The Journal of Chemical Physics

RESEARCH ARTICLE | JUNE 15 2023

Using classifiers to understand coarse-grained models and their fidelity with the underlying all-atom systems

Special Collection: Machine Learning Hits Molecular Simulations

Aleksander E. P. Durumeric ; Gregory A. Voth

J. Chem. Phys. 158, 234101 (2023)
https://doi.org/10.1063/5.0146812

AIP Publishing

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Using classifiers to understand coarse-grained models and their fidelity with the underlying all-atom systems

Cite as: J. Chem. Phys. 158, 234101 (2023); doi: 10.1063/5.0146812
Submitted: 16 February 2023 • Accepted: 26 May 2023 •
Published Online: 15 June 2023

Aleksander E. P. Durumeric[a] and Gregory A. Voth[b]

AFFILIATIONS

Department of Chemistry, Chicago Center for Theoretical Chemistry, James Franck Institute, and Institute for Biophysical Dynamics, The University of Chicago, 5735 S. Ellis Ave., Chicago, Illinois 60637, USA

Note: This paper is part of the JCP Special Topic on Machine Learning Hits Molecular Simulations.
[a]**Current address:** Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 12, 14195 Berlin, Germany
[b]**Author to whom correspondence should be addressed:** gavoth@uchicago.edu

ABSTRACT

Bottom-up coarse-grained (CG) molecular dynamics models are parameterized using complex effective Hamiltonians. These models are typically optimized to approximate high dimensional data from atomistic simulations. However, human validation of these models is often limited to low dimensional statistics that do not necessarily differentiate between the CG model and said atomistic simulations. We propose that classification can be used to variationally estimate high dimensional error and that explainable machine learning can help convey this information to scientists. This approach is demonstrated using Shapley additive explanations and two CG protein models. This framework may also be valuable for ascertaining whether allosteric effects at the atomistic level are accurately propagated to a CG model.

I. INTRODUCTION

Atomistic molecular dynamics (MD) has provided scientific insight into many problems.[1–5] Despite improvements in computing hardware, however, atomistic MD is still computationally limited in terms of the time- and space-scales it can access. These limitations have motivated the development of coarse-grained (CG) MD models that simulate a chemical system at a minimal resolution, aiming to reduce computational costs while maintaining quantitative accuracy.[6–16] The properties of these CG MD models are typically controlled via an effective Hamiltonian. For example, a CG MD model could simulate a solvated protein by propagating only the center of mass of each amino acid; the equilibrium distribution would then be controlled by a Hamiltonian defined at the resolution of these centers of mass. The behavior of these models can be divided into dynamic and thermodynamic properties. Thermodynamics here refers to long-time behavior related to the equilibrium distribution of the model and includes both issues related to estimating thermodynamic quantities, such as pressure, and averages of functions of microstates.[16,17] While accurately reproducing the dynamics of a reference system is an area of current interest,[18] the remainder of this article focuses on thermodynamic issues, and more specifically, the configurational distribution produced by a CG model. We limit our discussion to systems in the canonical ensemble and focus on the configurational portion of this effective Hamiltonian (which we term the CG force-field).

There are many ways to create the force-field that characterizes a CG model.[6–17] These various approaches often result in different equilibrium configurational distributions. Top-down force-field parameterization methods typically aim to reproduce observables that are coarser than the effective Hamiltonian, such as partition coefficients or interfacial tension. These low dimensional observables are obtainable from either experiment or reference all-atom MD simulation. In contrast, bottom-up methods tend to require molecular trajectories from a reference atomistic simulation, which are mapped to the resolution of the CG model and used as a high dimensional target for parameterization.[16,19,20] The variety of possible parameterization techniques makes it valuable to characterize how a proposed CG force-field approximates a reference atomistic simulation. While comparing the distribution of chosen low

dimensional observables is straightforward, it is also desirable to perform this comparison at the resolution of the CG model itself (i.e., at the resolution of its configurational phase space), as said low dimensional observables may not fully describe how the CG model configurationally differs from the given reference data.

However, while a CG model is intrinsically coarser than the atomistic model it represents, it is still often highly dimensional. For example, the CG molecules in this paper are relatively small but easily reach 36 configurational dimensions, which is well beyond what generic data visualizations (e.g., scatter plots or histograms) can communicate to humans. These models are often still amenable to visualization as groups of particles in three dimensional space[21] as the systems preferentially occupy a small portion of the possible phase space (for example, a protein does not dissociate into its constituent atoms during simulation—its behavior is dictated by its primary and higher order structures). However, while visual inspection can detect some differences between two simulations, it can be difficult to discern local but important discrepancies that may not be salient (e.g., bond lengths or tetrahedral order parameters). Alternatively, established dimensional reduction techniques[14,15,22] can be used to individually summarize the emergent behavior present in the model and reference data; unfortunately, these approaches are not typically optimized to isolate differences between two datasets and may similarly miss discrepancies present in the CG model.

Discerning the error present in a CG simulation has clear connections to the parameterization of a CG force-field, as this procedure is typically designed to optimize a chosen measure of error. In the case of top-down parameterization, optimization involves low dimensional statistics that are readily interpretable by computational scientists; however, these observables are coarser than the resolution of the effective Hamiltonian. In the case of bottom-up parameterization, while the considered resolution is ideal, said optimization uses specific (and often opaque) computational algorithms to optimize the Hamiltonian such that its high dimensional configurational statistics approximate those of the reference model. Collectively, these existing approaches do not make it straightforward to understand how the full high dimensional configurational behavior produced by a CG force-field deviates from an atomistic reference, even when appropriate validation data are available.

This article will provide a way to compare two equilibrium samples from differing high dimensional free energy surfaces. The analysis presented here does not specify a particular process to be used to generate these samples. However, the examples we will study have been created using bottom-up coarse-graining techniques and we will borrow terminology from the bottom-up coarse-graining literature to express our ideas: for example, we will refer to the function that maps each atomistic configuration to its CG counterpart as the CG map, and we will refer to the ideal effective force-field perfectly reproducing the mapped configurational atomistic statistics as the many-body potential of mean force (many-body-PMF). Furthermore, as the techniques we describe provide a way to compare two samples from differing high dimensional free energy surfaces, these approaches are especially pertinent to bottom-up strategies as a mapped atomistic reference sample is already required for parameterization.

As noted previously, the high dimensional nature of the data produced by CG models makes it difficult to directly analyze and validate their full configurational behavior. However, when parameterizing *atomistic* force-fields using quantum mechanical reference data, computational scientists can often compare the individual energies or forces produced by the reference method to those produced by the proposed atomistic force-field and use this per-configuration error (or per-atom decompositions of this error) to identify problematic areas of phase space (e.g., Bartok *et al.*[23]). This is difficult to do with CG force-fields as the corresponding reference data typically do not have energies or forces: a point-wise evaluation of the many-body-PMF is often not available, and only a noisy estimate may be present[24] for its gradient.[25] An attractive workaround is to train two CG force-fields, one of which has more complex many-body interactions. Treating the higher-order potential as if it were the true many-body-PMF allows one to estimate the conditional free energy and force errors present in the less complex force-field. However, this approach requires multiple CG force-fields to be created and exceedingly high-order interactions to be considered.

In lieu of training multiple CG force-fields, the error analysis proposed in this article uses classification to determine which portions of a CG model trajectory and reference atomistic trajectory differ from each other. The algorithm is trained to predict whether an arbitrary CG configuration is more typical of the model or reference configurational distribution. As shown in the supplementary material, classification performed in this manner can be mathematically understood as a two-ensemble variational statement that estimates the offset arithmetic difference between the studied CG force-field and the many-body PMF. This information allows analyses that would typically be reserved for atomistic force-fields to be performed on their CG counterparts. The proposed approach requires both a reference trajectory and a trajectory generated using the CG force-field. No secondary higher-order CG force-field is trained to obtain this estimate of error. For the systems considered in this article, incorporating high-order interactions is straightforward in the proposed classification-based approach.

While this classification-based analysis alone allows a computational scientist to systematically and variationally characterize problematic configurations based on force-field error, the second contribution of this work is the realization that this classifier-based approach, when combined with methods from explainable machine learning/artificial intelligence (ML/AI) (collectively referred to as XAI in this article), allows one to convey force-field errors in an alternative manner. XAI is a subfield of AI under active development (for an overview of the corresponding definitions, see Refs. 26–31). The issue of algorithmic transparency is not new (see, e.g., Refs. 32 and 33); however, as computational decisions become more common in everyday life, an increasing amount of scrutiny has been placed on providing justifications for the output of automated systems.[34,35] This is required for a variety of reasons, including regulatory compliance, ethical analysis, debugging, or further comprehension of the data used to train the algorithm. For the purposes of this article, we divide the algorithms in this field into two categories: transparent (or interpretable) models and those with *post hoc* explanations. Transparent models are algorithms that, once trained, can be intrinsically understood by a particular audience; examples include shallow decision trees or rule lists. Methods of this type are generally simpler than more opaque algorithms such as deep neural networks. *Post hoc* explanations, on the other hand, are methods that digest and summarize information from an already optimized

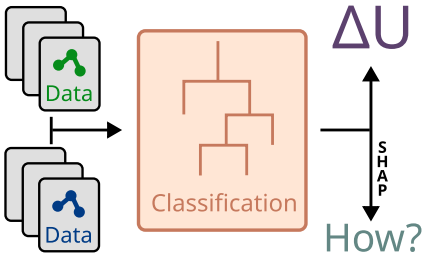

FIG. 1. The proposed analysis process. Configurations are generated for the model and reference systems and compared via classification. This produces an estimate of the microscopic error (ΔU) and interpretations of the features causing said error [Shapley additive explanations (SHAP) values].

external algorithm. While transparent models are intrinsically interpretable, explainable models are those which have additional *post hoc* explanation methods to provide reduced representations of the knowledge present in the trained model.

Using XAI to extract knowledge present in a classifier trained to estimate force-field error provides an explanation of the errors present in the original force-field. The nature of the insight provided depends on the particular methods from XAI we adopt. To illustrate this concept, we take a particular modern *post hoc* explanation technique, Shapley additive explanations (SHAP values),[36–38] and demonstrate how they can isolate which physical aspects of two CG peptides or proteins (dodecaalanine and actin monomer) are problematic for specific CG force-fields. The goal of this paper is not to model these two proteins accurately; instead, it is to show that non-ideal force-fields can be detected and characterized. The behavior of the produced SHAP values is analyzed as a function of pairwise distances and collective variables derived using nonlinear dimensional reduction. These insights are shown to be useful when considering the behavior of CG force-fields and provide a conceptual basis for future bottom-up error analysis through classification. A diagram outlining this analysis is shown in Fig. 1: Configurations are generated for a model and compared to those from a reference dataset via classification to obtain an estimate of the microscopic error (referred to as ΔU) and a characterization of observed differences.

The remainder of this article proceeds as follows: Sec. II discusses the theory behind the classification-based approach, including the method selected for producing explanations; Sec. III discusses the concrete methodological details of the classification and related CG models; Sec. IV summarizes the results of applying the proposed method; and Sec. V discusses the future applications and consequences.

II. THEORY

The techniques presented in this article compare two high dimensional free energy surfaces, which we refer to as $U_{\mathrm{PMF}}(x)$ and $U_{\mathrm{FF}}(x)$, where x represents a sample on the free energy surface. Via the canonical ensemble, these free energy surfaces define probability densities as

$$p_{\mathrm{PMF}}(x) = Z_{\mathrm{PMF}}^{-1} e^{-\beta U_{\mathrm{PMF}}(x)} \tag{1}$$

and

$$p_{\mathrm{FF}}(x) = Z_{\mathrm{FF}}^{-1} e^{-\beta U_{\mathrm{FF}}(x)} \tag{2}$$

with Z_{PMF} and Z_{FF} defined as the integral of $e^{-\beta U_{\mathrm{PMF}}}$ and $e^{-\beta U_{\mathrm{FF}}}$ over all possible x, and β defined as the inverse temperature scaled by the Boltzmann constant. These free energy surfaces could be defined by any collective variables. However, in the analysis that follows, we will assume that x is the configurational variable that comprises the domain of the CG force-field (typically the CG Cartesian coordinates); in this case, U_{FF} refers to the configurational CG force-field and U_{PMF} refers to the many-body-PMF. The remainder of this section provides an intuitive explanation of how classification relates to estimating the difference in conditional free energies; a precise mathematical connection may be found in the supplementary material.

Classification is the machine learning task of predicting the most probable class, or label, associated with a data point.[33] For example, one might want to predict the particular number present in a picture of a handwritten digit. Algorithms in supervised classification are trained to complete this task by studying an already labeled dataset: in this example, said data would be a set of pictures that have had the correct number already associated with each picture. In certain learning contexts, various samples may not have a clear correct class. For example, certain handwritten digits may be so messy that only the original writer knows the digit truly intended. In order to naturally adapt to ambiguous samples, classifiers can be designed to output a guess of the probability of each possible class.[39–45] Focusing on the case where we only have two possible classes to predict (for example, if we were only considering pictures of 4 or 5), this probabilistic estimate can be quantified by a single real number for each sample between 0 and 1.

This probabilistic approach can be used to describe the distributions typical to such a classification task. Again, focusing on classifying a picture as containing a 4 or 5, we first define the distribution of all possible pictures we could consider (i.e., containing 4 *or* 5); this characterizes the fact that certain images (such as one comprised of random pixels) are not typical of pictures of 4 nor 5. We refer to this overall probability density as $M(x)$. Then, for each possible picture, we refer to the conditional probability that a specific picture contains a 4 as $\eta(x)$; the probability of 5 is then $1-\eta(x)$. Together, $M(x)$ and $\eta(x)$ define a joint probability density over pictures and possible classes. A classifier that is directed to produce probabilities attempts to estimate η on samples produced from M.

This probabilistic description can be reframed by considering the pictures that are characteristic of each class individually. This would involve first considering the probability distribution of pictures that have a ground truth of containing 4 and then those with a ground truth of 5. The ambiguous pictures above will be then described as areas in which these two distributions (termed the class conditional distributions) overlap. When combined with the overall balance between the classes, the class conditional description in this paragraph and the example-conditional viewpoint in the previous paragraph are equivalent specifications of a classification problem. The classification problems constructed in this paper have equal overall populations of each label by design, and our expressions make this assumption throughout the article.

Classification is used in this work by setting these class conditional distributions to p_{FF} and p_{PMF} (see the supplementary material). Using the connections between the two views of classification, η then takes the following form:

$$\eta(x) = \frac{p_{PMF}(x)}{p_{PMF}(x) + p_{FF}(x)}. \tag{3}$$

A calibrated classifier can be used to approximate η using only samples from each free energy surface. Critically, this formulation implies that it is not necessary to know the values of U_{FF} or U_{PMF} evaluated on samples in order to estimate η on each sample. This approach is implemented by the following algorithm:

1. Generate N samples from U_{PMF}
2. Generate N samples from U_{FF}
3. Label all samples from U_{PMF} with "A"
4. Label all samples from U_{FF} with "B"
5. Train a calibrated classifier to predict "A" or "B" for each sample via η

η, when combined with β, can be transformed into an offset pointwise difference between U_{FF} and U_{PMF}, which we refer to as ΔU,

$$\Delta U(x) := k_b T \log \frac{\eta(x)}{1-\eta(x)} = U_{FF}(x) - U_{PMF}(x) + k_b T \log \frac{Z_{FF}}{Z_{PMF}}. \tag{4}$$

When U_{FF} corresponds to the configurational distribution of a CG simulation and U_{PMF} corresponds to the many-body-PMF, we typically can evaluate U_{FF} on an arbitrary configuration but are unable to evaluate U_{PMF}. Additively combining ΔU and U_{FF} provides an estimate of U_{PMF} that would be exact with a perfect classifier and learning procedure. In practice, limited data and imperfect classification algorithms make this estimate only approximately correct. Using ΔU to form an additive update to U_{FF} has been performed in the past.[46,47] This estimation duality, combined with the realization that classification is a variational process with respect to the proposed hypothesis, establishes that classification can here be viewed as a variational technique to estimate the many-body-PMF using U_{FF} as a reference (see supplementary material). Note that ΔU is defined here such that there is no unknown global offset.

As ΔU precisely characterizes the pointwise difference between two free energy surfaces, evaluating it at a particular configuration quantifies the difference in the conditional free energies at that point. This property makes ΔU a useful descriptor of the "microscopic" error present in a model. Areas of high ΔU imply that one ensemble has considerably more population in said area than the other ensemble, while a largely negative ΔU implies the opposite. Equivalently, when used as a structural collective variable (CV), ΔU describes which configurations occupy areas of high distributional overlap and which are specific to either free energy surface. In the context of U_{FF} approximating U_{PMF}, linking ΔU to other intuitive structural variables can characterize what errors a CG model is committing. For example, if ΔU is negative whenever a particular bond distance is small, this implies that the CG model (U_{FF}) is over-stabilizing small bond distances. The advantage relative to direct visualization of configurational statistics from either ensemble is that configurations occupying areas of high distributional overlap may be discarded prior to analysis. This approach, however, is still relatively tedious: it again requires human analysis of the resulting configurations, incurring all of the difficulties discussed in the introduction. Furthermore, as demonstrated in later sections, ΔU may not correlate with any intuitive physical feature of the system under study. An appealing alternative is to understand the algorithmic form of ΔU itself: for example, if it is a linear function, its learned coefficients may offer insight. However, if U_{FF} is composed of low order n-body terms while U_{PMF} contains higher-order terms, ΔU will contain higher-order terms as well, and it may be difficult for interpretable models to provide a good estimate of ΔU. In this article, we use techniques from XAI to overcome this difficulty and extract configurational information from a complex estimate of ΔU.

It is important to note that the classification performed provides a variational estimate to ΔU and not ΔU itself. The quality of this approximation is a function of the algorithm used and the amount of data available; if the approximation is poor, the provided estimate of ΔU may be heavily biased. We note that the classifiers used in the current context regularly achieve test accuracies of 80%–95%, suggesting they are able to discern clear patterns in the provided data. Nevertheless, it is currently difficult to assess this possible discrepancy. We hope the preliminary success shown in this manuscript motivates studies targeted at quantifying this source of error.

A. Feature attribution

XAI includes a large variety of methods. This article will focus on the use of a single method from this field: SHAP values. SHAP values are a feature attribution method[26,32,33] with a strong mathematical underpinning. Feature attribution methods, or feature importance measures, provide a quantification of how informative a feature, or particular input variable, is to an algorithm. Some feature attribution methods are global, meaning that stated feature values are related to the aggregate behavior of the classifier over the entire dataset. Other feature attribution methods, such as SHAP values, are local: every prediction made by a classifier can be associated with a particular set of SHAP values. When estimating ΔU, this means that a prediction of a large positive or negative ΔU can be analyzed to determine which features lead the classifier to that conclusion.

The classification examples in this article quantify the error present in various CG models of proteins. The classification algorithm is trained on the ordered distance matrix derived from each configuration. As a result, applying a feature attribution method to explain ΔU isolates which pairwise distances are most connected to the estimated error and in doing so clarifies which configurational aspects of the protein are not reproduced by the CG model.

B. Shapley and SHAP values

SHAP values are based on Shapley values[26,48] from cooperative game theory. We first explain Shapley values and how they relate to classification and then provide a description of SHAP values.

1. Shapley values

Shapley values are a method to fairly distribute a reward among a group of individuals. Suppose a group of five scientists decides to create a product to bring to market. These five people have joined

together because their individual knowledge, when combined, produces a better product than they could individually. However, suppose one of the five people has knowledge that is vital to the product: if they were not present, the total amount of profit would be greatly diminished. In contrast, the expertise of the remaining four people is largely, but not completely, shared. As a result, losing one of those four people would reduce the possible profit but would not do so substantially. In this situation, how should the profit be fairly divided among the scientists? Allocation in these circumstances can be addressed by Shapley values.

The central calculation needed to define Shapley values is the ability to estimate the reward in the absence of some of the individuals in the group. Suppose the five people present are referred to as A, B, C, D, and E. We then denote the reward when everyone is present $K(\{A,B,C,D,E\})$. In order to calculate Shapley values, we must be able to calculate, as an example, $K(\{A,B,D,E\})$: the reward had individual C not been present. Shapley values then consider growing the number of present individuals incrementally, such as the (ordered) sequence $K(\{B\})$, $K(\{B,A\})$, $K(\{B,A,D\})$..., and associating the term with individual D $K(\{B,A,D\})$–$K(\{B,A\})$: the incremental increase that was seen when D was added in this particular sequence. The Shapley value averages over all such sequences of adding people to a group. Mathematically, the Shapley value for player i is defined as follows:

$$\varphi_i = \frac{1}{n!}\sum_R K\left(P_i^R \cup \{i\}\right) - K(P_i^R), \tag{5}$$

where n is the number of individuals, R iterates over all possible *orders* (not subsets) of players, and P_i^R is the subset of individuals that precedes player i in that particular R. It is important to note here that this sum is over all possible orderings, as where K only depends on the members present, not the order in which they were added. This calculation must be performed for each individual (or player) for whom we wish to calculate a Shapley value.

Shapley values satisfy a number of intuitive mathematical properties[36,48,49] and in some cases are the only allocation method that does so. The most important property of the current application is that summing together the Shapley values for all players provides the original reward when the entire group is present.

2. SHAP values

The connection between Shapley values and feature attribution is made by considering every individual prediction made by a classifier in a game in which each feature is a player. The output of the game, in analogy with the total profit in Subsection II B 1, is the numerical prediction of the classifier. However, one important detail is absent when considering feature attribution: what does it mean for a feature to be *missing*? It is possible in some cases to retrain a model with only a subset of the original features;[50] however, the number of models required quickly becomes infeasible. Instead, Shapley Additive Explanation (SHAP) values train a single model and average the full model's predictions over missing variables to represent missing features.[36] For example, consider the hypothetical classifier $f(w,x,y,z)$ with four input variables. Suppose we wished to calculate the SHAP value of w for configuration (w_0,x_0,y_0,z_0): this would include calculating $f_{wz}(x_0,y_0)$ where w and z are "missing." SHAP values dictate that $f_{wz}(x_0,y_0) := \int f(w,x_0,y_0,z)p(w,z|x_0,y_0)dwdz$, where $p(w,z|x_0,y_0)$ is the distribution of w and z conditioned on $x = x_0$ and $y = y_0$. Some implementations of SHAP values, instead, use the marginal expectation value, some label the choice of the marginal expectation as an approximation, and some argue that the marginal value is the correct one to use from an interventional perspective.[36,50,51] The implementation and physical interpretation presented in this paper use the conditional expectation value. With this definition of "missing values," Shapley values are applied as before to give SHAP values for each feature. The efficient algorithmic calculation of these values for arbitrary classification techniques is far from trivial but is possible for tree ensembles such as those used in this article.[38]

Despite the abstract description above, SHAP values can be physically interpretable in the current study. All the examples in this paper perform classification using distance matrices as input. The signal additively explained by the given SHAP values corresponds to ΔU. The individual terms in Eq. (5) can be understood as follows: $C\left(P_i^R \cup \{i\}\right)$ corresponds to the mean ΔU found when holding the distances specified by $P_i^R \cup \{i\}$ constant and letting the remainder of the protein freely explore the canonical ensemble, and $C\left(P_i^R\right)$ is the same but letting distance i also freely explore (SHAP values include a single global offset in their additive explanations; however, this distinction does not matter for applications presented in this article). In this way, SHAP values isolate which parts of a specific protein configuration contribute to its specified ΔU. The same idea can be applied to a feature set of an arbitrary free energy surface: the system is allowed to explore conditional to the given features under investigation.

SHAP values provide a real number for each feature and configuration considered. In this way, they can be considered a new set of descriptors for each protein configuration. This set of descriptors has equal dimensionality as the original dataset (the pairwise distances characterizing the protein configuration) and may seem to provide little advantage relative to the original distances. However, three properties of SHAP coordinates are more desirable than the original distances:

1. SHAP values additively relate to ΔU.
2. SHAP coordinates are of the same scale for each configuration and feature. As a result, the relative importance of an inter-domain distance and an intra-domain distance can be directly compared in SHAP space.
3. SHAP values individually reflect many-body correlations in the original data. As a result, quickly inspecting the individual distributions of SHAP values can produce insight into problematic aspects of U_{FF} in the original coordinate system.

In order to summarize the types of error present in our CG molecular trajectories, we may additionally reduce the dimensionality of the generated SHAP values to produce a lower dimensional set of CVs. Collectively, these analyses may suggest the source of microscopic errors in the CG model's force-field.

It is important to note that SHAP values using conditional expectations, such as those in this article, assign feature attributions using both correlations present in the learning distribution (here, the combined model and reference ensembles) and information in ΔU.[50,51] For example, consider the more generic case of the function $f(x,y) = 2x$ analyzed over a distribution of x and y where x and

y are highly correlated. Conditional SHAP values will assign importance to both x and y. Informally, this is because because knowing that y is large also implies that x is large due to the high level of correlation: y contains information about x. As a result, it is difficult to infer the structure of a hypothetical f using said SHAP values. Other feature attribution methods based on Shapley values circumvent this limitation at the cost of considering unrepresentative samples;[50] we leave investigating these alternative methods to a future study.

While this article uses tree ensembles for classification and SHAP values for interpretation, it is important to note that the concepts presented apply more broadly. Estimation of ΔU is applicable to any classifier that uses a proper loss,[40,42] ranging from other generic classifiers such as neural networks or logistic regression to variational approaches based on low-body-order expansions to ΔU. Each of these different approaches in turn may have a different degree of intrinsic interpretability and explanation methods available (which may not include SHAP values). The performance of each classification approach, as well as the utility of the related explanation techniques, represents a new direction for research analogous to the design of force-field bases.

III. METHODS

The methods used in this article are implemented in a publicly available Python module found at https://github.com/alekepd/ClassE, including an interactive notebook[52] and command line executable. The classification was performed using the DART (boosted decision trees with dropout[53]) algorithm in the Light Gradient Boosting Machine (lightgbm[54]) library. Hyperparameters used may be found in the supplementary material. Further analysis and visualization used the theano,[55] numpy,[56] scikit-learn,[57] umap,[58,59] pandas,[60] ggplot2,[61] data.table,[62] and pracma[63] libraries. The dimensional reduction was performed on SHAP values using principal component analysis[39] (PCA) combined with the Uniform Manifold Approximation and Projection (UMAP) method;[58,59] details may be found in the supplementary material. This technique was selected due to observed computational efficiency and separation of clusters. Classifiers trained to distinguish the difference between model and reference ensembles at a 12 site resolution regularly achieved accuracies of above 95% for both actin and dodecaalanine; classifiers trained at the four site resolution achieved accuracies of ~80% (see supplementary material).

Conclusions made through CVs (both ΔU and those generated via dimensional reduction) were stable to choices of hyperparameters. Estimated pointwise free energy differences at the extremes of ΔU were found to be sensitive to choices of hyperparameters. This is expected: large absolute differences in ΔU imply that a comparison is being made at a location with very low configurational density in one ensemble; the precise level of population is difficult to accurately estimate without using enhanced sampling. As a result, if one wishes to make a quantitative comparison between models based on the pointwise free energy differences in regions of poor overlap, it is likely that enhanced sampling methods must be used. In order to avoid the effect of outlying values of ΔU on reported aggregate statistics, medians are used instead of means, and box plots are shown without outliers; for visualization of such outliers, see supplementary material.

A. Dodecaalanine

Dodecaalanine (DDA) is a short polypeptide that adopts a variety of conformations in solution: a hairpin like conformation, a helical conformation, and an extended conformation (see, for example, Rudzinski and Noid[64]). DDA was simulated at the solvated atomistic resolution using Amber18[65] and the Charmm36m[66] force-field. Samples were extracted every 50 ps from a 5.1 μs trajectory propagated using a 2 fs timestep in the constant NVT ensemble at 303 K using a Langevin[67] thermostat with a damping parameter of 0.5 ps. Additional details on atomistic simulations may be found in the supplementary material. Radius of gyration and Q-helicity were used to quantify the behavior of DDA; their formulation can be found in Rudzinski and Noid.[64] Q-helicity quantifies the similarity of a given configuration to a helix: 0 corresponds to no helical character, while 1 corresponds to a completely helical polypeptide. Similar CVs are present in enhanced sampling studies, see Piana and Laio[68] and Prakash *et al.*[69]

DDA was modeled using five different CG force-fields at the resolution of one CG site per amino acid (Fig. 2). Each force-field was composed solely of pairwise spline interactions at the bonded and nonbonded level and parameterized using relative entropy minimization.[20] Sites adjacent along the backbone were connected via bonds; sites separated by a single site were connected via an additional bond to emulate an angle potential. Within each type of interaction (bonded, angle-bonded, and nonbonded), a unique interaction type was defined for each possible pair of site types. Each model discussed can be considered as extending the model before it. The first model was composed of a single site type. The pairwise nonbonded interactions were set to be constant, i.e., nonbonded pairwise forces were uniformly zero. This resulted in two unique pairwise interactions (one bonded and one angle bonded). The second model was identical to the first model but included pairwise

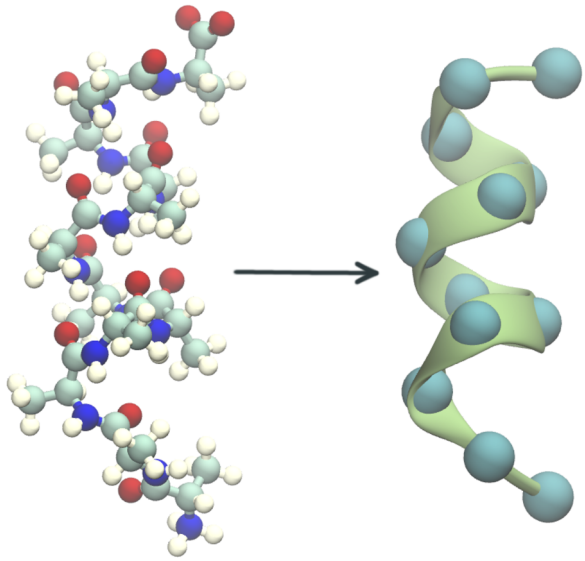

FIG. 2. The map used to coarse-grain dodecaalanine in the current study. Each amino acid was mapped to a single CG site via a center of mass mapping.

nonbonded interactions. The third model included a single additional collective site type for the termini, resulting in two site types total. The fourth extended the model by considering the site types at each terminus to be distinct, resulting in three site types. The final model was extended to five CG site types by providing additional unique site types to each CG bead adjacent to a terminal bead. Additional details are found in the supplementary material. Only a subset of these models is analyzed in certain sections for brevity.

B. Actin

Actin monomer (G-actin) was simulated at the solvated atomistic resolution using GROMACS 5.1.4.[70] Equilibration details may be found in Hocky *et al.*[71] Production simulations were performed for 1 μs using CHARMM27+CMAP[72] in the constant NPT ensemble at 310 K and 1 bar using the stochastic velocity rescaling thermostat[73] with a coupling parameter of 0.1 ps and a Parrinello–Rahman barostat[74] with a coupling parameter of 2 ps. Samples were collected every 100 ps. The protein was observed to maintain its tertiary structure throughout the simulation. Additional information on atomistic simulations may be found in the supplementary material.

Actin was modeled at a 12 site and four site CG resolution using the configurational map found in Saunders and Voth.[75] Briefly, sites indexed 1–4 represent actin's four main subdomains, which are approximately arranged at the four corners of a square; site 9 represents the nucleotide adenosine diphosphate (ADP) situated at the center of these four subdomains, and site 5 represents the D-loop, a semistructured region connected to site 2. The map is characterized in Fig. 3 (adapted from Saunders and Voth[75] with permission). All CG models of actin are harmonic networks parameterized using the heteroENM methodology described by Lyman *et al.*[76] Two different elastic network models were created: one model at the full 12 site resolution and the other using only sites indexed 1 through 4. In order to better understand the model error, the samples produced by the 12 site model were additionally mapped to the four site resolution for comparison to the four site model for specific analyses. Note that the utilized elastic framework creates force-fields that preserve tertiary structure by design.

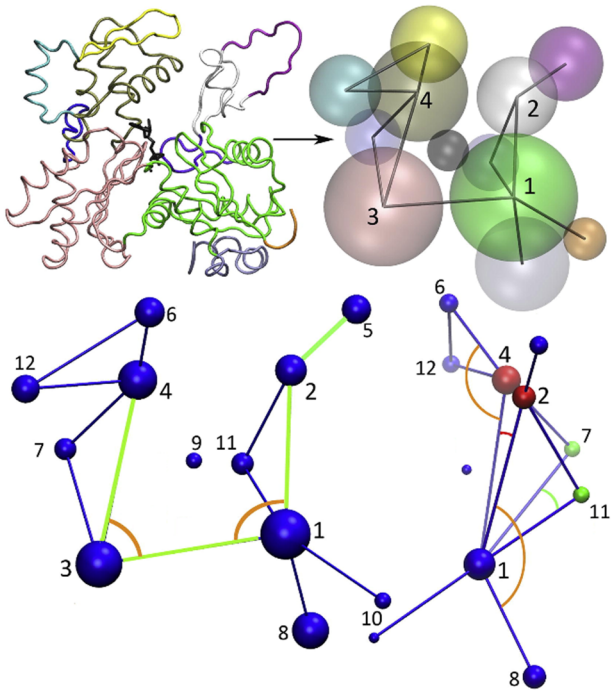

FIG. 3. The map used to coarse-grain actin in the current study. Each set of atoms was mapped to a position via a center of mass mapping. CG models of actin were simulated and modeled at both the resolution of all 12 sites presented and at the resolution of sites 1–4.

IV. RESULTS

The ability to estimate microscopic error using classification creates multiple avenues for the characterization of CG models. In Subsection IV A, we compare the behavior of DDA and actin models to reference data using only the transformed output of a classifier (ΔU), demonstrating the general properties of classification-based analysis. We then continue in Subsection IV B by discussing the inherent difficulties in ΔU-based analysis and demonstrating how SHAP values may be used to overcome these challenges.

A. ΔU-based analysis

Classification performed between a force-field and a reference estimates the microscopic error present via ΔU. In contrast, traditional configurational validation of CG force-fields focuses on low dimensional free energy surfaces defined on physically meaningful collective variables. The overlap observed in these low dimensional free energy surfaces may also be viewed as a quantification of error; however, these two measures of error do not necessarily align.

To demonstrate these discrepancies, DDA, a polypeptide that transiently forms a helix in solution, was modeled using a variety of CG force-fields using increasingly complex pairwise force-field basis sets (Sec. III A). Free energy surfaces along Q-helicity and radius of gyration, two CVs that characterize the transient folding behavior of the polypeptide, are shown for a variety of models and the reference data in Fig. 4. While it is clear that no CG force-field achieves states of high Q-helicity (signifying helix formation), the observed difference between the various models is minimal.

In contrast, when each of these models is analyzed via classification, the increasingly complex force-field bases reduce microscopic error. Box plots of the microscopic error (ΔU) for each comparison of the DDA model to reference data are shown in Fig. 5; the medians of the distributions, signified by a horizontal line in each box, trend toward zero. Note that configurations from the CG model trajectory are preferentially associated with negative values of ΔU as said configurations are over stabilized by the CG force-field, causing the sum in Eq. (4) to become negative. An analogous relationship associates configurations from the reference trajectory to positive values of ΔU. The discrepancy between our physical CVs and microscopic error may also be visualized by plotting ΔU as a function of said CVs (Fig. 6), where is clear that the selected CVs do not effectively disentangle configurations based on their microscopic error (ΔU). In contrast, using ΔU itself as a collective variable naturally does distinguish states based on their microscopic error. While the microscopic error is an important error metric in itself, the trends of ΔU

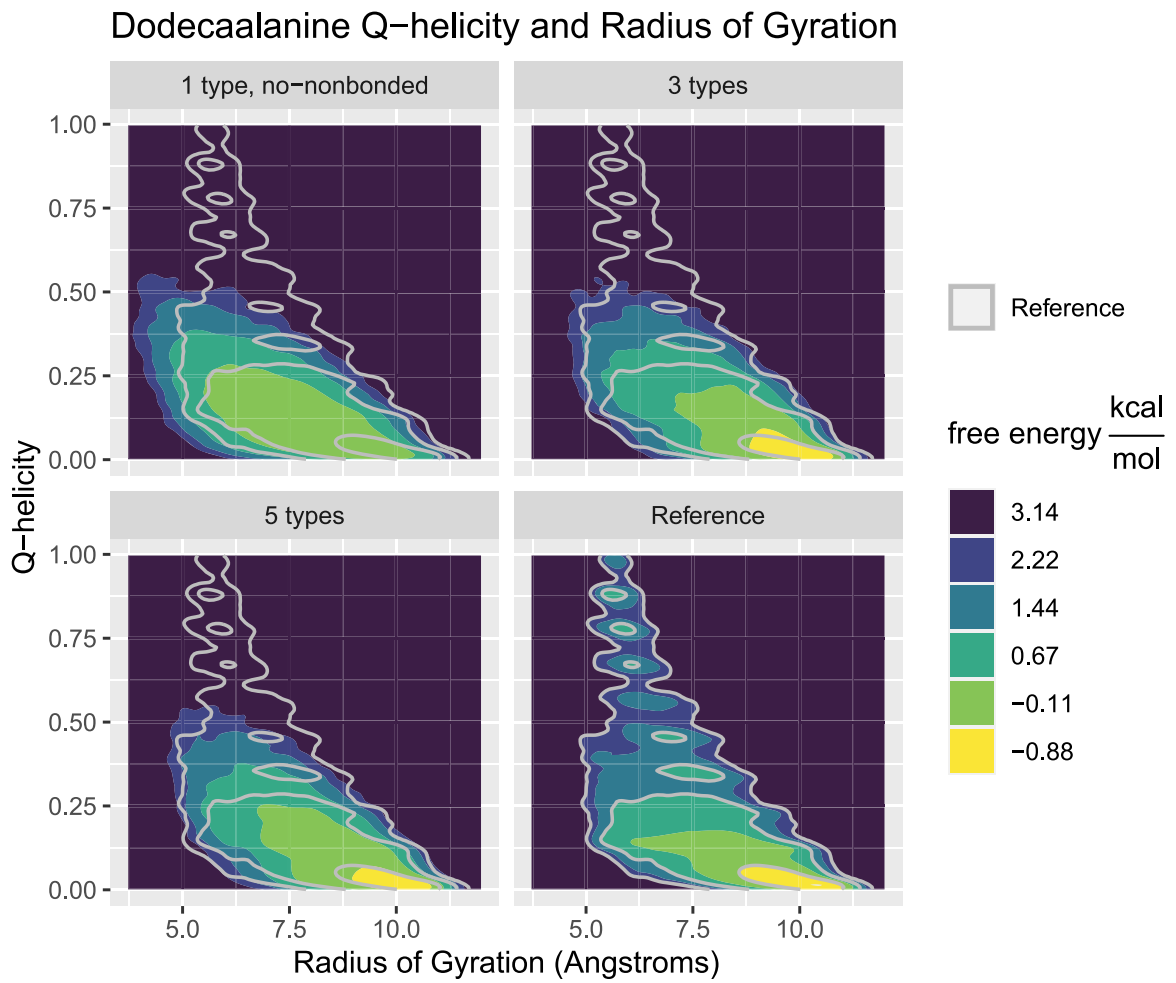

FIG. 4. Free energy surfaces of the radius of gyration and Q-helicity for various models of DDA and the reference distribution. The gray overlay (and filled regions in the lower right panel) is given by the reference density and is present as a visual guide. Clear differences exist between the models and the reference distribution, but little difference can be seen between various CG models.

visualized in Fig. 5 provide little physical insight; however, we defer extracting conclusions from ΔU until Sec. IV B.

Similar to traditional CV analyses, the classification-based analysis may be performed after reducing the resolution of the model or reference data. This procedure allows the resolution of a proposed CG model to be validated in novel ways. To demonstrate this capability, an actin monomer was modeled using heterogeneous harmonic networks at resolutions of 12 sites and four sites, with the four site model defined to be a subset of the CG beads present in the 12 site model (Sec. III B). The 12 site model distinguishes between many parts of the protein, such as nucleotides and disordered regions, while the four site model only considers the movement of the main four subdomains. When compared to the reference data, the 12 site model exhibits larger absolute values of ΔU than the four site model (Fig. 7). However, when the 12 site model is mapped to the resolution of the four site model for analysis of its microscopic error, it exhibits less microscopic error than the four site model; i.e., the microscopic error present in the 12 site model is not evident in the motion of sites 1–4. If a hypothetical application of this model requires only the correlations of the four main subdomains, the 12 site model may provide a higher level of accuracy than the model that only preserves only the four sites of interest. The physical insight underlying the trends observed in Fig. 7 is discussed in Sec. IV B.

B. SHAP-based analysis

While ΔU allows one to probe microscopic errors, it does not necessarily directly translate into scientifically useful information. In the context of the models studied in Sec. IV A, it is desirable to understand how the addition of force-field complexity improves the

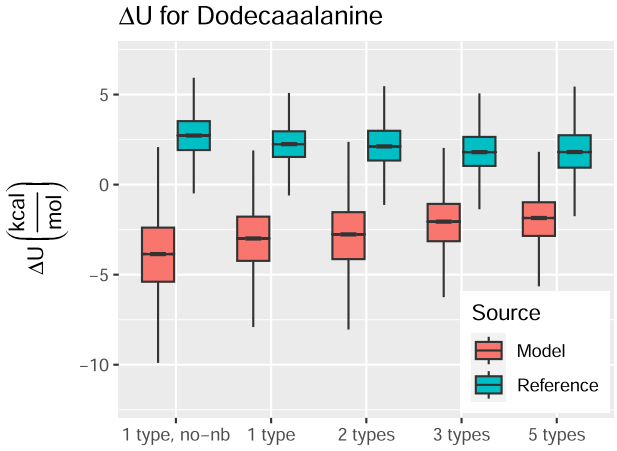

FIG. 5. Box plots of ΔU for multiple models of DDA. Each model presents its own samples, and the reference samples are projected along its ΔU. The median is marked by the center line in each box. Note that different models have different forms of ΔU, changing the shape of the projected shared reference distribution.

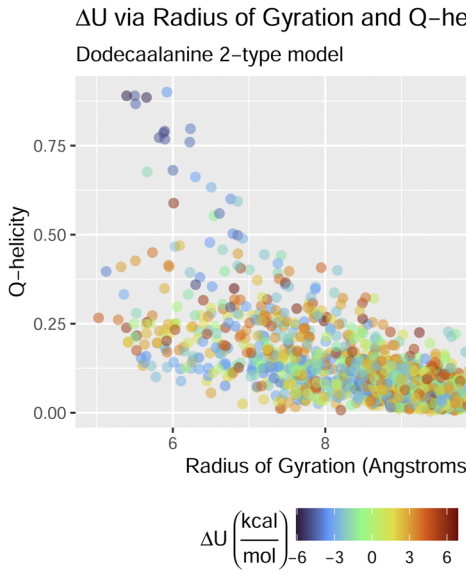

FIG. 6. ΔU as a function of the radius of gyration and Q-helicity for the two-type DDA model. No clear pattern is evident, demonstrating that these CVs do not reflect trends in the microscopic error.

emulation of DDA and how increasing the force-field resolution of actin improves performance at the four site resolution. As demonstrated in this subsection, directly correlating ΔU with traditional structural features is difficult; however, by utilizing SHAP values, the physical causes underlying the trends observed in Figs. 5 and 7 can be understood.

The classifiers trained in this work are functions of the pairwise distances between all CG particles; as a result, SHAP attributions are associated with pairwise distance values. The interparticle distances associated with the top six largest median absolute values of

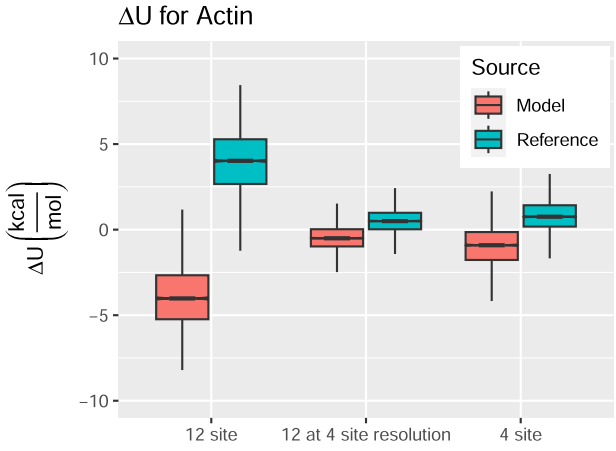

FIG. 7. Box plots of ΔU for multiple models and resolutions of actin divided along the reference and model ensembles. The "12 at 4 site resolution" and "four site" models are compared to the reference at the 4 site resolution, while "12 site" is compared at the 12 site resolution. The median is marked by the center line in each box. Note that varying models/resolutions change the form of ΔU, which changes the shape of the projected reference distribution.

SHAP attributions generated from analyzing two-site DDA are presented in Table I in the supplementary material; the largest value is associated with the terminal bond between sites 11 and 12 at the C terminus. As discussed in Sec. II B 2, the size of SHAP attributions is directly related to their importance. Plotting the SHAP values for distance 11–12 against the distance itself (Fig. 8) shows a clear dependence, whereas plotting ΔU as a function of distance 11–12 does not, i.e., the many body microscopic error is difficult to associate with physical features of the system, but its SHAP trend exhibits a clear relationship. The trend seen in Fig. 8 is representative of the error present in the 11–12 bond length distribution seen between the two-type DDA model and the reference data (see supplementary material). Similar trends are seen for the SHAP values associated with the 1–2 distance (see supplementary material). In summary, while the dependence of ΔU is difficult to associate with physical distances, SHAP values may show clear patterns and may be used to isolate problematic degrees of freedom: here, we did so by selecting SHAP values by their maximum median absolute value, investigating how said SHAP value varied, and locating a problematic marginal distribution. Accounting for this marginal disagreement by increasing the number of site types present at the termini of the model results in a drop in median ΔU as visualized in Fig. 5 (see supplementary material).

Classifiers that are trained to estimate microscopic error (and the SHAP values produced) reflect differences in multibody correlations between the various models analyzed and their associated reference data. In certain cases, summarized SHAP attributions may not have clear outlying values, or these values may be dependent on one another, impairing the creation of informative plots such as Fig. 8. For example, in the case of the four site model of actin, the leading SHAP values have three features of approximately equal SHAP magnitude, and plots such as Fig. 8 do not show clear trends (see supplementary material). However, the interdependence of SHAP values may be visualized using CVs derived

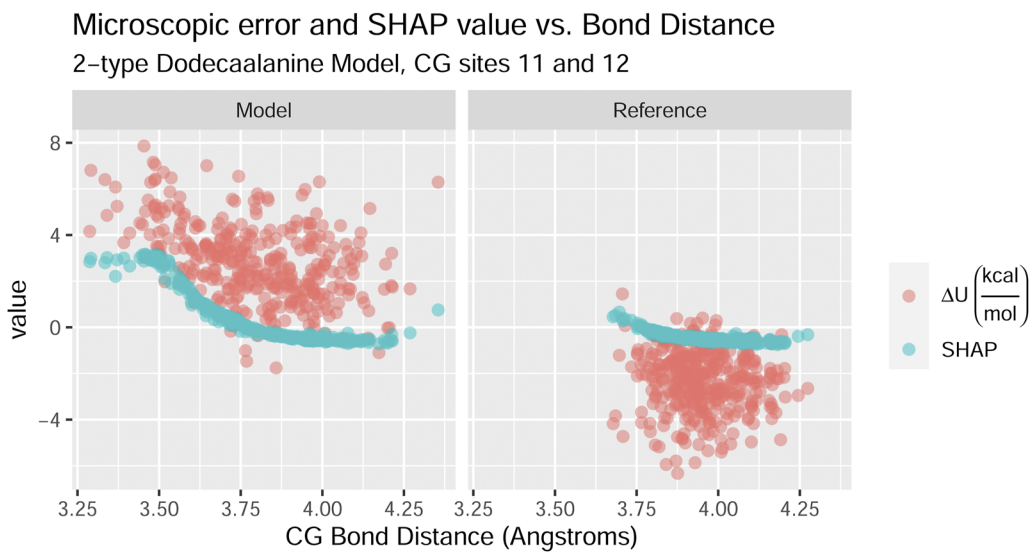

FIG. 8. Scatter plot of ΔU and the appropriate SHAP value as a function of the bond distance associated with the largest median absolute SHAP value for the two-type DDA model. Panels separate data from the model and reference ensembles. The SHAP values follow a clear trend as a function of bond distance, while ΔU values do not.

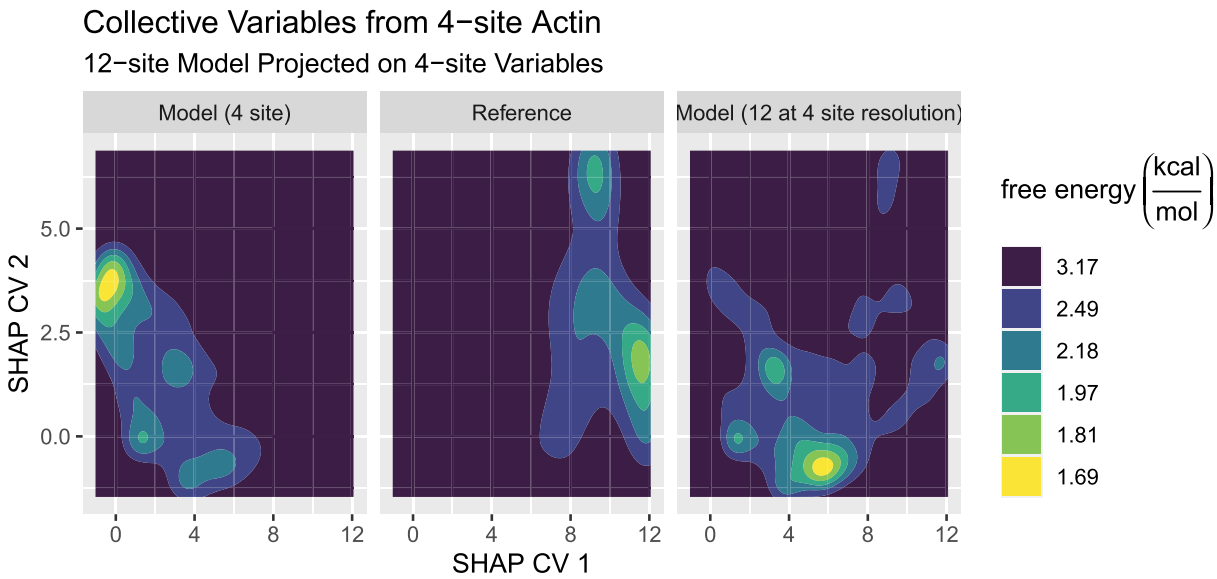

FIG. 9. Free energy surface produced along the SHAP variables generated by comparing a four site elastic network model to mapped reference data. The colored regions represent the individual model and reference densities. The left and middle panels contain the four site data used to generate the CVs; the right panel visualizes the 12 site model projected onto the CVs derived using the four site model.

from the SHAP values themselves. Using UMAP-based dimensional reduction (Sec. III and supplementary material), the trajectories of the four site model and reference can be projected onto variables that summarize the various SHAP values present in the trajectory (Fig. 9), and these derived collective variables can be used to visualize the interdependence of leading SHAP values (Fig. 10): relative to the four site model, the 12 site model reduces density in the top left corner and increases density in the low left corner, suggesting that errors related to the distances between site 2–3 and 2–4 are both altered. Direct visualization of these distances confirms that their cross correlations are improved in the 12 site model (Fig. 11), likely leading to the accuracy improvements shown in Fig. 7. Similar analyses may be performed for the DDA model discussed above; however, as the discussed errors are associated with individual bonds along the backbone and are effectively independent, no additional insight is gained.

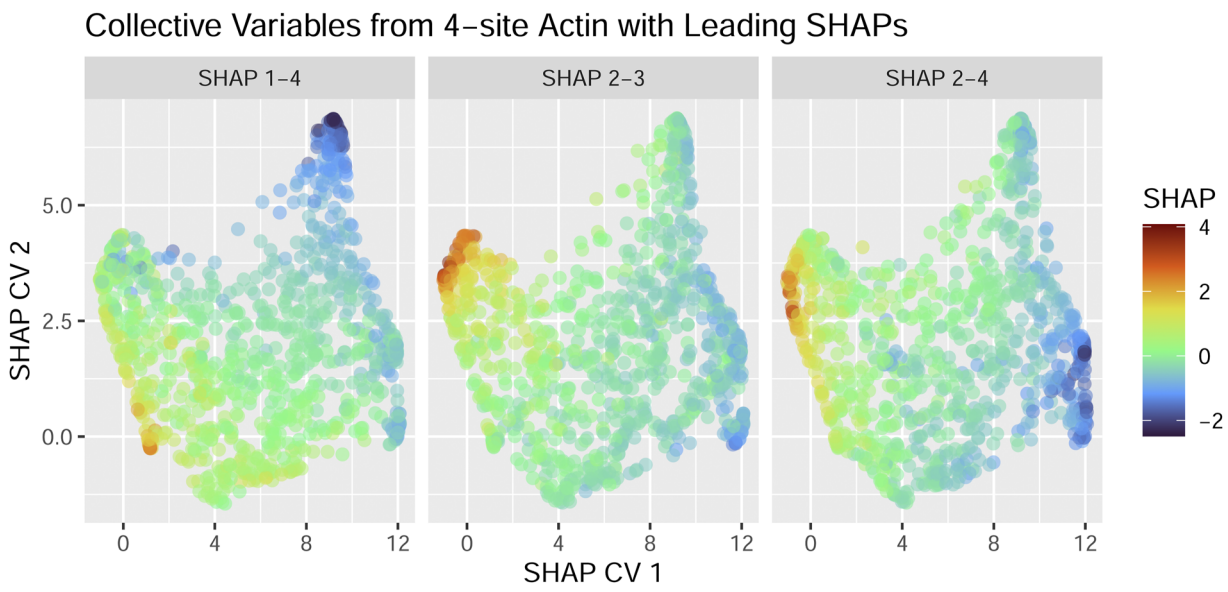

FIG. 10. The dependence of the SHAP values with the largest median absolute value for the four site actin model plotted as a function of UMAP-PCA-derived CVs. Note that SHAP 2–3 and SHAP 2–4 both concentrate in the top left corner.

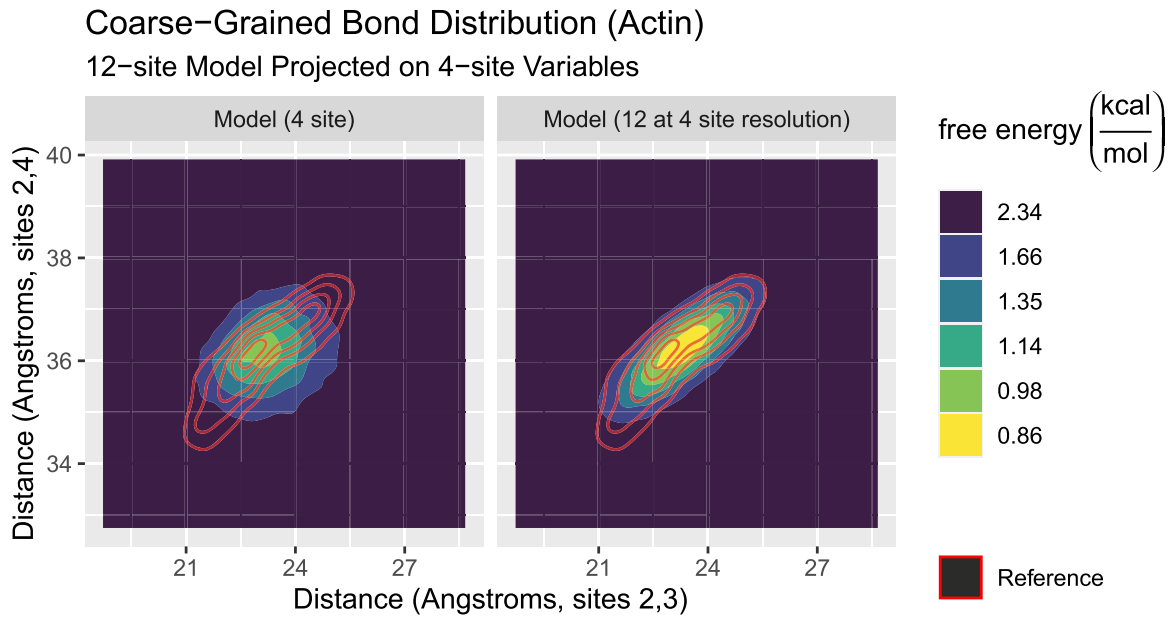

FIG. 11. Free energy surface produced the distance between sites 2 and 4 and the distance between sites 2 and 3. Filled contour panels represent the two model ensembles, while the red line contour overlay corresponds to statistics from the reference ensemble.

V. DISCUSSION AND CONCLUSIONS

The results presented here demonstrate that classification can be used to estimate microscopic error and that this error can be interpreted using SHAP values. The derived microscopic error does not necessarily correlate with traditional characterizations based on existing CVs and was compatible with comparing models at differing resolutions. In the case of DDA, SHAP values attributed the largest errors to bond disagreements. This is intuitive given the large conditional free energies found in effective bonded interactions that may result in areas of phase space that are effectively not traversed by either the model or reference data, as a lack of overlap in any

dimension implies a lack of overlap at the full phase space (see supplementary material).

In the case of actin, ΔU directly showed that the microscopic error present in the 12 site model was worse than that of the four site model when both were considered at their native resolution but that the 12 site model exhibited higher accuracy when considered at the four site resolution. SHAP values and derived CVs facilitated connecting this observation to a concrete cross correlation. This observation suggests that approaches that select CG mappings without considering the limited force-field basis that is eventually applied may not produce CG mapping operators that are necessarily useful in practice: in order to achieve accuracy for a given resolution, it may be helpful to model the system at a finer resolution. Similarly, if the proposed approach is combined with a series of maps at varying resolutions, the effect of model resolution can be validated in a novel and rigorous way. The interplay between ΔU and resolution supports the idea that the correctness of an approximate CG model may be difficult to consider without an implied resolution for its analysis.

A. Invariance and adversarial learning

The relationship between η and ΔU [Eq. (4)] provides important insight into classification in the current context. First, if the force-fields generating both ensembles are known, the information provided by classification can alternatively be gleaned by calculating the overall differences in free energies using, for example, the Bennet Acceptance Ratio.[77] Similarly, if performing classification between a CG model and a mapped reference ensemble, estimating the ideal classifier is analogous to estimating the true PMF along with the corresponding free energy difference. Additionally, Eq. (4) directly implies that any symmetry or locality shared by U_{FF} and U_{PMF} is shared by η. This has important consequences when considering applying the proposed XAI approach to novel molecular systems: the corresponding classification problem will contain physical symmetries and locality, and approaches that do not take this into account will likely provide poor estimates of η. A straightforward example is a homogeneous liquid where an asymmetric classifier is trained on the Cartesian coordinates: sampling sufficient to converge various local correlations (e.g., radial distribution functions) may be insufficient to parameterize such a classifier. On the other hand, functions that are formulated to obey permutational and rotational–translational symmetries, while already widely investigated as techniques for atomistic and CG force-field development,[14,16,78–80] will require appropriate explanation methods to take advantage of the approach described in this article.

Systematic coarse-graining methodologies typically define a numerical measure of error and then return the force-field that minimizes said error. Classification suggests similar ideas of global error based on the average accuracy achievable when performing classification between the ensemble implied by the force-field and the reference ensemble: a lower level of mean accuracy implies better emulation of the reference statistics (the accuracy is minimized if the reference and model are indistinguishable; this results in a constant η of 0.5). A natural question is then to consider force-fields that are optimized using this particular measure of quality. These force-field optimization approaches lead to adversarial learning, an approach firmly established by generative adversarial networks,[81] which, when applied to CG force-field development, is termed Adversarial Residual Coarse-Graining (ARCG).[82] The properties of η described in the previous paragraph additionally often apply to adversarial learning. The error estimation present in ARCG[82] can resultingly be viewed as simultaneously providing an estimate of U_{PMF} and the difference in configurational free energy. Conversely, the variational error in ARCG can be calculated without performing any classification: if a higher order force-field is used to approximate U_{PMF} and supplemented with a free energy difference method, the derivatives updating the parameters are similarly calculable through Eq. (4). Additionally, as η is central to adversarial residuals,[42,82] the explanations proposed in this article are fundamentally related to global residuals such as the relative entropy and the Hellinger distance. These various divergences provide a quantification of the overlap of ΔU described in Figs. 5 and 8; however, the numerical values of these divergences are difficult to interpret without context.

CG models are often created to study specific phenomena, and it may not be necessary to perfectly produce the microscopic behavior of the mapped atomistic system. In this case, the proposed methodology can be adapted to a certain extent by customizing the resolution at which it is performed, as in the example of actin. However, more broadly, the concept of microscopic error analysis as presented here may not be appropriate for these models. The modeler must decide whether to view the model as a way to reproduce specific emergent phenomena or whether to view the model as a drop in quantitative replacement for atomistic simulation. Certain coarse-graining strategies[82,83] can parameterize a force-field to reproduce the many-body behavior of a subset of the particles present in the CG system. However, doing so incorporates additional human influence into the creation of said CG model: as the resolution becomes coarser, the approach begins to resemble top-down parameterization strategies. We note that machine-learned atomistic force-fields are often quantified using values similar to ΔU;[24,84] if CG models are to be eventually considered as accurate as their fine-grained counterparts, utilizing similar measures of quality is critical.

B. XAI and future directions

The analysis in this article focuses on using classification and SHAP values to describe the behavior produced by CG potentials. The proposed approach trains a classifier to estimate ΔU and then uses techniques from XAI to explain said estimate. Interpretable models and explanations intrinsically provide a way to understand the high dimensional differences characterizing the quality of a proposed CG force-field, and we fully expect that other methods from the rapidly developing field of XAI will find similar utility. Furthermore, the study of explanations and interpretability is fundamentally relevant to the creation of CG models: CG force-fields are rarely created solely to reproduce the many-body-PMF of the training ensemble. They are, instead, often created either to extract knowledge from the system under study or to investigate new physical settings, tasks that intrinsically require human understanding of the limitations and workings of the utilized CG model. Any technique for bottom-up CG model creation that uses external human validation is a candidate for using explainable techniques. We hope that this work will serve as an initial example of a new approach to CG model validation.

SUPPLEMENTARY MATERIAL

Computational details, a formal mathematical connection between binary classification and microscopic error, tables, and additional figures are found in the supplementary material.

ACKNOWLEDGMENTS

This material is based upon the work supported by the National Science Foundation (NSF Grant No. CHE-2102677). A.E.P.D. thanks Dr. Alexander Pak, Dr. Glen Hocky, and Dr. Sriramvignesh Mani for insight, guidance, and simulation data. Simulations were performed using computational resources provided by the University of Chicago Research Computing Center (RCC).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Aleksander E. P. Durumeric: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Writing – original draft (equal). **Gregory A. Voth**: Conceptualization (supporting); Formal analysis (supporting); Funding acquisition (lead); Project administration (lead); Resources (lead); Supervision (lead); Writing – review & editing (lead).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," Nat. Struct. Biol. **9**, 646–652 (2002).
[2]D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, "Atomic-level characterization of the structural dynamics of proteins," Science **330**, 341–346 (2010).
[3]I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, "High-throughput all-atom molecular dynamics simulations using distributed computing," J. Chem. Inf. Model. **50**, 397–403 (2010).
[4]K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," Science **334**, 517–520 (2011).
[5]M. Karplus and R. Lavery, "Significance of molecular dynamics simulations for life sciences," Isr. J. Chem. **54**, 1042–1051 (2014).
[6]G. A. Voth, *Coarse-graining of Condensed Phase and Biomolecular Systems* (Taylor & Francis, , 2008).
[7]E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero, and N. F. A. van der Vegt, "Systematic coarse-graining methods for soft matter simulations—A review," Soft Matter **9**, 2108–2119 (2013).
[8]M. G. Saunders and G. A. Voth, "Coarse-graining methods for computational biology," Annu. Rev. Biophys. **42**, 73–93 (2013).
[9]W. G. Noid, "Perspective: Coarse-grained models for biomolecular systems," J. Chem. Phys. **139**, 090901 (2013).
[10]W. G. Noid, *Biomolecular Simulations* (Springer, 2013).
[11]S. J. Marrink and D. P. Tieleman, "Perspective on the Martini model," Chem. Soc. Rev. **42**, 6801–6822 (2013).
[12]R. Potestio, C. Peter, and K. Kremer, "Computer simulations of soft matter: Linking the scales," Entropy **16**, 4199–4245 (2014).
[13]A. J. Pak and G. A. Voth, "Advances in coarse-grained modeling of macromolecular complexes," Curr. Opin. Struct. Biol. **52**, 119–126 (2018).
[14]P. Gkeka, G. Stoltz, A. Barati Farimani, Z. Belkacemi, M. Ceriotti, J. D. Chodera, A. R. Dinner, A. L. Ferguson, J.-B. Maillet, H. Minoux, C. Peter, F. Pietrucci, A. Silveira, A. Tkatchenko, Z. Trstanova, R. Wiewiora, and T. Lelièvre, "Machine learning force fields and coarse-grained variables in molecular dynamics: Application to materials and biological systems," J. Chem. Theory Comput. **16**, 4757–4775 (2020).
[15]F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," Annu. Rev. Phys. Chem. **71**, 361–390 (2020).
[16]J. Jin, A. J. Pak, A. E. P. Durumeric, T. D. Loose, and G. A. Voth, "Bottom-up coarse-graining: Principles and perspectives," J. Chem. Theory Comput. **18**, 5759–5791 (2022).
[17]W. G. Noid, "Perspective: Advances, challenges, and insight for predictive coarse-grained models," J. Phys. Chem. B **127**, 4174 (2023).
[18]J. F. Rudzinski, "Recent progress towards chemically-specific coarse-grained simulation models with consistent dynamical properties," Computation **7**, 42 (2019).
[19]W. G. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, "Multiscale coarse-graining and structural correlations: Connections to liquid-state theory," J. Phys. Chem. B **111**, 4116–4127 (2007).
[20]M. S. Shell, "The relative entropy is fundamental to multiscale and inverse thermodynamic problems," J. Chem. Phys. **129**, 144108 (2008).
[21]W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," J. Mol. Graph. **14**, 33–38 (1996).
[22]M. Ceriotti, "Unsupervised machine learning in atomistic simulations, between predictions and understanding," J. Chem. Phys. **150**, 150901 (2019).
[23]A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, "Machine learning a general-purpose interatomic potential for silicon," Phys. Rev. X **8**, 041048 (2018).
[24]A. E. P. Durumeric, N. E. Charron, C. Templeton, F. Musil, K. Bonneau, A. S. Pasos-Trejo, Y. Chen, A. Kelkar, F. Noé, and C. Clementi, "Machine learned coarse-grained protein force-fields: Are we there yet?," Curr. Opin. Struct. Biol. **79**, 102533 (2023).
[25]G. Stoltz, M. Rousset, and T. Lelièvre, *Free Energy Computations: A Mathematical Perspective* (World Scientific, 2010).
[26]C. Molnar, *Interpretable Machine Learning*, 2 edn (Lulu.com, 2019).
[27]A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion **58**, 82–115 (2020).
[28]V. Arya *et al.*, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," arXiv:1909.03012 (2019).
[29]W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," Proc. Natl. Acad. Sci. U. S. A. **116**, 22071–22080 (2019).
[30]C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning—A brief history, state-of-the-art and challenges," in *ECML PKDD 2020 Workshops*, edited by I. Koprinska *et al.* (Springer, Cham, 2020), pp. 417–431.
[31]A. Holzinger, P. Kieseberg, E. Weippl, and A. M. Tjoa, "Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI," *Lecture Notes in Computer Science* (Springer, Cham, 2018), Vol. 11015, pp. 1–8.
[32]Y. Kodratoff, "The comprehensibility manifesto," https://www.kdnuggets.com/news/94/n9.txt.
[33]S. Rüping, "Learning interpretable models," Ph.D. thesis, Dortmund University, 2006.
[34]C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell. **1**, 206–215 (2019).

[35]B. Mittelstadt, C. Russell, and S. Wachter, *Explaining Explanations in AI*, edited by D. Boyd, J. Morgenstern, A. Chouldechova and F. Diaz (Association for Computing Machinery, New York, 2019), pp. 279–288.
[36]S. M. Lundberg and S.-I. Lee, in *A Unified Approach to Interpreting Model Predictions*, edited by I. Guyon *et al.* (Red Hook, NY), pp. 4768–4777.
[37]S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," Nat. Biomed. Eng. **2**, 749–760 (2018).
[38]S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell. **2**, 56–67 (2020).
[39]J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer, 2001), Vol. 1.
[40]A. Buja, W. Stuetzle, and Y. Shen, "Loss functions for binary class probability estimation and classification: Structure and applications of work," Technical report, University of Pennsylvania, 2005.
[41]T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," J. Am. Stat. Assoc. **102**, 359–378 (2007).
[42]M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," J. Mach. Learn. Res. **12**, 731–817 (2011).
[43]A. Niculescu-Mizil and R. Caruana, in *Predicting good probabilities with supervised learning*, edited by S. Dzeroski, L. De Raedt, and S. Wrobel (Association for Computing Machinery, New York, 2005), pp. 625–632.
[44]C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," Proc. Mach. Learn. Res. **70**, 1321–1330 (2017).
[45]V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," Proc. Mach. Learn. Res. **80**, 2796–2804 (2018).
[46]T. Lemke and C. Peter, "Neural network based prediction of conformational free energies—A new route toward coarse-grained simulation models," J. Chem. Theory Comput. **13**, 6213–6221 (2017).
[47]X. Ding and B. Zhang, "Contrastive learning of coarse-grained force fields," J. Chem. Theory Comput. **18**, 6334–6344 (2022).
[48]L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games (AM-28)* (Princeton University Press, 1953), Vol. 2, pp. 307–318.
[49]H. P. Young, "Monotonic solutions of cooperative games," Int. J. Game Theory **14**, 65–72 (1985).
[50]M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in *Proceedings of the 37th International Conference on Machine Learning* (PMLR, 2020), Vol. 119, pp. 9269–9278.
[51]I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in *Proceedings of the 37th International Conference on Machine Learning* (PMLR, 2020), Vol. 119, pp. 5491–5500.
[52]T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, Jupyter development team. *Jupyter Notebooks—A publishing format for reproducible computational workflows*, edited by F. Loizides and B. Scmidt (IOS Press BV: Amsterdam, Netherlands, 2016), pp. 87–90.
[53]R. K. Vinayak and R. Gilad-Bachrach, "DART: Dropouts meet multiple additive regression trees," Proc. Mach. Learn. Res. **38**, 489–497 (2015).
[54]G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon et al. (Curran Associates, 2017).
[55]R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky *et al.*, "Theano: A Python framework for fast computation of mathematical expressions," arXiv:1605.02688 (2016).
[56]C. R. Harris *et al.*, "Array programming with NumPy," Nature **585**, 357–362 (2020).
[57]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[58]L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv:1802.03426 (2018).
[59]L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," J. Open Source Softw. **3**, 861 (2018).
[60]Pandas Development Team T. pandas-dev/pandas: Pandas 1.1.3 version v. 1.1.3, Pandas, 2020.
[61]H. Wickham, *Ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2016).
[62]M. Dowle and A. Srinivasan, *data.table: Extension of "data.frame" v. 1.12.8*, 2019.
[63]H. W. Borchers. *pracma: Practical Numerical Math Functions v. 2.2.9*, 2019.
[64]J. F. Rudzinski and W. G. Noid, "Bottom-up coarse-graining of peptide ensembles and helix-coil transitions," J. Chem. Theory Comput. **11**, 1278–1291 (2015).
[65]D. A. Case *et al.*, AMBER 2018, University of California, San Francisco, 2018.
[66]J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell, Jr., "CHARMM36m: An improved force field for folded and intrinsically disordered proteins," Nat. Methods **14**, 71–73 (2017).
[67]T. Schneider and E. Stoll, "Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions," Phys. Rev. B **17**, 1302–1322 (1978).
[68]S. Piana and A. Laio, "A bias-exchange approach to protein folding," J. Phys. Chem. B **111**, 4553–4559 (2007).
[69]A. Prakash, M. D. Baer, C. J. Mundy, and J. Pfaendtner, "Peptoid backbone flexibilility dictates its interaction with water and surfaces: A molecular dynamics investigation," Biomacromolecules **19**, 1006–1015 (2018).
[70]M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX **1–2**, 19–25 (2015).
[71]G. M. Hocky, J. L. Baker, M. J. Bradley, A. V. Sinitskiy, E. M. De La Cruz, and G. A. Voth, "Cations stiffen actin filaments by adhering a key structural element to adjacent subunits," J. Phys. Chem. B **120**, 4558–4567 (2016).
[72]A. D. Mackerell, Jr., M. Feig, and C. L. Brooks III, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," J. Comput. Chem. **25**, 1400–1415 (2004).
[73]G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," J. Chem. Phys. **126**, 014101 (2007).
[74]M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," J. Appl. Phys. **52**, 7182–7190 (1981).
[75]M. G. Saunders and G. A. Voth, "Comparison between actin filament models: Coarse-graining reveals essential differences," Structure **20**, 641–653 (2012).
[76]E. Lyman, J. Pfaendtner, and G. A. Voth, "Systematic multiscale parameterization of heterogeneous elastic network models of proteins," Biophys. J. **95**, 4183–4192 (2008).
[77]C. H. Bennett, "Efficient estimation of free energy differences from Monte Carlo data," J. Comput. Phys. **22**, 245–268 (1976).
[78]A. Goscinski, G. Fraux, G. Imbalzano, and M. Ceriotti, "The role of feature space in atomistic learning," Mach. Learn.: Sci. Technol. **2**, 025028 (2021).
[79]B. Anderson, T.-S. Hy, and R. Kondor, "Cormorant: Covariant molecular neural networks," in Advances in Neural Information Processing Systems Vol. 32, edited by H. Wallach et al. (Curran Associates, 2019).
[80]O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," Chem. Rev. **121**, 10142–10186 (2021).
[81]I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, in *Generative Adversarial Nets*, edited by Z. Ghahramani *et al.* (Red Hook, NY), pp. 2672–2680.
[82]A. E. P. Durumeric and G. A. Voth, "Adversarial-residual-coarse-graining: Applying machine learning theory to systematic molecular coarse-graining," J. Chem. Phys. **151**, 124110 (2019).

[83]P. G. Sahrmann, T. D. Loose, A. E. P. Durumeric, and G. A. Voth, "Utilizing machine learning to greatly expand the range and accuracy of bottom-up coarse-grained models through virtual particles," J. Chem. Theory Comput. (published online) (2023).

[84]X. Fu, Z. Wu, W. Wang, T. Xie, S. Keten, R. Gomez-Bombarelli, and T. Jaakkola, "Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations," arXiv:2210.07237 (2022).