Title

Exploring the landscape of model representations

Journal

Proceedings of the National Academy of Sciences of the United States of America, 117(39)

ISSN

0027-8424

Authors

Foley, Thomas T
Kidder, Katherine M
Shell, M Scott
et al.

Publication Date

2020-09-29

DOI

10.1073/pnas.2000098117

Peer reviewed

Exploring the landscape of model representations

Thomas T. Foley[a,b], Katherine M. Kidder[a], M. Scott Shell[c,1], and W. G. Noid[a,1]

[a]Department of Chemistry, The Pennsylvania State University, University Park, PA 16802; [b]Department of Physics, The Pennsylvania State University, University Park, PA 16802; and [c]Department of Chemical Engineering, University of California, Santa Barbara, CA 93106

Edited by Valeria Molinero, The University of Utah, Salt Lake City, UT, and accepted by Editorial Board Member Peter J. Rossky August 7, 2020 (received for review January 6, 2020)

The success of any physical model critically depends upon adopting an appropriate representation for the phenomenon of interest. Unfortunately, it remains generally challenging to identify the essential degrees of freedom or, equivalently, the proper order parameters for describing complex phenomena. Here we develop a statistical physics framework for exploring and quantitatively characterizing the space of order parameters for representing physical systems. Specifically, we examine the space of low-resolution representations that correspond to particle-based coarse-grained (CG) models for a simple microscopic model of protein fluctuations. We employ Monte Carlo (MC) methods to sample this space and determine the density of states for CG representations as a function of their ability to preserve the configurational information, I, and large-scale fluctuations, $\mathcal{Q}$, of the microscopic model. These two metrics are uncorrelated in high-resolution representations but become anticorrelated at lower resolutions. Moreover, our MC simulations suggest an emergent length scale for coarse-graining proteins, as well as a qualitative distinction between good and bad representations of proteins. Finally, we relate our work to recent approaches for clustering graphs and detecting communities in networks.

multiscale modeling | entropy | networks | information theory | proteins

Introduction

Remarkably simple models explain many physical phenomena (1). This is clearly true of thermodynamic models for macroscopic systems (2). It is also true of simulation models for soft materials, such as polymers and proteins. While atomistic models provide exquisite detail, they are often computationally intractable. Moreover, unnecessary atomic details tend to obscure basic physical insight. Consequently, simulations of soft materials often adopt simplified, coarse-grained (CG) models that provide much greater computational efficiency and more transparent insight (3, 4).

Just as thermodynamic models rely upon identifying appropriate order parameters (1, 2), one expects the success of a CG model will critically hinge upon the quality of the CG representation, i.e., the degrees of freedom the CG model retains. However, it is often difficult to discern the essential degrees of freedom for complex phenomena. Historically, researchers have generally relied upon physical intuition to determine CG representations (5). For instance, generic bead-spring models of polymers often represent each monomer with a single sphere (4). More recent studies have proposed various methods for optimizing the representation of CG models for specific chemical systems (6–18).

Unfortunately, it is quite nontrivial to directly assess the intrinsic quality of a CG representation since the performance of a CG model will generally reflect various approximations introduced, e.g., when parameterizing its potential (5). Consequently, there remain many basic questions regarding the choice of CG representations. For instance, it is often far from obvious whether there exist significant distinctions between good and bad representations. Moreover, assuming such distinctions exist, it remains unclear whether good representations share certain common features or whether they are easy to find. These questions are also of considerable importance for the closely related problem of identifying order parameters or collective variables for accelerating, analyzing, and interpreting calculations with high-resolution models (19). Even more generally, these basic questions are of fundamental importance for developing reduced models for the large datasets that are relevant to, e.g., modern materials science (20).

In this work, we develop and apply a statistical physics framework for addressing these questions. Rather than optimizing the CG representation according to some specific metric, we seek to explore and characterize the entire landscape of representations (21). As an instructive case study, we start from a simple microscopic model of protein conformational fluctuations. There are an essentially infinite variety of ways to represent the protein in CG detail. Each representation corresponds to a different set of order parameters for characterizing the fluctuations of the underlying microscopic model. In particular, we consider representations that replace connected atomic groups with discrete CG particles. We introduce quantitative metrics for assessing the intrinsic quality of each representation. We employ Monte Carlo (MC) simulations to sample the space of representations and estimate a density of states quantifying the number of representations with a given quality. Interestingly, this density of states suggests the emergence of a phase transition distinguishing good and bad representations beyond a certain characteristic resolution. Finally, we also

Significance

Physical phenomena can often be described by surprisingly few order parameters. Unfortunately, it is challenging to identify these essential degrees of freedom. Here we develop a statistical physics framework for exploring the landscape of order parameters, or coarse-grained representations, for a microscopic protein model. We employ Monte Carlo methods to statistically characterize this landscape. We define metrics assessing the intrinsic quality of each representation for preserving the configurational information and large-scale motions of the underlying microscopic model. Interestingly, these metrics are anticorrelated in low-resolution representations. Moreover, below a critical resolution, a phase transition qualitatively distinguishes superior and inferior representations. Finally, we relate our work to recent approaches for clustering graphs and detecting communities in networks.

Author contributions: T.T.F., M.S.S., and W.G.N. designed research; T.T.F. and K.M.K. performed research; T.T.F., K.M.K., M.S.S., and W.G.N. analyzed data; and T.T.F., K.M.K., M.S.S., and W.G.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. V.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: wnoid@chem.psu.edu or shell@engineering.ucsb.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2000098117/-/DCSupplemental.

First published September 14, 2020.

relate this work to research on community detection in complex networks.

Results

Microscopic Model. We adopt the Gaussian network model (GNM) as a simple microscopic model of protein fluctuations about a single equilibrium conformation (22, 23). The GNM describes a protein as an isotropic network of n atoms, each corresponding to the α carbon of an amino acid residue. The reduced potential for the GNM is $u(\mathbf{q}) = \frac{1}{2}\mathbf{q}^\dagger \boldsymbol{\kappa} \mathbf{q}$, where $\mathbf{q}^\dagger = (q_1, \ldots, q_n)$ specifies the displacements of the n atoms from their equilibrium positions, while $\boldsymbol{\kappa}$ is a symmetric matrix that connects nearby atoms with linear springs and determines the corresponding covariance matrix, $\mathbf{c} \propto \boldsymbol{\kappa}^{-1}$. The equilibrium distribution is then $p(\mathbf{q}) \propto \exp[-\beta u(\mathbf{q})]$, where β is the inverse temperature. Despite its simplicity, the GNM has proven remarkably useful for investigating functional motions in proteins and biological complexes (24).

In the following, we primarily focus on the small helical protein 2ERL, although *SI Appendix* indicates that our conclusions are robust with respect to variations in the protein sequence, structure, and size. Fig. 1*A* presents the equilibrium structure of 2ERL, while Fig. 1 *B* and *C* present $\boldsymbol{\kappa}$, $\mathbf{c}$, and the corresponding vibrational density of states. We consider two metrics for characterizing the microscopic model: 1) the information content, $h \propto \ln\det \boldsymbol{\kappa}$, determines the protein-dependent contribution to the configurational entropy of the microscopic GNM (25) and quantifies the information stored in its equilibrium distribution (26) and 2) the vibrational power, $\sigma \propto \mathrm{Tr}\,\mathbf{c}$, quantifies the magnitude of conformational fluctuations sampled by the protein. Note that h emphasizes high-frequency motions, while σ emphasizes biologically important, low-frequency motions (24). *Materials and Methods* and *SI Appendix* describe the model and metrics in greater detail.

Characterizing and Sampling Representations. There exist many ways to represent the microscopic model in reduced detail. We codify each representation in terms of a mapping, $\mathbf{M}$, that specifies the CG configuration $\mathbf{Q} = (Q_1, \ldots, Q_N)$ for N CG particles, or sites, as a function of the microscopic coordinates, i.e., $\mathbf{M} : \mathbf{q} \to \mathbf{Q} = \mathbf{M}(\mathbf{q})$ (27). For simplicity, we assume that 1) this mapping corresponds to partitioning the n atoms into N disjoint groups of $R = n/N$ connected atoms and 2) each CG coordinate corresponds to the mass center for the associated atomic group. For example, Fig. 1 *D* and *E* illustrate the block map, which partitions the protein sequence into N contiguous fragments of R consecutive residues and associates a site with each fragment.

Given the microscopic equilibrium ensemble, $\mathbf{M}$ determines a mapped ensemble with the distribution $P(\mathbf{Q};\mathbf{M}) = \int \mathrm{d}\mathbf{q}\, p(\mathbf{q}) \delta(\mathbf{Q} - \mathbf{M}(\mathbf{q}))$. The covariance matrix for the CG coordinates in this mapped ensemble is $\mathbf{C}_\mathbf{M} = \mathbf{M}\mathbf{c}\mathbf{M}^\dagger \propto \mathbf{K}_\mathbf{M}^{-1}$ (25). Importantly, the information content, $H(\mathbf{M}) \propto \ln\det \mathbf{K}_\mathbf{M}$, and vibrational power, $\Sigma(\mathbf{M}) \propto \mathrm{Tr}\,\mathbf{C}_\mathbf{M}$, within the mapped ensemble are functions of $\mathbf{M}$. We quantitatively assess each CG representation, $\mathbf{M}$, based upon the fraction of information, $I(\mathbf{M}) = H(\mathbf{M})/h$, and vibrational power, $\mathcal{Q}(\mathbf{M}) = \Sigma(\mathbf{M})/\sigma$, that are preserved in the mapped ensemble. While many metrics may prove useful, I and $\mathcal{Q}$ exemplify metrics that emphasize high-frequency, localized motions and low-frequency, global motions, respectively. Importantly, these metrics directly assess the quality of the CG representation and can be analytically calculated for the GNM (25).

We seek to explore the landscape of CG mappings and to investigate the thermodynamics of selecting maps. Accordingly, we define an energy function $\mathcal{E}(\mathbf{M}) \propto I(\mathbf{M})$ or $\mathcal{E}(\mathbf{M}) = 1 - \mathcal{Q}(\mathbf{M})$ and perform MC simulations that sample maps according to a canonical distribution, $\mathcal{P}_\mathbf{M} \propto e^{-\beta_\mathcal{E}\mathcal{E}(\mathbf{M})}$, at a conjugate inverse temperature, $\beta_\mathcal{E}$. Starting from the block map, the simulations

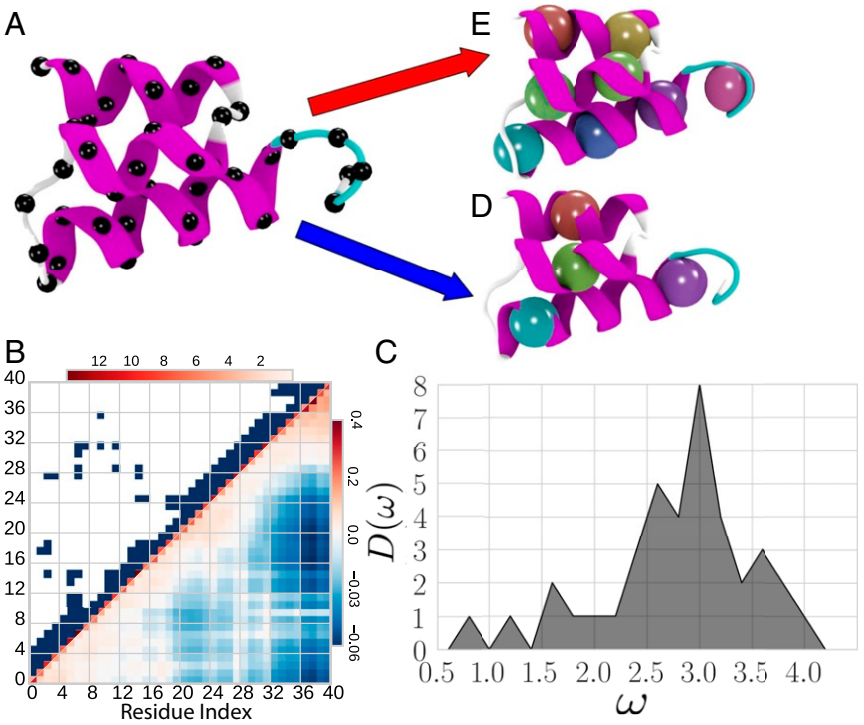

Fig. 1. Characterization of the model protein 2ERL. (*A*) Cartoon representation of the equilibrium folded structure with black spheres indicating α carbons. (*B*) Intensity plots of the upper and lower halves of the symmetric connectivity, $\boldsymbol{\kappa}$, and covariance, $\mathbf{c} \propto \boldsymbol{\kappa}^{-1}$, matrices. (*C*) Vibrational densities of states for the high resolution GNM of 2ERL. (*D* and *E*) CG representations with spheres representing the location of the CG sites for block maps with N = 4 and 8 sites, respectively. Figure employed VMD (66).

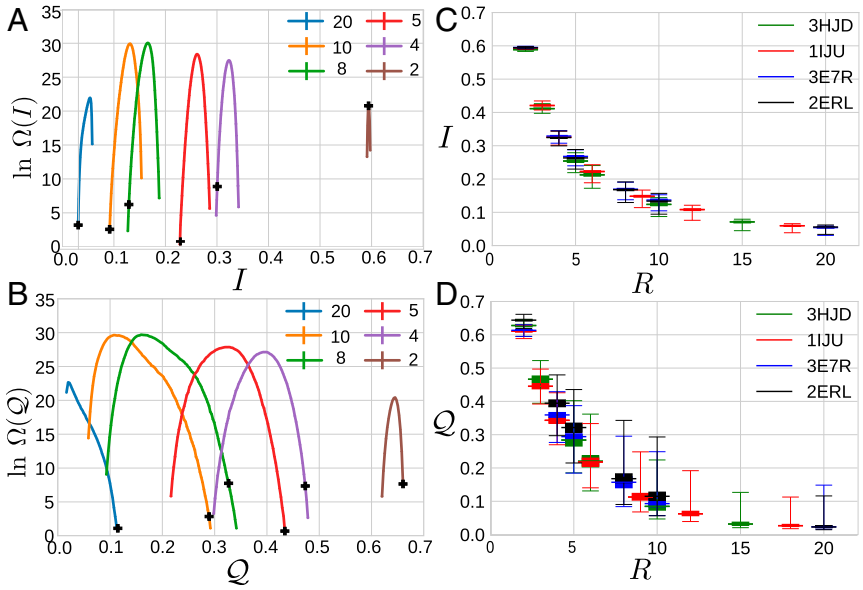

Fig. 2. Statistical analysis of mapping space. (*A* and *B*) The natural logarithm of the density of states, $\ln \Omega$, quantifying the number of maps, **M**, with given information content, *I*, or spectral quality, $\mathcal{Q}$, for 2ERL at varying degrees of coarsening, $R = n/N$, indicated by the colors of the legend. The black crosses indicate *I* and $\mathcal{Q}$ for the block map at each resolution. (*C* and *D*) Box plots indicating the mean (widest bar), extrema (top and bottom bars), and the 25 and 75% quantiles (shaded box) characterizing these densities of states for 2ERL (black) and for three other small proteins.

diffuse through mapping space by swapping atoms between pairs of CG sites, while ensuring that each site remains connected. Each move is accepted or rejected based upon a Metropolis criterion ensuring detailed balance (28). Given the maps sampled at a wide range of conjugate temperatures, we estimate densities of states that quantify the number of maps with a given information content, $\Omega(I)$, and spectral quality, $\Omega(\mathcal{Q})$.

Densities of States. Fig. 2 presents the natural logarithm of the resulting density of states, $\ln \Omega$, for different degrees of coarsening, R. Since $\ln \Omega$ exhibits pronounced peaks for each R, each resolution is characterized by a very large number of typical maps with a characteristic information content, I, and spectral quality, $\mathcal{Q}$. As expected, the characteristic values for I and $\mathcal{Q}$ systematically decrease with increased coarsening, although they decrease less rapidly than might be naïvely expected. For instance, when each site represents two atoms, i.e., $R = 2$, typical representations preserve $\sim$60% of the information and vibrational power present in the microscopic ensemble. Interestingly, $\ln \Omega(I)$ is similarly narrow at each resolution $R > 2$. In contrast, $\ln \Omega(\mathcal{Q})$ becomes increasingly broad with coarsening. In particular, at a given resolution, there exist rare mappings that provide significantly higher spectral quality than typical maps.

Fig. 2 *C* and *D* present box plots summarizing these statistics as a function of resolution for 2ERL and three other small proteins. The I distributions are not only narrow but also almost identical for different proteins. Thus, the information content depends strongly upon resolution but appears relatively insensitive to the details of the mapping or the particular proteins. In contrast, the $\mathcal{Q}$ distributions are much broader, demonstrate greater protein dependence, and demonstrate long tails toward relatively high spectral quality. Certain proteins and certain resolutions appear particularly amenable for preserving the low-frequency motions of the microscopic model.

The crosses in Fig. 2 *A* and *B* indicate I and $\mathcal{Q}$ for the block map at each resolution. When compared to typical representations, the block map tends to exhibit relatively low I and relatively high $\mathcal{Q}$. Due to the simplicity and symmetry of the 2ERL structure, the block map provides nearly minimal I and maximal $\mathcal{Q}$ at almost every resolution for this protein. For some proteins and resolutions, though, block maps do not optimize either metric but instead exhibit more typical values of I and $\mathcal{Q}$.

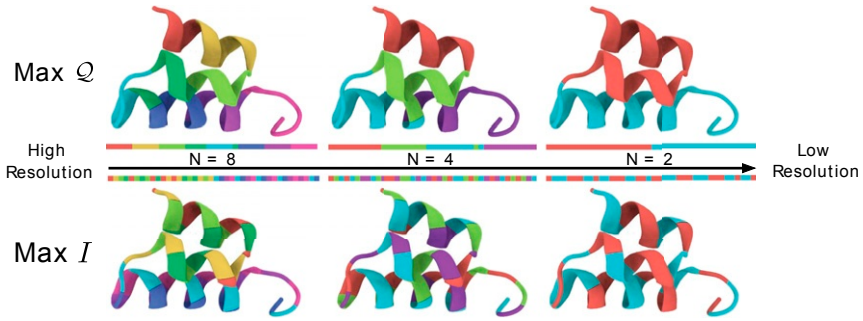

Fig. 3. Maps that maximize $\mathcal{Q}$ (*Top*) and *I* (*Bottom*) among maps with *N* = 8, 4, or 2 sites. The ribbon and line diagrams are colored to indicate the atoms that are grouped together in the three-dimensional structure and in the one-dimensional amino acid sequence, respectively. Figure employed VMD (66).

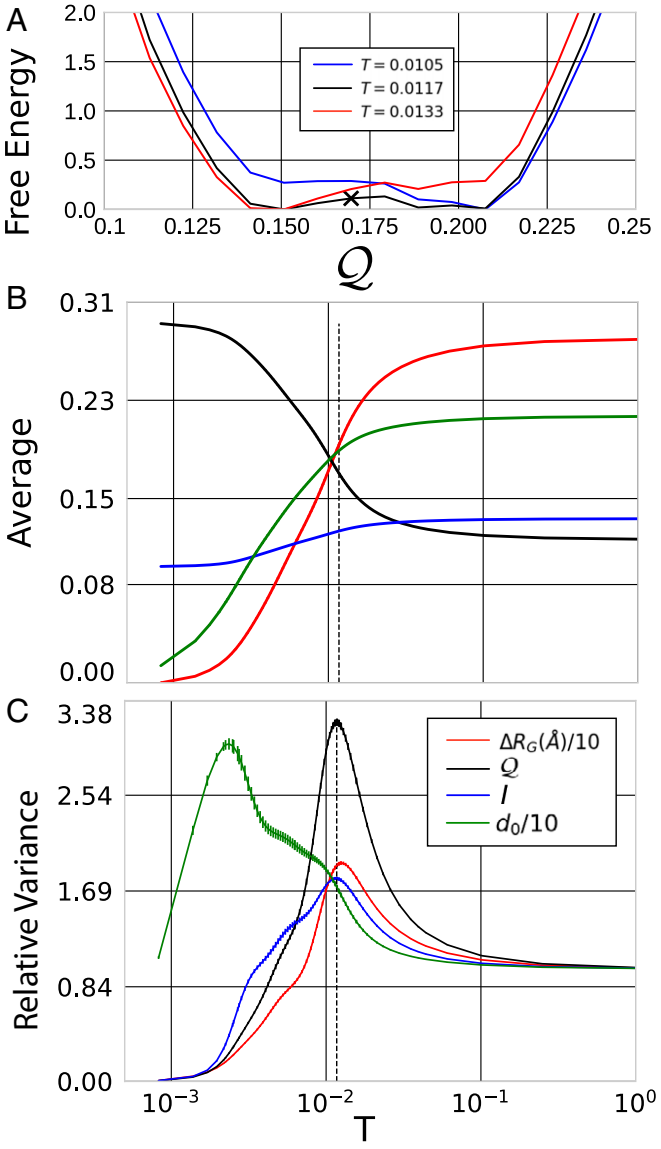

Fig. 4. Characterization of the apparent transition for $N = 4$ site representations of 2ERL. (*A*) The dimensionless free energy, $\beta_{\mathcal{Q}}F$, at the transition temperature (black) and at temperatures above (red) and below (blue) the transition. The black X indicates the separatrix, $\mathcal{Q}_*$, for which $\mathcal{P}(\mathcal{Q} < \mathcal{Q}_*) = 1/2$ at the transition temperature. (*B* and *C*) The averages and variances, respectively, for several metrics. The metric $d_0(\mathbf{M})$ quantifies the difference in the atomic groups defined by the map, **M**, and the ground state map, $\mathbf{M}_0$, while $R_G(\mathbf{M})$ quantifies the compactness of the associated atomic groups. For convenience, we have shifted R_G such that $\Delta R_G(\mathbf{M})$ vanishes as $T_{\mathcal{Q}} \to 0$ and have normalized variances relative to their $T_{\mathcal{Q}} \to \infty$ limit. Error bars estimate statistical uncertainty. The dashed vertical line indicates the transition temperature, which is defined by the variance peak in $\mathcal{Q}$. T denotes the fictitious temperature, $T_{\mathcal{Q}}$, conjugate to $\mathcal{E}(\mathbf{M}) = 1 - \mathcal{Q}(\mathbf{M})$.

Optimal Representations. Fig. 2 indicates that I and $\mathcal{Q}$ are not equivalent measures of model quality. Fig. 3 illuminates the difference between these two metrics by comparing the representations of 2ERL that maximize $\mathcal{Q}$ and I at various resolutions. *SI Appendix* presents analogous comparisons for several additional proteins. Representations that maximize the spectral quality, $\mathcal{Q}$, preserve low-frequency fluctuations by grouping atoms into densely packed sites that move coherently. Accordingly, the block map generally has relatively high spectral quality and, in the case of 2ERL, even maximizes $\mathcal{Q}$ at certain resolutions. In contrast, representations that maximize the information content, I, form sites by grouping atoms that are distributed across the protein in order to preserve high-frequency motions. Because these high-frequency motions are usually localized and often physically uninteresting in soft materials, we focus on the spectral quality, $\mathcal{Q}$, in the remainder of this work.

Apparent Phase Transition. Given a mechanical model for a finite physical system with energy E, an inflection point in the corresponding density of states, $\Omega(E)$, implies the existence of a first-order phase transition (29–31). Interestingly, the densities of states, $\Omega(\mathcal{Q})$, in Fig. 2*B* also exhibit inflection points at sufficiently coarse resolutions, which suggest the existence of analogous phase transitions in the space of CG representations. In order to characterize these transitions, we define a dimensionless free energy, $\beta_{\mathcal{Q}}F(\mathcal{Q};\beta_{\mathcal{Q}}) = -\ln \mathcal{P}(\mathcal{Q};\beta_{\mathcal{Q}})$, where $\mathcal{P}(\mathcal{Q};\beta_{\mathcal{Q}})$ is the probability of sampling a map **M** with spectral quality $\mathcal{Q}$ at the inverse temperature $\beta_{\mathcal{Q}}$ that is conjugate to $\mathcal{E}(\mathbf{M}) = 1 - \mathcal{Q}(\mathbf{M})$. Similarly, we define averages and variances as a function of the temperature $T_{\mathcal{Q}} = {\beta_{\mathcal{Q}}}^{-1}$. Fig. 4 characterizes the suggested transition in the space of maps for $N = 4$ site representations of 2ERL.

Fig. 4*A* demonstrates that at the transition temperature the free energy surface features two shallow minima, corresponding to maps with relatively high and low $\mathcal{Q}$. These minima are separated by a relatively small barrier, as might be expected for a weak first-order transition in a small system. The black cross indicates the separatrix, which is estimated based upon equal population for the two states. Fig. 4*B* demonstrates that near this transition the spectral quality increases and the information content decreases, as expected. Fig. 4*C* demonstrates that the variance in these metrics also peaks at or near this transition (32). In particular, we define the transition temperature by the variance peak for $\mathcal{Q}$.

Fig. 4 *B* and *C* also present two additional metrics characterizing this transition. For each resolution, $R = n/N$, we define the ground state map, $\mathbf{M}_0$, as the N-site map with maximum spectral quality. We define the distance, d_0, of a map **M** from $\mathbf{M}_0$,

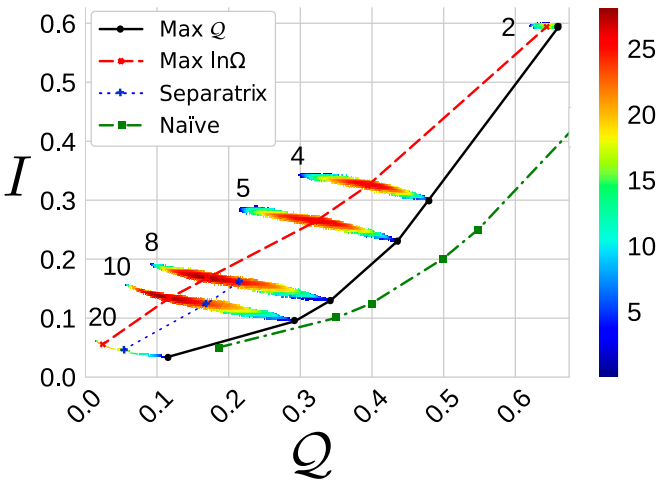

Fig. 5. Global perspective on mapping space for 2ERL. The heat map colors indicate the magnitude of the 2D ln densities of states, $\ln \Omega(\mathcal{Q}, I)$, for CG maps with resolutions R = 2, 4, 5, 8, 10, and 20. The dashed red and solid black curves indicate the maxima of $\ln \Omega$ and $\mathcal{Q}$, respectively, at each resolution. The dashed-dotted green curve presents a naïve estimate of the expected information content at each resolution, i.e., N/n, and the optimal spectral quality, $\mathcal{Q}_{N;\max}$, which corresponds to reproducing perfectly the $N - 1$ lowest vibrational frequencies of the high-resolution model. The dotted blue curve and crosses indicate the separatrices of transitions that are observed at sufficiently low resolutions.

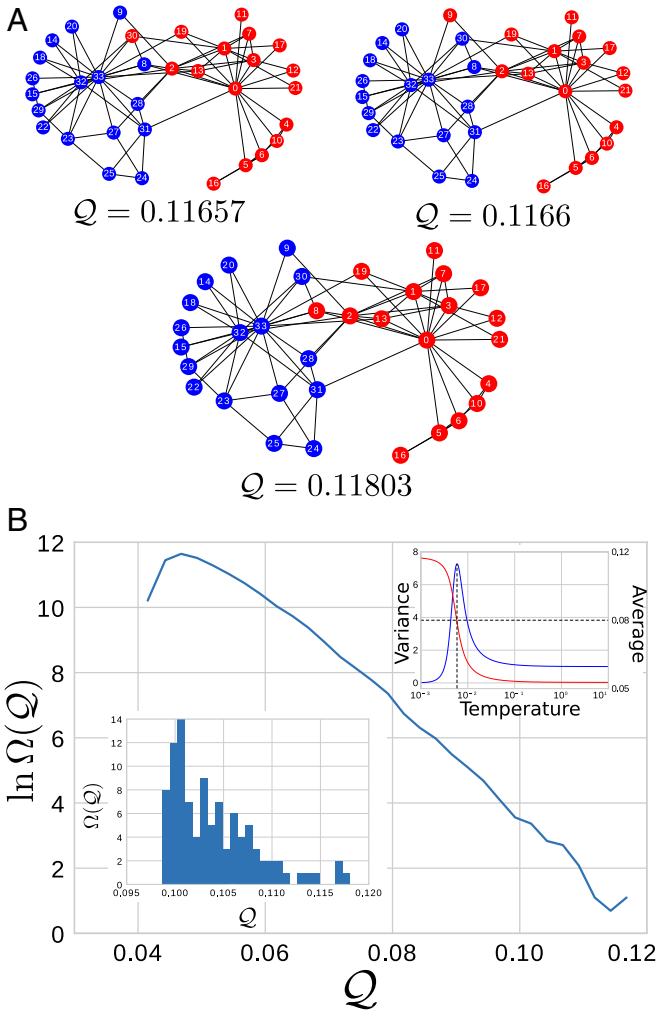

Fig. 6. Coarse-graining the GNM defined by Zachary's karate club network. (*A*) The $N = 2$ CG map with optimal spectral quality ($\mathcal{Q}_{2;\max}$ = 0.11803), as well as the first two excited states with slightly lower spectral quality. (*B*) $\ln \Omega(\mathcal{Q})$. (*Lower Left Inset*) The spectra for the first 100 maps. (*Upper Right Inset*) As a function of conjugate temperature, the average spectral quality (red curve, right scale) and the corresponding variance (blue curve, left scale), which has been normalized relative to its $\beta_{\mathcal{Q}} \to 0$ limit. The vertical line in *Upper Right Inset* indicates the transition temperature, while the horizontal line indicates the corresponding mean.

based upon the variation of information (VI) (33), which quantifies the dissimilarity of the corresponding atomic partitions, i.e., $d_0(\mathbf{M}) = \mathrm{VI}(\mathbf{M}, \mathbf{M}_0)$. Additionally, we define $\mathrm{R_G}(\mathbf{M})$, by the average gyration radius of the partitioned atomic groups in the equilibrium protein conformation. Fig. 4 *B* and *C* demonstrate that the sampled maps become more compact and also more similar to $\mathbf{M}_0$ at the observed transition. Interestingly, the variance in d_0 peaks at a noticeably lower temperature, which indicates considerable variation among the clusterings of maps with high spectral quality. Nevertheless, *SI Appendix* demonstrates that these metrics correlate quite well with the spectral fitness.

Consequently, Fig. 4 demonstrates that this transition in the space of representations bears consider similarity to a physical phase transition between different phases of matter. Moreover, this transition suggests a qualitative distinction between good and bad representations of a protein at a given resolution. In particular, good maps are characterized by compact sites that ensure high spectral quality and by quantitatively similar partitions of atoms.

Global Perspective. Fig. 5 provides a broader perspective on the space of representations by presenting an intensity plot of the natural logarithm of the joint density of states, $\ln \Omega(\mathcal{Q}, I)$ for CG representations of 2ERL. In particular, this two-dimensional density of states indicates the correlation between $\mathcal{Q}$ and I. While $\mathcal{Q}$ and I are essentially uncorrelated for the highest-resolution CG representations, they become highly anticorrelated for lower resolutions. As noted above, at a given resolution, the CG representations sample a fairly narrow range in I but a much broader range in $\mathcal{Q}$. For comparison, the green curve indicates, for each resolution, the maximum possible spectral fitness, $\mathcal{Q}_{\max}$, which would be achieved if the N-site CG representation perfectly preserved the $N-1$ lowest vibrational frequencies of the atomic model, as well as a naïve scaling expectation for the information content $I = 1/R$. Fig. 5 demonstrates that the CG representations almost always preserve more information than might be naïvely expected. Moreover, the best maps achieve $\sim$80% of $\mathcal{Q}_{\max}$.

The blue crosses in Fig. 5 indicate the separatrices for the transitions that are observed at lower resolutions. In the case of 2ERL, we only observe these transitions for resolutions grouping at least eight residues per site, corresponding to approximately two turns of an α helix. *SI Appendix* presents $\ln \Omega(\mathcal{Q})$ for six additional proteins of up to 72 amino acid residues with varying secondary structures and topologies, which suggest that these trends are quite common among proteins. In almost every case, the densities of states indicate the onset of a phase transition past a certain threshold resolution, which suggests a characteristic length scale for coarse-graining these proteins.

Relation to Networks. The process of determining CG representations for molecular systems bears striking similarity to clustering or detecting communities in complex networks. The GNM makes this analogy particularly transparent. The GNM defines an interaction network for a single protein by connecting nearby residues with linear springs. The protein residues correspond to the nodes of the network (or, equivalently, to the vertices of a graph) that are connected by edges corresponding to linear springs. The curvature of the GNM potential, κ, corresponds to the graph Laplacian, L, which specifies the edges of the protein interaction network (34). The process of grouping atoms into CG sites then corresponds to clustering nodes in a graph or defining communities in a network.

Consequently, the present work bears considerable similarity to several leading approaches for clustering and community detection (35, 36). For instance, *SI Appendix* demonstrates that the spectral quality, $\mathcal{Q}$, of a CG representation is quite correlated with the modularity (37), which quantifies the strength of the corresponding communities based upon the fraction of edges connecting the nodes within each cluster. Thus, the ground state representation that maximizes $\mathcal{Q}$ should be quite similar to the clustering obtained in simulations of Potts models that optimize the modularity (38, 39). The present work is also related to spectral clustering approaches that, e.g., partition nodes according to the lowest eigenvalues of L (40, 41) or that identify communities based upon the stability of random walks on the graph (42, 43).

However, the present work bears two crucial distinctions with respect to prior investigations of network communities. First, while many prior studies have sought a single clustering that achieves a specific objective (38, 40, 42) or an ensemble of graphs with certain characteristics (44, 45), we have focused on the space of representations for a single GNM, which corresponds to an ensemble of clusterings for a single graph (46). Second, and more importantly, the process of coarse-graining does not simply correspond to grouping nodes but rather to the process of viewing the fluctuations of an underlying microscopic model for a physical system through a particular coarse lens. This physical process

corresponds to a rigorous thermodynamic projection that renormalizes the underlying microscopic potential (47), such that the resulting effective springs connecting CG sites vary in strength and even in sign (25, 48).

Accordingly, it is intriguing to apply the physical coarse-graining process to Zachary's karate club network (49, 50), which is illustrated in Fig. 6. Zachary's karate club provides a particularly simple archetype of networks considered by community detection algorithms and is known to have two meaningful communities. We defined a corresponding GNM by defining $\boldsymbol{\kappa}$ as the graph Laplacian of the network. In this case, we can exhaustively enumerate all representations of the GNM since the network includes only 34 nodes with local connections. Fig. 6*A* presents the three CG representations that maximize $\mathcal{Q}$. Indeed, the ground state representation of the GNM corresponds to the known communities, while the first two excited state representations correspond to very similar clusterings. Fig. 6*B* presents the density of states, $\Omega(\mathcal{Q})$, for two-site representations of this GNM. Interestingly, a similar phase transition is also observed between good and bad representations of the GNM for Zachary's karate club network.

Conclusions

We have presented a statistical thermodynamic formalism and computational investigation of the landscape of CG representations for physical systems. In this first investigation, we adopted the GNM as a high-resolution model since it provides a qualitatively useful description of protein fluctuations and is amenable to theoretical analysis. We considered CG representations that are both linear and local since we defined CG coordinates as linear combinations of the atomic coordinates for connected groups. By employing an analytic coarse-graining of the GNM (25), we quantitatively and exactly assessed the intrinsic quality of CG representations without introducing any approximations, e.g., due to approximating the interactions between CG particles.

The present work focused on characterizing CG representations according to two metrics, I and $\mathcal{Q}$, which quantify the ability of the CG representation to preserve the information and large-scale fluctuations, respectively, contained in the microscopic ensemble. A priori, one might anticipate that these metrics would both prove useful for optimizing CG representations. Our numerical studies demonstrate that both I and $\mathcal{Q}$ decrease in a similar fashion with coarsening for typical maps. However, while I appears relatively insensitive to the details of the CG representation or the protein, $\mathcal{Q}$ appears more sensitive to variations in representation and protein structure. Furthermore, these metrics appear uncorrelated for high-resolution representations but become highly anticorrelated for low-resolution representations. Representations that maximize I feature loosely connected sites that preserve the information associated with the many localized and, thus, relatively informative high-frequency vibrations. Conversely, representations that maximize $\mathcal{Q}$ correspond to densely connected sites that preserve few low-frequency vibrations.

These considerations explain why block maps that group residues consecutively in sequence tend to be information-poor but provide a good description of low-frequency fluctuations. This intuition also underlies the connection between principal component analysis and the renormalization group (51), as well as current strategies for optimizing CG representations (6, 13) and order parameters (52, 53). In particular, *SI Appendix* demonstrates that $\mathcal{Q}$ is (anti-) correlated with the objective function, χ^2, which is minimized in the essential dynamics coarse-graining methodology for determining good CG representations (7). Moreover, our estimates for $\Omega(\chi^2)$ indicate that similar phase behavior would be observed if χ^2 were adopted as a metric for characterizing CG representations. Since $\mathcal{Q}$ emphasizes the large-magnitude, low-frequency motions that define the essential dynamics subspace of the mapped covariance matrix (54), we expect $\mathcal{Q}$ is representative of many metrics employed to identify coherent structural domains in proteins.

Quite generally, one expects that physical models of soft materials will often demonstrate relatively few low-frequency motions that correspond to important physical transitions and comparatively many high-frequency modes that correspond to uninteresting localized motions. In other words, most of the information contained in high-resolution models describes uninteresting noise, while a comparatively small fraction of the information describes interesting physics. For this reason, physical models are often "sloppy" in the sense of predicting large-scale phenomena that are insensitive to most of the parameters defining the model (55, 56). Moreover, these results suggest that it may be unwise to optimize representations of physical systems by naïvely maximizing their information content. Similarly, it may be unwise to optimize CG representations for backmapping to atomic resolution, i.e., for reintroducing high-resolution details into low-resolution structures. Rather, it is important to consider the physical variables of interest when determining the CG representation of a particular system.

Most interestingly, our numerical results suggest the emergence of a characteristic resolution for coarse-graining proteins. For relatively high resolutions, all CG representations are qualitatively similar. Below this characteristic resolution, a phase transition indicates a qualitative distinction between good and bad representations that becomes increasingly significant with further coarsening. Good representations reflect similar partitions of atoms into spatially compact, highly modular sites with relatively many stabilizing intrasite interactions.

In the case of 2ERL, this phase transition first emerges in $N=5$ site representations for which $R=8$ amino acids are grouped into each CG site. Fig. 3 indicates that just below this critical resolution, the ground state map represents each of the two small helices with distinct sites, while splitting the larger helix into contiguous fragments. This suggests that the critical resolution corresponds to the emergence of distinct, modular subunits that are stabilized by many internal interactions with relatively few interactions between different subunits. It also indicates that the details of this transition may depend somewhat upon the specific interactions included in the microscopic model. In the extreme limit that the microscopic GNM only includes nearest-neighbor interactions along the backbone, then only block maps are allowed, and no phase transition will be observed. However, *SI Appendix* demonstrates that similar transitions and critical resolutions are observed when the length scale defining interactions in the microscopic GNM is either decreased or increased by 30%. Thus, our findings appear fairly robust with respect to variations in the microscopic model.

Additionally, our work also highlights the similarity between the selection of CG representations for physical systems and recent work in clustering and detecting communities in complex networks (35). In particular, our work bears striking similarity to a variety of spectral approaches based upon the eigenvalues of the graph Laplacian (40–43). Importantly, though, the process of coarse-graining the microscopic GNM reweights the edges of the reduced graph to reflect the effective interactions at the CG resolution. Interestingly, this physical coarse-graining approach identifies the known communities for an archetypal network. Consequently, the present landscape approach may prove fruitful for considering ensembles of clusterings of a single graph (46) and for characterizing the effective interactions between communities. Conversely, the tools developed for community detection may prove useful for developing CG representations of physical systems (15, 16).

In closing, we note several promising directions for future work. First of all, future studies should further investigate the sensitivity of the observed phase transition and critical resolution

to protein size and structure. Moreover, while the present work assumed that each CG particle corresponded to an equal number of atoms, we anticipate that it may be fruitful to relax this assumption. Similarly, while the present work considered linear, local CG representations, future studies should consider more general nonlinear, nonlocal order parameters. Additionally, we anticipate further exploring the relation to network community detection in future studies. Finally, it would be most interesting to extend this approach to simple model potentials with multiple metastable states (57) and, ultimately, to more realistic potentials that allow for folding–unfolding transitions (58). In this case, the quality of a given CG representation may vary among these metastable states (17, 59, 60). Nevertheless, we hope that this first study provides a useful framework for systematically constructing representations of complex physical systems.

Materials and Methods

High-Resolution Model. The GNM represents a protein as a network of n atoms with linear springs connecting nearby atoms. The dimensionless GNM potential is $u(\mathbf{q}) = \frac{1}{2}\mathbf{q}^\dagger \boldsymbol{\kappa} \mathbf{q}$ where the dimensionless configuration $\mathbf{q}^\dagger = (q_1, \ldots, q_n)$ specifies the displacement of the atoms from equilibrium, and † denotes the transpose. The present approach can be readily adopted for anisotropic network models (61) or for quasiharmonic approximations to more general nonlinear models (62). The symmetric matrix, $\boldsymbol{\kappa}$, corresponds to the graph Laplacian (34) for a protein interaction network that is formed by representing each atom with a vertex and introducing edges between nearby atoms. Because the protein is connected, the null space of $\boldsymbol{\kappa}$ is spanned by a single vector corresponding to uniform translation of all n atoms. Consequently, we consider matrix inverses and determinants in the complementary image space.

We employ the ProDy server to determine $\boldsymbol{\kappa}$ for the high-resolution GNM (63). This GNM treats the n α carbons associated with the n residues of the protein and includes interactions between each pair of α carbons that are within a cutoff of R_c = 7.5 Å.

The (dimensionless) excess configurational entropy, s, of the GNM is

$$s = -\int d\mathbf{q}\, p(\mathbf{q}) \ln \left[L^n p(\mathbf{q})\right] = (n-1)s_0 - \frac{1}{2}\ln t_\kappa, \qquad [1]$$

where $s_0 = \frac{1}{2}\left(1 + \ln[2\pi/\beta L^2]\right)$ is a protein-independent constant, while $t_\kappa = n^{-1} \det \boldsymbol{\kappa}$ is the number of spanning trees for the protein interaction network (34). Consequently, we define $h = h(\boldsymbol{\kappa}) = \frac{1}{2}\ln t_\kappa$ as the nontrivial information in the high-resolution model. Additionally, we consider the mass-weighted fluctuations about the equilibrium configuration:

$$\sigma = \left\langle \sum_{i=1}^{n} m q_i^2 \right\rangle = \mathrm{Tr}_n m\mathbf{c} = \beta^{-1} \sum_{i=1}^{n-1} \omega_i^{-2}, \qquad [2]$$

where $\mathbf{c} = (\beta\boldsymbol{\kappa})^{-1}$ is the covariance matrix describing correlated fluctuations, the angular brackets denote an equilibrium average according to $p(\mathbf{q})$, and $\omega_i > 0$ is the i^{th} vibrational frequency. For simplicity, we assume that all atoms have equal mass, m.

Coarse Representation. We codify CG representations with a mapping, $\mathbf{M}$, that specifies the CG configuration, $\mathbf{Q} = (Q_1, \ldots, Q_N)$, as a function of the microscopic configuration, $\mathbf{Q} = \mathbf{M}(\mathbf{q})$. We consider mappings that partition the n atoms into N mutually disjoint subsets, $\{S_1, \ldots, S_N\}$, each of which contains $R = n/N$ atoms that form a connected subgraph of the high-resolution protein interaction network, i.e., the bonds of the GNM must connect the atoms within each site. We associate a CG site, I, with each atomic group, S_I, and we define the CG coordinate, Q_I, by the mass center of the atomic group.

The mapping, $\mathbf{M}$, along with the microscopic configuration distribution, $p(\mathbf{q})$, determines the configuration distribution for the mapped ensemble:

$$P(\mathbf{Q};\mathbf{M}) = \int d\mathbf{q}\, p(\mathbf{q})\, \delta(\mathbf{Q} - \mathbf{M}(\mathbf{q})) \propto \exp\left[-\frac{1}{2}\beta \mathbf{Q}^\dagger \mathbf{K}_\mathbf{M} \mathbf{Q}\right], \qquad [3]$$

where $\mathbf{K}_\mathbf{M}^{-1} = \mathbf{M}\boldsymbol{\kappa}^{-1}\mathbf{M}^\dagger$ for the maps we consider (25). The excess entropy of the mapped ensemble is

$$S(\mathbf{M}) = -\int d\mathbf{Q}\, P(\mathbf{Q};\mathbf{M}) \ln \left[L^N P(\mathbf{Q};\mathbf{M})\right], \qquad [4]$$

$$= (N-1)s_0 - \frac{1}{2}\ln T_{\mathbf{K}_\mathbf{M}}, \qquad [5]$$

where $T_\mathbf{K} = N^{-1} \det \mathbf{K}$. Accordingly, $H = H(\mathbf{M}) = \frac{1}{2}\ln T_{\mathbf{K}_\mathbf{M}}$ quantifies the nontrivial information preserved in the mapped ensemble. We also consider the mass-weighted fluctuations in the mapped ensemble:

$$\Sigma(\mathbf{M}) = \left\langle \sum_{I=1}^{N} M Q_I^2 \right\rangle = \mathrm{Tr}_N M \mathbf{C}_\mathbf{M} = k_B T \sum_{I=1}^{N-1} \Omega_I^{-2}, \qquad [6]$$

where $\mathbf{C}_\mathbf{M} = \mathbf{M}\mathbf{c}\mathbf{M}^\dagger$, $M = mn/N$ is the CG mass, and $\Omega_I > 0$ is the I^{th} vibrational frequency of the CG model.

Metrics for Characterizing Representations. We consider two metrics for quantitatively assessing the quality of a CG representation. We define the information quality

$$I = I(\mathbf{M}) = H(\mathbf{M})/h = \ln T_{\mathbf{K}_\mathbf{M}} / \ln t_\kappa \qquad [7]$$

as the fraction of information preserved by the mapping. We define the spectral quality

$$\mathcal{Q} = \mathcal{Q}(\mathbf{M}) = \Sigma(\mathbf{M})/\sigma = \sum_{I=1}^{N-1} \Omega_I^{-2} \Big/ \sum_{i=1}^{n-1} \omega_i^{-2} \qquad [8]$$

as the fraction of vibrational power preserved by the CG representation. Both metrics satisfy $0 \le I, \mathcal{Q} \le 1$, vanish in the limit $N \to 0$, and equal unity only in the limit $N = n$.

Fig. 4 considers two additional metrics: 1) Given the folded structure of a protein, we define the physical size of a CG site as the three-dimensional radius of gyration for the α carbons that are grouped into the site. We define the radius of gyration, $\mathrm{R_G}(\mathbf{M})$, for the map, $\mathbf{M}$, as the average gyration radius of the corresponding sites. 2) Given the ground state map, $\mathbf{M}_0$, which maximizes $\mathcal{Q}$, we define the distance of a map, $\mathbf{M}$, from the ground state as $d_0(\mathbf{M}) = \mathrm{VI}(\mathbf{M}, \mathbf{M}_0)$, where VI is the variation of information (33), which is a distance metric commonly employed for distinguishing clusterings on graphs and is explicitly defined in *SI Appendix*.

Exploring Representations. We employ MC simulations to sample the space of connected CG maps at different resolutions, R. These simulations treat the CG mapping, $\mathbf{M}$, as the microstate and employ a dimensionless energy function $\mathcal{E} = \mathcal{E}(\mathbf{M}) = 1 - \mathcal{Q}(\mathbf{M})$ or $2H(\mathbf{M})$ to define an equilibrium Boltzmann distribution:

$$\mathcal{P}_\mathbf{M} \propto \exp\left[-\beta_\mathcal{E}\, \mathcal{E}(\mathbf{M})\right], \qquad [9]$$

where $\beta_\mathcal{E}$ is the conjugate inverse temperature. Starting from a map defined by N connected atomic subgroups, i.e., $\mathbf{M} = \{S_1, \ldots, S_N\}$, we consider two move sets for generating a new trial map, $\mathbf{M}'$. Both move sets select a pair of sites $S_I, S_J \in \mathbf{M}$ that are replaced with a new pair of sites, S_I' and S_J', while leaving the remaining $N-2$ sites unchanged. 1) The swap-based move set swaps a pair of atoms between the two sites, i.e., one atom is moved from site I to site J, while a second atom is moved from site J to site I. 2) The site-based move set merges the two sites to form a supersite $S_{IJ} = S_I \cup S_J$ of $2R$ atoms and then partitions S_{IJ} into two new sites, S_I' and S_J', each of which contains R atoms. Both move sets require that the resulting sites S_I' and S_J' are connected subgraphs of the high-resolution protein interaction network. Note that we employed the swap-based move set to exhaustively enumerate the set of maps for Zachary's karate club. It is possible that the swap-based move set is not ergodic under certain conditions, although we have obtained numerically identical results with the less restrictive site-based move set. In cases that the move set is not ergodic, our results strictly apply to the subset of mapping space that is reachable from the block map.

The restriction to connected maps significantly reduces the size of mapping space but also complicates sampling. Operationally, given a connected map, $\mathbf{M}$, and a specific move set, we first determine the number, $C_\mathbf{M}$, of connected maps that can be reached in one move from $\mathbf{M}$. We then select one of these connected maps, $\mathbf{M}'$, according to a uniform probability distribution and determine the number of maps, $C_{\mathbf{M}'}$, that can be reached in one move from $\mathbf{M}'$. We accept or reject the move $\mathbf{M} \to \mathbf{M}'$ according to the acceptance probability (28)

$$\mathrm{Acc}(\mathbf{M} \to \mathbf{M}') = \frac{C_\mathbf{M}}{\max\{C_\mathbf{M}, C_{\mathbf{M}'}\}} \min\left\{1, \mathcal{P}_{\mathbf{M}'}/\mathcal{P}_\mathbf{M}\right\}. \qquad [10]$$

Because in general, $C_{\mathbf{M}} \neq C_{\mathbf{M}'}$, a prefactor is necessary to preserve detailed balance, although other prefactors are possible.

We performed MC simulations using either energy function $\mathcal{E} = 1 - \mathcal{Q}$ or $2H$ at a range of positive and negative conjugate (inverse) temperatures, $\beta_{\mathcal{E}}$. Given the CG maps, **M**, sampled from these MC simulations, we employed the multistate Bennett acceptance ratio method to estimate their statistical weights for various energy functions and conjugate temperatures (64, 65). We estimated the density of states for each energy function $\mathcal{E}$ from the $\beta_{\mathcal{E}} \to 0$ limit of these statistical weights.

Data Availability. Software, Python notebooks, and text data files have been deposited in http://www.datacommons.psu.edu with DOI 10.26208/139c-8x65.

ACKNOWLEDGMENTS The authors gratefully acknowledge financial support from the National Science Foundation (Grants MCB-1053970 and CHE-1856337 to W.G.N. and CHE-1800344 to M.S.S.). Portions of this research were conducted with Advanced CyberInfrastructure computational resources provided by The Institute for CyberScience at The Pennsylvania State University (http://ics.psu.edu). In addition, parts of this research were conducted with XSEDE resources awarded by Grant TG-CHE170062. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation (Grant ACI-1548562). Figs. 1 and 3 employed VMD. VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana–Champaign.

1. N. Goldenfeld, L. P. Kadanoff, Simple lessons from complexity. *Science* **284**, 87–89 (1999).
2. H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics* (Wiley, 1985).
3. M. Levitt, A. Warshel, Computer simulation of protein folding. *Nature* **253**, 694–698 (1975).
4. C. Peter, K. Kremer, Multiscale simulation of soft matter systems. *Faraday Discuss.* **144**, 9–24 (2010).
5. W. G. Noid, Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **139**, 090901 (2013).
6. H. Gohlke, M. F. Thorpe, A natural coarse graining for simulating large biomolecular motion. *Biophys. J.* **91**, 2115–2120 (2006).
7. Z. Y. Zhang *et al.*, A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys. J.* **95**, 5073–5083 (2008).
8. Z. Y. Zhang, G. A. Voth, Coarse-grained representations of large biomolecular complexes from low-resolution structural data. *J. Chem. Theor. Comput.* **6**, 2990–3002 (2010).
9. A. V. Sinitskiy, M. G. Saunders, G. A. Voth, Optimal number of coarse-grained sites in different components of large biomolecular complexes. *J. Phys. Chem. B* **116**, 8363–8374 (2012).
10. N. Guttenberg *et al.*, Minimizing memory as an objective for coarse-graining. *J. Chem. Phys.* **138**, 094111 (2013).
11. J. F. Rudzinski, W. G. Noid, Investigation of coarse-grained mappings via an iterative generalized yvon-born-green method. *J. Phys. Chem. B* **118**, 8295–8312 (2014).
12. M. Li, J. Z. Zhang, F. Xia, Constructing optimal coarse-grained sites of huge biomolecules by fluctuation maximization. *J. Chem. Theory Comput.* **12**, 2091–2100 (2016).
13. S. Orioli, P. Faccioli, Dimensional reduction of Markov state models from renormalization group theory. *J. Chem. Phys.* **145**, 124120 (2016).
14. J. J. Madsen, A. V. Sinitskiy, J. Li, G. A. Voth, Highly coarse-grained representations of transmembrane proteins. *J. Chem. Theory Comput.* **13**, 935–944 (2017).
15. M. Chakraborty, C. Xu, A. D. White, Encoding and selecting coarse-grain mapping operators with hierarchical graphs. *J. Chem. Phys.* **149**, 134106 (2018).
16. M. A. Webb, J. Y. Delannoy, J. J. de Pablo, Graph-based approach to systematic molecular coarse-graining. *J. Chem. Theory Comput.* **15**, 1199–1208 (2018).
17. L. Boninsegna, R. Banisch, C. Clementi, A data-driven perspective on the hierarchical assembly of molecular structures. *J. Chem. Theory Comput.* **14**, 453–460 (2018).
18. P. Diggins, C. Liu, M. Deserno, R. Potestio, Optimal coarse-grained site selection in elastic network models of biomolecules. *J. Chem. Theory Comput.* **15**, 648–664 (2019).
19. F. Noé, C. Clementi, Collective variables for the study of long-time kinetics from molecular trajectories: Theory and methods. *Curr. Opin. Struct. Biol.* **43**, 141–147 (2017).
20. A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater.* **4**, 053208 (2016).
21. T. T. Foley, "Statistical mechanics of coarse-graining," PhD thesis, Pennsylvania State University, University Park, PA (2017).
22. P. J. Flory, M. Gordon, N. G. McCrum, Statistical thermodynamics of random networks [and discussion]. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **351**, 351–380 (1976).
23. T. Haliloglu, I. Bahar, B. Erman, Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **79**, 3090–3093 (1997).
24. I. Bahar, T. R. Lezon, A. Bakan, I. H. Shrivastava, Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. *Chem. Rev.* **110**, 1463–1497 (2010).
25. T. T. Foley, M. S. Shell, W. G. Noid, The impact of resolution upon entropy and information in coarse-grained models. *J. Chem. Phys.* **143**, 243104 (2015).
26. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley Interscience, ed. 2, 2006).
27. W. G. Noid *et al.*, The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **128**, 244114 (2008).
28. M. E. J. Newman, G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon Press, 1999).
29. D. H. E. Gross, *Microcanonical Thermodynamics* (World Scientific, 2001).
30. D. H. E. Gross, J. F. Kenney, The microcanonical thermodynamics of finite systems: The microscopic origin of condensation and phase separations, and the conditions for heat flow from lower to higher temperatures. *J. Chem. Phys.* **122**, 224111 (2005).
31. S. Schnabel, D. T. Seaton, D. P. Landau, M. Bachmann, Microcanonical entropy inflection points: Key to systematic understanding of transitions in finite systems. *Phys. Rev. E* **84**, 011127 (2011).
32. D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).
33. M. Meilă, Comparing clusterings—An information based distance. *J. Multivar. Anal.* **98**, 873–895 (2007).
34. J. M. Harris, J. L. Hirst, M. J. Mossinghoff, *Combinatorics and Graph Theory* (Springer, 2010).
35. S. Fortunato, Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
36. T. P. Peixoto, Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X* **4**, 011047 (2014).
37. M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).
38. J. Reichardt, S. Bornholdt, Statistical mechanics of community detection. *Phys. Rev. E* **74**, 016110 (2006).
39. R. Peter, N. Zohar, Local resolution-limit-free Potts model for community detection. *Phys. Rev. E* **81**, 046114 (2010).
40. D. Gfeller, P. De Los Rios, Spectral coarse graining of complex networks. *Phys. Rev. Lett.* **99**, 038701 (2007).
41. D. Gfeller, P. De Los Rios, Spectral coarse graining and synchronization in oscillator networks. *Phys. Rev. Lett.* **100**, 174104 (2008).
42. J. C. Delvenne, S. N. Yaliraki, M. Barahona, Stability of graph communities across time scales. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12755–12760 (2010).
43. M. T. Schaub, J. C. Delvenne, S. N. Yaliraki, M. Barahona., Markov dynamics as a zooming lens for multiscale community detection: Non clique-like communities and the field-of-view limit. *PLoS One* **7**, e32210 (2012).
44. G. Bianconi, Entropy of network ensembles. *Phys. Rev. E* **79**, 036114 (2009).
45. M. E. J. Newman, T. P. Peixoto, Generalized communities in networks. *Phys. Rev. Lett.* **115**, 088701 (2015).
46. C. P. Massen, J. P. K. Doye, Thermodynamics of community structure. arXiv:0610077 (3 October 2006).
47. L. P. Kadanoff, *Statistical Physics* (World Scientific, 2000).
48. T. R. Lezon, I. Bahar, Using entropy maximization to understand the determinants of structural dynamics beyond native contact topology. *PLoS Comput. Biol.* **6**, e1000816 (2010).
49. W. W. Zachary, An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
50. T. P. Peixoto, Reconstructing networks with unknown and heterogeneous errors. *Phys. Rev. X* **8**, 041011 (2018).
51. S. Bradde, W. Bialek, PCA meets RG. *J. Stat. Phys.* **167**, 462–475 (2017).
52. G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, F. Noé, Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 015102 (2013).
53. P. Tiwary, B. J. Berne, Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2839–2844 (2016).
54. A. Amadei, A. B. M. Linssen, H. J. C. Berendsen, Essential dynamics of proteins. *Proteins* **17**, 412–425 (1993).
55. B. B. Machta, R. Chachra, M. K. Transtrum, J. P. Sethna, Parameter space compression underlies emergent theories and predictive models. *Science* **342**, 604–607 (2013).
56. M. K. Transtrum *et al.*, Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys.* **143**, 010901 (2015).
57. J. W. Chu, G. A. Voth, Coarse-grained free energy functions for studying protein conformational changes: A double-well network model. *Biophys. J.* **93**, 3860–3871 (2007).
58. P. E. M. Lopes, O. Guvench, A. D. MacKerell, *Current Status of Protein Force Fields for Molecular Dynamics Simulations* (Springer, New York, 2015).
59. J. F. Dama *et al.*, The theory of ultra-coarse-graining. 1. General principles. *J. Chem. Theory Comput.* **9**, 2466–2480 (2013).
60. T. Bereau, J. F. Rudzinski, Accurate structure-based coarse graining leads to consistent barrier-crossing dynamics. *Phys. Rev. Lett.* **121**, 256002 (2018).
61. A. R. Atilgan *et al.*, Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80**, 505–515 (2001).
62. R. M. Levy, A. R. Srinivasan, W. K. Olson, J. A. McCammon, Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers* **23**, 1099–1112 (1984).
63. A. Bakan, L. M. Meireles, I. Bahar, ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).
64. M. R. Shirts, J. D. Chodera, Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105 (2008).
65. J. Towns *et al.*, Xsede: Accelerating scientific discovery. *Comput. Sci. Eng.* **16**, 62–74 (2014).
66. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).