2014).

[144] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu, PLOS Computational Biology **12**, 1 (2016).

[145] J. W. Wagner, J. F. Dama, A. E. P. Durumeric, and G. A. Voth, The Journal of Chemical Physics **145**, 044108 (2016).

[146] T. D. Potter, J. Tasche, and M. R. Wilson, Phys. Chem. Chem. Phys. **21**, 1912 (2019).

[147] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, The Journal of Chemical Physics **134**, 085101 (2011).

[148] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).

[149] W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, and O. Hahn, Acta Polymerica **49**, 61 (1998).

[150] A. Soper, Chemical Physics **202**, 295 (1996).

[151] A. P. Lyubartsev, M. Karttunen, I. Vattulainen, and A. Laaksonen, Soft Materials **1**, 121 (2002).

[152] G. S. Pawley, R. H. Swendsen, D. J. Wallace, and K. G. Wilson, Phys. Rev. B **29**, 4030 (1984).

[153] R. H. Swendsen, Phys. Rev. Lett. **42**, 859 (1979).

[154] A. P. Lyubartsev, A. Naômé, D. P. Vercauteren, and A. Laaksonen, The Journal of Chemical Physics **143**, 243120 (2015).

[155] E. Ising, Zeitschrift für Physik **31**, 253 (1925).

[156] N. Ashcroft and N. Mermin, *Solid State Physics* (Saunders College, Philadelphia, 1976).

[157] A. Savelyev and G. A. Papoian, Biophysical Journal **96**, 4044 (2009).

[158] P. Koehl, F. Poitevin, R. Navaza, and M. Delarue, Journal of Chemical Theory and Computation **13**, PMID: 28170254, 1424 (2017).

[159] B. E. Clements, C. E. Campbell, P. J. Samsel, and F. J. Pinski, Phys. Rev. A **44**, 1139 (1991).

[160] A. A. Louis, P. G. Bolhuis, J. P. Hansen, and E. J. Meijer, Phys. Rev. Lett. **85**, 2522 (2000).

[161] S. Jain, S. Garde, and S. K. Kumar, Industrial & Engineering Chemistry Research **45**, 5614 (2006).

[162] C. Scherer and D. Andrienko, The European Physical Journal Special Topics **225**, 1441 (2016).

[163] V. Agrawal, G. Arya, and J. Oswald, Macromolecules **47**, 3378 (2014).

[164] T. C. Moore, C. R. Iacovella, and C. McCabe, The Journal of Chemical Physics **140**, 224104 (2014).

[165] J. G. Kirkwood and F. P. Buff, The Journal of Chemical Physics **19**, 774 (1951).

[166] P. Krüger, S. K. Schnell, D. Bedeaux, S. Kjelstrup, T. J. H. Vlugt, and J.-M. Simon, The Journal of Physical Chemistry Letters **4**, PMID: 26283427, 235 (2013).

[167] D. Rosenberger, M. Hanke, and N. F. van der Vegt, The European Physical Journal Special Topics **225**, 1323 (2016).

[168] Z. Xie, D. Chai, Y. Wang, and H. Tan, The Journal of Physical Chemistry B **120**, PMID: 27766876, 11834 (2016).

[169] T. Murtola, E. Falck, M. Karttunen, and I. Vattulainen, The Journal of Chemical Physics **126**, 075101 (2007).

[170] R. Henderson, Physics Letters A **49**, 197 (1974).

[171] C.-C. Fu, P. M. Kulkarni, M. Scott Shell, and L. Gary Leal, The Journal of Chemical Physics **137**, 164106 (2012).

[172] R. Evans, Molecular Simulation **4**, 409 (1990).

[173] L. Larini, L. Lu, and G. A. Voth, The Journal of Chemical Physics **132**, 164107 (2010).

[174] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, Journal of Chemical Theory and Computation **5**, PMID: 26602505, 3211 (2009).

[175] F. Frommer, M. Hanke, and S. Jansen, Journal of Mathematical Physics **60**, 093303 (2019).

[176] S. Izvekov and G. A. Voth, The Journal of Physical Chemistry B **109**, PMID: 16851243, 2469 (2005).

[177] S. Izvekov and G. A. Voth, The Journal of Chemical Physics **123**, 134105 (2005).

[178] M. S. Shell, The Journal of Chemical Physics **129**, 144108 (2008).

[179] A. Chaimovich and M. S. Shell, Phys. Chem. Chem. Phys. **11**, 1901 (2009).

[180] A. Chaimovich and M. S. Shell, The Journal of Chemical Physics **134**, 094112 (2011).

[181] M. S. Shell, "Coarse-graining with the relative entropy", in *Advances in chemical physics* (John Wiley & Sons, Ltd, 2016), pp. 395–441.

[182] F. Ercolessi and J. B. Adams, Europhysics Letters (EPL) **26**, 583 (1994).

[183] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, The Journal of Chemical Physics **120**, 10896 (2004).

[184] W. G. Noid, P. Liu, Y. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, The Journal of Chemical Physics **128**, 244115 (2008).

[185] A. Das and H. C. Andersen, The Journal of Chemical Physics **136**, 194113 (2012).

[186] J. W. Mullinax and W. G. Noid, The Journal of Chemical Physics **131**, 104110 (2009).

[187] H. M. Cho and J.-W. Chu, The Journal of Chemical Physics **131**, 134107 (2009).

[188] L. Lu, J. F. Dama, and G. A. Voth, The Journal of Chemical Physics **139**, 121906 (2013).

[189] E. Kalligiannaki, V. Harmandaris, M. A. Katsoulakis, and P. Plecháč, The Journal of Chemical Physics **143**, 084105 (2015).

[190] F. Sittel and G. Stock, The Journal of Chemical Physics **149**, 150901 (2018).

[191] A. L. Gibbs and F. E. Su, International Statistical Review / Revue Internationale de Statistique **70**, 419 (2002).

[192] M. D. Reid and R. C. Williamson, Journal of Machine Learning Research **12**, 731 (2011).

[193] T. van Erven and P. Harremos, IEEE Transactions on Information Theory **60**, 3797 (2014).

[194] I. Bilionis and N. Zabaras, The Journal of Chemical Physics **138**, 044313 (2013).

[195] M. P. Hodges, A. J. Stone, and S. S. Xantheas, The Journal of Physical Chemistry A **101**, 9163 (1997).

[196] A. Grosberg, *Statistical physics of macromolecules* (AIP Press, New York, 1994).

[197] D. Feldman, Journal of Polymer Science Part C: Polymer Letters **27**, 239 (1989).

[198] H. S. Ashbaugh and L. R. Pratt, Rev. Mod. Phys. **78**, 159 (2006).

[199] Y. Han, J. F. Dama, and G. A. Voth, The Journal of Chemical Physics **149**, 044104 (2018).

[200] J. F. Rudzinski and W. G. Noid, The Journal of Chemical Physics **135**, 214101 (2011).

[201] J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth, Journal of Chemical Theory and Computation **9**, PMID: 26583735, 2466 (2013).

[202] A. Davtyan, J. F. Dama, A. V. Sinitskiy, and G. A. Voth, Journal of Chemical Theory and Computation **10**, PMID: 26583210, 5265 (2014).

[203] J. F. Dama, J. Jin, and G. A. Voth, Journal of Chemical Theory and Computation **13**, PMID: 28112956, 1010 (2017).

[204] T. Lelièvre, M. Rousset, and G. Stoltz, *Free energy computations* (IMPERIAL COLLEGE PRESS, 2010).

[205] I. Bilionis and P. Koutsourelakis, Journal of Computational Physics **231**, 3849 (2012).

[206] M. A. Katsoulakis, A. J. Majda, and D. G. Vlachos, Journal of Computational Physics **186**, 250 (2003).

[207] A. Chatterjee, D. G. Vlachos, and M. A. Katsoulakis, The Journal of Chemical Physics **121**, 11420 (2004).

[208] M. A. Katsoulakis, P. Plecháč, and A. Sopasakis, SIAM Journal on Numerical Analysis **44**, 2270 (2006).

[209] M. A. Katsoulakis, P. Plecháč, and L. Rey-Bellet, Journal of Scientific Computing **37**, 43 (2008).

[210] M. A. Katsoulakis and P. Plecháč, The Journal of Chemical Physics **139**, 074115 (2013).

[211] E. Kalligiannaki, M. A. Katsoulakis, P. Plecháč, and D. G. Vlachos, Journal of Computational Physics **231**, 2599 (2012).

[212] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, Faraday Discuss. **144**, 301 (2010).

[213] P. Español, M. Serrano, I. Pagonabarraga, and I. Zúñiga, Soft Matter **12**, 4821 (2016).

[214] A. Davtyan, G. A. Voth, and H. C. Andersen, The Journal of Chemical Physics **145**, 224107 (2016).

[215] V. Harmandaris, E. Kalligiannaki, M. Katsoulakis, and P. Plecháč, Journal of Computational Physics **314**, 355 (2016).

[216] E. Kalligiannaki, M. Katsoulakis, P. Plechac, and V. Harmandaris, Procedia Computer Science **136**, 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece, 331 (2018).

[217] G. Baxevani, E. Kalligiannaki, and V. Harmandaris, Procedia Computer Science **156**, 8th International Young Scientists Conference on Computational Science, YSC2019, 24-28 June 2019, Heraklion, Greece, 59 (2019).

[218] S. Kaltenbach and P.-S. Koutsourelakis, *Incorporating physical constraints in a deep probabilistic machine learning framework for coarse-graining dynamical systems*, 2019.

[219] P. Español and I. Zúñiga, Phys. Chem. Chem. Phys. **13**, 10538 (2011).

[220] J. F. Rudzinski, Computation **7** (2019).

[221] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).

[222] M. A. Raihan, N. Goli, and T. M. Aamodt, in 2019 ieee international symposium on performance analysis of systems and software (ispass) (Mar. 2019), pp. 79–92.

[223] L. Zhang, J. Han, H. Wang, R. Car, and W. E, The Journal of Chemical Physics **149**, 034101 (2018).

[224] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé, and C. Clementi, ACS Central Science **5**, 755 (2019).

[225] S. Russell and P. Norvig, *Artificial intelligence: a modern approach*, 3rd (Prentice Hall Press, Upper Saddle River, NJ, USA, 2009).

[226] P. Auer, H. Burgsteiner, and W. Maass, Neural Networks **21**, 786 (2008).

[227] M. Ceriotti, The Journal of Chemical Physics **150**, 150901 (2019).

[228] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, in *Advances in neural information processing systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014), pp. 2672–2680.

[229] S. Mohamed and B. Lakshminarayanan, *Learning in implicit generative models*, 2016.

[230] A. E. P. Durumeric and G. A. Voth, The Journal of Chemical Physics **151**, 124110 (2019).

[231] M. S. Shell, The Journal of Chemical Physics **129**, 144108 (2008).

[232] F. H. Stillinger, The Journal of Physical Chemistry **74**, 3677 (1970).

[233] L. Vlcek and A. A. Chialvo, The Journal of Chemical Physics **143**, 144110 (2015).

[234] W. Wang and R. Gómez-Bombarelli, npj Computational Materials **5**, 125 (2019).

[235] D. P. Kingma and M. Welling, Foundations and Trends® in Machine Learning **12**, 307 (2019).

[236] E. Jang, S. Gu, and B. Poole, in 5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings (2017).

[237] A. Rajeswaran, C. Finn, S. Kakade, and S. Levine, *Meta-learning with implicit gradients*, 2019.

[238] L. Dinh, D. Krueger, and Y. Bengio, in 3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, workshop track proceedings (2015).

[239] D. P. Kingma and P. Dhariwal, in *Advances in neural information processing systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), pp. 10215–10224.

[240] F. Noé, S. Olsson, J. Köhler, and H. Wu, Science **365** (2019).

[241] F. Noé, G. D. Fabritiis, and C. Clementi, *Machine learning for protein folding and dynamics*, 2019.

[242] Y. Zhu, N. Zabaras, P.-S. Koutsourelakis, and P. Perdikaris, (2019).

[243] N. Geneva and N. Zabaras, (2019).

[244] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Science Advances **3** (2017).

[245] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, Nature Communications **9**, 3887 (2018).

[246] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, Computer Physics Communications **240**, 38 (2019).

[247] K. Farrell, J. T. Oden, and D. Faghihi, Journal of Computational Physics **295**, 189 (2015).

[248] I. J. Myung and M. A. Pitt, Psychonomic Bulletin & Review **4**, 79 (1997).

[249] J. Zavadlav, G. Arampatzis, and P. Koumoutsakos, Scientific Reports **9**, 99 (2019).

[250] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning (adaptive computation and machine learning)* (The MIT Press, 2005).

[251] T. T. Foley, M. S. Shell, and W. G. Noid, The Journal of Chemical Physics **143**, 243104 (2015).

[252] V. A. Harmandaris, D. Reith, N. F. A. van der Vegt, and K. Kremer, Macromolecular Chemistry and Physics **208**, 2109 (2007).

[253] M. Dallavalle and N. F. A. van der Vegt, Phys. Chem. Chem. Phys. **19**, 23034 (2017).

[254] Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfaendtner, and G. A. Voth, Biophysical Journal **95**, 5073 (2008).

[255] I. T. Jolliffe, "Principal component analysis and factor analysis", in *Principal component analysis* (Springer New York, New York, NY, 1986), pp. 115–128.

[256] J. R. López-Blanco and P. Chacón, Current Opinion in Structural Biology **37**, Theory and simulation • Macromolcular machines, 46 (2016).

[257] L. Yang, G. Song, and R. L. Jernigan, Proceedings of the National Academy of Sciences **106**, 12347 (2009).

[258] M. Li, J. Z. Zhang, and F. Xia, Journal of Chemical Theory and Computation **12**, PMID: 26930392, 2091 (2016).

[259] M. Chakraborty, C. Xu, and A. D. White, The Journal of Chemical Physics **149**, 134106 (2018).

[260] A. Mowshowitz and M. Dehmer, Entropy **14**, 559 (2012).

[261] L. Molgedey and H. G. Schuster, Phys. Rev. Lett. **72**, 3634 (1994).

[262] B. Schölkopf and A. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, Adaptive Computation and Machine Learning, Parts of this book, including an introduction to kernel methods, can be downloaded here. (MIT Press, Cambridge, MA, USA, Dec. 2002), p. 644.

[263] M. A. Rohrdanz, W. Zheng, and C. Clementi, Annual Review of Physical Chemistry **64**, PMID: 23298245, 295 (2013).

[264] M. A. Webb, J.-Y. Delannoy, and J. J. de Pablo, Journal of Chemical Theory and Computation **15**, 1199 (2019).

[265] F. R. K. Chung, *Spectral graph theory* (American Mathematical Society, 1997).

[266] K. H. Kanekal and T. Bereau, The Journal of Chemical Physics **151**, 164106 (2019).

[267] J. Lin, IEEE Transactions on Information theory **37**, 145 (1991).

[268] V. Bruni and D. Vitulano, in Advanced concepts for intelligent vision systems, edited by J. Blanc-Talon, C. Distante, W. Philips, D. Popescu, and P. Scheunders (2016), pp. 311–323.

[269] S. J. Marrink, A. H. de Vries, and A. E. Mark, The Journal of Physical Chemistry B **108**, 750 (2004).

[270] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, The Journal of Physical Chemistry B **111**, PMID: 17569554, 7812 (2007).

[271] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, Journal of chemical theory and computation **4 5**, 819 (2008).

[272] P. Liu, Q. Shi, H. Daumé, and G. A. Voth, The Journal of Chemical Physics **129**, 214114 (2008).

[273] A. V. Sinitskiy, M. G. Saunders, and G. A. Voth, The Journal of Physical Chemistry B **116**, PMID: 22276676, 8363 (2012).

[274] J. F. Rudzinski and W. G. Noid, The Journal of Physical Chemistry B **118**, PMID: 24684663, 8295 (2014).

[275] J. D. Lee, J. Li, Z. Zhang, and L. Wang, "Sequential and concurrent multi-scale modeling of multiphysics: from atoms to continuum", in *Micromechanics and nanomechanics of composite solids*, edited by S. A. Meguid and G. J. Weng (Springer International Publishing, Cham, 2018), pp. 1–38.

[276] R. K. Kalia, A. Nakano, P. Vashishta, C. L. Rountree, L. Van Brutzel, and S. Ogata, International Journal of Fracture **121**, 71 (2003).

[277] H. Talebi, M. Silani, and T. Rabczuk, Advances in Engineering Software **80**, Civil-Comp, 82 (2015).

[278] M. Praprotnik, L. Delle Site, and K. Kremer, The Journal of Chemical Physics **123**, 224106 (2005).

[279] M. Praprotnik, L. Delle Site, and K. Kremer, Phys. Rev. E **73**, 066701 (2006).

[280] M. Praprotnik, L. D. Site, and K. Kremer, Annual Review of Physical Chemistry **59**, PMID: 18062769, 545 (2008).

[281] C. Krekeler, A. Agarwal, C. Junghans, M. Praprotnik, and L. Delle Site, The Journal of Chemical Physics **149**, 024104 (2018).

[282] J. Zavadlav, S. J. Marrink, and M. Praprotnik, Journal of Chemical Theory and Computation **12**, PMID: 27409519, 4138 (2016).

[283] J. Zavadlav and M. Praprotnik, The Journal of Chemical Physics **147**, 114110 (2017).

[284] M. E. Johnson, T. Head-Gordon, and A. A. Louis, The Journal of Chemical Physics **126**, 144509 (2007).

[285] R. Baron, A. H. de Vries, P. H. Hünenberger, and W. F. van Gunsteren, The Journal of Physical Chemistry B **110**, PMID: 16623533, 8464 (2006).

[286] N. Singh and W. Li, International Journal of Molecular Sciences **20** (2019).

[287] N. J. H. Dunn and W. G. Noid, The Journal of Chemical Physics **143**, 243148 (2015).

[288] C. Caccamo and G. Pellicane, The Journal of Chemical Physics **117**, 5072 (2002).

[289] A. Das and H. C. Andersen, The Journal of Chemical Physics **132**, 164106 (2010).

[290] J. Hu, T. Chen, M. Wang, H. S. Chan, and Z. Zhang, Phys. Chem. Chem. Phys. **19**, 13629 (2017).

[291] W. Chen, A. R. Tan, and A. L. Ferguson, The Journal of Chemical Physics **149**, 072312 (2018).

[292] J. M. A. Grime, J. F. Dama, B. K. Ganser-Pornillos, C. L. Woodward, G. J. Jensen, M. Yeager, and G. A. Voth, Nature Communications **7**, 11568 (2016).

[293] A. Kendall and Y. Gal, in *Advances in neural information processing systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), pp. 5574–5584.

[294] N. Geneva and N. Zabaras, (2018).

[295] P. Koutsourelakis, N. Zabaras, and M. Girolami, Journal of Computational Physics **321**, 1252 (2016).

[296] M. I. Jordan, Statist. Sci. **19**, 140 (2004).

[297] Z. Ghahramani, Nature **521**, 452 (2015).

[298] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, Second, Texts in Statistical Science Series (Chapman & Hall/CRC, Boca Raton, FL, 2004), pp. xxvi+668.

[299] V. K. Mansinghka, T. D. Kulkarni, Y. N. Perov, and J. Tenenbaum, in *Advances in neural information processing systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., 2013), pp. 1520–1528.

[300] M.-H. Chen and Q.-M. Shao, Journal of Computational and Graphical Statistics **8**, 69 (1999).

[301] H. N. Najm, Annual Review of Fluid Mechanics **41**, 35 (2009).

[302] C. Li and H. Chen, in 2014 ieee symposium on computational intelligence in big data (cibd) (Dec. 2014), pp. 1–7.

[303] X. Yang, W. Pan, and Y. Guo, PLOS ONE **12**, 1 (2017).

[304] W. Janke, *Rugged free energy landscapes: common computational approaches to spin glasses, structural glasses and biological macromolecules*, en (Springer, Dec. 2007).

[305] M. Raissi, P. Perdikaris, and G. Karniadakis, Journal of Computational Physics **378**, 686 (2019).

[306] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, The Journal of Chemical Physics **21**, 1087 (1953).

[307] E. Paquet and H. L. Viktor, en, Biomed Res. Int. **2015**, 183918 (2015).

[308] E. Cancès, F. Legoll, and G. Stoltz, ESAIM: Mathematical Modelling and Numerical Analysis **41**, 351 (2007).

[309] L. Boltzmann, *Vorlesungen über gastheorie*, (J.A. Barth, Leipzig, 1896).

[310] R. D. Skeel, SIAM journal on scientific computing : a publication of the Society for Industrial and Applied Mathematics **31**, PMC2800798[pmcid], 1363 (2009).

[311] L. Verlet, Phys. Rev. **159**, 98 (1967).

[312] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, The Journal of Chemical Physics **76**, 637 (1982).

[313] M. P. Allen and F. Schmid, Molecular Simulation **33**, 21 (2007).

[314] X. Yong and L. T. Zhang, The Journal of Chemical Physics **138**, 084503 (2013).

[315] S. Plimpton, Journal of Computational Physics **117**, 1 (1995).

[316] H. Berendsen, D. van der Spoel, and R. van Drunen, Computer Physics Communications **91**, 43 (1995).

[317] E. Lindahl, B. Hess, and D. van der Spoel, Molecular modeling annual **7**, 306 (2001).

[318] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, Journal of Computational Chemistry **26**, 1701 (2005).

[319] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, Journal of Chemical Theory and Computation **4**, 435 (2008).

[320] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, Bioinformatics **29**, 845 (2013).

[321] S. Páll, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl, in Solving software challenges for exascale, edited by S. Markidis and E. Laure (2015), pp. 3–27.

[322] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1-2**, 19 (2015).

[323] B. J. Leimkuhler, S. Reich, and R. D. Skeel, "Integration methods for molecular dynamics", in *Mathematical approaches to biomolecular structure and dynamics*, edited by J. P. Mesirov, K. Schulten, and D. W. Sumners (Springer New York, New York, NY, 1996), pp. 161–185.

[324] N. Bou-Rabee, Entropy **16**, 138 (2014).

[325] F. Niederhöfer and J. Wackerfuß, PAMM **12**, 47 (2012).

[326] A. Z. Panagiotopoulos, "Gibbs ensemble techniques", in *Observation, prediction and simulation of phase transitions in complex fluids*, edited by M. Baus, L. F. Rull, and J.-P. Ryckaert (Springer Netherlands, Dordrecht, 1995), pp. 463–501.

[327] J. Wereszczynski and J. A. McCammon, Quarterly reviews of biophysics **45**, S0033583511000096[PII], 1 (2012).

[328] J. Weiner, *Statistical mechanics of elasticity; 2nd ed.* Dover books on physics (Dover Publications, New York, NY, 2012).

[329] D. C. Rapaport, *The art of molecular dynamics simulation*, 2nd (Cambridge University Press, USA, 2004).

[330] D. P. Landau and K. Binder, *A guide to monte carlo simulations in statistical physics*, en (Cambridge University Press, Nov. 2014).

[331] H.-X. Zhou and M. K. Gilson, Chemical reviews **109**, PMC3329805[pmcid], 4092 (2009).

[332] S. A. Adcock and J. A. McCammon, Chemical reviews **106**, PMC2547409[pmcid], 1589 (2006).

[333] H. Schreiber and O. Steinhauser, Biochemistry **31**, PMID: 1610828, 5856 (1992).

[334] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, The Journal of Physical Chemistry **91**, 6269 (1987).

[335] P. G. Kusalik and I. M. Svishchev, Science **265**, 1219 (1994).

[336] P. Mark and L. Nilsson, The Journal of Physical Chemistry A **105**, 9954 (2001).

[337] E. J. Sorin and V. S. Pande, Biophys J **88**, 2472 (2005).

[338] A. J. DePaul, E. J. Thompson, S. S. Patel, K. Haldeman, and E. J. Sorin, Nucleic Acids Res **38**, 4856 (2010).

[339] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids* (Clarendon Press, New York, NY, USA, 1989).

[340] M. S. Shell, R. Ritterson, and K. A. Dill, The journal of physical chemistry. B **112**, PMC2699260[pmcid], 6878 (2008).

[341] P. Español, M. Serrano, I. Pagonabarraga, and I. Zúñiga, Soft Matter **12**, 4821 (2016).

[342] L. Lu and G. A. Voth, The Journal of Chemical Physics **134**, 224107 (2011).

[343] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[344] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).

[345] S. Kullback and R. A. Leibler, The Annals of Mathematical Statistics **22**, 79 (1951).

[346] S. Kullback, *Information theory and statistics* (Wiley, New York, 1959).

[347] T. Bayes, Phil. Trans. of the Royal Soc. of London **53**, 370 (1763).

[348] E. T. Jaynes, The Mathematical Intelligencer **27**, 83 (2005).

[349] M. Dashti and A. M. Stuart, "The bayesian approach to inverse problems", in *Handbook of uncertainty quantification*, edited by R. Ghanem, D. Higdon, and H. Owhadi (Springer International Publishing, Cham, 2017), pp. 311–428.

[350] R. T. Cox, Am. J. Phys. **14**, 1 (1946).

[351] J. Y. Halpern, J. Artif. Int. Res. **10**, 67 (1999).

[352] E.-J. Wagenmakers, M. Lee, T. Lodewyckx, and G. J. Iverson, "Bayesian versus frequentist inference", in *Bayesian evaluation of informative hypotheses*, edited by H. Hoijtink, I. Klugkist, and P. A. Boelen (Springer New York, New York, NY, 2008), pp. 181–207.

[353] F. J. Samaniego, *A comparison of the bayesian and frequentist approaches to estimation*, en (Springer Science & Business Media, June 2010).

[354] A. Caticha, Brazilian Chapter of the International Society for Bayesian Analysis-ISBrA, Sao Paulo, Brazil (2012).

[355] G. Larry Bretthorst, *Bayesian spectrum analysis and parameter estimation*, en (Springer Science & Business Media, Mar. 2013).

[356] C. M. Bishop, *Pattern recognition and machine learning (information science and statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).

[357] M. J. Beal, "Variational algorithms for approximate bayesian inference", PhD thesis (Gatsby Computational Neuroscience Unit, University College London, 2003).

[358] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 2008 (2019).

[359] D. V. Lindley, O. Barndorff-Nielsen, G. Elfving, E. Harsaae, D. Thorburn, A. Hald, and E. Spjötvoll, Scandinavian Journal of Statistics **5**, 1 (1978).

[360] A. Gelman, Bayesian Anal. **3**, 445 (2008).

[361] A. Gelman, en, Bayesian Anal. **1**, 515 (2006).

[362] I. Leenen, I. Van Mechelen, A. Gelman, and S. De Knop, en, Psychometrika **73**, 39 (2008).

[363] D. J. C. MacKay, "Bayesian methods for adaptive models", PhD thesis (California Institute of Technology, 1992).

[364] D. J. C. MacKay, in *Maximum entropy and bayesian methods: santa barbara, california, U.S.A., 1993*, edited by G. R. Heidbreder (Springer Netherlands, Dordrecht, 1996), pp. 43–59.

[365] A. Gelman, Statistical Science **24**, 176 (2009).

[366] J. E. Griffin and P. J. Brown, Bayesian Anal. **8**, 691 (2013).

[367] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, Statist. Sci. **14**, 382 (1999).

[368] R. E. Kass and A. E. Raftery, Journal of the American Statistical Association **90**, 773 (1995).

[369] R. M. Neal, *Bayesian learning for neural networks*, en (Springer Science & Business Media, Dec. 2012).

[370] W. von der Linden, V. Dose, and U. von Toussaint (Cambridge University Press, 2014), p. 649.

[371] M. W. Berry, A. Mohamed, and B. W. Yap, eds., *Supervised and unsupervised learning for data science* (Springer, Cham, 2020).

[372] M. I. Jordan, ed., *Learning in graphical models* (MIT Press, Cambridge, MA, USA, 1999).

[373] A. Y. Ng and M. I. Jordan, in *Advances in neural information processing systems 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, 2002), pp. 841–848.

[374] S. R. Hare, L. A. Bratholm, D. R. Glowacki, and B. K. Carpenter, Chem. Sci. **10**, 9954 (2019).

[375] M. Schöberl, N. Zabaras, and P.-S. Koutsourelakis, Journal of Computational Physics **333**, 49 (2017).

[376] A. Cichocki and S.-i. Amari, Entropy **12**, 1532 (2010).

[377] S.-H. Cha, in, Vol. 1 (2007).

[378] J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernandez-Lobato, and R. Turner, in Proceedings of the 33rd international conference on machine learning, Vol. 48, edited by M. F. Balcan and K. Q. Weinberger, Proceedings of Machine Learning Research (20–22 Jun 2016), pp. 1511–1520.

[379] , The Mathematical Intelligencer **27**, 83 (2005).

[380] R. J. Rossi, "Likelihood-based estimation", in *Mathematical statistics* (John Wiley & Sons, Ltd, 2018) Chap. 5, pp. 223–279.

[381] R. Bassett and J. Deride, Mathematical Programming **174**, 129 (2019).

[382] M. J. Wainwright and M. I. Jordan, Foundations and Trends® in Machine Learning **1**, 1 (2008).

[383] J. Besag, P. Green, D. Higdon, and K. Mengersen, Statist. Sci. **10**, 3 (1995).

[384] V. Plagnol and S. Tavaré, in Monte carlo and quasi-monte carlo methods 2002, edited by H. Niederreiter (2004), pp. 99–113.

[385] J. L. W. V. Jensen, Acta Math. **30**, 175 (1906).

[386] L. E. Baum, T. Petrie, G. Soules, and N. R. Weiss, in (1970).

[387] A. P. Dempster, N. M. Laird, and D. B. Rubin, JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B **39**, 1 (1977).

[388] Z. Ghahramani and M. I. Jordan, in *Advances in neural information processing systems 8*, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (MIT Press, 1996), pp. 472–478.

[389] R. M. Neal, Artificial Intelligence **56**, 71 (1992).

[390] G. E. Hinton and R. S. Zemel, in *Advances in neural information processing systems 6*, edited by J. D. Cowan, G. Tesauro, and J. Alspector (Morgan-Kaufmann, 1994), pp. 3–10.

[391] Z. Ghahramani and G. E. Hinton, Neural Comput. **12**, 831 (2000).

[392] L. Younes, Stochastics and Stochastic Reports **65**, 177 (1999).

[393] C. Andrieu, É. Moulines, and P. Priouret, SIAM Journal on Control and Optimization **44**, 283 (2005).

[394] G. Fort, E. Moulines, A. Schreck, and M. Vihola, SIAM Journal on Control and Optimization **54**, 866 (2016).

[395] H. Kahn and T. E. Harris, National Bureau of Standards applied mathematics series **12**, 27 (1951).

[396] P. Moral, A. Doucet, and A. Jasra, Statistics and Computing **22**, 1009 (2012).

[397] F. Doshi, K. Miller, J. V. Gael, and Y. W. Teh, in Proceedings of the twelth international conference on artificial intelligence and statistics, Vol. 5, edited by D. van Dyk and M. Welling, Proceedings of Machine Learning Research (16–18 Apr 2009), pp. 137–144.

[398] J. Sung, Z. Ghahramani, and S. Bang, IEEE Transactions on Pattern Analysis and Machine Intelligence **30**, 2236 (2008).

[399] R. M. Neal and G. E. Hinton, in *Learning in graphical models*, edited by M. I. Jordan (Springer Netherlands, Dordrecht, 1998), pp. 355–368.

[400] X.-L. Meng and D. B. Rubin, Biometrika **80**, 267 (1993).

[401] D. M. Blei, A. Y. Ng, and M. I. Jordan, J. Mach. Learn. Res. **3**, 993 (2003).

[402] M. J. Johnson and A. S. Willsky, in Icml, Vol. 32, JMLR Workshop and Conference Proceedings (2014), pp. 1854–1862.

[403] T. S. Jaakkola and M. I. Jordan, "Variational methods for inference and estimation in graphical models", AAI0598367, PhD thesis (USA, 1997).

[404] H. Attias, Adv. Neural Inf. Process. Syst. **12** (2000).

[405] Z. Ghahramani and M. J. Beal, in *Advances in neural information processing systems 13*, edited by T. K. Leen, T. G. Dietterich, and V. Tresp (MIT Press, 2001), pp. 507–513.

[406] R. Ranganath, S. Gerrish, and D. Blei, in Proceedings of the seventeenth international conference on artificial intelligence and statistics, Vol. 33, edited by S. Kaski and J. Corander, Proceedings of Machine Learning Research (22–25 Apr 2014), pp. 814–822.

[407] D. P. Kingma and M. Welling, in 2nd international conference on learning representations, ICLR 2014, banff, ab, canada, april 14-16, 2014, conference track proceedings (2014).

[408] D. J. Rezende, S. Mohamed, and D. Wierstra, in Proceedings of the 31st international conference on machine learning, Vol. 32, edited by E. P. Xing and T. Jebara, Proceedings of Machine Learning Research 2 (22–24 Jun 2014), pp. 1278–1286.

[409] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, J. Mach. Learn. Res. **14**, 1303 (2013).

[410] H. Robbins and S. Monro, Ann. Math. Statist. **22**, 400 (1951).

[411] S. J. W. Jorge Nocedal, *Numerical optimization*, edited by J. Nocedal and S. J. Wright (Springer, New York, NY, 1999).

[412] J. C. Spall, *Introduction to stochastic search and optimization*, 1st ed. (John Wiley & Sons, Inc., USA, 2003).

[413] C. Peterson and J. R. Anderson, Complex Systems **1**, 995 (1987).

[414] G. Parisi and J. Machta, American Journal of Physics **57**, 286 (1989).

[415] M. Schöberl, N. Zabaras, and P.-S. Koutsourelakis, The Journal of Chemical Physics **150**, 024109 (2019).

[416] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, Mach. Learn. **37**, 183 (1999).

[417] P.-A. Mattei and J. Frellsen, in *Advances in neural information processing systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), pp. 3859–3870.

[418] L. L. Cam, International Statistical Review / Revue Internationale de Statistique **58**, 153 (1990).

[419] M. West, in Bayesian statistics (2003), pp. 723–732.

[420] M. A. T. Figueiredo, IEEE Trans. Pattern Anal. Mach. Intell. **25**, 1150 (2003).

[421] D. J. C. MacKay and R. M. Neal, *Automatic relevance determination for neural networks*, tech. rep. (University of Cambridge, 1994).

[422] M. E. Tipping, J. Mach. Learn. Res. **1**, 211 (2001).

[423] D. J. C. MacKay, *Information theory, inference, and learning algorithms* (Cambridge University Press, 2003).

[424] H. Ritter, A. Botev, and D. Barber, in International conference on learning representations (2018).

[425] M. A. T. Figueiredo, in NIPS (2001), pp. 697–704.

[426] G. E. Hinton, S. Osindero, and Y.-W. Teh, en, Neural Comput. **18**, 1527 (2006).

[427] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 EP (2015).

[428] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, in , edited by D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group (MIT Press, Cambridge, MA, USA, 1986) Chap. Learning Internal Representations by Error Propagation, pp. 318–362.

[429] C. Van Der Malsburg, in Brain theory, edited by G. Palm and A. Aertsen (1986), pp. 245–248.

[430] S. Haykin, *Neural networks: a comprehensive foundation*, 2nd (Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998).

[431] A. M. D'Amato, A. J. Ridley, and D. S. Bernstein, Statistical Analysis and Data Mining **4**, 446 (2010).

[432] S. Chowdhury, A. Mehmani, W. Tong, and A. Messac, in (2016).

[433] I. Kobyzev, S. Prince, and M. A. Brubaker, (2019).

[434] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, (2019).

[435] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, in *Advances in neural information processing systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 4743–4751.

[436] D. Rezende and S. Mohamed, in Proceedings of the 32nd international conference on machine learning, Vol. 37, edited by F. Bach and D. Blei, Proceedings of Machine Learning Research (July 2015), pp. 1530–1538.

[437] L. Dinh, J. Sohl-Dickstein, and S. Bengio, in 5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings (2017).

[438] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, in *Advances in neural information processing systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014), pp. 2672–2680.

[439] Y. Zhu and N. Zabaras, Journal of Computational Physics **366**, 415 (2018).

[440] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, IEEE Signal Processing Magazine **35**, 53 (2018).

[441] L. Gonog and Y. Zhou, in 2019 14th ieee conference on industrial electronics and applications (iciea) (June 2019), pp. 505–510.

[442] J. Liu, *Monte carlo strategies in scientific computing*, edited by J. Liu (Springer Verlag, New York, Berlin, Heidelberg, 2008), p. 344.

[443] D. P. Landau and K. Binder, *A guide to monte carlo simulations in statistical physics*, en (Cambridge University Press, Nov. 2014).

[444] G. C. Tiao, G. C. Reinsel, D. Xu, J. H. Pedrick, X. Zhu, A. J. Miller, J. J. DeLuisi, C. L. Mateer, and D. J. Wuebbles, Journal of Geophysical Research: Atmospheres **95**, 20507 (1990).

[445] A. W. Marshall, (1953).

[446] M. B. Priestley, *Spectral analysis and time series*, English (Academic Press London ; New York, 1981), 2 v. (xvii, [45], 890 p.) :

[447] D. P. Landau and K. Binder, *A guide to monte carlo simulations in statistical physics* (Cambridge University Press, Nov. 2014).

[448] W. K. Hastings, Biometrika **57**, 97 (1970).

[449] S. Geman and D. Geman, IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-6**, 721 (1984).

[450] U. Grenander and M. I. Miller, Journal of the Royal Statistical Society. Series B (Methodological) **56**, 549 (1994).

[451] G. O. Roberts and R. L. Tweedie, Bernoulli **2**, 341 (1996).

[452] D. S. Lemons and A. Gythiel, American Journal of Physics **65**, 1079 (1997).

[453] G. Storvik, in (2011).

[454] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth, Physics Letters B **195**, 216 (1987).

[455] R. M. Neal, Handbook of Markov Chain Monte Carlo **54**, 113 (2010).

[456] M. Betancourt, *A conceptual introduction to hamiltonian monte carlo*, arxiv:1701.02434, 2017.

[457] S. Lan, J. Streets, and B. Shahbaba, Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence **2014**, PMC4386289[pmcid], 1953 (2014).

[458] E. Marinari and G. Parisi, Europhysics Letters (EPL) **19**, 451 (1992).

[459] Y. Li, M. Mascagni, and A. Gorin, Parallel Comput. **35**, 269 (2009).

[460] G. I. Valderrama-Bahamóndez and H. Fröhlich, Frontiers in Applied Mathematics and Statistics **5**, 55 (2019).

[461] M. Girolami and B. Calderhead, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**, 123 (2011).

[462] A. Doucet, N. de Freitas, and N. Gordon, "An introduction to sequential monte carlo methods", in *Sequential monte carlo methods in practice*, edited by A. Doucet, N. de Freitas, and N. Gordon (Springer New York, New York, NY, 2001), pp. 3–14.

[463] P. Del Moral, A. Doucet, and A. Jasra, J. R. Stat. Soc. Series B Stat. Methodol. **68**, 411 (2006).

[464] P. Del Moral, *Feynman-Kac formulae: genealogical and interacting particle systems with applications* (Springer New York, Dec. 2011).

[465] Swendsen and Wang, Physical review letters **57 21**, 2607 (1986).

[466] C. J. Geyer, (1991).

[467] R. M. Neal, Statistics and Computing **11**, 125 (2001).

[468] D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).

[469] P. Del Moral, A. Doucet, and J. Ajay, BAYESIAN STATISTICS **8**, 1 (2007).

[470] F. Locatello, R. Khanna, J. Ghosh, and G. Ratsch, in Proceedings of the twenty-first international conference on artificial intelligence and statistics, Vol. 84, edited by A. Storkey and F. Perez-Cruz, Proceedings of Machine Learning Research (Sept. 2018), pp. 464–472.

[471] P. Hennig, in Proceedings of the 30th international conference on machine learning, Vol. 28, edited by S. Dasgupta and D. McAllester, Proceedings of Machine Learning Research 1 (17–19 Jun 2013), pp. 62–70.

[472] D. P. Kingma and J. Ba, in 3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings (2015).

[473] J. Duchi, E. Hazan, and Y. Singer, J. Mach. Learn. Res. **12**, 2121 (2011).

[474] T. Tieleman, "Optimizing neural networks that generate images", PhD thesis (2014).

[475] S. Ruder, *An overview of gradient descent optimization algorithms*, 2016.

[476] R. Adams, H. Wallach, and Z. Ghahramani, in Proceedings of the thirteenth international conference on artificial intelligence and statistics, Vol. 9, edited by Y. W. Teh and M. Titterington, Proceedings of Machine Learning Research (13–15 May 2010), pp. 1–8.

[477] C. K. Wikle, "Hierarchical models for uncertainty quantification: an overview", in *Handbook of uncertainty quantification*, edited by R. Ghanem, D. Higdon, and H. Owhadi (Springer International Publishing, Cham, 2017), pp. 193–218.

[478] M. A. Katsoulakis and J. Trashorras, Journal of Statistical Physics **122**, 115 (2006).

[479] E. Ising, Zeitschrift für Physik **31**, 253 (1925).

[480] G. Ciccotti, K. Binder, and M. Ferrario, "Introduction: condensed matter theory by computer simulation", in *Computer simulations in condensed matter systems: from materials to chemical biology volume 1*, edited by M. Ferrario, G. Ciccotti, and K. Binder (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006), pp. 1–11.

[481] J. Lu, Y. Qiu, R. Baron, and V. Molinero, Journal of Chemical Theory and Computation **10**, PMID: 26588552, 4104 (2014).

[482] C. M. Bishop, in Proceedings of the 11th international conference on neural information processing systems, NIPS'98 (1998), pp. 382–388.

[483] M. E. Tipping and C. M. Bishop, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**, 611 (1999).

[484] M. E. Tipping and C. Bishop, Journal of the Royal Statistical Society, Series B **21**, Available from http://www.ncrg.aston.ac.uk/Papers/index.html, 611 (1999).

[485] H. Hotelling, J. Educ. Psych. **24** (1933).

[486] J. Zhang, Y. I. Yang, and F. Noé, The Journal of Physical Chemistry Letters **10**, PMID: 31522495, 5791 (2019).

[487] S. Ravi and A. Beatson, in International conference on learning representations (2019).

[488] J. Marino, Y. Yue, and S. Mandt, in Proceedings of the 35th international conference on machine learning, Vol. 80, edited by J. Dy and A. Krause, Proceedings of Machine Learning Research (Oct. 2018), pp. 3403–3412.

[489] R. Shu, H. H. Bui, S. Zhao, M. J. Kochenderfer, and S. Ermon, in *Advances in neural information processing systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), pp. 4393–4402.

[490] Z. Ghahramani and M. J. Beal, in *Advances in neural information processing systems 12*, edited by S. A. Solla, T. K. Leen, and K. Müller (MIT Press, 2000), pp. 449–455.

[491] M. E. Tipping, J. Mach. Learn. Res. **1**, 211 (2001).

[492] C. M. Bishop and M. E. Tipping, Artif. Intell., 46 (2002).

[493] C. Bishop, in Proceedings ninth international conference on artificial neural networks, icann'99, Vol. 1 (Jan. 1999), pp. 509–514.

[494] C. Bishop, in Advances in neural information processing systems, Vol. 11 (Jan. 1999), pp. 382–388.

[495] P. Moritz, R. Nishihara, and M. I. Jordan, (2015).

[496] R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal, SIAM Journal on Optimization **21**, 977 (2011).

[497] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer, (2014).

[498] C. Chen, D. Carlson, Z. Gan, C. Li, and L. Carin, (2015).

[499] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, Vol. 35 (Springer Science & Business Media, 2003).

[500] R. Vargas, J. Garza, B. P. Hay, and D. A. Dixon, The Journal of Physical Chemistry A **106**, 3213 (2002).

[501] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, Journal of Molecular Biology **7**, 95 (1963).

[502] N. Metropolis and S. Ulam, J. Am. Stat. Assoc. **44**, 335 (1949).

[503] M. Griebel, *Sparse grids and related approximation schemes for higher dimensional problems* (2005).

[504] A. Barducci, M. Bonomi, and M. Parrinello, WIREs Computational Molecular Science **1**, 826 (2011).

[505] J. R. Perilla, B. C. Goh, C. K. Cassidy, B. Liu, R. C. Bernardi, T. Rudack, H. Yu, Z. Wu, and K. Schulten, Current Opinion in Structural Biology **31**, 64 (2015).

[506] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, Journal of chemical theory and computation **6**, PMC3633568[pmcid], 583 (2010).

[507] F. Pietrucci and W. Andreoni, Phys. Rev. Lett. **107**, 085504 (2011).

[508] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, The Journal of Chemical Physics **134**, 135103 (2011).

[509] W. Zheng, M. A. Rohrdanz, and C. Clementi, The Journal of Physical Chemistry B **117**, PMID: 23865517, 12769 (2013).

[510] O. Valsson and M. Parrinello, Phys. Rev. Lett. **113**, 090601 (2014).

[511] W. Chen and A. L. Ferguson, Journal of Computational Chemistry **39**, 2079 (2018).

[512] P.-Y. Chen and M. E. Tuckerman, The Journal of Chemical Physics **148**, 024106 (2018).

[513] A. Mitsutake, Y. Mori, and Y. Okamoto, "Enhanced sampling algorithms", in *Biomolecular simulations: methods and protocols*, edited by L. Monticelli and E. Salonen (Humana Press, Totowa, NJ, 2013), pp. 153–195.

[514] C. Bierig and A. Chernov, Journal of Computational Physics **314**, 661 (2016).

[515] J. Luque and X. Barril, eds., *Physico-chemical and computational approaches to drug discovery*, RSC Drug Discovery (The Royal Society of Chemistry, 2012), FP001–418.

[516] M. A. Rohrdanz, W. Zheng, and C. Clementi, Annual Review of Physical Chemistry **64**, PMID: 23298245, 295 (2013).

[517] R. T. McGibbon, B. E. Husic, and V. S. Pande, The Journal of Chemical Physics **146**, 044109 (2017).

[518] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, Science **290**, 2319 (2000).

[519] M. Ceriotti, G. A. Tribello, and M. Parrinello, Proceedings of the National Academy of Sciences **108**, 13023 (2011).

[520] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, The Journal of Chemical Physics **134**, 124116 (2011).

[521] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proceedings of the National Academy of Sciences **102**, 7426 (2005).

[522] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proceedings of the National Academy of Sciences **102**, 7432 (2005).

[523] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, Applied and Computational Harmonic Analysis **21**, Special Issue: Diffusion Maps and Wavelets, 113 (2006).

[524] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Multiscale Modeling & Simulation **7**, 842 (2008).

[525] F. Noé and F. Nüske, Multiscale Modeling & Simulation **11**, 635 (2013).

[526] F. Noé, R. Banisch, and C. Clementi, Journal of Chemical Theory and Computation **12**, PMID: 27696838, 5620 (2016).

[527] J. McCarty and M. Parrinello, The Journal of Chemical Physics **147**, 204109 (2017).

[528] W. Zheng, A. V. Vargiu, M. A. Rohrdanz, P. Carloni, and C. Clementi, The Journal of Chemical Physics **139**, 145102 (2013).

[529] M. Balasubramanian and E. L. Schwartz, Science **295**, 7 (2002).

[530] D. L. Donoho and C. Grimes, Proceedings of the National Academy of Sciences **100**, 5591 (2003).

[531] H. Risken and T. Frank, *The fokker-planck equation: methods of solution and applications (springer series in synergetics)* (Springer, 1996).

[532] Y. Kim, S. Wiseman, A. Miller, D. Sontag, and A. Rush, in Proceedings of the 35th international conference on machine learning, Vol. 80, edited by J. Dy and A. Krause, Proceedings of Machine Learning Research (Oct. 2018), pp. 2678–2687.

[533] R. Ranganath, D. Tran, and D. Blei, in Proceedings of the 33rd international conference on machine learning, Vol. 48, edited by M. F. Balcan and K. Q. Weinberger, Proceedings of Machine Learning Research (20–22 Jun 2016), pp. 324–333.

[534] A. Lecchini-visintini, J. Lygeros, and J. Maciejowski, in *Advances in neural information processing systems 20*, edited by J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Curran Associates, Inc., 2008), pp. 865–872.

[535] Y. Li, V. A. Protopopescu, N. Arnold, X. Zhang, and A. Gorin, Appl. Math. Comput. **212**, 216 (2009).

[536] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, Nature physics **6**, PMC3032381[pmcid], 751 (2010).

[537] D. L. Ensign and V. S. Pande, Biophysical journal **96**, S0006-3495(09)00495-0[PII], L53 (2009).

[538] Y.-L. Chen and M. Habeck, PLOS ONE **12**, 1 (2017).

[539] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).

[540] C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, and V. S. Pande, in (2017).

[541] C. Wehmeyer and F. Noé, The Journal of Chemical Physics **148**, 241703 (2018).

[542] M. M. Sultan, H. K. Wayment-Steele, and V. S. Pande, Journal of Chemical Theory and Computation **14**, PMID: 29529369, 1887 (2018).

[543] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Structure, Function, and Bioinformatics **17**, 412 (1993).

[544] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Chemical Physics Letters **509**, 1 (2011).

[545] J. C. Pinheiro and D. M. Bates, Statistics and Computing **6**, 289 (1996).

[546] D. Kingma and M. Welling, in Proceedings of the 31st international conference on machine learning, Vol. 32, edited by E. P. Xing and T. Jebara, Proceedings of Machine Learning Research 2 (22–24 Jun 2014), pp. 1782–1790.

[547] F. R. Ruiz, M. Titsias RC AUEB, and D. Blei, in *Advances in neural information processing systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 460–468.

[548] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Nature **323**, 533 (1986).

[549] R. Chandra, K. Jain, R. V. Deo, and S. Cripps, (2018).

[550] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, in *Advances in neural information processing systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), pp. 971–980.

[551] A. Onufriev, D. Bashford, and D. A. Case, Proteins: Structure, Function, and Bioinformatics **55**, 383 (2004).

[552] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, Journal of the American Chemical Society **112**, 6127 (1990).

[553] J. M. Troyer and F. E. Cohen, Proteins: Structure, Function, and Bioinformatics **23**, 97 (1995).

[554] F. Liang, Journal of Statistical Computation and Simulation **80**, 1007 (2010).

[555] R. Galvelis and Y. Sugita, Journal of Chemical Theory and Computation **13**, PMID: 28437616, 2489 (2017).

[556] S. P. Carmichael and M. S. Shell, The Journal of Physical Chemistry B **116**, PMID: 22300263, 8383 (2012).

[557] T. P. Minka, "A family of algorithms for approximate bayesian inference", PhD thesis (Massachusetts Institute of Technology, Cambridge, MA, USA, 2001).

[558] J. A. T. Thomas M. Cover, *Elements of information theory*, 2nd ed, Wiley Series in Telecommunications and Signal Processing (Wiley-Interscience, 2006).

[559] H. Kahn and T. E. Harris, Journal of Research of the National Bureau of Standards. Applied Mathematics Series **12**, 27 (1951).

[560] J. P. Valleau and D. N. Card, The Journal of Chemical Physics **57**, 5457 (1972).

[561] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, (2017).

[562] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, `http://www.deeplearningbook.org` (MIT Press, 2016).