Analysis of Mapping Atomic Models to Coarse-Grained Resolution

Katherine M. Kidder and W. G. Noid[a]

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

(Dated: 21 April 2025)

Low-resolution coarse-grained (CG) models provide significant computational and conceptual advantages for simulating soft materials. However, the properties of CG models depend quite sensitively upon the mapping, $\mathbf{M}$, that maps each atomic configuration, $\mathbf{r}$, to a CG configuration, $\mathbf{R}$. In particular, $\mathbf{M}$ determines how the configurational information of the atomic model is partitioned between the mapped ensemble of CG configurations and the lost ensemble of atomic configurations that map to each $\mathbf{R}$. In this work, we investigate how the mapping partitions the atomic configuration space into CG and intra-site components. We demonstrate that the corresponding coordinate transformation introduces a nontrivial Jacobian factor. This Jacobian factor defines a labelling entropy that corresponds to the uncertainty in the atoms that are associated with each CG site. Consequently, the labelling entropy effectively transfers configurational information from the lost ensemble into the mapped ensemble. Moreover, our analysis highlights the possibility of resonant mappings that separate the atomic potential into CG and intra-site contributions. We numerically illustrate these considerations with a Gaussian Network model for the equilibrium fluctuations of actin. We demonstrate that the spectral quality, $\mathcal{Q}$, provides a simple metric for identifying high quality representations for actin. Conversely, we find that neither maximizing nor minimizing the information content of the mapped ensemble results in high quality representations. However, if one accounts for the labelling uncertainty, $\mathcal{Q}(\mathbf{M})$ correlates quite well with the adjusted configurational information loss, $\hat{\mathrm{I}}_{\mathrm{map}}(\mathbf{M})$, that results from the mapping.

[a] Electronic mail: wgn1@psu.edu

I. INTRODUCTION

Richard Hamming famously asserted that "the purpose of computing is insight not numbers."[1] According to this premise, low resolution coarse-grained (CG) models provide a uniquely powerful framework for studying complex systems.[2,3] By eliminating unnecessary details, CG models provide the necessary computational efficiency for simulating length- and time-scales that cannot be effectively addressed with, e.g., conventional all-atom (AA) models.[4–6] Perhaps even more importantly, CG models provide researchers the opportunity to eliminate unnecessary details and precisely focus their intellectual resources on the features that are essential for understanding a particular phenomenon.[7–9] Unfortunately, it is not always easy to design CG models that properly distinguish "unnecessary details" from "essential features." Consequently, many recent studies have investigated the choice of CG representation, i.e., the degrees of freedom that are explicitly treated by the CG model.[10,11]

There exist many coarse-graining approaches with varying advantages and limitations.[12–14] In this work, we focus on bottom-up CG models that are based upon an underlying atomistic model.[11,15] In this case, the CG representation is precisely defined by a mapping, $\mathbf{M}$, that determines a unique CG configuration, $\mathbf{R} = \mathbf{M}(\mathbf{r})$, for each AA configuration, $\mathbf{r}$. Because the properties of bottom-up models can depend quite sensitively upon the CG mapping,[16–27] recent studies have proposed various metrics for optimizing $\mathbf{M}$.[10,11,28] These methods have often employed network-based[29–33] or machine-learning tools.[34–41] In particular, one class of studies has focused on preserving the large-amplitude, low-frequency motions of the AA model.[29,31,32,42–47] For instance, the essential dynamics coarse-graining (ED-CG) method of Voth and coworkers[48–52] first employs principle component analysis (PCA) to identify important collective motions[53] and then identifies CG sites with rigid atomic groups that preserve these essential dynamics. Recently, they have extended the ED-CG method with K-means clustering.[54] Conversely, a second class of studies has focused on preserving the configurational information of the AA model. In particular, Potestio and coworkers have proposed minimizing the mapping entropy,[38,55–58] which quantifies the configurational information that is lost when viewing the AA model at the CG resolution.[59,60]

Very recently, Foley, Kidder, and coworkers have adopted a complementary approach for investigating CG representations.[28,61–63] Specifically, they adopted the Gaussian Network Model (GNM) as an analytically tractable high resolution model for the equilibrium fluctuations of globular proteins about their folded conformation.[64–66] They employed Monte Carlo (MC) methods to systematically explore and statistically characterize the entire space of CG representations for the high resolution GNM. In particular, they focused on two metrics for assessing the quality of a given mapping, $\mathbf{M}$, based upon the mapped ensemble that results from viewing the high resolution ensemble at the CG resolution: (1) the spectral quality, $\mathcal{Q}(\mathbf{M})$, quantifies the mass-weighted covariance of the mapped ensemble; (2) the information content, $I(\mathbf{M})$, quantifies the information content of the mapped ensemble. In the case of the GNM, I is perfectly anti-correlated with the mapping entropy, $\mathrm{I}_{\mathrm{map}}$, considered by Potestio and coworkers: $\mathrm{I}_{\mathrm{map}}(\mathbf{M}) = mI(\mathbf{M}) + b$, where $m < 0$ and b are both independent of $\mathbf{M}$.[63] CG representations that maximized $\mathcal{Q}$ were highly consistent with the physical intuition that CG sites should correspond to distinct structural features that move coherently.[62,63] Conversely, CG representations that minimized the mapping information loss, $\mathrm{I}_{\mathrm{map}}$, were not consistent with this intuition. Interestingly, $\mathcal{Q}$ and $\mathrm{I}_{\mathrm{map}}$ were negatively correlated among high-resolution representations, but positively correlated at lower resolutions.[62,63] This suggests that it may be beneficial to design low-resolution representations that maximize the information lost from the AA model.

While this proposal may initially seem counter-intuitive, it perhaps can be rationalized.[67] The vibrational density of states for soft materials typically contains many high frequency modes. These high frequency modes are information-rich in that they describe localized motions that highly constrain the system. Conversely, the vibrational density of states typically contains relatively few low frequency modes. These low frequency modes are information-poor in that they describe delocalized motions that only weakly constrain the system. In this sense, most of the information in high-resolution models is high-frequency "noise," while relatively little is low-frequency "physics."[67] Thus, representations that minimize information loss may focus on preserving noise at the expense of physics.

We consider this proposal more closely in the present work. We first analyze the relationship between the mapping, $\mathbf{M}$, and the information content of the corresponding mapped ensemble. This analysis reveals a new source of information loss - the labelling entropy, H_{L} - that quantifies the uncertainty associated with the partitioning of atoms into CG particles. This analysis also suggests the notion of a 'resonance' between a family of high resolution potentials and a special CG mapping, $\mathbf{M}_{*}$. We numerically illustrate the consequences of the labelling entropy with a GNM for actin, which is considerably more complex than the proteins we have previously considered. We demonstrate that the spectral quality, $\mathcal{Q}$, identifies CG representations for actin that are consistent with our physical intuition. In contrast, we do not obtain physically reasonable representations by either maximizing or minimizing the mapping information loss, $\mathrm{I}_{\mathrm{map}}$. However, by accounting for the labelling entropy, the adjusted information loss, $\hat{\mathrm{I}}_{\mathrm{map}} = \mathrm{I}_{\mathrm{map}} + \mathrm{H}_{\mathrm{L}}$, is highly correlated with $\mathcal{Q}$. Finally, we briefly investigate resonant mappings by "atomizing" an idealized CG model and examining how the properties of the CG model vary as the mapping moves off of resonance.

The remainder of this manuscript is organized as follows. Section II reviews the Kullback-Leibler divergence[68,69] as a quantitative metric for information loss, analyzes a coordinate transformation associated with the mapping, and introduces the labelling entropy, H_{L}. Section III develops simple approximate models that allow us to illustrate H_{L} and its consequences. Section IV summarizes our computational methods, while section V presents calculations that numerically illustrate the analysis of Sections II and III. Section VI summarizes our findings and provides concluding comments. Finally, one appendix considers the impact of coarse-graining upon symmetries present in AA models, while a second appendix calculates the Jacobian associated with the coordinate transformation that is defined by the CG mapping.

II. THE MAPPING ENTROPY

A. Quantifying information content

We consider the canonical ensemble for an AA model with n atoms in a D-dimensional spatial region $\mathcal{D}(V)$ with volume $V = L^D$. We denote the AA potential by $u(\mathbf{r})$ and the AA configuration integral by $z = \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r})]$. The AA model is characterized by the configurational probability density $p_{\mathrm{r}}(\mathbf{r}) = \exp[-\beta u(\mathbf{r})]/z$. We quantify the information content of the AA canonical ensemble by

$$\mathrm{I}_{\mathrm{AA}} = \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r}\, p_{\mathrm{r}}(\mathbf{r}) \ln\left[p_{\mathrm{r}}(\mathbf{r}) / \, q_{\mathrm{r}}(\mathbf{r})\right], \tag{1}$$

which is the Kullback-Leibler divergence between $p_{\mathrm{r}}(\mathbf{r})$ and the corresponding uniform distribution for n atoms $q_{\mathrm{r}}(\mathbf{r}) = 1/V^n$.[68,69] I_{AA} is nonnegative and is proportional to the (negative) of the excess configurational entropy of the AA model.[70]

We define a mapping function, $\mathbf{M}(\mathbf{r})$, that determines a CG representation of each AA configuration, $\mathbf{R} = \mathbf{M}(\mathbf{r})$. The probability density for sampling a CG configuration, $\mathbf{R}$, in the resulting "mapped ensemble" is[71]

$$p_{\mathrm{R}}(\mathbf{R}) = z_{\mathrm{R}}(\mathbf{R})/z, \tag{2}$$

where

$$z_{\mathrm{R}}(\mathbf{R}) = \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r} \exp[-u(\mathbf{r})/k_B T]\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \tag{3}$$

is the restricted configuration integral, i.e., the total Boltzmann weight that is mapped to the CG configuration, $\mathbf{R}$. The mapped probability density, $p_{\mathrm{R}}(\mathbf{R})$, determines the configurational information that is present in the mapped ensemble:

$$\mathrm{I}_{\mathrm{CG}} = \int_{\mathcal{D}^N(V)} \mathrm{d}\mathbf{R}\, p_{\mathrm{R}}(\mathbf{R}) \ln\left[p_{\mathrm{R}}(\mathbf{R})/\, q_{\mathrm{R}}(\mathbf{R})\right], \tag{4}$$

where $q_{\mathrm{R}}(\mathbf{R}) = 1/V^N$ is the uniform distribution for the mapped configuration space. I_{CG} is proportional to (the negative of) the "apparent excess configurational entropy" present in the mapped ensemble.[60] The restricted configuration integral also determines the many-body potential of mean force (PMF)

$$W(\mathbf{R}) = -k_B T \ln\left[V^{-n_{\mathrm{x}}} z_{\mathrm{R}}(\mathbf{R})\right], \tag{5}$$

where we have defined $n_{\mathrm{x}} \equiv n - N$ as the number of particles that have been eliminated from the CG model.[72–74] The PMF is the AA free energy expressed as a function of CG coordinates. Consequently, the PMF is the appropriate CG potential for reproducing both structural and thermodynamic properties of the AA model.[28,75]

For each CG configuration, $\mathbf{R}$, the mapping also determines a "lost" subensemble, $\mathcal{S}_{\mathbf{R}}$, of AA configurations that map to $\mathbf{R}$:

$$\mathcal{S}_{\mathbf{R}} = \{\mathbf{r} \in \mathcal{D}^n(V) |\, \mathbf{M}(\mathbf{r}) = \mathbf{R}\}. \tag{6}$$

This subensemble is characterized by the conditioned distribution,

$$p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) = \exp[-u(\mathbf{r})/k_B T]\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})\,/z_{\mathrm{R}}(\mathbf{R}). \tag{7}$$

We quantify the configurational information contained in $\mathcal{S}_{\mathbf{R}}$ by

$$\mathrm{I}_{\mathrm{map}}(\mathbf{R}) = \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r}\, p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) \ln\left[p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})/\, q_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})\right], \tag{8}$$

where $q_{\mathrm{r|R}}(\mathbf{r}|\mathbf{R}) = V^{-n_{\mathrm{x}}}\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$ is the uniform conditioned distribution. This lost information determines the entropic contribution to the PMF:

$$W(\mathbf{R}) = E_{\mathrm{W}}(\mathbf{R}) + k_B T \mathrm{I}_{\mathrm{map}}(\mathbf{R}), \tag{9}$$

where $E_{\mathrm{W}}(\mathbf{R}) \equiv \langle u(\mathbf{r})\rangle_{\mathbf{R}}$ is a conditioned average of the atomic potential over $\mathcal{S}_{\mathbf{R}}$ evaluated according to $p_{\mathrm{r|R}}(\mathbf{r}|\mathbf{R})$.[28,75] Moreover, $\mathrm{I}_{\mathrm{map}}(\mathbf{R})$ determines the temperature-dependence of the PMF:

$$\left(\frac{\partial W(\mathbf{R})}{\partial T}\right)_{\mathbf{R}} = +k_B \mathrm{I}_{\mathrm{map}}(\mathbf{R}). \tag{10}$$

Note that Eq. (10) does not assume that either E_{W} or $\mathrm{I}_{\mathrm{map}}$ are independent of temperature, since quite generally $(\partial E_{\mathrm{W}}(\mathbf{R})/\partial T)_{\mathbf{R}} = -k_B T\left(\partial \mathrm{I}_{\mathrm{map}}(\mathbf{R})/\partial T\right)_{\mathbf{R}}$.[61] (See SM.)

We define

$$\mathrm{I}_{\mathrm{map}} = \int_{\mathcal{D}^N(V)} \mathrm{d}\mathbf{R}\, p_{\mathrm{R}}(\mathbf{R}) \mathrm{I}_{\mathrm{map}}(\mathbf{R}). \tag{11}$$

as the average of $\mathrm{I}_{\mathrm{map}}(\mathbf{R})$ over the mapped ensemble. Importantly, the KL divergence satisfies a chain rule[69]

$$\mathrm{I}_{\mathrm{AA}} = \mathrm{I}_{\mathrm{CG}} + \mathrm{I}_{\mathrm{map}} \tag{12}$$

that partitions the configurational information of the AA model between the mapped ensemble and the "lost" subensembles of AA configurations that map to each CG configuration.[28,61] Thus, the configurational information that is eliminated by the mapping is stored in conditioned distributions for the lost subensembles. In particular, the lost subensembles become more informative as the mapped ensemble becomes less informative.

We note that our notation slightly departs from prior studies. In particular, we have previously defined $S_W(\mathbf{R}) = -k_B \mathrm{I}_{\mathrm{map}}(\mathbf{R})$.[61] Similarly, $\mathrm{I}_{\mathrm{map}}$ has been previously denoted $\mathrm{H}_{\mathrm{map}}$ or $-\mathrm{S}_{\mathrm{map}}$.[55,59,60]

B. Mapping AA to CG configurations

We now develop a formalism for analyzing the CG mapping. It is often convenient to represent the atomic configuration with an $n \times D$ matrix, $\mathbf{r} = [r_{i\alpha}]$, where $r_{i\alpha}$ is the α Cartesian coordinate of atom i. Column α of this $n \times D$ matrix corresponds to the Cartesian direction that is specified by $\mathbf{e}_\alpha$. Row i of this matrix corresponds to the Cartesian coordinates of atom i: $\mathbf{r}_i = \sum_{\alpha=1}^{D} r_{i\alpha}\mathbf{e}_\alpha$. For each atom, $i = 1, \ldots, n$, we now introduce a "label" vector, $\mathbf{e}_i$, that identifies the corresponding row of the configuration matrix: $\mathbf{e}_i^\dagger \mathbf{r} = \mathbf{r}_i$. The n label vectors $\{\mathbf{e}_i\}_{i=1,\ldots,n}$ form a complete orthonormal basis for an n dimensional "AA label space," $\mathcal{V}_{\mathrm{AA}} = \mathrm{span}\{\mathbf{e}_i\}_{i=1,\ldots,n}$, with $\mathbf{e}_i^\dagger \mathbf{e}_j = \delta_{ij}$ and $\mathbb{1}_n = \sum_{i=1}^n \mathbf{e}_i \mathbf{e}_i^\dagger$. The AA configuration may then be denoted:

$$\mathbf{r} = \sum_{i=1}^{n} \mathbf{e}_i \otimes \mathbf{r}_i = \sum_{i\alpha=1}^{nD} r_{i\alpha} \mathbf{e}_{i\alpha}. \tag{13}$$

In the second expression, we have defined the set of nD orthonormal basis vectors $\{\mathbf{e}_{i\alpha} \equiv \mathbf{e}_i \otimes \mathbf{e}_\alpha\}$ such that $\mathbf{e}_{i\alpha}^\dagger \mathbf{r} = r_{i\alpha}$.

Similarly, it is often convenient to represent the CG configuration as an $N \times D$ matrix, $\mathbf{R} = [R_{I\alpha}]$. Accordingly, for each CG site $I = 1, \ldots, N$, we define a label vector, $\mathbf{e}_I$, and a corresponding N-dimensional "CG label space:" $\mathcal{V}_{\mathrm{CG}} = \mathrm{span}\{\mathbf{e}_I\}_{I=1,\ldots,N}$ with $\mathbf{e}_I^\dagger \mathbf{e}_J = \delta_{IJ}$ and $\mathbb{1}_N = \sum_{I=1}^N \mathbf{e}_I \mathbf{e}_I^\dagger$. Thus, we may express the CG configuration in analogy to Eq. (13)

$$\mathbf{R} = \sum_{I=1}^{N} \mathbf{e}_I \otimes \mathbf{R}_I = \sum_{I\alpha=1}^{ND} R_{I\alpha} \mathbf{e}_{I\alpha}. \tag{14}$$

In the first expression, $\mathbf{R}_I = \mathbf{e}_I^\dagger \mathbf{R} = \sum_{\alpha=1}^D R_{I\alpha} \mathbf{e}_\alpha$ specifies the Cartesian coordinates of site I. In the second expression, we have defined the set of ND orthonormal basis vectors $\{\mathbf{e}_{I\alpha} \equiv \mathbf{e}_I \otimes \mathbf{e}_\alpha\}$ such that $\mathbf{e}_{I\alpha}^\dagger \mathbf{R} = R_{I\alpha}$. Note that $i = 1, \ldots, n$ indicate AA labels, $I = 1, \ldots, N$ indicate CG labels, and $\alpha = 1, \ldots, D$ indicate Cartesian directions.

As in most particle-based CG models, we consider linear mappings $\mathbf{M} : \mathbf{r} \to \mathbf{R} = \mathbf{M}\mathbf{r}$

$$\mathbf{R}_I = \sum_{i=1}^{n} c_{Ii} \mathbf{r}_i, \tag{15}$$

where $c_{Ii} \geq 0$ for all $I = 1, \ldots, N$ and $i = 1, \ldots, n$. Note that the mapping coefficients do not depend upon α and act equivalently on each Cartesian direction. Consequently, $\mathbf{M}$ may be considered a transformation from AA label space to CG label space, $\mathbf{M} : \mathcal{V}_{\mathrm{AA}} \to \mathcal{V}_{\mathrm{CG}}$,

$$\mathbf{M} = \sum_{I=1}^{N} \sum_{i=1}^{n} \mathbf{e}_I c_{Ii} \mathbf{e}_i^\dagger = \sum_{I=1}^{N} \mathbf{e}_I \mathbf{c}_I^\dagger, \tag{16}$$

where we have defined a mapping vector, $\mathbf{c}_I = \sum_{i=1}^n c_{Ii} \mathbf{e}_i \in \mathcal{V}_{\mathrm{AA}}$, for each CG site, I. In the following, we shall not distinguish between $\mathbf{M}$ and its extension to the AA configuration space, $\tilde{\mathbf{M}} \equiv \mathbf{M} \otimes \mathbb{1}_D$, where $\mathbb{1}_D \equiv \sum_{\alpha=1}^D \mathbf{e}_\alpha \mathbf{e}_\alpha^\dagger$ is the identity operator for Cartesian space.

We impose several restrictions upon the CG mapping. In order to simply express these restrictions, we define $\mathbf{J}_n \equiv \sum_{i=1}^n \mathbf{e}_i \in \mathcal{V}_{\mathrm{AA}}$ and $\mathbf{J}_N = \sum_{I=1}^N \mathbf{e}_I \in \mathcal{V}_{\mathrm{CG}}$ as label vectors that act equivalently on each atom and site, respectively. In particular, we require that the N mapping vectors, $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$, are linearly independent such that each site moves independently of the others. Moreover, we require that the mapping coefficients for each site I are normalized according to

$$\mathbf{J}_n^\dagger \mathbf{c}_I = \sum_{i=1}^{n} c_{Ii} = 1. \tag{17}$$

This implies that $\mathbf{M}\mathbf{J}_n = \mathbf{J}_N$ and ensures that, for any $\mathbf{v} \in \mathbb{R}^D$,

$$\mathbf{M}\left(\mathbf{r} + \mathbf{J}_n \otimes \mathbf{v}\right) = \mathbf{M}\mathbf{r} + \mathbf{J}_N \otimes \mathbf{v}, \tag{18}$$

i.e., if we displace each atom by $\mathbf{v}$, then the mapping also displaces each site by $\mathbf{v}$.

Note that if the AA distribution is invariant with respect to uniform translation of all atoms, then Eq. (18) implies that the mapped ensemble will also preserve this symmetry. More generally, Appendix A demonstrates that the mapped ensemble will be invariant with respect to any symmetry, $\hat{T}$, that is present in the AA ensemble as long as $\hat{T}$ and $\mathbf{M}$ commute, i.e., $\hat{T}\mathbf{M} = \mathbf{M}\hat{T}$.

In the following, we shall consider maps that partition the n atoms into N disjoint subsets. More precisely, we define $V_{\mathrm{AA}} = \{1, \ldots, n\}$ as the set of atoms and $V_I = \{i | c_{Ii} > 0\}$ as the subset that contributes to site I. We require that $\cup_{I=1}^{N} V_I = V_{\mathrm{AA}}$ and that $V_I \cap V_J = \varnothing$ for all $I \neq J$. Note that this requirement excludes 'decimation' and 'slicing' maps that associate each CG site with a single atom.[47,55] We expect that it is straight-forward to relax this restriction. Finally, we shall also assume that the mapping associates each site with a single molecule, i.e., atoms in distinct molecules are not grouped together. This last assumption becomes necessary for developing simple approximations in Section III.

C. Backmapping and projection operators

For each site, $I = 1, \ldots, N$, we define a vector, $\mathbf{j}_I = \sum_{i \in V_I} \mathbf{e}_i \in \mathcal{V}_{\mathrm{AA}}$, that corresponds to uniformly displacing all the atoms that contribute to site I. Because we require the atomic groups, V_I, to be disjoint, it follows that $\mathbf{j}_I^\dagger \mathbf{j}_J = n_I \delta_{IJ}$ where $n_I = |V_I|$ is the number of atoms that contribute to site I. Moreover, Eq. (17) implies that

$$\mathbf{c}_I^\dagger \mathbf{j}_J = \delta_{IJ} \qquad \text{for all } I, J = 1, \ldots, N \tag{19}$$

When the mapping coefficients correspond to the center of geometry (cog) for the corresponding atomic group, then $\mathbf{c}_I$ and $\mathbf{j}_I$ are parallel: $\mathbf{c}_{I;\text{cog}} = n_I^{-1}\mathbf{j}_I$. More generally, there is no simple relationship between $\mathbf{c}_I$ and $\mathbf{j}_I$. Nevertheless, Eq. (19) holds for any disjoint mapping.

We now define a "backmapping" operator from the CG particle space back to the AA particle space:

$$\mathbf{B} \equiv \sum_{I=1}^{N} \mathbf{j}_I \mathbf{e}_I^\dagger, \tag{20}$$

which is a simple example of a right inverse for $\mathbf{M}$.[76] Because of Eq. (19), the combination $\mathbf{MB}$ acts as the identity operator in $\mathcal{V}_{\text{CG}}$: $\mathbf{MB} = \sum_{I=1}^{N} \mathbf{e}_I \mathbf{e}_I^\dagger = \mathbb{1}_N$. More importantly, Eq. (19) implies that the combination $\mathbf{BM}$ acts as an oblique projection operator[77] in $\mathcal{V}_{\text{AA}}$:

$$\mathbb{P} \equiv \mathbf{BM} = \sum_{I=1}^{N} \mathbf{j}_I \mathbf{c}_I^\dagger. \tag{21}$$

In contrast to projection operators that are familiar from quantum mechanics, $\mathbb{P}$ is not generally Hermitian, i.e., symmetric. Nevertheless, $\mathbb{P}$ is idempotent, $\mathbb{P}^2 = \mathbb{P}$, and projects arbitrary elements of AA particle space, $\mathbf{v} \in \mathcal{V}_{\text{AA}}$, onto a "CG" subspace that is spanned by $\{\mathbf{j}_1, \ldots, \mathbf{j}_N\}$. We define the complementary projection operator, $\mathbb{Q} \equiv \mathbb{1}_n - \mathbb{P}$, such that $\mathbb{P} + \mathbb{Q} = \mathbb{1}_n$, $\mathbb{PQ} = \mathbb{QP} = \mathbb{0}$, and $\mathbb{Q}^2 = \mathbb{Q}$.

While we have defined $\mathbb{P}$ as a projection operator acting in AA particle space, $\mathcal{V}_{\text{AA}}$, this also trivially defines a projection operator in the AA configuration space. For any AA displacement, $\delta\mathbf{r} = \sum_{i=1}^{n} \mathbf{e}_i \otimes \delta\mathbf{r}_i$, $\mathbb{P}$ defines corresponding displacements in the CG subspace of the AA configuration space. Specifically, each term, $\mathbf{j}_I \mathbf{c}_I^\dagger$, in Eq. (21) determines a displacement $\delta\mathbf{R}_I = \sum_{i=1}^{n} c_{Ii} \delta\mathbf{r}_i$ for CG site I and then moves each atom associated with site I by $\delta\mathbf{R}_I$:

$$\mathbb{P}\delta\mathbf{r} = \sum_{I=1}^{N} \mathbf{j}_I \otimes \delta\mathbf{R}_I. \tag{22}$$

We now introduce dual bases for $\mathcal{V}_{\text{AA}}$ in order to obtain an explicit expression for $\mathbb{Q}$. Accordingly, we let $\{\mathbf{x}_{N+k}\} \equiv \{\mathbf{x}_{N+1}, \ldots, \mathbf{x}_n\}$ be a basis for null($\mathbb{P}$) such that $\mathbf{c}_I^\dagger \mathbf{x}_{N+k} = 0$ for all $I = 1, \ldots, N$ and $k = 1, \ldots, n_{\text{x}}$. The rank-nullity theorem implies that $\{\mathbf{x}_i\} = \{\mathbf{j}_I, \mathbf{x}_{N+k}\} = \{\mathbf{j}_1, \ldots, \mathbf{j}_N, \mathbf{x}_{N+1}, \ldots, \mathbf{x}_n\}$ forms a basis for $\mathcal{V}_{\text{AA}}$.[77] We define a corresponding $n \times n$ matrix $\mathbf{X} = [\mathbf{j}_I | \mathbf{x}_{N+k}] = [\mathbf{X}_{\text{CG}} | \mathbf{X}_{\text{AA}}]$ where $\mathbf{X}_{\text{CG}} = [\mathbf{j}_1 \cdots \mathbf{j}_N]$ is an $n \times N$ matrix and $\mathbf{X}_{\text{AA}} = [\mathbf{x}_{N+1} \cdots \mathbf{x}_n]$ is an $n \times n_{\text{x}}$ matrix. We define $\mathbf{Z}^\dagger = \mathbf{X}^{-1}$. Since $\mathbf{c}_I^\dagger \mathbf{j}_J = \delta_{IJ}$

and $\mathbf{c}_I^\dagger \mathbf{x}_{N+k} = 0$ for all I and k, it follows that $\mathbf{Z} = [\mathbf{Z}_{\mathrm{CG}}|\mathbf{Z}_{\mathrm{AA}}]$ where $\mathbf{Z}_{\mathrm{CG}} = [\mathbf{c}_1 \cdots \mathbf{c}_N]$ and $\mathbf{Z}_{\mathrm{AA}} = [\mathbf{z}_{N+1} \cdots \mathbf{z}_n]$ such that $\mathbf{Z}_{\mathrm{AA}}^\dagger \mathbf{X}_{\mathrm{CG}} = \mathbf{0}$ and $\mathbf{Z}_{\mathrm{AA}}^\dagger \mathbf{X}_{\mathrm{AA}} = \mathbb{1}_{n_\mathrm{x}}$. We shall find it convenient to assume that the set of $\{\mathbf{z}_{N+k}\}$ are orthonormal with respect to each other such that $\mathbf{z}_{N+k}^\dagger \mathbf{z}_{N+k'} = \delta_{k,k'}$ for all $k, k' = 1, \ldots, n_\mathrm{x}$. This is always possible, e.g., by applying the Gram-Schmidt procedure[77] to $\{\mathbf{z}_{N+k}\}$ and the inverse transformation to $\{\mathbf{x}_{N+k}\}$. The resulting set of n vectors $\{\mathbf{z}_i\} = \{\mathbf{c}_I, \mathbf{z}_{N+k}\}$ form a dual basis with $\{\mathbf{x}_i\} = \{\mathbf{j}_I, \mathbf{x}_{N+k}\}$ such that $\sum_{i=1}^n \mathbf{x}_i \mathbf{z}_i^\dagger = \mathbb{1}_n$ is the identity operator for $\mathcal{V}_{\mathrm{AA}}$ and $\mathbf{z}_i^\dagger \mathbf{x}_j = \delta_{ij}$ for all $i, j = 1, \ldots n$. Finally, it follows that $\mathbb{P} = \mathbf{X}_{\mathrm{CG}} \mathbf{Z}_{\mathrm{CG}}^\dagger$ and the complementary projection operator may be expressed

$$\mathbb{Q} = \mathbf{X}_{\mathrm{AA}} \mathbf{Z}_{\mathrm{AA}}^\dagger = \sum_{k=1}^{n_\mathrm{x}} \mathbf{x}_{N+k} \mathbf{z}_{N+k}^\dagger. \tag{23}$$

The SM explicitly illustrates this dual basis for both label space and configuration space.

D. The labelling entropy

We now employ the n linearly independent vectors, $\{\mathbf{z}_i\} = \{\mathbf{c}_I, \mathbf{z}_{N+k}\}$, to define

$$\bar{\mathbf{r}}_I \equiv \mathbf{c}_I^\dagger \mathbf{r} \qquad \text{for all } I = 1, \ldots, N \tag{24}$$

$$\hat{\mathbf{r}}_k \equiv \mathbf{z}_{N+k}^\dagger \mathbf{r} \qquad \text{for all } k = 1, \ldots, n_\mathrm{x} \tag{25}$$

such that

$$\mathbf{r} = (\mathbb{P} + \mathbb{Q})\, \mathbf{r} = \mathbf{B}\bar{\mathbf{r}} + \mathbf{X}_{\mathrm{AA}} \hat{\mathbf{r}} \tag{26}$$

where $\bar{\mathbf{r}} = \mathbf{M}\mathbf{r}$ and $\hat{\mathbf{r}} = \mathbf{Z}_{\mathrm{AA}}^\dagger \mathbf{r}$. Since the mapping coefficients are normalized according to Eq. (17) it follows that the mapped coordinates, $\bar{\mathbf{r}}_I \in \mathcal{D}(V)$. Moreover, since the CG model explicitly represents each molecule and each site is associated with a single molecule, in the next section we shall interpret the $\hat{\mathbf{r}}_k$ coordinates as intrasite coordinates.

By construction there exists a 1-1 relationship between the n atomic coordinates $\mathbf{r}$ and the set of n coordinates $\tilde{\mathbf{r}} = (\bar{\mathbf{r}}, \hat{\mathbf{r}})$: $\tilde{\mathbf{r}} = \mathbf{Z}^\dagger \mathbf{r}$ and $\mathbf{r} = \mathbf{X}\tilde{\mathbf{r}}$. However, this transformation is not volume-preserving. In particular, Eq. (17) implies that $|\mathbf{c}_I| \equiv \sqrt{\sum_{i=1}^n c_{Ii}^2} < 1$ whenever site I is associated with more than one atom. Appendix B proves that the Jacobian associated with this transformation is

$$\left|\left|\frac{\partial \tilde{\mathbf{r}}}{\partial \mathbf{r}}\right|\right| = \left|\left|\mathbf{Z}^\dagger\right|\right|^D = \left|\left|\Delta_N\right|\right|^{-D/2} \tag{27}$$

where we have defined a diagonal "participation" matrix[40]

$$\Delta_N \equiv \sum_{I=1}^{N} \mathbf{e}_I n_I \mathbf{e}_I^{\dagger}. \tag{28}$$

We can now obtain a relatively simple expression for z_{R}. We define $\tilde{u}(\bar{\mathbf{r}}, \hat{\mathbf{r}}) = u(\mathbf{r} = \mathbf{B}\bar{\mathbf{r}} + \mathbf{X}_{\mathrm{AA}}\hat{\mathbf{r}})$. It then follows that

$$z_{\mathrm{R}}(\mathbf{R}) = \int_{\mathcal{D}^N(V)} \mathrm{d}\bar{\mathbf{r}} \int_{\hat{\mathcal{D}}_{n_{\mathrm{x}}}(V;\bar{\mathbf{r}})} \mathrm{d}\hat{\mathbf{r}} \;\; ||\Delta_N||^{D/2} \exp[-\beta \tilde{u}(\bar{\mathbf{r}}, \hat{\mathbf{r}})]\, \delta(\bar{\mathbf{r}} - \mathbf{R})\,, \tag{29}$$

where the second integral is over

$$\hat{\mathcal{D}}_{n_{\mathrm{x}}}(V; \bar{\mathbf{r}}) \equiv \{\hat{\mathbf{r}} \in \mathbb{R}^{n_{\mathrm{x}} \times D} | \mathbf{B}\bar{\mathbf{r}} + \mathbf{X}_{\mathrm{AA}}\hat{\mathbf{r}} \in \mathcal{D}^n(V)\}, \tag{30}$$

i.e., the set of atomic displacements, $\hat{\mathbf{r}}$, such that, $\mathbf{r}(\bar{\mathbf{r}}, \hat{\mathbf{r}}) \equiv \mathbf{B}\bar{\mathbf{r}} + \mathbf{X}_{\mathrm{AA}}\hat{\mathbf{r}}$, is in the AA configuration space. The first integral may be trivially evaluated for all $\mathbf{R} \in \mathcal{D}^N(V)$ to obtain

$$z_{\mathrm{R}}(\mathbf{R}) \;=\; ||\Delta_N||^{D/2}\, \hat{z}_{\mathrm{R}}(\mathbf{R}) \tag{31}$$

$$\hat{z}_{\mathrm{R}}(\mathbf{R}) \;\equiv\; \int_{\hat{\mathcal{D}}_{n_{\mathrm{x}}}(V;\mathbf{R})} \mathrm{d}\hat{\mathbf{r}} \;\; \exp[-\beta \tilde{u}(\mathbf{R}, \hat{\mathbf{r}})] \tag{32}$$

Because the transformation $\mathbf{r} \leftrightarrow \tilde{\mathbf{r}}$ is 1-1, the factor, $\hat{z}_{\mathrm{R}}(\mathbf{R})$ gives the total Boltzmann weight for all the AA configurations that map to $\mathbf{R}$.

Equations (5) and (31) imply that the PMF may be decomposed

$$W(\mathbf{R}) = \hat{W}(\mathbf{R}) - k_B T \mathrm{H}_{\mathrm{L}} \tag{33}$$

where

$$\hat{W}(\mathbf{R}) = -k_B T \ln\left[V^{-n_{\mathrm{x}}} \hat{z}_{\mathrm{R}}(\mathbf{R})\right] \tag{34}$$

and we have defined a "labelling entropy"

$$\mathrm{H}_{\mathrm{L}} \equiv \frac{1}{2} D \ln ||\Delta_N|| = \frac{1}{2} D \sum_{I=1}^{N} \ln n_I \geq 0, \tag{35}$$

which corresponds to the degeneracy of atoms associated with the CG sites. According to Eq. (8), the information present in the lost subensemble $\mathcal{S}_{\mathbf{R}}$ may be expressed:

$$\mathrm{I}_{\mathrm{map}}(\mathbf{R}) = \hat{\mathrm{I}}_{\mathrm{map}}(\mathbf{R}) - \mathrm{H}_{\mathrm{L}}, \tag{36}$$

where

$$\hat{\mathrm{I}}_{\mathrm{map}}(\mathbf{R}) = \int_{\hat{\mathcal{D}}_{n_\mathrm{x}}(V;\mathbf{R})} \mathrm{d}\hat{\mathbf{r}}\ p_{\hat{\mathrm{r}}|\mathrm{R}}(\hat{\mathbf{r}}|\mathbf{R}) \ln\left[V^{n_\mathrm{x}} p_{\hat{\mathrm{r}}|\mathrm{R}}(\hat{\mathbf{r}}|\mathbf{R})\right] \tag{37}$$

and we have defined

$$p_{\hat{\mathrm{r}}|\mathrm{R}}(\hat{\mathbf{r}}|\mathbf{R}) = \exp[-\beta\tilde{u}(\mathbf{R},\hat{\mathbf{r}})]/\hat{z}_\mathrm{R}(\mathbf{R}) \tag{38}$$

such that for any function $f(\mathbf{r},\mathbf{R})$,

$$\int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r}\ p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) f(\mathbf{r},\mathbf{R}) = \int_{\hat{\mathcal{D}}_{n_\mathrm{x}}(V;\mathbf{R})} \mathrm{d}\hat{\mathbf{r}}\ p_{\hat{\mathrm{r}}|\mathrm{R}}(\hat{\mathbf{r}}|\mathbf{R})\tilde{f}(\hat{\mathbf{r}},\mathbf{R}), \tag{39}$$

where $\tilde{f}(\hat{\mathbf{r}},\mathbf{R}) = f(\mathbf{r}(\mathbf{R},\hat{\mathbf{r}}),\mathbf{R})$.

Equation (36) decomposes the information lost in CG configuration $\mathbf{R}$ into two contributions. The first contribution, $\hat{\mathrm{I}}_{\mathrm{map}}(\mathbf{R})$, reflects the distribution, $p_{\hat{\mathrm{r}}|\mathrm{R}}(\hat{\mathbf{r}}|\mathbf{R})$, of internal displacements, $\hat{\mathbf{r}}$. However, the second contribution reflects the uncertainty associated with the partitioning of atoms between CG sites. Since this uncertainty reduces $\mathrm{I}_{\mathrm{map}}$, it effectively increases the configurational information present in the mapped ensemble.

The labelling entropy attains its global minimum $\mathrm{H}_{\mathrm{L;min}} = 0$ for decimation maps in which each site corresponds to a single atom, i.e., $n_I = 1$ for all $I = 1, \ldots, N$.[32,47] In the case that $N < n$ and each atom contributes to a single site, the minimum value of the labelling entropy is $\frac{1}{2}D\ln(n_\mathrm{x} - 1)$. For a fixed number of atoms, n, and CG sites, $N < n$, the labelling entropy increases as the partitioning of atoms between CG sites becomes increasingly uniform. The labelling entropy achieves its maximum $\mathrm{H}_{\mathrm{L;max}} = \frac{1}{2}DN\ln(n/N)$ when each site is associated with an equal number of atoms, $n_I = n/N$. If we quantify the resolution of the CG model by $r = N/n \in [0,1]$, then $\mathrm{H}_{\mathrm{L;max}}(r) = -\frac{1}{2}Dnr\ln r$, which attains its maximum at the resolution $r_* = e^{-1} \approx 0.37$, i.e., when the CG model preserves approximately 37% of the AA degrees of freedom.

III. SIMPLE APPROXIMATIONS AND MODELS

A. Local harmonic approximation

To this point our treatment has been exact. We now consider a very simple local harmonic approximation for Eq. (32). Since we have required that the CG model explicitly represents each molecule and since $\mathbf{X}_{\mathrm{AA}}\hat{\mathbf{r}}$ describes intramolecular displacements about the back-mapped configuration, $\mathbf{BR}$, we expect that $\tilde{u}(\mathbf{R}, \hat{\mathbf{r}}) \to \infty$ for large intra-site displacements. Accordingly, for each CG configuration, $\mathbf{R}$, we define $u_0(\mathbf{R})$ as the minimum of the AA potential within the subensemble, $\mathcal{S}_{\mathbf{R}}$, of AA configurations that map to $\mathbf{R}$:

$$u_0(\mathbf{R}) = \min_{\mathbf{r} \in \mathcal{S}_{\mathbf{R}}} u(\mathbf{r}) = \min_{\hat{\mathbf{r}} \in \hat{\mathcal{D}}_{n_{\mathrm{x}}}(V;\mathbf{R})} \tilde{u}(\mathbf{R}, \hat{\mathbf{r}}). \tag{40}$$

For simplicity, we assume that this minimum corresponds to a unique AA configuration, $\mathbf{r}_{\mathbf{R}}$, and define the corresponding intra-site displacements, $\hat{\mathbf{r}}_{\mathbf{R}} = \mathbf{Z}^{\dagger}_{\mathrm{AA}} \mathbf{r}_{\mathbf{R}}$, such that $\mathbf{r}_{\mathbf{R}} = \mathbf{BR} + \mathbf{X}_{\mathrm{AA}} \hat{\mathbf{r}}_{\mathbf{R}}$. We expand the AA potential quadratically about this minimum:

$$\tilde{u}(\mathbf{R}, \hat{\mathbf{r}}) \approx u_0(\mathbf{R}) + \frac{1}{2} \delta\hat{\mathbf{r}}^{\dagger} \hat{\mathbf{h}}_{\mathrm{AA}} \delta\hat{\mathbf{r}}, \tag{41}$$

where $\delta\hat{\mathbf{r}} = \hat{\mathbf{r}} - \hat{\mathbf{r}}_{\mathbf{R}}$ and $\hat{\mathbf{h}}_{\mathrm{AA}} \equiv \hat{\mathbf{h}}_{\mathrm{AA}}(\mathbf{R}) \equiv \mathbf{X}^{\dagger}_{\mathrm{AA}} \mathbf{h}(\mathbf{r}_{\mathbf{R}}) \mathbf{X}_{\mathrm{AA}}$ is the projection of the AA Hessian matrix, $\mathbf{h} \equiv \mathbf{h}(\mathbf{r}_{\mathbf{R}}) \equiv \partial^2 u / \partial \mathbf{r} \partial \mathbf{r}'|_{\mathbf{r}_{\mathbf{R}}}$, into the subspace of intra-site displacements. Since we have assumed that $\mathbf{r}_{\mathbf{R}}$ is the unique minimizer of $u(\mathbf{r})$ in $\mathcal{S}_{\mathbf{R}}$, we assume that $\hat{\mathbf{h}}_{\mathrm{AA}}$ is positive definite. Consequently, we can evaluate the resulting Gaussian integrals to obtain:

$$\hat{z}_{\mathrm{R}}(\mathbf{R}) \approx \sqrt{\frac{(2\pi)^{n_{\mathrm{x}} D}}{\left|\left|\beta \hat{\mathbf{h}}_{\mathrm{AA}}(\mathbf{R})\right|\right|}} \exp[-\beta u_0(\mathbf{R})]. \tag{42}$$

Note that the local harmonic approximation does not apply to implicit solvent CG models that eliminate entire molecules from the CG representation. In order to apply this approximation to implicit solvent models, $u(\mathbf{r})$ must be considered the free energy for the AA solute coordinates after the solvent molecules have already been integrated out.[78]

In this local harmonic approximation, the conditioned distribution of intra-site displacements within the lost subensemble, $\mathcal{S}_{\mathbf{R}}$, is simply Gaussian

$$p_{\hat{\mathrm{r}}|\mathrm{R}}(\hat{\mathbf{r}}|\mathbf{R}) \approx \sqrt{(2\pi)^{-n_{\mathrm{x}} D} \left|\left|\mathbf{C}_{\delta\hat{\mathbf{r}}}(\mathbf{R})\right|\right|^{-1}} \exp\left[-\frac{1}{2} \delta\hat{\mathbf{r}}^{\dagger} \mathbf{C}^{-1}_{\delta\hat{\mathbf{r}}}(\mathbf{R}) \delta\hat{\mathbf{r}}\right], \tag{43}$$

where

$$\mathbf{C}_{\delta\hat{\mathbf{r}}}(\mathbf{R}) \equiv \left\langle \delta\hat{\mathbf{r}} \delta\hat{\mathbf{r}}^{\dagger} \right\rangle_{\mathbf{R}} = \left(\beta \hat{\mathbf{h}}_{\mathrm{AA}}(\mathbf{R})\right)^{-1} \tag{44}$$

is the conditioned covariance matrix describing fluctuations in the vibrational intra-site degrees of freedom about the given CG configuration. The mapped distribution is

$$p_{\mathrm{R}}(\mathbf{R}) \approx z^{-1} \sqrt{(2\pi)^{n_{\mathrm{x}} D} \left|\left|\Delta_N\right|\right|^D \left|\left|\mathbf{C}_{\delta\hat{\mathbf{r}}}(\mathbf{R})\right|\right|} \exp[-\beta u_0(\mathbf{R})]. \tag{45}$$

As expected, the mapped probability density, $p_{\mathrm{R}}(\mathbf{R})$, is proportional to both the Boltzmann weight of the most probable configuration in the lost subensemble $\mathcal{S}_{\mathbf{R}}$, as well as to the magnitude of the AA fluctuations in $\mathcal{S}_{\mathbf{R}}$. Additionally, p_{R} is uniformly scaled by the Jacobian factor defining the labelling entropy, which does not relate to AA interactions, but is simply a consequence of how atoms are grouped into CG sites. In this local harmonic approximation, the PMF may be expressed according to Eq. (34) with an energetic component

$$E_{\mathrm{W}}(\mathbf{R}) \approx u_0(\mathbf{R}) + \frac{1}{2} n_{\mathrm{x}} D k_B T \tag{46}$$

that reflects both the temperature-independent minimizing energy, $u_0(\mathbf{R})$, and also the temperature-dependent average energy of the $n_{\mathrm{x}} D$ internal vibrations. This approximation also gives

$$\begin{aligned} \mathrm{I}_{\mathrm{map}}(\mathbf{R}) &\approx \frac{1}{2} \ln \left[\left(\frac{L^2}{2\pi} \right)^{n_{\mathrm{x}} D} ||\mathbf{C}_{\delta\hat{\mathbf{r}}}(\mathbf{R})||^{-1} \right] - \frac{1}{2} n_{\mathrm{x}} D - \mathrm{H}_{\mathrm{L}} & (47) \\ &= \frac{1}{2} \ln \left[\left(\frac{\beta L^2}{2\pi} \right)^{n_{\mathrm{x}} D} \frac{\left|\left| \hat{\mathbf{h}}_{\mathrm{AA}}(\mathbf{R}) \right|\right|}{||\Delta_N||^D} \right] - \frac{1}{2} n_{\mathrm{x}} D, & (48) \end{aligned}$$

where the volume is $V = L^D$. As expected $\mathrm{I}_{\mathrm{map}}(\mathbf{R})$ increases as the lost subensemble becomes increasingly constrained, i.e., as $\left|\left| \hat{\mathbf{h}}_{\mathrm{AA}}(\mathbf{R}) \right|\right|$ increases and $||\mathbf{C}_{\delta\hat{\mathbf{r}}}(\mathbf{R})||$ decreases. However, $\mathrm{I}_{\mathrm{map}}(\mathbf{R})$ is also reduced by the labelling entropy.

B. Harmonic model

We now specialize to harmonic AA potentials for which the preceding approximation is exact:

$$u_{\mathrm{harm}}(\mathbf{r}) = \frac{1}{2} \Delta\mathbf{r}^\dagger \mathbf{h} \Delta\mathbf{r} = \frac{1}{2} \sum_{i\alpha} \sum_{j\beta} \Delta r_{i\alpha} h_{i\alpha;j\beta} \Delta r_{j\beta}. \tag{49}$$

Here $\Delta\mathbf{r} = \mathbf{r} - \mathbf{r}^*$ describes the displacement from a reference configuration, $\mathbf{r}^*$, that minimizes the AA potential and $h_{i\alpha;j\beta} = \partial^2 u_{\mathrm{harm}} / \partial r_{i\alpha} \partial r_{j\beta}|_{\mathbf{r}^*}$ is the Hessian of u_{harm}. We assume the Hessian matrix, $\mathbf{h}$, is positive semi-definite with a nullspace, $\mathrm{null}(\mathbf{h}) = \mathrm{span}\{\boldsymbol{\eta}_\varphi\}$, that is associated with the uniform translation and rotation of all n atoms. This type of potential naturally arises, e.g., in normal mode analysis when approximating a nonlinear molecular mechanics potential about $\mathbf{r}^*$[79,80] or when defining an anisotropic network model[81] from the Tirion elastic network model.[82]

By adopting Eq. (26), the harmonic potential may be explicitly expressed

$$\tilde{u}_{\mathrm{harm}}(\overline{\mathbf{r}}, \hat{\mathbf{r}}) = \frac{1}{2} \left\{ \Delta\overline{\mathbf{r}}^\dagger \overline{\mathbf{h}}_{\mathrm{CG}} \Delta\overline{\mathbf{r}} + 2\Delta\overline{\mathbf{r}}^\dagger \mathbf{h}_{\mathrm{x}} \Delta\hat{\mathbf{r}} + \Delta\hat{\mathbf{r}}^\dagger \hat{\mathbf{h}}_{\mathrm{AA}} \Delta\hat{\mathbf{r}} \right\}. \tag{50}$$

Here $\Delta\overline{\mathbf{r}} = \mathbf{M}\Delta\mathbf{r}$, $\Delta\hat{\mathbf{r}} = \mathbf{Z}_{\mathrm{AA}}^{\dagger}\Delta\mathbf{r}$, and we have partitioned the AA Hessian into a CG component, $\overline{\mathbf{h}}_{\mathrm{CG}} = \mathbf{B}^{\dagger}\mathbf{h}\mathbf{B}$, an AA component, $\hat{\mathbf{h}}_{\mathrm{AA}} = \mathbf{X}_{\mathrm{AA}}^{\dagger}\mathbf{h}\mathbf{X}_{\mathrm{AA}}$, and a coupling component, $\mathbf{h}_{\mathrm{x}} = \mathbf{B}^{\dagger}\mathbf{h}\mathbf{X}_{\mathrm{AA}}$. We assume that the CG mapping preserves the translational and rotational symmetries of the AA potential, such that $\dim \mathrm{span}\{\mathbf{M}\boldsymbol{\eta}_{\varphi}\} = \dim \mathrm{null}(\mathbf{h})$. The SM demonstrates that, as a consequence, $\hat{\mathbf{h}}_{\mathrm{AA}}$ is positive definite and, thus, invertible.

Given a fixed CG configuration, $\mathbf{R}$, the AA potential, $\tilde{u}_{\mathrm{harm}}(\mathbf{R}, \hat{\mathbf{r}})$, is minimized by

$$\hat{\mathbf{r}}_{\mathbf{R}} = \hat{\mathbf{r}}^{*} - \hat{\mathbf{h}}_{\mathrm{AA}}^{-1}\mathbf{h}_{\mathrm{x}}^{\dagger}\Delta\mathbf{R}, \tag{51}$$

where $\Delta\mathbf{R} = \mathbf{R} - \mathbf{M}\mathbf{r}^{*}$. Because u_{harm} is bilinear in Cartesian coordinates, the local harmonic approximation is exact:

$$\tilde{u}_{\mathrm{harm}}(\mathbf{R}, \hat{\mathbf{r}}) = u_0(\mathbf{R}) + \frac{1}{2}\delta\hat{\mathbf{r}}^{\dagger}\hat{\mathbf{h}}_{\mathrm{AA}}\delta\hat{\mathbf{r}}, \tag{52}$$

where $\delta\hat{\mathbf{r}} = \hat{\mathbf{r}} - \hat{\mathbf{r}}_{\mathbf{R}}$, the minimizing AA potential is

$$u_0(\mathbf{R}) = \frac{1}{2}\Delta\mathbf{R}^{\dagger}\mathbf{H}\Delta\mathbf{R} \tag{53}$$

and the renormalized Hessian matrix is the Schur complement[83,84]

$$\mathbf{H} = \overline{\mathbf{h}}_{\mathrm{CG}} - \mathbf{h}_{\mathrm{x}}\hat{\mathbf{h}}_{\mathrm{AA}}^{-1}\mathbf{h}_{\mathrm{x}}^{\dagger}, \tag{54}$$

which is independent of $\mathbf{R}$. Equation (54) explicitly demonstrates how intra-site interactions impact the mapped ensemble and the CG potential through the coupling component, $\mathbf{h}_{\mathrm{x}}$. Moreover, Eq. (54) suggests that it may be possible to identify "resonant" maps that eliminate this coupling component, such that the atomic potential can be separated into independent CG and intra-site components. Section III D considers this possibility further.

Equation (54) corresponds to a previous result of Potestio and coworkers.[32] The SM explicitly demonstrates that Eq. (54) is also consistent with the generalization of our prior result[61] for the Gaussian Network model: $\mathbf{H}^{\mathrm{I}} = \mathbb{Q}_{\mathbf{H}}\mathbf{M}\mathbf{h}^{\mathrm{I}}\mathbf{M}^{\dagger}\mathbb{Q}_{\mathbf{H}}$, where $^{\mathrm{I}}$ denotes the Moore-Penrose pseudo-inverse and $\mathbb{Q}_{\mathbf{H}}$ is the projector orthogonal to the nullspace of $\mathbf{H}$.[77] Since $\hat{\mathbf{h}}_{\mathrm{AA}}$ is full rank, the null spaces of $\mathbf{h}$ and $\mathbf{H}$ have the same dimension.[84] The SM demonstrates that each distinct null-vector, $\boldsymbol{\eta}_{\varphi}$, of $\mathbf{h}$ maps onto a distinct null-vector $\overline{\boldsymbol{\eta}}_{\varphi} \equiv \mathbf{M}\boldsymbol{\eta}_{\varphi}$ of $\mathbf{H}$.

C. Gaussian Network Model

In order to numerically illustrate this framework, we further specialize to the Gaussian network model (GNM).[65,66,85] Here we briefly summarize the key aspects of coarse-graining the GNM. Ref. 63 provides a much more detailed presentation.

The high resolution GNM represents each residue in a protein with its α carbon. The GNM potential introduces a linear isotropic spring between each pair of residues that is in contact (i.e., within a given cut-off, r_{c}) in the equilibrium folded structure, $\mathbf{r}^*$. The resulting one-dimensional potential ($D = 1$) is given by Eq. (49) with $\mathbf{h} = \Gamma\boldsymbol{\kappa}$ where Γ is a dimensional factor with units of energy/length^2 and $\boldsymbol{\kappa}$ is

$$\kappa_{ij} = d_i\delta_{ij} - \theta_{ij}, \tag{55}$$

where $\theta_{ij} = 1$ if residues i and j contact in $\mathbf{r}^*$ and 0 otherwise, while $d_i = \sum_{j(\neq i)} \theta_{ij}$. The GNM corresponds to a graph describing the network of springs: d_i is the degree of residue i, θ_{ij} corresponds to the adjacency matrix, and κ_{ij} defines the Kirchhoff or Laplacian matrix for this graph.[86,87] The null space of $\boldsymbol{\kappa}$ is spanned by $\mathbf{J}_n$ and we define $\mathbb{Q}_{\boldsymbol{\kappa}} \equiv \mathbb{1}_n - \mathbf{J}_n n^{-1}\mathbf{J}_n^\dagger$ as the projector orthogonal to this null-space. The information content of the AA model may be expressed

$$\mathrm{I}_{\mathrm{AA}} = (n-1)h_1 + \frac{1}{2}\ln t_{\boldsymbol{\kappa}}. \tag{56}$$

Here $h_1 = \ln(L/L_{\mathrm{vib}}) - \frac{1}{2}$ may be interpreted as the information gained when replacing a free translational degree of freedom by a vibrational degree of freedom with a characteristic length-scale $L_{\mathrm{vib}} = \sqrt{2\pi k_B T/\Gamma}$. In the second term of Eq. (56) we have defined $t_{\boldsymbol{\kappa}} \equiv n^{-1}\det_1\boldsymbol{\kappa}$ where $\det_1\boldsymbol{\kappa}$ is the product of the $n-1$ positive eigenvalues of $\boldsymbol{\kappa}$. The Kirchhoff matrix-tree theorem states that $t_{\boldsymbol{\kappa}}$ is the number of spanning trees that are present in the AA GNM graph.[88,89] Additionally, we define $\widetilde{\boldsymbol{\kappa}} \equiv \mathbf{g}^{-1}\boldsymbol{\kappa}\mathbf{g}^{-1}$ in terms of the diagonal mass-weighting matrix $\mathbf{g} \equiv \sum_{i=1}^n \mathbf{e}_i m_i^{1/2}\mathbf{e}_i^\dagger$, where m_i is the mass of atom i. Finally, the mass-weighted vibrational covariance matrix of the AA model may be expressed

$$\mathbf{c}_{\mathrm{v}} = (\beta\Gamma\widetilde{\boldsymbol{\kappa}})^{\mathrm{I}}. \tag{57}$$

Given the CG mapping, $\mathbf{M}$, the renormalized Hessian matrix is $\mathbf{H} = \Gamma\mathbf{K}$ where $\mathbf{K}$ is positive semi-definite with a one-dimensional nullspace spanned by $\mathbf{J}_N = \mathbf{M}\mathbf{J}_n$ and we define the projector $\mathbb{Q}_{\mathbf{K}} = \mathbb{1}_N - \mathbf{J}_N N^{-1}\mathbf{J}_N^\dagger$ orthogonal to this nullspace. According to Eq. (54)

$$\mathbf{K} = \overline{\boldsymbol{\kappa}}_{\mathrm{CG}} - \boldsymbol{\kappa}_{\mathrm{x}}\hat{\boldsymbol{\kappa}}_{\mathrm{AA}}^{-1}\boldsymbol{\kappa}_{\mathrm{x}}^\dagger = \left(\mathbb{Q}_{\mathbf{K}}\mathbf{M}\boldsymbol{\kappa}^{\mathrm{I}}\mathbf{M}^\dagger\mathbb{Q}_{\mathbf{K}}\right)^{\mathrm{I}}. \tag{58}$$

The information content of the mapped ensemble is

$$\mathrm{I}_{\mathrm{CG}} = (N-1)h_1 + \frac{1}{2}\ln T_{\mathbf{K}}, \tag{59}$$

where $T_{\mathbf{K}} = N^{-1}\mathrm{det}_1\mathbf{K}$ and $\mathrm{det}_1\mathbf{K}$ is the product of the $N-1$ positive eigenvalues of $\mathbf{K}$. In analogy to the AA case, we define $\widetilde{\mathbf{K}} \equiv \mathbf{G}^{-1}\mathbf{K}\mathbf{G}^{-1}$ in terms of the mass-weighting matrix $\mathbf{G} \equiv \sum_{I=1}^{N} \mathbf{e}_I M_I^{1/2} \mathbf{e}_I^\dagger$ where M_I is the mass of site I. The mass-weighted covariance matrix for the mapped ensemble is then

$$\mathbf{C}_{\mathrm{v}} = \left(\beta\Gamma\widetilde{\mathbf{K}}\right)^{\mathrm{I}}. \tag{60}$$

The SM demonstrates that $t_{\boldsymbol{\kappa}}$ and $T_{\mathbf{K}}$ are related according to

$$t_{\boldsymbol{\kappa}}/T_{\mathbf{K}} = ||\hat{\boldsymbol{\kappa}}_{\mathrm{AA}}|| \,/\, ||\Delta_N||\,. \tag{61}$$

Consequently, the information loss due to the mapping may be expressed

$$\mathrm{I}_{\mathrm{map}} = n_{\mathrm{x}} h_1 + \frac{1}{2}\ln\left(||\hat{\boldsymbol{\kappa}}_{\mathrm{AA}}|| \,/\, ||\Delta_N||\right). \tag{62}$$

Since h_1 is independent of the CG mapping, Eq. (62) makes it particularly clear that the information lost by the CG mapping increases with the stiffness of the intra-site vibrations, $||\hat{\boldsymbol{\kappa}}_{\mathrm{AA}}||$, but is reduced by the labelling degeneracy, $||\Delta_N||$.

We define the spectral quality, $\mathcal{Q}$, to quantify the ability of the CG mapping to preserve the large scale motions of the AA model:

$$\mathcal{Q} \equiv \mathrm{Tr}_N\, \mathbf{C}_{\mathrm{v}} / \,\mathrm{Tr}_n\, \mathbf{c}_{\mathrm{v}} = \frac{\mathrm{Tr}_N\, \mathbb{Q}_{\widetilde{\mathbf{K}}} \mathbf{V}^\dagger \left(\mathbf{g}\boldsymbol{\kappa}^{\mathrm{I}}\mathbf{g}\right) \mathbf{V} \mathbb{Q}_{\widetilde{\mathbf{K}}}}{\mathrm{Tr}_n\, \mathbb{Q}_{\widetilde{\boldsymbol{\kappa}}} \left(\mathbf{g}\boldsymbol{\kappa}^{\mathrm{I}}\mathbf{g}\right) \mathbb{Q}_{\widetilde{\boldsymbol{\kappa}}}}. \tag{63}$$

In the second expression, we have defined $\mathbb{Q}_{\widetilde{\boldsymbol{\kappa}}} = \mathbb{1}_n - m_t^{-1}\mathbf{g}\mathbf{J}_n\mathbf{J}_n^\dagger\mathbf{g}$ and $\mathbb{Q}_{\widetilde{\mathbf{K}}} = \mathbb{1}_N - M_t^{-1}\mathbf{G}\mathbf{J}_N\mathbf{J}_N^\dagger\mathbf{G}$ as projection operators orthogonal to the nullspace of $\widetilde{\boldsymbol{\kappa}}$ and $\widetilde{\mathbf{K}}$, respectively, where $m_t = \sum_{i=1}^{n} m_i$ and $M_t = \sum_{I=1}^{N} M_I$ are the total mass of the AA and CG models, respectively. Thus, $\mathcal{Q}$ appears very similar to a Rayleigh quotient for aligning $\mathbf{V} \equiv \mathbf{g}^{-1}\mathbf{M}^\dagger\mathbf{G}$ with the subspace corresponding to the largest eigenvalues of $\mathbf{g}\boldsymbol{\kappa}^{\mathrm{I}}\mathbf{g}$, while accounting for the zero eigenvalue associated with free translational motion. The spectral quality appears qualitatively similar to the scoring function employed in the variational approach for Markov processes.[47,90–92] Appendix C compares the spectral quality with the ED-CG metric.[48]

D. Resonance between AA and CG models

Subsection III B suggested the possibility of resonant mappings that perfectly eliminate the coupling between the CG and intra-site degrees of freedom in an underlying atomic model. In this subsection, we construct an atomic GNM that allows for such a resonance. We first specify the special CG mapping, $\mathbf{M}_* = \sum_{I=1}^N \mathbf{e}_I \mathbf{c}_I^\dagger$, and construct a CG network model that perfectly aligns with $\mathbf{M}_*$. We then "atomize" this CG model such that the mapping, $\mathbf{M}_*$, is resonant with the resulting atomic potential.

As before, we assume that the specified CG mapping, $\mathbf{M}_*$, partitions the n atoms into N disjoint subsets, $V_I = \{i | c_{Ii} > 0\}$, that are associated with each site. The mapping also determines a corresponding back-mapping, $\mathbf{B}_* = \sum_{I=1}^N \mathbf{j}_I \mathbf{e}_I^\dagger$, and a corresponding participation matrix, $\Delta_N = \sum_{I=1}^N \mathbf{e}_I n_I \mathbf{e}_I^\dagger$, where $n_I = |V_I|$ is the number of atoms that map to site I.

We construct the CG model by first constructing a simple, connected CG graph, $G_{\text{CG}} = (V_{\text{CG}}, E_{\text{CG}})$. This graph represents each site with a vertex, $V_{\text{CG}} = \{1, \dots, N\}$, and introduces edges, $e_{IJ} \in E_{\text{CG}}$, between the sites. This graph determines a CG adjacency matrix, $\Theta = \sum_{I,J=1}^N \mathbf{e}_I \Theta_{IJ} \mathbf{e}_J^\dagger$, where $\Theta_{IJ} = 1$ for distinct sites I and J that are connected ($e_{IJ} \in E_{\text{CG}}$); otherwise $\Theta_{IJ} = 0$. We now weight each edge according to the number of atoms associated with the corresponding sites, $w(e_{IJ}) = n_I \Theta_{IJ} n_J$. This determines a weighted adjacency matrix:

$$\mathbf{A} \equiv \Delta_N \mathbf{\Theta} \Delta_N = \sum_{I,J=1}^N \mathbf{e}_I A_{IJ} \mathbf{e}_J^\dagger, \tag{64}$$

where $A_{IJ} = w(e_{IJ})$. Similarly, we define a corresponding weighted degree matrix:

$$\mathbf{D} = \sum_{I=1}^N \mathbf{e}_I D_I \mathbf{e}_I^\dagger \tag{65}$$

where $D_I = n_I N_I$ gives the weighted degree of vertex I in terms of $N_I \equiv \sum_{J=1}^N \Theta_{IJ} n_J$, which is the number of atoms that are associated with sites J that connect to site I. The Laplacian matrix for the weighted graph is then

$$\mathbf{K}_* \equiv \mathbf{D} - \mathbf{A}, \tag{66}$$

which is semi-positive definite with a 1-dimensional null space that is spanned by $\mathbf{J}_N$. The generalized Kirchhoff-matrix tree theorem states that $T_{\mathbf{K}_*} = N^{-1}\text{det}_1 \mathbf{K}_*$ is the sum of weights for all the spanning trees in the weighted CG graph.[88,89] We define a GNM-like potential associated with the weighted CG graph:

$$W_*(\mathbf{R}) = \frac{1}{2}\Gamma \delta\mathbf{R}^\dagger \mathbf{K}_* \delta\mathbf{R} + \text{const} \tag{67}$$

where $\delta\mathbf{R} = \mathbf{R} - \mathbf{R}^*$ for an arbitrary reference CG configuration, $\mathbf{R}^* = \mathbf{M}\mathbf{r}^*$, and const is configuration-independent constant. The resulting CG distribution is $P_{\mathrm{R}}(\mathbf{R}) \propto \exp\left[-\frac{1}{2}\beta\Gamma\delta\mathbf{R}^\dagger\mathbf{K}_*\delta\mathbf{R}\right]$.

We now atomize the CG potential, W_*. We first transform $\mathbf{K}_*$ from the CG configuration space into the AA configuration space:

$$\boldsymbol{k}_* \equiv \mathbf{M}_*^\dagger\mathbf{K}_*\mathbf{M}_* = \sum_{I,J=1}^{N} \mathbf{c}_I \left(D_I\delta_{IJ} - A_{IJ}\right)\mathbf{c}_J^\dagger, \tag{68}$$

such that $\delta\mathbf{r}^\dagger\boldsymbol{k}_*\delta\mathbf{r} = \delta\overline{\mathbf{r}}^\dagger\mathbf{K}_*\delta\overline{\mathbf{r}}$. Note that $\boldsymbol{k}_*$ accounts for atomic interactions between CG sites but not for interactions within sites. Consequently, we define

$$\delta\boldsymbol{\kappa} \equiv \mathbf{Z}_{\mathrm{AA}}\hat{\boldsymbol{\kappa}}_*\mathbf{Z}_{\mathrm{AA}}^\dagger, \tag{69}$$

where $\hat{\boldsymbol{\kappa}}_*$ is an arbitrary $n_{\mathrm{x}} \times n_{\mathrm{x}}$ matrix describing intra-site interactions and $\mathbf{Z}_{\mathrm{AA}} = [\mathbf{z}_{N+1} \cdots \mathbf{z}_n]$ is the $n \times n_{\mathrm{x}}$ matrix defined in Section II C. We define an AA spring matrix

$$\boldsymbol{\kappa} \equiv \boldsymbol{k}_* + \delta\boldsymbol{\kappa}, \tag{70}$$

and a corresponding AA potential

$$u(\mathbf{r}) = \frac{1}{2}\Gamma\delta\mathbf{r}^\dagger\boldsymbol{\kappa}\delta\mathbf{r}, \tag{71}$$

where $\delta\mathbf{r} = \mathbf{r} - \mathbf{r}^*$. By construction, $\overline{\boldsymbol{\kappa}}_{\mathrm{CG}} \equiv \mathbf{B}_*^\dagger\boldsymbol{\kappa}\mathbf{B}_* = \mathbf{K}_*$ and $\boldsymbol{\kappa}_{\mathrm{x}} \equiv \mathbf{B}_*^\dagger\boldsymbol{\kappa}\mathbf{X}_{\mathrm{AA}} = \mathbf{0}$. Consequently, Eq. (71) can be exactly decomposed into independent contributions from CG and intra-site degrees of freedom. Moreover, every AA potential of the form given by Eqs. (68) – (71) corresponds to the CG potential given by Eq. (67). Equivalently, every such AA model gives rise to the same mapped ensemble, $p_{\mathrm{R}}(\mathbf{R}) = P_{\mathrm{R}}(\mathbf{R})$. Thus, we see explicitly that the information lost from this AA model due to coarse-graining corresponds to the intra-site spring matrix, $\hat{\boldsymbol{\kappa}}_*$. However, Eq. (71) does not necessarily correspond to an atomic GNM.

We can gain additional insight by specializing to geometric-center mappings, $\mathbf{c}_I \equiv n_I^{-1}\mathbf{j}_I$. In this case,

$$\boldsymbol{k}_* = \sum_{I=1}^{N} N_I\mathbb{P}_I - \sum_{I,J=1}^{N} \mathbf{j}_I\Theta_{IJ}\mathbf{j}_J^\dagger, \tag{72}$$

where $\mathbb{P}_I = \mathbf{j}_I n_I^{-1}\mathbf{j}_I^\dagger$ is a projection operator describing the coarse-grained motion of the atomic group associated with site I. Moreover, in this case we can explicitly construct an atomic network, G_{AA}, that is consistent with the CG network, G_{CG}.

For each site, I, we construct a simple connected intra-site graph, $G_I = (V_I, E_I)$, by introducing edges $e_{ij} \in E_I$ between the atoms $i, j \in V_I$ that are associated with the site. For each pair of atoms associated with the site, we set $\theta_{ij} = 1$ if $e_{ij} \in E_I$; otherwise $\theta_{ij} = 0$. For each atom $i \in V_I$, we define $d_{Ii} = \sum_{j \in V_I} \theta_{ij}$ as the number of intrasite edges to atom i. The Laplacian matrix for the intra-site graph, G_I, is

$$\boldsymbol{\kappa}_{I;\mathrm{in}} = \sum_{i,j \in V_I} \mathbf{e}_i (d_{Ii}\delta_{ij} - \theta_{ij}) \mathbf{e}_j^\dagger. \tag{73}$$

We form the atomic graph, G_{AA}, by connecting the intra-site graphs, G_I, according to the original CG network, G_{CG}. Specifically, for each distinct pair of sites, $I \neq J$, we require that all of the associated atoms, $i \in V_I$ and $j \in V_J$, are either connected or not connected according to Θ_{IJ}: $\theta_{ij} = \Theta_{IJ}$. The Laplacian for G_{AA} is $\boldsymbol{\kappa} = \mathbf{d} - \boldsymbol{\theta}$ where $\boldsymbol{\theta} = \sum_{i,j=1}^n \mathbf{e}_i \theta_{ij} \mathbf{e}_j^\dagger$ is the adjacency matrix and the degree matrix is

$$\mathbf{d} = \sum_{I=1}^N \sum_{i \in V_I} \mathbf{e}_i \left(d_{Ii} + N_I\right) \mathbf{e}_i^\dagger. \tag{74}$$

It then follows that $\boldsymbol{\kappa}$ is given by Eq. (70) with

$$\delta\boldsymbol{\kappa} \; = \; \sum_{I=1}^N \delta\boldsymbol{\kappa}_I \tag{75}$$

$$\delta\boldsymbol{\kappa}_I \; = \; \boldsymbol{\kappa}_{I;\mathrm{in}} + N_I \mathbb{Q}_I \tag{76}$$

and $\mathbb{Q}_I = \mathbb{1}_I - \mathbb{P}_I$ is a projection operator onto the internal motions of site I. The AA GNM potential for G_{AA} is then given by Eq. (71). This AA potential maps to the same CG Kirchhoff matrix, $\mathbf{K}_*$, for every choice of $\delta\boldsymbol{\kappa}$.

As might be expected, $\delta\boldsymbol{\kappa}$ reflects independent, additive contributions for each site. However, these contributions reflect not only the intra-site bonding network, $\boldsymbol{\kappa}_{I;\mathrm{in}}$, but also the effects of inter-site bonds. Note that the Kirchhoff matrix, $\boldsymbol{\kappa}_{I;\mathrm{c}}$, for a fully connected intra-site graph, $G_{I;\mathrm{c}}$ is proportional to the projection operator $\mathbb{Q}_I$: $\boldsymbol{\kappa}_{I;\mathrm{c}} = n_I \mathbb{Q}_I$. Thus, Eq. (76) indicates that the inter-site bonds have been uniformly smeared across the intrasite network in $\delta\boldsymbol{\kappa}_I$. Moreover, it follows that the information lost by coarse-graining reflects both the intra-site bond networks, as well as the atomic bonds between CG sites:

$$||\hat{\boldsymbol{\kappa}}_{\mathrm{AA}}|| = \prod_{I=1}^N \det{}_1 \delta\boldsymbol{\kappa}_I. \tag{77}$$

Equation (76) demonstrates that the inter-site connections systematically increase $\det_1 \delta\boldsymbol{\kappa}_I$. In particular, if the intra-site network is fully connected, $\boldsymbol{\kappa}_{I;\text{in}} \to n_I \mathbb{Q}_I$, then $\det_1 \delta\boldsymbol{\kappa}_I$ achieves its maximum: $\det_1 \delta\boldsymbol{\kappa}_I \to (n_I + N_I)^{n_I - 1}$. Moreover, it is interesting that the ratio of weighted spanning trees for the AA and CG graphs can be expressed:

$$\frac{t_{\boldsymbol{\kappa}}}{T_{K_*}} = \prod_{I=1}^{N} n_I^{-1} \det_1 \delta\boldsymbol{\kappa}_I. \tag{78}$$

IV. METHODS

A. High and low resolution models for actin

In section V A we adopt a Gaussian Network Model (GNM) as a simple model for the equilibrium fluctuations of actin about its folded conformation. We defined the actin equilibrium structure by the three-dimensional coordinates for the 369 residues in the PDB structure 1J6Z, including the coordinates of the methylated histidine 73.[93] Although adenosine diphosphate (ADP) is present in this PDB structure, we did not explicitly represent ADP in the GNM. The high resolution GNM represents each amino acid with its α carbon. We employed ProDy version 3.0.4[94] to determine the Kirchhoff matrix, $\boldsymbol{\kappa}$, for the high resolution GNM, while adopting a cut-off of $r_{\text{c}} = 7.5$ A to identify contacting residues. We assigned the same mass, m, to each residue in the high resolution GNM.

We determined the normal mode frequencies, ω_i, of the high resolution GNM from the eigenvalue equation $|\Gamma\boldsymbol{\kappa} - \omega_i^2 \mathbf{g}^2| = 0$, where Γ is the GNM spring constant, and $\mathbf{g} = \text{diag}(m^{1/2})$ is the $n \times n$ mass-weighting matrix.[63,79] We define the frequency scale by $\omega_0 = \sqrt{\Gamma/m}$ and report dimensionless scaled frequencies $\tilde{\omega} \equiv \omega/\omega_0 \to \omega$ in the following.

We determined the CG coordinate, R_I, of each site, I, by the geometric center of the n_I atoms associated with the site. We defined the mass, m_I, of site I by the net mass of the associated atoms, i.e., $m_I = n_I m$. We determined the normal mode frequencies, ω_I, of the low resolution GNM from the eigenvalue equation $|\Gamma\mathbf{K} - \omega_I^2 \mathbf{G}^2| = 0$, where $\mathbf{K}$ is the CG spring matrix, and $\mathbf{G} = \text{diag}(m_I^{1/2})$ is the $N \times N$ mass-weighting matrix. We report dimensionless frequencies for the CG model by scaling with respect to the same constant, ω_0.

Equation (62) expresses the mapping information loss as a sum of two terms, $\text{I}_{\text{map}}(\mathbf{M}) = n_{\text{x}} h_1 + \frac{1}{2} \ln\left(||\hat{\boldsymbol{\kappa}}_{\text{AA}}|| \,/\, ||\Delta_N|| \right)$. While the second term depends upon the details of the mapping,

$\mathbf{M}$, the first term depends only upon the number of degrees of freedom, $n_{\mathrm{x}} = n - N$, that have been eliminated from the high resolution model. This first term is proportional to the dimensional constant, $h_1 \equiv \ln[L/L_{\mathrm{vib}}] - 1/2$, where L is the length of the system enclosing the protein, $L_{\mathrm{vib}} = \sqrt{2\pi/\beta\Gamma}$ is a characteristic length-scale for thermal vibrations, and $L/L_{\mathrm{vib}} \gg 1$ in order to analytically treat the GNM. For a fixed number of CG sites, N, this first term, $n_{\mathrm{x}} h_1$, only introduces an overall shift defining the baseline for $\mathrm{I}_{\mathrm{map}}$. In the following numerical calculations, we adopted $\beta\Gamma = 1$ A^{-2}, which is qualitatively consistent with the experimentally measured B-factors for actin,[93] and $L/L_{\mathrm{vib}} \approx 79.8$.

B. Mapping space

We consider N-site CG representations that partition the n atoms into N disjoint connected subsets and associate a CG site with the geometric center of each subset. Each N-site mapping, $\mathbf{M}$, is in one-to-one correspondence with an atomic partition $(V_1, \ldots, V_N)$ where $V_I = \{i | c_{Ii} > 0\}$ such that $\cup_{I=1}^{N} V_I = \{1, \ldots, n\}$ and $V_I \cap V_J = \emptyset$ for all $I \neq J$. (In order to identify a unique partition, we order the sets V_I such that atom 1 is in V_1 and set V_I contains the first atom that is not in the sets $V_1, \ldots, V_{I-1}$.) We require that the atoms, V_I, associated with each site, I, are connected by the springs of the high resolution GNM. We defined the N-site mapping space, $\mathcal{M}_N$, as the set of all such N-site maps.

We employed Monte Carlo (MC) simulations to explore the space, $\mathcal{M}_N$, of N-site CG representations for actin. Each MC simulation sampled a Boltzmann distribution

$$\mathcal{P}(\mathbf{M}; \beta, \lambda, \mathcal{E}_{\mathrm{bias}}) \propto \exp\left[-\beta(\mathcal{E}(\mathbf{M}) + \lambda\sigma^2(\mathbf{M}) + \mathcal{E}_{\mathrm{bias}}(\mathbf{M}))\right], \tag{79}$$

where $\mathcal{E}(\mathbf{M})$ is the base energy function, $\sigma^2(\mathbf{M}) = \mathrm{var}\{n_1, \ldots, n_N\}$ is the variance in the size of the N sites, and $\mathcal{E}_{\mathrm{bias}}(\mathbf{M})$ is a bias energy, while β and λ are sampling parameters analogous to the inverse temperature and external pressure in a constant NPT simulation. Here we defined our base energy function as the non-trivial part of I_{CG} and $\mathrm{I}_{\mathrm{map}}$: $\mathcal{E}(\mathbf{M}) = \ln T_{\mathbf{K}}(\mathbf{M}) = \ln t_{\boldsymbol{\kappa}} - \ln\left(||\hat{\boldsymbol{\kappa}}_{\mathrm{AA}}(\mathbf{M})|| \,/\, ||\Delta_N(\mathbf{M})||\right)$. The bias potential is defined

$$\mathcal{E}_{\mathrm{bias}}(\mathbf{M}; \mathcal{Q}_k, \mathcal{E}_k, \sigma_k^2) = \frac{1}{2}k_{\mathcal{Q}}\left(\mathcal{Q}(\mathbf{M}) - \mathcal{Q}_k\right)^2 + \frac{1}{2}k_{\mathcal{E}}\left(\mathcal{E}(\mathbf{M}) - \mathcal{E}_k\right)^2 + \frac{1}{2}k_{\sigma^2}\lambda\left(\sigma^2(\mathbf{M}) - \sigma_k^2\right)^2, \tag{80}$$

where $\mathcal{Q}_k$, $\mathcal{E}_k$, σ_k^2, and the corresponding spring constants were chosen to target specific regions of mapping space. In the majority of simulations $k_{\mathcal{Q}} = 0$.

As in our previous works, each MC simulation in $\mathcal{M}_N$ started from the same block map, $\mathbf{M}_{\mathrm{B}N}$. Given a fixed number, N, of CG sites, we define the block size $n_{\mathrm{B}N} = \mathrm{floor}(n/N)$. We define the block map, $\mathbf{M}_{\mathrm{B}N}$, by associating CG sites $I = 1, \ldots, N-1$ with the first $N-1$ blocks of $n_{\mathrm{B}N}$ consecutive residues in the protein sequence. We associated the last CG site with the remaining $n - n_{\mathrm{B}N}(N-1)$ residues. Starting from $\mathbf{M}_{\mathrm{B}N}$, we employed a steal move set to perform a random walk through mapping space. Given a map, $\mathbf{M}$, a steal move proposes a new map, $\mathbf{M}'$, by moving a single atom between two sites in such a way that both modified sites remain connected. The move is accepted or rejected according to a criterion that satisfies detailed balance. The MC simulations employed NetworkX to analyze the graph associated with each map.[95] We performed each MC simulation for 2.5 $\times 10^5$ steps, while discarding the first 5 $\times 10^3$ MC steps as equilibration and sampling every 10^{th} map from the remainder of the simulation. Ref. 63 provides a much more detailed description of both mapping space and our MC methods.

C. CG Bond Distributions

In section V A we present CG bond length distributions for different N = 2-site CG representations of the high resolution actin GNM. For these calculations, we define the high resolution configuration by the x-coordinates of the α carbons for the n = 369 residues in the actin sequence, $\mathbf{r} = (x_1, \ldots, x_n)$. Similarly, we define the high resolution reference configuration, $\mathbf{r}^*$, by the corresponding x-coordinates in the PDB structure 1J6Z.[93] The N = 2-site CG representation specifies the x-coordinates for the 2 sites, $\mathbf{R} = (X_1, X_2)$, while the mapped reference structure $\mathbf{R}^* = \mathbf{M}\mathbf{r}^* = (X_1^*, X_2^*)$ explicitly depends upon the mapping, $\mathbf{M}$. The N = 2-site mapped distribution is

$$p_{\mathrm{R}}(\mathbf{R}; \mathbf{M}) \propto \exp\left[-\frac{1}{2}\beta\Gamma\delta\mathbf{R}^{\dagger}\mathbf{K}\delta\mathbf{R}\right] = \exp\left[-\frac{1}{4}\beta\Gamma\Lambda\left(R - R^*\right)^2\right] \propto p_{\mathrm{R}}(R; \mathbf{M}). \qquad (81)$$

In the first Gaussian expression $\delta\mathbf{R} = \mathbf{R} - \mathbf{R}^*$ and $\mathbf{K} = \mathbf{K}(\mathbf{M})$ is the CG Kirchhoff matrix, which depends upon $\mathbf{M}$ according to Eq. (58). In the second Gaussian $\Lambda = \Lambda(\mathbf{M})$ is the positive eigenvalue of $\mathbf{K}(\mathbf{M})$, $R = X_1 - X_2$ is the CG bond length, and $R^* = X_1^* - X_2^*$ is the CG bond length in the mapped reference structure, $\mathbf{R}^*$. This second Gaussian determines the mapped bond length distribution, $p_{\mathrm{R}}(R; \mathbf{M})$.

Section V A also presents bond distributions for randomly selected maps that are representative of particular values for the spectral quality and labelling entropy. For each target value of the spectral quality, $\mathcal{Q}_j$, we identified the set, $\mathcal{S}(\mathcal{Q}_j)$, of all sampled maps, $\mathbf{M}$, with $\mathcal{Q}_j - 0.0005 \leq \mathcal{Q}(\mathbf{M}) \leq \mathcal{Q}_j + 0.0005$. Similarly, given the set of labelling entropies, $\{\mathrm{H}_\mathrm{L}(\mathbf{M})\}$, for the sampled maps, we selected 7 representative values, $\mathrm{H}_{\mathrm{L};k}$. Since the spectrum for the labelling entropy is discrete, we associated each representative value, $\mathrm{H}_{\mathrm{L};k}$, also with the adjacent values in the spectrum. This allows us to identify a set, $\mathcal{S}(\mathcal{Q}_j, \mathrm{H}_{\mathrm{L};k})$, of sampled maps with corresponding values of the spectral quality and labelling entropy. We randomly selected one map, $\mathbf{M}$, from this set, $\mathcal{S}(\mathcal{Q}_j, \mathrm{H}_{\mathrm{L};k})$, i.e., according to a uniform distribution. We presented the bond distribution for each sampled map as a function of the displacement from equilibrium, $\delta R = R - R^*$.

D. Perturbing AA spring matrices

In Section V B, we consider the impact of perturbing the underlying spring matrix, $\boldsymbol{\kappa}_*$, either by shuffling or deleting randomly selected springs. Let $G = (V_\mathrm{AA}, E)$ be the graph associated with $\boldsymbol{\kappa}_*$, where $V_\mathrm{AA} = \{1, \ldots, n\}$ is the set of n atoms and $E = \{e_{ij}\}$ is the set of springs defined by $\boldsymbol{\kappa}_*$. We define the set of backbone springs, $E_\mathrm{b} = \{e_{ij} \in E || i - j| = 1\}$, and the set of long-ranged springs $E_\mathrm{n} = E - E_\mathrm{b} = \{e_{ij} \in E || i - j| > 1\}$. We determined a set of springs, E_x, to perturb by randomly sampling a fraction, f, of the springs in E_n without replacement. In the case of shuffling experiments, we first randomly selected one atom k of the pair $\{i, j\}$ for each sampled spring $e_{ij} \in E_\mathrm{x}$. We then randomly selected a new atom $k' \notin \{i, j\}$ that was not connected to atom k. We replaced the spring $e_{ij} \in E$ with a new spring $e_{kk'} \notin E$ with $k \in \{i, j\}$ and $k' \notin \{i, j\}$. In the case of deletion experiments, we simply deleted the springs in E_x from E. In both cases, we repeated this process 100 times for each fraction, f, of edges.

E. Distance in mapping space

As in our previous studies,[62,63] we adopt the variation of information (VI) as a formal metric for measuring the distance between representations based upon the overlap between the corresponding atomic partitions.[96] Consider a mapping, $\mathbf{M}$, that corresponds to the atomic partition $(V_1, \ldots, V_N)$, where $V_I = \{i | c_{Ii} > 0\}$ is the set of atoms associated site I. We define $n_I = |V_I|$ as the number of atoms in the set V_I and $P_I(\mathbf{M}) = n_I / n$ as the probability of randomly selecting an atom in V_I. The entropy of this partition[69] is then

$$H_1(\mathbf{M}) = -\sum_{I=1}^{N} P_I(\mathbf{M}) \ln P_I(\mathbf{M}). \tag{82}$$

Now consider a second mapping, $\mathbf{M}' \sim (V_1', \ldots, V_N')$, where $V_{I'}' = \{i | c_{I'i} > 0\}$ is the set of atoms associated with site I' in $\mathbf{M}'$. We define $n_{II'}$ as the number of atoms in the set $V_I \cap V_{I'}'$. We then define $P_{II'}(\mathbf{M}, \mathbf{M}') = n_{II'}/n$ 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 joint entropy, $H_2(\mathbf{M}, \mathbf{M}')$, and the mutual information, $\mathrm{MI}(\mathbf{M}, \mathbf{M}')$, associated with the corresponding partitions[69] by

$$\begin{aligned} H_2(\mathbf{M}, \mathbf{M}') &= -\sum_{I=1}^{N}\sum_{I'=1}^{N} P_{II'}(\mathbf{M}, \mathbf{M}') \ln P_{II'}(\mathbf{M}, \mathbf{M}') & (83) \\ \mathrm{MI}(\mathbf{M}, \mathbf{M}') &= -\sum_{I=1}^{N}\sum_{I'=1}^{N} P_{II'}(\mathbf{M}, \mathbf{M}') \ln \left[\frac{P_{II'}(\mathbf{M}, \mathbf{M}')}{P_I(\mathbf{M}) P_{I'}(\mathbf{M}')} \right]. & (84) \end{aligned}$$

The VI quantifies the information in $P_{II'}(\mathbf{M}, \mathbf{M}')$ that is not shared between the two mappings[96]

$$\mathrm{VI}(\mathbf{M}, \mathbf{M}') = \mathrm{H}_2(\mathbf{M}, \mathbf{M}') - \mathrm{MI}(\mathbf{M}, \mathbf{M}') = \mathrm{H}_1(\mathbf{M}) + \mathrm{H}_1(\mathbf{M}') - 2\mathrm{MI}(\mathbf{M}, \mathbf{M}'). \tag{85}$$

In Section V B, we employ VI to measure the distance of a map, $\mathbf{M}$, from the resonant mapping, $\mathbf{M}_*$, according to $d_*(\mathbf{M}) \equiv \mathrm{VI}(\mathbf{M}, \mathbf{M}_*)$.

V. RESULTS AND DISCUSSION

A. Labelling entropy

We first investigate the impact of the CG mapping, $\mathbf{M}(\mathbf{r})$, upon the partitioning of atomic configurational information between the mapped distribution, $p_{\mathrm{R}}(\mathbf{R})$, and the conditioned distribution, $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})$, that describes the atomic degrees of freedom that are eliminated from the CG model. We adopt a GNM as a simple high resolution model for the equilibrium fluctuations of actin about its folded conformation. The high resolution GNM represents each amino acid with its α carbon and introduces an isotropic linear spring between each pair of contacting residues that are within $r_{\mathrm{c}} = 7.5$ A in the folded reference structure, $\mathbf{r}^*$. For simplicity, we assign the same mass, m, to each amino acid.

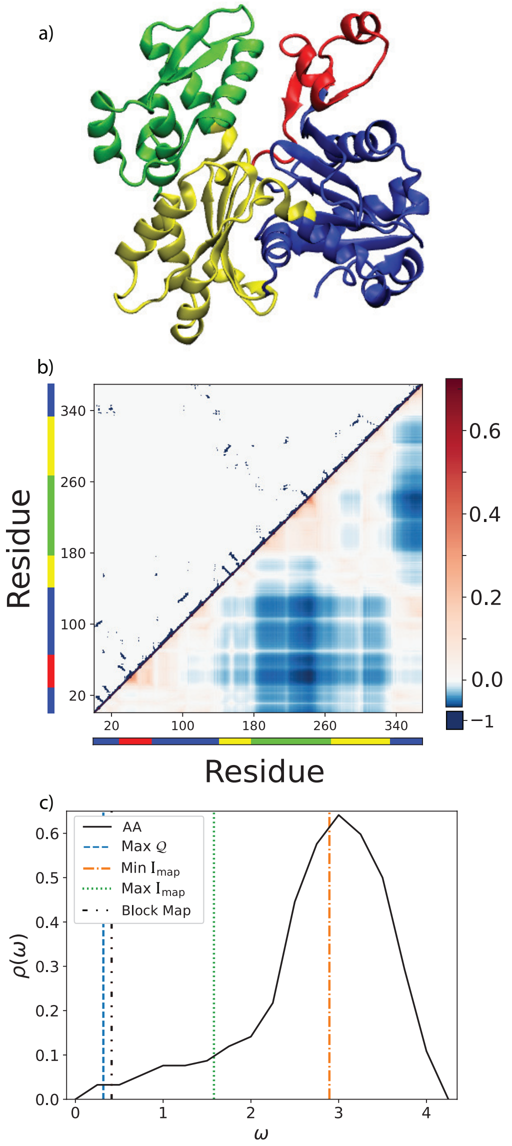

FIG. 1. Analysis of the high resolution GNM for actin. Panel a presents a ribbon cartoon of the reference folded structure, $\mathbf{r}^*$. Each amino acid is colored according to its biochemical domain assignment.[93] Domains 1, 2, 3, and 4 are colored blue, red, yellow, and green, respectively. Panel b presents intensity plots of the AA Kirchhoff matrix, $\boldsymbol{\kappa}$, and the scaled vibrational covariance matrix, $\beta\Gamma\mathbf{c}_{\mathrm{v}} = \boldsymbol{\kappa}^{\mathrm{I}}$, above and below the diagonal, respectively. The horizontal bars adjacent to each axis indicate the domain assignment of each residue. Panel c presents the density of vibrational states for the AA GNM, which has been normalized to integrate to 1. The vertical lines in panel c present the vibrational frequency for various N = 2-site CG representations. The dashed blue, dotted green, and dashed-dotted orange lines correspond to the representations that maximize $\mathcal{Q}$, maximize $\mathrm{I}_{\mathrm{map}}$, and minimize $\mathrm{I}_{\mathrm{map}}$, respectively. The dotted-dashed black line corresponds to the block map, $\mathbf{M}_{\mathrm{B2}}$, that was employed as the initial map for the MC simulations described in Sec. IV B.

Figure 1a presents a ribbon cartoon of this reference actin structure, $\mathbf{r}^*$. The protein secondary structure primarily consists of α-helices and β-strands that are connected by turns and coils. Previous biochemical studies have decomposed the actin structure into four domains that are indicated by the colors in Fig. 1a.[93,97]

The top half of Fig. 1b presents the upper half of the Kirchhoff matrix, $\boldsymbol{\kappa}$, for the high resolution actin GNM. Each black mark in Fig. 1b identifies a pair of contacting residues in the reference structure. The color bars parallel to the two axes indicate the domain assignment of each residue. The black marks that are slightly above the diagonal of Fig. 1b indicate contacts between residues that are close in sequence, while black marks that are further above the diagonal indicate contacts between distinct secondary structures. The large majority of contacts correspond to residues within the same domain. The Kirchhoff matrix also indicates significant inter-domain contacts between domains 1 and 2, between domains 1 and 3, and between domains 3 and 4.

The bottom half of Fig. 1b presents the lower half of the scaled covariance matrix, $\beta\Gamma\mathbf{c}_{\mathrm{v}} = \boldsymbol{\kappa}^{\mathrm{I}}$. The covariance matrix highlights strong positive correlations within domains 2 and 4. The covariance matrix also emphasizes that the motion of domains 1 and 2 are strongly anti-correlated with the motion of domains 3 and 4.

The bottom panel of Fig. 1c presents the normalized density of vibrational states for the high resolution GNM. As expected, the density of states includes a few low frequency normal modes and many high frequency normal modes.

We consider N-site maps that partition the 369 α carbons into N disjoint sets, $V_1, \ldots, V_N$, of connected atoms. We define the CG coordinate, R_I, of site I by the geometric center of the corresponding atomic set, V_I. We explored the space of CG maps by performing Monte Carlo (MC) simulations with an ergodic "steal" move set. Starting from a given map, $\mathbf{M}$, the steal move set creates a new map, $\mathbf{M}'$, by moving a single atom to a new site.[63]

Figure 2 characterizes three particular $N = 2$-site mappings that were sampled during these MC simulations. The top panel presents the mapped probability distribution, $p_{\mathrm{R}}(R; \mathbf{M})$, for the CG bond-length, $R = R_1 - R_2$, defined by each CG representation, $\mathbf{M}$. Clearly, the CG bond distributions are very different for the three mappings. The ribbon cartoons indicate the corresponding atomic partitions, which are also indicated by the horizontal bars below the distributions. In the cartoons and the horizontal bars, each residue is colored according to the associated CG site. The vertical lines in Fig. 1c indicate the vibra-

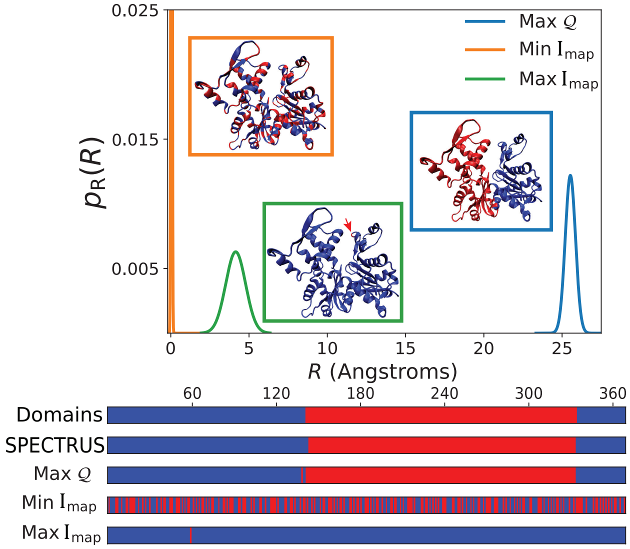

FIG. 2. Mapped bond distributions for three different N=2-site representations of the high resolution GNM. The red arrow indicates the single residue that is assigned to the second site in the map, $\mathbf{M}_{\mathrm{I}^+_{\mathrm{map}}}$, with maximal information loss. The five horizontal bars present various partitions of the 369 amino acids. The first bar colors the biochemical domains 1 and 2 blue, while coloring domains 3 and 4 red.[93,97] The remaining 4 bars assign residues according to (2) the two rigid domains identified by the SPECTRUS webserver;[98] (3) the map, $\mathbf{M}_{\mathcal{Q}}$, that maximizes $\mathcal{Q}$; (4) the map, $\mathbf{M}_{\mathrm{I}^-_{\mathrm{map}}}$, that minimizes $\mathrm{I}_{\mathrm{map}}$; and (5) the map, $\mathbf{M}_{\mathrm{I}^+_{\mathrm{map}}}$, that maximizes $\mathrm{I}_{\mathrm{map}}$.

tional frequencies for these three representations. The first three entries of Table I further characterize these 2-site maps. In particular, Table I reports the spectral quality, $\mathcal{Q}(\mathbf{M})$, and mapping information loss, $\mathrm{I}_{\mathrm{map}}(\mathbf{M})$, for each mapping, $\mathbf{M}$. The spectral quality, which is defined by Eq. (63), quantifies the extent to which a given mapping preserves the low frequency, large amplitude motions of the high resolution model. Conversely, $\mathrm{I}_{\mathrm{map}}$ quantifies the configurational information that is lost due to the mapping according to Eq. (12).

The blue distribution in Fig. 2 corresponds to the sampled map, $\mathbf{M}_{\mathcal{Q}}$, with maximal spectral quality. Because $\mathbf{M}_{\mathcal{Q}}$ optimally preserves the mass-weighted covariance, the associated atomic partitioning nicely aligns with the lowest frequency breathing mode of the high resolution GNM. Consequently, $\mathbf{M}_{\mathcal{Q}}$ effectively associates the first CG site with domains 1 and 2, while associating the second CG site with domains 3 and 4. This mapping is highly consistent with our physical intuition and, indeed, almost perfectly aligns with the two most

TABLE I. Extreme representations of the actin GNM with N= 2-, 4-, and 12-sites. For each representation, we report the spectral quality ($\mathcal{Q}$), mapping information loss ($\mathrm{I}_{\mathrm{map}}$), and labelling entropy (H_{L}). We also report the number of intra-site and inter-site bonds, $n_{\mathrm{b;Intra}}$ and $n_{\mathrm{b;Inter}}$, respectively. Consider a high-resolution bond between two atoms, i and j, that are associated with CG sites, I and J, respectively. We classify the bond as intra-site if $I = J$ (i.e., the atoms are assigned to the same site) and inter-site if $I \neq J$ (i.e., the atoms are assigned to different sites).

N	Mapping	$\mathcal{Q}$	$\mathrm{I}_{\mathrm{map}}$	H_{L}	$n_{\mathrm{b;Intra}}$	$n_{\mathrm{b;Inter}}$
2	max $\mathcal{Q}$: $\mathbf{M}_{\mathcal{Q}}$	0.112	1795.2	5.2	1637	43
2	max $\mathrm{I}_{\mathrm{map}}$: $\mathbf{M}_{\mathrm{I}^{+}_{\mathrm{map}}}$	0.005	1795.8	3.0	1675	5
2	min $\mathrm{I}_{\mathrm{map}}$: $\mathbf{M}_{\mathrm{I}^{-}_{\mathrm{map}}}$	0.001	1793.0	5.2	767	913
4	max $\mathcal{Q}$: $\mathbf{M}_{\mathcal{Q}}$	0.222	1785.3	8.8	1578	102
4	max $\mathrm{I}_{\mathrm{map}}$: $\mathbf{M}_{\mathrm{I}^{+}_{\mathrm{map}}}$	0.024	1787.4	3.5	1652	28
4	min $\mathrm{I}_{\mathrm{map}}$: $\mathbf{M}_{\mathrm{I}^{-}_{\mathrm{map}}}$	0.004	1779.4	9.0	404	1276
12	max $\mathcal{Q}$: $\mathbf{M}_{\mathcal{Q}}$	0.362	1743.7	20.2	1296	384
12	max $\mathrm{I}_{\mathrm{map}}$: $\mathbf{M}_{\mathrm{I}^{+}_{\mathrm{map}}}$	0.052	1751.1	4.0	1598	82
12	min $\mathrm{I}_{\mathrm{map}}$: $\mathbf{M}_{\mathrm{I}^{-}_{\mathrm{map}}}$	0.144	1732.6	20.1	387	1293

rigid regions identified by the SPECTRUS webserver.[98]

Because $\mathbf{M}_{\mathcal{Q}}$ defines the two sites by splitting the actin structure into two distinct halves, the equilibrium bond length in the mapped ensemble is quite long. Moreover, because $\mathbf{M}_{\mathcal{Q}}$ preserves the low-frequency breathing mode, the corresponding mapped distribution is quite broad and relatively uninformative. Conversely, $\mathbf{M}_{\mathcal{Q}}$ partitions the actin residues such that the overwhelming majority of the high resolution GNM bonds are intra-site bonds, i.e., between residues that are associated with the same CG site. This results in a rather sharp conditioned distribution, $p_{\mathrm{r}|\mathrm{R}}$, governing the intra-site degrees of freedom in the lost ensemble. Accordingly, the mapping information loss, $\mathrm{I}_{\mathrm{map}}(\mathbf{M}_{\mathcal{Q}})$, for this mapping is rather large.

The orange distribution corresponds to the sampled mapping, $\mathbf{M}_{\mathrm{I}^{-}_{\mathrm{map}}}$, with minimal information loss, i.e., minimal $\mathrm{I}_{\mathrm{map}}$. The two sites are again quite similar in size. However, this mapping is not consistent with our physical intuition because the two sites do not correspond to coherent structural features. Rather the two sites appear to form alternating stripes on both the protein sequence and the folded structure. Consequently, $\mathbf{M}_{\mathrm{I}^{-}_{\mathrm{map}}}$ maps the two sites almost on top of each other, which results in a very short equilibrium bond length in the mapped ensemble. For the same reason, more than half of the bonds in the underlying GNM are now inter-site bonds, i.e., between residues associated with distinct sites. The many inter-site bonds strongly constrain the motion of the CG sites. This results in a very narrow and, thus, highly informative mapped probability density. In this case, $\mathbf{M}_{\mathrm{I}^{-}_{\mathrm{map}}}$ appears to maximize configurational information in the mapped ensemble by preserving localized high frequency "noise" from the AA model.

The two preceding cases suggest that it may be advantageous to adopt CG maps that *maximize* the lost configurational information, $\mathrm{I}_{\mathrm{map}}$. This would appear to simplify the mapped ensemble, $p_{\mathrm{R}}(\mathbf{R})$, by minimizing the high frequency noise that is preserved from the AA model. Moreover, this approach would maximize the information contained in the conditioned distribution, $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})$, which should minimize the effective degeneracy of each CG configuration, $\mathbf{R}$, and render back-mapping efforts more meaningful. However, this intuition fails.

The green distribution in Fig. 2 corresponds to the sampled mapping, $\mathbf{M}_{\mathrm{I}^{+}_{\mathrm{map}}}$, that maximizes $\mathrm{I}_{\mathrm{map}}$. This mapping associates one site with a single residue in a flexible loop, while representing the remainder of the protein with a single site. As a result, only 5 of the underlying GNM bonds connect residues that have been assigned to different sites. Consequently, the corresponding mapped distribution is very broad and, thus, information-poor. Nevertheless, $\mathbf{M}_{\mathrm{I}^{+}_{\mathrm{map}}}$ is clearly inconsistent with our physical intuition.

In order to understand these observations, we analyze the space of 2-site CG representations. Figure 3b presents a scatter plot of $\{\mathcal{Q}(\mathbf{M}), \mathrm{I}_{\mathrm{map}}(\mathbf{M})\}$ for the 2-site representations that we sampled during our MC simulations of mapping space. The blue, orange, and green stars indicate the three maps $\mathbf{M}_{\mathcal{Q}}$, $\mathbf{M}_{\mathrm{I}^{-}_{\mathrm{map}}}$, and $\mathbf{M}_{\mathrm{I}^{+}_{\mathrm{map}}}$, respectively, that were considered in Fig. 2. The remaining points are colored according to the labelling entropy, $\mathrm{H}_{\mathrm{L}} = \frac{1}{2}\sum_{I} \ln n_{I}$.

The colors appear to form stripes on the scatter plot in Fig. 3b. The stripes of a given color are consistent with our initial intuition. Among maps with a given labelling entropy, increasing $\mathrm{I}_{\mathrm{map}}$ reduces I_{CG}, which results in a broader mapped distribution that better preserves large scale motions and, thus, increases $\mathcal{Q}$. Equation (62) clearly explains this observation. Fixing H_{L} corresponds to fixing $||\Delta_{N}|| = \exp[2\mathrm{H}_{\mathrm{L}}]$. Consequently, increasing $\mathrm{I}_{\mathrm{map}}$ corresponds to increasing $||\hat{\boldsymbol{\kappa}}_{\mathrm{AA}}||$, which effectively transfers atomic bonds from the

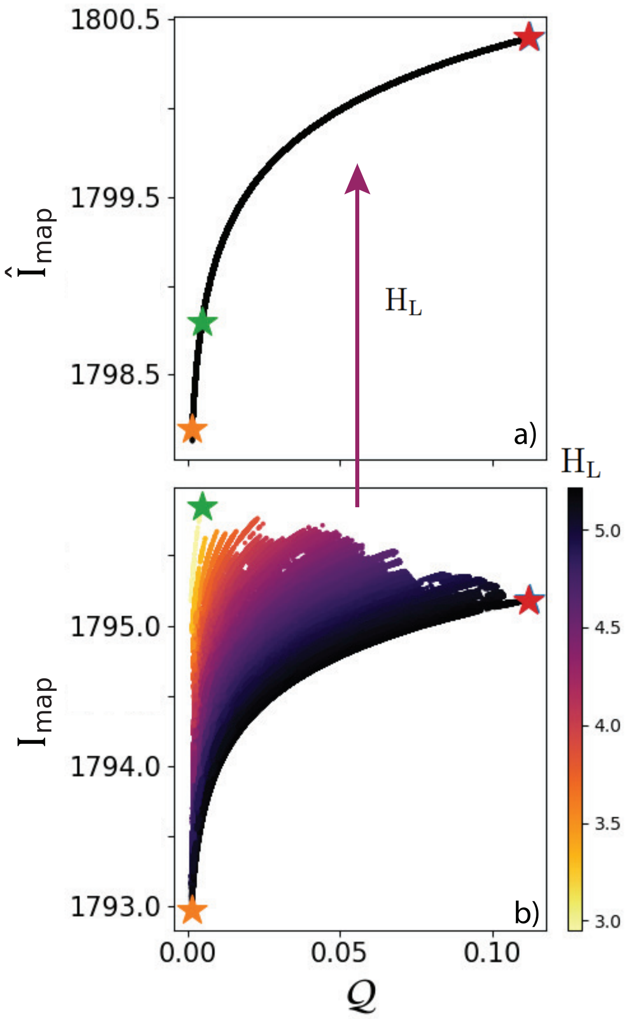

FIG. 3. Analysis of N=2-site representations for actin. Panel b presents a scatter plot of sampled representations as a function of $\mathcal{Q}$ and $\mathrm{I_{map}}$. Each point of this scatter plot is colored according to $\mathrm{H_L}$. Panel a presents a scatter plot of the same representations as a function $\mathcal{Q}$ and $\hat{\mathrm{I}}_{\mathrm{map}} = \mathrm{I_{map}} + \mathrm{H_L}$. The blue, orange and green stars indicate the maps $\mathbf{M}_{\mathcal{Q}}$, $\mathbf{M}_{\mathrm{I}^-_{\mathrm{map}}}$, $\mathbf{M}_{\mathrm{I}^+_{\mathrm{map}}}$, respectively, from Fig. 2. The red star, which obscures the blue star, indicates the CG representation that associates site 1 with the first two actin domains and site 2 with the second two actin domains.

mapped ensemble into the lost ensemble. This results in larger displacements in the mapped ensemble and, thus, increases $\mathcal{Q}$.

Figure 3 also reveals why our initial intuition failed. The colors in Fig. 3b demonstrate that $\mathrm{I_{map}}$ systematically increases as $\mathrm{H_L}$ decreases, as indicated by Eq. (62). In particular, the minimally informative mapping, $\mathbf{M}_{\mathrm{I}^+_{\mathrm{map}}}$, maximizes $\mathrm{I_{map}}$ by minimizing $\mathrm{H_L}$ with a representation that associates one site with a single residue. Conversely, the maximally informative mapping, $\mathbf{M}_{\mathrm{I}_{\mathrm{map}}^{-}}$, and the mapping with maximal spectral quality, $\mathbf{M}_{\mathcal{Q}}$, both correspond to rather homogeneous mass distributions and, thus, relatively high labelling entropy.

This suggests defining a modified mapping information loss, $\hat{\mathrm{I}}_{\mathrm{map}} = \mathrm{I}_{\mathrm{map}} + \mathrm{H}_{\mathrm{L}} = n_{\mathrm{x}} h_1 + \frac{1}{2} \ln ||\hat{\boldsymbol{\kappa}}_{\mathrm{AA}}||$, that may provide a better predictor for the spectral quality, $\mathcal{Q}$, by accounting for the labelling entropy. Figure 3 demonstrates that this is indeed the case: $\hat{\mathrm{I}}_{\mathrm{map}}$ and $\mathcal{Q}$ appear perfectly correlated among $N = 2$-site maps.

These results indicate that the information content of the mapped ensemble, $\mathrm{I}_{\mathrm{CG}} = \mathrm{I}_{\mathrm{AA}} - \mathrm{I}_{\mathrm{map}} = \mathrm{I}_{\mathrm{AA}} - \hat{\mathrm{I}}_{\mathrm{map}} + \mathrm{H}_{\mathrm{L}}$, systematically increases as the mapping becomes more homogeneous, as quantified by the labelling entropy, $\mathrm{H}_{\mathrm{L}} = \frac{1}{2} \ln ||\Delta_N||$. Section II motivated this effect via the Jacobian associated with the transformation to CG coordinates in Eq. (27), while Appendix B derived this effect for the GNM via the determinant identity in Eq. (B1). This effect can also be motivated by simple statistical considerations.

We have partitioned the n atomic coordinates $\mathbf{r} = (r_1, \ldots, r_n)$, into two disjoint sets, V_1 and V_2, and defined the CG coordinates $R_I = n_I^{-1} \sum_{i \in V_I} r_i$. The central limit theorem suggests that, as $n_I \to \infty$, the variance, $\sigma_{R_I}^2$, in the CG coordinate, R_I, should scale as n_I^{-1}. For simplicity, we approximate the variance in both site coordinates by $\sigma_{R_I}^2 \approx \sigma_r^2 / n_I$, where σ_r^2 corresponds to the variance in the coordinates of a characteristic atom. Assuming that the CG coordinates are weakly correlated, the variance in the CG bond length, $R = R_1 - R_2$, is

$$\sigma_R^2 \approx \sigma_{R_1}^2 + \sigma_{R_2}^2 \approx \frac{\sigma_r^2}{n_1} + \frac{\sigma_r^2}{n_2} = \frac{n \sigma_r^2}{||\Delta_N||}. \tag{86}$$

Thus, given these simplifying approximations, the width of the mapped ensemble scales inversely with $||\Delta_N|| = n_1 n_2$. If we define $\phi \equiv n_1 / n$ as the fraction of atoms assigned to site 1, then $\sigma_R^2 \approx n^{-1} \sigma_r^2 / f(\phi)$ where $f(\phi) = \phi(1 - \phi)$. On the interval $0 \leq \phi \leq 1$, $f(\phi)$ achieves its maximum at $\phi = 1/2$ and approaches its minimum as $\phi \to 0$ or 1. Therefore, at least in this simple example, the uncertainty in the mapped ensemble systematically decreases as the mapping becomes increasingly uniform.

The preceding analysis considered an extremely simple case. Nevertheless, Fig. 4 demonstrates that these considerations qualitatively apply for the actin GNM. Figure 4 presents mapped bond displacement distributions, $p_{\mathrm{R}}(\delta R; \mathbf{M})$, for representative 2-site CG representations with varying labelling entropies, H_{L}. The top, middle, and bottom panels compare representations with relatively low, intermediate, and high spectral quality, respectively.

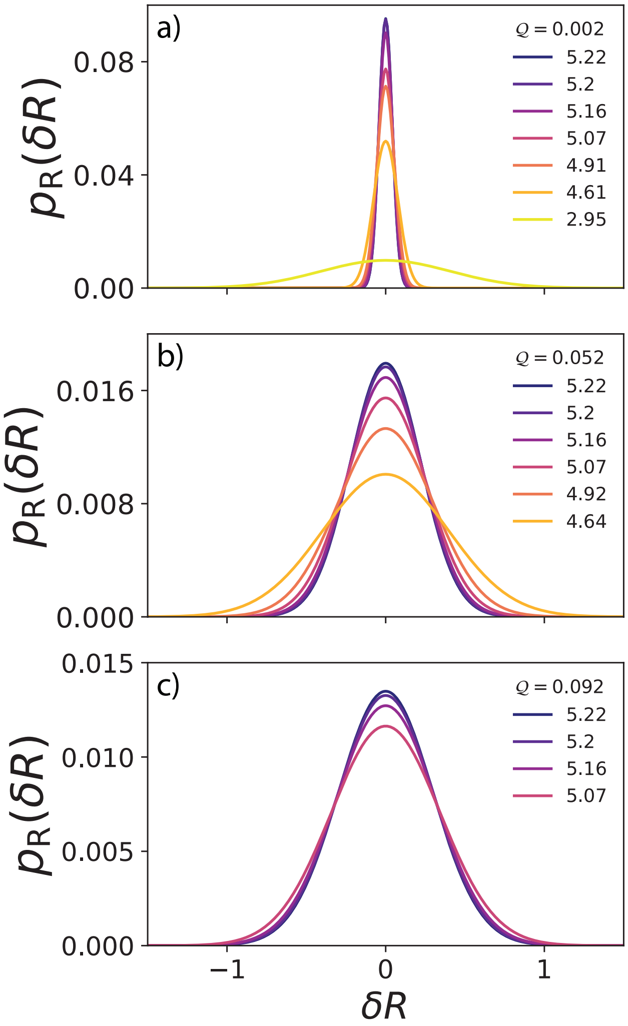

FIG. 4. Mapped bond displacement distributions, $p_{\mathrm{R}}(\delta R; \mathbf{M})$, for representative N=2-site maps, $\mathbf{M}$. The top, middle, and bottom panels correspond to representations with relatively low ($\mathcal{Q} = 0.002$), moderate ($\mathcal{Q} = 0.052$), and high ($\mathcal{Q} = 0.092$) spectral quality. Each curve is colored according to the labelling entropy, H_{L}, of the corresponding mapping. Section IV C describes these calculations in greater detail.

As expected, the mapped ensemble generally broadens as $\mathcal{Q}$ increases. Moreover, among representations with a given spectral quality, the uncertainty in the mapped ensemble systematically increases as the labelling entropy decreases, i.e., as the site size distribution becomes increasingly heterogeneous. Consequently, there exist maps with very heterogeneous site distributions that are characterized by both very low spectral quality and also very broad, uninformative mapped ensembles.

To this point we have focused on two-site CG representations. We now briefly consider

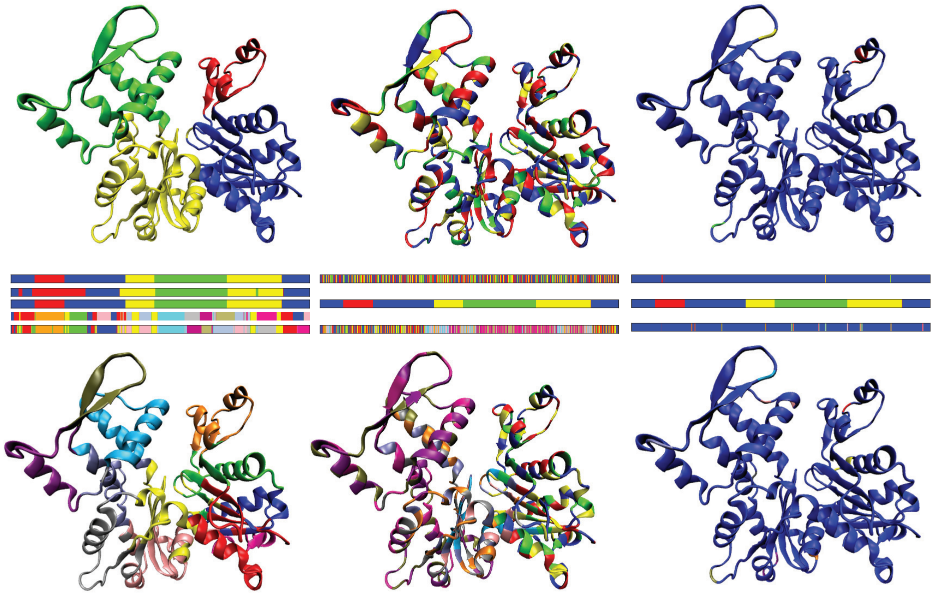

FIG. 5. CG representations that maximize $\mathcal{Q}$ (left), minimize $\mathrm{I}_{\mathrm{map}}$ (center), and maximize $\mathrm{I}_{\mathrm{map}}$ (right) for N = 4-site (top) and 12-site (bottom) representations. The horizontal bars between the two rows of representations indicate corresponding residue assignments. In each vertical stack of horizontal bars, the top and bottom bars indicate the corresponding 4- and 12-site CG representations, respectively, while the central bar indicates the 4 domains identified in the biochemical literature.[93] In the left-most stack, the second and fourth horizontal bars indicate the rigid domain decomposition identified by the SPECTRUS webserver[98] for 4 and 12 domains, respectively.

slightly higher resolution representations of actin. We again performed MC simulations to explore mapping space for N = 4- and 12-sites.

The left column of Fig. 5 presents the sampled maps, $\mathbf{M}_{\mathcal{Q}}$, that maximize $\mathcal{Q}$ for N = 4- and 12-site representations. The dashed blue curves in Fig. 6 demonstrate that these maps nicely preserve the lowest frequency modes of the AA model, while filtering out the high frequency modes. Table I indicates that these maps are characterized by a relatively high labelling entropy and, thus, a relatively uniform mass distribution. Moreover, these maps are characterized by a relatively large number of intrasite bonds, which corresponds to a broad mapped distribution, $p_{\mathrm{R}}(\mathbf{R})$, and a narrow conditioned distribution, $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})$. At both resolutions, the map $\mathbf{M}_{\mathcal{Q}}$ is consistent with our physical intuition, as it assigns the CG sites to distinct structural motifs. These representations are also quite consistent with the rigid domains identified by the SPECTRUS webserver.[98] Interestingly, the 4-site map with maximal spectral quality aligns almost perfectly with the four actin domains that are discussed in the biochemical literature.[93] Moreover, this 4-site representation appears similar to the CG representation recently identified by combining the ED-CG method with K-means clustering.[54] Thus, the spectral quality appears to be a reasonable metric for identifying high quality CG representations of actin with $N = 2$, 4, or 12 sites.

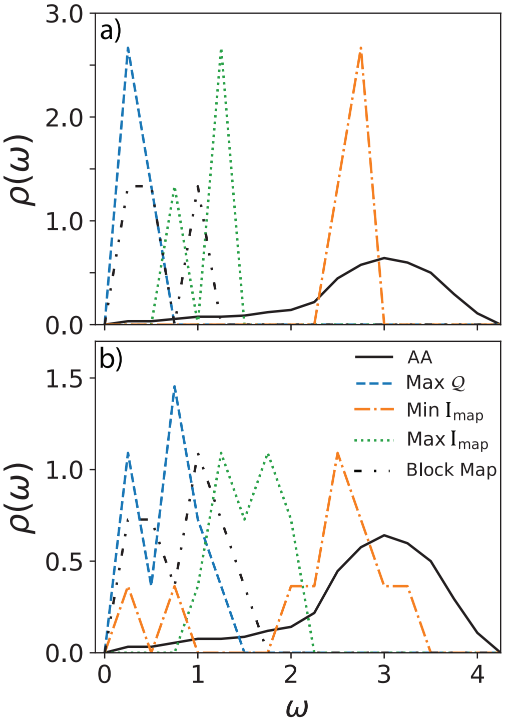

FIG. 6. Normalized vibrational DoS for various CG representations of actin. The solid black curve presents the AA DoS, while the dotted-dashed black curve present results for the block map, $\mathbf{M}_{\mathrm{B}N}$. The dashed blue, dotted green, and dashed-dotted orange curves present results for the representations that maximize $\mathcal{Q}$, maximize $\mathrm{I}_{\mathrm{map}}$, and minimize $\mathrm{I}_{\mathrm{map}}$, respectively. Panels a and b present results for $N = 4$ and 12 -site representations, respectively.

The center and right columns of Fig. 5 present the sampled maps, $\mathbf{M}_{\mathrm{I}^-_{\mathrm{map}}}$ and $\mathbf{M}_{\mathrm{I}^+_{\mathrm{map}}}$, that minimize and maximize the mapping information loss, $\mathrm{I}_{\mathrm{map}}$, respectively. The dashed-dotted orange and dotted green curves in Fig. 6 present the vibrational densities of states for these representations. The representations that minimize $\mathrm{I}_{\mathrm{map}}$ do not associate CG sites with coherent groups. Rather, these representations partition residues in such a way that more than 75 % of the atomic springs link residues in distinct CG sites. As a result, these representations generate very narrow mapped ensembles that reflect the localized and, thus, informative high-frequency motions of the high resolution model. Conversely, the representations that maximize $\mathrm{I}_{\mathrm{map}}$ represent the overwhelming majority of the protein with a single residue, which results in a relatively small value for the labelling entropy, H_{L}. Consequently, Fig. 5 indicates that neither minimizing nor maximizing $\mathrm{I}_{\mathrm{map}}$ is consistent with our physical intuition.

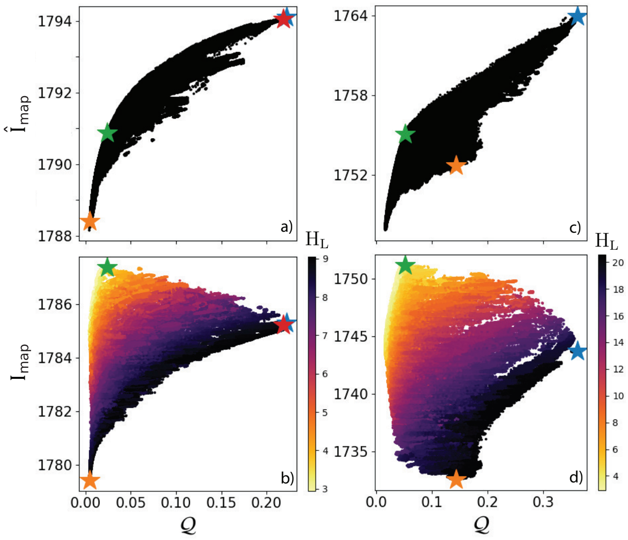

FIG. 7. Analysis of $N = 4$-site (left) and 12-site (right) representations for actin in analogy to Fig. 3. Panels b and d present a scatter plot of sampled representations as a function of $\mathcal{Q}$ and $\mathrm{I}_{\mathrm{map}}$ with each point colored according to H_{L}. Panels a and c present scatter plots of the same representations as a function $\mathcal{Q}$ and $\hat{\mathrm{I}}_{\mathrm{map}} = \mathrm{I}_{\mathrm{map}} + \mathrm{H}_{\mathrm{L}}$. The blue, orange, and green stars indicate the three maps from Fig. 5 that maximize $\mathcal{Q}$, minimize $\mathrm{I}_{\mathrm{map}}$, and maximize $\mathrm{I}_{\mathrm{map}}$, respectively. The blue stars in panels a and b are obscured by the red star, which indicates the mapping defined by the biochemical domain structure.

Figure 7 demonstrates that the labelling entropy also plays a significant role for higher resolution representations. The bottom row of Fig. 7 presents a scatter plot of $\mathcal{Q}$ and $\mathrm{I}_{\mathrm{map}}$ for sampled CG representations with $N = 4$- and 12-sites. As in Fig. 3, we have colored the points according to the labelling entropy, $\mathrm{H}_{\mathrm{L}}(\mathbf{M})$, for the corresponding CG representation, $\mathbf{M}$. There appears to be a significant correlation between $\mathcal{Q}$ and $\mathrm{I}_{\mathrm{map}}$ among maps with a given H_{L}, but $\mathrm{I}_{\mathrm{map}}$ systematically increases as H_{L} decreases. Consequently, there is little correlation between $\mathcal{Q}$ and $\mathrm{I}_{\mathrm{map}}$ across the ensemble of sampled maps. The top row of Fig. 7 presents corresponding scatter plots of $\mathcal{Q}$ and $\hat{\mathrm{I}}_{\mathrm{map}} = \mathrm{I}_{\mathrm{map}} + \mathrm{H}_{\mathrm{L}}$ that account for the information loss associated with the site assignments. As in Fig. 3 for $N = 2$-site representations, Fig. 7 demonstrates a very strong positive correlation between $\mathcal{Q}$ and $\hat{\mathrm{I}}_{\mathrm{map}}$ for $N = 4$ and 12, although this correlation is no longer perfectly 1-to-1 at these higher resolutions.

B. Resonance between AA and CG models

In Section III D, we considered the possibility of a "resonance" between a high resolution model and a CG mapping. In this case, the high resolution potential is separable and does not couple the AA and CG subspaces of the high resolution configuration space. We construct this resonance by first specifying a CG spring matrix, $\mathbf{K}_*$, along with a corresponding CG mapping, $\mathbf{M}_*$. We use $\mathbf{M}_*$ to project $\mathbf{K}_*$ into the AA configuration space, $\mathbf{K}_* \rightarrow \boldsymbol{k}_* = \mathbf{M}_*^\dagger \mathbf{K}_* \mathbf{M}_*$. We define the high resolution spring matrix, $\boldsymbol{\kappa}$, by decorating $\boldsymbol{k}_*$ with atomic details, $\delta\boldsymbol{\kappa}$: $\boldsymbol{\kappa} = \boldsymbol{k}_* + \delta\boldsymbol{\kappa}$. By construction, the mapping, $\mathbf{M}_*$, perfectly preserves the underlying CG component of the AA model, while eliminating these atomic details. In this section, we briefly consider how robust this resonance is to the choice of mapping and to the details of the atomic potential.

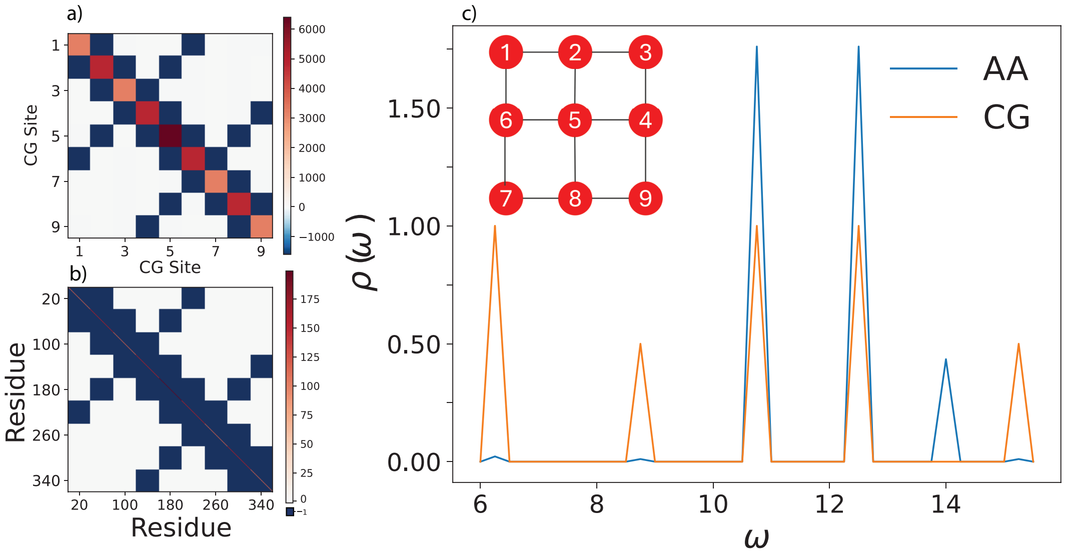

FIG. 8. Toy model illustrating "resonance" between an atomic model and the CG mapping. Panels a and b present the corresponding CG and AA spring matrices, $\mathbf{K}_*$ and $\boldsymbol{\kappa}_*$, respectively. Panel c presents the normalized vibrational density of states for both models.

Figure 8 illustrates a toy model for this notion of resonance. Figure 8a presents the spring matrix, $\mathbf{K}_*$, for an $N = 9$-site CG model. This spring matrix is reminiscent of the GNM for three anti-parallel β strands. For simplicity, we assume that the mapping, $\mathbf{M}_*$, defines the coordinates of each CG site, I, from the geometric center for $n_I = 40$ consecutive atoms in the protein sequence. As above, we assume that each atom has the same mass, m, and that each CG site has a mass $m_I = 40m$.

The CG spring matrix, $\mathbf{K}_*$, is the Laplacian matrix for the weighted graph, G_{CG}, that is shown in the inset of Fig. 8c. We weight each CG edge, e_{IJ}, in G_{CG} by $w(e_{IJ}) = n_I \times n_J = 1600$. Figure 8b presents the corresponding all-atom (AA) spring matrix, $\boldsymbol{\kappa}_*$. As discussed in Section III D, there exists a family of AA spring matrices, $\boldsymbol{\kappa} = \boldsymbol{k}_* + \delta\boldsymbol{\kappa}$, that are all resonant with $\mathbf{K}_*$ but that differ in atomic details, $\delta\boldsymbol{\kappa}$. For simplicity, we have selected the AA spring matrix, $\boldsymbol{\kappa}_*$, in this family that is maximally connected, i.e., $\delta\boldsymbol{\kappa}_{I*} = (n_I + N_I)\mathbb{Q}_I$. Each spring in the CG network is 1600 times stronger than the atomic springs, but the AA network compensates for this by introducing 1600 springs between each pair of connected CG sites. Note also that, while the toy AA model has a similar number of atoms to the actin GNM, this toy model is much more strongly coupled.

TABLE II. Normal mode frequencies, ω, and degeneracies, $\Omega(\omega)$, for the AA and CG toy models.

ω	$\Omega_{\mathrm{AA}}(\omega)$	$\Omega_{\mathrm{CG}}(\omega)$
6.325	2	2
8.944	1	1
10.954	158	2
12.649	158	2
14.142	39	0
15.492	1	1

The blue curve in Fig. 8c presents the normalized vibrational density of states, $\rho(\omega)$, for the AA toy model, which is also summarized in Table II. The AA density of states (DoS) contains 359 finite modes, but these are distributed across only 6 finite frequencies due to the high symmetry of $\boldsymbol{\kappa}$. In particular, the AA DoS contains two symmetric modes at the fundamental frequency, $\omega \approx 6$, as well as a single mode at $\omega \approx 9$. As expected the AA DoS is overwhelmingly dominated by higher frequency modes, $\omega > 10$.

The orange curve in Fig. 8c presents the normalized vibrational DoS for the CG model. While the CG DoS contains only 8 modes, it perfectly preserves 5 of the 6 finite frequencies in the AA DoS.[87] Moreover, the CG model perfectly preserves both of the symmetric fundamental modes at $\omega \approx 6$, as well as the slightly higher nondegenerate mode at $\omega \approx 9$. However, the shape of the CG DoS is dramatically different from the AA DoS. Whereas less than 1 % of the AA modes have frequencies below 10, nearly 40 % of the CG modes have frequencies below 10. Due to the high symmetry of the toy model and the strength of the CG springs, the high frequency AA mode at $\omega \approx 16$ actually lies in the CG subspace and is preserved by the mapping. Nevertheless, the resonant mapping perfectly preserves all low frequency modes of the AA model and completely filters out the high frequency modes that reflect atomic decorations.

Given the fixed AA spring matrix, $\boldsymbol{\kappa}$, in Fig. 8b, we now consider how the properties of the CG model deteriorate as the CG mapping, $\mathbf{M}$, moves off resonance. Specifically, starting from the resonant mapping, $\mathbf{M}_*$, we consider each neighboring map, $\mathbf{M}$, that differs by the assignment of a single atom. We select the map, $\mathbf{M}_{\mathcal{Q}1}$, that has lowest spectral quality within this neighborhood. We repeat this process to step through mapping space in order to generate a sequence of maps of decreasing spectral quality, $\mathcal{Q}(\mathbf{M}_*) > \mathcal{Q}(\mathbf{M}_{\mathcal{Q}1}) > \mathcal{Q}(\mathbf{M}_{\mathcal{Q}2}) > \cdots > \mathcal{Q}(\mathbf{M}_{\mathcal{Q}\infty})$, until the walk terminates when we reach a map, $\mathbf{M}_{\mathcal{Q}\infty}$, that has lower spectral quality than any of its neighbors. While this walk provides some local information about moving off resonance, it does not address the statistical properties of mapping space. It may be beneficial to statistically characterize the neighborhood of $\mathbf{M}_*$ in future work.

Figure 9 characterizes this walk away from the resonant mapping, $\mathbf{M}_*$. During the first 45 steps in this walk, site 9 grows by stealing atoms from site 1. This results in the formation of new CG springs that couple site 9 to sites 1, 2, and 6. These new springs break the degeneracy of the lowest frequency modes. While the frequency of mode 1 remains near its initial value, the frequency of mode 2 rapidly increases. Consequently, the spectral quality decreases from $\mathcal{Q} = 0.0373$ to 0.0322. Nevertheless, the CG spring matrix, $\mathbf{K}(\mathbf{M}_{\mathcal{Q}45})$, preserves much of the original structure in $\mathbf{K}_*$ after 45 steps in this walk.

During the next 45 steps, site 9 continues to grow by stealing atoms from sites 3 and 7. This further strengthens the CG springs from site 9 to sites 2 and 6, while also introducing new CG springs from site 9 to sites 3, 5, and 7. During these steps, the frequencies of the first and third CG normal modes rapidly increase, while the spectral quality decreases to

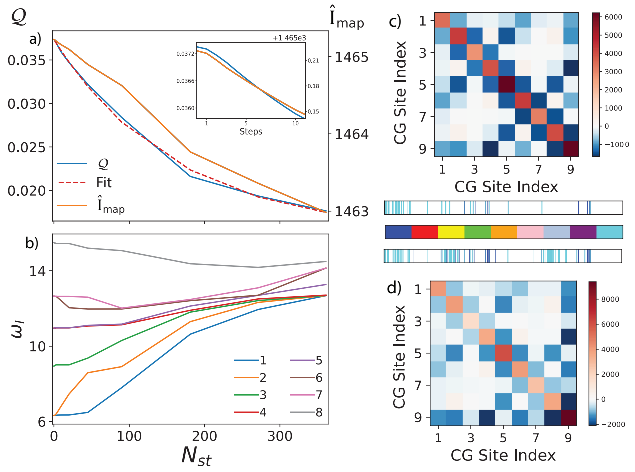

FIG. 9. Walk though mapping space by perturbing the resonant mapping, $\mathbf{M}_*$, to reduce $\mathcal{Q}$. Panel a presents $\mathcal{Q}$ (blue) and $\hat{I}_{\text{map}}$ (orange) as a function of the number of steps, N_{st}, on this walk. The dashed red curve in panel a presents an exponential fit to $\mathcal{Q}(N_{\text{st}})$. Panel b presents the corresponding 8 finite normal mode frequencies, ω_I, of the CG model. The legend indicates the order of the normal mode frequencies, $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_8$. Panels c and d present the CG spring matrix, $\mathbf{K}(\boldsymbol{\kappa}_*, \mathbf{M})$, after 45 and 90 steps along this walk. The middle horizontal color bar between panels c and d indicate the initial assignment of the 360 atoms into 9 CG sites, i.e., site I is initially associated with the I^{th} block of 40 amino acids. The horizontal color bars above and below this middle color bar indicate the assignment of atoms that have been moved from their initial partition during the first 45 and 90 steps, respectively, of this walk.

0.0284. After 90 steps, the CG spring matrix, $\mathbf{K}(\mathbf{M}_{\mathcal{Q}90})$, bears relatively little resemblance to $\mathbf{K}_*$.

As the walk proceeds further, site 9 continues to grow by stealing atoms from other sites. In these later stages of the walk, the unusually high CG frequency, ω_8, slightly decreases from ≈ 16 to ≈ 14. The frequencies of the other CG modes all systematically increase. All of the CG modes are in the high frequency range, $\omega_I \geq 10$, by the end of the walk. Interestingly, the spectral quality decreases less rapidly as the walk progresses. Consequently, as we move off resonance, $\mathcal{Q}(N_{\text{st}})$ appears to decay in a manner that is qualitatively similar to an exponential. Conversely, the adjusted mapping information, $\hat{\text{I}}_{\text{map}}$, decreases in a more nearly linear manner.

Figure 10 characterizes a second walk away from the resonant mapping, $\mathbf{M}_*$. Each step

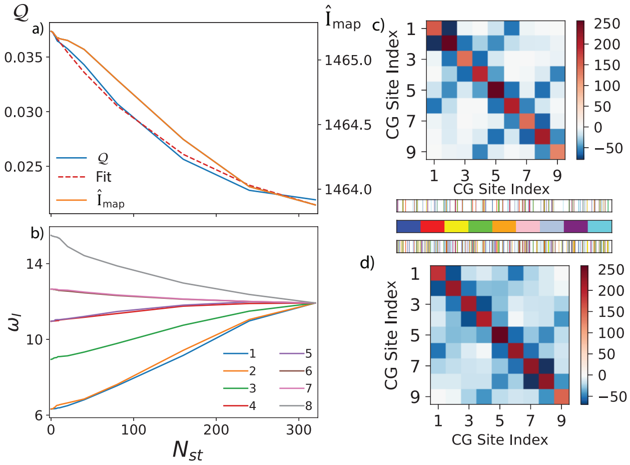

FIG. 10. Walk though mapping space by perturbing the resonant mapping, $\mathbf{M}_*$, to maximize $d(\mathbf{M}, \mathbf{M}_*)$. Panel a presents $\mathcal{Q}$ (blue) and $\hat{I}_{\mathrm{map}}$ (orange) as a function of the number of steps, N_{st}, on this walk. The dashed red curve in panel a presents an exponential fit to $\mathcal{Q}(N_{\mathrm{st}})$. Panel b presents the corresponding 8 finite normal mode frequencies, ω_I, of the CG model. The legend indicates the order of the normal mode frequencies, $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_8$. Panels c and d present the CG spring matrix, $\mathbf{K}(\boldsymbol{\kappa}_*, \mathbf{M})$, after 80 and 160 steps along this walk. The middle horizontal color bar between panels c and d indicate the initial assignment of the 360 atoms into 9 CG sites, i.e., site I is initially associated with the I^{th} block of 40 amino acids. The horizontal color bars above and below this middle color bar indicate the assignment of atoms that have been moved from their initial partition during the first 80 and 160 steps, respectively, of this walk.

of this walk selects the neighboring map, $\mathbf{M}_{dt+1}$, that is farthest from $\mathbf{M}_*$, while using the variation of information (VI) to define the distance, $d(\mathbf{M}, \mathbf{M}_*) = \mathrm{VI}(\mathbf{M}, \mathbf{M}_*)$, between maps based upon the similarity in the corresponding partitions. As the walk proceeds, the atoms appear to be "randomly" re-assigned among the 9 sites and the CG spring matrix, $\mathbf{K}(\boldsymbol{\kappa}_*, \mathbf{M})$, becomes increasingly blurred. The underlying CG spring matrix, $\mathbf{K}_*$, is easily visible after 80 steps, but has become much less clear after 160 steps. Interestingly, this walk preserves the degeneracy present in the density of states for the original CG spring matrix. However, the frequencies of the CG normal modes all converge towards a single high frequency, $\omega_I \approx 12$. This walk terminates when the final CG spring matrix connects all of the sites with similarly weak springs, which results in a final spectral quality of $\mathcal{Q}_{d\infty} \approx 0.0219$. The spectral quality again appears to exponentially decay along this walk.

To this point, we have considered the sensitivity of this resonance to the details of the

CG mapping, $\mathbf{M}$, for a given high resolution GNM with spring matrix, $\boldsymbol{\kappa}_*$. Figures 9 and 10 demonstrate that $\mathcal{Q}$ rapidly decreases as we move away from the resonant mapping, $\mathbf{M}_*$. Interestingly, though, the features of the original CG spring matrix, $\mathbf{K}_* = \mathbf{K}_*(\boldsymbol{\kappa}_*, \mathbf{M}_*)$, remain visible in the resulting CG spring matrix, $\mathbf{K}(\boldsymbol{\kappa}_*, \mathbf{M})$, even after 45 steps away from the resonant mapping, $\mathbf{M}_*$. We now briefly consider how robust this resonance is to the details of the AA spring matrix, $\boldsymbol{\kappa}$, for the fixed CG mapping, $\mathbf{M}_*$.

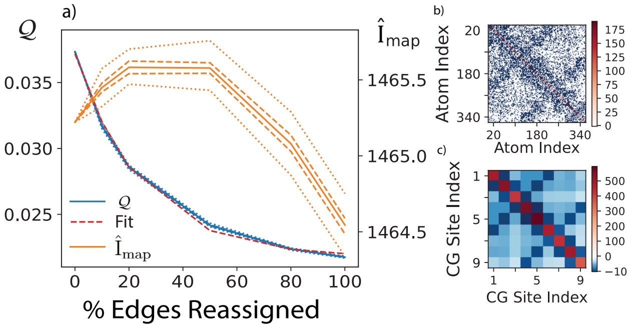

FIG. 11. Impact of randomly shuffling springs in the original high resolution spring matrix, $\boldsymbol{\kappa}_*$, according to the protocol of Sec. IV D. Panel a presents statistics for the spectral quality, $\mathcal{Q}(\mathbf{M}_*, f)$, and adjusted mapping information, $\hat{\mathrm{I}}_{\mathrm{map}}(\mathbf{M}_*, f)$, of the resonant mapping, $\mathbf{M}_*$, as a function of the fraction, f, of high-resolution springs that have been re-assigned. The red curve presents an exponential fit to $\mathcal{Q}(\mathbf{M}_*, f)$. The solid, dashed, and dotted lines present the mean, quartiles, and extrema obtained from 100 independent numerical experiments. Panel b presents a typical AA spring matrix, $\boldsymbol{\kappa}(f)$, when $f = 1/2$ of the original springs have been reassigned. Panel c presents the resulting CG spring matrix, $\mathbf{K}(f) = \mathbf{K}(\boldsymbol{\kappa}(f), \mathbf{M}_*)$.

We first consider the effect of randomly re-assigning a fraction, f, of the springs in the high resolution spring matrix, $\boldsymbol{\kappa}_*$, according to the protocol described in Sec. IV D. In order to identify statistically significant trends, we repeat this procedure 100 times for each f. Figure 11a demonstrates that $\mathcal{Q}(f)$ decays nearly exponentially with f. Interestingly, while $\mathcal{Q}$ appears quite sensitive to f, $\mathcal{Q}$ appears surprisingly insensitive to the identity of the reassigned springs. Conversely, $\hat{\mathrm{I}}_{\mathrm{map}}(f)$ initially increases as springs are randomly reassigned, but begins to decrease with f after half of the springs have been reassigned. Moreover, in comparison to $\mathcal{Q}$, $\hat{\mathrm{I}}_{\mathrm{map}}$ appears much more sensitive to the identity of the springs that are reshuffled.

Figure 11b presents the corresponding AA spring matrix $\boldsymbol{\kappa}(f)$ for $f = 0.50$ from one trial of this numerical experiment. Once half of the springs have been re-assigned, the spectral quality of the model has decreased from 0.037 to 0.024. At this point, the pattern of the original AA spring matrix is barely perceptible. As a consequence of randomly reassigning the AA springs, the connections in the CG spring matrix, $\mathbf{K}(f)$, are weaker. Moreover, $\mathbf{K}(f)$ now includes effective connections that were not present in $\mathbf{K}_*$. Nevertheless, Fig. 11c demonstrates that the $\mathbf{K}(f)$ remains quite similar to the underlying CG spring matrix, $\mathbf{K}_*$. Thus, it appears that this resonance remains quite robust with respect to even half of the AA springs being reassigned.

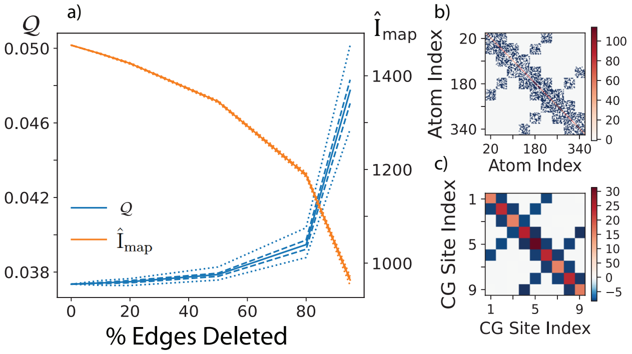

FIG. 12. Impact of randomly deleting springs in the original high resolution spring matrix, $\boldsymbol{\kappa}_*$, according to the protocol of Sec. IV D. Panel a presents statistics for the spectral quality, $\mathcal{Q}(\mathbf{M}_*, f)$, and adjusted mapping information, $\hat{\mathrm{I}}_{\mathrm{map}}(\mathbf{M}_*, f)$, of the resonant mapping, $\mathbf{M}_*$, as a function of the fraction, f, of high-resolution springs that have been deleted. The solid, dashed, and dotted lines present the mean, quartiles, and extrema obtained from 100 independent numerical experiments. Panel b presents a typical AA spring matrix, $\boldsymbol{\kappa}(f)$, when $f = 1/2$ of the original springs have been deleted. Panel c presents the resulting CG spring matrix, $\mathbf{K}(f) = \mathbf{K}(\boldsymbol{\kappa}(f), \mathbf{M}_*)$.

Finally, Figure 12 considers the effect of randomly deleting a fraction, f, of the springs in the high resolution spring matrix, $\boldsymbol{\kappa}_*$, according to the protocol described in Sec. IV D. In this case, $\mathcal{Q}(f)$ initially increases slowly as springs are deleted, but then increases more rapidly for $f \geq 0.50$. Conversely, $\mathrm{I}_{\mathrm{map}}(f)$ monotonically decreases increasingly rapidly as springs are deleted. In both cases, we expect that this reflects a general shift of the AA density of states to lower frequencies as springs are removed.

Figure 12b presents the AA spring matrix $\boldsymbol{\kappa}(f)$ from one trial after $f = 0.50$ of the springs have been deleted. In this case, the pattern of $\boldsymbol{\kappa}_*$ is unperturbed, although it is much fainter. Conversely, the CG Kirchoff matrix, $\mathbf{K}(f)$, in Figure 12c perfectly preserves the connectivity of $\mathbf{K}_*$, although the effective CG springs are much weaker.

VI. CONCLUSIONS

The mapping, $\mathbf{M}$, profoundly impacts CG models. In particular, $\mathbf{M}$ determines both the mapped distribution, $p_{\mathrm{R}}(\mathbf{R})$, of CG configurations and also the conditioned distribution, $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})$, describing the lost subensemble of AA configurations that map to each CG configuration, $\mathbf{R}$. The information content of this lost subensemble, $\mathrm{I}_{\mathrm{map}}(\mathbf{R})$, determines the degeneracy of AA configurations that map to $\mathbf{R}$ and, thus, governs the physical significance and computational feasibility of back-mapping approaches.[11,99,100] Moreover, $\mathrm{I}_{\mathrm{map}}(\mathbf{R})$ determines both the entropic component and also the temperature-dependence of the exact CG potential, $W(\mathbf{R})$. In particular, any estimates of thermodynamic energies or entropies with CG models should account for $\mathrm{I}_{\mathrm{map}}(\mathbf{R})$.

Accordingly, in this work we have investigated the relationship between the mapping and the CG model. Our analysis of the mapping identifies a simple back-mapping operator and a corresponding projection operator for relating the motion of AA and CG models. This analysis also provides a simple partitioning of the AA configuration space into CG and intra-site subspaces. In order to preserve translational motion between AA and CG models, the coefficients defining $\mathbf{M}$ must sum to 1 for each site, i.e., they must be L^1 normalized. Consequently, the mapped distribution and the PMF must both be invariant with respect to any translational or rotational symmetries present in the AA model. More generally, Appendix A demonstrates that the mapped distribution and PMF will be invariant with respect to any symmetry operator, $\hat{T}$, that commutes with the mapping, $\mathbf{M}$.

The partitioning of AA coordinates implies a formal partitioning of the underlying high resolution potential into CG, intra-site, and coupling components. In the case of linear models, one can readily see how the coupling between CG and intra-site coordinates impacts the mapped ensemble and the CG effective potential.[32] This partitioning suggests the general possibility of "resonant" mappings that eliminate the coupling between the CG and intra-site coordinates. More generally, resonant mappings arise when the AA potential can be separated into independent, additive contributions governing the CG and intra-site coordinates. In this case, the intra-site interactions do not impact either the mapped ensemble or the CG effective potential. Consequently, resonant mappings seem like an idealization of "perfect" coarse-graining. These considerations are certainly not new and perhaps intuitively obvious. Nevertheless, the present work hopefully provides additional insight.

Because the mapping coefficients are L^1 normalized, the partitioning into CG and intra-site coordinates introduces a nontrivial Jacobian factor that equals the determinant of the participation matrix, $||\Delta_N||$. This Jacobian determines a "labelling entropy," $\mathrm{H_L} = \ln ||\Delta_N||$, that systematically increases as the CG sites become increasingly uniform in size. Because it quantifies the uncertainty in the atoms associated with each CG site, $\mathrm{H_L}$ effectively reduces the information content of the conditioned distribution, $p_{\mathrm{r|R}}$, describing the lost ensemble. Conversely, $\mathrm{H_L}$ effectively increases the information content of the mapped ensemble. While the labelling entropy is perhaps unexpected, we show that it can be qualitatively motivated by simple statistical considerations for weakly correlated CG coordinates. Moreover, $\mathrm{H_L}$ can be explicitly derived for linear models as an identity relating the determinants of the Hessian matrices describing the AA and CG models. We speculate that $\mathrm{H_L}$ may arise naturally in a coarse-graining formalism that explicitly treated the indistinguishability of equivalent particles.

We numerically illustrated these considerations with a Gaussian Network Model (GNM) for the equilibrium fluctuations of actin about its folded conformations. Our calculations indicated that the spectral quality, $\mathcal{Q}$, provides a good metric for identifying CG representations that are consistent with our physical intuition. Since it attempts to preserve low-frequency, large-amplitude motions, the spectral quality is qualitatively similar to many metrics that have been previously developed for identifying rigid protein domains that move coherently.[29,31,32,42–52,54,101–103] Representations with high spectral quality associate CG sites with compact, highly connected atomic groups that generate broad mapped ensembles because the sites are weakly constrained by relatively few inter-site bonds. In particular, the 4-site representation, $\mathbf{M}_{\mathcal{Q}}$, that maximized $\mathcal{Q}$ aligns very nicely with the four rigid domains identified by the Spectrus webserver,[98] as well as the four domains that have been previously identified in the biochemical literature.[93] Moreover, this representation appears quite similar to the 4-site representation that was identified by combining the ED-CG method with K-means clustering to analyze microsecond molecular dynamics simulations of an AA model for actin. In comparison, minimal resources are required to identify $\mathbf{M}_{\mathcal{Q}}$ via steepest descent of $\mathcal{Q}$ for the actin GNM. Thus, we anticipate that $\mathcal{Q}$ may be a useful metric for identifying high quality CG representations of systems that fluctuate about an equilibrium conformation. We anticipate that it may be possible to generalize $\mathcal{Q}$ for more complex systems that transition between multiple conformations by generalizing the Rayleigh-type quotient of Eq. (63) or by considering linear discriminant analsysis.[104]

In contrast, neither minimizing nor maximizing the mapping information loss, $\mathrm{I}_{\mathrm{map}}$, identifies CG representations that are consistent with our physical intuition. Representations that minimize $\mathrm{I}_{\mathrm{map}}$ associate CG sites with diffuse, interspersed atomic groups. The resulting mapped ensemble is very narrow and, thus, highly informative because the sites are highly constrained by many inter-site bonds. Conversely, maps that maximize $\mathrm{I}_{\mathrm{map}}$ tend to represent the large majority of the protein with a single site, while associating the remaining sites with individual residues. These maps generate very broad mapped ensembles by minimizing the number of inter-site bonds, but also minimize the labelling entropy. By accounting for the labelling entropy, the adjusted information loss, $\hat{\mathrm{I}}_{\mathrm{map}} = \mathrm{I}_{\mathrm{map}} + \mathrm{H}_{\mathrm{L}}$, correlates very well with the spectral quality, $\mathcal{Q}$, for the present CG representations of the GNM.

We also numerically illustrated a notion of resonance between an AA model and a CG mapping. In this case, we specified an underlying CG spring matrix, $\mathbf{K}_*$, and a corresponding CG mapping, $\mathbf{M}_*$. We then atomized $\mathbf{K}_*$ in order to determine a family of AA spring matrice, $\{\boldsymbol{\kappa}_*\}$, that are all resonant with $\mathbf{M}_*$. By construction, the resonant mapping perfectly preserved the low frequency modes of the AA spring matrix, while eliminating the irrelevant high resolution details. Given a fixed AA spring matrix, $\boldsymbol{\kappa}_*$, the spectral quality, $\mathcal{Q}(\mathbf{M})$, exponentially decreased as the mapping, $\mathbf{M}$, moved away from resonance. Nevertheless, the CG spring matrix, $\mathbf{K}(\mathbf{M})$, remained quite similar to the underlying spring matrix, $\mathbf{K}_*$, even after 45 steps away from resonance. Conversely, given the fixed CG mapping, $\mathbf{M}_*$, the spectral quality, $\mathcal{Q}(f)$, exponentially decreased with the fraction, f, of springs that were randomly reassigned from the original AA spring matrix, $\boldsymbol{\kappa}_*$. Interestingly, the CG spring matrix, $\mathbf{K}(\boldsymbol{\kappa})$, remained quite similar to the original underlying CG spring matrix, $\mathbf{K}_*$, even when half of the original AA springs were randomly reassigned.

Of course, we do not anticipate finding a perfect resonance when coarse-graining soft materials. Given a realistic high resolution model, it may be possible to identify nearly resonant mappings by minimizing the memory kernel describing the dynamics of the CG variables, as suggested by Voth, Dinner, and coworkers.[105] The present results suggest that the spectral quality may also be a particularly simple metric for finding nearly resonant mappings. Moreover, the present results suggest that the idealized CG representation of the system remains visible rather far from resonance. Thus, systematic coarse-graining may generate "sloppy" models that preserve robust, underlying features that are often obscured by high resolution details.[67]

The present work also indicates many directions for future work. While the GNM provides a qualitatively reasonable description for equilibrium fluctuations about a single free energy minimum, it has many significant limitations. For instance, the present GNM considers a single energy scale, a single mass scale, a single length scale, and, most importantly, a single free energy minima. Clearly, future investigations should investigate the impact of the mapping upon more complex models. In particular, it will be interesting to generalize $\mathcal{Q}$ for more complex models that transition between multiple free energy minimum. We anticipate that it may be useful to explore the relationship between $\mathcal{Q}$ and the VAMP score employed in Markov state models.[47,90–92] Similarly, it will be interesting to investigate the importance of $\mathrm{H_L}$ and $\mathrm{I_{map}}$ for systems of interacting molecules and for systems with multiple mass, length, and energy scales. Finally, it would also be interesting to consider the ramifications of the present transformation for modeling dynamical properties.[106–109] Nevertheless, we hope that this work may provide useful insight for considering the mapping and how it influences the properties of CG models.

SUPPLEMENTARY MATERIAL

See the supplementary material for additional results and analysis, including an explicit illustration of the dual basis introduced in Sec. II C and the derivation of the identity, Eq. (61).

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). KMK and WGN also gratefully acknowledge very fruitful collaborations with M. Scott Shell that lead to this work. WGN gratefully acknowledges enlightening conversations with Xiantao Li, Raffaello Potestio, and Markus Deserno, as well as helpful discussions with Nick Jackson, Ben Curlee, Sanchay Agarwal, Lucus Mussi, and Maria Lesniewski. 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),[110] 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.[111] VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of 2510 Illinois at Urbana-Champaign. We employed Matplotlib for all figures[112] and performed all matrix calculations with NumPy.[113,114]

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 A: MAPPING SYMMETRIES TO COARSE-GRAINED RESOLUTION

We briefly consider the impact of the CG mapping upon symmetries that are present in an AA model. Specifically, we consider symmetries that correspond to an operator, $\hat{T}$, that acts as a bijective, volume-preserving mapping of the AA configuration space $\mathcal{D}^n(V)$ onto itself. We also assume that a corresponding operator exists on the CG configuration space,

$\mathcal{D}^N(V)$.

Any symmetry, $\hat{T}$, of the AA potential for which $u(\mathbf{r}) = u(\hat{T}\mathbf{r})$ will leave the mapped ensemble, $p_{\mathrm{R}}(\mathbf{R})$, and, consequently, the PMF, $W(\mathbf{R})$, invariant as long as the AA model ergodically samples configuration space and $\hat{T}$ commutes with the mapping operator, $\mathbf{M}\hat{T} = \hat{T}\mathbf{M}$.[115] This follows because

$$\begin{aligned} z_{\mathrm{R}}(\mathbf{R}) &\equiv \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r})]\delta(\mathbf{M}\mathbf{r} - \mathbf{R}) \\ &= \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r} \exp[-\beta u(\hat{T}\mathbf{r})]\delta(\mathbf{M}\mathbf{r} - \mathbf{R}) \\ &= \int_{\hat{T}\mathcal{D}^n(V)} |\hat{T}^{-1}| \mathrm{d}\mathbf{r}' \exp[-\beta u(\mathbf{r}')]\delta(\mathbf{M}\hat{T}^{-1}\mathbf{r}' - \mathbf{R}) \\ &= \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r}' \exp[-\beta u(\mathbf{r}')]\delta(\mathbf{M}\hat{T}^{-1}\mathbf{r}' - \mathbf{R}) \\ &= \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r}' \exp[-\beta u(\mathbf{r}')]\delta(\hat{T}^{-1}\mathbf{M}\mathbf{r}' - \mathbf{R}) \\ &= \int_{\mathcal{D}^n(V)} \mathrm{d}\mathbf{r}' \exp[-\beta u(\mathbf{r}')]\delta(\mathbf{M}\mathbf{r}' - \hat{T}\mathbf{R}) = z_{\mathrm{R}}(\hat{T}\mathbf{R}) \end{aligned} \tag{A1}$$

The second line follows because $u(\hat{T}\mathbf{r}) = u(\mathbf{r})$ for the symmetry operator, $\hat{T}$. The third line follows by transforming variables $\mathbf{r} \to \mathbf{r}' = \hat{T}\mathbf{r}$, while the fourth line relies upon the symmetry being volume preserving and bijective. The fifth line follows because we have assumed that $\hat{T}$ and $\mathbf{M}$ commute, while the sixth line follows because $\hat{T}$ is bijective and volume preserving. Of course, symmetries that are present in the AA potential may be broken because the boundary conditions of the AA model are not consistent with the symmetry (e.g., periodic boundary conditions are not commensurate with rotational symmetry) or because the AA model does not ergodically sample configuration space (e.g., simulations of lipid bilayers break symmetry).

Here we focus on translational and rotational symmetries in $D = 3$ dimensions because they are most commonly relevant to AA models. For each Cartesian direction, α, and each distance, d, we define a translational symmetry operator, $\hat{T}_{\mathrm{tr};\alpha}(d)\mathbf{r} \to \mathbf{r} + d\mathbf{J}_n \otimes \mathbf{e}_\alpha$, that displaces each atom a distance d along $\mathbf{e}_\alpha$. Similarly, we define a rotational symmetry operator, $\hat{T}_{\mathrm{rot};\alpha}(\theta)\mathbf{r} = \sum_{i=1}^{n} \mathbf{e}_i \otimes \mathbf{\Omega}_\alpha(\theta)\mathbf{r}_i$, where $\mathbf{\Omega}_\alpha(\theta) = \exp[\mathbf{G}_\alpha\theta]$ corresponds to the $D \times D$ matrix describing a rotation of θ about the α Cartesian axis and $\mathbf{G}_\alpha$ is the corresponding generator for infinitesimal rotations.[116] These continuous symmetry operators define an infinitesimal displacement, $\boldsymbol{\eta}t$, as $t \to 0$, such that

$$\hat{T}(t)\mathbf{r} \xrightarrow{t\to 0} \mathbf{r} + \boldsymbol{\eta}t + \mathcal{O}(t^2). \tag{A2}$$

In particular, $\boldsymbol{\eta}_{\text{tr};\alpha} = \mathbf{J}_n \otimes \mathbf{e}_\alpha$ and $\boldsymbol{\eta}_{\text{rot};\alpha} = \sum_{i=1}^n \mathbf{e}_i \otimes \mathbf{G}_\alpha \mathbf{r}_i^*$ correspond to infinitesimal translational and rotational displacements about the minimum of the AA potential, $\mathbf{r}^*$. Assuming that the AA potential is invariant with respect to the continuous symmetry operator, $\hat{T}(t)$, it then follows that

$$u(\mathbf{r}^*) = u(\hat{T}(t)\mathbf{r}^*) \xrightarrow{t\to 0} u(\mathbf{r}^*) + \frac{1}{2}\boldsymbol{\eta}^\dagger \mathbf{h} \boldsymbol{\eta} t^2 + \mathcal{O}(t^3) \tag{A3}$$

where $\mathbf{h} \equiv \mathbf{h}(\mathbf{r}^*)$ is the Hessian of the AA potential about its minimum. Since this identity holds for all $t \to 0$, it follows that the symmetry operator, $\hat{T}$, defines an element, $\boldsymbol{\eta}$ in the nullspace of $\mathbf{h}$. We assume that the only symmetries of the AA potential correspond to uniform rotations and translations, such that $\text{null}(\mathbf{h}) = \text{span}\{\boldsymbol{\eta}_{\text{tr};\alpha}, \boldsymbol{\eta}_{\text{rot};\alpha}\}_{\alpha=1,2,3}$.

Note that both translational and rotational symmetry operators commute with the CG mapping, $\mathbf{M}$. In the following, we will neglect the effect of boundary conditions and assume that these operators provide a bijective, volume-preserving mapping of the configuration space. Consequently, Eq. (A1) implies that the PMF will be invariant with respect to uniform rotations and translations. Therefore, $\overline{\boldsymbol{\eta}}_{\text{tr};\alpha} \equiv \mathbf{M}\boldsymbol{\eta}_{\text{tr};\alpha} = \mathbf{J}_N \otimes \mathbf{e}_\alpha$ and $\overline{\boldsymbol{\eta}}_{\text{rot};\alpha} \equiv \mathbf{M}\boldsymbol{\eta}_{\text{rot};\alpha} = \sum_{I=1}^N \mathbf{e}_I \otimes \mathbf{G}_\alpha \mathbf{R}_I^*$ lie in the nullspace of the CG Hessian, $\mathbf{H}$. In most cases, one expects that the corresponding set of 6 vectors, $\{\overline{\boldsymbol{\eta}}_{\text{tr};\alpha}, \overline{\boldsymbol{\eta}}_{\text{rot};\alpha}\}_{\alpha=1,2,3}$, will be linearly independent. However, in certain special cases, e.g., for $N = 2$ site representations, the three mapped rotational eigenvectors, $\{\overline{\boldsymbol{\eta}}_{\text{rot};\alpha}\}$, can become linearly dependent such that the dimensionality of null($\mathbf{H}$) may be smaller than the dimensionality of null($\mathbf{h}$).

The argument that leads to Eq. (A1) also holds for more general symmetries, e.g., particle permutations or discrete rotations about bonds. In this case, one may need to more carefully distinguish the action of the symmetry upon the AA and CG configurations, $\hat{T}_{\text{AA}}$ and $\hat{T}_{\text{CG}}$, respectively. In particular, if $\mathbf{M}$ "coarse-grains" over an AA symmetry, $\hat{T}_{\text{AA}}$, then one expects that the corresponding CG symmetry operator simply reduces to an identity operator, $\hat{T}_{\text{CG}} = \hat{\mathbb{1}}$. Conversely, if $\mathbf{M}$ is not commensurate with an AA symmetry, $\hat{T}_{\text{AA}}$, then the CG model may preserve a remnant of the symmetry, $\hat{T}_{\text{CG}}$, that satisfies, $\hat{T}_{\text{CG}}\mathbf{M} = \mathbf{M}\hat{T}_{\text{AA}}$, such that the mapped distribution is invariant with respect to $\hat{T}_{\text{CG}}$. However, $\hat{T}_{\text{AA}}$ and $\hat{T}_{\text{CG}}$ may have rather different forms.

APPENDIX B: ANALYSIS OF THE JACOBIAN $||\mathbf{Z}||$

We are interested in the determinant of the matrix, $\mathbf{Z} = [\mathbf{z}_i]$, where $\{\mathbf{z}_i\} = \{\mathbf{c}_I, \mathbf{z}_{N+k}\}$ is a set of n linearly independent vectors that span $\mathcal{V}_{\mathrm{AA}} \sim \mathbb{R}^n$. We have defined the $n_{\mathrm{x}} = n - N$ vectors $\{\mathbf{z}_{N+k}\}$ to be orthonormal $\mathbf{z}_{N+k}^{\dagger}\mathbf{z}_{N+k'} = \delta_{kk'}$. Moreover, the set $\{\mathbf{z}_i\}$ are dual to the n linearly independent vectors $\{\mathbf{x}_i\} = \{\mathbf{j}_I, \mathbf{x}_{N+k}\}$ such that $\mathbf{z}_i^{\dagger}\mathbf{x}_j = \delta_{ij}$ for all $i, j = 1, \dots n$. In particular, this implies that

$$\mathbf{z}_{N+k}^{\dagger}\mathbf{j}_I = 0 \qquad \text{for all } k = 1, \dots, n_{\mathrm{x}}; I = 1, \dots, N$$

We define

$$\mathbf{z}_{\varnothing I} = n_I^{-1/2}\mathbf{j}_I$$

for $I = 1, \dots, N$. Because the mapping corresponds to disjoint atomic groups it follows that $\mathbf{z}_{\varnothing I}^{\dagger}\mathbf{z}_{\varnothing J} = \delta_{IJ}$ for all $I, J = 1, \dots, N$. Consequently, the set $\{\mathbf{z}_{\varnothing I}, \mathbf{z}_{N+k}\}$ forms a complete orthonormal basis for $\mathcal{V}_{\mathrm{AA}}$. Since $\mathbf{z}_{\varnothing I}^{\dagger}\mathbf{c}_J = n_I^{-1/2}\delta_{IJ}$, it follows that

$$\mathbf{c}_I = n_I^{-1/2}\mathbf{z}_{\varnothing I} + \sum_{k=1}^{n_{\mathrm{x}}} \gamma_{Ik}\mathbf{z}_{N+k},$$

where $\gamma_{Ik} = \mathbf{z}_{N+k}^{\dagger}\mathbf{c}_I$. This decomposition allows us to determine the desired determinant:

$$\begin{aligned} ||\mathbf{Z}|| = \left|\left|\mathbf{Z}^{\dagger}\right|\right| &= \left|\left| \begin{matrix} n_I^{-1/2}\mathbf{z}_{\varnothing I}^{\dagger} + \sum_{k=1}^{n_{\mathrm{x}}} \gamma_{Ik}\mathbf{z}_{N+k}^{\dagger} \\ \mathbf{z}_{N+k'}^{\dagger} \end{matrix} \right|\right| \\ &= \left|\left| \begin{matrix} n_I^{-1/2}\mathbf{z}_{\varnothing I}^{\dagger} \\ \mathbf{z}_{N+k'}^{\dagger} \end{matrix} \right|\right| = \left(\prod_{I=1}^{N} n_I^{-1/2}\right) \left|\left| \begin{matrix} \mathbf{z}_{\varnothing I}^{\dagger} \\ \mathbf{z}_{N+k'}^{\dagger} \end{matrix} \right|\right| \\ &= \left(\prod_{I=1}^{N} n_I^{-1/2}\right). \end{aligned}$$

The second line follows because determinants are unchanged by the addition of rows, while the third row follows because $\{\mathbf{z}_{\varnothing I}, \mathbf{z}_{N+k}\}$ form a complete orthonormal basis. We then have the desired result:

$$||\mathbf{Z}||^{-1} = ||\Delta_N||^{1/2} \tag{B1}$$

where $\Delta_N = \sum_{I=1}^{N} \mathbf{e}_I n_I \mathbf{e}_I^{\dagger}$ is a diagonal participation matrix.

APPENDIX C: COMPARISON WITH ED-CG METHOD

Here we briefly compare the ED-CG metric with the spectral quality. The ED-CG method was originally developed for coarse-graining simulations of complex biomolecules.[48] Let $\mathbf{r}(t) = \{\mathbf{r}_1(t), \ldots, \mathbf{r}_n(t)\}$ be the coordinates for n atoms in D dimensions at time t after eliminating overall rotational and translational motion. Given n_t configurations, we define the mean position of atom i by $\bar{\mathbf{r}}_i \equiv n_t^{-1} \sum_{t=1}^{n_t} \mathbf{r}_i(t)$ and the displacement by $\Delta\mathbf{r}_i(t) = \mathbf{r}_i(t) - \bar{\mathbf{r}}_i$. We define the covariance matrix by

$$\mathbf{C} \equiv \frac{1}{n_t} \sum_{t=1}^{n_t} \Delta\mathbf{r}(t) \Delta\mathbf{r}^\dagger(t) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\dagger, \tag{C1}$$

where the last expression is the SVD decomposition of $\mathbf{C}$: $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_{nD}\}$ is a diagonal matrix of eigenvalues that are sorted in decreasing order (i.e., $\lambda_1 \geq \lambda_2 \geq \cdots$) and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{nD}]$ is the matrix of corresponding eigenvectors. The essential dynamics subspace[53] is defined by the first n_{ED} eigenvectors of $\mathbf{C}$: $\mathbf{U}_{\mathrm{ED}} = [\mathbf{u}_1, \ldots, \mathbf{u}_{n_{\mathrm{ED}}}]$. We define $\Delta\mathbf{r}_{\mathrm{ED}}(t) = \mathbf{U}_{\mathrm{ED}}^\dagger \Delta\mathbf{r}(t)$ and $\mathbf{C}_{\mathrm{ED}} = \mathbf{U}_{\mathrm{ED}}\mathbf{C}\mathbf{U}_{\mathrm{ED}}^\dagger$ as projections onto this subspace. Given a mapping, $\mathbf{M}$, that partitions the n atoms into N disjoint atomic groups, $V_1, \ldots, V_N$, the ED-CG metric may be expressed

$$\chi^2(\mathbf{M}) \equiv \frac{1}{n_t} \sum_{t=1}^{n_t} \frac{1}{ND} \sum_{I=1}^{N} \sum_{i \in V_I} \sum_{j(\geq i) \in V_I} \left| \Delta\mathbf{r}_{i;\mathrm{ED}}(t) - \Delta\mathbf{r}_{j;\mathrm{ED}}(t) \right|^2 \tag{C2}$$

$$= \frac{1}{ND} \sum_{I=1}^{N} \sum_{i \in V_I} \sum_{j(\geq i) \in V_I} \left\{ C_{\mathrm{ED};ii} - 2C_{\mathrm{ED};ij} + C_{\mathrm{ED};jj} \right\}. \tag{C3}$$

In Eq. (C3) $C_{\mathrm{ED};ij} \equiv \sum_{\alpha=1}^{D} C_{\mathrm{ED};i\alpha|j\alpha}$ traces $\mathbf{C}_{\mathrm{ED}}$ over Cartesian directions. The ED-CG method identifies the optimal map by minimizing χ^2. According to Eq. (C2), the ED-CG method attempts to define CG sites that correspond to atomic groups that move rigidly within the ED subspace.

In the case of linear network models, we can analytically evaluate $\chi^2(\mathbf{M})$ from Eq. (C3). In the following calculations, we define the ED-CG subspace by the first $n_{\mathrm{ED}} = 10$ eigenvectors. Once C_{ED} has been determined, calculating $\chi^2(\mathbf{M})$ requires approximately half the time of computing $\mathcal{Q}(\mathbf{M})$ for the maps that we consider in this work.

Figure 13a presents a scatter plot comparing $\chi^2(\mathbf{M})$ and $\mathcal{Q}(\mathbf{M})$ for sampled N-site maps of actin. As expected, the spectral quality and ED-CG metrics are anti-correlated. We

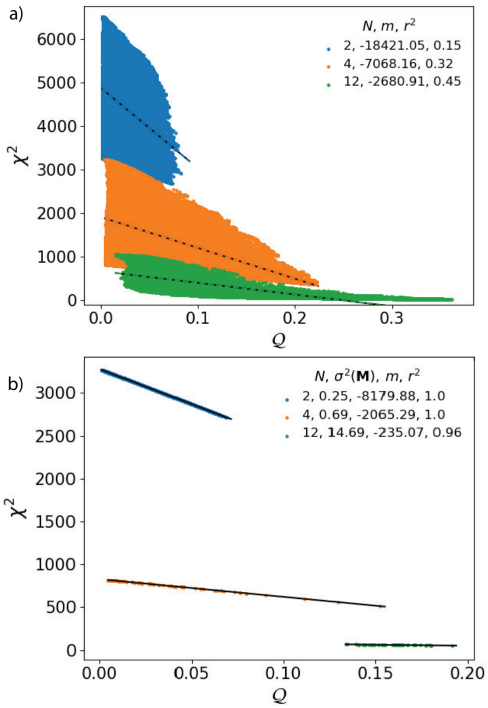

FIG. 13. Scatter plot of $\mathcal{Q}(\mathbf{M})$ and $\chi^2(\mathbf{M})$ for CG representations of actin sampled by MC simulations in mapping space. Blue, orange, and green points correspond to $N = 2$, 4, and 12-site representations, respectively. The dashed lines indicate best fit lines to the scatter plots. The legend indicates the slopes, m, and quality of fit parameters, r^2, for these lines. Panel (a) presents a scatter plot for all sampled maps at each resolution. Panel (b) presents sampled maps that are nearly uniform with the specified variance, $\sigma^2(\mathbf{M})$.

intuitively expect that sites corresponding to rigid atomic groups (i.e., relatively low χ^2) will tend to undergo relatively large amplitude motion (i.e., relatively high $\mathcal{Q}$). However, Fig. 13a demonstrates that this (anti-)correlation is quite weak when considered across the entirety of mapping space.

Figure 13b presents a scatter plot for a subset of the sampled maps at each resolution with minimal site-size variance, $\sigma^2(\mathbf{M}) = \mathrm{var}\{n_I\}$, where n_I is the number of residues that $\mathbf{M}$ associates with site I. Among these nearly uniform maps, χ^2 and $\mathcal{Q}$ are nearly perfectly (anti-)correlated.

We observed similar trends in our prior study of mapping space for ubiquitin.[63] We proposed there that Eq. (C3) can be used to rationalize these trends. First, note that each term, $\chi^2_{ij} \equiv \{C_{\mathrm{ED};ii} - 2C_{\mathrm{ED};ij} + C_{\mathrm{ED};jj}\}$, in Eq. (C3) is large and positive because the diagonal elements of the covariance matrix tend to be much larger than off-diagonal elements. Moreover, we note that Eq. (C3) contains $W(\mathbf{M}) = \frac{1}{2}N \times n_I \times (n_I - 1)$ such terms. Importantly, $W(\mathbf{M})$ grows linearly with $\sigma^2(\mathbf{M})$. For these reasons, χ^2 may tend to favor nearly uniform maps with small $\sigma^2(\mathbf{M})$.

REFERENCES

[1] R. W. Hamming. *Numerical methods for scientists and engineers.* McGraw-Hill, New York, 1962.

[2] C. Peter and K. Kremer. Multiscale simulation of soft matter systems. *Faraday Discuss.*, 144:9–24, 2010.

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

[4] 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.

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

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

[7] M. Muller, K. Katsov, and M. Schick. Biological and synthetic membranes: What can be learned from a coarse-grained description? *Phys. Rep.*, 434(5-6):113–176, 2006.

[8] M. Deserno. Mesoscopic membrane physics: Concepts, simulations, and selected applications. *Macromol. Rapid Comm.*, 30(9-10):752–771, 2009.

[9] F. Schmid. Toy amphiphiles on the computer: What can we learn from generic models? *Macromol. Rapid Comm.*, 30(9-10):741–751, 2009.

[10] 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.

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

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

[13] Helgi I. Ingólfsson, Cesar A. Lopez, Jaakko J. Uusitalo, Djurre H. de Jong, Srinivasa M. Gopal, Xavier Periole, and Siewert J. Marrink. The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 4(3):225–248, 2014.

[14] Luís Borges-Araújo, Ilias Patmanidis, Akhil P. Singh, Lucianna H. S. Santos, Adam K. Sieradzan, Stefano Vanni, Cezary Czaplewski, Sergio Pantano, Wataru Shinoda, Luca Monticelli, Adam Liwo, Siewert J. Marrink, and Paulo C. T. Souza. Pragmatic Coarse-Graining of Proteins: Models and Applications. *Journal of Chemical Theory and Computation*, 19(20):7112–7135, October 2023.

[15] 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.

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

[17] 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.

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

[19] 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.

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

bles and helix-coil transitions. *J. Chem. Theory Comput.*, 11(3):1278–1291, 2015.

[24] 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.

[25] 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.

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

[27] 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.

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

[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] J Charlie Maier, Chun-I Wang, and Nicholas E Jackson. Distilling coarse-grained representations of molecular electronic structure with continuously gated message passing. *The Journal of Chemical Physics*, 160(2), 2024.

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

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] Jiangbo Wu, Weizhi Xue, and Gregory A Voth. K-means clustering coarse-graining (kmc-cg): A next generation methodology for determining optimal coarse-grained mappings of large biomolecules. *Journal of Chemical Theory and Computation*, 19(23):8987–8997, 2023.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] M. Scott Shell. The relative entropy is fundamental to multiscale and inverse thermody-

namic problems. *J. Chem. Phys.*, 129:144108, 2008.

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

[61] 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.

[62] 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.

[63] Katherine M. Kidder, M. Scott Shell, and W. G. Noid. Surveying the energy landscape of coarse-grained mappings. *J. Chem. Phys.*, 160(5):054105, February 2024.

[64] 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.

[65] Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, 79:3090–3093, Oct 1997.

[66] 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.

[67] Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys.*, 143(1):010901, July 2015.

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

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

[70] Mark E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation.* Oxford University Press, Oxford, Great Britain, 2013.

[71] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128:244114, 2008.

[72] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3(5):300–313, 1935.

[73] Christos N. Likos. Effective interactions in soft condensed matter physics. *Phys. Rep.*, 348(4–5):267 – 439, 2001.

[74] Reinier L. C. Akkermans and W. J. Briels. A structure-based coarse-grained model for polymer melts. *J. Chem. Phys.*, 114(2):1020–1031, 2001.

[75] Nicholas J. H. Dunn, Thomas T. Foley, and William G. Noid. van der waals perspective on coarse-graining: progress toward solving representability and transferability problems. *Acc. Chem. Res.*, 49(12):2832–2840, 2016.

[76] Vagelis Harmandaris, Evangelia Kalligiannaki, Markos Katsoulakis, and Petr Plecháăc. Path-space variational inference for non-equilibrium coarse-grained systems. *J. Comp. Phys.*, 314:355–383, June 2016.

[77] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 1 edition, 2000.

[78] Benoit Roux and Thomas Simonson. Implicit solvent models. *Biophysical Chemistry*, 78(1-2):1–20, App. Phys. Rev. 1999.

[79] Q. Cui and I. Bahar, editors. *Normal mode analysis: Theory and applications to biological and chemical systems*. Chapman & Hall: CRC Press, Boca Raton, FL USA, 2006.

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

[81] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.

[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] Diane Valérie Ouellette. Schur complements and statistics. *Linear Algebra and its Applications*, 36:187–295, March 1981.

[84] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 1996.

[85] 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.

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

[87] Dragos Cvetkovic, Peter Rowlinson, and Slobodan Simic. *An Introduction to the Theory of Graph Spectra*. London Mathematical Society Student Texts. Cambridge University

Press, 2009.

[88] Richard P. Stanley. *Enumerative Combinatorics*, volume 2. Cambridge University Press, 1999.

[89] Art M. Duval, Caroline J. Klivans, and Jeremy L. Martin. Simplicial matrix-tree theorems. *Transactions of the American Mathematical Society*, 361(11):6073–6114, June 2009.

[90] Frank Noé and Feliks Nüske. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Modeling & Simulation*, 11(2):635–655, January 2013.

[91] Feliks Nüske, Lorenzo Boninsegna, and Cecilia Clementi. Coarse-graining molecular systems by spectral matching. *The Journal of Chemical Physics*, 151(4):044116, July 2019.

[92] Hao Wu and Frank Noé. Variational Approach for Learning Markov Processes from Time Series Data. *Journal of Nonlinear Science*, 30(1):23–66, February 2020.

[93] Ludovic R Otterbein, Philip Graceffa, and Roberto Dominguez. The crystal structure of uncomplexed actin in the adp state. *Science*, 293(5530):708–711, 2001.

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

[95] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

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

[97] Wolfgang Kabsch, Hans Georg Mannherz, Dietrich Suck, Emil F Pai, and Kenneth C Holmes. Atomic structure of the actin: Dnase i complex. *Nature*, 347(6288):37–44, 1990.

[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] Marc Stieffenhofer, Michael Wand, and Tristan Bereau. Adversarial Reverse Mapping of Equilibrated Condensed-Phase Molecular Structures. *Machine Learning: Science and Technology*, 1:045014, 2020.

[100] Wujie Wang, Minkai Xu, Chen Cai, Benjamin Kurt Miller, Tess Smidt, Yusu Wang, Jian Tang, and Rafael Gómez-Bombarelli. Generative coarse-graining of molecular conforma-

tions, 2022.

[101] 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.

[102] 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.

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

[104] Subarna Sasmal, Martin McCullagh, and Glen M. Hocky. Reaction Coordinates for Conformational Transitions Using Linear Discriminant Analysis on Positions. *Journal of Chemical Theory and Computation*, 19(14):4427–4435, July 2023.

[105] Nicholas Guttenberg, James F Dama, Marissa G Saunders, Gregory A Voth, Jonathan Weare, and Aaron R Dinner. Minimizing memory as an objective for coarse-graining. *J. Chem. Phys.*, 138(9):094111, Mar 2013.

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

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

[108] 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.

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

[110] 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.

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

[112] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engi-*

neering, 9(3):90–95, 2007.

[113] Travis E. Oliphant. Python for scientific computing. *Comput. Sci. Eng.*, 9:10–20, 2007.

[114] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

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

[116] W. Appel. *Mathematics for Physics and Physicists*. Princeton University Press, 2007.