# Multi-body effects in a coarse-grained protein force field



Citation for the original published paper (version of record):
Wang, J., Charron, N., Husic, B. et al (2021). Multi-body effects in a coarse-grained protein force field. Journal of Chemical Physics, 154(16). http://dx.doi.org/10.1063/5.0041022



(article starts on next page)

The Journal of Chemical Physics

# Multi-body effects in a coarse-grained protein force field

Cite as: J. Chem. Phys. **154**, 164113 (2021); https://doi.org/10.1063/5.0041022
Submitted: 18 December 2020 . Accepted: 01 April 2021 . Published Online: 29 April 2021

Jiang Wang, Nicholas Charron, Brooke Husic, Simon Olsson, Frank Noé, and Cecilia Clementi

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

# Multi-body effects in a coarse-grained protein force field

Cite as: J. Chem. Phys. 154, 164113 (2021); doi: 10.1063/5.0041022
Submitted: 18 December 2020 • Accepted: 1 April 2021 •
Published Online: 29 April 2021

View Online | Export Citation | CrossMark

Jiang Wang,[1,2,3] Nicholas Charron,[1,4,5] Brooke Husic,[6] Simon Olsson,[6,7] Frank Noé,[1,2,5,6,a)] and Cecilia Clementi[1,2,4,5,b)]

## AFFILIATIONS

[1] Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, USA
[2] Department of Chemistry, Rice University, Houston, Texas 77005, USA
[3] College of Science, Guizhou Institute of Technology, Guiyang, Guizhou 550000, China
[4] Department of Physics, Rice University, Houston, Texas 77005, USA
[5] Department of Physics, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany
[6] Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany
[7] Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden

[a)] **Electronic mail:** frank.noe@fu-berlin.de
[b)] **Author to whom correspondence should be addressed:** cecilia.clementi@fu-berlin.de

## ABSTRACT

The use of coarse-grained (CG) models is a popular approach to study complex biomolecular systems. By reducing the number of degrees of freedom, a CG model can explore long time- and length-scales inaccessible to computational models at higher resolution. If a CG model is designed by formally integrating out some of the system's degrees of freedom, one expects multi-body interactions to emerge in the effective CG model's energy function. In practice, it has been shown that the inclusion of multi-body terms indeed improves the accuracy of a CG model. However, no general approach has been proposed to systematically construct a CG effective energy that includes arbitrary orders of multi-body terms. In this work, we propose a neural network based approach to address this point and construct a CG model as a multi-body expansion. By applying this approach to a small protein, we evaluate the relative importance of the different multi-body terms in the definition of an accurate model. We observe a slow convergence in the multi-body expansion, where up to five-body interactions are needed to reproduce the free energy of an atomistic model.



## I. INTRODUCTION

Molecular dynamics (MD) is a well-established tool for studying biomolecular systems. Atomistic MD can be used to characterize protein configurational changes, folding, and binding of small to intermediate-sized proteins (hundreds of residues) on timescales of milliseconds.[1–3] When combined with recent methodological and algorithmic advances, even longer timescales can be reached.[4–6] However, despite this progress, MD is still limited to relatively fast and localized processes, when considering biological length- and timescales.

Various methods have been developed to push the boundaries of the time- and length-scales accessible by MD. Enhanced sampling methods promote the exploration of conformation space and the transition between long-lived (metastable) states in order to obtain converged estimates of free energy landscapes that are beyond the reach of naive MD simulations. Examples include umbrella sampling,[7,8] parallel tempering,[9–12] or adaptive sampling.[5,13–15] Alternatively, instead of expediting free energy landscape exploration, the free energy landscape itself can be simplified by defining and applying a molecular model with a coarse-grained (CG) representation.[16–21] By using coarse-graining to reduce the degrees of freedom in the molecular system, simulations can be significantly sped up. Because coarse-graining necessarily omits some physicochemical details, it is crucial to choose a CG strategy designed to preserve the properties of interest to the researcher. Even if some chemical details are missing in the CG representation, one can argue that a successful CG model allows us to focus on the most important physical factors associated with the system behavior.[19]

In practice, the definition of a CG model consists of two steps: first, multiple atoms are mapped onto CG sites, often referred to as "beads." Then, an effective CG Hamiltonian is defined as a function of the bead coordinates. These steps are interconnected and are both important for the success of a CG model.[22] Although multiple algorithmic strategies have been proposed for the CG mapping,[23–26] they are most commonly based on physical and chemical intuition, and the optimization of the CG mapping is still an open area of research.[25] The definition of an effective Hamiltonian for a given CG mapping depends on the goal of the coarse-grained model. As some of the information is necessarily lost upon coarse-graining, CG models must be designed such that certain targeted properties of the molecular system are retained and can be computed from both the all-atom and the CG ensembles. Depending on the subject of study, this can be accomplished by using top-down, bottom-up, or knowledge-based models.[22] In top-down methods, a CG model is defined to optimally reproduce a set of global (macroscopic) observables.[17,20,27–30] In contrast, bottom-up CG methods are designed to preserve specific microscopic properties of an atomistic or first-principles model while coarse-graining, e.g., thermodynamic properties[16,18,31–34] or kinetic properties.[35]

When designing a CG model, one of the main difficulties is that the CG Hamiltonian should, in principle, include multi-body interaction terms among the beads in the system.[36–38] Traditionally, CG Hamiltonians for protein systems have been defined as a combination of functional forms similar to the ones used in atomistic forcefields, that is, harmonic bonds, angle and dihedral terms, and non-bonded two-body interactions (see, e.g., Refs. 29, 39, and 40). Multi-body terms have been added in CG models as a correction[41–45] or by considering the physical nature of the corresponding interactions, as for instance, water-mediated potentials.[20,46] However, the multi-body terms are often crucial to reproduce the system's behavior correctly. For instance, an early study with a CG water model including three-body interactions showed the importance of many-body terms.[47] Larini *et al.* designed a CG model by parameterizing specific forms of two-body and three-body energy functions and have shown that they perform significantly better than the ones parameterized with only two-body potentials.[42] A more general and systematic approach, which does not require the choice of a specific three-body functional form, was developed by Das and Andersen.[48,49] When tested on the modeling of SPC/E water, this approach obtained a significant improvement in the model accuracy.[49] Additionally, Andrienko and co-workers have shown that the inclusion of many-body terms into a CG model can result in substantial changes in the two-body interactions, making them much more attractive at short distances.[50] By first parameterizing the two-body potential and then introducing three-body terms as a correction, these authors have demonstrated that three-body interactions are essential to reproduce structural properties of liquid water.[50] Four-body terms have also been taken into account in a CG model energy, by considering dihedral potentials between sets of four atoms,[51,52] and higher-body terms have been built by means of statistical contact potentials.[53–55]

Many-body terms can also be included in a CG potential by using kernel-based machine learning methods, such the Gaussian approximation potentials pioneered by Csańyi and co-workers. It has been shown that CG molecular models designed using this framework are able to describe many-body interactions and are much more accurate than models only using pair potentials.[56] Scherer and co-workers have also described a number of kernel-based strategies to parameterize the traditional force field of molecular liquids, and they have showed that a model with two- and three-body machine learned potentials is computationally efficient and correctly recovers two- and three-body distribution functions.[57]

Alternative approaches have been proposed to take into account multi-body terms. For instance, multi-body corrections can be included by utilizing virtual sites in a CG mapping scheme.[58] In the ultra-coarse-grained (UCG) theory, standard single-state CG beads are mixed with "special" CG beads with rapidly adjusting internal states. In theory, this approach can effectively account for multi-body effects in the CG model, but it is limited by the choices for the functional forms of the interactions.[59] UCG representations are equivalent to interacting-particle reaction dynamics (iPRD),[60] which were obtained as fine-grained representations of particle-based reaction dynamics rather than coarse-grained representations of molecular systems.

In recent work, by our group[61,62] and others,[26,63,64] a different philosophy has been employed to take into account multi-body effects in CG modeling: namely, taking advantage of the ability of modern machine learning techniques to approximate arbitrary complex multi-body functions. Given the recent success in the use of these techniques for the definition of the classical energy function from quantum mechanical calculations,[65–84] a similar idea has been applied for coarse-graining. To this end, we have used both neural networks (CGnets)[61,62] and kernel methods[85] as universal function approximators that can represent complex many-body terms on top of lower order terms. We have demonstrated on several simple systems and a mini-protein that, thanks to the general modeling of the full $n$-body interaction potential allowed by these techniques, it is possible to design CG models that accurately reproduce the free energy landscape of atomistic models.[61,62,85]

In the present work, we employ general CGnets and a multi-body CGnet architecture in order to analyze to which degree multi-body interactions are required to represent accurate coarse-grained force fields. We find that, on a test miniprotein, correction terms limited to three-body interactions are not sufficient for a CG model to reproduce the free energy of an atomistic model. Even if they are very small with respect to the two-body and three-body contributions, four-body terms and chirality information are needed to at least qualitatively reproduce the free energy landscape of the miniprotein considered here. Furthermore, five-body terms are necessary to quantitatively reproduce the free energy. Surprisingly, not only do additional terms beyond five-body not further improve the CG model (according to the mean square distance from the reference free energy landscape), but a model including up to five-body interactions outperforms our original CGnet model, which in theory allows for interactions at any order.[61] We believe that this architectural restriction of the multi-body interactions up to a certain order acts as an implicit regularization on the CG model that reduces overfitting, producing a smoother free energy landscape and better agreement with the atomistic model.

Although this approach is applied to a single system and we cannot directly generalize the results to different systems, the multi-body decomposition presented here opens the way to the formulation of more general and accurate CG models and to the understanding of the key physical ingredients, shaping the energy landscape of a CG protein model.

## II. THEORY AND METHODS

### A. Coarse-graining with thermodynamic consistency

Bottom-up methods are said to be "thermodynamically consistent" if the free energy landscape of the resulting CG model matches the corresponding free energy landscape of the fine-grained model (when projected in the same space). One approach to enforce thermodynamic consistency is the so-called multi-scale coarse-graining method (MS-CG) proposed by Noid, Voth, and colleagues.[36,86] It has been proven that, under certain restrictions of the CG map, the thermodynamically consistent CG model can be uniquely identified from the set of all possible CG energy functions by minimizing the mean square error (MSE) between the instantaneous atomistic forces projected onto the CG space and the CG forces.[22] This procedure, originally developed in the atomistic context for *ab initio* data,[87] is called "force matching" and was first employed in the CG context in Ref. 18. However, the force matching MSE can never be reduced to 0. This is because each point in CG space corresponds to an *ensemble* of atomistic configurations, and the CG force cannot match each instantaneous force of an all-atom configuration mapping to the same CG configuration. Thus, the forces calculated from the thermodynamically consistent CG model correspond to the *mean* atomistic force computed on that ensemble of atomistic configurations weighted by their Boltzmann factors.[36] Therefore, the minimum MSE obtained through force matching in the CG context is strictly larger than 0 for any non-trivial CG mapping. In a statistical machine learning framework, this minimum MSE corresponds to the estimator noise.[61] In this section, we briefly describe the theory behind the force matching method for coarse-graining.

A configuration of an all-atom protein system consisting of $N_a$ atoms can be represented by a vector $\mathbf{r} \in \mathbb{R}^{3N_a}$. If the representation is coarse-grained, the atomistic configuration is mapped into a lower dimensional vector $\mathbf{x}$ via the mapping function,

$$\mathbf{x} = \xi(\mathbf{r}) \in \mathbb{R}^{3n_{CG}}, \tag{1}$$

where $n_{CG} < N_a$ is the number of beads in the CG system. The mapping scheme $\xi$ depends on the specific system under study.[37] Here, we assume that the CG mapping is given and that it captures the most important structural features of the molecule. We further assume that $\xi$ is linear, i.e., the mapping is a linear transformation of variables defined by the coarse-graining matrix $\Xi \in \mathbb{R}^{3n\times 3N_a}$, so that $\mathbf{x} = \Xi\mathbf{r}$. In the following, we use a matrix $\Xi$ whose elements are only zeros and ones: in other words, the CG mapping is a slicing of the configurational space of the atomistic model.

We indicate the CG energy function as $U(\mathbf{x};\, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are parameters that need to be optimized. The parameterization of $U(\mathbf{x};\, \boldsymbol{\theta})$ can be realized in different ways, i.e., through the combination of fixed functional forms or through machine learning approaches.[61,62,85]

Here, we employ a neural network architecture to represent $U(\mathbf{x};\, \boldsymbol{\theta})$ and we seek to optimize the network parameters $\boldsymbol{\theta}$ to obtain the thermodynamically consistent energy function for a given CG system. This means that we wish to identify the parameters that best satisfy the following equation:

$$U(\mathbf{x};\boldsymbol{\theta}) \equiv -k_B T \ln p^{CG}(\mathbf{x}) + \text{const}, \tag{2}$$

where $k_B$ is the Boltzmann constant, $T$ is the temperature, and $p^{CG}(\mathbf{x})$ is the probability density distribution in the CG space, given by the marginalization of the probability density of the atomistic model,

$$p^{CG}(\mathbf{x}) = \frac{\int \mu(\mathbf{r})\delta(\mathbf{x} - \xi(\mathbf{r}))d\mathbf{r}}{\int \mu(\mathbf{r})d\mathbf{r}}, \tag{3}$$

where $\mu(\mathbf{r}) = \exp(-V(\mathbf{r})/k_B T)$ is the Boltzmann weight associated with the atomistic energy $V(\mathbf{r})$.[88]

It is important to note that even if the atomistic energy $V(\mathbf{r})$ contains mostly pairwise interactions, by enforcing Eq. (2), multi-body terms emerge in the thermodynamically consistent energy of a CG model, by effect of the dimensionality reduction.

Various methods have been proposed to satisfy Eq. (3) as best as possible, such as relative entropy[34] and force matching.[18,37] In this work, we design protein CG models by means of the force matching method.

In practice, force matching optimizes the parameters $\theta$ in the CG potential $U(\mathbf{x};\boldsymbol{\theta})$ through the minimization of the functional,[18,37]

$$\chi^2(\boldsymbol{\theta}) = \left\langle \|\Xi_F(\mathbf{F}(\mathbf{r})) + \nabla U(\Xi\mathbf{r},\boldsymbol{\theta})\|^2 \right\rangle_{\mathbf{r}}, \tag{4}$$

where $-\nabla U(\mathbf{x};\, \boldsymbol{\theta})$ is the CG force field, $\Xi_F(\mathbf{F}(\mathbf{r}))$ is the instantaneous all-atom force projected onto the CG space,[88] and $\langle\cdot\rangle_{\mathbf{r}}$ is the weighted average over the equilibrium distribution of the atomistic model, i.e., $\mathbf{r} \sim \mu(\mathbf{r})$. With our assumptions on the CG mapping, the projection of the forces becomes $\Xi_F = \Xi$.

It can be proven that the CG potential minimizing (4) in the space of all possible functions satisfies thermodynamical consistency (2) in the limit of infinite sampling.[18,37]

### B. Multi-body terms in the CG potential

In principle, the MS-CG approach allows us to find the correct thermodynamically consistent CG energy function if the MSE is minimized in the space of all possible functions, including multi-body terms. However, in practice, the CG model is usually optimized variationally in the space spanned by only two-body (or few-body) functions.[86] For an all-atom system with $N_a$ atoms, the atomistic potential energy of the system $V(\mathbf{r})$ is usually expressed as the sum of non-bonded pairwise interactions (e.g., Lennard-Jones and Coulomb) and local terms such as harmonic bond, angle, and dihedral potentials. In general, the CG energy $U(\mathbf{x},\, \theta)$ can then be decomposed as follows:

$$U(\mathbf{x},\theta) = U(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n;\theta) = \sum_{k=2}^{n} U^{(k)}(\mathbf{x},\theta), \tag{5}$$

where $\mathbf{x}_i$, with $i \in \{1, 2, \ldots, n\}$, indicates the coordinate of the $i$th CG bead. $U^{(k)}(\mathbf{x},\, \theta)$ indicates a functional form involving $k$-body interactions and can be further decomposed as

$$U^{(k)}(\mathbf{x},\theta) = U^{(k)}(\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_n};\theta) = \sum_{\{i_1,i_2,\ldots,i_k\}} f_{i_1 i_2\ldots i_k}(\mathbf{x_{i_1}},\mathbf{x_{i_2}},\ldots,\mathbf{x_{i_k}}), \tag{6}$$

where $\{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$ indicates all possible combinations of $k$ indexes chosen from the full set $\{1, 2, \ldots, n\}$ and $f_{i_1 i_2\ldots i_k}$ are non-decomposable $k$-body functions (i.e., they cannot be written as the sum of lower order functions). Note that we do not include $k = 1$ in (5) as (i) only relative energies matter, and thus, the energy $U$ is defined up to an additive constant, and (ii) we assume the absence of external forces, and thus, the energy only depends on internal coordinates.

Traditional CG approaches only include lower order terms for the non-bonded interactions, that is, the effective CG potential usually contains only $U^{(2)}$ and sometimes $U^{(3)}$ terms in the decomposition (6), as discussed in the Introduction. In this work, we focus on the effect of higher order terms in CG protein models and extend our previously proposed deep learning framework (CGnets)[61] to explicitly learn the terms $U^{(k)}$ for any $k$ in a thermodynamically consistent potential $U(\mathbf{x}, \theta)$.

### C. Constructing a multi-body CG model using neural networks

In our previous work,[61] we proposed CGnet, a deep learning framework, to model a thermodynamically consistent CG potential from all-atom molecular dynamics trajectories. In this work, we extend CGnet to extract the different $n$-body contributions explicitly, i.e., by means of a multi-body expansion.

We create a set of several different models to explore the contributions of the multi-body terms in the CG energy. We start the multi-body expansion (5) by considering the two-body contribution. A similar two-body network was also employed in our previous work[61] to define what we called "the spline model," which was used as a comparison to CGnet. In practice, as shown in Fig. 1(a), the Cartesian coordinates $\mathbf{x}$ of the CG system are first transformed into a set of roto-translational invariant features $\mathbf{y}$. Then, each single feature is passed separately through an individual network [(indicated as "two-body unit" in Fig. 1(a)]. The features considered here consist of all the pairwise distances and $\cos(\phi)$ and $\sin(\phi)$ of each dihedral angle $\phi$ spanned by four continuous beads. In addition, all the angles defined by three adjacent beads, all the bonds between two adjacent beads, and excluded volume terms are passed through the prior energy unit, which precomputes a prior potential as previously described.[61] The excluded volume repulsive terms take the form $\sum_{ij}(\frac{\sigma}{r_{ij}})^c$, where $r_{ij}$ is the distance between CG beads $i$ and $j$ for all pairs $(i, j)$ connected by $\geqslant 3$ bonds. The excluded volume radius $\sigma$ and exponent $c$ are fixed to the values obtained in previous work:[61] $\sigma = 5.5$ Å and $c = 6$. The prior energy terms serve as a physical constraint and act as a regularization, ensuring that trajectories generated by the learned potential lie within physical meaningful regions. The sum of the energy resulting from all the two-body units, the dihedral units, and the prior energy is the CG energy of the two-body model.

We then consider the next term in expansion (5), by including three-body terms. That is, once the two-body CGnet is trained, we fix its weights, and three-body contributions are added to the fixed two-body CGnet. The three-body contributions are added by defining three-body unit networks (see Fig. 1, with $k = 3$): each unit network takes as input the three pairwise distances among a set of three beads for each of the $\binom{n}{3}$ sets that can be selected from the total

**TABLE I.** Hyperparameters of the neural network for different models.

| Model | Network depth | No. of neurons/ layer | Lipschitz reg. strength |
|---|---|---|---|
| CGnet | 5 | 250 | 4.0 |
| 2-body | 3 | 60 | 5.0 |
| 2, 3-body | 3 | 60 | 5.0 |
| 2, 3, 4-body | 3 | 80 | 5.0 |
| 2, 3, 4C-body | 4 | 80 | 5.0 |
| 2, 3, 4C, 5-body | 3 | 70 | 5.0 |

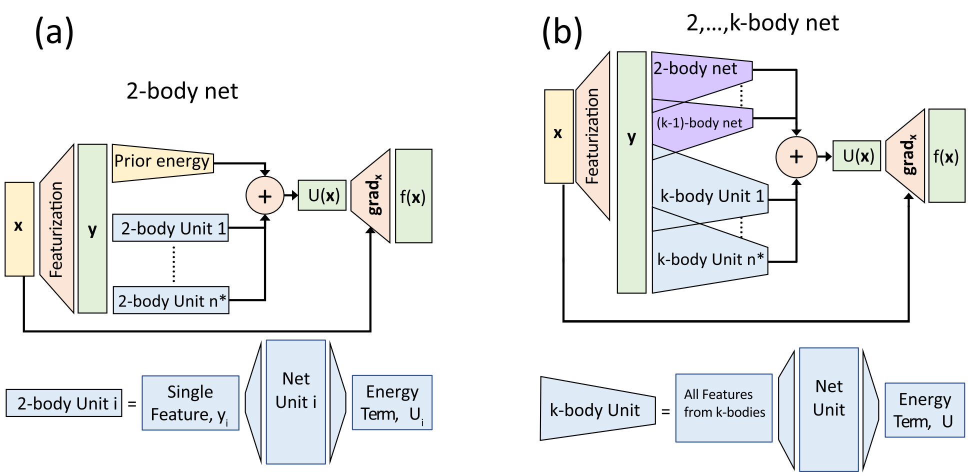


**FIG. 1**. (a) Neural network structure for the two-body CG potential. The input to a given two-body unit is a single pairwise distance between a pair of CG beads. (b) Neural network structure for a multi-body CG potential up to order $k$. The inputs to a given $k$-body unit are all pairwise distances within a given set of $k$ CG beads.

$n$ beads of the system. The training of the network is similar to the two-body model. However, now, only the three-body unit networks are trained, while the pre-trained two-body network is kept fixed, in order to obtain a three-body correction to the two-body model. The whole network defines the $U^{(n,2)}(\mathbf{x}) + U^{(n,3)}(\mathbf{x})$ terms of the multi-body expansion (5), and we refer to it as the "2, 3-body model" in the following. Note that in the 2, 3-body model, we do not include a prior energy unit explicitly because it is already included in the two-body model.

We continue to extend the network this way to model expansion (5) by adding higher order corrections, as shown in Fig. 1(b). At the $k$-body order, the purple colored blocks in Fig. 1(b) indicate the previously trained networks, up to the $k-1$-order, that are kept fixed when training the $k$-body network units. At the $k$th order, there are $\binom{n}{k}$ $k$-body units in the network (light blue blocks in Fig. 1). Each of the units takes as input the $k(k-1)/2$ pairwise distances among a set of $k$ CG beads selected among the $n$ beads of the system. The sum of the outputs of all the $k$-body network units captures the $U^{(n,k)}(\mathbf{x})$ term in Eq. (5), and each k-body unit captures a $f_{i_1,i_2,\ldots,i_k}$ non-decomposable function in Eq. (6). The entire model captures the $U^{(n,2)}(\mathbf{x}) + \cdots + U^{(n,k)}(\mathbf{x})$ terms of the multi-body expansion Eq. (5). We refer to the entire model as "2, …, $k$-body model" in the following.

For the special case of 2, 3, 4-body models, we define both a *non-chiral* model and *a chiral* model. If we consider the six pairwise

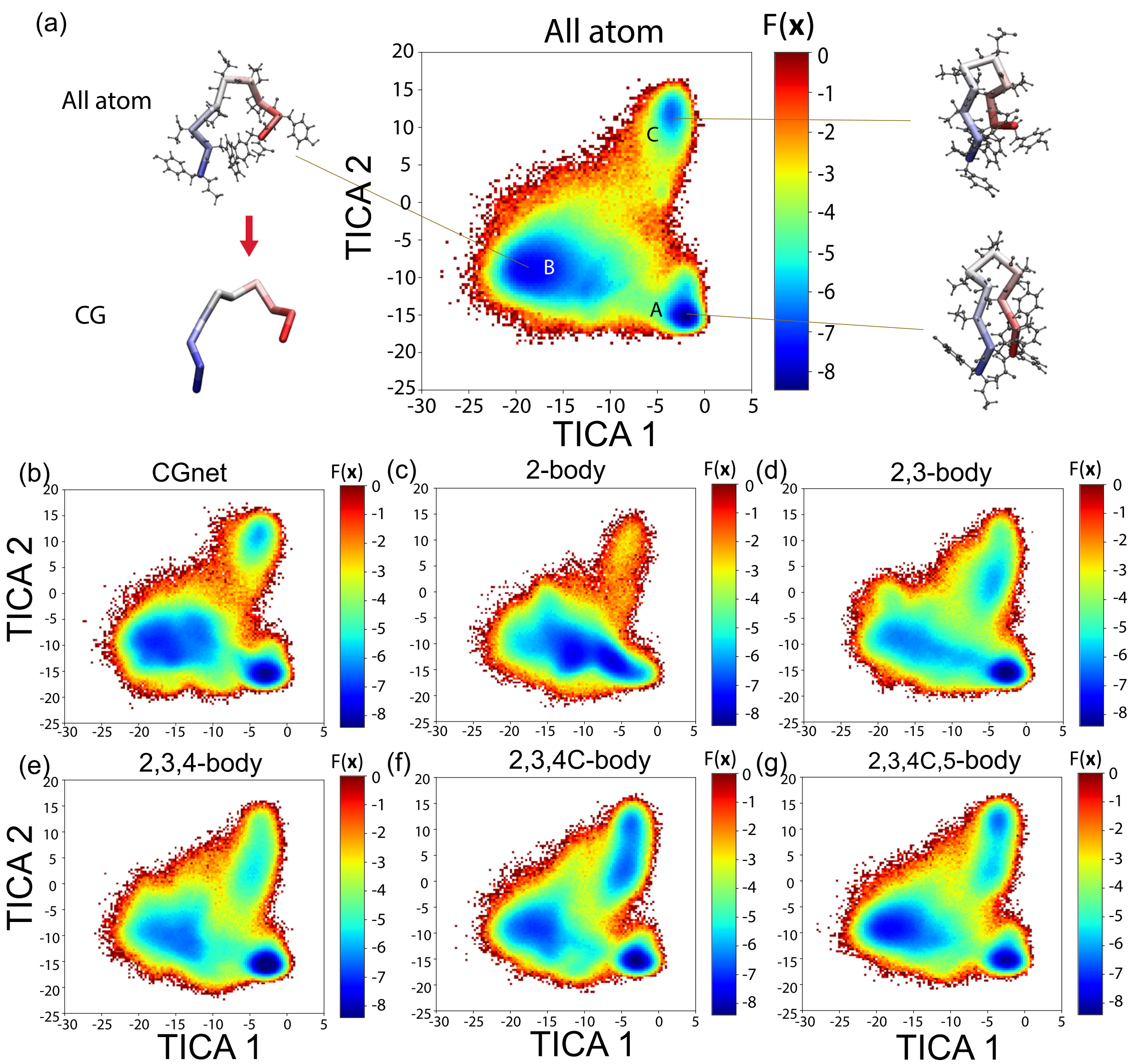


**FIG. 2**. Free energy landscapes associated with the trained multi-body CG models as a function of the two slowest time-lagged independent components of the all-atom simulation. (a) Reference free energy from the all-atom model and representative molecular structures for each of the three metastable states. (b) Full CGnet, (c) two-body model, (d) 2, 3-body model, (e) 2, 3, 4-body model with no chirality, (f) chiral 2, 3, 4*C*-body model, and (g) 2, 3, 4*C*, 5-body model. The free energies are reported in units of $k_BT$.

distances corresponding to a group of four beads (quadrupole), they do not uniquely define the configuration of the beads as arrangements of different chiralities are consistent with the same set of distances. We define a chiral four-body network including up to four-body interactions but with additional information that encodes chirality: we also consider the dihedral angles defined by each set of four CG beads. For each dihedral angle $\phi$ spanned by four beads, we include $\cos(\phi)$ and $\sin(\phi)$ as additional features entering the corresponding four-body unit. We refer to the four-body chiral model as the "2, 3, 4$C$-body model" in the rest of this paper.

With this approach, we could, in principle, construct models at any order, up to $k = n$, as shown in Fig. 1: once a 2, 3, ..., $(k-1)$-body model is trained, we keep it fixed and add $\binom{n}{k}$ different $k$-body unit networks to it. However, the number of $k$-body units increases rapidly with $k$, and the memory requirements for the training become quickly prohibitive. In this work, we consider up to five-body terms in expansion (5) and compare the results with the "vanilla" CGnet where interactions up to any order are included (as all the input features enter one large dense neural network).

## D. Training and simulation of the multi-body models

For the two-body CG model, a three-stage fivefold cross-validation is conducted to find the optimal hyperparameters (network depth, the number of nodes per layer, and Lipschitz regularization strength) of the network units as follows. In the first stage, we fix the number of neural per layer and Lipschitz regularization strength as some finite value, only sweeping the number of network layers. For each hyperparameter combination, we conduct fivefold cross-validation and we identify the optimal hyperparameters as the ones associated with the smallest cross-validation score. By sweeping the number of network layers, we identify the optimal network depth. In the second stage, we fix the number of network layers to the optimal value and sweep the number of nodes per layer, while fixing the Lipschitz regularization strength. We then identify the optimal network width as the minimum cross-validation score. In the third stage, we proceed similarly by fixing the network depth and width to the optimal values and sweeping on the value of the Lipschitz regularization strength and identifying its optimal value with the one associated with the minimum cross-validation score. The Adam optimizer with a mini-batch stochastic gradient descent is used to train the network units.[89,90] The hyperparameters resulting from cross-validation are reported in Table I. When the optimal hyperparameters are selected, the final energy model is defined as the average of the five models corresponding to each fold in the cross-validation at the optimal values of the hyperparameters.

For the multi-body models, we follow the same cross-validation procedure as for the two-body model to obtain the optimal values of the hyperparameters. However, at each order, only the hyperparameters for the network unit that is added are optimized by cross-validation, while the underlying lower order networks are kept the same as in the lower order models. For example, the 2, 3-body model contains the two-body network previously trained and additional three-body units (see Fig. 1). In the 2, 3-body model, the hyperparameters of the additional three-body network units are optimized (the same hyperparameters are used in each unit), while the two-body component is kept fixed. As higher order units describe more complex interactions, their network structures are not necessarily the same at every order.

After a multi-body CGnet model has been obtained, we simulate it by numerically integrating the overdamped Langevin dynamics equation with the corresponding CG potential to generate trajectories and explore the free energy landscape,

$$\mathbf{x}_{t+\tau} = \mathbf{x}_t - \tau \frac{D}{k_B T} \nabla U(\mathbf{x}_t) + \sqrt{2\tau D}\eta, \tag{7}$$

where $\mathbf{x}_t$ ($\mathbf{x}_{t+\tau}$) is a CG configuration at time $t$ ($t+\tau$), $\tau$ is the time step, $D$ is the diffusion constant, and $\eta$ is a Gaussian random variable with zero-mean and unit-standard deviation. As in our previous work,[61] to sample the effective multi-body potential more efficiently, we generate 100 independent trajectories in parallel, with initial configurations randomly sampled from the original dataset. In order to visualize and compare the results, in the following, we

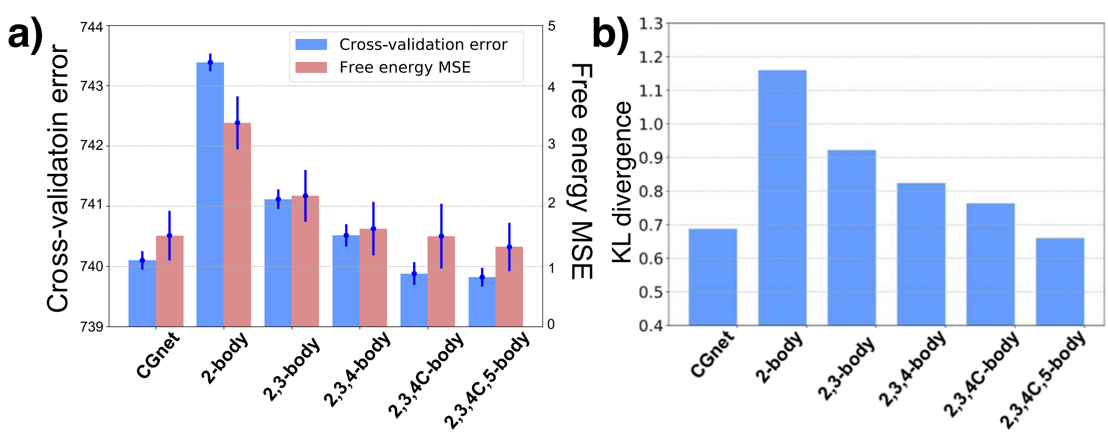


**FIG. 3**. (a) Cross-validation error (blue bars) and free energy mean square error, MSE (red bars), for the different CG models studied. The units for the CV-error are $[kcal/(mol \cdot \AA)]^2$, while the free energy MSE is measured in $(k_B T)^2$. (b) KL divergence between the equilibrium distribution of the different CG models studied and the reference atomistic model.

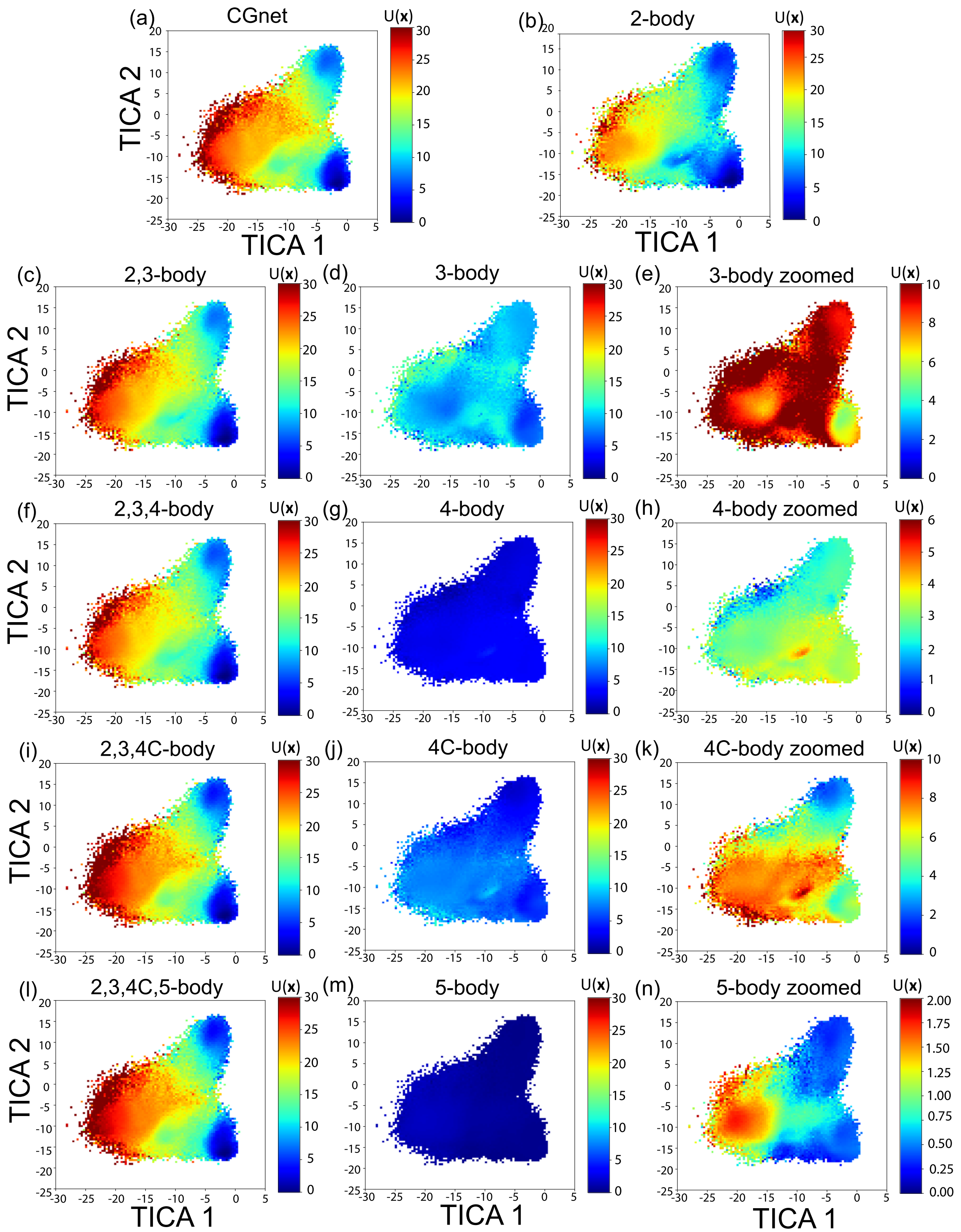


**FIG. 4**. CG potential energy for the different multi-body models: (a) CGnet, (b) two-body, (c) 2, 3-body, (f) 2, 3, 4-body, (i) 2, 3, 4*C*-body, and (l) 2, 3, 4*C*, 5-body model. The energy contributions of the different terms in the multibody expansions are also reported for (d) three-body, (g) four-body, (j) 4*C*-body, and (m) five-body term, with the same energy scale as used for the total CG energy. As the energy differences are relatively small with respect to the energy gap between folded and unfolded states, the same multibody contributions are shown with a color scale zooming in a smaller energy range for the (e) three-body, (h) four-body, (k) 4*C*-body, and (n) five-body term.

project the trajectories of each model onto the space spanned by two collective coordinates: the first two TICA coordinates[91,92] of the all-atom system. Free energy surfaces (Fig. 2) are then computed as the negative logarithm of a two-dimensional histogram over the TICA coordinates.

## III. RESULTS

We apply the multi-body decomposition described above to study a series of CG models for chignolin, a 10 amino acid mini-protein [Fig. 2(a)].[93] The reference all-atom trajectories were obtained by simulating the system with 1881 water molecules for a total of 5820 atoms. All the CG models studied here consist of 10 beads, located at the position of the $C_\alpha$ atoms along the protein backbone [Fig. 2(a)]. The reference all-atom free energy, shown in Fig. 2(a), exhibits three minima, where minimum A corresponds to the folded state, B corresponds to the unfolded state, and C corresponds to the misfolded state. Typical configurations from these three states are shown in Fig. 2(a).

Figure 3 reports the cross-validation error, the free energy MSE, and the KL divergence (computed as in our previous work[62]) of each model with respect to the reference atomistic model [Fig. 2(a)]. Several notable results are apparent from Fig. 3: first, the multi-body CG expansion converges slowly, while the 2, 3-body model presents a significant improvement over the two-body model; the errors are still larger than the corresponding ones for the "vanilla" CGnet up to the 2, 3, 4*C*-body model. More quantitatively, while the addition of three-body interactions lowers the MSE of a little more than one $k_BT$, the difference in MSE between the 2, 3, 4*C*, 5-body model and the 2, 3-body model is still about 1/2 $k_BT$. The slow convergence is quite evident also in the gradual reduction of the KL divergence with the addition of multi-body terms [Fig. 3(b)].

Additionally, in all three measures of the error reported, the errors associated with the 2, 3, 4*C*, 5-body model are *smaller* than for the vanilla CGnet model: the neural network potential including up to five-body terms appears to outperform the neural network potential where higher order interactions are included. The same trend appears in the learning curves of these models: the validation error for the 2, 3, 4*C*, 5-body model is smaller than the validation error of CGNet at every step of the training. This result suggests that the original CGnet overfits the training data and that the multi-body expansion acts as a useful form of implicit regularization or inductive bias.

Free energy landscapes associated with the different models are obtained by means of overdamped Langevin simulations (Sec. II D) and are shown in Fig. 2, together with the reference free energy landscape of the all-atom model [Fig. 2(a)] and of the original CGnet [Fig. 2(b)]. The comparison of these free energy landscapes echoes the results illustrated by Fig. 3. While the two-body model [Fig. 4(c)] does not show a separation between the folded and unfolded states of the protein, the addition of three-body terms in the 2, 3-body model allows us to clearly identify the folded, unfolded, and misfolded states on the free energy landscape [Fig. 4(d)]. The free energy landscape becomes progressively closer to the reference one when four-body [Fig. 4(e)] and chiral four-body interactions [Fig. 4(f)] are added, and it is in good quantitative agreement for the 2, 3, 4*C*, 5-body model [Fig. 4(g)]. The free energy landscape of the latter is smoother and visually closer to the reference than the original CGnet model, supporting the hypothesis of a regularizing role of the multi-body expansion.

It is interesting to consider the CG potential energy contribution to the free energy for the different models. In Fig. 4, we report the values of the CG energy for all configurations sampled by the all-atom model, projected onto the same two-dimensional space of the first two TICA coordinates obtained from the all-atom data. Figures 4(a)–(c), 4(f), 4(i), and 4(l) show the total CG energy of the different models, including the original CGnet [Fig. 4(a)]. The different energy landscapes appear all surprisingly similar, with only the two-body model being clearly different from the others: all energy surfaces show a significant energy minimum corresponding to the folded state and an additional minimum corresponding to the misfolded state, while the configurations in the unfolded state have significantly higher energy. However, the small differences among these energy landscapes are associated with markedly different *free* energy landscapes. Figures 4(d), 4(g), 4(j), and 4(m) show the incremental differences in CG energy corresponding to the different terms in the multi-body expansion (5). As the energy differences are relatively small with respect to the energy gap between folded and unfolded states, Figs. 4(e), 4(h), 4(k), and 4(n) show the same CG energies with a color scale zooming in a smaller energy range.

## IV. CONCLUSIONS

We have presented the results of a multi-body expansion of a CG model for a small protein, chignolin. This is made possible by constructing a neural network architecture for the CG potential in a manner resembling a multi-body energy expansion. Using this approach, we can separate the different terms in the multi-body expansion and evaluate their contribution to the energy and free energy landscapes. Not surprisingly, CG potentials including only pairwise interactions (in addition to angle and dihedral terms between adjacent CG beads) fail to reproduce the correct folding landscape of the protein, even at the qualitative level.

Perhaps more surprisingly, the CG multi-body expansion converges slowly for our model miniprotein: Only when the CG potential includes up to five-body terms, the free energy associated with the CG model appears remarkably similar to the reference all-atom free energy as a function of the same collective coordinates. As only one model system is studied here, we cannot easily generalize these results. However, at least for the case of the system considered here, such a slow convergence of the multi-body expansion is in contrast to the fast convergence of the multi-body expansion, capturing the behavior of the Born–Oppenheimer potential energy surface (PES). In the latter, three-body terms can be as large as 15%–20% of the total interaction energy, while four-body terms provide on average only a 1% energy contribution.[94] This fast convergence has allowed for the development of very accurate analytical models for the PES of water from high-level quantum mechanical calculations for low-order interactions. On the other hand, if a slow convergence of the multi-body expansion also holds for other CG protein models, it makes it more challenging to obtain explicit analytical expressions for their effective energy functions. In principle, we expect that, when extended to the recently proposed transferable neural network architecture for the design of CG models,[62,63] multi-body expansion could disentangle the different contributions of interactions between different groups of residues and analytical expression could then be considered (e.g., by means of permutationally invariant polynomials[95]). However, in addition to the slow convergence of the multi-body expansion, the large number of combinations for the different residues' clusters makes this task much more daunting than what has been possible for the characterization of bulk water PES.[94,96] It is important to note the different nature of the multi-body terms in Born–Oppenheimer PES or CG models. In PES, such terms are directly linked to the quantum mechanical description of the system and the delocalized nature of the electronic structure. In CG models, they emerge as a result of the renormalization of atomistic degrees of freedom. We expect multi-body terms to play a more significant role when the dimensionality reduction is strong. Here, we have used a pretty aggressive coarse-graining scheme from a solvated atomistic description to a $C_\alpha$-only resolution. It is likely that even for the same system, a different coarse-graining mapping could change the relative importance of the multi-body terms. We believe that the absence or limited presence of multi-body terms in traditional CG models has hindered the design of transferable CG models. It remains to be seen how the inclusion of multi-body terms in the form of neural network potentials changes the delicate balance between accuracy and transferability in CG models.

It is also worth noting that, at least for the chignolin system considered here, a CG neural network model truncating the multi-body expansion to five-body terms performs better than a model with a CG energy built as a fully connected neural network, which, in principle, includes interactions at any order. This result suggests a possible overfitting of the CGnet potential and the implicit regularizing effect of the multi-body expansion.

In terms of computational cost, MD simulations with CG neural network potentials are still slower (by about a factor of 10) than MD simulations with conventional CG potentials where the potential energy is composed of pairwise interactions with a given functional form or tabulated values. However, CG neural network models are already faster (by a factor 5–50 depending on the system and on the network architecture) than the corresponding atomistic model in explicit solvent. The studies that have so far been presented with CG neural network potentials are still exploratory in nature, and no significant effort has been devoted to the code optimization for simulation speed. We believe that this problem will be addressed in the near future as this field matures.

This study was performed using the small protein chignolin as a model system. We can only speculate if the conclusions from this study will also apply to larger proteins; future work will address this question.

## ACKNOWLEDGMENTS

We thank Francesco Paesani for insightful discussions and Gianni de Fabritiis and Adria Perez for the chignolin atomistic trajectories. This work was supported by the National Science Foundation (Grant Nos. CHE-1738990, CHE-1900374, and PHY-1427654), the Welch Foundation (Grant No. C-1570), the German MATH+ Excellence Cluster (Grant Nos. AA1-6 and EF1-2), the Deutsche Forschungsgemeinschaft (Grant Nos. SFB 1114/C03, SFB/TRR 186/A12, and SFB 1078/C7), the European Commission (Grant No. ERC CoG 772230 "ScaleCell"), and the Einstein Foundation Berlin. Simulations have been performed on the computer clusters of the Center for Research Computing at Rice University, supported, in part, by the Big-Data Private-Cloud Research Cyberinfrastructure MRI-award (NSF, Grant No. CNS-1338099).

## DATA AVAILABILITY

The data used for the training of the models are available online at https://figshare.com/articles/dataset/Chignolin_Simulations/13858898, while the source codes used to produce the results presented in this paper are available online at https://github.com/ClementiGroup/CGmanybody.git.

## REFERENCES


[1] R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D. E. Shaw, "Pathway and mechanism of drug binding to G-protein-coupled receptors," Proc. Natl. Acad. Sci. U. S. A. **108**, 13118–13123 (2011).

[2] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," Science **334**(6055), 517–520 (2011).

[3] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, "Activation pathway of Src kinase reveals intermediate states as targets for drug design," Nat. Commun. **5**, 3397 (2014).

[4] N. Plattner and F. Noé, "Protein conformational plasticity and complex ligand binding kinetics explored by atomistic simulations and Markov models," Nat. Commun. **6**, 7653 (2015).

[5] N. Plattner, S. Doerr, G. D. Fabritiis, and F. Noé, "Protein-protein association and binding mechanism resolved in atomic detail," Nat. Chem. **9**, 1005–1011 (2017).

[6] F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schöneberg, J. Clarke, C. Freund, T. R. Weikl, and F. Noé, "Protein-ligand kinetics on the seconds timescale from atomistic simulations," Nat. Commun. **8**, 1095 (2017).

[7] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," J. Comput. Phys. **23**(2), 187–199 (1977).

[8] J. Kästner, "Umbrella sampling," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**(6), 932–942 (2011).

[9] R. H. Swendsen and J.-S. Wang, "Replica Monte Carlo simulation of spin-glasses," Phys. Rev. Lett. **57**, 2607–2609 (1986).

[10] R. M. Neal, "Sampling from multimodal distributions using tempered transitions," Stat. Comput. **6**(4), 353–366 (1996).

[11] J. W. Pitera and W. Swope, "Understanding folding and design: Replica-exchange simulations of 'Trp-cage' miniproteins," Proc. Natl. Acad. Sci. U. S. A. **100**, 7587–7592 (2003).

[12] L. S. Stelzl and G. Hummer, "Kinetics from replica exchange molecular dynamics simulations," J. Chem. Theory Comput. **13**(8), 3927–3935 (2017).

[13] G. R. Bowman, D. L. Ensign, and V. S. Pande, "Enhanced modeling via network theory: Adaptive sampling of Markov state models," J. Chem. Theory Comput. **6**(3), 787–794 (2010).

[14] J. Preto and C. Clementi, "Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics," Phys. Chem. Chem. Phys. **16**, 19181–19191 (2014).

[15] E. Hruska, J. R. Abella, F. Nüske, L. E. Kavraki, and C. Clementi, "Quantitative comparison of adaptive sampling methods for protein dynamics," J. Chem. Phys. **149**(24), 244119 (2018).

[16] F. Müller-Plathe, "Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back," ChemPhysChem **3**(9), 754–769 (2002).

[17] S. J. Marrink, A. H. de Vries, and A. E. Mark, "Coarse grained model for semiquantitative lipid simulations," J. Phys. Chem. B **108**(2), 750–760 (2004).

[18] S. Izvekov and G. A. Voth, "A multiscale coarse-graining method for biomolecular systems," J. Phys. Chem. B **109**(7), 2469–2473 (2005).

[19]C. Clementi, "Coarse-grained models of protein folding: Toy-models or predictive tools?," Curr. Opin. Struct. Biol. **18**, 10–15 (2008).
[20]A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, "AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing," J. Phys. Chem. B **116**(29), 8494–8503 (2012).
[21]M. G. Guenza, M. Dinpajooh, J. McCarty, and I. Y. Lyubimov, "Accuracy, transferability, and efficiency of coarse-grained models of molecular liquids," J. Phys. Chem. B **122**(45), 10257–10278 (2018).
[22]W. G. Noid, "Perspective: Coarse-grained models for biomolecular systems," J. Chem. Phys. **139**(9), 090901 (2013).
[23]A. V. Sinitskiy, M. G. Saunders, and G. A. Voth, "Optimal number of coarse-grained sites in different components of large biomolecular complexes," J. Phys. Chem. B **116**(29), 8363–8374 (2012).
[24]L. Boninsegna, R. Banisch, and C. Clementi, "A data-driven perspective on the hierarchical assembly of molecular structures," J. Chem. Theory Comput. **14**(1), 453–460 (2018).
[25]Z. Li, G. P. Wellawatte, M. Chakraborty, H. A. Gandhi, C. Xu, and A. D. White, "Graph neural network based coarse-grained mapping prediction," Chem. Sci. **11**, 9524–9531 (2020).
[26]W. Wang and R. Gómez-Bombarelli, "Coarse-graining auto-encoders for molecular dynamics," npj Comput. Mater. **5**(1), 125 (2019).
[27]S. O. Nielsen, C. F. Lopez, G. Srinivas, and M. L. Klein, "A coarse grain model for *n*-alkanes parameterized from surface tension data," J. Chem. Phys. **119**(14), 7043–7049 (2003).
[28]S. Matysiak and C. Clementi, "Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: How far can a minimalist model go?," J. Mol. Biol. **343**, 235–248 (2004).
[29]S. Matysiak and C. Clementi, "Minimalist protein model as a diagnostic tool for misfolding and aggregation," J. Mol. Biol. **363**, 297–308 (2006).
[30]J. Chen, J. Chen, G. Pinamonti, and C. Clementi, "Learning effective molecular models from experimental observables," J. Chem. Theory Comput. **14**(7), 3849–3858 (2018).
[31]A. P. Lyubartsev and A. Laaksonen, "Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach," Phys. Rev. E **52**(4), 3730–3737 (1995).
[32]M. Praprotnik, L. D. Site, and K. Kremer, "Multiscale simulation of soft matter: From scale bridging to adaptive resolution," Annu. Rev. Phys. Chem. **59**(1), 545–571 (2008).
[33]Y. Wang, W. G. Noid, P. Liu, and G. A. Voth, "Effective force coarse-graining," Phys. Chem. Chem. Phys. **11**(12), 2002 (2009).
[34]M. S. Shell, "The relative entropy is fundamental to multiscale and inverse thermodynamic problems," J. Phys. Chem. **129**(14), 144108 (2008).
[35]F. Nüske, L. Boninsegna, and C. Clementi, "Coarse-graining molecular systems by spectral matching," J. Chem. Phys. **151**(4), 044116 (2019).
[36]W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, "The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models," J. Chem. Phys. **128**(24), 244114 (2008).
[37]H. Wang, C. Junghans, and K. Kremer, "Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining?," Eur. Phys. J. E **28**(2), 221–229 (2009).
[38]L. Larini and J.-E. Shea, "Coarse-grained modeling of simple molecules at different resolutions in the absence of good sampling," J. Phys. Chem. B **116**(29), 8337–8349 (2012).
[39]C. Clementi, H. Nymeyer, and J. N. Onuchic, "Topological and energetic factors: What determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? Investigation for small globular proteins," J. Mol. Biol. **298**(5), 937–953 (2000).
[40]S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, "The MARTINI force field: Coarse grained model for biomolecular simulations," J. Phys. Chem. B **111**(27), 7812–7824 (2007).
[41]M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin, "Three-body interactions improve the prediction of rate and mechanism in protein folding models," Proc. Natl. Acad. Sci. U. S. A. **101**(42), 15088–15093 (2004).
[42]L. Larini, L. Lu, and G. A. Voth, "The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials," J. Chem. Phys. **132**(16), 164107 (2010).
[43]H. Kobayashi, P. B. Rohrbach, R. Scheichl, N. B. Wilding, and R. L. Jack, "Correction of coarse-graining errors by a two-level method: Application to the Asakura-Oosawa model," J. Chem. Phys. **151**(14), 144108 (2019).
[44]J. W. Wagner, T. Dannenhoffer-Lafage, J. Jin, and G. A. Voth, "Extending the range and physical accuracy of coarse-grained models: Order parameter dependent interactions," J. Chem. Phys. **147**(4), 044113 (2017).
[45]A. Tsourtis, V. Harmandaris, and D. Tsagkarogiannis, "Parameterization of coarse-grained molecular interactions through potential of mean force calculations and cluster expansion techniques," Entropy **19**(8), 395 (2017).
[46]G. A. Papoian, J. Ulander, and P. G. Wolynes, "Role of water mediated interactions in protein-protein recognition landscapes," J. Am. Chem. Soc. **125**(30), 9170–9178 (2003).
[47]V. Molinero and E. B. Moore, "Water modeled as an intermediate element between carbon and silicon," J. Phys. Chem. B **113**(13), 4008–4016 (2009).
[48]A. Das and H. C. Andersen, "The multiscale coarse-graining method. VIII. Multiresolution hierarchical basis functions and basis function selection in the construction of coarse-grained force fields," J. Chem. Phys. **136**(19), 194113 (2012).
[49]A. Das and H. C. Andersen, "The multiscale coarse-graining method. IX. A general method for construction of three body coarse-grained force fields," J. Chem. Phys. **136**(19), 194114 (2012).
[50]C. Scherer and D. Andrienko, "Understanding three-body contributions to coarse-grained force fields," Phys. Chem. Chem. Phys. **20**(34), 22387–22394 (2018).
[51]G. DAdamo, A. Pelissetto, and C. Pierleoni, "Coarse-graining strategies in polymer solutions," Soft Matter **8**(19), 5151 (2012).
[52]J. Yang, G. Zhu, D. Tong, L. Lu, and Z. Shen, "B-spline tight frame based force matching method," J. Comput. Phys. **362**, 208–219 (2018).
[53]S. P. Leelananda, Y. Feng, P. Gniewek, A. Kloczkowski, and R. L. Jernigan, "Statistical contact potentials in protein coarse-grained modeling: From pair to multi-body potentials," in *Multiscale Approaches to Protein Modeling* (Springer, New York, 2010), pp. 127–157.
[54]M. T. Zimmermann, S. P. Leelananda, P. Gniewek, Y. Feng, R. L. Jernigan, and A. Kloczkowski, "Free energies for coarse-grained proteins by integrating multibody statistical contact potentials with entropies from elastic network models," J. Struct. Funct. Genomics **12**(2), 137–147 (2011).
[55]P. Gniewek, S. P. Leelananda, A. Kolinski, R. L. Jernigan, and A. Kloczkowski, "Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models," Proteins **79**(6), 1923–1929 (2011).
[56]S. T. John and G. Csányi, "Many-body coarse-grained interactions using Gaussian approximation potentials," J. Phys. Chem. B **121**(48), 10934–10949 (2017).
[57]C. Scherer, R. Scheid, D. Andrienko, and T. Bereau, "Kernel-based machine learning for efficient simulations of molecular liquids," J. Chem. Theory Comput. **16**(5), 3194–3204 (2020).
[58]J. Jin, Y. Han, and G. A. Voth, "Coarse-graining involving virtual sites: Centers of symmetry coarse-graining," J. Chem. Phys. **150**(15), 154103 (2019).
[59]J. F. Dama, J. Jin, and G. A. Voth, "The theory of ultra-coarse-graining. 3. Coarse-grained sites with rapid local equilibrium of internal states," J. Chem. Theory Comput. **13**(3), 1010–1022 (2017).
[60]J. Schöneberg and F. Noé, "ReaDDy—A software for particle based reaction diffusion dynamics in crowded cellular environments," PLoS ONE **8**, e74261 (2013).
[61]J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé, and C. Clementi, "Machine learning of coarse-grained molecular dynamics force fields," ACS Cent. Sci. **5**, 755 (2019).
[62]B. E. Husic, N. E. Charron, D. Lemm, J. Wang, A. Pérez, M. Majewski, A. Krämer, Y. Chen, S. Olsson, G. de Fabritiis, F. Noé, and C. Clementi, "Coarse graining molecular dynamics with graph neural networks," J. Chem. Phys. **153**(19), 194101 (2020).
[63]J. Ruza, W. Wang, D. Schwalbe-Koda, S. Axelrod, W. H. Harris, and R. Gómez-Bombarelli, "Temperature-transferable coarse-graining of ionic liquids

with dual graph convolutional neural networks," J. Chem. Phys. **153**(16), 164501 (2020).
[64]L. Zhang, J. Han, H. Wang, R. Car, and W. E, "DeePCG: Constructing coarse-grained models via deep neural networks," J. Chem. Phys. **149**(3), 034101 (2018).
[65]K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, "Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions," Nat. Commun. **10**(1), 5024 (2019).
[66]J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[67]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**(13), 136403 (2010).
[68]A. P. Bartók, M. J. Gillan, F. R. Manby, and G. Csányi, "Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water," Phys. Rev. B **88**(5), 054104 (2013).
[69]A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules," Sci. Adv. **3**(12), e1701816 (2017).
[70]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**(5), 058301 (2012).
[71]J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost," Chem. Sci. **8**(4), 3192–3203 (2017).
[72]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys. **148**(24), 241733 (2018).
[73]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 13890 (2017).
[74]K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**(24), 241722 (2018).
[75]A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, "Symmetry-adapted machine learning for tensorial properties of atomistic systems," Phys. Rev. Lett. **120**(3), 036002 (2018).
[76]G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials," J. Chem. Phys. **148**(24), 241730 (2018).
[77]T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, "Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions," J. Chem. Phys. **148**(24), 241725 (2018).
[78]L. Zhang, J. Han, H. Wang, W. A. Saidi, R. Car, and E. Weinan, "End-to-end symmetry preserving inter-atomic potential energy model for finite and extended," in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31; available at https://proceedings.neurips.cc/paper/2018/file/e2ad76f2326fbc6b56a45a56c59fafdb-Paper.pdf .
[79]L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).
[80]T. Bereau, R. A. DiStasio, A. Tkatchenko, and O. A. V. Lilienfeld, "Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning," J. Chem. Phys. **148**, 241706 (2018).
[81]S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, e1603015 (2017).
[82]S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," Nat. Commun. **9**, 3887 (2018).
[83]S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, "sGDML: Constructing accurate and data efficient molecular force fields using machine learning," Comput. Phys. Commun. **240**, 38–45 (2019).
[84]F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," Annu. Rev. Phys. Chem. **71**(1), 361–390 (2020).
[85]J. Wang, S. Chmiela, K.-R. Müller, F. Noé, and C. Clementi, "Ensemble learning of coarse-grained molecular dynamics force fields with a kernel approach," J. Chem. Phys. **152**(19), 194106 (2020).
[86]W. G. Noid, P. Liu, Y. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, "The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models," J. Chem. Phys. **128**(24), 244115 (2008).
[87]F. Ercolessi and J. B. Adams, "Interatomic potentials from first-principles calculations: The force-matching method," Europhys. Lett. **26**(8), 583–588 (1994).
[88]G. Ciccotti, T. Lelievre, and E. Vanden-Eijnden, "Projection of diffusions on submanifolds: Application to mean force computation," Commun. Pure Appl. Math. **61**(3), 371–408 (2008).
[89]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 (2014).
[90]J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I: Fundamentals* (Springer, Heidelberg, 1993).
[91]G. Perez-Hernandez, F. Paul, T. Giorgino, G. D. Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys. **139**, 015102 (2013).
[92]C. R. Schwantes and V. S. Pande, "Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9," J. Chem. Theory Comput. **9**, 2000–2009 (2013).
[93]S. Honda, T. Akiba, Y. S. Kato, Y. Sawada, M. Sekijima, M. Ishimura, A. Ooishi, H. Watanabe, T. Odahara, and K. Harata, "Crystal structure of a ten-amino acid protein," J. Am. Chem. Soc. **130**(46), 15327–15331 (2008).
[94]G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero, and F. Paesani, "Modeling molecular interactions in water: From pairwise to many-body potential energy functions," Chem. Rev. **116**(13), 7501–7528 (2016).
[95]B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," Int. Rev. Phys. Chem. **28**(4), 577–606 (2009).
[96]V. Babin, G. R. Medders, and F. Paesani, "Development of a 'first principles' water potential with flexible monomers. II: Trimer potential energy surface, third virial coefficient, and small clusters," J. Chem. Theory Comput. **10**(4), 1599–1607 (2014).