Rheologica Acta
https://doi.org/10.1007/s00397-023-01397-w

ORIGINAL CONTRIBUTION

Bayesian coarsening: rapid tuning of polymer model parameters

Hansani Weeratunge[1] · Dominic Robe[1] · Adrian Menzel[2] · Andrew W. Phillips[2] · Michael Kirley[3] · Kate Smith-Miles[4] · Elnaz Hajizadeh[1]

Received: 16 March 2023 / Revised: 30 April 2023 / Accepted: 4 May 2023

Abstract
A protocol based on Bayesian optimization is demonstrated for determining model parameters in a coarse-grained polymer simulation. This process takes as input the microscopic distribution functions and temperature-dependent density for a targeted polymer system. The process then iteratively considers coarse-grained simulations to sample the space of model parameters, aiming to minimize the discrepancy between the new simulations and the target. Successive samples are chosen using Bayesian optimization. Such a protocol can be employed to systematically coarse-grained expensive high-resolution simulations to extend accessible length and time scales to make contact with rheological experiments. The Bayesian coarsening protocol is compared to a previous machine-learned parameterization technique which required a high volume of training data. The Bayesian coarsening process is found to precisely and efficiently discover appropriate model parameters, in spite of rough and noisy fitness landscapes, due to the natural balance of exploration and exploitation in Bayesian optimization.

Keywords Polymers · Molecular dynamics · Machine learning · Bayesian optimization

Introduction

Coarse-grained (CG) modeling is a valuable technique to bridge the gaps in time and length scales between atomically accurate molecular dynamics simulations and experimentally observable rheological behavior (Prathumrat et al., 2021). Particularly, low frequency responses, glassy dynamics, and finite size effects have been unrelenting challenges for computer models. CG modeling provides a lower-resolution depiction of a complicated system by grouping atoms into a representative particle. However, the interactions between atoms must be accurately represented by a CG model for it to yield correct rheology (Hajizadeh et al., 2014a, b, 2015). The strategy is to replace sets of atoms with a single point-like pseudo-atom or bead, reducing the degrees of freedom in the simulation. The beads interact with potentials that are chosen to reproduce the structure and dynamics brought about by the underlying atoms. In our case, we consider monomers to be connected by Hookean springs defined by a stiffness k_l and a rest length l_0, and a harmonic angular potential with stiffness k_θ and rest angle θ_0. Additionally, the interaction between nearby monomers which are not connected by a bond is represented with a Lennard-Jones potential with length scale σ and energy scale ϵ. These six parameters specify a CG model, but there is no general technique to specify a priori what the specific parameters for the interaction potentials should be to reproduce the rheology of a material of interest.

To ensure an accurate representation of molecular interactions, force field parameterization must be validated. This is usually accomplished by determining how well a set of candidate force field parameters reproduces essential physical phenomena, such as ab initio quantum calculation, or experimental data on structure and rheology (Prathumrat et al.,

Hansani Weeratunge and Dominic Robe contributed equally to this work

✉ Elnaz Hajizadeh
ellie.hajizadeh@unimelb.edu.au

Dominic Robe
nick.robe@unimelb.edu.au

1 Department of Mechanical Engineering, Faculty of Engineering and Information Technology , The University of Melbourne, 700 Swanston St., Melbourne, Victoria, Australia

2 Platforms Division, Defence Science and Technology Group, Port Melbourne, Victoria, Australia

3 School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia

4 School of Mathematics and Statistics, The University of Melbourne, Melbourne, Victoria, Australia

Published online: 05 July 2023

2021). In our case we will rely on the density as a function of temperature, as well as probability distributions of bond lengths and angles, to evaluate model fitness. Iteratively refining model parameters to navigate toward the best approximation of the system being modeled is time consuming due to the expensive simulations required for each iteration. With technological advancements and increased computational power, efficient optimization algorithms and data-driven techniques can be employed to automate the parameterization process (Sestito et al. 2020; Kanada et al. 2020). In general, force field calibration can be defined as an optimization process in which the force field parameters are tuned to minimize the difference between a predicted property from the coarse-grained (CG) simulation and the reference value from high-resolution atomistic trajectories (Liu et al. 2008) or experimental data.

Direct search methods, gradient-based approaches (Hajizadeh and Garmabi, 2008), and machine learning (ML) approaches (Hansani et al. 2022) have recently emerged as effective tools for exploring the complex solution space of material design problems (Chen et al. 2021; Solomou et al. 2018). These methods can systematically discover the optimum of "black box" functions. However, these methods tend to require a large set of training data to perform such a task. CG simulations are most valuable when experiments or detailed simulations are prohibitively expensive, which means that even though coarse-graining brings a problem into feasibility, they still require non-trivial resources to execute. Furthermore, modern material development workflows involve comparison between a variety of molecular components, requiring model parameterization to be carried out many times for different chemistries. Bayesian optimization is an active learning algorithm that can reduce the number of evaluations needed to locate the optimum of an expensive black box function. This work therefore will apply Bayesian optimization to a modern method of coarse-graining.

Several studies (Dequidt and Solano 2015; Fröhlking et al. 2020; Ye et al. 2021) have optimized force field parameters for molecular dynamics using ML and Bayesian models through "bottom-up" approaches where microscopic structural properties obtained from atomistic simulations form the objective function. Furthermore, classical approaches such as iterative Boltzmann inversion (Agrawal et al. 2014; Bayramoglu et al. 2012; Liu and Oswald 2019; Ohkuma and Kremer 2020), inverse Monte Carlo (Korolev et al. 2014; Lyubartsev et al. 2015), and relative entropy (Foley et al. 2015; Shell 2016) approximate the microscopic configuration at a single state point to parameterize CG force fields. As a result, models calibrated with a bottom-up objective frequently suffer from issues of transferability, i.e., applicability in other state points. Furthermore, these parameterizations are also likely to provide limited accuracy in determining thermodynamic properties (Dunn and Noid 2015).

In contrast, "top-down approaches" calibrate model parameters using macroscopic phenomena or mechanical properties (such as density, glass transition temperature, and elastic modulus). These models generally demonstrate better transferability for modeling a wide range of thermodynamic conditions. They may, however, provide a poor description of microscopic structural properties. Therefore, many studies have recently embraced a hybrid approach, integrating both bottom-up and top-down calibrations (Shireen et al. 2022; Huang et al. 2018; Hsu et al. 2015; Moradzadeh and Aluru 2019; Duan et al. 2019). As a result of combining the constraints from both methodologies, the hybrid strategy yields CG models capable of replicating microscopic features and macroscopic properties with significant transferability and representability (Joshi and Deshmukh 2020). In a recent study, Shireen et al. (2022) demonstrated that neural networks (NNs) can be trained to receive data from a high-resolution simulation as bottom-up input and experimental density versus temperature data as top-down input, then return CG model parameters which accurately reproduced the target data (Shireen et al., 2023). This strategy for determining temperature-transferable CG parameters using NNs was effective, but costly, requiring several thousand CG simulations as training data to achieve viable accuracy.

The aim of this work is therefore to maintain the accuracy and temperature-transferability of CG parameter choices, but reduce the number of CG simulations needed to discover the parameters that best fit a target. Measuring the accuracy of model parameters is challenging because different applications require different properties to be reproduced, and a choice of coarse-grained force field forms and parameters will inevitably introduce trade-offs, so the "correct" model parameters not well-defined. A separate challenge is that merely constructing a fine-grained simulation of a particular material is a time-consuming expert task which itself requires validation. A statistically significant comparison of the number of simulations needed by different strategies to resolve CG parameters would require many such targets. However, since Shireen et al. (2022) generated a set of thousands of CG simulations for training and validation of the NN based coarse-graining method, this data set provides a test-bed for measuring the improved data requirements of a new method.

In general, optimization techniques search for the maximum or the minimum of a function by evaluating selected locations in the search space. These approaches must balance the exploitation of the knowledge gained from previous evaluations and the exploration of unknown regions that might hold a better solution. This balance is crucial when there is a limited budget available for sample evaluations. Among these methods, Bayesian optimization (BO) is an efficient tool for optimizing expensive black box functions with the least amount of direct evaluation of the objective function (Liang et al. 2021). BO is a ML approach capable of effi-

ciently balancing the exploration-exploitation dilemma that is common in optimization problems. This can be categorized as an active learning methodology that builds a probabilistic surrogate model of the objective function to account for the uncertainty. This is usually generated using a Bayesian model known as a Gaussian process, though other models, such as Bayesian neural networks (Fortuin 2022), have also been used successfully. An acquisition function, in our case expected improvement, is applied to the surrogate model to determine the next point to sample in the solution space. The posterior distribution improves as the number of observations increases, and the algorithm becomes more certain of which regions of parameter space are worth exploring and exploiting. Under some conditions, expected improvement can be guaranteed to converge to the global optimum (Bull 2011).

In this present work, coarse-grained force field parameters are identified efficiently by applying Bayesian Optimization to a hybrid parameterization process. First, six parameters are approximated using physical principles laid out in the "Parameter estimations" section. Then, the two bond extension parameters are refined using BO, using the accuracy of the bond length distribution (a microscopic, bottom-up target) as an objective. Then, the two bond angle parameters are similarly refined using the bond angle distribution (again, bottom-up). Finally, the two pairwise parameters are refined using the temperature-dependent density (a macroscopic, top-down objective). The relative computational cost of identifying the correct parameters using this process is then compared to a previous technique (Shireen et al. 2022).

Other choices of objective function for the non-bonded interactions were considered for this study. It should be noted that, though the ultimate goal of such a coarse-graining process is to construct a model that can predict rheological properties at practically relevant length and time scales, those rheological properties are not generally an effective metric from which to derive an objective function. Coarse-grained simulations are most useful for studying novel materials for which detailed rheological data is not available. Generating rheological data from detailed chemical models is the prohibitively expensive task which CG models are meant to circumvent. Previous work has shown that this CG model reproduces the mean squared displacement, glass transition temperature, and relaxation spectrum of an all-atom model (Shireen et al., 2022) if the parameters are correct, even though the parameters are calibrated using structural properties. Therefore, though intuitively obvious, using rheological properties to build an objective function is not as effective as utilizing the measurements used here and elsewhere. One common strategy for determining the non-bonded interaction parameters is to reproduce the radial distribution function $g(r)$ (Agrawal et al., 2014; Bayramoglu et al., 2012; Liu and Oswald, 2019; Ohkuma and Kremer, 2020). This bottom-up approach yields parameters which are only valid at a particular temperature. Previous work has demonstrated that using $\rho(T)$ as a signal for tuning the non-bonded interactions may yield a less precise reproduction of $g(r)$, but reproduces other macroscopic properties accurately over a range of temperatures, even outside of the range used for tuning. This could be considered as an over-fitting situation, where the details of $g(r)$ are not vital to the macroscopic rheology, so fitting it precisely sacrifices the broader validity of a model to new conditions. This is not the first effort to use Bayesian optimization to facilitate parameterization, but this work integrates and extends reported strategies. Sestito et al. (2020) calibrated bonded interaction parameters for a CG model of polycaprolactone using the linear elastic modulus and Fikian diffusion coefficient as objective functions. McDonagh et al. (2019) used BO to parameterize the non-bonded interactions in DPD models of an assortment of alkanes and primary alcohols, using a top-down objective function to reproduce partition coefficients in water. Befort et al. (2012) applied BO to optimize atomistic force field parameters for specific use cases. These studies demonstrate the viability of a Bayesian approach to parameter calibration under diverse training conditions, but they do not address the common pitfall of transferability. That is to say, parameters have been determined efficiently to reproduce any particular targeted rheological measurement under any particular conditions, but the parameterization must be carried out separately for each property or condition of interest. Work with the energy renormalization method (Giuntoli et al. 2020; Xia et al. 2017; Hsu et al. 2015) has combined bottom-up and top-down information and Gaussian process models to specify temperature-dependent CG parameters, but did not actively select new trial simulations based on existing data. This work applies BO to accelerate a parameterization strategy that has been shown to reproduce many rheological properties under diverse conditions with a single transferable parameter set. The strategy presented here also requires low implementation cost since all parameters are determined by applications of the same Bayesian optimization workflow, just with different objective functions.

The framework demonstrated here possesses some qualitative advantages over other parameterization strategies beyond minimizing the necessary simulations. Bayesian Optimization selects new parameters to query which provide the maximum novel information about the search space, which tends to mitigate the over-fitting issue in most machine learning processes. BO also produces an estimate of the uncertainty around the landscape of model parameter fitness instead of just an inscrutable declaration of the recommended parameters.

The structure of this manuscript is as follows. The "Coarse-graining framework" section introduces the coarse-grained model to be parameterized and a novel workflow for carrying out that parameterization. The "Polymer models" section details the CG model and its parameters. The "Bayesian optimization" section explains the features of Bayesian optimization. The "Bayesian coarsening" section details the optimization routine used to search the parameter space. In particular, the "Parameter estimations" section describes a general technique for estimating model parameters to kick-start the tuning. The "Bottom-up tuning of bonded interactions" section specifies the bottom-up objective functions used to calibrate the bonded interactions and the "Top-down tuning of pairwise interactions" section specifies the top-down objective for pairwise interactions. The outcome of this calibration process is discussed in the "Results" section. The "Resource requirements" section discusses the accuracy of this workflow as a function of computational effort in comparison with a previous strategy which used neural networks to identify appropriate model parameters (Shireen et al. 2022). Finally, the "Conclusion" section contains some concluding remarks on the viability of this method for efficiently expanding the window of rheological observations.

Coarse-graining framework

The automated Bayesian parameterization approach developed in this study requires the integration of numerous elements in a workflow. As shown in Fig. 1, the workflow consists of microscopic target data, an iterated coarse-grained simulation, macroscopic target data, a regression model to estimate the fitness landscape, and an acquisition function to select new points on that landscape. The following subsections will explain the components of this workflow.

Polymer models

As suggested in the top right of Fig. 1, fine-grained atomic models represent atoms or small groups of atoms explicitly with an assortment of stretching, bending, and dihedral potentials. These models are able to reproduce the molecular structures at the atomic scale, but require evaluation of dozens of forces per monomer. Here we consider a generic coarse-grained (CG) model with just three interactions. It has been shown that this class of model is able to reproduce microscopic structural and dynamic properties of atomistic models (above the monomer scale), as well as material properties

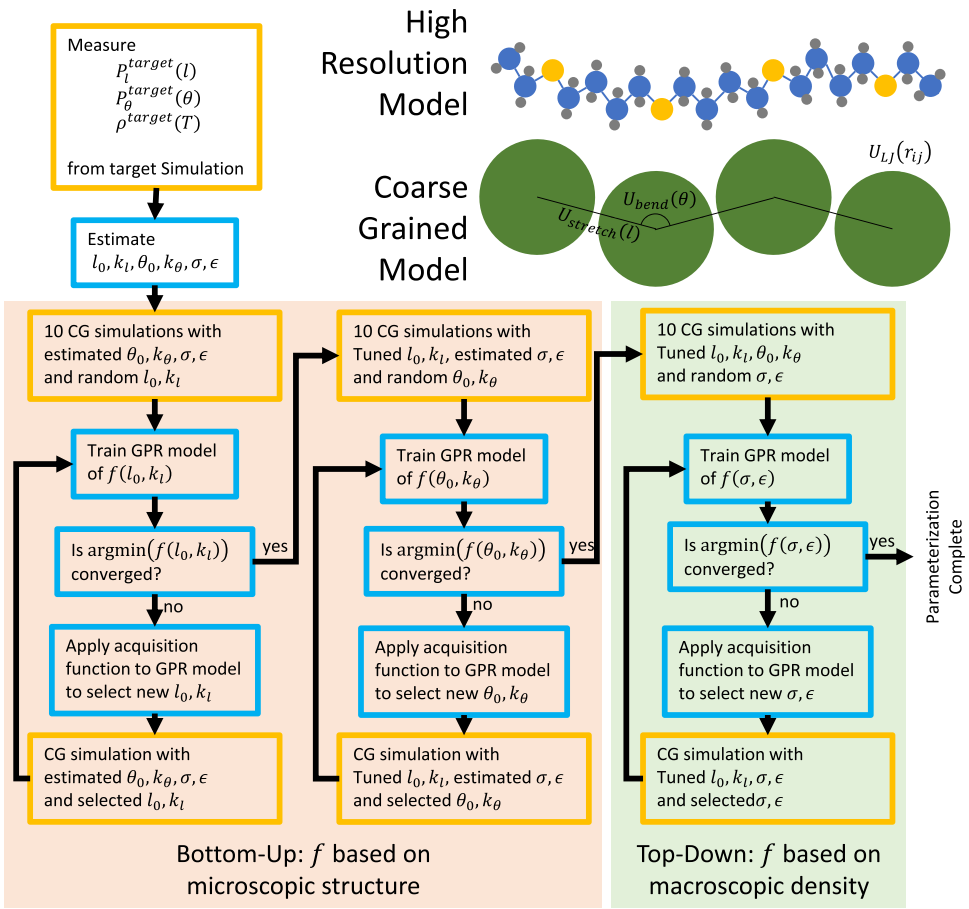

Fig. 1 Overview of this work's protocol for identifying appropriate coarse-grained model parameters. Top Right: Schematics comparing the CG and atomistic models. Gray, blue, and yellow circles represent different hydrogen, carbon, and oxygen atoms, respectively. Every type of atom and bond requires unique force field parameters. Flow Chart: Each column shows a similar application of a Bayesian Optimization process to a pair of model parameters

such as dynamic moduli and the glass transition temperature (Shireen et al. 2022). The potential due to stretching of the bond between two beads is (Zhu et al., 2016; Xiang et al., 2021)

$$U_{\text{stretch}}(r) = k_l \cdot (l - l_0)^2, \tag{1}$$

where l_0 is the rest length and k_l is a spring constant. The potential due to bending of the angle between any three consecutive beads along a polymer chain is

$$U_{\text{bend}}(\theta) = k_\theta \cdot (\theta - \theta_0)^2, \tag{2}$$

where θ_0 is the rest angle and k_θ is a spring constant. These potentials represent the monomers in a chain with a softened "freely rotating rod" model. In addition to these bonded interactions, there is a non-bonded Lennard-Jones interaction (Hajizadeh and Larson, 2017) between all pairs of nearby CG beads. The form of this force field is

$$U_{\text{nonb}}(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right], \tag{3}$$

where ϵ is the potential well depth and σ is the length scale.

The four bonded parameters $k_l, l_0, k_\theta, \theta_0$ and the two non-bonded parameters ϵ, σ must be set correctly for a CG simulation to accurately represent a particular material. The correct parameters will of course depend on the material being modeled, so the parameterization process would need to be repeated for every material of interest. It is therefore vital to develop a strategy to minimize the computational effort needed for such a parameterization.

Bayesian optimization

BO is an essential tool for optimizing objective functions that lack known functional forms and are expensive to evaluate. BO is characterized by a more efficient exploration and exploitation of the design space than other black box optimization techniques (Shahriari et al. 2016). This approach has shown substantial influence on current scientific discoveries, particularly autonomous calibration of force fields, which can be formulated as an optimization problem aiming at finding the maximum (or minimum) of an objective function (McDonagh et al. 2019).

Bayesian optimization works by generating a stochastic approximation of the expensive objective function via a probabilistic surrogate model. This stochastic predictive model is usually, as in this work, built using Gaussian process regression (GPR). The GPR model is initially trained using a small set of data prior to the optimization. In the present work, the GPR is initialized with 10 random points in the space of parameters within certain ranges presented in Table 1.

Table 1 Parameter ranges used to generate CG simulation data

Parameter	Unit	Minimum	Maximum
Rest Length l_0	Å	3.00	7.0
Stretching stiffness k_l	kcal/mol·Å^2	0.01	50.0
Rest Angle θ_0	degrees	100.00	180.0
Bending stiffness k_θ	kcal/mol·rad^2	0.01	5.0
Non-bonded length σ	Å	1.00	10.0
Non-bonded energy ϵ	kcal/mol	0.10	1.2

Since BO is a stochastic process, the performance of the algorithm depends on the initial data set. To investigate the robustness of this initialization, the optimization process is repeated for 200 different targets to ensure statistical significance, using independent initial training sets for each target. The objective functions used here are the root mean squared error (RMSE) of a property of interest between a trial CG simulation with known parameters and a target taken from the set of pre-existing simulations. The GPR model, trained on the accumulated set of points in parameter space, predicts this objective function as a continuous function throughout the parameter space.

The GPR surrogate model of the objective function is much cheaper to evaluate than running new simulations, so it is used to estimate the location of the minimum of the objective function in parameter space. As more simulations are provided to the GPR model, it is updated and the location of the minimum is re-evaluated. When new data points stop moving the location of the minimum, the parameter values which minimize the surrogate objective function are accepted.

A critical component of a Bayesian optimization process is the algorithm by which new parameter values are chosen to be evaluated. This choice is called the "acquisition function". The GPR model produces both a predicted value of the objective function, and the variance associated with that prediction at any point in the parameter space. The exploration/exploitation problem posed by optimization tasks is formulated explicitly in BO through these values and uncertainty estimations. An exploitative strategy would search near locations with high value regardless of uncertainty. An exploratory strategy would search in areas of high uncertainty, hoping for new heights. Effective acquisition functions combine the value and uncertainty estimates to pinpoint test parameters that will provide novel information about the optimum of the objective. Numerous studies have specifically focused on different acquisition functions and how they trade off between exploration and exploitation (Pawar and Warbhe 2021; De Ath et al. 2021). Common acquisition functions include the upper confidence bound, probability of improvement, and expected improvement (EI). Here, we use the EI

matrix as the acquisition function that selects the next sample point where the highest magnitude of progress is expected. EI is derived from the prediction and uncertainty reported by the GPR model. If $\mathbf{x}$ is the vector of CG parameters, then the GPR prediction for the objective function evaluated at $\mathbf{x}$ is $\mu_{\mathrm{GPR}}(\mathbf{x})$ and the uncertainty of the same is $\sigma_{\mathrm{GPR}}(\mathbf{x})$. EI is then calculated as (De Ath et al. 2021; Jones et al. 1998)

$$EI(\mathbf{x}) = \sigma_{\mathrm{GPR}}(\mathbf{x})(s\Phi(s)\phi(s)), \tag{4}$$

where $s = (\mu_{\mathrm{GPR}}(\mathbf{x}) - f^*)/\sigma_{\mathrm{GPR}}(\mathbf{x})$ is the modeled improvement over the best true objective function evaluation so far f^*, normalized by the model uncertainty, and ϕ and Φ are the Gaussian probability and cumulative density functions. EI is a well-established criterion in Bayesian global optimization that is less likely to converge to a local optimum solution compared to other acquisition functions (Wu et al. 2019). In addition, EI has been shown to avoid samples that are dominated by another choice with a similar prediction but worse variance or vice versa (De Ath et al. 2021).

As shown in Fig. 1, the acquisition function is applied to the surrogate model, and the point which maximizes the acquisition function is simulated to produce a new data point. The GPR model is re-trained with the larger data set, and the process is repeated until the iteration budget is exhausted. The specific objective functions used will be discussed in the "Bottom-up tuning of bonded interactions" section and "Top-down tuning of pairwise interactions" section.

Bayesian coarsening

Here, we present our hybrid approach to coarse-graining using Bayesian optimization, combining bottom-up and top-down information. Since optimization routines are more efficient in lower dimensional parameter spaces, the process is decomposed into three stages, where subsets of the parameters are tuned in turn. This strategy is viable because the three different interactions are assumed to not be strongly coupled, so minor inaccuracies in a parameter for one type of interaction are assumed not to disrupt the objective function for a different interaction. Departures from this assumption will be discussed in the "Results" and "Conclusion" sections.

The precision of model parameters identified by the protocol detailed below, as well as the rate of improvement with successive iterations will be characterized by using CG simulations with hidden parameters as target data. Since the true parameters used to generate the target data are known, but not available to the protocol, the accuracy of the resulting parameterizations can be measured. We carried out the parameterization on 200 unique target CG simulations, allowing the Bayesian optimization routine to query 400 samples for the stretching potential, and 100 points in parameter space each for the bending and non-bonded interactions. To mitigate the computational load of carrying out all of the required CG simulations, a pre-sampling strategy was employed. 2000 CG simulations were run using LAMMPS (Plimpton 1995) with parameters drawn randomly from the parameter space outlined in Table 1 and temperatures ranging from 313K to 453K as part of a previous investigation (Shireen et al. 2022). Here each target simulation was removed from this pre-sampled data set, and nine initial samples were selected from the set at random. The tenth initial sample was selected using the parameter estimations discussed in the "Parameter estimations" section. The Bayesian optimization routine then computes the objective functions for each of these sample points against the target. It then fits a GPR model to the RMSE as a function of the parameters. The expected improvement of this GPR model is then computed for the parameter values of the remaining pre-sampled data set. The pre-sampled point with the highest expected improvement is added to the model's data set, and the process is repeated.

Parameter estimations

In any optimization process, initial parameter estimates and imposed ranges for each parameter must be determined before any optimization algorithm can run. Ranges that are too broad can yield slow convergence, while a narrow range might exclude the optimal parameter values. Determination of these ranges sometimes requires expert knowledge of a particular system. An approach that aspires to generalize to novel materials must have a robust systematic strategy for determining practical ranges for parameters without relying on expert experience with a specific material. To that end, parameter estimates for the CG model are deduced from the target data and general principles of molecular mechanics. This procedure may be applied to any material without detailed experience with that material.

Given the stretching potential U_{stretch} in Eq. 1, the Boltzmann distribution for the length l of a particular bond is

$$P(l) \propto \Omega(l) \cdot \exp\left[-\frac{U_{\mathrm{stretch}}(l)}{k_B T}\right] \tag{5}$$

where $\Omega(l) \propto l^2$ is the number of microstates which exhibit a particular value of l. Here it is proportional to the surface area of a sphere with radius l. The approximate probability distribution over l is therefore

$$P(l) = Cl^2 \exp\left[-\frac{k_l}{k_B T}(l - l_0)^2\right]. \tag{6}$$

This form was used to fit the bond length distribution data from the target simulation at T=453 K, using k_l, l_0 as fit

parameters. Some examples of these distributions are shown in Fig. 2. The resulting values were used to select one of the initial samples of the parameter space at the start of the BO process, and nine other samples were chosen at random from the data set.

The probability distribution for the bond angle is very similar to that for the bond length, except that the number of microstates is determined by the area of a cone with interior angle $180° - \theta$, which is proportional to $\sin\theta$. Therefore the probability distribution over θ has the form

$$P(l) = C \sin\theta \exp\left[-\frac{k_\theta}{k_B T}(\theta - \theta_0)^2\right]. \quad (7)$$

Using this form to fit the target bond angle distribution, the resulting parameter values were used to select one of the ten initial samples for the bending potential phase of the parameterization. Examples of the bond angle distributions from CG simulations are shown in Fig. 3.

Because the target data set was drawn from the parameter space outlined in Table 1, the searchable parameter space is known in advance. However, this is not the case in a genuine coarse-graining task. We therefore note that the bounds of the parameter search could be set much narrower, based on the fit to the target distribution, without foreknowledge of the correct parameters. From our observations of several hundred target simulations, the parameter estimations using

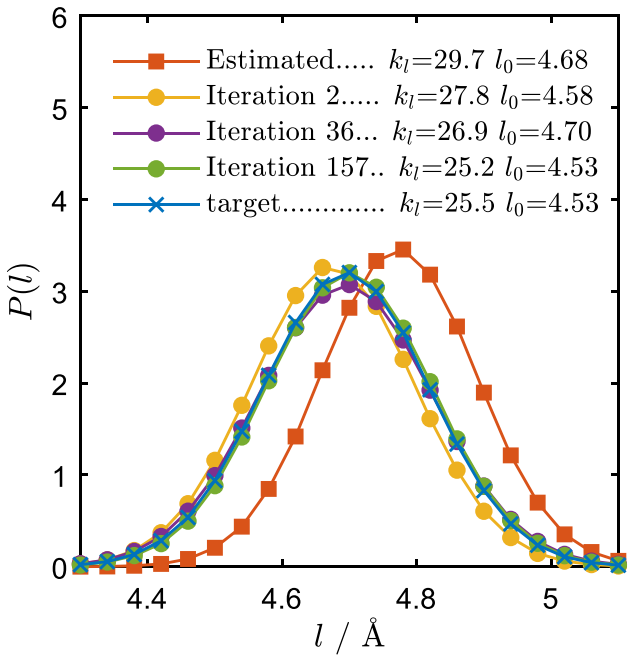

Fig. 2 Convergence of the probability distribution for the bond length to an example target. The "estimated" and "iteration" curves correspond to the square and circle points in Fig. 4. True parameters for this target were $k_l = 25.5, l_0 = 4.53, k_\theta = 7.43, \theta_0 = 104.2, \epsilon = 0.48, \sigma = 9.49$

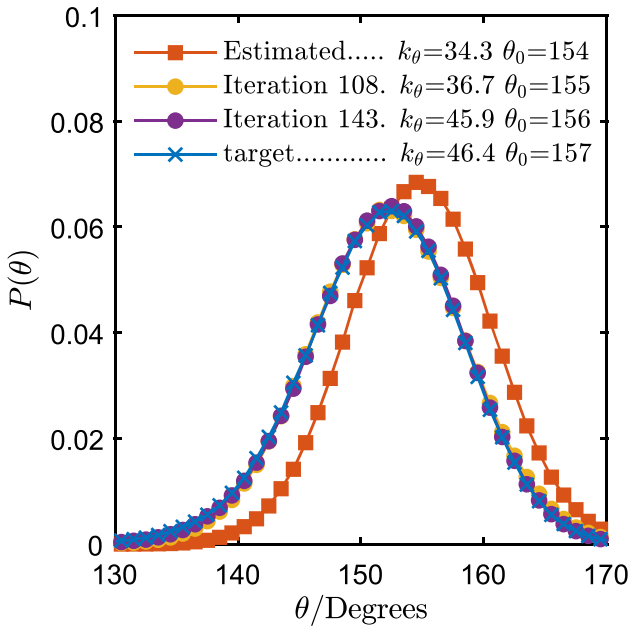

Fig. 3 Convergence of the probability distribution for the bond angle to an example target. The estimated and iteration curves correspond to the square and circle points in Fig. 5.True parameters for this target were $k_l = 41.3, l_0 = 5.79, k_\theta = 46.4, \theta_0 = 156.6, \epsilon = 1.19, \sigma = 4.15$

the Boltzmann distributions are usually accurate to within a few percent. Further, l_0 and θ_0 are almost always within 10% of the target, and k_l and k_θ are almost always within 50% of the target. The use of the full range of parameters listed in Table 1 to generate the initial samples and carry out the parameter search is therefore more of a challenge than an advantage for the search due to the unnecessarily wide search space.

Regarding the non-bonded interaction, the pair distribution function $g(r)$ from the target simulation provides estimates for the length scale σ and energy scale ϵ as these are correlated with the location and height of the first peak, respectively. The location of the first peak of $g(r)$ provided a sufficient estimate of σ to initialize the parameter search, but the energy scale ϵ is sometimes challenging to estimate. Various estimations of ϵ as a function of peak height were tested, and none achieved consistent accuracy. Perhaps a more sophisticated analysis of $g(r)$ could extract a more useful estimate of ϵ, but in practice, the Bayesian search can discover the true ϵ even with a random initialization. The non-bonded interaction represents the collective Van der Waals interaction between the atoms represented by a CG bead. Van der Waals forces typically have a strength on the order of 0.1 kcal/mol, so the non-bonded interaction energy was limited to a range of $0.01 < \epsilon < 1$ kcal/mol. Typical length scales for this interaction between monomers are on the order of a few angstroms, so the protocol presented here was investigated in a range of $1 < \sigma < 10$Å.

It is noted that this use of the Boltzmann distribution to estimate appropriate parameter values can be directly applied to any material for which the ground-truth probability distributions have been measured. This strategy generalizes to other systems or CG models where different microscopic structural measurements are relevant. The specific probability distributions calculated here apply to the CG model used in this work. If a different model were needed for a specific application, the Boltzmann distributions for that model could be calculated in a similar exercise. This use of fundamental principles of statistical mechanics avoids the need for material-specific expert knowledge to initialize the parameterization process.

Bottom-up tuning of bonded interactions

Bottom-up approaches employ information on atomistic structural properties from target data to parameterize the interactions in the CG model. In this study, we first carried out the bottom-up method to determine the bonded potential parameters between the CG beads by considering the bond length and angle distributions. The statistical optimization framework based on BO was formulated to find the bonded potential parameters $(k_l, l_0, k_\theta, \theta_0)$ by minimizing the RMSE between the CG simulation and target data for bond and angle distributions. The objective function for the bond length was

$$f(k_l, l_0)\text{: min}\sqrt{\frac{\sum_{l=1}^{N}(P_l^{\mathrm{CG}}(k_l, l_0) - P_l^{\mathrm{target}})}{N}}, \tag{8}$$

and the objective function for the bond angle is given by

$$f(k_\theta, \theta_0)\text{: min}\sqrt{\frac{\sum_{\theta=1}^{N}(P_\theta^{\mathrm{CG}}(k_\theta, \theta_0) - P_\theta^{\mathrm{target}})}{N}}. \tag{9}$$

Figure 2 contains examples of differences between distributions $P_l^{\mathrm{CG}}(k_l, l_0)$ and the target data P_l^{target}. These curves demonstrate the protocol's ability to rapidly discover an approximate match for the model parameters. After just two iterations, the model k_l and l_0 are within 9% and 1% of the target, respectively. Then the parameters are iteratively improved it until an indistinguishable result is found. Figure 4 presents the evolution of the GPR model of Eq. 8 (blue background indicates a lower prediction of f, yellow higher)

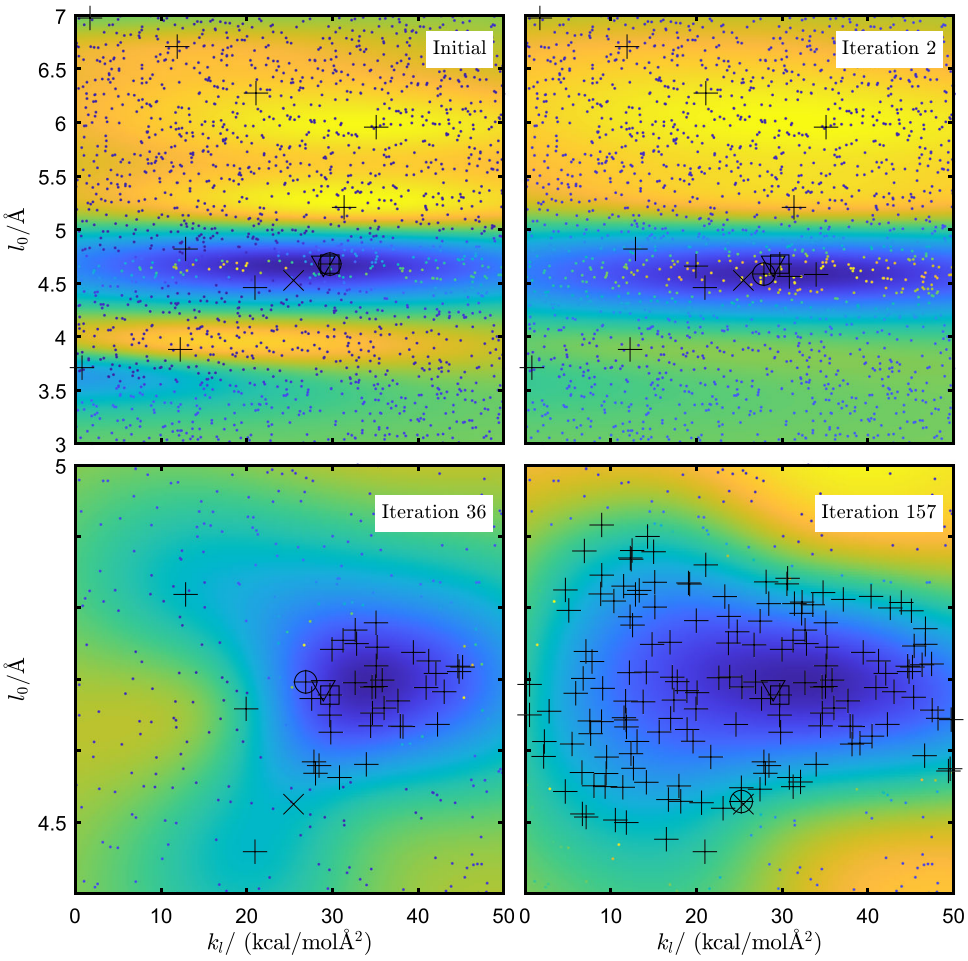

Fig. 4 Exploration of the space of stretching potential parameters k_l, l_0 driven by Bayesian optimization. The black + marks represent the sample points included in the GPR model up to a particular iteration. The black X marks the true parameters which generated the target data. The black triangle marks the estimated parameters using the Boltzmann distribution. The black square marks the initial sample closest to these estimated parameters. The black circle identifies the queried sample with the minimum objective. The colored background field represents the GPR model fit to the queried samples (black +). The colored points represent the expected improvement for candidate parameters for the next iteration. For both the GPR field and the EI points, blue indicates lower values, and yellow higher values. True parameters for this target were $k_l = 25.5, l_0 = 4.53, k_\theta = 7.43, \theta_0 = 104.2, \epsilon = 0.48, \sigma = 9.49$

with successive samples. By iteration 36, the correct value of k_l has been identified, and at iteration 157 the sample in the data set with the values of k_l and l_0 closest to the target has been discovered. The optimal region demonstrates that the discrepancy between the target data and the samples is much more sensitive to l_0 than it is to k_l. That is, the dark blue band spans a range of 0.5 Å in l_0, or about 10%, but a range of at least 20 kcal/molÅ^2 in k_l, almost a factor of 2 around the target. The concentration of potential samples with high EI (yellow points), near the target, demonstrates protocol's ability to identify regions of interest.

Because this study was carried out using a pre-existing set of CG simulations with random parameters, the exploration of the two-dimensional space of l_0 and k_l ignores the random changes in $\theta_0, k_\theta, \sigma$, and ϵ between simulations with similar bond stretching parameters. In a coarse-graining application, the acquisition function would be calculated over the whole parameter space, and a new simulation would be run with novel parameters. We have relied on the pre-existing data set here to avoid running hundreds of new simulations, each taking about a CPU hour, for each of our 200 targets. The disadvantage of this strategy is that we don't probe the rate that this workflow would converge if it had full control of all parameters.

It should be noted that the example target used for Figs. 2 and 4 was selected deliberately from the worst cases of slow convergence to illustrate this convergence process. Typically the initial approximation from the Boltzmann distribution is much more accurate, and the Bayesian search mostly validates that this solution is correct. This example target ($k_l = 25.5, l_0 = 4.53, k_\theta = 7.43, \theta_0 = 104.2, \epsilon = 0.48, \sigma = 9.49$) was chosen because the initial estimate was somewhat inaccurate due to the influence of the non-bonded interaction. This situation creates both an opportunity to illustrate the convergence process, and to demonstrate that the parameter search can be robust against such irregularities. The deceptiveness of the objective function is seen clearly in iteration 36 in Fig. 4. There are several sample points that are closer to the target than the lowest-RMSE point (identified with a circle). This happens because mismatched values for the non-bonded parameters are perceived by the GPR model as noise. So while the circled point may have less accurate stretching parameters than the closer samples, its particular non-bonded parameters make its distributions a better match to the target data. As the protocol continues to explore the parameter space, it eventually discovers best-fitting parameters in spite of this challenge.

Figure 3 presents the convergence of the bond angle distribution as the angle potential parameters are refined. Once again this target was selected intentionally from among the worst cases of initial estimate to illustrate the process of parameter space exploration. In this case, BO initially prefers a local minimum, seen as the area with high EI (yellow points) in the "Initial" state in Fig. 5. As samples are accumulated, the GPR model variance within that basin is reduced, and BO begins seeking new information in areas of high variance. New samples discover the global optimum, and the sampling preference shifts to the more optimal region (yellow points at iteration 143).

Top-down tuning of pairwise interactions

Developing transferable force field parameters in classical CG MD simulations has long been a challenge, especially for complex macromolecules. Classical techniques include optimizing the radial distribution function, which frequently necessitates an extra correction term to account for pressure variations in order to accurately model the thermodynamic

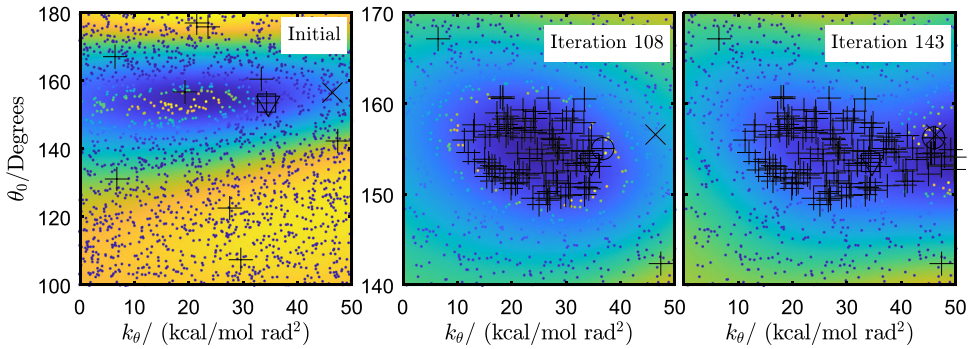

Fig. 5 Exploration of the space of bending potential parameters k_θ, θ_0 driven by Bayesian optimization. The black + marks represent the sample points included in the GPR model up to a particular iteration. The black X marks the true parameters which generated the target data. The black triangle marks the estimated parameters using the Boltzmann distribution. The black square marks the initial sample closest to these estimated parameters. The black circle identifies the queried sample with the minimum objective. The colored background field represents the GPR model fit to the queried samples (black +). The colored points represent the expected improvement for candidate parameters for the next iteration. For both the GPR field and the EI points, blue indicates lower values, and yellow higher values. True parameters for this target were $k_l = 41.3, l_0 = 5.79, k_\theta = 46.4, \theta_0 = 156.6, \epsilon = 1.19, \sigma = 4.15$

behavior (Bayramoglu et al. 2012; Reith et al. 2023). In our framework, after bonds and angle parameters have been determined, the non-bonded interaction parameters (σ and ϵ) are tuned by minimizing the RMSE of the density vs. temperature curve using

$$f(\sigma,\epsilon)\text{: min}\sqrt{\frac{\sum_{T=1}^{N}(\rho_T^{\text{CG}}(\sigma,\epsilon)-\rho_T^{\text{target}})}{N}}. \tag{10}$$

By asserting temperature-independent values of σ and ϵ, but tuning their values based on a range of temperatures, the CG model has been shown to reproduce viscoelastic response $G(t)$, the mean squared displacement, and the full dependence of density on temperature, even outside of the training range, capturing the glass transition temperature (Shireen et al. 2022).

The convergence to an example target ($k_l = 23.1, l_0 = 4.05, k_\theta = 7.42, \theta_0 = 179.8, \epsilon = 0.247, \sigma = 8.31$) is illustrated in Fig. 6. The initial estimates of σ and, particularly, ϵ are poor, yielding density measurements far from the target. After 2 iterations, a sample with a much closer value of ϵ is found which has the correct density at least at lower temperatures. After 105 iterations, a sample has been found with more accurate σ, but less accurate ϵ, which has better density agreement across the whole temperature range. Subsequent iterations refine the parameters to improve the density match.

Figure 7 visualizes the predictions of the objective function from the GPR for varying ϵ and σ. As seen in the broad scatter of potential samples with high expected improvement (yellow points), there is a very high uncertainty in the model parameters. However, the uncertainty decreases with iterations by gaining more knowledge of the search space as the GPR model obtains more data. These figures also illustrate the trade-off between exploration and exploitation. The exploitation aspect of BO can be clearly seen in the clustering of selected points, i.e., more data is collected in the region where the predicted RMSE is the lowest. However, The broad region of parameter space with high EI suggests that the GPR model is not confident in locations that merit further sampling, even after 350 iterations. This is in contrast to Fig. 4, in which high EI is tightly constrained to a narrow band in l_0, and even exploration along the k_l axis is slow. Relative to that focused exploitation, the broad scatter of samples in Fig. 7 is due to a weak dependence of the objective function on σ and ϵ, so the variance in the GPR model is high, and BO automatically favors a more exploratory search in this case.

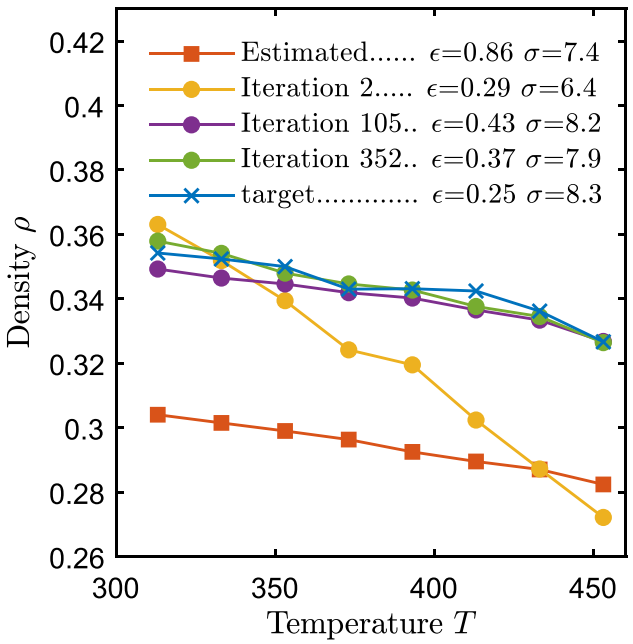

Fig. 6 Convergence of the temperature-dependent density to the target data. The estimated and iteration curves correspond to the square and circle points in Fig. 7. True parameters for this target were $k_l = 23.1, l_0 = 4.05, k_\theta = 7.42, \theta_0 = 179.8, \epsilon = 0.247, \sigma = 8.31$

Results

Resource requirements

After each iteration of the protocol, the model accuracy is evaluated in two ways. One is to find the parameter values which minimize the GPR model. The other is to find the sampled point with the minimum RMSE relative to the target data. These two measurements (presented in Fig. 8) roughly decrease with the number of samples at approximately the same rate, while the favored sample tends to be a better estimate of the model parameters than the minimum of the GPR model. Note that, due to the restriction to the pre-sampled set of parameter points, the stagnation of the stretching parameters curve at MAE near 0.03 after 200 samples is due to the limit of how close the closest pre-sampled point lies to the target. In some cases, the initial estimate using the Boltzmann distribution was more accurate than the closest available pre-sampled point. The available points appear to densely sample the parameter space in Figs. 4, 5, and 7, but these are 2D projections of a 6D space in which the true distance between simulations is much larger. If new simulations were run with fully customized parameters, the error floors in Fig. 8 could be significantly lower. Note also that the minimum of the GPR model is on average a less accurate prediction of the correct model parameters than the sample point with the lowest RMSE. This is likely because the GPR model assumes the samples of the objective function are noisy, so it doesn't strictly place the minimum of the surrogate model at the best sample.

A previous method for systematic coarse-graining relied on a dense neural network (NN), which was trained with

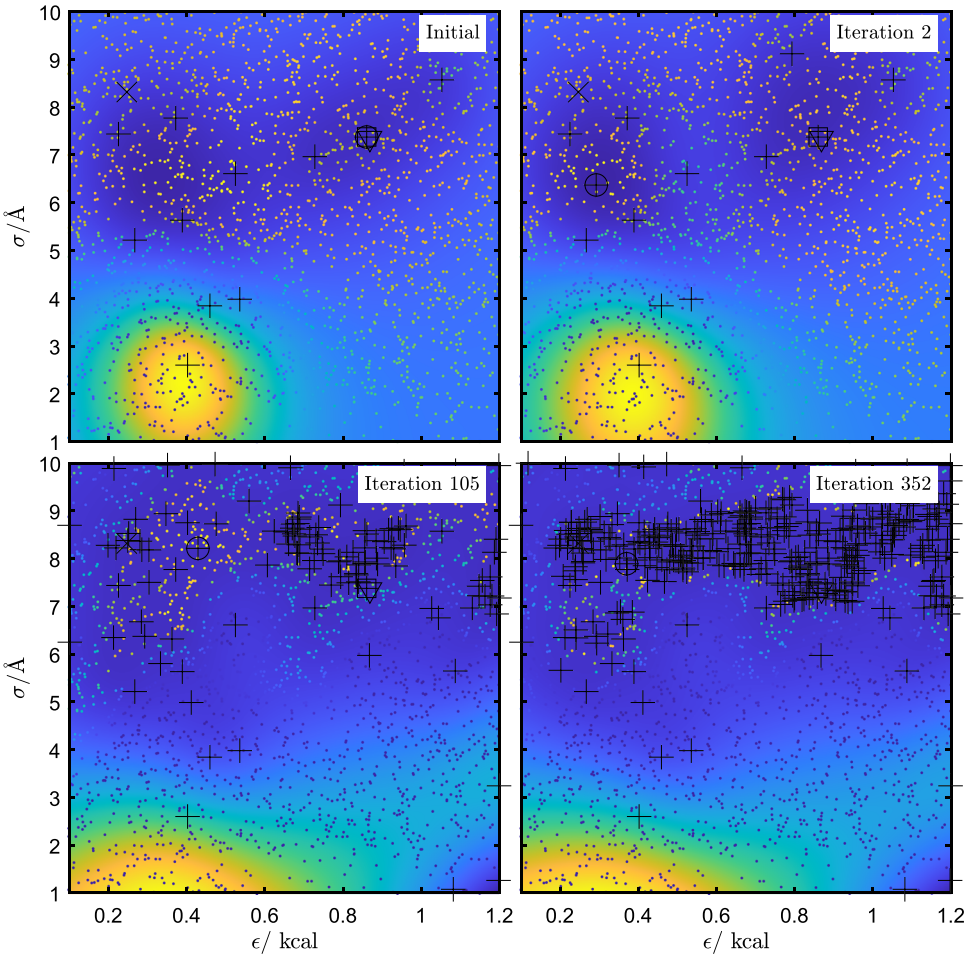

Fig. 7 Exploration of the space of non-bonded potential parameters k_θ, θ_0 driven by Bayesian optimization. The black + marks represent the sample points included in the GPR model up to a particular iteration. The black X marks the true parameters which generated the target data. The black triangle marks the estimated parameters using the Boltzmann distribution. The black square marks the initial sample closest to these estimated parameters. The black circle identifies the queried sample with the minimum objective. The colored background field represents the GPR model fit to the queried samples (black +). The colored points represent the expected improvement for candidate parameters for the next iteration. For both the GPR field and the EI points, blue indicates lower values, and yellow higher values. True parameters for this target were $k_l = 23.1, l_0 = 4.05, k_\theta = 7.42, \theta_0 = 179.8, \epsilon = 0.247, \sigma = 8.31$

distribution data as inputs and parameter values as outputs (Shireen et al. 2022). It then receives a novel distribution and returns an estimate of the model parameters which would produce that distribution. Data for the accuracy of this previous method as a function of the training set size are also included in Fig. 8 (filled blue triangles) to demonstrate the improved efficiency of the protocol presented here. Note that the accuracy of the bending and non-bonded parameters have not been analyzed as a function of training set size, but the mean absolute error for the bending parameters was 0.05 with 1300 samples. This suggests that the inaccuracy of the bending parameters in Fig. 8 relative to those for stretching is not a deficiency of the Bayesian protocol, but a quality of the CG parameter space. That is, it seems that $P(\theta)$ is just not as sensitive to k_θ and θ_0 as $P(l)$ is to k_l and l_0.

Pitfalls

Regarding the non-bonded interactions, Fig. 8 might seem to suggest that BO is performing poorly, but this may simply be a more severe case of insensitive parameters. Note that the optimal region of the parameter space in Fig. 7 (low RMSE, illustrated as blue background), is very broad and shallow. The $\rho(T)$ curves in Fig. 6 demonstrate that even when σ and ϵ are not perfect matches to the target, the density may still be reproduced faithfully. This may mean that the top-down approach to parameterization doesn't tightly constrain the non-bonded parameters, at least at the temperature range studied here. This could be advantageous as it leaves flexibility if other rheological properties need to be matched in addition to density in the future.

Another potential pitfall for the protocol as specified is the possibility that the different objective functions are not fully independent from the ignored parameters. That is, $P(l)$ for instance could be distorted by the influence of non-bonded interactions. In particular, large values of σ>7Å sometimes result in misidentified parameters, particularly if k_l or $k_\theta < 1$. Figure 2 demonstrates a mild case of this, as the peak of $P(l)$ is at $l \approx 4.7$ Å, noticeably higher than the true rest length of 4.5 Å, leading to the initial over-estimation of l_0. In these cases, the effect of the non-bonded interaction can

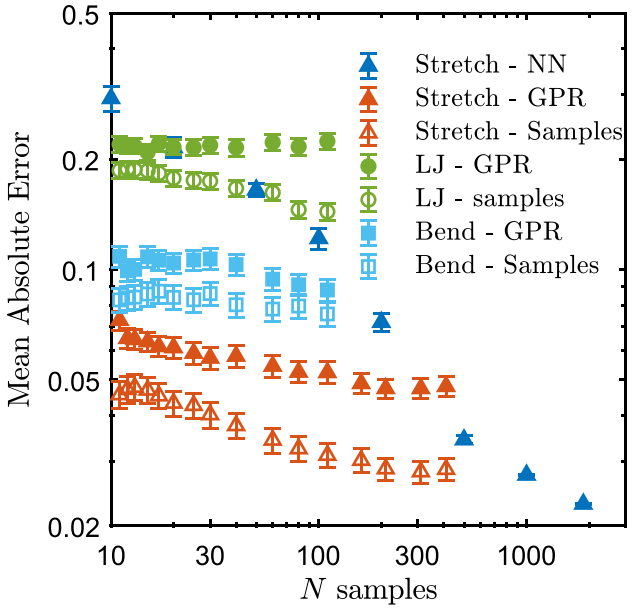

Fig. 8 Accuracy of model parameters as a function of the number of CG simulations used by the protocol. Absolute error is here measured by mapping the parameter ranges in Table 1 to the range [0,1] as in Shireen et al. (2022) to enable comparison between parameters with different scales. Triangles represent the parameters for the bond stretching potential. Squares represent the bond bending potential. Circles represent the Lennard-Jones non-bonded potential. Filled symbols represent the minimum of the GPR (or NN) model, while open symbols represent the sampled simulation with the lowest RMSE compared to the target. The filled blue triangles represent the accuracy of the NN developed in Shireen et al. (2022), with different training set sizes. Results for our protocol are averaged over 200 independent targets with independent sets of initial samples

overwhelm the bonded interactions. For example, sometimes $P(l)$ becomes bimodal. In principle, a similar optimization routine could overcome this complexity, but the assumption in this workflow that the interactions are independent makes this case challenging. As with any ML based protocol, solutions should be sanity-checked where possible. Nevertheless, the protocol seems to navigate toward fitting solutions. Ultimately, for the purpose of systematic coarse-graining, no correct solution is defined a priori, so as long as the optimizer can reduce the discrepancy between target and CG data systematically, the resulting parameterization could be useful. This challenge is particularly important when multiple rheological properties are to be measured and inaccuracies in parameters could have different effects on different properties.

Conclusion

We have described a novel protocol that leverages the properties of Bayesian inference to efficiently tune parameters for a polymer model to reproduce previously generated data. This work reduces the cost of developing models for which simulating time scales needed to measure experimentally relevant rheological properties is tractable. For this investigation, the target data set was generated using the same polymer model so that the accuracy of the identified parameters could be validated. This protocol could now be applied to data from more costly higher resolution models or experiments to identify appropriate parameters to represent a system using the cheaper coarse-grained model.

In this study, accuracy was measured as a function of the computational investment in exploring the parameter space, so the exploration was not terminated at a particular threshold accuracy. In a production context, one would likely identify the necessary precision relative to the target data, and halt the optimizer when sufficient precision is achieved.

The same broad framework could be executed with minor variations in the objective functions. The probability distributions for bonded monomers are an obvious choice for polymers, but different metrics could be used to evaluate the discrepancy between target and sample distributions. For instance, the correlation could be used instead of the RMSE. This would have the advantage of restricting the range of the objective to [-1,1], which could enable a simultaneous multi-objective optimization. Alternatively, the three optimization phases could be interwoven, updating each parameter once in turn to tune them in parallel. Another important consideration is the variety of available properties from which to derive an objective function. Previous work has considered properties such as The Debye-Waller factor, Young's modulus, and yield stress in addition to density (Giuntoli et al., 2020). While the temperature-dependent density seems to be sufficient as a top-down objective to capture a variety of material properties (Shireen et al., 2022), a systematic comparison of the efficiency of different objective functions and the trade-offs in accuracy of the various properties of interest would benefit the field of rheological simulations. The best case would be to find one objective function, possibly a multi-objective construction, that yields adequate accuracy on all properties, but perhaps different models will be necessary for applications that demand high precision for a particular property.

For some parameters investigated here, parameter ranges could be narrowed down significantly. Particularly, σ and l_0 can usually be read off from $P(l)$ and $g(r)$ data as long as $\sigma < l_0$ (a physically meaningful constraint that distance from one monomer to the next is not shorter than the monomer size). A systematic survey of the effect of individual parameters one at a time, with the other parameters set at estimated values could provide practical limits for the ranges without incurring the combinatorial cost of varying multiple parameters.

The parameter spaces explored in this investigation clearly exhibit some local optima which are distinct from the global optimum, especially when the sample data set is sparse, so

the ability of Bayesian optimization to escape local optima is critical to the success of this protocol. However, the ruggedness of the landscape seems to be limited to large scales. For instance, k_l=30 is not dramatically different than k_l=31, all other parameters being equal. A potential improvement could be to develop a heuristic for identifying when the basin of the global optimum has been discovered, and switch to a greedier search algorithm to refine the precision of the parameters.

We note in closing that the proposed framework could also bear utility for other classes of simulation than polymers. In many contexts, there are models with parameters that do not map analytically to more realistic data. When such models have several parameters, or rugged parameter landscapes, determining appropriate parameters can be a combinatorially difficult problem. An adaptation of the workflow presented here could be useful in such cases.

Funding This research is supported by the Commonwealth of Australia as represented by the Defence Science and Technology Group of the Department of Defence through the multi-disciplinary materials sciences stream of the Next Generation Technologies Fund. This research was partially supported by the Australian Government through the Australian Research Council Industrial Transformation Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Project ID IC200100009. Open Access funding enabled and organized by CAUL and its Member Institutions

Data availability All data that support the findings of this study are included within the article (and any supplementary files)

Declarations

Conflict of interest The authors declare no competing interests.

References

Agrawal V, Arya G, Oswald J (2014) Simultaneous iterative boltzmann inversion for coarse-graining of polyurea. Macromolecules 47(10):3378–3389

Bayramoglu B, Faller R (2012) Coarse-grained modeling of polystyrene in various environments by iterative boltzmann inversion. Macromolecules 45(22):9205–9219

Befort BJ, DeFever RS, Tow GM et al (2021) Machine learning directed optimization of classical molecular modeling force fields. J Chem Inf Model 61(9):4400–4414

Bull AD (2011) Convergence rates of efficient global optimization algorithms. J Mach Learn Res 12(10)

Chen L, Pilania G, Batra R et al (2021) Polymer informatics Current status and critical next steps. Materials Science and Engineering: R: Reports 144(100):595. https://doi.org/10.1016/jmser.2020.100595 https://www.sciencedirect.com/science/article/pii/S0927796X2030053X

De Ath G, Everson RM, Rahat AAM, et al (2021) Greed is good Exploration and exploitation trade-offs in bayesian optimisation. ACM Trans Evol Learn Optim 1(1). https://doi.org/10.1145/3425501

Dequidt A, Solano C JG (2015) Bayesian parametrization of coarsegrain dissipative dynamics models. J Chem Phys 143(8):084,122. https://doi.org/10.1063/1.4929557

Duan K, He Y, Li Y et al (2019) Machine-learning assisted coarse-grained model for epoxies over wide ranges of temperatures and cross-linking degrees. Mater Design 183(108):130. https://doi.org/10.1016/.matdes.2019.108130 https://www.sciencedirect.com/science/article/pii/S0264127519305684

Dunn NJH, Noid WG (2015) Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids. J Chem Phys 143(24):243,148. https://doi.org/10.1063/1.4937383

Foley TT, Shell MS, Noid WG (2015) The impact of resolution upon entropy and information in coarse-grained models. J Chem Phys 143(24):243–104

Fortuin V (2022) Priors in bayesian deep learning: A review. International Statistical Review 90(3):563–591. https://doi.org/10.1111/insr.12502 https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12502

Fröhlking T, Bernetti M, Calonaci N et al (2020) Toward empirical force fields that match experimental observables. J Chem Phys 152(23):230–902 arXiv:5001.1346

Giuntoli A, Hansoge NK, van Beek A, et al (2021) Systematic coarse-graining of epoxy resins with machine learning-informed energy renormalization. Npj Comput Mater 7(1)

Hajizadeh E, Garmabi H (2008) Response surface based optimization of toughness of hybrid polyamide 6 nanocomposites. Soft Matter 1:40–44

Hajizadeh E, Larson RG (2017) Stress-gradient-induced polymer migration in taylor–couette flow. Soft Matter 13(35):5942–5949

Hajizadeh E, Todd BD, Daivis PJ (2014a) Nonequilibrium molecular dynamics simulation of dendrimers and hyperbranched polymer melts undergoing planar elongational flow. J Rheol 58(2):281–305

Hajizadeh E, Todd BD, Daivis PJ (2014b) Shear rheology and structural properties of chemically identical dendrimer-linear polymer blends through molecular dynamics simulations. J Chem Phys 141(19):194,905

Hajizadeh E, Todd BD, Daivis PJ (2015) A molecular dynamics investigation of the planar elongational rheology of chemically identical dendrimer-linear polymer blends. J Chem Phys 142(17):174,911

Hansani W, Shireen Z, Iyer S et al (2022) A machine learning accelerated inverse design of underwater acoustic polyurethane coatings. Structural and Multidisciplinary Optimization 65:213

Hsu DD, Xia W, Arturo SG et al (2015) Thermomechanically consistent and temperature transferable coarse-graining of atactic polystyrene. Macromolecules 48(9):3057–3068

Huang H, Wu L, Xiong H et al (2018) A transferrable coarse-grained force field for simulations of polyethers and polyether blends. Macromolecules 52(1):249–261

Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. J Global Optim 13(4):455

Joshi SY, Deshmukh SA (2020) A review of advancements in coarse-grained molecular dynamics simulations. Mol Simul 47:786–803

Kanada R, Tokuhisa A, Tsuda K, et al (2020) Exploring successful parameter region for coarse-grained simulation of biomolecules by bayesian optimization and active learning. Biomolecules 10(3). https://www.mdpi.com/2218-273X/10/3/482

Korolev N, Luo D, Lyubartsev A et al (2014) A coarse-grained DNA model parameterized from atomistic simulations by inverse monte carlo. Polymers (Basel) 6(6):1655–1675

Liang Q, Gongora AE, Ren Z, et al (2021) Benchmarking the performance of bayesian optimization across multiple experimental materials science domains. npj Computational Materials 7(1):188. https://doi.org/10.1038/s41524-021-00656-9

Liu M, Oswald J (2019) Coarse-grained molecular modeling of the microphase structure of polyurea elastomer. Polymer (Guildf) 176:1–10

Liu P, Shi Q, Daumé H, et al (2008) A bayesian statistics approach to multiscale coarse graining. The Journal of Chemical Physics 129(21):214,114. https://doi.org/10.1063/1.3033218

Lyubartsev AP, Naômé A, Vercauteren DP, et al (2015) Systematic hierarchical coarse-graining with the inverse monte carlo method. J Chem Phys 143(24):243,120

McDonagh JL, Shkurti A, Bray DJ et al (2019) Utilizing machine learning for efficient parameterization of coarse grained molecular force fields. J Chem Inf Model 59(10):4278–4288. https://doi.org/10.1021/acs.jcim.9b00646

Moradzadeh A, Aluru NR (2019) Transfer-learning-based coarse-graining method for simple fluids: Toward deep inverse liquid-state theory. J Phys Chem Lett 10(6):1242–1250. https://doi.org/10.1021/acs.jpclett.8b03872

Ohkuma T, Kremer K (2020) A composition transferable and time-scale consistent coarse-grained model for cis-polyisoprene and vinyl-polybutadiene oligomeric blends. J Phys Mater 3(3):034–007

Pawar AA, Warbhe U (2021) Optimizing bayesian acquisition functions in gaussian processes. https://doi.org/10.48550/ARXIV.2111.04930

Plimpton S (1995) Fast parallel algorithms for short-range molecular dynamics. J Comput Phys 117(1):1–19

Prathumrat P, Sbarski I, Hajizadeh E et al (2021) A comparative study of force fields for predicting shape memory properties of liquid crystalline elastomers using molecular dynamic simulations. J Appl Phys 129(15):155–101

Reith D, Pütz M, Müller-Plathe F (2003) Deriving effective mesoscale potentials from atomistic simulations. J Comput Chem 24(13):1624–1636

Sestito JM, Thatcher ML, Shu L et al (2020) Coarse-grained force field calibration based on multiobjective bayesian optimization to simulate water diffusion in poly-ε-caprolactone. J Phys Chem A 124(24):5042–5052. https://doi.org/10.1021/acs.jpca.0c01939 , pMID 32452682

Shahriari B, Swersky K, Wang Z et al (2016) Taking the human out of the loop A review of bayesian optimization. Proceedings of the IEEE 104(1):148–175. https://doi.org/10.1109/JPROC.2015.2494218

Shell MS (2016) Coarse-graining with the relative entropy. Adv Chem Phys. John Wiley & Sons Inc, Hoboken, NJ, USA, pp 395–441

Shireen Z, Weeratunge H, Menzel A, et al (2022) A machine learning enabled hybrid optimization framework for efficient coarse-graining of a model polymer. npj Computational Materials 8(1):224. https://doi.org/10.1038/s41524-022-00914-4

Shireen Z, Hajizadeh E, Daivis PJ et al (2023) Linear viscoelastic shear and bulk relaxation moduli in poly (tetramethylene oxide)(ptmo) using united-atom molecular dynamics. Comput Mater Sci 216(111):824

Solomou A, Zhao G, Boluki S et al (2018) Multi-objective bayesian materials discovery: Application on the discovery of precipitation strengthened niti shape memory alloys through micromechanical modeling. Materials and Design 160:810–827. https://doi.org/10.1016/j.matdes.2018.10.014 . https://www.sciencedirect.com/science/article/pii/S026412751830769X

Wu J, Chen XY, Zhang H, et al (2019) Hyperparameter optimization for machine learning models based on bayesian optimizationb. J Electron Sci Technol 17(1):26–40. https://doi.org/10.11989/JEST.1674-862X.80904120. https://www.sciencedirect.com/science/article/pii/S1674862X19300047

Xia W, Song J, Jeong C et al (2017) Energy-renormalization for achieving temperature transferable coarse-graining of polymer dynamics. Macromolecules 50(21):8787–8796

Xiang J, Hajizadeh E, Larson RG et al (2021) Predictions of polymer migration in a dilute solution between rotating eccentric cylinders. J Rheol 65(6):1311–1325

Ye H, Xian W, Li Y (2021) Machine learning of coarse-grained models for organic molecules and polymers: Progress, opportunities, and challenges. ACS Omega 6(3):1758–1772. https://doi.org/10.1021/acsomega.0c05321 pMID: 33521417

Zhu G, Rezvantalab H, Hajizadeh E et al (2016) Stress-gradient-induced polymer migration: Perturbation theory and comparisons to stochastic simulations. J Rheol 60(2):327–343