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

Citation for the original published paper (version of record):

N.B. When citing this work, cite the original published paper.
Multi-body effects in a coarse-grained protein force field

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

ARTICLES YOU MAY BE INTERESTED IN

Machine learning meets chemical physics
The Journal of Chemical Physics 154, 160401 (2021); https://doi.org/10.1063/5.0051418

Coarse graining molecular dynamics with graph neural networks
The Journal of Chemical Physics 153, 194101 (2020); https://doi.org/10.1063/5.0026133

When do short-range atomistic machine-learning models fall short?
The Journal of Chemical Physics 154, 034111 (2021); https://doi.org/10.1063/5.0031215
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

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 Clementi1,2,4,5,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.

Published under license by AIP Publishing. https://doi.org/10.1063/5.0041022

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 parallel tempering,8-12 or adaptive sampling.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.\textsuperscript{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.\textsuperscript{11,12,26} Although multil algorithmic strategies have been proposed for the CG mapping,\textsuperscript{27–29} they are most commonly based on physical and chemical intuition, and the optimization of the CG mapping is still an open area of research.\textsuperscript{9} The definition of an effective CG model 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.\textsuperscript{27–29} In top-down methods, a CG model is defined to optimally reproduce a set of global (macroscopic) observables.\textsuperscript{11,12,26} 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\textsuperscript{30–33} or kinetic properties.\textsuperscript{34}

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.\textsuperscript{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\textsuperscript{11–14,40} or by considering the physical nature of the corresponding interactions, as for instance, water-mediated potentials.\textsuperscript{26,28} 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.\textsuperscript{17} Larini \textit{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.\textsuperscript{27} 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.\textsuperscript{48,49} When tested on the modeling of SPC/E water, this approach obtained a significant improvement in the model accuracy.\textsuperscript{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.\textsuperscript{25,26} 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.\textsuperscript{25,26} Four-body terms have also been taken into account in a CG model energy, by considering dihedral potentials between sets of four atoms,\textsuperscript{20,32} and higher-body terms have been built by means of statistical contact potentials.\textsuperscript{20–33}

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 Csa\'nyi and co-workers.\textsuperscript{51,52,61} 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.\textsuperscript{53–55} 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.\textsuperscript{56}

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.\textsuperscript{57} 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.\textsuperscript{58–60} UCG representations are equivalent to interacting-particle reaction dynamics (iPRD),\textsuperscript{61} 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\textsuperscript{1,25} 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,\textsuperscript{65–67} a similar idea has been applied for coarse-graining. To this end, we have used both neural networks (CGnets)\textsuperscript{1,26} and kernel methods\textsuperscript{68} 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.\textsuperscript{68,69}

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.\textsuperscript{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, Voit, and colleagues.26,27 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,22 is called “force matching” and was first employed in the CG context in Ref. 18. However, the force matching MSE cannot directly generalize the results to different systems, the 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 entropy34 and force matching.13,37 In this work, we design protein CG models by means of the force matching method.

In practice, force matching optimizes the parameters θ in the CG potential U(x; θ) through the minimization of the functional,37

\[ \chi^2(\theta) = \left( \| \mathcal{E}_\mathcal{F}(\mathbf{r}(\theta)) + \nabla U(\mathbf{r}, \theta) \|_2^2 \right)_F, \]

where −∇U(x; θ) is the CG force field, \( \mathcal{E}_\mathcal{F}(\mathbf{r}) \) is the instantaneous all-atom force projected onto the CG space,23 and \( \langle \cdot \rangle_F \) is the weighted average over the equilibrium distribution of the atomistic model, i.e., \( \mathbf{r} \sim p(\mathbf{r}) \).

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.13,37

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 multibody terms. However, in practice, the CG model is usually optimized variationally in the space spanned by only two-body (or few-body) functions. 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 \( U(x, \theta) \) can then be decomposed as follows:

\[ U(x, \theta) = U(x_1, x_2, \ldots, x_n; \theta) = \sum_{k=2}^{n} U^{(k)}(x, \theta), \]

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

\[
U^{(k)}(x, \theta) = U^{(2)}(x_1, x_2, \ldots, x_n; \theta)
= \sum_{i_1, i_2, \ldots, i_k} f_{i_1i_2\ldots i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}),
\]

where \( \{i_1, i_2, \ldots, i_k\} \subseteq \{1, 2, \ldots, n\} \) indicates all possible combinations of \( k \) indexes chosen from the full set \( \{1, 2, \ldots, n\} \) and \( f_{i_1i_2\ldots i_k} \) 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 (3) 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)\(^7\) to explicitly learn the terms \( U^{(k)} \) for any \( k \) in a thermodynamically consistent potential \( U(x, \theta) \).

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

In our previous work,\(^6\) 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\(^5\) 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 \( x \) of the CG system are first transformed into a set of roto-translational invariant features \( y \). 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.\(^7\) The excluded volume repulsive terms take the form \( \sum_i \sigma_i \), where \( \sigma_i \) is the distance between CG beads \( i \) and \( j \) for all pairs \( (i, j) \) connected by \( >3 \) bonds. The excluded volume radius \( \sigma \) and exponent \( c \) are fixed to the values obtained in previous work.\(^6\)

<table>
<thead>
<tr>
<th>Model</th>
<th>Network depth</th>
<th>No. of neurons/ layer</th>
<th>Lipschitz reg. strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGnet</td>
<td>5</td>
<td>250</td>
<td>4.0</td>
</tr>
<tr>
<td>2-body</td>
<td>3</td>
<td>60</td>
<td>5.0</td>
</tr>
<tr>
<td>2, 3-body</td>
<td>3</td>
<td>60</td>
<td>5.0</td>
</tr>
<tr>
<td>2, 3, 4-body</td>
<td>3</td>
<td>80</td>
<td>5.0</td>
</tr>
<tr>
<td>2, 3, 4C-body</td>
<td>4</td>
<td>80</td>
<td>5.0</td>
</tr>
<tr>
<td>2, 3, 4C, 5-body</td>
<td>3</td>
<td>70</td>
<td>5.0</td>
</tr>
</tbody>
</table>

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.
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 $U^{(n,2)}(x) + U^{(n,3)}(x)$ 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 $U^{(n,k)}(x)$ term in Eq. (5), and each $k$-body unit captures a $f_{i_1,i_2,...,i_k}$ non-decomposable function in Eq. (6). The entire model captures the $U^{(n,2)}(x) + \cdots + U^{(n,k)}(x)$ 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](https://www.elsevier.com)

**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, 4C-body model, and (g) 2, 3, 4C, 5-body model. The free energies are reported in units of $k_B T$. 


Published under license by AIP Publishing
of Chemical Physics

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. 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 two, 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,

$$x_{t+\tau} = x_t - \tau \frac{D}{k_B T} \nabla U(x_t) + \sqrt{2\tau D} \eta, \tag{7}$$

where $x_t$ ($x_{t+\tau}$) 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, 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 $(\text{kcal/mol} \cdot \AA)^2$, 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.](scitation.org/journal/jcp)
FIG. 4. CG potential energy for the different multi-body models: (a) CGnet, (b) two-body, (c) 2, 3-body, (d) 2, 3-body, (e) 2, 3-body, (f) 2, 3, 4-body, (g) 2, 3, 4-body, (h) 2, 3, 4-body, (i) 2, 3, 4, 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-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-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 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)]. 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) 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-body model. More quantitatively, while the addition of three-body interactions lowers the MSE of a little more than one $k_B T$, the difference in MSE between the 2, 3, 4C, 5-body model and the 2, 3-body model is still about $1/2 k_B T$. 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, 4C, 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, 4C, 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, 4C, 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 (3). 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. 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, 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, multi-body expansion could disentangle the different contributions of interactions between different groups of residues and analytical expression could then be
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).

REFERENCES


