Surveying the energy landscape of coarse-grained mappings

Katherine M. Kidder,[1] M. Scott Shell,[2, a)] and W. G. Noid[1, b)]

[1)] *Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*

[2)] *Department of Chemical Engineering, University of California, Santa Barbara, California 93106, USA*

(Dated: 11 April 2024)

Simulations of soft materials often adopt low-resolution coarse-grained (CG) models. However, the CG representation is not unique and its impact upon simulated properties is poorly understood. In this work, we investigate the space of CG representations for ubiquitin, which is a typical globular protein with 72 amino acids. We employ Monte Carlo methods to ergodically sample this space and to characterize its landscape. By adopting the gaussian network model as an analytically tractable atomistic model for equilibrium fluctuations, we exactly assess the intrinsic quality of each CG representation without introducing any approximations in sampling configurations or in modeling interactions. We focus on two metrics, the spectral quality and the information content, that quantify the extent to which the CG representation preserves low-frequency, large-amplitude motions and configurational information, respectively. The spectral quality and information content are weakly correlated among high resolution representations, but become strongly anti-correlated among low-resolution representations. Representations with maximal spectral quality appear consistent with physical intuition, while low-resolution representations with maximal information content do not. Interestingly, quenching studies indicate that the energy landscape of mapping space is very smooth and highly connected. Moreover, our study suggests a critical resolution below which a "phase transition" qualitatively distinguishes good and bad representations.

Keywords: multiscale modeling; entropy; relative entropy; information theory; Gaussian Network Model; proteins

[a)] Electronic mail: shell@engineering.ucsb.edu

[b)] Electronic mail: wnoid@chem.psu.edu

I. INTRODUCTION

An essential first step in modeling any physical phenomena is determining how to represent the system of interest. The choice of representation is critical to the success of any model because it immediately determines the physical mechanisms that can be described and the questions that can be answered. More pragmatically, the choice of representation also determines the model's range of validity and the computational effort that it requires.

These considerations are particularly important for low resolution, particle-based coarse-grained (CG) models that are widely employed to study soft materials.[1–5] By averaging over atomic details, these CG models enable simulations of length- and time-scales that cannot be effectively addressed with conventional all-atom (AA) models.[6–8] The representation of a CG model is often specified by a "mapping" that projects the AA configuration onto relatively few CG degrees of freedom.[9] Tremendous effort has been invested in developing potentials that accurately reproduce the mapped distribution for these CG degrees of freedom.[10–16] In comparison, relatively few studies have investigated the CG mapping itself. Although one expects that it may profoundly impact the accuracy and utility of CG models,[17] most researchers rely upon their intuition to determine the CG representation.[2]

In recent years, a growing number of studies have investigated the impact of the mapping upon the structural fidelity of CG models that describe interactions with relatively simple molecular mechanics potentials. These comparative case studies suggest that "good" CG maps generate relatively simple mapped distributions for the CG degrees of freedom. In particular, good CG maps should avoid complex intramolecular correlations that cannot be modeled with additive bond, angle, and torsional potentials.[18–23] Similarly, several studies suggest that lower resolution representations often improve the structural fidelity of bottom-up models.[13,24–28] One intuitively expects that, by reducing the density of CG sites, lower resolution CG models reduce the importance of higher order many-body correlations that cannot be accurately modeled with simple pair additive potentials.[5] These case studies provide important insight for guiding the choice of CG mappings for practical calculations. However, they provide relatively little insight into the "intrinsic" quality of a CG mapping. Rather, they assess the consistency of the mapping with the approximations that are introduced by modeling CG interactions with simple molecular mechanics potentials.

Recent studies have also developed automated procedures for determining the CG repre-

sentation, e.g., by graph-based[29–33] and machine learning (ML) approaches.[34–40] In particular, several studies have optimized the CG mapping in order to minimize the error associated with reconstructing AA configurations.[35,36,40] More generally, Potestio and coworkers have proposed optimizing the CG mapping by minimizing the mapping entropy,[38,41–44] which quantifies the configurational information discarded by the mapping.[14,45] While the mapping entropy is very challenging to calculate for realistic AA models, they developed cumulant-[41] and ML-based[38] approximations that, along with enhanced sampling methods,[42] allowed them to explore the entire space of "decimation" mappings that associate each CG site with a single atom. In their very interesting journey through mapping space,[42] they introduced a scalar product and distance for characterizing the metric and topological properties of this mapping space.

Conversely, other studies have proposed optimizing the CG representation in order to preserve the low-frequency, large-amplitude motions that are sampled by the AA model.[46–49] The essential dynamics coarse-graining (ED-CG) method of Voth and coworkers exemplifies this approach.[50–54] The ED-CG methodology first projects an AA trajectory onto the "essential dynamics" subspace that is identified by principal component analysis.[55] The ED-CG methodology then defines the CG mapping by partitioning the atoms into rigid atomic groups that move coherently within this subspace. Similarly, several studies have employed graph-based approaches or network models to identify CG representations that preserve low frequency motions.[29,31,32] While these studies have primarily focused on modeling fluctuations about a single equilibrium structure, Clementi and coworkers have employed diffusion maps and Markov state methods to identify CG sites that move coherently in AA simulations of protein folding.[56] More recently, they employed a variational approach for Markov processes (VAMP) to determine CG representations that accurately described the dynamics of slow transitions.[57]

Our prior work has provided complementary insight into the space of CG mappings.[58–60] By employing the Gaussian Network Model (GNM) for globular proteins,[61–63] we previously derived an exact, analytic expression for the mapped ensemble as a function of the CG mapping.[58] Based upon this result, we introduced two metrics to assess the intrinsic quality of CG mappings.[59] The spectral quality, $\mathcal{Q}$, quantifies the extent to which the CG mapping preserves large-amplitude motions and is closely related to the fitness metric employed in the ED-CG method. Conversely, the information content, I, quantifies the fraction of non-

trivial configurational information that is preserved by the mapping and is closely related to the mapping entropy considered by Potestio and coworkers. By employing these metrics to define an effective energy for the CG mapping, we employed Monte Carlo methods to explore the space of CG maps and to calculate a density of states quantifying the number of maps with a given quality.[59] In distinction to Potestio and coworkers,[38,41–43] we considered CG maps that represented groups of connected amino acids with their mass center. Furthermore, we restricted our sampling to "homogeneous" maps for which each site corresponded to an equal number of amino acids. We observed that $\mathcal{Q}$ and I were almost uncorrelated among the sampled high-resolution maps, but became anti-correlated among low-resolution maps. Most interestingly, the calculated density of states for the spectral quality, $\Omega(\mathcal{Q})$, suggested the existence of a "critical resolution" below which a "phase transition" qualitatively distinguishes maps of high and low quality. Remarkably, Potestio and coworkers employed an analogy with lattice gas models to discover a similar phase transition in the space of decimation mappings.[42]

In the present work, we extend our prior studies in several key aspects. Most importantly, while our prior study was restricted to a "canonical" ensemble of homogeneous maps, here we consider a vastly larger "semi-grand" ensemble of inhomogeneous maps in which sites can correspond to different numbers of amino acids. We introduce a more general algorithm for sampling this semi-grand ensemble and we prove its ergodicity. Moreover, we generalize our definition of the spectral quality, $\mathcal{Q}$, to account for the inhomogeneity of site sizes. Given this ergodic sampling algorithm and an appropriate fitness metric, we examine the landscape of mapping space by characterizing its connectivity and exploring its local and global minimum. We again observe the characteristics of a phase transition distinguishing good and bad maps below a critical resolution in this much more complex semi-grand ensemble.

We note that many prior studies have employed normal mode analysis (NMA), often in combination with the GNM or other network models,[64] to efficiently investigate protein dynamics, folding, and conformational changes,[65–75] as well as to refine low resolution structural data.[76–78] NMA is itself a form of coarse-graining protein dynamics that considers the simplified harmonic potential defined by the Hessian of a potential energy surface.[79] Many studies have further simplified atomic NMA calculations, e.g., by treating blocks of amino acids as rigid bodies[80,81] or by adopting simplified elastic network models to describe atomic interactions.[82,83] The lowest frequency normal modes of network models have often been

employed to identify rigid functional domains.[84,85] Moreover, several studies have developed hierarchies of network models for modeling proteins at various resolutions.[86–89] The quality of such models has been assessed based upon the overlap between observed conformational changes and the lowest frequency normal modes of the coarse-grained network.[64,90] Furthermore, information-theoretic metrics, such as the mode-concentration, have been employed to demonstrate that protein conformational changes are often described by surprisingly few low-frequency normal modes.[91]

While these studies have provided important insights into protein biophysics, the present study differs in several respects. In particular, we exactly coarse-grain the underlying network model by analytically integrating out atomic degrees of freedom. Moreover, rather than assessing our CG models against experimentally observed conformational transitions, we assess the quality of the CG representation by comparing the corresponding mapped ensemble with the underlying AA ensemble. Most importantly, we employ Monte Carlo methods to exhaustively explore and quantitatively assess the entire space of CG representations. Our primary goal is to understand the impact of the CG representation upon the resulting mapped ensemble.

The remainder of this manuscript is organized as follows. Section II reviews the relevant statistical mechanics for coarse-graining the GNM and generalizes the spectral quality to account for variations in site sizes. Section III defines mapping space and describes the sampling methods that we employ to investigate the semi-grand ensemble. Section IV provides additional details of our computational methods. Section V presents the results of our computational studies, while Section VI discusses these results and provides concluding comments. The Appendix provides a brief, self-contained summary of key graph concepts that are useful for this work and also proves the ergodicity of our sampling algorithm.

II. GAUSSIAN NETWORK MODEL

A. Atomic description

1. AA equilibrium distribution

The Gaussian Network Model (GNM) provides a simple description of equilibrium protein dynamics about a reference folded structure.[61–63] The AA GNM represents each amino acid

with its α carbon and introduces an isotropic linear spring between each pair of α carbons that are within a distance, r_c, in the reference structure. Due to the isotropy of these linear springs, the GNM potential separates into independent contributions that govern the motion in each Cartesian direction:

$$u(\mathbf{r}) = \frac{1}{2}\Gamma\delta\mathbf{r}^\dagger\boldsymbol{\kappa}\delta\mathbf{r} \geq 0. \tag{1}$$

Here $\mathbf{r} = (r_1, \ldots, r_n)$ indicates the coordinates for the n atoms in the specified direction, $\delta\mathbf{r} = \mathbf{r} - \mathbf{r}_0$ is the corresponding displacement from the reference coordinates $\mathbf{r}_0$, Γ is a dimensional factor with units of energy/length^2, and † denotes the transpose. The Kirchhoff matrix, $\boldsymbol{\kappa}$, is an $n \times n$ symmetric, positive semi-definite matrix with elements

$$\kappa_{ij} = n_i\delta_{ij} - \theta_{ij}, \tag{2}$$

where n_i is the number of springs connected to atom i, while θ_{ij} is a contact matrix: $\theta_{ij} = 1$ if i and j are connected by springs and otherwise $\theta_{ij} = 0$. Assuming that the protein is connected, the nullspace of $\boldsymbol{\kappa}$ is spanned by the vector $\mathbf{J}_n = (1, \ldots, 1)^\dagger \in \mathbb{R}^n$, which describes free uniform translation of all atoms. The spectral resolution theorem[92] implies that

$$\boldsymbol{\kappa} = \sum_{i=1}^{n-1} \mathbf{u}_i\lambda_i\mathbf{u}_i^\dagger, \tag{3}$$

where $\lambda_i > 0$ for $i = 1, \ldots, n-1$, $\mathbf{u}_0 = n^{-1/2}\mathbf{J}_n$, and $\{\mathbf{u}_0, \ldots, \mathbf{u}_{n-1}\}$ forms a complete orthonormal basis for the space of AA displacements.

The equilibrium distribution for the AA GNM is

$$p_{\mathrm{r}}(\mathbf{r}) \propto \exp[-\beta u(\mathbf{r})] = \exp\left[-\frac{1}{2}\beta\Gamma\delta\mathbf{r}^\dagger\boldsymbol{\kappa}\delta\mathbf{r}\right], \tag{4}$$

where $\beta = 1/k_BT$. Thus, the AA GNM has one free translational degree of freedom and $n-1$ internal degrees of freedom that are described by Gaussian random variables. In the following, averages over AA configurations, $\mathbf{r}$, are defined

$$\langle a(\mathbf{r})\rangle = \int \mathrm{d}\mathbf{r}\, p_{\mathrm{r}}(\mathbf{r})a(\mathbf{r}). \tag{5}$$

2. Configurational information

We quantify the information content of the AA GNM, H_{AA}, by the Kullback-Leibler (KL) divergence[93,94] between the AA equilibrium distribution, $p_{\mathrm{r}}(\mathbf{r}) \propto \exp[-\beta u(\mathbf{r})]$, and the

corresponding uniform distribution, $q_{\rm r}(\mathbf{r}) = L^{-n}$, where L is the length of the system in the specified direction:

$$H_{\rm AA} \equiv {\rm D}_{\rm KL}[p_{\rm r}||q_{\rm r}] \equiv \int {\rm d}\mathbf{r}\, p_{\rm r}(\mathbf{r}) \ln[p_{\rm r}(\mathbf{r})/q_{\rm r}(\mathbf{r})] = (n-1)h_1 + \frac{1}{2}\ln t_{\boldsymbol{\kappa}}. \tag{6}$$

Here $h_1 = \frac{1}{2}\left\{\ln(\beta\Gamma L^2/2\pi) - 1\right\}$ and $t_{\boldsymbol{\kappa}} = n^{-1}\det{}_1\,\boldsymbol{\kappa}$, where $\det_1\boldsymbol{\kappa} = \lambda_1\cdots\lambda_{n-1}$ is the product of the $n-1$ positive eigenvalues of $\boldsymbol{\kappa}$. Note that h_1 is a trivial protein-independent constant accounting for replacing a translational degree of freedom with a vibrational degree of freedom. Moreover, h_1 depends upon the system size, L, which is not only irrelevant to the protein vibrational motion but also must be large, $\beta\Gamma L^2 \gg 1$, in order to analytically evaluate Eq. (6) and other integrals over configuration space. Conversely, $t_{\boldsymbol{\kappa}}$ depends upon the connectivity of the protein that is encoded in the AA Kirchhoff matrix, $\boldsymbol{\kappa}$, and is independent of L. Consequently, we consider $I_{\rm AA} = \frac{1}{2}\ln t_{\boldsymbol{\kappa}}$ to quantify the non-trivial information present in the AA GNM for a given protein.

3. Mass-weighted vibrations

In our prior work,[59] we also considered the magnitude of the mass-weighted vibrations, π, sampled by the AA model, which we referred to as the "vibrational power." In this previous work, we assumed that all CG sites corresponded to an equal number of atoms and, therefore, had equal mass. Consequently, the geometric center coincided with the mass center. In the present work, we relax this restriction and treat CG sites with different mass, such that the geometric and mass centers no longer coincide. Accordingly, we now generalize our previous definition of the vibrational power to allow for atoms of different mass, as in normal mode analysis.[95,96]

The displacement of the geometric center may be computed $\delta r_{\rm cg} = n^{-1}\mathbf{J}_n^\dagger\delta\mathbf{r}$. The projection operator $\mathbb{P}_n = n^{-1}\mathbf{J}_n\mathbf{J}_n^\dagger$ and its orthogonal complement, $\mathbb{Q}_n = \mathbb{1}_n - \mathbb{P}_n$, may be used to decompose the displacement $\delta\mathbf{r}$ into a translation of the geometric center, $\mathbb{P}_n\delta\mathbf{r} = \delta r_{\rm cg}\mathbf{J}_n$, and an internal displacement, $\delta\mathbf{r}_{\rm int} = \mathbb{Q}_n\delta\mathbf{r} = \delta\mathbf{r} - \delta r_{\rm cg}\mathbf{J}_n$. The covariance matrix describing the correlation among these internal displacements is then

$$\mathbf{c}_{\rm int} \equiv \left\langle \delta\mathbf{r}_{\rm int}\delta\mathbf{r}_{\rm int}^\dagger \right\rangle = (\beta\Gamma\boldsymbol{\kappa})^{\rm I}, \tag{7}$$

where $^{\rm I}$ denotes the pseudo-inverse, such that $\boldsymbol{\kappa}^{\rm I} = \sum_{i=1}^{n-1}\mathbf{u}_i\lambda_i^{-1}\mathbf{u}_i^\dagger$.

In the present work it is important to distinguish between the mass and geometric centers. The displacement of the mass center is defined $\delta r_{\rm cm} \equiv m_t^{-1} \sum_{i=1}^{n} m_i \delta r_i$, where m_i is the mass of atom i and $m_t = \sum_{i=1}^{n} m_i$ is the total mass. We define vibrational displacements with respect to the mass center:

$$\delta \mathbf{r}_{\rm v} \equiv \delta \mathbf{r} - \delta r_{\rm cm} \mathbf{J}_n. \tag{8}$$

In order to distinguish fast and slow vibrations, it is convenient to introduce a diagonal mass-weighting matrix $\mathbf{g}_n \equiv \mathrm{diag}(m_i^{1/2})$ for defining mass-weighted coordinates $\delta \overline{\mathbf{r}} = \mathbf{g}_n \delta \mathbf{r}$. When considering mass-weighted coordinates, the vector describing free uniform translation of all atoms becomes $\overline{\mathbf{J}}_n \equiv \mathbf{g}_n \mathbf{J}_n$ such that $\delta r_{\rm cm} = m_t^{-1} \overline{\mathbf{J}}_n^{\dagger} \delta \overline{\mathbf{r}}$. The projection operator for mass-weighted translational motion is $\overline{\mathbb{P}}_n \equiv m_t^{-1} \overline{\mathbf{J}}_n \overline{\mathbf{J}}_n^{\dagger}$, while the projection operator for vibrational motion is $\overline{\mathbb{Q}}_n \equiv \mathbb{1}_n - \overline{\mathbb{P}}_n$. The mass-weighted vibrational displacement is then

$$\delta \overline{\mathbf{r}}_{\rm v} \equiv \mathbf{g}_n \delta \mathbf{r}_{\rm v} = \overline{\mathbb{Q}}_n \delta \overline{\mathbf{r}} = \overline{\mathbb{Q}}_n \mathbf{g}_n \delta \mathbf{r}_{\rm int}, \tag{9}$$

where the last equality follows because $\overline{\mathbb{Q}}_n \mathbf{g}_n = \overline{\mathbb{Q}}_n \mathbf{g}_n \mathbb{Q}_n$. The covariance matrix describing mass-weighted vibrations is

$$\mathbf{c}_{\rm v} \equiv \left\langle \delta \overline{\mathbf{r}}_{\rm v} \delta \overline{\mathbf{r}}_{\rm v}^{\dagger} \right\rangle = \overline{\mathbb{Q}}_n \mathbf{g}_n \mathbf{c}_{\rm int} \mathbf{g}_n \overline{\mathbb{Q}}_n. \tag{10}$$

We define

$$\overline{\boldsymbol{\kappa}} \equiv \mathbf{g}_n^{-1} \boldsymbol{\kappa} \mathbf{g}_n^{-1}, \tag{11}$$

such that

$$\mathbf{c}_{\rm v} = \left(\beta \Gamma \overline{\boldsymbol{\kappa}}\right)^{\rm I}, \tag{12}$$

which can be proved by using the identity $\mathbb{Q}_n \mathbf{g}_n^{-1} = \mathbb{Q}_n \mathbf{g}_n^{-1} \overline{\mathbb{Q}}_n$.

Note that $\Gamma \boldsymbol{\kappa}$ is the Hessian of the GNM potential, while $\mathbf{g}_n^2$ is the Hessian of the GNM kinetic energy with respect to the corresponding velocities. Consequently, the GNM normal mode frequencies, ω_i, satisfy

$$\left| \Gamma \boldsymbol{\kappa} - \omega_i^2 \mathbf{g}_n^2 \right| = \left| \mathbf{g}_n \right|^2 \left| \Gamma \overline{\boldsymbol{\kappa}} - \omega_i^2 \mathbb{1}_n \right| = 0. \tag{13}$$

Thus, the $n-1$ positive eigenvalues of $\Gamma \overline{\boldsymbol{\kappa}}$ are given by the square of the corresponding normal mode frequency, ω_i^2. We define the vibrational power, π, by the mass-weighted vibrations:

$$\pi \equiv \left\langle \sum_{i=1}^{n} m_i \delta r_{{\rm v}i}^2 \right\rangle = \mathrm{Tr}_n \, \mathbf{c}_{\rm v} = k_B T \sum_{i=1}^{n-1} \omega_i^{-2}, \tag{14}$$

where Tr_n denotes the trace over the n AA coordinates.

B. Coarse-grained description

1. *Mapped distribution*

The N-site CG representation of the n-atom AA GNM is defined by a linear mapping, $\mathbf{M} = (c_{Ii}) : \mathbf{r} \to \mathbf{R} = \mathbf{M}(\mathbf{r})$, that determines the CG coordinates by

$$R_I = \sum_{i=1}^{n} c_{Ii} r_i, \tag{15}$$

for each CG site, $I = 1, \ldots, N$. We require that a valid map satisfies two properties: (1) We require that the mapping coefficients are linearly independent such that $\sum_{I=1}^{N} t_I c_{Ii} = 0$ for all $i = 1, \ldots, n$ if and only if $t_I = 0$ for all $I = 1, \ldots, N$. This ensures that the mapped CG coordinates are linearly independent. (2) We also require that the mapping coefficients are nonnegative and properly normalized such that $\sum_{i=1}^{n} c_{Ii} = 1$ for all $i = 1, \ldots, n$, i.e., the mapped coordinates correspond to a convex combination of atomic coordinates. This normalization ensures that the mapping preserves translational motion between the AA and CG representations, i.e., for any $t \in \mathbb{R}$, $\mathbf{M}(\mathbf{r} + t\mathbf{J}_n) = \mathbf{M}(\mathbf{r}) + t\mathbf{J}_N$, where $\mathbf{J}_N \equiv \mathbf{M}\mathbf{J}_n = (1, \ldots, 1)^\dagger \in \mathbb{R}^N$ is a corresponding vector describing translational motion of the N CG sites. As for the AA model, we shall find it convenient to define CG projection operators for translational motion, $\mathbb{P}_N = N^{-1}\mathbf{J}_N\mathbf{J}_N^\dagger$, and for internal motion, $\mathbb{Q}_N = \mathbb{1}_N - \mathbb{P}_N$.

Given the mapping, $\mathbf{M}$, the probability (density) for observing an AA configuration, $\mathbf{r}$, that maps to the CG configuration, $\mathbf{R}$, is given by

$$p_{\mathrm{R}}(\mathbf{R}) \equiv \int \mathrm{d}\mathbf{r}\, p_{\mathrm{r}}(\mathbf{r}) \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})). \tag{16}$$

Foley et al. previously proved that[58]

$$p_{\mathrm{R}}(\mathbf{R}) \propto \exp\left[-\frac{1}{2}\beta\Gamma \delta\mathbf{R}^\dagger \mathbf{K} \delta\mathbf{R}\right], \tag{17}$$

where

$$\mathbf{K} \equiv \mathbf{K}(\mathbf{M}) \equiv \left(\mathbb{Q}_N \mathbf{M} \boldsymbol{\kappa}^{\mathrm{I}} \mathbf{M}^\dagger \mathbb{Q}_N\right)^{\mathrm{I}}, \tag{18}$$

explicitly depends upon the mapping, $\mathbf{M}$. $\mathbf{K}$ does not have exactly the same form as Eq. (2) and, e.g., can have positive off-diagonal elements. Nevertheless, $\mathbf{K}$ preserves several key properties of the AA Kirchhoff matrix, $\boldsymbol{\kappa}$. In particular, Eq. (18) implies that $\mathbf{K}$ is symmetric and positive semi-definite. Moreover, because the mapping coefficients are linearly

independent, $\mathbf{J}_N$ spans the nullspace of $\mathbf{K}$. Therefore, the spectral resolution theorem implies that $\mathbf{K} = \sum_{I=1}^{N-1} \mathbf{v}_I \Lambda_I \mathbf{v}_I^\dagger$, where $\Lambda_I > 0$ for $I = 1, \ldots, N-1$, $\mathbf{v}_0 = N^{-1/2}\mathbf{J}_N$, and $\{\mathbf{v}_0, \ldots, \mathbf{v}_{N-1}\}$ forms a complete orthonormal basis for the space of CG displacements. Thus, the CG representation of the AA GNM preserves one free translational degree of freedom and $N-1$ internal degrees of freedom that are Gaussian random variables, in perfect analogy with Eq. (4). In the following, averages of CG configurations, $\mathbf{R}$, are defined over the mapped ensemble:

$$\langle A(\mathbf{R})\rangle \equiv \int \mathrm{d}\mathbf{R}\, p_{\mathrm{R}}(\mathbf{R}) A(\mathbf{R}) = \langle A(\mathbf{M}(\mathbf{r}))\rangle\,, \tag{19}$$

where the last expression is an average over $\mathbf{r}$ according to Eq. (5). Given the analogy between Eqs. (3)-(5) for the AA model and Eqs. (16)-(19) for its CG representation, the relevant averages of CG observables can be directly determined from Section II A.

2. *Information content*

We quantify the information content of the mapped ensemble, H_{CG}, by the KL divergence between the mapped distribution, p_{R}, and the minimally informative uniform distribution, $q_{\mathrm{R}} = L^{-N}$:

$$H_{\mathrm{CG}} \equiv \mathrm{D}_{\mathrm{KL}}[p_{\mathrm{R}}||q_{\mathrm{R}}] \equiv \int \mathrm{d}\mathbf{R}\, p_{\mathrm{R}}(\mathbf{R}) \ln[p_{\mathrm{R}}(\mathbf{R})/q_{\mathrm{R}}(\mathbf{R})] = (N-1)h_1 + \frac{1}{2}\ln T_{\mathbf{K}}, \tag{20}$$

where $T_{\mathbf{K}} = N^{-1}\mathrm{det}_1\mathbf{K}$. In analogy to the AA model, we define $I_{\mathrm{CG}} \equiv \frac{1}{2}\ln T_{\mathbf{K}}$ as the non-trivial information preserved in the mapped ensemble. The mapping entropy is

$$H_{\mathrm{map}} \equiv H_{\mathrm{AA}} - H_{\mathrm{CG}} = (n-N)h_1 + \frac{1}{2}\ln t_{\boldsymbol{\kappa}}/T_{\mathbf{K}}. \tag{21}$$

While the mapping entropy is an important metric for characterizing CG models, H_{map} includes a "trivial" contribution, $(n-N)h_1$. This trivial contribution is independent of the properties of the protein itself and is simply determined by the number of degrees of freedom that have been eliminated from the model. Consequently, in the present work we consider the metric

$$I(\mathbf{M}) \equiv I_{\mathrm{CG}}/\,I_{\mathrm{AA}} = \ln T_{\mathbf{K}}/\ln t_{\boldsymbol{\kappa}} = \ln\left[N^{-1}\mathrm{det}_1\mathbf{K}\right]/\ln\left[n^{-1}\mathrm{det}_1\boldsymbol{\kappa}\right], \tag{22}$$

which quantifies the fraction of (non-trivial) information that is preserved by the CG representation, $\mathbf{M}$. Note that the mapping entropy may be expressed

$$H_{\mathrm{map}} = (n-N)h_1 + \frac{1}{2}(1-I)\ln t_{\boldsymbol{\kappa}}. \tag{23}$$

Because $\ln t_{\boldsymbol{\kappa}}$ is fixed by the AA model, it follows that minimizing $H_{\rm map}$ (at fixed N) corresponds to maximizing I.

3. Mass-weighted vibrations

As for the AA model, the displacement of the geometric center of the CG configuration is $\delta R_{\rm cg} \equiv N^{-1}\mathbf{J}_N^\dagger \delta\mathbf{R}$. The CG internal displacement is then $\delta\mathbf{R}_{\rm int} \equiv \delta\mathbf{R} - \delta R_{\rm cg}\mathbf{J}_n = \mathbb{Q}_N\delta\mathbf{R}$ and the corresponding covariance matrix is

$$\mathbf{C}_{\rm int} \equiv \left\langle \delta\mathbf{R}_{\rm int}\delta\mathbf{R}_{\rm int}^\dagger \right\rangle = (\beta\Gamma\mathbf{K})^{\rm I} \tag{24}$$

As in our prior studies with the GNM,[58,59] we again assume that $\mathbf{M}$ partitions the set of n atoms, $V = \{1,\ldots,n\}$, such that each atom contributes to a single site. For each site, I, we define a set of atoms, $V_I = \{i|c_{Ii} > 0\}$, that contribute to the site, such that $\cup_{I=1}^N V_I = V$ and $V_I \cap V_J = \varnothing$ if $I \neq J$. We then define the mass, M_I, of site I as the total mass of the atoms in V_I, i.e., $M_I = \sum_{i\in V_I} m_i$. Note that this partitioning preserves the total atomic mass, $M_t = \sum_{I=1}^N M_I = m_t$.

In analogy to the AA model, we define a CG mass-weighting matrix, $\mathbf{G}_N \equiv \mathrm{diag}(M_I^{1/2})$, and CG mass-weighted coordinates, $\delta\overline{\mathbf{R}} \equiv \mathbf{G}_N\delta\mathbf{R}$. Similarly, we define $\overline{\mathbf{J}}_N \equiv \mathbf{G}_N\mathbf{J}_N$ and corresponding projection operators, $\overline{\mathbb{P}}_N \equiv M_t^{-1}\overline{\mathbf{J}}_N\overline{\mathbf{J}}_N^\dagger$, and $\overline{\mathbb{Q}}_N \equiv \mathbb{1}_N - \overline{\mathbb{P}}_N$. The CG mass center displacement is $\delta R_{\rm cm} \equiv M_t^{-1}\sum_{I=1}^N M_I\delta R_I = M_t^{-1}\overline{\mathbf{J}}_N^\dagger\delta\overline{\mathbf{R}}$ and the CG vibrations are defined $\delta\mathbf{R}_{\rm v} \equiv \delta\mathbf{R} - \delta R_{\rm cm}\mathbf{J}_N$. The CG mass-weighted vibrations are then

$$\delta\overline{\mathbf{R}}_{\rm v} \equiv \mathbf{G}_N\delta\mathbf{R}_{\rm v} = \overline{\mathbb{Q}}_N\delta\overline{\mathbf{R}} = \overline{\mathbb{Q}}_N\mathbf{G}_N\delta\mathbf{R}_{\rm int}, \tag{25}$$

where the last expression employs $\overline{\mathbb{Q}}_N\mathbf{G}_N = \overline{\mathbb{Q}}_N\mathbf{G}_N\mathbb{Q}_N$. The mass-weighted covariance matrix is

$$\mathbf{C}_{\rm v} \equiv \left\langle \delta\overline{\mathbf{R}}_{\rm v}\delta\overline{\mathbf{R}}_{\rm v}^\dagger \right\rangle = \overline{\mathbb{Q}}_N\mathbf{G}_N\mathbf{C}_{\rm int}\mathbf{G}_N\overline{\mathbb{Q}}_N = \left(\beta\Gamma\overline{\mathbf{K}}\right)^{\rm I}, \tag{26}$$

where

$$\overline{\mathbf{K}} \equiv \mathbf{G}_N^{-1}\mathbf{K}\mathbf{G}_N^{-1} \tag{27}$$

and we have used the analogous identity $\mathbb{Q}_N\mathbf{G}_N^{-1} = \mathbb{Q}_N\mathbf{G}_N^{-1}\overline{\mathbb{Q}}_N$. The $N-1$ positive eigenvalues of $\Gamma\overline{\mathbf{K}}$ are given by the square of the CG normal mode frequencies, Ω_I^2. We define the vibrational power, Π, of the CG representation, $\mathbf{M}$, by

$$\Pi(\mathbf{M}) \equiv \left\langle \sum_{I=1}^N M_I\delta R_{{\rm v}I}^2 \right\rangle = \mathrm{Tr}_N\,\mathbf{C}_{\rm v} = k_BT\sum_{I=1}^{N-1}\Omega_I^{-2}, \tag{28}$$

where Tr_N denotes the trace. We define the spectral quality, $\mathcal{Q}$, of the CG representation, $\mathbf{M}$, by the fraction of vibrational power that it preserves:

$$\mathcal{Q}(\mathbf{M}) \equiv \Pi/\pi = \mathrm{Tr}_N\, \overline{\mathbf{K}}^{\mathrm{I}} \Big/ \mathrm{Tr}_n\, \overline{\boldsymbol{\kappa}}^{\mathrm{I}} = \sum_{I=1}^{N-1} \Omega_I^{-2} \Bigg/ \sum_{i=1}^{n-1} \omega_i^{-2}. \tag{29}$$

It is important to note that the mass-weighted vibrational covariance matrix, $\mathbf{C}_{\mathrm{v}}$, is not simply obtained by mass-weighting the covariance matrix of internal displacements, $\mathbf{C}_{\mathrm{int}}$. Specifically, if we define $\overline{\mathbf{C}}_{\mathrm{int}} \equiv \mathbf{G}_N \mathbf{C}_{\mathrm{int}} \mathbf{G}_N$ as the matrix obtained by mass-weighting $\mathbf{C}_{\mathrm{int}}$, then $\mathbf{C}_{\mathrm{v}} = \overline{\mathbb{Q}}_N \overline{\mathbf{C}}_{\mathrm{int}} \overline{\mathbb{Q}}_N \neq \overline{\mathbf{C}}_{\mathrm{int}}$ when the CG sites have different masses. Physically, the two differ because $\overline{\mathbf{C}}_{\mathrm{int}}$ mass-weights displacements with respect to the geometric center of the mapped configuration, while $\mathbf{C}_{\mathrm{v}}$ mass-weights vibrations with respect to the mass center of the mapped configuration. The mass centers of AA and mapped configurations always coincide by definition. However, the geometric centers of AA and mapped configurations do not necessarily coincide. In particular, when the CG sites correspond to different numbers of amino acids, the geometric centers of the AA and mapped configurations can dramatically differ. While this distinction may seem a minor technical detail, it has important ramifications for determining the spectral quality of CG representations.

Figure 1 presents a scatter plot comparing $\Pi \equiv \mathrm{Tr}_N\, \mathbf{C}_{\mathrm{v}}$ and $\mathrm{Tr}_N\, \overline{\mathbf{C}}_{\mathrm{int}}$ for 3-site maps sampled from Monte Carlo simulations in the mapping space for ubiquitin. The orange star indicates the map that maximizes Π, which is illustrated at the top right of Fig. 1. This map associates each site with a coherent structural motif and agrees well with our physical intuition. Conversely, the blue star indicates the map that maximizes $\mathrm{Tr}_N\, \overline{\mathbf{C}}_{\mathrm{int}}$. This map, which is illustrated at the top left of Fig. 1, does not agree well with our physical intuition. In particular, this map represents almost the entire protein with a single site, while the remaining sites correspond to individual amino acids. Thus, it is important to properly account for the inhomogeneous mass distribution when evaluating the spectral quality of a CG mapping. Moreover, Eq. (28) indicates that $\Pi(\mathbf{M})$ increases when $\mathbf{M}$ reduces the normal mode frequencies of the mapped Hamiltonian. Consequently, we anticipate that the spectral quality, $\mathcal{Q}(\mathbf{M})$, defined by Eq. (29) is quite similar to the VAMP metric proposed by Clementi and coworkers.[57]

The spectral quality, $\mathcal{Q}$, is also similar to a number of other metrics that have been previously proposed to identify CG sites with coherently moving atomic groups[46,50] and to define quasi-rigid protein domains.[85,97–100] While it does not explicitly employ essential dy-

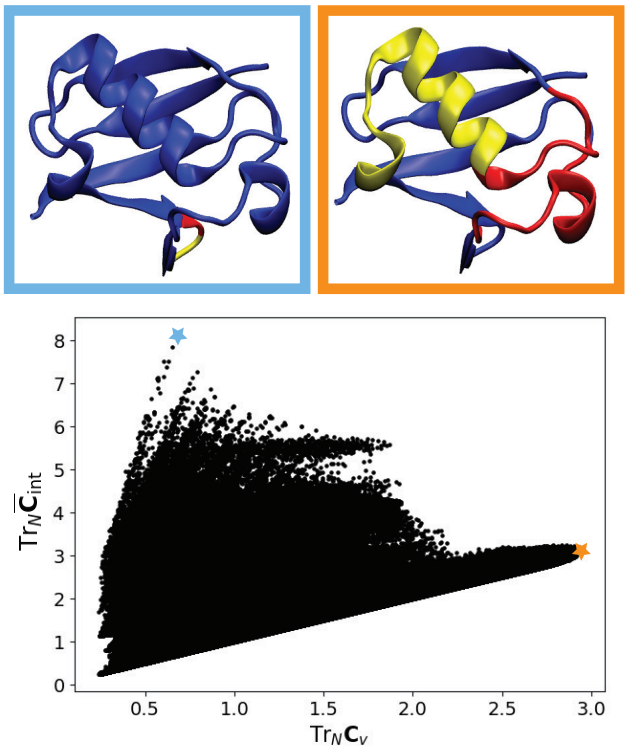

FIG. 1. Scatter plot of $\Pi \equiv \mathrm{Tr}_N\, \mathbf{C}_{\mathrm{v}}$ and $\mathrm{Tr}_N\, \overline{\mathbf{C}}_{\mathrm{int}}$ for 3-site maps sampled from Monte Carlo simulations of the mapping space for ubiquitin. The blue and orange stars indicate the 3-site maps that maximize $\mathrm{Tr}_N\, \overline{\mathbf{C}}_{\mathrm{int}}$ and $\mathrm{Tr}_N\, \mathbf{C}_{\mathrm{v}}$, respectively. The top left (blue) and right (orange) panels present the corresponding maps.

namics analysis,[50,97] Eqs. (28) and (29) demonstrate that $\mathcal{Q}$ similarly emphasizes large scale motions of an underlying AA model. Conversely, although it does not explicitly employ rigidity analysis,[46,50,97] $\mathcal{Q}$ gives little weight to localized, high-frequency motions that are characteristic of quasi-rigid intra-domain motions. Similarly, by emphasizing low-frequency motions, one expects that $\mathcal{Q}$ will also emphasize persistent, slow motions without explicitly considering time-correlation functions.[99] Finally, it is worth noting that $\mathcal{Q}$ emphasizes the spectral properties of a mass-weighted covariance matrix that can be derived from the Laplacian matrix of a graph, as in the SPECTRUS method.[98,100]

III. SAMPLING MAPPING SPACE

A. Graph description of GNM

Graph theoretic ideas provide a useful framework for exploring mapping space. The GNM determines a simple graph for describing the network of interactions within a globular protein. This graph includes a labelled vertex, v_i, for each atom, $i = 1, \ldots, n$, in the AA GNM. For simplicity, we shall use i to identify both the atom index and also the corresponding vertex, v_i. This graph includes an edge, e_{ij}, between two distinct vertices, $i(\neq)j$, if the corresponding atoms are connected by a spring in the AA GNM potential, i.e., if $\theta_{ij} = 1$. The resulting vertex set, $V = \{1, \ldots, n\}$, and edge set, $E = \{e_{ij} | i, j \in V \text{ and } \theta_{ij} = 1\}$, then define a graph, $G = (V, E)$, that we shall refer to as the AA protein graph.

The edge set, E, includes edges that reflect, e.g., hydrogen-bonding, salt-bridge, or dispersive interactions between amino acids. Because the protein is physically connected along the backbone, E also includes edges, $e_{i,i+1}$, between each successive pair of atoms. Moreover, when using a cut-off, $r_c \geq 7.5$ Å, to define GNM bonds, we find that E also includes an edge $e_{i,i+2}$ whenever $i, i+2 \in V$. Consequently, we shall assume that valid protein graphs are singly and doubly connected along the backbone, such that $e_{ij} \in E$ whenever $i, j \in V$ and $|i - j| \leq 2$.

B. Allowed maps as graph partitions

Given an N-site mapping, $\mathbf{M} = (c_{Ii})$, we define a site vertex set, $V_I = \{i | c_{Ii} > 0\}$, by the set of atoms associated with CG site I. We define the size of site I by the number of elements in V_I, i.e., $|V_I|$. We define the site edge set, $E_I = \{e_{ij} \in E | i, j \in V_I\}$, by the set of bonds among the atoms in the vertex set V_I. This then defines a subgraph $G_I = (V_I, E_I)$ for each CG site I.

We consider the space, $\mathcal{M}_N$, of N-site CG representations, $\mathbf{M}$, that satisfy the following properties:

1. $\mathbf{M}$ partitions V into N disjoint, nonempty vertex sets, $V_1, \ldots, V_N$, such that $|V_I| \geq 1$, $\cup_{I=1}^{N} V_I = V$, and $V_I \cap V_J = \varnothing$ if $I \neq J$.

2. Each atomic group, V_I, is "connected" in the sense that each pair of atoms $i, j \in V_I$

can be connected by a sequence of edges in E_I.

3. The mapping coefficients correspond to the mass center for each site I, i.e.,

$$c_{Ii} = \begin{cases} m_i/M_I & \text{if } i \in V_I \\ 0 & \text{otherwise} , \end{cases} \tag{30}$$

where $M_I = \sum_{i \in V_I} m_i$.

Each allowed mapping, $\mathbf{M} \in \mathcal{M}_N$, is then in one-to-one relationship with a partitioning, $[G_1, \ldots, G_N]$, of the AA protein graph, i.e., $\mathbf{M} \sim [G_1, \ldots, G_N]$.

Note that any of the $N!$ permutations of the site labels, $I = 1, \ldots, N$, generates an equivalent CG mapping, $\mathbf{M}$. In order to break this degeneracy, we order the sites based upon the minimal elements in the corresponding atomic groups, i.e., atom $1 \in V_1$, V_2 includes the first atom that is not in V_1, and so on.

C. Equilibrium ensemble

We are interested in characterizing the statistical properties of this N-site mapping space and, in particular, the number of allowed N-site maps, $\Omega_N(\mathcal{Q}, I)$, with a given spectral quality, $\mathcal{Q}$, and information content, I. In order to calculate, $\Omega_N(\mathcal{Q}, I)$, it useful to define a "semi-grand ensemble" in which the probability of sampling an allowed map, $\mathbf{M}$, is

$$\mathcal{P}_N(\mathbf{M}; \beta, \lambda) \propto \exp\left[-\beta\left(\mathcal{E}(\mathbf{M}) + \lambda\sigma^2(\mathbf{M})\right)\right] . \tag{31}$$

Here $\mathcal{E}(\mathbf{M}) = 1 - \mathcal{Q}(\mathbf{M})$ and β is a conjugate inverse temperature such that the "ground state" sampled at $\beta = \infty$ maximizes $\mathcal{Q}$. Additionally, we have defined $\sigma^2(\mathbf{M}) \equiv \text{var}\{|V_1|, |V_2|, \ldots, |V_N|\}$ as the variance in the size of the N sites, while λ is a pressure-like conjugate variable. Equilibrium MC simulations at a given state point (β, λ) will primarily sample CG representations with a corresponding spectral quality, $\overline{\mathcal{E}}(\beta, \lambda)$, and variance, $\overline{\sigma}^2(\beta, \lambda)$. By performing MC simulations for a wide range of β and λ, we can sample the entirety of mapping space, $\mathcal{M}_N$, for a given resolution, N. We can then estimate $\Omega_N(\mathcal{Q}, I)$ (to within a constant prefactor) by reweighting the sampled configurations in the limit $\beta, \lambda \to 0$ such that $\mathcal{P}_N(\mathcal{Q}, I) \propto \Omega_N(\mathcal{Q}, I)$. Previously, we considered a "canonical" ensemble of connected maps, $\mathbf{M}$, in which each site was the same size, i.e., $\sigma^2(\mathbf{M}) = 0$, and which corresponds to $\beta\lambda \to \infty$. The present semi-grand ensemble is vastly larger than this canonical ensemble.

D. Steal move set

We employ a "steal" move set to explore mapping space. Starting from an allowed map, $\mathbf{M} \sim [G_1, \ldots, G_N]$, a steal move generates a new map, $\mathbf{M}'$, by moving an atom i from site I to a new site $J(\neq I)$, creating two new sites $G'_I = G_I - i$ and $G'_J = G_J + i$, while leaving the remaining $N-2$ sites unchanged, i.e., site J steals atom i from site I. The steal map is allowed if the new map, $\mathbf{M}'$, is allowed, i.e., if the modified sites, G'_I and G'_J, are both connected. This requires that removing atom i from site I does not disconnect G_I and that there exists an edge $e_{ij} \in E$ from atom i to an atom $j \in V_J$. Thus, atom i in site I is "stealable" if the following conditions are fulfilled: (1) $|V_I| \geq 2$ so that $V'_I = V_I - \{i\}$ is non-empty; (2) i is not a cut-vertex (i.e., articulation node) of G_I so that $G'_I = G_I - i$ remains connected; and (3) there exists a different site $J(\neq I)$ with an atom $j \in V_J$ that forms an "out-of-site" edge with i, $e_{ij} \in E$, so that $G'_J = G_J + i$ is connected. Note that allowed steal moves are reversible: site I' can steal atom i back from site J' to restore the original sites, G_I and G_J, which both are connected since $\mathbf{M}$ is an allowed map.

Importantly, the steal move set is ergodic and allows us to exhaustively explore all of mapping space. Our proof of this ergodicity proceeds by proving that steal moves can transform any allowable map, $\mathbf{M}$, into a special N-site "monster" map, $\mathbf{M}^*_{Nm}$. The appendix proves this proposition. It is then simple to prove that any two valid maps, $\mathbf{M}$ and $\mathbf{M}'$, can be connected by a series of steal moves. The proposition implies that there exist series of steal moves, $\mathbf{M} \to \mathbf{M}^*_{Nm}$ and $\mathbf{M}' \to \mathbf{M}^*_{Nm}$. Since steal moves can be reversed, the series $\mathbf{M} \to \mathbf{M}^*_{Nm} \to \mathbf{M}'$ connects $\mathbf{M}$ and $\mathbf{M}'$.

Given a map, $\mathbf{M} \sim [G_1, \ldots, G_N]$, we denote the steal move that transfers atom i to site J by the ordered pair (i, J). The set of allowed steal moves, $T = \{(i, J)\}$, and the number of allowed steal moves both vary with the mapping, $\mathbf{M}$, i.e., $T = T(\mathbf{M})$. We determine the set of allowed steal moves, $T(\mathbf{M})$, via the following algorithm:

1. For each distinct pair of sites, (I, J), we determine the corresponding set of intersite edges, $E_{IJ} = \{e_{ij} \in E | i \in V_I, j \in V_J\}$.

2. Each intersite edge, $e_{ij} \in E_{IJ}$, determines two potential steal moves, (i, J) and (j, I). We include each of these potential moves in the set $T^{(0)}_{IJ} = \{(i, J), (j, I) | e_{ij} \in E_{IJ}, i \in V_I, j \in V_J\}$.

3. We compile the set, $T_{IJ}^{(0)}$, for each pair of distinct sites, (I, J), to obtain $T^{(0)} = \cup_{(I,J)} T_{IJ}^{(0)}$.

4. Since each move, (i, J), included in $T^{(0)}$ corresponds to an intersite edge, it will be allowed if it satisfies two conditions: (1) i must not be the only vertex of a site graph G_I; and (2) i must not be a cut-vertex (articulation node) of this site graph G_I. Consequently, we then remove any pair, (i, J), from $T^{(0)}$ that violates these two conditions according to the following algorithm:

   (a) We first determine the size, $|V_I|$, of each site I. If site I contains only one atom, i, then we remove every pair (i, J) from $T^{(0)}$ that would transfer atom i.

   (b) We next identify the cut-vertices that would disconnect each site I. We remove from $T^{(0)}$ every pair, (c, J), that would transfer a cut vertex, c.

After eliminating the potential moves that would eliminate or disconnect a site, we obtain the set, $T(\mathbf{M})$, of valid steal moves. We define the coordination, $C_{\mathbf{M}}$, of the map, $\mathbf{M}$, as the number of maps, $\mathbf{M}'$, that can be reached from $\mathbf{M}$ via a single steal move, i.e., $C_{\mathbf{M}} = |T(\mathbf{M})|$.

E. Monte Carlo algorithm

Given a valid map, $\mathbf{M} \sim [G_1, \ldots, G_N]$, we perform each step of our MC simulation according to the following algorithm:

1. We construct the set of allowed steal moves, $T(\mathbf{M})$, from site $\mathbf{M}$.

2. We select one of the allowed steal moves, (i, J), according to the uniform distribution, $\alpha(\mathbf{M} \to \mathbf{M}') = C_{\mathbf{M}}^{-1} \delta_{\mathbf{M}\mathbf{M}'}$ where $\delta_{\mathbf{M}\mathbf{M}'} = 1$ if $\mathbf{M} \to \mathbf{M}'$ is an allowed steal move; otherwise $\delta_{\mathbf{M}\mathbf{M}'} = 0$.

3. We construct the trial map, $\mathbf{M}'$, by transferring atom i to site J.

4. We construct the set of allowed steal moves, $T(\mathbf{M}')$, from site $\mathbf{M}'$.

5. We accept the proposed move, $\mathbf{M} \to \mathbf{M}'$ with probability,

$$\mathrm{Acc}(\mathbf{M} \to \mathbf{M}') \equiv \frac{C_{\mathbf{M}}}{\max\{C_{\mathbf{M}}, C_{\mathbf{M}'}\}} \min\left\{1, \mathcal{P}_N(\mathbf{M}')/\mathcal{P}_N(\mathbf{M})\right\}, \tag{32}$$

while remaining at $\mathbf{M}$ with probability $1 - \mathrm{Acc}(\mathbf{M} \to \mathbf{M}')$. Note that, in order to satisfy the detailed balance condition, the modified Metropolis criterion must account for the difference in the coordination, $C_{\mathbf{M}} = |T(\mathbf{M})|$ and $C_{\mathbf{M}'} = |T(\mathbf{M}')|$, of the two maps.

IV. ADDITIONAL COMPUTATIONAL DETAILS

A. AA model

We adopted an α-carbon GNM as an AA model for equilibrium fluctuations of ubiquitin about its folded conformation. We defined the equilibrium structure for ubiquitin by the three-dimensional coordinates for the first 72 residues in the PDB structure 1UBQ, while discarding the coordinates for the 4 disordered residues at the C-terminus.[101] We employed ProDy version 3.0.4[102] to determine the Kirchhoff matrix, $\boldsymbol{\kappa}$, for the AA GNM, while employing the cut-off $r_c = 7.5$ Å to identify residues that are in contact.

B. Simulated annealing

We initially performed simulated annealing to find the range of β and λ that were relevant for sampling the mapping space, $\mathcal{M}_N$, of N-site representations. These simulations scanned a series of $T = \beta^{-1}$ and λ values on a log scale from 1 to 0.0001 and from -1 to -0.0001. Starting with $\lambda = 1$, we simulated 5000 equilibrium MC steps at $T = 1$ before reducing the temperature and performing an additional 5000 equilibrium MC steps. We continued this process with $\lambda = 1$ until we had performed 5000 equilibrium MC steps at the lowest temperature, 0.0001. We then reduced λ and repeated this process, again starting from $T = 1$. In order to fully explore $\mathcal{M}_N$ we repeated this process for both positive and negative T and λ values. We estimated $\overline{\mathcal{Q}}(T, \lambda) = \langle \mathcal{Q}(\mathbf{M}) \rangle_{T,\lambda}$ and $\overline{\sigma}^2(T, \lambda) = \langle \sigma^2(\mathbf{M}) \rangle_{T,\lambda}$ from these simulated annealing simulations and checked that the resulting averages attained plateaus at the limiting values for T and λ, indicating that we had sampled the bounds of $\mathcal{M}_N$.

C. Monte Carlo simulations

We primarily explored the mapping space, $\mathcal{M}_N$, for N-site representations by performing unbiased MC simulations that sampled $\mathcal{P}_N(\mathbf{M})$ at a range of state points, (β, λ). Each simulation in $\mathcal{M}_N$ started from the same "block" map, which is defined by associating a site with each consecutive group of n/N residues along the backbone. We performed each MC simulation for 1.5 $\times 10^6$ steps, while discarding the first 10^3 MC steps as equilibration and sampling every 10^{th} map from the remainder of the simulation.

After completing our initial equilibrium simulations, we visualized two-dimensional scatter plots of the sampled maps with respect to $(\mathcal{Q}, \sigma^2)$ and $(\mathcal{Q}, I)$. In some cases, we found that our canonical simulations failed to sample certain regions of mapping space. In order to specifically sample these regions, we performed additional biased simulations that supplemented $\mathcal{E}(\mathbf{M}) = 1 - \mathcal{Q}(\mathbf{M})$ with the umbrella potential:

$$\mathcal{E}_{\text{bias}}(\mathbf{M}; \mathcal{Q}_k, I_k, \sigma_k^2) = \frac{1}{2} k_{\mathcal{Q}} \left(\mathcal{Q}(\mathbf{M}) - \mathcal{Q}_k\right)^2 + \frac{1}{2} k_I \left(I(\mathbf{M}) - I_k\right)^2 + \frac{1}{2} k_{\sigma^2} \lambda \left(\sigma^2(\mathbf{M}) - \sigma_k^2\right)^2, \quad (33)$$

where $\mathcal{Q}_k$, I_k, σ_k^2, and the corresponding spring constants were chosen to target specific regions of mapping space.

D. Statistics and free energy calculations

Given the maps sampled from these biased and equilibrium simulations, we employed the multistate Bennett Acceptance Ratio (MBAR) method[103] to estimate the statistical weight of each map at a target state point (β, λ). We then estimated statistical properties of mapping space and, in particular, the density of states, Ω_N, from the calculated statistical weights for the infinite temperature, zero pressure limit, i.e., $\beta = \lambda = 0.01$. For all resolutions, we employed a bin spacing of $\delta\mathcal{Q} = 0.005$ and $\delta I = 0.001$ to represent Ω_N with respect to $\mathcal{Q}$ and I, respectively. Because σ^2 takes on discrete values over a range that varies with both protein size and CG resolution, we adopted a bin spacing $\delta\sigma^2$ that corresponds to twice as many bins for σ^2 as for $\mathcal{Q}$. We calculated temperature-dependent free energy surfaces, $\beta F(\mathcal{Q}) = \beta\mathcal{E}(\mathcal{Q}) - \ln \Omega_N(\mathcal{Q})$.

E. Representative maps and steepest descent simulations

Based upon our equilibrium MC simulations for $N = 3$ and 12, we constructed a histogram of the sampled maps as a function of spectral quality, $\mathcal{Q}$, while employing a bin spacing $\delta\mathcal{Q} = 0.005$. We randomly selected 5000 maps from each bin of the histogram with at least 5000 maps. From bins with fewer than 5000 maps, we selected all of the sampled maps. This process identified 185,112 distinct maps for $N = 3$ and 358,000 distinct maps for $N = 12$. We employed these representative maps to analyze the energy landscape of mapping space. In particular, we performed a steepest descent simulation with each of these selected maps. At each step in these simulations, we identified all the neighbors, $\mathbf{M}'$, of a given map, $\mathbf{M}$, and then moved to the neighbor, $\mathbf{M}'_*$, with greatest spectral quality. The quench ended when we reached a map, $\mathbf{M}_m$, with greater spectral quality than any of its neighbors.

F. Metrics

We characterize the physical size of the sites defined by a CG mapping, $\mathbf{M}$, by the radius of gyration, $R_{\rm g}(\mathbf{M})$. We denote $r^*_{i\alpha}$ as the α Cartesian coordinate of atom i in the atomically detailed PDB reference structure. Similarly, we denote $R^*_{I\alpha} = R^*_{I\alpha}(\mathbf{M}) = n_I^{-1}\sum_{i\in V_I} r^*_{i\alpha}$ as the α Cartesian coordinate of site I in the mapped representation of the PDB structure. For each site I, we define a gyration tensor, G_I, with elements:

$$\mathrm{G}_{I;\alpha\gamma} \equiv \mathrm{G}_{I;\alpha\gamma}(\mathbf{M}) \equiv n_I^{-1}\sum_{i\in V_I} \delta r^*_{i\alpha}\delta r^*_{i\gamma}, \tag{34}$$

where $\delta r^*_{i\alpha} = r^*_{i\alpha} - R^*_{I\alpha}$ and $1 \le \alpha,\gamma \le 3$. We define the gyration radius of site I by $R^2_{{\rm g};I}(\mathbf{M}) = \sum_{\alpha=1}^{3} \mathrm{G}_{I;\alpha\alpha}$ The gyration radius of the map is then $R_{\rm g}(\mathbf{M}) = N^{-1}\sum_{I=1}^{N} R_{{\rm g};I}(\mathbf{M})$.

We quantify the sequence similarity between two CG representations based upon the variation of information (VI), which quantifies the overlap between the corresponding atomic partitions.[104] Consider a mapping, $\mathbf{M} \sim [G_1,\ldots,G_N]$ where $G_I = (V_I, E_I)$ is the subgraph associated with site I. We define $P_I(\mathbf{M}) = |V_I|/n$ as the probability of randomly selecting an atom that is associated with site I, where $|V_I|$ is the number of atoms in V_I. We define

$$H_1(\mathbf{M}) = -\sum_{I=1}^{N} P_I(\mathbf{M}) \ln P_I(\mathbf{M}) \tag{35}$$

as the information stored in $P_I(\mathbf{M})$. Now consider a second mapping, $\mathbf{M}' \sim [G'_1,\ldots,G'_N]$, where $G'_{I'} = (V'_{I'}, E'_{I'})$ is the subgraph associated with site I' in $\mathbf{M}'$. We define $|V_I \cap V'_{I'}|$ as

the number of atoms in the set $V_I \cap V'_{I'}$. We then define $P_{II'}(\mathbf{M}, \mathbf{M}') = |V_I \cap V'_{I'}|/n$ as the probability of randomly selecting an atom that is associated with both site I in $\mathbf{M}$ and also site I' in $\mathbf{M}'$. Given the two representations, $\mathbf{M}$ and $\mathbf{M}'$, we define the total information, $H_2(\mathbf{M}, \mathbf{M}')$, and the mutual information, $\mathrm{MI}(\mathbf{M}, \mathbf{M}')$, associated with the corresponding partitions by

$$H_2(\mathbf{M}, \mathbf{M}') = -\sum_{I=1}^{N}\sum_{I'=1}^{N'} P_{II'}(\mathbf{M}, \mathbf{M}') \ln P_{II'}(\mathbf{M}, \mathbf{M}') \tag{36}$$

$$\mathrm{MI}(\mathbf{M}, \mathbf{M}') = -\sum_{I=1}^{N}\sum_{I'=1}^{N'} P_{II'}(\mathbf{M}, \mathbf{M}') \ln \left[\frac{P_{II'}(\mathbf{M}, \mathbf{M}')}{P_I(\mathbf{M}) P_{I'}(\mathbf{M}')}\right]. \tag{37}$$

The VI quantifies the information in $P_{II'}(\mathbf{M}, \mathbf{M}')$ that is not shared between the two mappings:

$$\mathrm{VI}(\mathbf{M}, \mathbf{M}') = H_2(\mathbf{M}, \mathbf{M}') - \mathrm{MI}(\mathbf{M}, \mathbf{M}') = H_1(\mathbf{M}) + H_1(\mathbf{M}') - 2\mathrm{MI}(\mathbf{M}, \mathbf{M}'). \tag{38}$$

Since VI defines a formal distance metric between partitions,[104] we quantify the distance of $\mathbf{M}$ from the ground state representation, $\mathbf{M}_0$, by $d_0(\mathbf{M}) \equiv \mathrm{VI}(\mathbf{M}, \mathbf{M}_0)$.

V. RESULTS

A. The AA Model

In the present work, we adopt an α-carbon Gaussian network model (GNM) as a high resolution model for the equilibrium fluctuations of ubiquitin. This GNM represents each of the $n = 72$ amino acids with its α-carbon and introduces a linear spring between each pair of amino acids that are in contact in the equilibrium folded structure. Figure 2a presents a ribbon structure of the equilibrium ubiquitin structure, while indicating the α carbons with yellow spheres.

The top half of Fig. 2b presents the upper half ($j > i$) of the corresponding Kirchhoff matrix, $\kappa_{ij} = -\theta_{ij}$, where θ_{ij} is the contact matrix for ubiquitin, i.e., $\theta_{ij} = 1$ if amino acids i and j are in contact and 0 otherwise. The blue marks along the diagonal, $j = i + 1$, correspond to consecutive amino acids along the protein backbone. We adopt a GNM cutoff, $r_c = 7.5$ Å, that also introduces a bond between each pair of next-nearest neighbors along the chain, $j = i + 2$. The thickened band along the diagonal for $20 < i < 33$ corresponds to

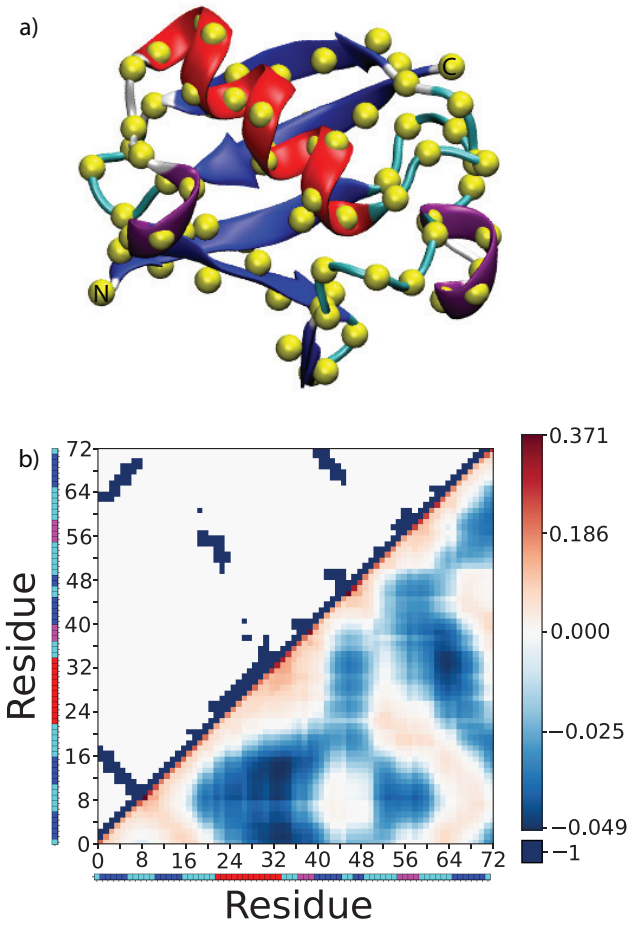

FIG. 2. High-resolution GNM for ubiquitin. Panel a presents a cartoon ribbon structure for the equilibrium folded structure. The amino acids are colored by secondary structure, while the yellow spheres indicate the α carbons. N and C indicate the α carbons for amino acids $i = 1$ and 72, respectively. Panel b presents intensity plots of the Kirchhoff matrix, $\boldsymbol{\kappa}$, (top half) and the dimensionless vibrational covariance matrix, $\beta\Gamma\mathbf{c}_{\rm v} = \boldsymbol{\kappa}^{\rm I}$, (bottom half) for the high-resolution GNM. The color bars along the axes of panel b indicate the secondary structure of each residue in panel a.

non-bonded contacts between successive turns of the α helix, while the band that is farther from the diagonal for $20 < i < 33$ reflects non-bonded contacts between the α helix and β sheet. Conversely, the longer bands that are parallel and anti-parallel to the main diagonal correspond to contacts between parallel and anti-parallel β strands, respectively.

The bottom half of Fig. 2b presents an intensity plot of the vibrational covariance matrix, $\mathbf{c}_{\rm v}$, for the high resolution GNM. The vibrational covariance matrix indicates strong corre-

lation along the backbone and within the α helical region, as well as between contacting β strands. Conversely, the covariance matrix indicates that the motion of the α helix and contacting β strands appear to be anti-correlated.

B. Mapping Space

In this work, we investigate the space, $\mathcal{M}_N$, of N-site CG representations for ubiquitin. Each CG representation corresponds to a mapping, $\mathbf{M}$, that partitions the $n = 72$ α carbons into N disjoint connected groups and associates a CG site with the mass center of each group. We explore $\mathcal{M}_N$ with an ergodic steal move set. Starting from valid representation $\mathbf{M}$, each steal move generates a new representation $\mathbf{M}'$ by moving a single atom between CG sites.

We employ several metrics to characterize the quality of a given map, $\mathbf{M}$. In particular, we focus on the spectral quality, $\mathcal{Q}(\mathbf{M})$, and the information content, $I(\mathbf{M})$, which are both defined in Section II. The spectral quality, $\mathcal{Q}(\mathbf{M})$, quantifies the extent to which $\mathbf{M}$ preserves low-frequency, large-amplitude motions. Conversely, the information content, $I(\mathbf{M})$, quantifies the fraction of non-trivial configurational information that is preserved by $\mathbf{M}$. We also define the site-size variance, $\sigma^2(\mathbf{M}) = \mathrm{var}\{|V_1|, \ldots, |V_N|\}$, where $|V_I|$ is the number of atoms associated with site I. This metric quantifies the heterogeneity of the mass-distribution defined by the mapping.

We define an equilibrium distribution for $\mathcal{M}_N$ according to

$$\mathcal{P}_N(\mathbf{M}; \beta, \lambda) \propto \exp\left[-\beta\left(\mathcal{E}(\mathbf{M}) + \lambda\sigma^2(\mathbf{M})\right)\right], \tag{39}$$

where $\mathcal{E}(\mathbf{M}) = 1 - \mathcal{Q}(\mathbf{M})$ biases sampling towards maps with high spectral quality and β is the conjugate inverse temperature. In order to facilitate sampling diverse representations, we supplemented $\mathcal{E}(\mathbf{M})$ with a term $\lambda\sigma^2(\mathbf{M})$, where λ is analogous to a pressure. For a given N, we define the "ground state" map, $\mathbf{M}_0$, as the N-site map that minimizes $\mathcal{E}$ and, equivalently, maximizes the spectral quality, $\mathcal{Q}$.

The top half of Fig. 3 presents the ground state representations, $\mathbf{M}_0$, for $N = 2$, 3, and 12. In the lowest resolution case, $N = 2$, $\mathbf{M}_0$ assigns the β sheet to one site, while assigning the α helix and loops that are above the β sheet to the second. This representation corresponds to a rather homogeneous mass distribution, $\sigma^2(\mathbf{M}_0) = 4$, with one site representing 38 residues and the second site representing 34 residues. Nevertheless, it would not be sampled

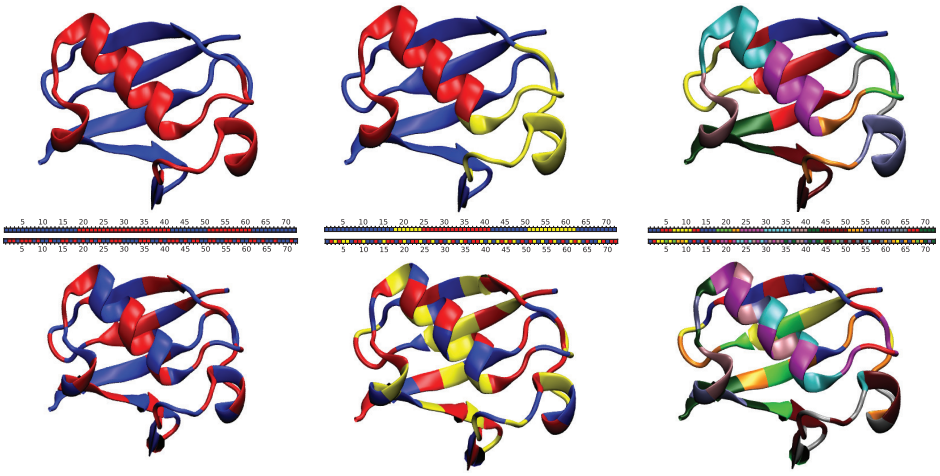

FIG. 3. CG representations of ubiquitin that maximize $\mathcal{Q}$ (top) and I (bottom). The left, center, and right columns correspond to $N = 2$-, 3-, and 12-site representations, respectively. The colors in the ribbon structures and in the linear sequence representation below each structure indicate the CG site that is associated with each amino acid.

in a canonical ensemble that required all sites to represent the same number of residues. Additionally, it is worth noting that the two CG sites do not correspond to consecutive residues in the protein sequence. Consequently, this representation would not be identified by algorithms that sampled representations by shifting the boundaries between contiguous atomic groups. In the case $N = 3$, the ground state representation again maps the β sheet and α helix to their own sites, while mapping the neighboring loop regions to the third site. Interestingly, the $N = 3$ ground state essentially retains the β-sheet site from the $N = 2$ ground state, while splitting the remaining residues approximately even between the second and third sites. This results in a rather inhomogeneous mass distribution, $\sigma^2(\mathbf{M}_0) = 72.67$, in which the blue site, red, and yellow sites correspond to 36, 19, and 17 amino acids, respectively. The ground state map is more complex for $N = 12$. Nevertheless, the different sites again correspond to coherent structural features.

These ground state maps partition atoms such that the majority of the bonds in the underlying AA model are "intra-site" bonds, i.e., bonds between two atoms that have been mapped to the same site. In particular, the ground state maps for $N = 2$ and 3 partition atoms such that 92% and 88%, respectively, of the AA bonds are intra-site bonds. Even the relatively high-resolution $N = 12$-site ground state partitions atoms such that 57% of the AA bonds are intra-site bonds. Conversely, these ground state maps are characterized by

relatively few "inter-site" bonds, i.e., AA bonds between atoms in distinct sites. Thus, the ground state maps maximize the mass-weighted displacements of the CG sites by associating them with rigid atomic groups that are minimally constrained by inter-site bonds.

The bottom half of Fig. 3 presents the representations, $\mathbf{M}_I$, that maximize $I(\mathbf{M})$ for $N =$ 2, 3, and 12. Even in the lowest resolution case, $N = 2$, $\mathbf{M}_I$ divides the α helix and β sheet rather evenly between the two CG sites. In all three cases, the sites in maximally informative mappings do not correspond to coherent structural features. These maps correspond to loosely connected atomic groups with relatively few intra-site bonds and relatively many inter-site bonds. The maximally informative maps for $N = 2$, 3, and 12 partition atoms such that 62%, 75%, and 77% of the AA bonds are inter-site bonds. These inter-site bonds strongly constrain the motion of the CG sites in the mapped ensemble, which results in a relatively narrow and, thus, highly informative mapped distribution.

These findings are consistent with the work of Giulini et al., who observed that minimizing H_{map} – which corresponds to maximizing I – resulted in CG representations that did not match their physical intuition.[41] Interestingly, they also observed that decimation CG representations with maximal H_{map} tended to highlight functionally important residues. However, while Giulini et al. considered decimation maps that associated CG sites with specific amino acids, here we have considered "partition" maps that preserve all of the amino acids and associate CG sites with connected groups of amino acids. The SM demonstrates that the partition maps with maximal information content, I, do not appear to highlight functionally important residues.

Our physical intuition suggests that "good" representations should generally associate CG sites with distinct structural features, e.g., secondary structures, that move coherently. Maps with maximal spectral quality appear quite consistent with this intuition, while maps with maximal information content do not. Consequently, we focus on exploring mapping space with the energy function, $\mathcal{E} = 1 - \mathcal{Q}$, that favors maps of high spectral quality.

C. The Landscape

We next consider the "energy landscape" of N-site mapping space, $\mathcal{M}_N$, that is specified by the energy function $\mathcal{E} = 1 - \mathcal{Q}$ and the steal move set. In this section, we focus on $N =$ 3 and 12 as two representative resolutions for ubiquitin. At each resolution, we performed

equilibrium Monte Carlo simulations according to the methods described in Sections III and IV. We then selected representative maps from these Monte Carlo simulations in order to investigate the minima and connectivity of mapping space.

In order to investigate the minima of the energy landscape, we quenched each representative map according to the energy function $\mathcal{E} = 1 - \mathcal{Q}$.[105] At each step in these steepest descent simulations, we identified all the neighbors of a map and then moved to the neighboring map with lowest energy. The quench ended when the simulation reached a map, $\mathbf{M}_m$, that had lower energy than any of its neighbors.

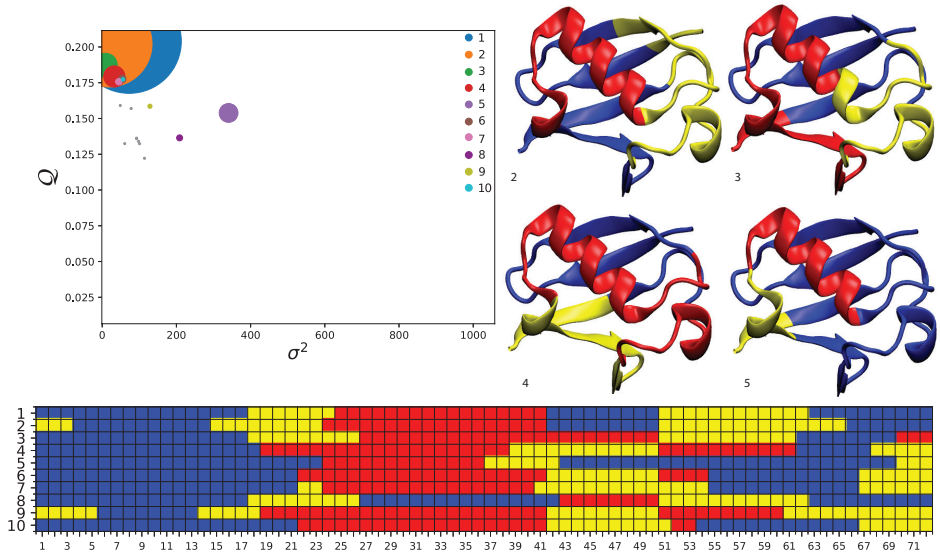

FIG. 4. Local minima in the energy landscape for $N = 3$-site representations of ubiquitin. The top left panel indicates $(\sigma^2(\mathbf{M}_m), \mathcal{Q}(\mathbf{M}_m))$ for the 10 minima, $\mathbf{M}_m$, most frequently obtained from quenching simulations. The size of each circle is proportional to the number of quenches that ended in the corresponding minimum. The top right panel presents the CG representations, $\mathbf{M}_m$, for $m = 2, \ldots, 5$. The bottom panel presents the partitioning associated with each of the top 10 minima, $\mathbf{M}_1, \ldots, \mathbf{M}_{10}$. In these two panels, the colors indicate the residues associated with each site. In particular, the blue, red, and yellow colors indicate the sites associated primarily with the β sheet, α helix, and remaining loops, respectively.

We performed steepest descent simulations for the 185,112 different $N = 3$-site representations that are described in Section IV E. These quenches ended in 18 distinct local minima, $\mathbf{M}_m$, in the energy landscape. Table I and Fig. 4 characterize the 10 minima, $\mathbf{M}_m$, that were most frequently obtained in these quenching simulations. Each circle in Fig. 4

TABLE I. Properties of the 10 local minima, $\mathbf{M}_m$, that are most frequently obtained from quenching 3- and 12-site representations of ubiquitin.

3					12				
Minima	Trajectories	$\mathcal{Q}$	σ^2	d_0	Minima	Trajectories	$\mathcal{Q}$	σ^2	d_0
1	101463	0.204	72.67	0.0	1	47379	0.469	1.83	0.0
2	69420	0.202	18.0	0.628	2	6692	0.469	4.67	0.375
3	5202	0.187	8.67	0.864	3	6581	0.466	3.17	0.488
4	4159	0.179	32.67	1.035	4	6132	0.467	4.67	0.508
5	3303	0.154	340.67	0.960	5	5626	0.469	3.5	0.538
6	529	0.178	54.0	1.015	6	5106	0.467	4.67	0.194
7	380	0.176	44.67	1.19	7	4683	0.466	4.33	0.313
8	333	0.136	208.67	0.828	8	3853	0.467	5.5	0.274
9	162	0.159	128.67	0.834	9	3462	0.467	3.67	0.674
10	117	0.177	56.0	1.072	10	3378	0.465	5.33	0.326

indicates the spectral quality, $\mathcal{Q}(\mathbf{M}_m)$, and site variance, $\sigma^2(\mathbf{M}_m)$, of a particular minimum, $\mathbf{M}_m$, while the size of the circle indicates the number of quenches that ended at $\mathbf{M}_m$. The overwhelming majority, ≈ 98 %, of quenches ended in local minima with very high spectral quality, $\mathcal{Q}(\mathbf{M}_m) > 0.175$. Moreover, the corresponding basins of attraction span the entirety of mapping space. Thus, it appears that $\mathcal{M}_3$ is highly connected with surprisingly few local minima.

The five minima with the largest basins accounted for 99% of the quenches and correspond to similar, physically reasonable representations. As shown in the top right panel of Fig. 4, these representations tend to map the β sheet and α helix to separate sites, while assigning the third site to different loop regions of the protein. In particular, the local minimum, $\mathbf{M}_1$, with the largest basin of attraction accounted for 55% of the quenches and corresponds to the 3-site ground state, $\mathbf{M}_0$, which is shown in Fig. 3. A second, nearly degenerate local minimum, $\mathbf{M}_2$, attracted 38% of the quenches. This second minima corresponded to a very similar representation in which part of the last β strand is mapped with the turn and loop regions. Interestingly, the fifth minimum, $\mathbf{M}_5$, is characterized by a particularly inhomogeneous mass distribution, $\sigma^2(\mathbf{M}_5) = 340.67$

Table I and the supplementary material present corresponding results for quenching 358,000 distinct $N = 12$-site representations. These quenches identified 8595 distinct minima. We analyzed 10 distinct minima that were each obtained from at least 3000 quenches. All of these 10 minima have very high and nearly equivalent spectral quality, $0.46 \leq \mathcal{Q}(\mathbf{M}_m) \leq 0.47$ and very homogeneous mass distributions, $\sigma^2(\mathbf{M}_m) \leq 6$. In particular, the 12-site local minimum with the largest basin of attraction accounted for 13.2% of quenches and again corresponds to the ground state, $\mathbf{M}_0$, with spectral quality $\mathcal{Q}(\mathbf{M}_0) = 0.4689$. More generally, 99.9% of the quenches resulted in representations with high spectral quality, $\mathcal{Q}(\mathbf{M}_m) > 0.44$. Consequently, the mapping space for 12-site representations also appears highly connected. Furthermore, steepest descent optimization appears very likely to determine physically reasonable CG representations.

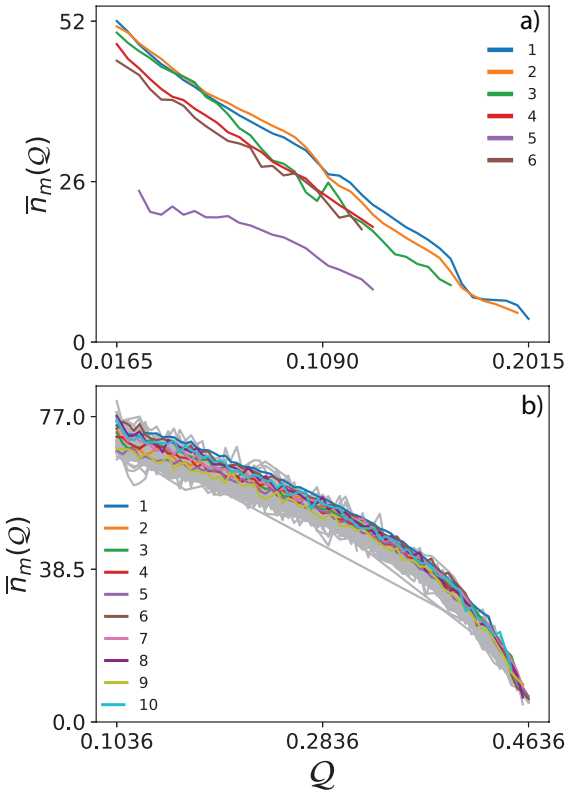

FIG. 5. Average length, $\overline{n}_m(\mathcal{Q})$, of quenching simulations that ended in the minimum, $\mathbf{M}_m$, when starting from initial maps with spectral quality, $\mathcal{Q}$. Each curve, $\overline{n}_m(\mathcal{Q})$, corresponds to a single minimum, $\mathbf{M}_m$, that is indicated by the legend. Panels a (top) and b (bottom) present results for $N = 3$ and 12, respectively.

Figure 5 characterizes the length of the quenching trajectories for 3- and 12-site repre-

sentations of ubiquitin. Each curve in Fig. 5 corresponds to a local minimum, $\mathbf{M}_m$, that was identified by at least 10^3 quenches. For each minimum, we plot the average number, $\overline{n}_m(\mathcal{Q})$, of quenching steps that were necessary to reach the local minimum, $\mathbf{M}_m$, when starting from an initial map with spectral quality $\mathcal{Q}$.

Figure 5 indicates that these steepest descent trajectories are quite short and, moreover, do not dramatically lengthen with increasing resolution. For instance, approximately 50 steps are required to reach a high-quality 3-site map when starting from a very poor 3-site map with minimal spectral quality $\mathcal{Q} \approx 0.016$. Similarly, approximately 70 steps are required to reach a high-quality 12-site map when starting from a low quality map.

The curves, $\overline{n}_m(\mathcal{Q})$, that describe the quenches to different local minima, $\mathbf{M}_m$, are remarkably similar. With the exception of the quenches to $\mathbf{M}_5$, $\overline{n}_m(\mathcal{Q})$ appears to be almost independent of m for $N = 3$-site representations. Moreover, the slopes, $\mathrm{d}\overline{n}_m/\mathrm{d}\mathcal{Q}$, appear nearly independent of both m and also $\mathcal{Q}$. In the case of $N = 12$-site representations, $\overline{n}_m(\mathcal{Q})$ again appears almost independent of m. Interestingly, though, the slope of these curves varies with $\mathcal{Q}$. Specifically, $\overline{n}_m(\mathcal{Q})$ decreases quite slowly with $\mathcal{Q}$ for initial maps with relatively low spectral quality, but decreases more rapidly for initial maps with relatively high spectral quality. Given a map with spectral quality, $\mathcal{Q}$, one expects that $\delta n \equiv |\mathrm{d}\overline{n}_m(\mathcal{Q})/\mathrm{d}\mathcal{Q}| \times \delta\mathcal{Q}$ quenching steps are necessary to increase the spectral quality by $\delta\mathcal{Q}$. Consequently, it requires relatively few steps to increase the spectral quality of relatively poor maps by $\delta\mathcal{Q}$, but requires more steps to improve high quality maps by the same increment. This suggests that, while the energy landscape is generally sloped towards the ground state, this slope decreases as one approaches the ground state.

Figure 6 characterizes the connectivity of mapping space for $N = 3$- and 12-site representations of ubiquitin. For each of the representative maps, $\mathbf{M}$, we identified and analyzed the properties of the neighboring maps, $\mathbf{M}'$, that are separated by a single steal move. As discussed in Sec. III D, a steal move is only allowed if (1) the stolen atom forms an inter-site connection to a new site; and (2) the stolen atom is not an articulation node that would disconnect its original site when stolen. The panels of Fig. 6 present box plots characterizing the distribution of neighbor properties as a function of the spectral quality, $\mathcal{Q} = \mathcal{Q}(\mathbf{M})$, of the map, $\mathbf{M}$. In each panel, the solid curve presents the mean of this distribution as a function of $\mathcal{Q}$. The dashed curves present the first and third quartiles of the distribution, while the dotted curves present the extrema.

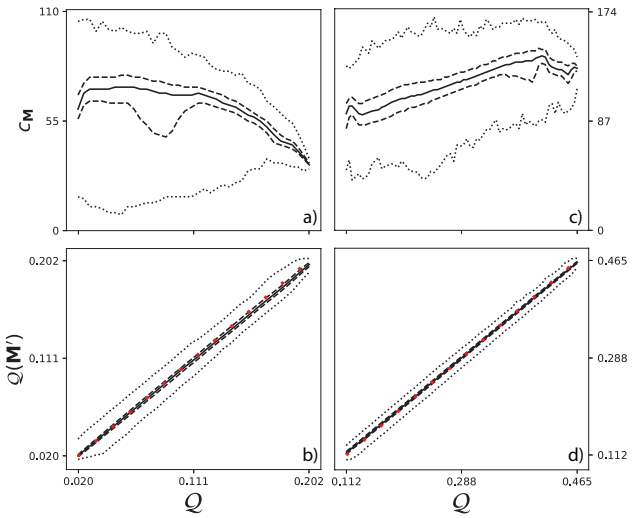

FIG. 6. Neighbor analysis of maps, $\mathbf{M}$, as a function of their spectral quality, $\mathcal{Q}(\mathbf{M}) = \mathcal{Q}$. The top panels characterize the number of neighbors, $C_{\mathbf{M}}$, while the bottom panels characterize the spectral quality of these neighbors. The left and right panels present results for $N = 3$- and 12-site representations of ubiquitin, respectively. Each panel presents a box plot of the corresponding distribution as a function of $\mathcal{Q}$. The solid curves indicate the mean, the dashed curves indicate the first and third quartile, and the dotted lines indicate the extrema of each distribution. The dotted red curve in the bottom panels indicates the line $y = x$.

We define the coordination number, $C_{\mathbf{M}}$, of a map, $\mathbf{M}$, as the number of its neighbors or, equivalently, as the number of possible steal moves starting from $\mathbf{M}$. The mean coordination number of low-resolution 3-site maps is 59 with a standard deviation of approximately 16. Conversely, the mean coordination number for high-resolution 12-site maps is 116 with a standard deviation of approximately 19. The top panels of Fig. 6 characterize the coordination number, $C_{\mathbf{M}}$, of maps as a function of their spectral quality, $\mathcal{Q} = \mathcal{Q}(\mathbf{M})$. The coordination number, $C_{\mathbf{M}}$, appears to vary relatively little with the spectral quality of $\mathbf{M}$. Interestingly, the average coordination number, $\overline{C}(\mathcal{Q})$, tends to slightly decrease with spectral quality among $N = 3$-site representations but tends to slightly increase with spectral quality among $N = 12$-site representations. We hypothesize that this reflects a competition between surface and volume properties of CG sites. The following subsection demonstrates that representations with relatively high spectral quality generally correspond to spatially compact sites. Because compact sites are characterized by relatively little surface area, they form relatively few inter-site connections that are required for stealing atoms. Similarly,

because compact sites tend to be densely connected, they contain relatively few articulation nodes that would disconnect the site if removed. We hypothesize that the surface effect dominates at low resolution: $C_{\mathbf{M}}$ tends to decrease with increasing $\mathcal{Q}$ because higher quality sites include fewer surface atoms that form the necessary inter-site connections for being stolen. Conversely, we hypothesize that the volume effect dominates at high resolution: $C_{\mathbf{M}}$ tends to increase with increasing $\mathcal{Q}$ because higher quality sites have fewer articulation nodes that cannot be stolen.

The bottom panels of Fig. 6 characterize the distribution of spectral quality, $\mathcal{Q}' = \mathcal{Q}(\mathbf{M}')$, for the neighbors, $\mathbf{M}'$, of a map, $\mathbf{M}$, with spectral quality, $\mathcal{Q} = \mathcal{Q}(\mathbf{M})$. Clearly, each map, $\mathbf{M}$, and its neighbors, $\mathbf{M}'$, tend to have extremely similar spectral quality. The mean of these distributions is very close to the line $y = x$, which is indicated by the dotted red line. Moreover, the first and third quartiles of this distribution span a spectral quality range of approximately .005, which is a very small fraction of the range sampled across mapping space.

D. Statistical Properties

We next investigate the statistical properties of mapping space for $N = 3$- and 12-site representations of ubiquitin. Figure 7 presents intensity plots of the corresponding two-dimensional (log) densities of states, $\ln \Omega_N(\sigma^2, \mathcal{Q})$, as a function of the spectral quality, $\mathcal{Q}$, and the site-size variance, σ^2. While our prior investigation of mapping space considered only a "canonical" ensemble with $\sigma^2 = 0$,[59] here we have ergodically sampled the entire space of maps with N connected sites.

In the case of low resolution $N = 3$-site representations, mapping space is dominated by maps with relatively low spectral quality and low site-size variance. In particular, the canonical slice, $\sigma^2 = 0$, spans almost the entire range of spectral qualities sampled by $N = 3$-site representations. Conversely, in the case of higher resolution $N = 12$-site representations, mapping space is dominated by maps with moderate spectral quality and comparatively greater site-size variance. The canonical slice, $\sigma^2 = 0$, for $N = 12$ only includes maps with relatively high spectral quality. Thus, 12-site representations with low spectral quality appear to be characterized by relatively inhomogeneous mass distributions.

Figure 7 demonstrates that the canonical slice, $\sigma^2 = 0$, includes representations with very

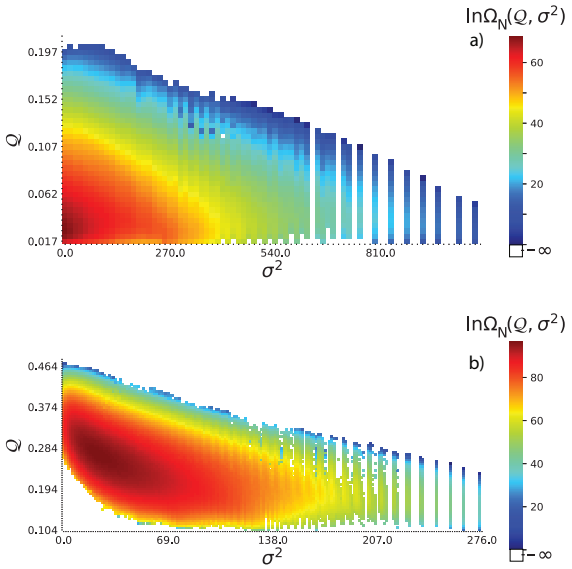

FIG. 7. Intensity plot of the two-dimensional (log) density of states, $\ln \Omega_N(\sigma^2, \mathcal{Q})$, for CG representations of ubiquitin as a function of site-size variance, σ^2, and spectral quality, $\mathcal{Q}$. The top and bottom panels present results for $N = 3$- and 12-site representations, respectively. We have shifted $\ln \Omega_N$ to 0 for the point $(\sigma^2, \mathcal{Q})$ with minimal sampling. White regions correspond to points for which no maps have been sampled. In particular, the spectrum of the discrete variable σ^2 becomes increasingly sparse as σ^2 increases.

high spectral quality at both resolutions. Moreover, Fig. 7 indicates that $\mathcal{Q}$ and σ^2 tend to be negatively correlated, i.e., maps with uniform mass distributions tend to have relatively high spectral quality. However, this correlation is not as strong as might be expected. There exist maps at every resolution with rather inhomogeneous mass distributions and also relatively high spectral quality. In particular, the $N = 3$ ground state corresponded to a rather heterogeneous mass distribution with $\sigma^2 = 72.67$, while the local minima in $\mathcal{M}_3$ were characterized by a rather wide range of σ^2. Furthermore, as just noted, the overwhelming majority of $N = 3$ representations are characterized by homogeneous mass distributions and low spectral quality.

Figure 8 investigates the correlation between the spectral quality, $\mathcal{Q}$, and three other metrics that are defined in Section IV. The left, center, and right columns present intensity plots of two-dimensional (log) densities of states, $\ln \Omega_N$, for $\mathcal{Q}$ with (1) the information content, $I(\mathbf{M})$; (2) the average radius of gyration for the corresponding CG sites, $R_g(\mathbf{M})$; and (3) the distance, $d_0(\mathbf{M}) = ||\mathbf{M} - \mathbf{M}_0||$, of $\mathbf{M}$ from the ground state, $\mathbf{M}_0$. In particular,

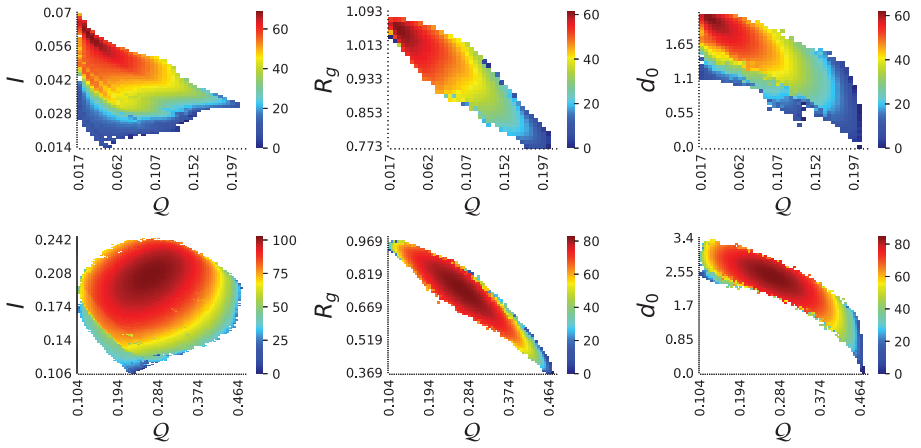

FIG. 8. Intensity plots of two-dimensional (log) densities of states, $\ln \Omega_N$, for the spectral quality, $\mathcal{Q}$, with three other metrics: (1) the information content, I (left); (2) the CG site size, $R_{\rm g}$ (center); and (3) the distance, d_0, from the ground state, $\mathbf{M}_0$ (right). The top and bottom panels correspond to $N = 3$- and 12-site representations of ubiquitin.

the distance, $d_0(\mathbf{M})$, quantifies the overlap of the partitions (i.e., the atomic groups) defined by $\mathbf{M}$ and $\mathbf{M}_0$. For calibration, $N = 3$- and 12-site maps, $\mathbf{M}$, are characterized by an average distance $\overline{d} = \langle d(\mathbf{M}, \mathbf{M}') \rangle = 0.12$ and 0.083, respectively, from their neighbors, $\mathbf{M}'$.

While the information content, I, and spectral quality, $\mathcal{Q}$, appear quite strongly anti-correlated among low-resolution representations, they appear weakly correlated among high-resolution representations. Conversely, $R_{\rm g}$ is strongly anti-correlated with $\mathcal{Q}$ at all resolutions. Thus, maps with high spectral quality are characterized by spatially compact sites. Finally, the right column demonstrates that the spectral quality is also anti-correlated with the distance, d_0, from the ground state. However, the set of maps that are a fixed distance from the ground state is characterized by a very wide range of spectral qualities. Similarly, the set of maps with high spectral quality are characterized by a rather wide range of distances from the ground state, i.e., a very diverse set of partitions. We hypothesize that this diversity of high-quality representations may be the source of the variation in the slope, $\mathrm{d}\overline{n}_m(\mathcal{Q})/\mathrm{d}\mathcal{Q}$, that was observed in Fig. 5. Some high-quality representations may correspond to partitions that are very different from the ground state and, consequently, may require many quenching steps to reach a local minimum in the energy landscape.

E. Resolution Variation

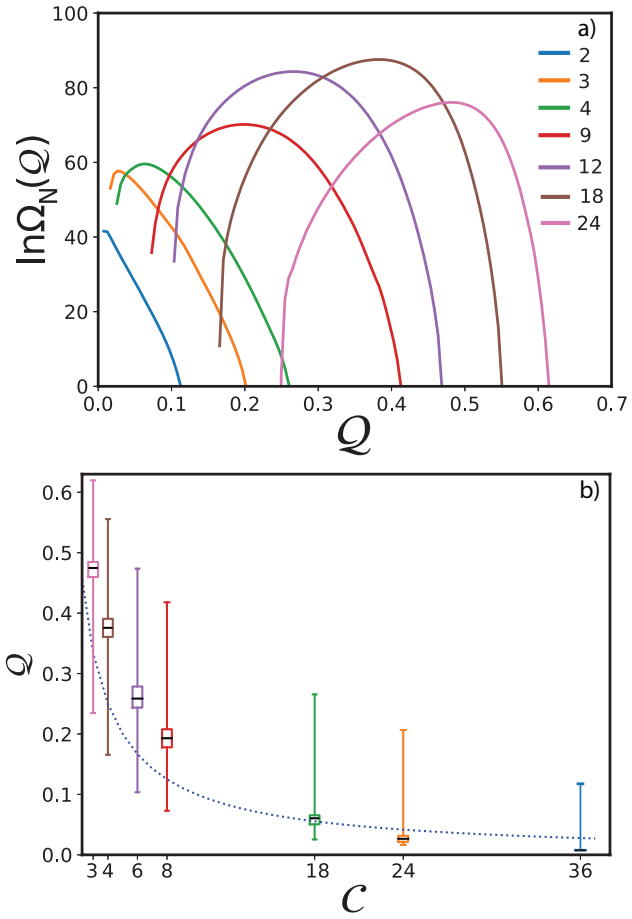

FIG. 9. Distribution of spectral quality, $\mathcal{Q}$, for CG representations of ubiquitin. The top panel presents the one-dimensional (log) density of states, $\ln\Omega_N(\mathcal{Q})$ for all simulated resolutions. The bottom panel presents box plots characterizing the distribution of $\mathcal{Q}$ for each N as a function of $\mathcal{C} = n/N$. In these box plots the horizontal black line indicates the mean, the box indicates the first and third quartiles of the distribution, and the whiskers indicate the extrema. The dotted blue curve in the bottom panel plots the naïve expectation $1/\mathcal{C}$. The legend indicates the number, N, of CG sites for each resolution.

Having examined mapping space, $\mathcal{M}_N$, for two representative resolutions, $N = 3$ and 12, we now systematically examine the impact of resolution, N, upon the properties of $\mathcal{M}_N$. In particular, Fig. 9 investigates the impact of resolution upon the spectral quality, $\mathcal{Q}$. Figure 9a plots the logarithm of the one-dimensional density of states, $\ln\Omega_N(\mathcal{Q})$, with respect to the spectral quality, $\mathcal{Q}$, for resolutions ranging from $N = 2$ (blue) to $N = 24$

(pink). Figure 9b presents a box plot indicating the mean, quartiles, and extrema of the corresponding distributions as a function of coarsening, $\mathcal{C} = n/N$, which corresponds to the average number of amino acids associated with each site. The dotted blue curve in Fig. 9b presents a naïve expectation for the typical spectral quality of N-site representations, i.e., $1/\mathcal{C}$,

The density of states, $\ln \Omega_N(\mathcal{Q})$, features a single prominent maximum at each resolution, N. Consequently, the mapping space, $\mathcal{M}_N$, for each resolution is characterized by a very large number of maps with a typical spectral quality, $\mathcal{Q}_N^*$. This typical spectral quality systematically decreases with resolution. For relatively high resolutions, $\mathcal{C} \leq 8$, the typical spectral quality, $\mathcal{Q}_N^*$, is significantly greater than the naïve expectation $1/\mathcal{C}$. For instance, when $N = 18$ and $\mathcal{C} = 4$, $\mathcal{Q}_N^* \approx 0.4$, which is significantly larger than $1/\mathcal{C} = 0.25$. However, at very low resolutions, $\mathcal{C} > 18$, the typical spectral quality is less than naïvely expected. This reflects an interesting "tilt" in Fig. 9a, as the maximum of $\ln \Omega_N(\mathcal{Q})$ shifts from relatively high to relatively low spectral quality with decreasing resolution.

The one-dimensional densities of states, $\ln \Omega_N(\mathcal{Q})$, are remarkably broad. There are very rare maps at each resolution with much greater spectral quality than $\mathcal{Q}_N^*$. In particular, the ground state map always has significantly greater spectral quality than naïvely expected. Moreover, because the densities of states overlap significantly, higher resolution maps do not necessarily have higher spectral quality. For instance, there exist rare $N = 4$-site representations with higher spectral quality than either typical $N = 9$-site representations or poor $N = 24$-site representations. These observations are generally consistent with our prior study in the canonical ensemble.[59]

It is worth noting that the one-dimensional densities of states in Fig. 9 appear to systematically broaden with increasing resolution. In contrast, our prior canonical study indicated that the one-dimensional densities of states, $\ln \Omega_N(\mathcal{Q})$, became more narrow with increasing resolution. This distinction reflects the difference in the two-dimensional densities of states, $\ln \Omega_N(\sigma^2, \mathcal{Q})$, for $N = 3$ and $N = 12$ in Fig. 7. The canonical slice, $\sigma^2 = 0$, essentially spans the entire spectral quality range for low resolution representations. However, the canonical slice omits a large range of spectral qualities for higher resolution representations.

Figure 10 similarly analyzes the one-dimensional densities of states, $\ln \Omega_N(I)$, quantifying the number of N-site maps with a given information content, I. (The SM presents a corresponding density of states for the mapping entropy, H_{map}.) Figure 10 demonstrates

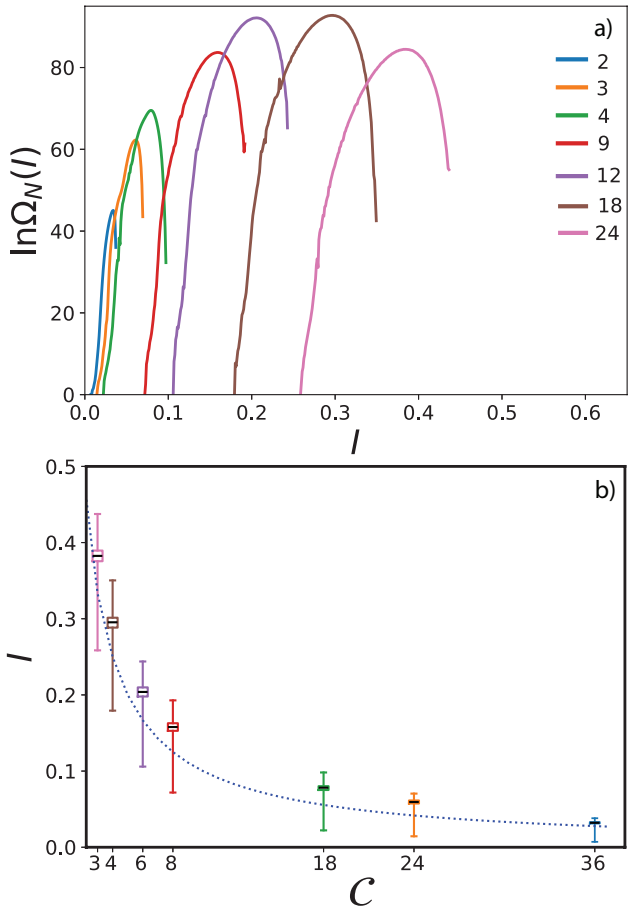

FIG. 10. Distribution of information content, I, for CG representations of ubiquitin. The top panel presents the one-dimensional (log) density of states, $\ln \Omega_N(I)$ for all simulated resolutions. The bottom panel presents box plots characterizing the distribution of I for each N as a function of $\mathcal{C} = n/N$. In these box plots the horizontal black line indicates the mean, the box indicates the first and third quartiles of the distribution, and the whiskers indicate the extrema. The dotted blue curve in the bottom panel plots the naïve expectation $1/\mathcal{C}$. The legend indicates the number, N, of CG sites for each resolution.

that $\mathcal{M}_N$ is also characterized by a large number of maps with typical information content. This typical information content, I_N^*, decreases with coarsening but is always larger than the naïve expectation of $1/\mathcal{C}$. Interestingly, $\ln \Omega_N(I)$ is considerably more narrow than $\ln \Omega_N(\mathcal{Q})$. Thus, in comparison to the spectral quality, the information content appears less sensitive to the details of the particular mapping.

Figure 11 investigates the impact of resolution upon the correlation between spectral quality, $\mathcal{Q}$, and information content, I. Because the densities of states overlap, we plot

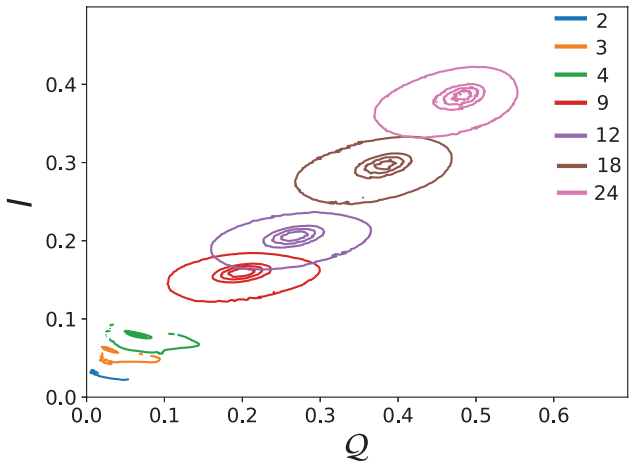

FIG. 11. Contour plots of the two-dimensional (log) density of states, $\Omega_N(\mathcal{Q}, I)$, for all simulated resolutions of ubiquitin. The plots are colored according to the number, N, of CG sites. For each N, the inner three contours indicate where Ω_N decreases to 75%, 50%, and 25% of its maximum value, while the outer-most contour corresponds to the boundaries of sampled maps.

contours of $\ln \Omega_N(\mathcal{Q}, I)$ for each resolution. As just noted, the mapping space, $\mathcal{M}_N$, for each resolution is characterized by a very wide range of spectral qualities, but a comparatively narrow range of information content.

For low resolution representations, $N \leq 4$, $\mathcal{Q}$ and I appear quite strongly anti-correlated, as already noted in Fig. 8 and in our prior study. Interestingly, typical maps for these low resolutions (i.e., maps near the peak of $\ln \Omega_N(\mathcal{Q}, I)$) are characterized by $I > \mathcal{Q}$. Conversely, maps with relatively high spectral quality are characterized by minimal information content. Consequently, it is not possible to simultaneously optimize both $\mathcal{Q}$ and I for low-resolution representations.

The situation is rather different for higher resolution representations with $N \geq 9$. At these higher resolutions, typical maps are characterized by $\mathcal{Q} > I$. Moreover, $\mathcal{Q}$ and I demonstrate a weak positive correlation among higher resolution maps. This somewhat differs from our prior study,[59] which did not notice a significant correlation between $\mathcal{Q}$ and I in the canonical ensemble for higher resolutions. Consequently, it may be possible to simultaneously optimize both $\mathcal{Q}$ and I for high-resolution representations. More importantly, though, we again observe a qualitative distinction between mapping space for low- and high-resolutions.

Figure 12 characterizes the thermodynamics associated with sampling mapping space at

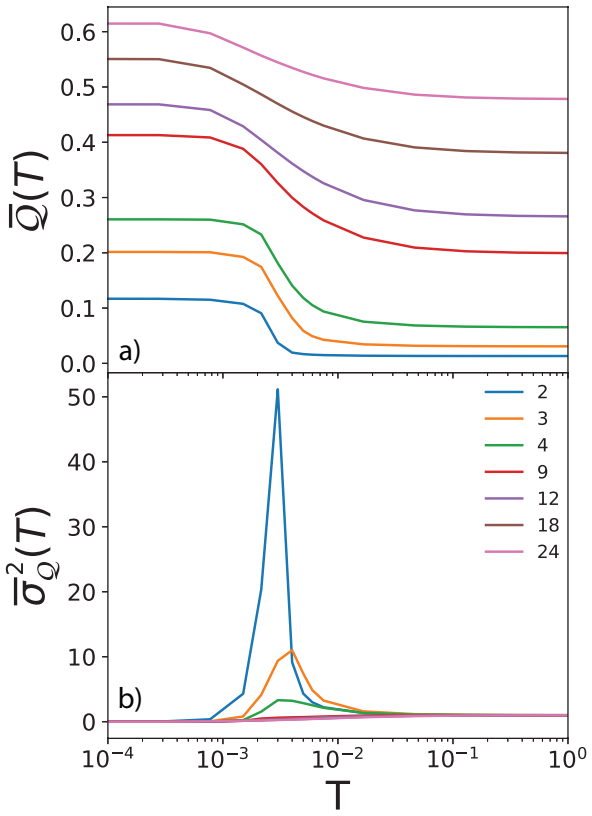

FIG. 12. Thermodynamics of sampling mapping space. The top panel presents the mean spectral quality, $\overline{\mathcal{Q}}(T) = \langle \mathcal{Q} \rangle_T$, as a function of temperature, $T = \beta^{-1}$, for $\lambda = 0.01 \approx 0$. The bottom panel presents the scaled variance, $\overline{\sigma}^2_{\mathcal{Q}}(T) = \sigma^2_{\mathcal{Q}}(T)/\sigma^2_{\mathcal{Q}}(T_\infty)$, where $\sigma^2_{\mathcal{Q}}(T) = \left\langle \left(\mathcal{Q} - \overline{\mathcal{Q}}(T)\right)^2 \right\rangle_T$ and $T_\infty = 100$.

each resolution. In particular, the top panel presents the average spectral quality, $\overline{\mathcal{Q}}(T) = \langle \mathcal{Q} \rangle_T$, as a function of the sampling temperature, $T = \beta^{-1}$, that is conjugate to $\mathcal{Q}$. Figure 9 indicates that the maximum in the one-dimensional density of states, $\Omega_N(\mathcal{Q})$, occurs at moderate spectral quality for high resolution maps, but occurs at very low spectral quality for low resolution maps. Consequently, the entropic $T \to \infty$ limit in Fig. 12 corresponds to moderate quality maps at high resolutions, but very low quality maps at low resolutions.

By definition, the spectral quality systematically increases with decreasing temperature. In the case of high-resolution representations, $N \geq 9$, the spectral quality gradually increases across a rather wide temperature range. In contrast, the spectral quality transitions from low

to high values over a relatively narrow temperature range for low-resolution representations, $N \leq 4$.

The bottom panel of Fig. 12 presents the scaled variance in the spectral quality, $\overline{\sigma}^2_{\mathcal{Q}}(T) = \sigma^2_{\mathcal{Q}}(T)/\sigma^2_{\mathcal{Q}}(T_\infty)$, where we define $T_\infty = 100$ as our infinite temperature limit. In the case of high-resolution representations, $\overline{\sigma}^2_{\mathcal{Q}}(T)$ monotonically decreases as T decreases. However, in the case of low-resolution representations, $\overline{\sigma}^2_{\mathcal{Q}}(T)$ peaks in the transition region, as would be expected for a physical phase transition in a mechanical model. This peak is first visible for $N = 4$ and grows with decreasing resolution, which suggests that $N = 4$ is near to a "critical resolution."

The Supplementary Material investigates the impact of this transition upon the physical size, R_{g}, of CG maps and their distance, d_0, from the ground state. These calculations demonstrate that as the temperature decreases through this transition, low-resolution maps become both more compact and also closer to the ground state. Moreover, the fluctuations in these metrics increase in the transition temperature range for low-resolution representations with $N \leq 4$ but not for high-resolution representations with $N \geq 9$. This transition and the suggestion of a critical resolution are both highly consistent with our prior study in the more restricted canonical ensemble.[59]

Finally, Fig. 13 analyzes the corresponding free energy surface for the two representative resolutions, $N = 3$ and 12, that are below and above the suggested critical resolution, respectively. Specifically, we present the dimensionless free energy, $\beta F(\mathcal{Q})$, as a function of $\mathcal{Q}$ for a range of temperatures, $T = \beta^{-1}$. In the high-resolution case, $N = 12$, the minimum of the free energy surface simply shifts to lower $\mathcal{Q}$ with decreasing temperature. Moreover, the high temperature limit, $T \to \infty$, corresponds to characteristic maps with modest spectral quality. In the low-resolution case, $N = 3$, the free energy surface appears more reminiscent of a physical phase transition. At high temperatures, the minimum of the free energy surface appears to correspond to a basin of maps with low spectral quality. Conversely at low temperatures, the minimum of the free energy surface appears to correspond to a basin of maps with high spectral quality and spatially compact sites. Near the transition temperature, the minimum quickly shifts from one basin to the other, as would be expected for a physical phase transition. However, in contrast to our prior study in the canonical ensemble, we do not find a barrier in the free energy surface at any temperature.

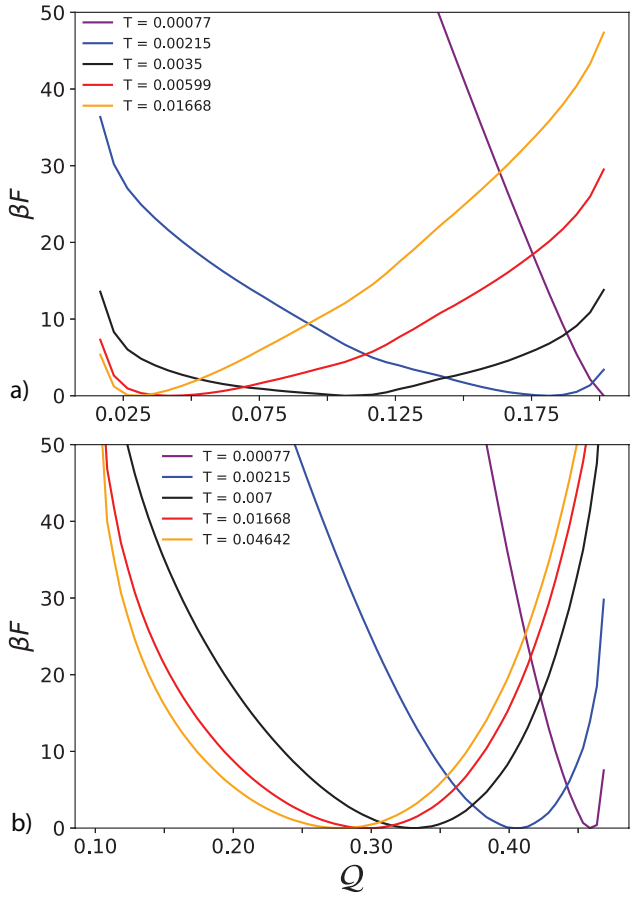

FIG. 13. Dimensionless free energy, βF, as a function of $\mathcal{Q}$ for a range of temperature, $T = \beta^{-1}$. The left and right panels present results for $N = 3$- and 12-site representations of ubiquitin, respectively.

VI. DISCUSSION

In this work we have systematically investigated and characterized the space of CG representations for ubiquitin, which is a typical globular protein. Specifically, we considered representations that partition the protein α carbons into connected groups and associate a CG site with the mass center of each group. We sampled mapping space by employing a Markov chain that employs steal moves to generate new maps by moving a single atom between sites. Importantly, we proved that this steal move set is ergodic. This allowed us to rigorously sample and statistically characterize the space, $\mathcal{M}_N$, of N-site CG representations. While our prior study investigated a canonical ensemble in which each site corresponded to an equal number of amino acids,[59] our present study investigated a much

larger semi-grand ensemble in which each site can represent an arbitrary number of amino acids.

By adopting the GNM as a simple high resolution model for equilibrium fluctuations about the protein folded conformation, we are able to exactly assess the intrinsic quality of each CG representation. We primarily focused on two metrics: the information content and the spectral quality. The information content, $I(\mathbf{M})$, quantifies the fraction of (non-trivial) configurational information that is preserved in the mapped ensemble. The information content is closely related to the mapping entropy, H_{map}, which has been previously investigated by Potestio and coworkers.[38,41–44] Conversely, the spectral quality, $\mathcal{Q}(\mathbf{M})$, quantifies the fraction of the mass-weighted covariance that is preserved in the mapped ensemble. Since the spectral quality favors maps that preserve large-amplitude, low frequency motions, it appears closely related to the ED-CG metric of Voth and coworkers,[50–54] the VAMP metric of Clementi and coworkers,[57] and many other metrics that have been developed to identify quasi-rigid domains in proteins.[46,85,97–100] However, it is important to properly account for the number of amino acids associated with each site when computing $\mathcal{Q}(\mathbf{M})$ for low resolution representations. In particular, Section II showed that $\mathcal{Q}$ should not be calculated by simply mass-weighting the mapped covariance matrix.

We observe striking differences between the maps that maximize I and $\mathcal{Q}$. Maps that maximize I associate CG sites with loosely connected atomic groups that are constrained by many inter-site bonds. These representations generate relatively narrow mapped distributions that preserve configurational information. Conversely, maps that maximize $\mathcal{Q}$ associate CG sites with compact, densely connected atomic groups that correspond to coherent structural features as observed in quasi-rigid domains.[46,97,98,100] Because these sites are constrained by relatively few inter-site bonds, these representations generate broad mapped distributions that preserve large-amplitude motions. Consequently, maps that maximize $\mathcal{Q}$ agree well with our physical intuition, while maps that maximize I do not. Interestingly, if one employs the distance metric, d_0, to quantify the similarity between atomic partitions, then representations with high spectral quality correspond to a rather diverse set of partitions. Moreover, while the spectral quality and the variance in the site-size distribution, σ^2, are anti-correlated, this anti-correlation is weaker than we had expected. In particular, the overwhelming majority of low-resolution representations are characterized by both low spectral quality and uniform mass distributions. Furthermore, we find representations

at every resolution that are characterized by both high spectral quality and also a rather inhomogeneous mass distribution.

In order to gain insight into $\mathcal{M}_N$, we investigated the landscape that is defined by the steal move set and the energy function, $\mathcal{E}(\mathbf{M}) = 1-\mathcal{Q}(\mathbf{M})$, for $N = 3$- and 12-site representations. Our analysis of neighboring maps demonstrated that the coordination number of typical $N = 3$- and 12- site maps are in the range 59 ± 32 and 116 ± 38, respectively. This coordination number, $C_{\mathbf{M}}$, depends relatively weakly upon the spectral quality of $\mathbf{M}$. Moreover, we find that neighboring representations have very similar spectral quality.

Our quenching studies revealed that the energy landscape for $\mathcal{M}_N$ is highly connected and characterized by relatively few local minima. The over-whelming majority of quenches ended in representations with very high spectral quality after relatively few steps. In particular, the ground state representations with maximal spectral quality had the largest basins of attraction for both $N = 3$ and 12. Thus, steepest descent optimization of $\mathcal{Q}$ is very likely to determine a high quality CG representation. Moreover, our quenching studies also suggest that this energy landscape generally slopes towards the ground state. This gradient appears independent of $\mathcal{Q}$ for $N = 3$. However, for $N = 12$, the bottom of the energy landscape appears relatively flat.

We employed Monte Carlo methods to statistically characterize the space of CG representations. At each resolution, $\mathcal{M}_N$ is dominated by a large number of characteristic maps with typical information content, I_N^*. This typical information content, I_N^*, decreases with coarsening, $\mathcal{C} = n/N$, but is always greater than the naïve expectation of $1/\mathcal{C}$. Similarly, $\mathcal{M}_N$ is also dominated by a large number of characteristic maps with a typical spectral quality, $\mathcal{Q}_N^*$, that also decreases with coarsening. Interestingly, $\mathcal{Q}_N^*$ is significantly greater than the naïve expectation, $1/\mathcal{C}$, for modest degrees of coarsening, $\mathcal{C} \leq 8$, but becomes less than $1/\mathcal{C}$ at lower resolutions, $\mathcal{C} \geq 18$. Moreover, the densities of states, $\Omega_N(\mathcal{Q})$, for the spectral quality are extremely broad. Thus, in comparison to I, $\mathcal{Q}$ appears considerably more sensitive to the details of the CG mapping. In particular, there exist very rare representations at each resolution with spectral quality that is much greater than typical. Consequently, the spectral quality of representations does not necessarily increase with resolution. For instance, the $N = 4$-site ground state representation has similar spectral quality to typical $N = 9$-site representations.

The two-dimensional density of states, $\Omega_N(\mathcal{Q}, I)$, reveals an interesting distinction be-

tween low- and high-resolution representations. In particular, for very low resolution representations, $N \leq 4$, I and $\mathcal{Q}$ appear strongly anti-correlated. Consequently, it appears impossible to simultaneously optimize I and $\mathcal{Q}$ for low-resolution representations. Conversely, for relatively high resolutions, $N \geq 9$, I and $\mathcal{Q}$ are characterized by a slightly positive correlation. We hypothesize that this reversal reflects the spectral features of the CG Kirchhoff matrix, $\mathbf{K} = (\mathbb{Q}_N \mathbf{M} \boldsymbol{\kappa}^{\mathrm{I}} \mathbf{M}^{\dagger} \mathbb{Q}_N)^{\mathrm{I}}$. The spectral quality, $\mathcal{Q}$, emphasizes the smallest eigenvalues of $\mathbf{K}$ that correspond to large amplitude, low frequency motions. (More precisely, $\mathcal{Q}$ emphasizes the lowest eigenvalues of the mass-weighted CG Kirchhoff matrix, $\overline{\mathbf{K}} = \mathbf{G}_N^{-1} \mathbf{K} \mathbf{G}_N^{-1}$.) In contrast, the information content, I, emphasizes the largest eigenvalues of $\mathbf{K}$ that correspond to localized, high frequency motions. Following the intuition of the essential dynamics formalism,[55] we expect that the spectrum of $\overline{\mathbf{K}}$ is dominated by relatively few low frequency modes. CG representations with high spectral fitness will tend to align CG sites to move along these few low frequency modes. We expect this objective strongly constrains low-resolution representations, $N \leq 4$. However, for higher resolution representations, $N \geq 9$, it appears possible to define CG sites that both lie along the few low frequency modes and also preserve higher frequency modes that contribute significantly to I.

The thermodynamics associated with sampling mapping space also suggests a qualitative distinction between high- and low-resolutions. At higher resolutions, $N \geq 9$, the average spectral quality, $\overline{\mathcal{Q}}(T)$, gradually increases as we decrease the temperature, T, of our Monte Carlo simulations. However, for sufficiently low resolutions, $N \leq 4$, the average spectral quality rather sharply transitions from relatively low values to relatively high values over rather narrow temperature range. Moreover, in this narrow temperature range, we observe large fluctuations in spectral quality that systematically grow with decreasing resolution. As the temperature decreases through this transition range, the minima of the corresponding free energy surface rapidly transitions from a basin with relatively low spectral quality to a basin that is close to the ground state. CG representations in this low temperature basin are characterized by more compact sites and by partitions that are more similar to the ground state partitioning. While we do not observe a free energy barrier associated with this transition, these features are otherwise quite reminiscent of a phase transition in a physical model for a finite-size system.[106–108] In this analogy, $N = 4$ appears close to a critical resolution that signals the onset of a phase transition that qualitatively distinguishes low and high quality representations. Intriguingly, the recent work on decimation mappings

by Potestio and coworkers also suggested the existence of a phase transition in mapping space based upon the analogy with the condensation transition in lattice-gas models.[42]

It is interesting that the mapped distribution and, thus, the renormalized potential for the GNM have the same form as the underlying AA distribution and potential, respectively.[58] This is in stark contrast to more general, nonlinear potentials, for which renormalization introduces higher-order, many-body interactions between the remaining CG degrees of freedom.[108] This suggests that it may be interesting for future studies to consider hierarchies of renormalized GNM's. Such a hierarchy of GNM's may possibly follow scaling laws, as recently reported for hierarchical CG models of molecular liquids.[109]

In closing, it is important to emphasize that our study has focused on CG representations for a very simple high resolution model. In particular, the high resolution GNM describes protein fluctuations by a simple harmonic potential with a single global minimum. This high resolution model cannot explicitly describe conformational transitions or any other nontrivial thermodynamic behavior.[56] However, the success of network models to study protein folding and conformational transitions,[65–75] suggests that the present approach may possibly be useful in determining CG representations for modeling these very complex and highly nonlinear processes, although this would certainly require a more sophisticated treatment of CG dynamics.[110–117] Consequently, future studies should certainly investigate the extent to which insights for the GNM generalize to more complex systems.

Nevertheless, we are optimistic that the GNM provides a qualitatively reasonable description of more complex nonlinear models that are characterized by equilibrium fluctuations about a single free energy minimum. Moreover, we speculate that the spectral quality and information content may prove rather robust metrics for assessing CG representations of more complex systems that undergo physical phase transitions. Furthermore, we have developed an ergodic algorithm for sampling the space of CG representations as a function of relevant order parameters. This algorithm may itself prove useful for optimizing or further exploring low-resolution representations. Thus, we hope that the present study will prove useful for developing CG models of soft materials. Much more generally, we hope that this study may provide useful insight for understanding and optimizing low-resolution representations of complex systems.

SUPPLEMENTARY MATERIAL

See the supplementary material for analysis of mapping space for additional resolutions.

ACKNOWLEDGMENTS

KMK acknowledges the financial support of a Marie Skłodowska-Curie science achievement graduate scholarship in Chemistry from Penn State. KMK and WGN acknowledge financial support from the National Science Foundation (Grant Nos. CHE-1856337 and CHE-2154433). MSS acknowledges financial support from the National Science Foundation through Award No. CHEM-1800344. Portions of this research were conducted with Advanced CyberInfrastructure computational resources provided by The Institute for Computational and Data Sciences at The Pennsylvania State University (http://icds.psu.edu). Additionally, parts of this research used the Expanse resource at the San Diego Supercomputer Center though allocation TG-CHE170062 from the Extreme Science and Engineering Discovery Environment (XSEDE),[118] which was supported by National Science Foundation grant number TG-CHE170062. This work also used allocation CHE170062 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Figures 1-4 employed VMD.[119] VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of 2510 Illinois at Urbana-Champaign.

AUTHOR DECLARATIONS

Conflict of interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

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

Appendix: Proof of ergodicity

In this appendix we provide additional background information regarding the graph treatment of CG representations. We also prove several results that imply the ergodicity of the steal move set. Our proof relies upon several basic definitions and properties of simple graphs that are found in Gross and Yellen.[120]

1. Basic graph definitions

A graph $G = (V, E)$ is defined by a set of vertices, $V = \{v\}$, and a set of edges, $E = \{e_{uv}\}$, that connect the vertices. We consider non-directed graphs such that e_{uv} and e_{vu} refer to the same edge. The degree of a vertex, v, is the number of edges that connect to v. Given a graph G, a walk is defined as a sequence of vertices $v_1, v_2, \ldots, v_n \in V$ such that there is an edge $e_{v_i,v_{i+1}} \in E$ connecting each successive pair of vertices, v_i and v_{i+1}, in the sequence. If there is a walk from u to v for every pair of vertices $u, v \in V$, then the graph G is connected.

Given $G = (V, E)$, a subgraph, $H = (V_H, E_H)$, is a graph defined by vertices and edges in G, i.e., $V_H \subset V$ and $E_H \subset E$. Given a vertex subset $U \subset V$, the subgraph induced on U, $G(U)$, is the subgraph defined by the vertex set U and the edge set $E_{G(U)}$ that includes all edges between the vertices of U, i.e., $E_{G(U)} = \{e_{uv} \in E | u, v \in U\}$. Given a vertex, $v \in V$, the deletion subgraph $G - v$ is defined by the vertex set $V_{G-v} = V - \{v\}$ and the edge-set E_{G-v} that is obtained by removing from E every edge that connects to v. Thus, $G - v$ is the subgraph of G that is induced by the vertex subset $V - \{v\}$.

Given a connected graph, $G = (V, E)$, a vertex $v \in V$ is a cut-vertex (articulation node) if the deletion graph $G - v$ is no longer connected. A graph G is k-connected if at least k nodes must be deleted from G in order to either disconnect the graph or reduce it to a single vertex.

Property 1 Any connected graph with more than one vertex contains at least two vertices that are not cut-vertices.

2. Atomic and CG graphs

The GNM defines a simple graph, $G = (V, E)$, for a protein with n amino acids. The vertex set $V = \{1, \ldots, n\}$ associates a labelled vertex with each α carbon in the protein. For

simplicity, we employ the integer, i, to identify both the atom and corresponding vertex. In particular, G includes two terminal vertices: the N-terminal α carbon is $i_{\rm N} = 1 \in V$, while the C-terminal α carbon is $i_{\rm C} = n \in V$. The edge set $E = \{e_{ij} | i, j \in V, \theta_{ij} = 1\}$ contains an edge between each pair of vertices that are connected by a linear spring in the GNM potential. We shall require that a **valid** protein graph is singly and doubly connected along the backbone, such that $e_{ij} \in E$ whenever $i, j \in V$ and $|i - j| \leq 2$. The Kirchhoff matrix, $\kappa_{ij} = n_i \delta_{ij} - \theta_{ij}$, is the Laplacian matrix for G, where n_i is the degree of vertex i and θ_{ij} is the adjacency matrix.

We assume that the CG mapping, $\mathbf{M} = (c_{Ii})$, partitions the atoms into disjoint sites and defines the coordinates of each site by the mass center of the associated atomic group. Thus, the mapping, $\mathbf{M}$, determines a vertex set $V_I = \{i \in V | c_{Ii} > 0\}$ for each site $I = 1, \ldots, N$ such that $\cup_I V_I = V$ and $V_I \cap V_J = \varnothing$ whenever $I \neq J$. The mapping also determines an edge set $E_I = \{e_{ij} \in E | i, j \in V_I\}$ for each site I that corresponds to the linear springs between atoms associated with the site. For each site I, we define a site graph, $G_I = (V_I, E_I)$, that is the subgraph of G induced on V_I. Therefore, the mapping, $\mathbf{M}$, corresponds to a partitioning of G, i.e., $\mathbf{M} \sim [G_1, \ldots, G_N]$. We define a CG mapping as **valid** when each site graph, G_I, is connected. If $e_{ij} \in E$ is an edge between two atoms i and j that are in different sites, then we refer to e_{ij} as an "out-of-site" edge. For each pair of distinct sites, $I \neq J$, we define $E_{IJ} = \{e_{ij} \in E | i \in V_I, j \in V_J\}$ as the set of out-of-site edges.

Lemma 1 Let $G = (V, E)$ be a valid protein graph for an n residue protein. Let $G_I = (V_I, E_I)$ be a site graph for which $2 \leq |V_I| \leq n - 1$. Then G_I contains at least one atom, $i_* \in V_I$, such that $i_* > 1$ and i_* forms an out-of-site edge. **Proof:** Define $i_+ \equiv \max\{i \in V_I\}$. Note that $i_+ > 1$ because $|V_I| \geq 2$. There exist two cases. Case 1: $i_+ \leq n - 1$. In this case, we define $j_+ \equiv i_+ + 1 \leq n$, such that $j_+ \in V$ and $e_{i_+ j_+} \in E$ because G is singly connected along the backbone. Since $j_+ > i_+$, $j_+ \notin V_I$ and $e_{i_+ j_+}$ is an out-of-site edge. Therefore, $i_* = i_+ \in V_I$ satisfies Lemma 1. Case 2: $i_+ = n$. We define $i_1 \equiv \max\{i \in V_I | i - 1 \notin V_I\}$. Because $|V_I| < n$ and $i_+ = n \in V_I$, it follows that $i_1 > 1$. We define $j_1 \equiv i_1 - 1 \geq 1$, such that $j_1 \in V$ and $e_{i_1 j_1} \in E$ because G is singly connected along the backbone. The definition of i_1 implies that $j_1 \notin V_I$ and, thus, $e_{i_1 j_1}$ is an out-of-site edge. Therefore, $i_* = i_1 \in V_I$ satisfies Lemma 1. Consequently, in either case, there exists at least one atom $i_* \in V_I$ such that $i_* > 1$ and i_* forms an out-of-site edge.

3. Steal move set

We employ a "steal" move set to explore mapping space. Starting from a valid map, $\mathbf{M} \sim [G_1, \ldots, G_N]$, we generate a new map, $\mathbf{M}'$, by transferring an atom $i \in V_I$ from site I to a new site $J(\neq I)$, while leaving the remaining $N-2$ sites unchanged. This steal move creates two new sites, $G'_I = G_I - i$ and $G'_J = G_J + i$, that are the subgraphs of G induced on $V'_I = V_I - \{i\}$ and $V'_J = V_J \cup \{i\}$, respectively. The steal move is **allowed** if the new map, $\mathbf{M}'$, is a valid map, i.e., if the modified sites, G'_I and G'_J, are both connected. Thus, atom $i \in V_I$ is **stealable** if two conditions are fulfilled: (1) i is not a cut-vertex (i.e., not an articulation node) of G_I and (2) i forms an out-of-site edge, $e_{ij} \in E$, with an atom $j \in V$ that is not in site I.

Lemma 2 Given a valid map, $\mathbf{M}$, an allowable steal move, $\mathbf{M} \to \mathbf{M}'$, is reversible, i.e., there exists an allowed steal move $\mathbf{M}' \to \mathbf{M}$. **Proof:** Since the map, $\mathbf{M} \sim [G_1, \ldots G_N]$, is valid, each of the site graphs $G_1, \ldots, G_N$ is connected. The allowed (forward) steal move, $\mathbf{M} \to \mathbf{M}'$, creates two new sites: $G_I \to G'_I = G_I - i$ and $G'_J = G_J + i$. Consider the (reverse) steal move that transfers atom i from site J' to site I'. This creates two new sites, $G''_I \equiv G'_I + i = G_I$ and $G''_J \equiv G'_J - i = G_J$, that correspond to the original connected sites in $\mathbf{M}$. The forward and reverse steal moves leave the remaining $N-2$ sites unchanged. Therefore, the reverse steal move is allowed and recreates the original map, $\mathbf{M}' \to \mathbf{M}$.

4. Block decomposition

Our proof of ergodicity employs the block decomposition of simple graphs, G, as described in Section 5.4 of Gross and Yellen.[120] A block $B_k = (V_k, E_k)$ is a maximally connected subgraph of G such that B_k has no cut-vertices. For a connected graph, G, with two or more vertices (and no loops), the blocks are either (1) pairs of vertices that are connected by an edge or (2) maximal 2-connected subgraphs of G with 3 or more vertices. A block B_k is a leaf block if it contains exactly one vertex, $v \in V_k$, that is a cut-vertex of G. The set of blocks, $\{B_k\}$, define the block decomposition of G.

This block decomposition, $\{B_k\}$, of a graph $G = (V, E)$ has the following useful properties:

2. Two blocks can share at most one vertex.

3. A vertex, $v \in V$, is a member of two or more blocks if and only if v is a cut-vertex of

G.

4. The block decomposition partitions the edges of G, i.e., $\cup_k E_k = E$ and $E_k \cap E_{k'} = \varnothing$ if $k \neq k'$.

5. All graphs with at least one cut-vertex contain at least two leaf blocks.

6. Let B_1 and B_2 be two blocks of G. Let v_1 and v_2 be vertices in B_1 and B_2, respectively, that are not cut-vertices of G. Then G does not contain an edge from v_1 to v_2.

The last property leads to a result that is particularly useful for proving the ergodicity of the steal move-set.

Lemma 3 Let $G = (V, E)$ be a valid atomic protein graph. Let $B_k = (V_k, E_k)$ be a block of a connected site graph, $G_I = (V_I, E_I)$. Suppose that $i \in V_k$ is not a cut-vertex of the site graph G_I. If $e_{ij} \in E$ and $j \notin V_k$, then i is stealable. **Proof:** Because $i \in V_k$ is not a cut-vertex of G_I, property 3 implies that i is not present in any other block of G_I. Because the block decomposition partitions E_I between blocks (property 4), any in-site edge, $e_{it} \in E_I$, must connect atom i to another atom t in the same block, B_k. Since $j \notin V_k$ by hypothesis, it follows that $e_{ij} \notin E_I$. Therefore, the edge, $e_{ij} \in E$ connects atom $i \in V_I$ to an atom j in a distinct site, $J(\neq I)$. Because i is not a cut-vertex of G_I and e_{ij} is an out-of-site edge, atom i is stealable.

5. Proposition 1

Let $G = (V, E)$ be a valid protein graph. Let $G_I = (V_I, E_I)$ be a connected site graph with $2 \leq |V_I| \leq n - 1$. Then G_I contains at least one stealable vertex, $i_s \in V_I$, that is not the N-terminal (i.e., $i_s > 1$).

Proof: If G_I does not contain a cut-vertex, then the proposition follows directly from Lemma 1. Conversely, if G_I does contain a cut-vertex, then the proposition follows by finding a leaf block that contains a stealable vertex, $i_s > 1$.

Suppose that G_I does not contain a cut-vertex. Lemma 1 states that G_I contains an atom $i_* \in V_I$ for which $i_* > 1$ and i_* forms an out-of-site edge. Because $i_* \in V_I$ and G_I does not contain a cut-vertex, i_* is stealable. Therefore, $i_s = i_*$ fulfills the proposition.

Alternatively, suppose that G_I does contain a cut-vertex. Property 5 above implies that the block decomposition, $\{B_k\}$, of G_I contains at least two leaf blocks. Since there exist at

least two leaf blocks and only 2 termini (i.e., $i_{\mathrm{N}} = 1$ and $i_{\mathrm{C}} = n$), there are only two cases to consider:

A: At least one leaf block does not contain a terminal vertex.

B: The block decomposition contains only two leaf blocks and both leaf blocks contain a terminal vertex.

In both cases, we select a single leaf block, $L_b = (V_b, E_b)$, of G_I. In case A, we select a leaf block that does not contain a terminal vertex. In case B, we select the leaf block that contains $i_{\mathrm{C}} = n$, but not $i_{\mathrm{N}} = 1$. In either case, L_b does not contain the N-terminal. Thus, if L_b contains a stealable vertex, $i_s \in V_b$, then i_s satisfies the proposition.

In both cases, the selected leaf block, L_b, contains at least two vertices, $|V_b| \geq 2$, and only one vertex, c, that is a cut-vertex of the site graph G_I. Consequently, the set $V_{b|c} = V_b - \{c\}$ is a non-empty set of vertices that are not cut-vertices of G_I. We define the maximal and minimal vertices in $V_{b|c}$: $i_+ = \max\{i \in V_{b|c}\}$ and $i_- = \min\{i \in V_{b|c}\}$. We also define $j_+ = i_+ + 1$ and $j_- = i_- - 1$, such that $j_+ \neq j_-$ and $j_+, j_- \notin V_{b|c}$.

case A In case A, we choose a leaf block, $L_b = (V_b, E_b)$, that does not contain a terminal vertex. Because $i_-, i_+ \in V_b$, it follows that $2 \leq i_- \leq i_+ \leq n-1$. Consequently, $1 \leq j_- < j_+ \leq n$, which implies that both $j_+, j_- \in V$. Because $j_+ \neq j_-$ and $j_+, j_- \notin V_{b|c}$, at least one vertex of the pair $\{j_+, j_-\}$ is not an element of V_b. We denote this vertex as $j' \notin V_b$ and define i' as the corresponding vertex in the pair $\{i_+, i_-\}$. Then $e_{i'j'} \in E$ because $i', j' \in V$, $|j' - i'| = 1$, and the valid protein graph, G, is singly connected along the backbone. Since $j' \notin V_b$ and $i' \in V_{b|c} \subset V_b$ is not a cut-vertex of the site graph G_I, Lemma 3 implies that $i_s = i' \in V_I$ is stealable.

case B In case B, we choose a leaf block, $L_b = (V_b, E_b)$, that contains the C-terminal vertex, but not the N-terminal vertex, i.e., $i_{\mathrm{C}} = n \in V_b$ and $i_{\mathrm{N}} = 1 \notin V_b$. In this case, $2 \leq i_-$ such that $j_- \geq 1 \in V$ and $e_{i_-j_-} \in E$, but i_+ may be the C-terminal vertex, $i_{\mathrm{C}} = n$. If $i_+ < n$, then $1 \leq j_- < j_+ \leq n$, such that $j_-, j_+ \in V$ and the proposition follows exactly as in case A. Conversely, if $i_+ = n$, then the proof is more cumbersome because $j_+ \notin V$ and we must focus on j_-. Note that $i_- \in V_{b|c}$ is not a cut-vertex of G_I and $j_- \notin V_{b|c}$. However, we must consider whether j_- is the cut-vertex, c, of G_I. If $j_- \neq c$, then $j_- \notin V_b$ and Lemma 3 implies that $i_s = i_- \in V_I$ is stealable. In contrast, if $j_- = c$, then $j_- \in V_b$. This implies that $j_- \geq 2$ and, moreover, j_- is the minimal element of V_b. We define $j_{2-} = j_- - 1 \geq 1$, such

that $j_{2-} \in V$ but $j_{2-} \notin V_b$. We now rely upon the double-connectivity along the backbone, such that $e_{i_- j_{2-}} \in E$ because $i_- - j_{2-} = 2$ and $i_-, j_{2-} \in V$. Consequently, $i_s = i_- \in V_I$ is stealable also if $j_- = c$.

Thus, in either case A or B, there exists a vertex $i_s \in V_I$ that is stealable and $i_s > 1$. Therefore, proposition 1 holds irrespective of whether G_I contains a cut-vertex.

6. Proposition 2

a. Preliminary considerations

Let $G = (V, E)$ be a valid protein graph with vertex set $V = \{1, 2, \ldots, n\}$. The edge set $E = \{e_{ij} | i, j \in V, \theta_{ij} = 1\}$ is singly and doubly connected along the backbone, i.e., $e_{ij} \in E$ for all $i, j \in V$ such that $|i - j| = 1$ or 2.

We define the N-site "monster" map $\mathbf{M}^*_{Nm} \sim [G^*_{1m}, \ldots G^*_{Nm}]$. Each of the first $N-1$ sites of $\mathbf{M}^*_{Nm}$ is associated with the corresponding single atom, i.e., $V^*_{im} = \{i\}$ for $i = 1, \ldots, N-1$. The last site is associated with the remaining $n-N+1$ atoms, i.e., $V^*_{Nm} = \{N, N+1, \ldots, n\}$.

Given G, we define a truncated (and shifted) protein graph, $G^{(k)} = (V^{(k)}, E^{(k)})$, by eliminating the first k N-terminal vertices from G and then shifting the labels of the remaining $n_k \equiv n - k$ vertices by k. The vertex set $V^{(k)} = \{1, 2, \ldots, n_k\}$ corresponds to the last n_k vertices of V, i.e., each $i \in V^{(k)}$ corresponds to a vertex $i_{\mathrm{R}} \equiv i + k \in V$ of the original graph, G. The edge set, $E^{(k)}$, is the set of edges among the corresponding vertices of the original graph, i.e., $E^{(k)} = \{e_{ij} | i, j \in V^{(k)}, e_{i+k,j+k} \in E\}$. Note that $E^{(k)}$ includes all of the edges among the vertices $i, j \geq k + 1$ of the original graph, G. In particular, the truncated graph $G^{(k)}$ is singly and doubly connected along the backbone. Consequently, $G^{(k)}$ is a valid protein graph for a protein with n_k amino acids.

Let $\mathbf{M}^{(k)} \sim [G^{(k)}_1, \ldots, G^{(k)}_{N_k}]$ be a map that represents the truncated graph, $G^{(k)}$, with $N_k \equiv N - k$ sites. The map, $\mathbf{M}^{(k)}$, corresponds to a unique map, $\mathbf{M}^{(k)}_{\mathrm{R}}$, that represents G with N sites:

$$\mathbf{M}^{(k)}_{\mathrm{R}} \sim [G^*_{1m}, \ldots, G^*_{km}, G^{(k)}_{k+1\mathrm{R}}, \ldots, G^{(k)}_{N\mathrm{R}}]. \tag{A.1}$$

The first k sites of $\mathbf{M}^{(k)}_{\mathrm{R}}$ correspond to the first k sites of $\mathbf{M}^*_{Nm}$, i.e., to the individual atoms $1, \ldots, k$ in the original protein graph. The last N_k sites of $\mathbf{M}^{(k)}_{\mathrm{R}}$ are obtained from the sites of $\mathbf{M}^{(k)}$. Specifically, for $I = 1, \ldots, N_k$, the site vertex set is $V^{(k)}_{k+I\mathrm{R}} = \{k + i | i \in V^{(k)}_I\}$,

while the site edge set is $E^{(k)}_{k+I\mathrm{R}} = \{e_{i+k,j+k} | e_{ij} \in E^{(k)}_I\}$. Importantly, the relabelled site $G^{(k)}_{k+I\mathrm{R}} = (V^{(k)}_{k+I\mathrm{R}}, E^{(k)}_{k+I\mathrm{R}})$ is a connected site graph for the original graph, G, if and only if $G^{(k)}_I = (V^{(k)}_I, E^{(k)}_I)$ is a connected site graph for the truncated graph, $G^{(k)}$. Consequently, $\mathbf{M}^{(k)}_{\mathrm{R}}$ is a valid map for G if and only if $\mathbf{M}^{(k)}$ is a valid map for $G^{(k)}$. Moreover, a steal move $\mathbf{M}^{(k)}_{\mathrm{R}} \to \mathbf{M}^{(k)'}_{\mathrm{R}}$ is allowed for G if and only if $\mathbf{M}^{(k)} \to \mathbf{M}^{(k)'}$ is allowed for $G^{(k)}$.

b. Statement of proposition 2

Starting from any allowed N-site CG mapping, $\mathbf{M} \sim [G_1, \ldots, G_N]$, it is possible to reach $\mathbf{M}^*_{Nm}$ via a series of steal moves.

c. Proof of proposition 2

We first provide a brief summary of the proof. Starting from $\mathbf{M} \sim [G_1, \ldots, G_N]$, we steal atoms from site 1 until it contains only the N-terminus of the protein. Proposition 1 implies that these steal moves are allowed. Then at each subsequent step $k \geq 1$, we hold the first k sites fixed and treat the remaining $n - k$ atoms as a truncated protein that we represent with $N-k$ sites. Because the truncated protein corresponds to a valid protein graph, we can apply proposition 1 in each iteration to reduce the next site, $I = k+1$, to the corresponding single atom, $i = k + 1$. After completing this for step $k = N - 2$, we have transformed $\mathbf{M}$ to $\mathbf{M}^*_{Nm}$. Below we give a more detailed proof.

Step 0: Given a valid protein graph, $G = (V, E)$, with $V = \{1, \ldots, n\}$, we define $G^{(0)} = (V^{(0)}, E^{(0)})$, where $V^{(0)} = V$ and $E^{(0)} = E$. Given the valid mapping, $\mathbf{M}$, we relabel the sites as necessary so that the N-terminus, $i_{\mathrm{N}} = 1$, is in site 1. This leads to $\mathbf{M}^{(0)}_1 \sim [G^{(0)}_{1;1}, \ldots, G^{(0)}_{N;1}]$, where $i_{\mathrm{N}} = 1 \in V^{(0)}_{1;1}$ and each site graph, $G^{(0)}_{I;1}$, is connected. Without loss of generality, we assume that the first site graph, $G^{(0)}_{1;1}$, contains more than one vertex, i.e., $|V^{(0)}_{1;1}| > 1$. Proposition 1 implies that we can steal an atom $i_1 \in V^{(0)}_{1;1}$ with $i_1 > 1$ from the first site. This steal move leads to a new allowed map, $\mathbf{M}^{(0)}_2 \sim [G^{(0)}_{1;2}, \ldots, G^{(0)}_{N;2}]$, where $G^{(0)}_{1;2} = G^{(0)}_{1;1} - i_1$ and all of the sites are connected. By proposition 1, we can continue to steal atoms from site 1 until its graph has been reduced to the N-terminal vertex, $i_{\mathrm{N}} = 1$. At this point, we relabel the CG sites such that site 2 contains the atomic vertex $i = 2$. We obtain a valid map $\mathbf{M}^{(0)}_\infty \sim [G^{(0)}_{1;\infty}, G^{(0)}_{2;\infty}, \ldots, G^{(0)}_{N;\infty}]$ with $G^{(0)}_{1;\infty} = G^*_{1m} = \{1\}$, vertex $i = 2 \in V^{(0)}_{2;\infty}$, and each

site is connected.

For each iteration, $k = 1, \ldots, N-2$, we define $n_k = n - k$ and $N_k = N - k$. We then perform the following two steps.

Step 1: We now eliminate vertex $i = 1$ (and its edges) from $G^{(k-1)}$ and shift the labels of the remaining n_k vertices $i \to i-1$ in order to obtain the truncated (and shifted) protein graph, $G^{(k)} = (V^{(k)}, E^{(k)})$. As described above, $G^{(k)}$ is a valid protein graph with $V^{(k)} = \{1, \ldots, n_k\}$ that is singly and doubly connected along the backbone. Similarly, we remove the first site from $\mathbf{M}_\infty^{(k-1)}$ while shifting the vertex and site labels by 1, i.e., $i \to i-1$ and $I \to I-1$. The resulting map, $\mathbf{M}_1^{(k)} \sim [G_{1;1}^{(k)}, \ldots, G_{N_k;1}^{(k)}]$, is a valid N_k-site map for the n_k-vertex graph, $G^{(k)}$, with the N-terminus of the truncated protein in the first site, $G_{1;1}^{(k)}$. Note that the N_k-site map, $\mathbf{M}_1^{(k)}$, for the truncated n_k-atom protein corresponds to a valid N-site map, $\mathbf{M}_{1\mathrm{R}}^{(k)}$, for the original n-vertex protein graph, G. The first k sites of $\mathbf{M}_{1\mathrm{R}}^{(k)}$ correspond to the first k atoms of G, i.e., $G_{1m}^*, \ldots, G_{km}^*$. The remaining N_k sites of $\mathbf{M}_{1\mathrm{R}}^{(k)}$ are obtained by relabelling the vertices and sites of $\mathbf{M}_1^{(k)}$.

Step 2: We are now in an analogous situation to step 0. $G^{(k)}$ is a valid protein graph with n_k vertices that is singly and doubly connected along the backbone. $\mathbf{M}_1^{(k)}$ is a valid N_k-site map for $G^{(k)}$ that assigns (shifted) vertex 1 to (shifted) site 1. Proposition 1 implies that we can again steal vertices from this first site until we have reduced $G_{1;1}^{(k)}$ to a single vertex that corresponds to the N-terminus of $G^{(k)}$. After performing these valid steal moves, we relabel the CG sites to obtain a valid map, $\mathbf{M}_\infty^{(k)} \sim [G_{1;\infty}^{(k)}, G_{2;\infty}^{(k)}, \ldots, G_{N_k;\infty}^{(k)}]$, with $V_{1;\infty}^{(k)} = \{1\}$, vertex $2 \in V_{2;\infty}^{(k)}$, and each site is connected. This series of steal moves for the truncated graph, $G^{(k)}$, corresponds to a series of valid steal moves for the original graph, G, in which the first k sites are unchanged, while site $k+1$ is reduced to the single vertex $k+1$. The resulting N-site map is then

$$\mathbf{M}_{\mathrm{R};\infty}^{(k)} \sim [G_{1m}^*, \ldots, G_{k+1m}^*, G_{k+2\mathrm{R};\infty}^{(k)}, \ldots, G_{N\mathrm{R};\infty}^{(k)}], \tag{A.2}$$

where, for $I = 2, \ldots, N_k$, $G_{k+I\mathrm{R};\infty}^{(k)}$ is obtained by relabelling $G_{I;\infty}^{(k)}$.

We iteratively apply steps 1 and 2 for $k = 1, \ldots, N-2$. Each successive iteration gives a series of valid steal moves that transform $\mathbf{M}$ by successively reducing site $k+1$ to the corresponding atom, while leaving the preceding k sites unchanged. After completing iteration $k = N-2$, we arrive at the desired monster map.

REFERENCES

[1]Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. Chem. Rev., 116:7898–7936, 2016.

[2]Marco Giulini, Marta Rigoli, Giovanni Mattiotti, Roberto Menichetti, Thomas Tarenzi, Raffaele Fiorentini, and Raffaello Potestio. From System Modeling to System Analysis: The Impact of Resolution Level and Resolution Distribution in the Computer-Aided Investigation of Biomolecules. Front. Mol. Biosci., 8:676976, June 2021.

[3]Jaehyeok Jin, Alexander J. Pak, Aleksander E. P. Durumeric, Timothy D. Loose, and Gregory A. Voth. Bottom-up Coarse-Graining: Principles and Perspectives. Journal of Chemical Theory and Computation, 18(10):5759–5791, October 2022.

[4]Friederike Schmid. Understanding and Modeling Polymers: The Challenge of Multiple Scales. ACS Polymers Au, 3(1):28–58, February 2023.

[5]W. G. Noid. Perspective: Advances, challenges, and insight for predictive coarse-grained models. J. Phys. Chem. B, 127:4174–4207, 2023.

[6]MG Guenza, M Dinpajooh, J McCarty, and IY Lyubimov. Accuracy, transferability, and efficiency of coarse-grained models of molecular liquids. J. Phys. Chem. B, 122(45):10257–10278, 2018.

[7]Thomas E. Gartner and Arthi Jayaraman. Modeling and simulations of polymers: A roadmap. Macromolecules, 52(3):755–786, 2019.

[8]Satyen Dhamankar and Michael A. Webb. Chemically specific coarse-graining of polymers: Methods and prospects. Journal of Polymer Science, 59(22):2613–2643, 2021.

[9]W. G. Noid. Perspective: coarse-grained models for biomolecular systems. J. Chem. Phys., 139(9):090901, 2013.

[10]A. P. Lyubartsev and A. Laaksonen. Calculation of effective interaction potentials from radial distribution functions: a reverse Monte Carlo approach. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics, 52:3730–3737, 1995.

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

[12]S. Izvekov and G. A. Voth. A multiscale coarse-graining method for biomolecular systems. J. Phys. Chem. B, 109:2469 – 2473, 2005.

[13]S. Izvekov and G. A. Voth. Multiscale coarse graining of liquid-state systems. J. Chem. Phys., 123:134105, 2005.

[14]M. Scott Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. J. Chem. Phys., 129:144108, 2008.

[15]Alexey Savelyev and Garegin A. Papoian. Molecular renormalization group coarse-graining of electrolyte solutions: Applications to aqueous NaCl and KCl. J. Phys. Chem. B, 113:7785–93, 2009.

[16]W G Noid. Systematic methods for structurally consistent coarse-grained models. Methods Mol Biol, 924:487–531, 2013.

[17]Sereina Riniker, Jane R Allison, and Wilfred F van Gunsteren. On developing coarse-grained models for biomolecular simulation: a review. Phys. Chem. Chem. Phys., 14(36):12423–12430, Sep 2012.

[18]V. A. Harmandaris, D. Reith, N. F. A. Van der Vegt, and K. Kremer. Comparison between coarse-graining models for polymer systems: Two mapping schemes for polystyrene. Macromol. Chem. Phys., 208:2109–2120, 2007.

[19]Takahiro Ohkuma and Kurt Kremer. Comparison of two coarse-grained models of cis-polyisoprene with and without pressure correction. Polymer, 130:88–101, 2017.

[20]V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko. Versatile object-oriented toolkit for coarse-graining applications. J. Chem. Theory Comput., 5(12):3211–3223, 2009.

[21]Avisek Das, Lanyuan Lu, Hans C. Andersen, and Gregory A. Voth. The multiscale coarse-graining method. X. Improved algorithms for constructing coarse-grained potentials for molecular systems. J. Chem. Phys., 136(19):194115, 2012.

[22]Joseph F. Rudzinski and William G. Noid. Investigation of coarse-grained mappings via an iterative generalized yvon-born-green method. J. Phys. Chem. B, 118(28):8295–8312, 2014.

[23]Joseph F. Rudzinski and William G. Noid. Bottom-up coarse-graining of peptide ensembles and helix-coil transitions. J. Chem. Theory Comput., 11(3):1278–1291, 2015.

[24]J. W. Mullinax and W. G. Noid. Extended ensemble approach for deriving transferable coarse-grained potentials. J. Chem. Phys., 131:104110, 2009.

[25]Marco Dallavalle and Nico FA van der Vegt. Evaluation of mapping schemes for systematic coarse graining of higher alkanes. Phys. Chem. Chem. Phys., 19(34):23034–23042, 2017.

[26]Jaehyeok Jin, Yining Han, and Gregory A. Voth. Ultra-Coarse-Grained Liquid State Models with Implicit Hydrogen Bonding. J. Chem. Theory Comput., 14(12):6159–6174, December 2018.

[27]Aditi Khot, Stephen B Shiring, and Brett M Savoie. Evidence of information limitations in coarse-grained models. J. Chem. Phys., 151(24):244105, 2019.

[28]Maghesree Chakraborty, Jinyu Xu, and Andrew D White. Is preservation of symmetry necessary for coarse-graining? Phys. Chem. Chem. Phys., 22(26):14998–15005, 2020.

[29]Patrice Koehl, Frederic Poitevin, Rafael Navaza, and Marc Delarue. The renormalization group and its applications to generating coarse-grained models of large biological molecular systems. J. Chem. Theory Comput., 13(3):1424–1438, 2017.

[30]Maghesree Chakraborty, Chenliang Xu, and Andrew D White. Encoding and selecting coarse-grain mapping operators with hierarchical graphs. J. Chem. Phys., 149(13):134106, 2018.

[31]Michael A. Webb, Jean-Yves Delannoy, and Juan J. de Pablo. Graph-Based Approach to Systematic Molecular Coarse-Graining. J. Chem. Theory Comput., December 2018.

[32]Patrick Diggins, Changjiang Liu, Markus Deserno, and Raffaello Potestio. Optimal Coarse-Grained Site Selection in Elastic Network Models of Biomolecules. J. Chem. Theory Comput., 15(1):648–664, January 2019.

[33]Xiang Fu, Tian Xie, Nathan J. Rebello, Bradley D. Olsen, and Tommi Jaakkola. Simulate Time-integrated Coarse-grained Molecular Dynamics with Geometric Machine Learning. June 2022. arXiv:2204.10348 [physics], doi:10.48550/ARXIV.2204.10348.

[34]A. Arkhipov, P.L. Freddolino, and K. Schulten. Stability and dynamics of virus capsids described by coarse-grained modeling. Structure, 129:1767–77, 2006.

[35]Wujie Wang and Rafael Gómez-Bombarelli. Coarse-graining auto-encoders for molecular dynamics. npj Comput. Mat., 5(1):125, December 2019.

[36]Jurgis Ruza, Wujie Wang, Daniel Schwalbe-Koda, Simon Axelrod, William H Harris, and Rafael Gómez-Bombarelli. Temperature-transferable coarse-graining of ionic liquids with dual graph convolutional neural networks. J. Chem. Phys., 153(16):164501, 2020.

[37]Zhiheng Li, Geemi P Wellawatte, Maghesree Chakraborty, Heta A Gandhi, Chenliang Xu, and Andrew D White. Graph neural network based coarse-grained mapping prediction. Chem., 11(35):9524–9531, 2020.

[38]Federico Errica, Marco Giulini, Davide Bacciu, Roberto Menichetti, Alessio Micheli, and

Raffaello Potestio. A deep graph network–enhanced sampling approach to efficiently explore the space of reduced representations of proteins. Frontiers in Molecular Biosciences, 8:637396, 2021.

[39]Keverne A. Louison, Ian L. Dryden, and Charles A. Laughton. GLIMPS: A Machine Learning Approach to Resolution Transformation for Multiscale Modeling. Journal of Chemical Theory and Computation, 17(12):7930–7937, December 2021.

[40]Shriram Chennakesavalu, David J. Toomer, and Grant M. Rotskoff. Ensuring thermodynamic consistency with invertible coarse-graining. The Journal of Chemical Physics, 158(12):124126, March 2023.

[41]Marco Giulini, Roberto Menichetti, M Scott Shell, and Raffaello Potestio. An information-theory-based approach for optimal model reduction of biomolecules. J. Chem. Theory Comput., 16(11):6795–6813, 2020.

[42]Roberto Menichetti, Marco Giulini, and Raffaello Potestio. A journey through mapping space: characterising the statistical and metric properties of reduced representations of macromolecules. The European Physical Journal B, 94(10):204, October 2021.

[43]Roi Holtzman, Marco Giulini, and Raffaello Potestio. Making sense of complex systems through resolution, relevance, and mapping entropy. Phys. Rev. E, 106:044101, Oct 2022.

[44]Margherita Mele, Roberto Covino, and Raffaello Potestio. Information-theoretical measures identify accurate low-resolution representations of protein configurational space. Soft Matter, 18(37):7064–7074, 2022.

[45]Joseph F Rudzinski and W G Noid. Coarse-graining entropy, forces, and structures. J. Chem. Phys., 135(21):214101, Dec 2011.

[46]H. Gohlke and M. F. Thorpe. A natural coarse graining for simulating large biomolecular motion. Biophys. J., 91:2115–20, 2006.

[47]Maria Stepanova. Dynamics of essential collective motions in proteins: Theory. Phys. Rev. E, 76:051918, Nov 2007.

[48]Min Li, John Z. H. Zhang, and Fei Xia. A new algorithm for construction of coarse-grained sites of large biomolecules. J. Comp. Chem., 37(9):795–804, App. Phys. Rev. 2016.

[49]Min Li, John Zenghui Zhang, and Fei Xia. Constructing Optimal Coarse-Grained Sites of Huge Biomolecules by Fluctuation Maximization. J. Chem. Theory Comput., 12(4):2091–2100, App. Phys. Rev. 2016.

[50]Z. Y. Zhang, L. Y. Lu, W. G. Noid, V. Krishna, J. Pfaendtner, and G. A. Voth. A

systematic methodology for defining coarse-grained sites in large biomolecules. Biophys. J., 95(11):5073–5083, 2008.

[51]Z. Y. Zhang, J. Pfaendtner, A. Grafmuller, and G. A. Voth. Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. Biophys. J., 97(8):2327–2337, 2009.

[52]Z. Zhang and G. A. Voth. Coarse-grained representations of large biomolecular complexes from low-resolution structural data. J. Chem. Theory Comput., 6:2990–3002, 2010.

[53]Anton V. Sinitskiy, Marissa G. Saunders, and Gregory A. Voth. Optimal number of coarse-grained sites in different components of large biomolecular complexes. J. Phys. Chem. B, 116(29):8363–8374, 2012.

[54]Jesper J. Madsen, Anton V. Sinitskiy, Jianing Li, and Gregory A. Voth. Highly Coarse-Grained Representations of Transmembrane Proteins. J. Chem. Theory Comput., 13(2):935–944, February 2017.

[55]A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. Proteins, 17:412 – 425, 1993.

[56]Lorenzo Boninsegna, Ralf Banisch, and Cecilia Clementi. A Data-Driven Perspective on the Hierarchical Assembly of Molecular Structures. J. Chem. Theory Comput., 14(1):453–460, January 2018. Publisher: American Chemical Society.

[57]Wangfei Yang, Clark Templeton, David Rosenberger, Andreas Bittracher, Feliks N´uske, Frank Noé, and Cecilia Clementi. Slicing and dicing: Optimal coarse-grained representation to preserve molecular kinetics. ACS Central Science, 2023.

[58]Thomas T. Foley, M. S. Shell, and W. G. Noid. The impact of resolution upon entropy and information in coarse-grained models. J. Chem. Phys., 143:243104, 2015.

[59]Thomas T Foley, Katherine M Kidder, M Scott Shell, and WG Noid. Exploring the landscape of model representations. Proc. Natl. Acad. Sci. U.S.A., 117(39):24061–24068, 2020.

[60]Katherine M Kidder, Ryan J. Szukalo, and W. G Noid. Energetic and entropic considerations for coarse-graining. Eur. Phys. J. B, 94:153, 2021.

[61]P. J. Flory, M. Gordon, and N. G. McCrum. Statistical thermodynamics of random networks [and discussion]. Proc. Roy. Soc. Lond. A: Math. Phys. Sci., 351(1666):351–380, 1976.

[62]Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian dynamics of folded proteins.

Phys. Rev. Lett., 79:3090–3093, Oct 1997.

[63]Ivet Bahar, Timothy R. Lezon, Ahmet Bakan, and Indira H. Shrivastava. Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. Chem. Rev., 110(3):1463–1497, 2010.

[64]I Bahar and AJ Rader. Coarse-grained normal mode analysis in structural biology. Curr. Opin. Struct. Biol., 15(5):586–592, OCT 2005.

[65]B Brooks and M Karplus. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. Proc. Natl. Acad. Sci. U.S.A., 82(15):4995–4999, August 1985.

[66]Jean-François Gibrat and Nobuhiro Gō. Normal mode analysis of human lysozyme: Study of the relative motion of the two domains and characterization of the harmonic motion. Proteins, 8(3):258–279, January 1990.

[67]Yasunobu Seno and Nobuhiro Gō. Deoxymyoglobin studied by the conformational normal mode analysis. J. Mol. Biol., 216(1):95–109, November 1990.

[68]Liliane Mouawad and David Perahia. Motions in Hemoglobin Studied by Normal Mode Analysis and Energy Minimization: Evidence for the Existence of Tertiary T-like, Quaternary R-like Intermediate Structures. J. Mol. Biol., 258(2):393–410, May 1996.

[69]F. Tama and Y.-H. Sanejouand. Conformational change of proteins arising from normal mode calculations. Protein Eng., 14(1):1–6, 2001.

[70]M Delarue and Y.-H Sanejouand. Simplified Normal Mode Analysis of Conformational Transitions in DNA-dependent Polymerases: the Elastic Network Model. J. Mol. Biol., 320(5):1011–1024, July 2002.

[71]Cristian Micheletti, Gianluca Lattanzi, and Amos Maritan. Elastic Properties of Proteins: Insight on the Folding Process and Evolutionary Selection of Native Structures. J. Mol. Biol., 321(5):909–921, August 2002.

[72]A.J. Rader and Ivet Bahar. Folding core predictions from network models of proteins. Polymer, 45(2):659–668, January 2004.

[73]A. J. Rader, Gülsüm Anderson, Basak Isin, H. Gobind Khorana, Ivet Bahar, and Judith Klein-Seetharaman. Identification of core amino acids stabilizing rhodopsin. Proc. Natl. Acad. Sci. U.S.A., 101(19):7246–7251, May 2004.

[74]Wenjun Zheng and Bernard R. Brooks. Normal-Modes-Based Prediction of Protein Conformational Changes Guided by Distance Constraints. Biophys. J., 88(5):3109–3117, May

2005.

[75]Wenjun Zheng, Bernard R. Brooks, and D. Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc. Natl. Acad. Sci. U.S.A., 103(20):7664–7669, May 2006.

[76]Florence Tama, Osamu Miyashita, and Charles L. Brooks III. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. J. Mol. Biol., 337(4):985 – 999, 2004.

[77]Marc Delarue and Philippe Dumas. On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. Proc. Natl. Acad. Sci. U.S.A., 101(18):6957–6962, May 2004.

[78]C. Gorba, O. Miyashita, and F. Tama. Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data. Biophys. J., 94(5):1589–1599, Mar 2008.

[79]Jianpeng Ma. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. Structure, 13(3):373–380, March 2005.

[80]Philippe Durand, Georges Trinquier, and Yves-Henri Sanejouand. A new approach for determining low-frequency normal modes in macromolecules. Biopolymers, 34(6):759–771, June 1994.

[81]Florence Tama, Florent Xavier Gadea, Osni Marques, and Yves-Henri Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. Proteins, 41(1):1–7, October 2000.

[82]Monique M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett., 77:1905–1908, Aug 1996.

[83]Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. Proteins, 33(3):417–429, 1998.

[84]Konrad Hinsen, Aline Thomas, and Martin J. Field. Analysis of domain motions in large proteins. Proteins, 34(3):369–382, February 1999.

[85]Sibsankar Kundu, Dan C. Sorensen, and George N. Phillips. Automatic domain decomposition of proteins by a gaussian network model. Proteins, 57(4):725–733, 2004.

[86]P. Doruker, R. L. Jernigan, and I. Bahar. Dynamics of large proteins through hierarchical levels of coarse-grained structures. J. Comp. Chem., 23(1):119–127, 2002.

[87]Chakra Chennubhotla, A J Rader, Lee-Wei Yang, and Ivet Bahar. Elastic network models

for understanding biomolecular machinery: from enzymes to supramolecular assemblies. Physical Biology, 2(4):S173–S180, November 2005.

[88]Chakra Chennubhotla and Ivet Bahar. Markov propagation of allosteric effects in biomolecular systems: application to GroEL–GroES. Molecular Systems Biology, 2(1):36, January 2006.

[89]Kelin Xia, Kristopher Opron, and Guo-Wei Wei. Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (mANM). J. Chem. Phys., 143(20):204106, November 2015.

[90]S. Nicolay and Y.-H. Sanejouand. Functional Modes of Proteins Are among the Most Robust. Phys. Rev. Lett., 96(7):078104, February 2006.

[91]W. G. Krebs, Vadim Alexandrov, Cyrus A. Wilson, Nathaniel Echols, Haiyuan Yu, and Mark Gerstein. Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. Proteins, 48(4):682–695, September 2002.

[92]H. T. Davis and K. T. Thomson. Linear Algebra and Linear Operators in Engineering. Academic Press, 2000.

[93]S. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Stat., 22(1):79–86, 1951.

[94]Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley Interscience, 2 edition, 2006.

[95]Bernard R. Brooks, Dusanka Janezic, and Martin Karplus. Harmonic analysis of large systems. I. Methodology. J. Comp. Chem., 16(12):1522–1542, December 1995.

[96]H. Goldstein, C. Poole, and J. Safko. Classical Mechanics. Adison Wesley, 2002.

[97]R. Potestio, F. Pontiggia, and C. Micheletti. Coarse-Grained Description of Protein Internal Dynamics: An Optimal Strategy for Decomposing Proteins in Rigid Subunits. Biophys. J., 96(12):4993–5002, June 2009.

[98]Luca Ponzoni, Guido Polles, Vincenzo Carnevale, and Cristian Micheletti. SPECTRUS: A Dimensionality Reduction Approach for Identifying Dynamical Domains in Protein Complexes from Limited Structural Datasets. Structure, 23(8):1516–1525, August 2015.

[99]Paolo Calligari, Marco Gerolin, Daniel Abergel, and Antonino Polimeno. Decomposition of Proteins into Dynamic Units from Atomic Cross-Correlation Functions. J. Chem. Theory Comput., 13(1):309–319, January 2017.

[100]Colin Brown, Anuradha Agarwal, and Antoni Luque. pyCapsid: Identifying dominant dynamics and quasi-rigid mechanical units in protein shells. preprint, Bioinformatics, March 2023.

[101]Senadhi Vijay-Kumar, Charles E Bugg, and William J Cook. Structure of ubiquitin refined at 1.8 åresolution. J. Mol. Biol., 194(3):531–544, 1987.

[102]Ahmet Bakan, Lidio M. Meireles, and Ivet Bahar. Prody: Protein dynamics inferred from theory and experiments. Bioinformatics, 27(11):1575–1577, 2011.

[103]M. R. Shirts and J. D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. J. Chem. Phys., 129(12):124105, 2008.

[104]Marina Meilă. Comparing clusterings—an information based distance. J. Multivar. Anal., 98(5):873–895, May 2007.

[105]Frank H. Stillinger and Thomas A. Weber. Hidden structure in liquids. Phys. Rev. A, 25:978–989, Feb 1982.

[106]Dieter H E Gross. Microcanonical Thermodynamics. WORLD SCIENTIFIC, 2001.

[107]David J.C. MacKay. Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003.

[108]Nigel Goldenfeld. Lectures on Phase Transitions and the Renormalization Group. Westview Press, 1992.

[109]Sergei Izvekov and Betsy M. Rice. Hierarchical Machine Learning of Low-Resolution Coarse-Grained Free Energy Potentials. J. Chem. Theory Comput., 19(14):4436–4450, July 2023.

[110]Joseph F Rudzinski. Recent progress towards chemically-specific coarse-grained simulation models with consistent dynamical properties. Comput., 7(3):42, 2019.

[111]Viktor Klippenstein, Madhusmita Tripathy, Gerhard Jung, Friederike Schmid, and Nico F. A. van der Vegt. Introducing Memory in Coarse-Grained Molecular Simulations. The Journal of Physical Chemistry B, 125(19):4931–4954, May 2021.

[112]Tanja Schilling. Coarse-grained modelling out of equilibrium. Physics Reports, 972:1–45, August 2022.

[113]Gene Lamm and Attila Szabo. Langevin modes of macromolecules. J. Chem. Phys., 85(12):7334–7348, December 1986.

[114]C. Hijon, P. Espanol, E. Vanden-Eijnden, and R. Delgado-Buscalioni. Mori-Zwanzig formalism as a practical computational tool. Faraday Discuss., 144:301–322, 2010.

[115]Nicodemo Di Pasquale, Thomas Hudson, and Matteo Icardi. Systematic derivation of hybrid coarse-grained models. Phys. Rev. E, 99(1):013303, January 2019.

[116]Sergei Izvekov. Mori-Zwanzig projection operator formalism: Particle-based coarse-grained dynamics of open classical systems far from equilibrium. Phys. Rev. E, 104(2):024121, August 2021.

[117]Jaehyeok Jin, Eok Kyun Lee, and Gregory A. Voth. Understanding dynamics in coarse-grained models. III. Roles of rotational motion and translation-rotation coupling in coarse-grained dynamics. J. Chem. Phys., 159(16):164102, October 2023.

[118]J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. Xsede: Accelerating scientific discovery. Computing in Science & Engineering, 16(5):62–74, Sept.-Oct. 2014.

[119]William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual Molecular Dynamics. J. Mol. Graph., 14:33–38, 1996.

[120]Jonathan L Gross and Jay Yellen. Graph theory and its applications. CRC press, 2005.