PNAS
www.pnas.org

Supplementary Information for

Exploring the landscape of model representations

Thomas T. Foley, Katherine M. Kidder, M. Scott Shell, W. G. Noid

W. G. Noid.
E-mail: wgn1@psu.edu

This PDF file includes:

Supporting Information Text

1. High resolution network models

We constructed a high resolution Gaussian Network Model (GNM) for each of 7 proteins.(1–3) In each case, the high resolution model represented each of the n amino acids with its α carbon. We employed the ProDy server to construct a contact matrix, θ_{ij}, from the folded structure, $\mathbf{r}^*$, for each protein.(4) For each distinct pair of atoms, i and j, in the high resolution GNM, the contact matrix is

$$\theta_{ij} = \begin{cases} 1 & \text{if } r^*_{ij} < R_c \\ 0 & \text{otherwise} \end{cases} \quad [1]$$

where r^*_{ij} is the distance between the pair in the folded reference structure and we define the parameter $R_c = 7.5$Å. The high resolution GNM potential is then defined

$$u_{\mathrm{GNM}}(\mathbf{r}|\mathbf{r}^*) = \frac{1}{2}\Gamma \sum_{i=1}^{n}\sum_{j>i}^{n} \theta_{ij}\left(\mathbf{r}_{ij} - \mathbf{r}^*_{ij}\right)^2, \quad [2]$$

where Γ is an irrelevant dimensional constant, $\mathbf{r}_{ij}$ is the vector from atom i to atom j in configuration $\mathbf{r}$, and $\mathbf{r}^*_{ij}$ is the corresponding vector in the folded structure, $\mathbf{r}^*$. Note that the main text presents results in terms of dimensionless quantities, while this SI explicitly treats the relevant dimensional factors. This GNM potential separates into independent potentials governing the fluctuations in each Cartesian direction, each of which is of the form:

$$u(\mathbf{q}) = \frac{1}{2}\Gamma \mathbf{q}^\dagger \boldsymbol{\kappa} \mathbf{q}, \quad [3]$$

where $\mathbf{q} = (q_1, \ldots, q_n)$ specifies the atomic displacements from equilibrium in one Cartesian direction, † denotes the transpose, and the curvature of the potential is

$$\kappa_{ij} = n_i \delta_{ij} - \theta_{ij}, \quad [4]$$

where $n_i = \sum_{j(\neq i)} \theta_{ij}$ is the total number of contacts formed by atom i. Because u_{GNM} is invariant under translation, $\boldsymbol{\kappa}$ possesses a one-dimensional null-space.

The thermodynamic and statistical properties of the atomic GNM can be analytically determined.(1, 5) The equilibrium probability distribution is given by

$$p(\mathbf{q}) = z^{-1} \exp\left[-\beta u(\mathbf{q})\right], \quad [5]$$

where $\beta = 1/k_B T$ is the inverse of the physical temperature, T, and

$$z = \int \mathrm{d}\mathbf{q}\, e^{-\beta u(\mathbf{q})} = n^{1/2} L \sqrt{(2\pi)^{n-1} \det \mathbf{c}}, \quad [6]$$

where L is the (one-dimensional) volume and $\mathbf{c} = (\beta\Gamma\boldsymbol{\kappa})^{-1}$ is the covariance matrix. In Eq. (6), the factor $n^{1/2}L$ comes from free translation, while the remaining factor comes from vibrational motion. Note that because $\boldsymbol{\kappa}$ is singular, we employ $\boldsymbol{\kappa}^{-1}$ to represent the Moore-Penrose pseudoinverse acting in the $n-1$ dimensional space of vibrations. Similarly, we consider the determinant of this projection: $\det \boldsymbol{\kappa} = \lambda_1 \cdots \lambda_{n-1}$, where $\lambda_1, \ldots, \lambda_{n-1}$ are the $n-1$ positive vibrational eigenvalues of $\boldsymbol{\kappa}$.

The (dimensionless) excess configurational entropy, s, is then computed

$$s = -\int \mathrm{d}\mathbf{q}\, p(\mathbf{q}) \ln\left[L^n p(\mathbf{q})\right] \quad [7]$$

$$= (n-1)s_0 - \frac{1}{2}\ln t_{\boldsymbol{\kappa}} \quad [8]$$

where $s_0 = \frac{1}{2}\left(1 + \ln[2\pi/\beta\Gamma L^2]\right)$ is a protein-independent constant and $t_{\boldsymbol{\kappa}} = n^{-1}\det\boldsymbol{\kappa}$. We employ $h = \frac{1}{2}\ln t_{\boldsymbol{\kappa}}$ to quantify the "non-trivial" information stored in the equilibrium distribution for the high-resolution GNM.

The covariance matrix describing equilibrium fluctuations is

$$\mathbf{c} = \left\langle \mathbf{q}\mathbf{q}^\dagger \right\rangle = (\beta\Gamma\boldsymbol{\kappa})^{-1} \quad [9]$$

We quantify the "vibrational power" of the high-resolution GNM in terms of the mass-weighted fluctuations:

$$\sigma = \left\langle \sum_{i=1}^{n} m q_i^2 \right\rangle = \mathrm{Tr}_n m\mathbf{c} = k_B T \sum_{i=1}^{n-1} \omega_i^{-2} \quad [10]$$

where we have assigned a mass m to each atom, $\omega_i = \sqrt{\Gamma\lambda_i/m} > 0$ is the i^{th} vibrational frequency, and Tr_n indicates the trace over the atomic degrees of freedom.

2. Coarse-grained Representations

The mapping, $\mathbf{M}$, specifies a CG representation of the high resolution GNM by determining the configuration $\mathbf{Q} = (Q_1, \ldots, Q_N)$ for N CG degrees of freedom as a function of the high resolution configuration, $\mathbf{q} = (q_1, \ldots, q_n)$, for $n \geq N$ atomic degrees of freedom:

$$\mathbf{M} : \mathbf{q} \to \mathbf{Q} = \mathbf{M}(\mathbf{q}). \quad [11]$$

In the present work, we consider mappings that partition the n atoms into N disjoint groups and associate a CG "site" with each atomic group. The mapping determines the coordinate of each site as the mass center for the associated atomic group. In the present work we consider only atomic groups with the additional properties: (1) each atom is associated with only one atomic group, (2) each atomic group contains $R = n/N$ atoms, and (3) the bonds between atoms in each group form a connected network.

The equilibrium distribution, $p(\mathbf{q})$, of the high-resolution model and the mapping, $\mathbf{M}$, then specify a "mapped" ensemble in which each CG configuration has probability

$$P(\mathbf{Q};\mathbf{M}) = \int \mathrm{d}\mathbf{q}\, p(\mathbf{q})\delta(\mathbf{Q} - \mathbf{M}(\mathbf{q})) = z^{-1}L^{-N+n}\exp\left[-\beta W(\mathbf{Q})\right] \quad [12]$$

where $W(\mathbf{Q}) = W(\mathbf{Q};\mathbf{M})$ is the "exact" CG potential obtained by renormalizing the microscopic potential(6–8)

$$\exp\left[-\beta W(\mathbf{Q})\right] = L^{N-n}\int \mathrm{d}\mathbf{q}\, \exp\left[-\beta u(\mathbf{q})\right]\delta(\mathbf{Q} - \mathbf{M}(\mathbf{q}))\,. \quad [13]$$

Because

$$L^{-N}\int \mathrm{d}\mathbf{Q}\, \exp\left[-\beta W(\mathbf{Q})\right] = L^{-n}\int \mathrm{d}\mathbf{q}\, \exp\left[-\beta u(\mathbf{q})\right], \quad [14]$$

this definition ensures that the excess free energies of the CG and high-resolution models are equal.(9) The exact CG potential can be decomposed into energetic and entropic components(10, 11)

$$W(\mathbf{Q}) = U_W(\mathbf{Q}) - TS_W(\mathbf{Q}). \quad [15]$$

For the GNM, Eq. (13) can be analytically calculated.(10) The energetic and entropic components of W are

$$U_W(\mathbf{Q}) = \frac{1}{2}\Gamma\mathbf{Q}^\dagger\mathbf{K}\mathbf{Q} + \frac{1}{2}(n-N)k_BT \quad [16]$$

$$S_W(\mathbf{Q}) = \frac{1}{2}(n-N)s_0 + \frac{1}{2}\left(\ln T_{\mathbf{K}} - \ln t_{\boldsymbol{\kappa}}\right) \quad [17]$$

where $\mathbf{K} = \left(\mathbf{M}\boldsymbol{\kappa}^{-1}\mathbf{M}^\dagger\right)^{-1}$ is the renormalized Hessian and $T_{\mathbf{K}} = N^{-1}\det\mathbf{K}$.

The (dimensionless) excess entropy of the mapped ensemble is

$$S = -\int \mathrm{d}\mathbf{Q}\, P(\mathbf{Q};\mathbf{M})\ln\left[L^N P(\mathbf{Q};\mathbf{M})\right] = (N-1)s_0 - \frac{1}{2}\ln T_{\mathbf{K}}. \quad [18]$$

As for the microscopic model, we define $H = H(\mathbf{M}) = \frac{1}{2}\ln T_{\mathbf{K}}$ as the non-trivial information preserved in the mapped ensemble. Consequently, we define the "information quality" of the representation $\mathbf{M}$ by

$$I = I(\mathbf{M}) = H(\mathbf{M})/h = \ln T_{\mathbf{K}}/\ln t_{\boldsymbol{\kappa}}, \quad [19]$$

i.e., the fraction of the information in the microscopic ensemble that is preserved by $\mathbf{M}$.

The covariance in the mapped ensemble is

$$\mathbf{C} = \mathbf{C}(\mathbf{M}) = (\beta\Gamma\mathbf{K})^{-1} = \mathbf{M}\mathbf{c}\mathbf{M}^\dagger. \quad [20]$$

The vibrational power of the mapped ensemble is then

$$\Sigma = \Sigma(\mathbf{M}) = \left\langle \sum_{I=1}^{N} M Q_I^2 \right\rangle = \mathrm{Tr}_N M\mathbf{C} = k_BT\sum_{I=1}^{N-1}\Omega_I^{-2}, \quad [21]$$

where we have assigned a mass $M = mn/N$ to each CG site, $\Omega_I = \sqrt{\Gamma\Lambda_I/M} > 0$ is the I^{th} vibrational frequency, Λ_I is the I^{th} positive eigenvalue of $\mathbf{K}$, and Tr_N indicates the trace over the CG degrees of freedom. Consequently, we quantify the spectral quality of the representation $\mathbf{M}$ by

$$\mathcal{Q} = \mathcal{Q}(\mathbf{M}) = \Sigma(\mathbf{M})/\sigma, \quad [22]$$

i.e., the fraction of vibrational power in the microscopic ensemble that is preserved by $\mathbf{M}$.

Note that the metrics I and $\mathcal{Q}$ are bounded between 0 and 1, only equalling 1 in the limit that $N \to n$.

3. Connection to basic graph and network concepts

The GNM has a particularly simple and informative connection to basic concepts in the theories of graphs(12) and networks.(13) For each atom $i = 1, \ldots, n$ treated by the GNM, one associates a vertex v_i of a graph (or equivalently a node of a network), which we shall simply indicate by i. Between each pair of atoms, i and j, that are connected by a spring in the GNM potential, one associates an edge $e_{ij} = e_{ji}$ connecting the i and j vertices. The resulting vertex set

$$V = \{1, 2, \ldots, n\} \quad [23]$$

and edge set

$$E = \{e_{ij} | i, j \in V, \theta_{ij} = 1\} \quad [24]$$

define a graph, $G = (V, E)$, which we refer to as the (intramolecular) protein interaction network for a specific protein. The contact matrix, θ_{ij}, which specifies which atoms of the GNM are connected by springs, corresponds to the adjacency matrix of the protein interaction network. The curvature of the GNM potential, κ_{ij}, is the corresponding graph Laplacian. Because the bonds of the GNM form a connected network, the protein interaction network is also connected, i.e., the edges in E provide a path between any two vertices in V. Moreover, the quantity $t_{\boldsymbol{\kappa}}$, which determines the protein-specific contribution to the GNM configurational entropy, equals the number of distinct trees that span the protein interaction network according to the Kirchhoff's matrix-tree theorem. A spanning tree is a subgraph of a connected graph that connects all of the vertices in V with a subset of the edges in E and that includes no cycles.(12)

Simple graph concepts are also useful for considering CG mappings. In the present work, we consider maps, $\mathbf{M}$, that partition the n atoms of the GNM into N disjoint and connected atomic groups that each contain $R = n/N$ atoms. This partitioning corresponds to defining a set of N equally sized communities in the protein interaction network.(13) It is then convenient to associate the mapping, $\mathbf{M}$, for an N site CG model with N atomic groups $\mathbf{M} = (S_1, \ldots, S_N)$ where S_I is the I^{th} atomic group. The requirements that the atomic groups are equally sized, disjoint, and account for all the atoms correspond to the following criteria

$$|S_I| = R \qquad \text{for all } I \quad [25]$$

$$S_I \cap S_J = \emptyset \qquad \text{for all } I \neq J \quad [26]$$

$$\bigcup_{I=1}^{N} S_I = V = \{1, 2, \ldots, n\}. \quad [27]$$

where $|S_I|$ indicates the number of elements in S_I, i.e., the number of atoms associated with site I. Given the mapping $\mathbf{M} = (S_1, \ldots, S_N)$, it is convenient to define for each I and J

$$E_{IJ} = E_{IJ}(\mathbf{M}) = \{e_{ij} \in E | i \in S_I, j \in S_J\}. \quad [28]$$

In the case that $I \neq J$, E_{IJ} is the set of edges connecting atoms in site S_I to atoms S_J. This set plays an important role in the move-sets developed for exploring mapping space.

Given the mapping, $\mathbf{M} = (S_1, \ldots, S_N)$, it is useful to define for each site I a subgraph of the protein interaction network, $G_I = (S_I, E_{II})$, that is formed by connecting the vertices $i \in S_I$ with the edges, E_{II}, that are internal to the site I. The restriction to connected maps implies that the corresponding subgraphs G_I must be connected for each site $I = 1, \ldots, N$.

It is also useful to define articulation nodes, which become important when swapping atoms between sites in the course of "swap-based" moves in mapping space. In brief, an articulation node is a vertex that causes a connected graph to become disconnected upon the removal of the vertex and all edges connecting (i.e., adjacent to) the vertex. More precisely, consider a move that, starting from a connected map $\mathbf{M} = (S_1, \ldots, S_N)$, generates a new map $\mathbf{M}' = (S'_1, \ldots, S'_N)$, by exchanging a pair of atoms i and j between two sites I and J. This creates two new sites

$$S_I \to S'_I = S_{I-i} \cup \{j\} \quad [29]$$

$$S_J \to S'_J = S_{J-j} \cup \{i\} \quad [30]$$

where $S_{I-i} = S_I - \{i\}$ and $S_{J-j} = S_J - \{j\}$ indicate the sets of $R - 1$ vertices remaining in S_I and S_J after vertices i and j, respectively, have been removed. Similarly, we define

$$E_{I-i} = \{e_{kl} \in E | k, l \in S_{I-i}\} \quad [31]$$

$$E_{J-j} = \{e_{kl} \in E | k, l \in S_{J-j}\} \quad [32]$$

as the sets of edges that remain in E_{II} and E_{JJ} after removing any edges that connect to vertices $i \in S_I$ and $j \in S_J$, respectively. The vertex $i \in S_I$ is an articulation vertex of G_I if the graph $G_{I-i} = (S_{I-i}, E_{I-i})$ is disconnected. Similarly, the vertex $j \in S_J$ is an articulation vertex of G_J if the graph $G_{J-j} = (S_{J-j}, E_{J-j})$ is disconnected.

4. Sampling representations

A. Mapping space. The mapping, $\mathbf{M}$, specifies a particular CG representation of the underlying microscopic model. As noted above, the mapping defines the coordinate of each CG site as the mass center of an associated atomic group. Each mapping, $\mathbf{M}$, then corresponds to a partitioning of the n atoms into N disjoint connected groups $\mathbf{M} = (S_1, S_2, \dots, S_N)$ where S_I is the I^{th} atomic group. More precisely, we consider maps that satisfy the following properties

1. Each atomic group includes $R = n/N$ atoms, i.e., $|S_I| = R$ for all I.

2. Each atom is included in only one atomic group, i.e., $S_I \cap S_J = \emptyset$ for all $I \neq J$.

3. The atoms in each group are connected by a network of bonds, i.e., $G_I = (S_I, E_{II})$ is a connected subgraph for all I.

We denote by $\mathcal{S}$ the set of mappings that satisfy these 3 properties. In particular, the "block map," $\mathbf{M}_{\text{bl}} \in \mathcal{S}$, is defined by assigning atoms $i = 1, 2, \dots, R$ to group 1, assigning atoms $i = R+1, R+2, \dots, 2R$ to group 2, etc.

The following subsection defines a "swap-based" move-set for exploring mapping space starting from the block map, $\mathbf{M}_{\text{bl}}$. However, we have not proved that this move-set is ergodic in $\mathcal{S}$. Consequently, it is possible that there exist some maps $\mathbf{M} \in \mathcal{S}$ that cannot be reached from $\mathbf{M}_{\text{bl}}$ via the swap move-set. Thus, our exploration of mapping space is limited to the set of connected maps that can be reached from the block map via swap-moves, i.e., to the set

$$\mathcal{S}_{\text{bl}} = \{\mathbf{M} \in \mathcal{S} | d_{\text{MS}}(\mathbf{M}, \mathbf{M}_{\text{bl}}) < \infty\} \quad [33]$$

where $d_{\text{MS}}(\mathbf{M}, \mathbf{M}_{\text{bl}})$ is the minimum number of swap moves necessary to reach the map $\mathbf{M}$ starting from $\mathbf{M}_{\text{bl}}$.

The following subsection also considers a less restrictive "site-based" move-set. Numerical calculations indicate that both move-sets provide equivalent sampling, which suggests that the swap-based move-set may be ergodic for the class of protein GNM's that we consider. Note, though, that the main text and SI only present results for the swap-based move-set.

B. Move-sets. We consider two move-sets for exploring mapping space. Both consider moves from one connected map, $\mathbf{M} = (S_1, \dots, S_N) \in \mathcal{S}_{\text{bl}}$, to a new connected map, $\mathbf{M}' = (S'_1, \dots, S'_N) \in \mathcal{S}_{\text{bl}}$, in which 2 of the N sites have been redefined, while the remaining $N-2$ sites are unchanged. Moreover, both move-sets are reversible in the sense that if $\mathbf{M} \to \mathbf{M}'$ is allowed, then $\mathbf{M}' \to \mathbf{M}$ is also allowed. Consequently, $d_{\text{MS}}(\mathbf{M}, \mathbf{M}') = d_{\text{MS}}(\mathbf{M}', \mathbf{M})$ for both move-sets. Additionally, given a move-set MS, we define two maps, $\mathbf{M}$ and $\mathbf{M}'$, as neighbors if $d_{\text{MS}}(\mathbf{M}, \mathbf{M}') = 1$.

B.1. Swap-based. Given the map $\mathbf{M} = (S_1, \dots, S_N)$, the swap-based move-set consists of all connected maps, $\mathbf{M}' \in \mathcal{S}$, that can be constructed by swapping a pair of atoms between a pair of sites, while leaving the remaining sites unchanged. Operationally, this move-set is constructed as follows:

1. For each pair of distinct sites, S_I and S_J, defined by $\mathbf{M}$, we construct the set, $E_{IJ} = E_{IJ}(\mathbf{M})$, defined in Eq. (28)

2. We then construct the set, $T_{IJ}(\mathbf{M})$, enumerating all (unordered) pairs of edges, $[e_{ij}, e_{i'j'}]$, formed by 4 distinct atoms connecting the two sites:

$$T_{IJ}(\mathbf{M}) = \{[e_{ij}, e_{i'j'}] | e_{ij}, e_{i'j'} \in E_{IJ}(\mathbf{M}) \text{ with } i, i' \in S_I, j, j' \in S_J, \text{ and } i \neq i', j \neq j'\} \quad [34]$$

3. For each pair of edges $[e_{ij}, e_{i'j'}] \in T_{IJ}$ we consider two swaps that define moves to two new possible maps, $\mathbf{M}_1$ and $\mathbf{M}_2$:

   (a) swap $(i \leftrightarrow j')$: Define $S'_I = S_I - \{i\} \cup \{j'\}$ by replacing atom i with atom j', define $S'_J = S_J - \{j'\} \cup \{i\}$ by replacing atom j' with atom i, and define $\mathbf{M}_1$ by replacing S_I and S_J with S'_I and S'_J, respectively, while leaving the remaining $N-2$ sites unchanged.

   (b) swap $(i' \leftrightarrow j)$: Define $S'_I = S_I - \{i'\} \cup \{j\}$ by replacing atom i' with atom j, define $S'_J = S_J - \{j\} \cup \{i'\}$ by replacing atom j with atom i', and define $\mathbf{M}_2$ by replacing S_I and S_J with S'_I and S'_J, respectively, while leaving the remaining $N-2$ sites unchanged.

   (c) Check that the proposed new maps, $\mathbf{M}_1$ and $\mathbf{M}_2$, remain connected. Note that the two swaps ensure that the moved atoms are connected to at least one atom in their new site. Consequently, if a proposed swap does not move an articulation node, then the resulting map is allowed as a new move. However, if the proposed swap does move an articulation node, then the resulting site may be disconnected. In this case, the resulting map is only allowed if the atom replacing the articulation node ensures connectivity of the new site.

By performing steps 1-3 for each distinct pair of sites S_I and S_J defined by $\mathbf{M}$, we identify all allowed maps $\mathbf{M}'$ that can be generated from $\mathbf{M}$ via swap-based moves.

B.2. Site-based. Given the connected mapping $\mathbf{M} = (S_1, \ldots, S_N) \in \mathcal{S}$, the site-based move-set consists of all connected maps, $\mathbf{M}' \in \mathcal{S}$, that can be constructed by first merging a pair of sites, S_I and S_J, to form a "super-site" $\hat{S}_{IJ}$, and then splitting $\hat{S}_{IJ}$ into 2 new connected sites S'_I and S'_J, while leaving the remaining $N-2$ sites unchanged. In order to apply this move-set we first determine the following prior to simulation

1. all possible connected sites of R atoms that can be formed from the protein interaction network.

2. all possible super-sites of $2R$ distinct atoms that can be formed from merging two of these connected sites

In this way, we determine all possible pairs of connected sites $[S_1, S_2]$ that can be formed by splitting any relevant super-site, $\hat{S}$ into two disjoint groups each containing R atoms. Then, during the course of the simulation, given the mapping, $\mathbf{M} = (S_1, \ldots, S_N) \in \mathcal{S}$, the site-based move-set identifies possible moves, $\mathbf{M} \to \mathbf{M}'$ as follows:

1. For each pair of distinct sites, S_I and S_J, defined by $\mathbf{M}$, we construct the super-site $\hat{S}_{IJ}$.

2. Using our precomputed list, we identify each pair of new connected sites, S'_I and S'_J, that can be formed by splitting the super-site $\hat{S}_{IJ}$.

3. Each such division of the super-site $\hat{S}_{IJ}$ determines a new map, $\mathbf{M}'$, defined by replacing S_I and S_J with S'_I and S'_J, respectively, while leaving the $N-2$ remaining sites unchanged.

By performing steps 1-3 for each distinct pair of sites S_I and S_J defined by $\mathbf{M}$, we identify all allowed maps $\mathbf{M}'$ that can be generated from $\mathbf{M}$ via site-based moves.

C. Exhaustive enumeration. For sufficiently small proteins with sufficiently simple interaction networks, it is possible to exhaustively enumerate all possible CG representations in $\mathcal{S}_{\text{bl}}$ via a "breadth-first" search. In this breadth-first search we enumerate successive generations of new maps by identifying neighbors of previously identified maps. In this calculation, we only employed swap-moves to identify neighbors.

The "zeroth generation" list includes only the block map, $\mathbf{M}_{\text{bl}}$. We then generate a "first generation" list of all maps, $\mathbf{M}'$, that are neighbors of $\mathbf{M}_{\text{bl}}$. We then identify the neighbors of each first generation map in order to identify a list of "second generation" maps. In generating this second generation, we ensure that each second generation map is unique and also exclude maps identified in previous generations, i.e., the block and first generation maps. We continue in this manner creating lists of unique n^{th} generation maps that are not included in prior generations, until all maps that can be reached have been previously identified. The union of these generations then corresponds to the complete set of maps that can be reached from $\mathbf{M}_{\text{bl}}$, i.e., the union corresponds to $\mathcal{S}_{\text{bl}}$.

D. Monte Carlo methods. In most cases, it is not feasible to exhaustively all possible maps in $\mathcal{S}_{\text{bl}}$. Consequently, we employ Monte Carlo methods to more effectively explore and characterize the statistical properties of $\mathcal{S}_{\text{bl}}$ at each resolution.

D.1. Energy and Temperature. Since we are particularly interested in characterizing the information content, I, and spectral quality, $\mathcal{Q}$, of CG maps, we performed Monte Carlo simulations to sample maps, $\mathbf{M}$, while employing these metrics to determine dimensionless energy functions $\mathcal{E} = \mathcal{E}(\mathbf{M})$. Equilibrium Monte Carlo simulations will then sample $\mathbf{M} \in \mathcal{S}_{\text{bl}}$ from the distribution(14)

$$\mathcal{P}_{\mathbf{M}} = \exp\left[-\beta_{\mathcal{E}}\mathcal{E}(\mathbf{M})\right] / \, Q_{\mathcal{E}}(\beta_{\mathcal{E}}) \qquad [35]$$

where $\mathcal{E}$ is either $2H$ or $1-\mathcal{Q}$, $\beta_{\mathcal{E}}$ is the (inverse) temperature conjugate to $\mathcal{E}$, and the normalization is

$$Q_{\mathcal{E}}(\beta_{\mathcal{E}}) = \sum_{\mathbf{M} \in \mathcal{S}_{\text{bl}}} \exp\left[-\beta_{\mathcal{E}}\mathcal{E}(\mathbf{M})\right]. \qquad [36]$$

Note that equilibrium MC simulations sample maps with varying values of $\mathcal{E}$ as $\beta_{\mathcal{E}}$ is varied:

- simulations with $\beta_{\mathcal{E}} \to \infty$ primarily sample maps that minimize $\mathcal{E}$
- simulations with $\beta_{\mathcal{E}} > 0$ primarily sample maps with relatively small values of $\mathcal{E}$
- simulations with $\beta_{\mathcal{E}} \to 0$ primarily sample maps with characteristic values of $\mathcal{E}$
- simulations with $\beta_{\mathcal{E}} < 0$ primarily sample maps with relatively large values of $\mathcal{E}$
- simulations with $\beta_{\mathcal{E}} \to -\infty$ primarily sample maps that maximize $\mathcal{E}$.

Thus, by performing MC simulations at a range of inverse temperatures, $\beta_{\mathcal{E}}$, we sample maps $\mathbf{M} \in \mathcal{S}_{\text{bl}}$ covering the entire range for $\mathcal{E}$.

D.2. Detailed balance. In order to ensure that the MC simulations sample the distribution given by Eq. (35), we require that the simulations satisfy the detailed balance condition:

$$\Pr(\mathbf{M} \to \mathbf{M}') = \Pr(\mathbf{M}' \to \mathbf{M}) \quad [37]$$

where, at equilibrium, the probability for moving from a given map $\mathbf{M}$ to a new map $\mathbf{M}'$ is given by

$$\Pr(\mathbf{M} \to \mathbf{M}') = \mathcal{P}_{\mathbf{M}}\, \pi(\mathbf{M} \to \mathbf{M}') \quad [38]$$

and $\pi(\mathbf{M} \to \mathbf{M}')$ is the transition probability. We decompose the transition probability(14)

$$\pi(\mathbf{M} \to \mathbf{M}') = g(\mathbf{M} \to \mathbf{M}')\, \mathrm{Acc}(\mathbf{M} \to \mathbf{M}') \quad [39]$$

where $g(\mathbf{M} \to \mathbf{M}')$ is the probability of proposing the move to $\mathbf{M}'$ and $\mathrm{Acc}(\mathbf{M} \to \mathbf{M}')$ is the probability for accepting this move. In our simulations, we propose all allowed moves with equal probability such that

$$g(\mathbf{M} \to \mathbf{M}') = C_{\mathbf{M}}^{-1}\, 1_{\mathbf{M},\mathbf{M}'} \quad [40]$$

where $C_{\mathbf{M}}$ is the number of maps that neighbor $\mathbf{M}$ and $1_{\mathbf{M},\mathbf{M}'}$ is an indicator function that equals 1 if $\mathbf{M}$ and $\mathbf{M}'$ are neighbors and vanishes otherwise. Note that $C_{\mathbf{M}}$ and $g(\mathbf{M} \to \mathbf{M}')$ both depend upon the move-set employed in the MC simulations. More importantly, since we restrict our sampling to connected maps, $C_{\mathbf{M}}$ is not a constant, but instead depends upon $\mathbf{M}$. Consequently, in order to ensure detailed balance, we accept allowed moves with probability

$$\mathrm{Acc}(\mathbf{M} \to \mathbf{M}') = \frac{C_{\mathbf{M}}}{\max\{C_{\mathbf{M}}, C_{\mathbf{M}'}\}} \min\left\{1, \mathcal{P}_{\mathbf{M}'}/\mathcal{P}_{\mathbf{M}}\right\}. \quad [41]$$

We have found this acceptance probability useful, although other acceptance probabilities are possible as long as they satisfy Eq. (37), while accounting for Eq. (40).

D.3. Monte Carlo simulations. We performed equilibrium MC simulations to sample and characterize mapping space, $\mathcal{S}_{\mathrm{bl}}$, for each protein at each resolution, R. The majority of our simulations employed the swap-based move-set described in B.1, although we also performed simulations with the site-based move-set in order to test the convergence of our simulations. Each MC simulation employed either $2H(\mathbf{M})$ or $1 - \mathcal{Q}(\mathbf{M})$, as an energy function, $\mathcal{E} = \mathcal{E}(\mathbf{M})$, at a fixed conjugate (inverse) temperature $\beta_{\mathcal{E}}$. The combination of a specific energy function $\mathcal{E}$ and specific $\beta_{\mathcal{E}}$ determine a "state point" for our simulations.(15) We employed a finer spacing of conjugate temperatures to sample near variance peaks in the corresponding energy, while employing a wider spacing of temperatures to sample simpler regions of the energy landscape.

Given a map, $\mathbf{M}$, each step of the simulation involved three steps.

1. We enumerated all $C_{\mathbf{M}}$ possible neighbors of the current map, $\mathbf{M}$, according to the specified move-set.
2. We randomly selected one of these neighbors, $\mathbf{M}'$, according to the uniform distribution $g(\mathbf{M} \to \mathbf{M}')$ given by Eq. (40).
3. We accepted the proposed move $\mathbf{M} \to \mathbf{M}'$ with probability $\mathrm{Acc}(\mathbf{M} \to \mathbf{M}')$ given by Eq. (41), while remaining at map $\mathbf{M}$ with probability $1 - \mathrm{Acc}(\mathbf{M} \to \mathbf{M}')$.

Each simulation started from the N site block map, $\mathbf{M}_{\mathrm{bl}}$. We treated at least the first 10^4 MC steps as an equilibration period. Subsequently, we sampled maps after every tenth MC step.

D.4. Simulated annealing. Prior to performing equilibrium MC simulations, we first performed simulated annealing in order to determine the relevant range for each energy function $\mathcal{E}$ and to estimate the appropriate conjugate (inverse) temperatures that should be employed in equilibrium simulations. These simulations began at very large positive temperature (i.e., inverse temperature $\beta_{\mathcal{E}} = +\epsilon$ for some very small, positive ϵ). The temperature was gradually decreased in log-based steps towards 0 (i.e., until the inverse temperature $\beta_{\mathcal{E}}$ reached a maximum value, M, for some very large constant M>0). After each temperature decrease, the simulation continued via equilibrium MC steps at constant temperature until a pseudo-equilibrium was reached when the energy plateaued. At this point the temperature was decreased again.

By performing multiple (i.e., order 10) independent simulated annealing calculations, we determined the ground state mapping, $\mathbf{M}_{\mathcal{E}0}$, that minimized the energy, $\mathcal{E}$, as well as a first estimate for the low-energy side of the density of states. We performed corresponding simulated annealing studies for negative temperatures to determine the maximum value of the energy and estimate the high-energy side of the density of states. In addition to determining the relevant range of energies, these simulations provided guidance for the appropriate conjugate temperatures that should be employed in equilibrium MC simulations.

D.5. Subsampling. Given the correlated time series of maps, $\mathbf{M}$, sampled from equilibrium MC simulations with the energy function $\mathcal{E}$ at a conjugate temperature, $\beta_{\mathcal{E}}$, we first estimated the correlation length of the time series from the PyMBAR time-series module.(15) We then subsampled the time series according to twice the estimated correlation length. In cases that we performed multiple MC simulations at the same state point, we determined the maximum correlation length among these simulations. We then employed this rate to subsample all simulations at this state point. Furthermore, we completely discarded any data from MC simulations that appeared trapped in basins in the free energy landscape. After this pruning process, we typically obtained approximately 10^6 statistically independent samples for each protein at each conjugate temperature for both energy functions.

D.6. Densities of states. Based upon this dataset, we employed the PyMBAR package to estimate the statistical weight of each sampled map at each state point of interest.(15) We estimated the density of states by discretizing the relevant range for the corresponding energy (i.e., $\mathcal{E} = \frac{1}{2}hI$ or $1 - \mathcal{Q}$) and summing the statistical weights for the samples assigned to each bin in the high temperature (i.e., $\beta_{\mathcal{E}} \to 0$) limit. We employed a bin spacing of $\delta I = .001$ and $\delta\mathcal{Q} = .0025$ for estimating the densities of states for all proteins except 1UBQ, for which $\delta I = .0025$ and $\delta\mathcal{Q} = .005$. We obtained similar estimates for the densities of states when using a more sophisticated kernel density estimator. The temperature-dependent free energy surfaces were estimated from (the logarithm of) the total statistical weight for the maps in each bin at the appropriate temperature. The 1D DoS for each energy was then shifted to generate the DoS, $\Omega(\mathcal{O})$, for the corresponding order parameter $\mathcal{O} = I$ or $\mathcal{Q}$. The 2D DoS's were estimated by creating a two dimensional array $(\mathcal{Q}, I)$ of bins with the above spacing $\delta\mathcal{Q}$, δI, and summing the statistical weights in each bin in the high temperature limit.

D.7. Alignment of densities of states. Since the MBAR calculations estimate statistical weights rather than absolute probabilities, they can only determine the densities of states $\Omega(\mathcal{E})$ to within an unknown constant. In order to estimate this constant, we attempted to exhaustively enumerate the best maps, i.e., the maps that correspond to the density of states near the minimum value of $\mathcal{E}$. The procedure is quite similar to the "breadth-first" procedure for exhaustively enumerating maps and employs swap-moves to identify neighbors, as described in subsection 4.C. This procedure starts from the optimal, ground state map, $\mathbf{M}_0$, minimizing the energy $\mathcal{E}(\mathbf{M}_0) = \mathcal{E}_0$, and requires an criterion, $\mathcal{E}_{\text{thr}}$, for enumerating maps.

Starting from $\mathbf{M}_0$, we identify a first generation of all unique maps, $\mathbf{M}$, that neighbor $\mathbf{M}_0$ and also lie below the threshold, $\mathcal{E}(\mathbf{M}) \leq \mathcal{E}_{\text{thr}}$. We then repeat the process for each map in this first generation in order to obtain a second generation of new maps for which $\mathcal{E}(\mathbf{M}) \leq \mathcal{E}_{\text{thr}}$. This procedure is repeated for successive generations until the next generation is empty because all potential members, $\mathbf{M}$, of the next generation have either been previously enumerated or they lie above the threshold, i.e., $\mathcal{E}(\mathbf{M}) > \mathcal{E}_{\text{thr}}$. This procedure is then successively repeated with increasing threshold, $\mathcal{E}_{\text{thr}}$, until the procedure does not terminate within a set time. The resulting enumerated maps are then used to estimate the density of states for energies $\mathcal{E}$ slightly greater than $\mathcal{E}_0$. For some range, $\mathcal{E}_0 \leq \mathcal{E} \leq \mathcal{E}_{\text{thr}}$, the enumerated density of states parallels the density of states obtained from the MBAR calculation. We then vertically shift the MBAR density of states to match the enumerated density of states in this range.

D.8. Statistical uncertainty. We estimated statistical uncertainties via bootstrapping. We resampled (with replacement) from the original data set of subsampled maps at each simulated state point. We repeated the MBAR calculation with this resampled data in order to obtain a new estimate for the statistical weight of each map at each state point. We then employed these new statistical weights to estimate the densities of states, as well as each observable of interest. We repeated this process 100 independent times. The reported uncertainties are the standard deviations from these 100 calculations of each observable.

5. Observables

A. Radius of gyration. Given a three-dimensional equilibrium PDB structure, $\mathbf{r}^*$, for a protein, we define $r^*_{i\alpha}$ as the α Cartesian coordinate of atom i in the PDB structure. A CG mapping, $\mathbf{M}$, then specifies a CG representation, $\mathbf{R}^*$, of the PDB structure. We define $R^*_{I\alpha}$ as the α Cartesian coordinate for the CG site S_I in the mapped structure. We then define the gyration tensor, $G_I = G_I(\mathbf{M})$, for CG site S_I:

$$G_{I;\alpha\gamma} = G_{I;\alpha\gamma}(\mathbf{M}) = \frac{N}{n}\sum_{i\in S_I} \delta r^*_{i\alpha}\delta r^*_{i\gamma} \qquad [42]$$

where $\delta r^*_{i\alpha} = r^*_{i\alpha} - R^*_{I\alpha}$ and $1 \leq \alpha, \gamma \leq 3$. The radius of gyration of site I in mapping $\mathbf{M}$ is given by

$$R^2_{G;I}(\mathbf{M}) = \sum_{\alpha=1}^{3} \lambda^2_{G_I;\alpha} \qquad [43]$$

where $\lambda_{G_I;\alpha}$ is the α eigenvalue of the gyration tensor, G_I. We then define

$$R_G = R_G(\mathbf{M}) = N^{-1}\sum_{I=1}^{N} R_{G;I}(\mathbf{M}). \qquad [44]$$

Finally, in the main text, we present the mean radius of gyration as a function of temperature:

$$R_G(T) = \sum_{\mathbf{M}} \mathcal{P}_{\mathbf{M}}(T) R_G(\mathbf{M}) \qquad [45]$$

where $T = T_{\mathcal{E}}$ is the temperature conjugate to $\mathcal{E} = 1 - \mathcal{Q}$.

B. Variation of information. As noted above, we consider CG mappings that correspond to partitions of the n atoms among N CG sites. The variation of information (VI) provides a formal metric for quantifying the "distance" between two mappings that is commonly used to compare different partitions of sets.(16)

Consider a mapping, $\mathbf{M} = (S_1, \ldots, S_N)$, of n atoms into N sites. We define $P_I(\mathbf{M})$ as the probability of randomly picking (according to a uniform distribution) an atom, i, that is associated with site I. Thus,

$$P_I(\mathbf{M}) = n^{-1}|S_I|, \quad [46]$$

where $|S_I|$ denotes the size of the set S_I, i.e., the number of atoms associated with site I. The information associated with this partitioning is then

$$H_1(\mathbf{M}) = -\sum_{I=1}^{N} P_I(\mathbf{M}) \log P_I(\mathbf{M}). \quad [47]$$

In the present work, we consider only maps for which all sites correspond to an equal number of atoms. Consequently, in this work $P_I(\mathbf{M}) = N^{-1}$ for all I and any $\mathbf{M}$, such that $H_1(\mathbf{M}) = \log N$ for any mapping $\mathbf{M}$ with N sites.

Now consider two distinct mappings, $\mathbf{M} = (S_1, \ldots, S_N)$ and $\mathbf{M}' = (S'_1, \ldots, S'_{N'})$ that map the n atoms to N and to N' sites, respectively. We define $P_{II'}(\mathbf{M}, \mathbf{M}')$ as the probability for randomly picking (according to a uniform distribution) an atom i that is associated with site I in mapping $\mathbf{M}$ and also associated with site I' in mapping $\mathbf{M}'$. Thus,

$$P_{II'}(\mathbf{M}, \mathbf{M}') = n^{-1}|S_I \cap S_{I'}| = P_{I'I}(\mathbf{M}', \mathbf{M}), \quad [48]$$

where $|S_I \cap S_{I'}|$ is the number of atoms that are mapped to site I by $\mathbf{M}$ and are also mapped to site I' by $\mathbf{M}'$. Note that

$$\sum_{I'=1}^{N'} P_{II'}(\mathbf{M}, \mathbf{M}') = P_I(\mathbf{M}). \quad [49]$$

The total information(17) stored in the distribution $P_{II'}$ is

$$H_2(\mathbf{M}, \mathbf{M}') = -\sum_{I=1}^{N}\sum_{I'=1}^{N'} P_{II'}(\mathbf{M}, \mathbf{M}') \log P_{II'}(\mathbf{M}, \mathbf{M}') \quad [50]$$

the mutual information, MI, shared between the two mappings is

$$\mathrm{MI}(\mathbf{M}, \mathbf{M}') = -\sum_{I=1}^{N}\sum_{I'=1}^{N'} P_{II'}(\mathbf{M}, \mathbf{M}') \log \left[\frac{P_{II'}(\mathbf{M}, \mathbf{M}')}{P_I(\mathbf{M}) P_{I'}(\mathbf{M}')} \right]. \quad [51]$$

We define the distance $d(\mathbf{M}, \mathbf{M}')$ between $\mathbf{M}$ and $\mathbf{M}'$ as VI:

$$d(\mathbf{M}, \mathbf{M}') \equiv \mathrm{VI}(\mathbf{M}, \mathbf{M}') \equiv H_2(\mathbf{M}, \mathbf{M}') - \mathrm{MI}(\mathbf{M}, \mathbf{M}') \quad [52]$$

$$= H_1(\mathbf{M}) + H_1(\mathbf{M}') - 2\mathrm{MI}(\mathbf{M}, \mathbf{M}'). \quad [53]$$

Note that VI allows one to quantify distances between mappings with different numbers of sites, i.e., for which $N \neq N'$. However, in the present work we only compare mappings with the same number of sites.

C. Modularity. As described in the main text and elaborated upon in Section 3 of this SI, the process of coarse-graining the GNM is very closely related to the process of clustering edges in a graph or defining communities in a network. The atoms and springs of the microscopic GNM correspond to the vertices and edges, respectively, of the graph that defines the underlying network. The Hessian of the microscopic GNM potential, $\kappa_{ij} = n_i\delta_{ij} - \theta_{ij}$, corresponds to the graph Laplacian, L_{ij}; the contact matrix of the GNM, θ_{ij}, corresponds to the adjacency matrix of the graph, A_{ij}; and the number of contacts formed by atom i, $n_i = \sum_{j(\neq i)} \theta_{ij}$, corresponds to the degree, k_i, of vertex i. The total number of edges in the network is then $m = \frac{1}{2}\sum_i n_i$.

The process of coarse-graining the GNM represents the n original atoms with N CG sites, which we shall denote here $C_1, \ldots, C_N$. In particular, we consider maps, $\mathbf{M}$, that associate each atom, i, with a unique CG site, which we shall denote $\widehat{C}_i \in \{C_1, \ldots, C_N\}$. This corresponds to partitioning the n vertices of the underlying graph into N communities. Newman and Girvan(18) proposed quantifying the "strength" of the resulting communities according to the modularity:

$$Q(\mathbf{M}) = \frac{1}{2m}\sum_{(i,j)} \left[\theta_{ij} - \frac{n_i n_j}{2m} \right] \delta(\widehat{C}_i, \widehat{C}_j), \quad [54]$$

where the sum is performed over all vertex pairs, while $\delta(\widehat{C}_i, \widehat{C}_j) = 1$ if the atoms (nodes) i and j are mapped to the same CG site (community) and otherwise vanishes.(13)

D. Essential dynamics coarse-graining. It is also instructive to compare the present work with the essential dynamics coarse-graining (EDCG) methodology.(19) The EDCG method partitions the n atoms into N coherently moving atomic groups based upon analyzing the "essential dynamics" (ED) subspace(20) that is defined from the covariance matrix, $\mathbf{c}_{\mathrm{MD}}$, of an atomically detailed molecular dynamics (MD) trajectory.

We label the atoms $i, j = 1, \ldots, n$ and Cartesian directions $d, d' = 1, 2, 3$. Given n_t configurations $\mathbf{r}(t) = (r_{id}(t))$ sampled from a trajectory, one eliminates any overall translation and rotational motion. The MD covariance matrix is a $3n \times 3n$ matrix, $\mathbf{c}_{\mathrm{MD}} \in \mathbb{R}^{3n} \times \mathbb{R}^{3n}$, with elements:

$$c_{\mathrm{MD}}(i_d, j_{d'}) = n_t^{-1} \sum_{t=1}^{n_t} \Delta r_{id}(t) \Delta r_{jd'}(t), \qquad [55]$$

where $\Delta r_{id}(t) = r_{id}(t) - \langle r_{id} \rangle$ quantifies the displacement of atom i from its average position (relative to the mass center) in configuration $\mathbf{r}(t)$. Because $\mathbf{c}_{\mathrm{MD}}$ is symmetric, its eigenvectors, $\{\boldsymbol{\eta}_q\}$, form a complete orthonormal basis and

$$\mathbf{c}_{\mathrm{MD}} = \sum_q \boldsymbol{\eta}_q \mu_q \boldsymbol{\eta}_q^\dagger, \qquad [56]$$

where we have sorted the corresponding eigenvalues, $\{\mu_q\}$, in order of decreasing magnitude. In practice, the eigenvalues quickly decay and a relatively small number, n_{ED}, of eigenvectors dominate $\mathbf{c}_{\mathrm{MD}}$. The ED subspace is then defined by these dominant eigenvectors. In particular, the projection operator

$$\mathbb{P}_{\mathrm{ED}} = \sum_{q=1}^{n_{\mathrm{ED}}} \boldsymbol{\eta}_q \boldsymbol{\eta}_q^\dagger \qquad [57]$$

defines motion in the ED subspace

$$\Delta \mathbf{r}_{\mathrm{ED}}(t) = \mathbb{P}_{\mathrm{ED}} \Delta \mathbf{r}(t). \qquad [58]$$

The EDCG methodology attempts to group atoms into CG sites such that each atomic group moves coherently in the ED subspace. In practice, the EDCG methodology minimizes the residual:

$$\chi^2(\mathbf{M}) = \frac{1}{3N} \sum_{I=1}^{N} \sum_{d=1}^{3} \frac{1}{n_t} \sum_{t=1}^{n_t} \left(\sum_{i \in S_I} \sum_{j \geq i \in S_I} | \Delta r_{\mathrm{ED};id}(t) - \Delta r_{\mathrm{ED};jd}(t) |^2 \right) \qquad [59]$$

$$= \frac{1}{3N} \sum_{I=1}^{N} \sum_{d=1}^{3} \sum_{i \in S_I} \sum_{j \geq i \in S_I} (c_{\mathrm{ED}}(i_d, i_d) - 2c_{\mathrm{ED}}(i_d, j_d) + c_{\mathrm{ED}}(j_d, j_d)) \qquad [60]$$

where $i \in S_I$ indicates the atoms i that are mapped to CG site I by the map, $\mathbf{M} = (S_1, \ldots, S_N)$, and $\mathbf{c}_{\mathrm{ED}} = \mathbb{P}_{\mathrm{ED}} \mathbf{c}_{\mathrm{MD}} \mathbb{P}_{\mathrm{ED}}$.

In the context of the present work, the MD covariance matrix, $\mathbf{c}_{\mathrm{MD}}$ in Eq. (55) is replaced by the GNM covariance matrix, $\mathbf{c} \in \mathbb{R}^n \times \mathbb{R}^n$, in Eq. (9):

$$\mathbf{c} = \left\langle \mathbf{q}\mathbf{q}^\dagger \right\rangle = \sum_q \boldsymbol{\eta}_q \left(\beta \Gamma \lambda_q\right)^{-1} \boldsymbol{\eta}_q^\dagger \qquad [61]$$

where $\boldsymbol{\eta}_q$ and λ_q are the eigenvectors and eigenvalues, respectively, of the Kirchoff matrix, $\boldsymbol{\kappa}$. These eigenvectors then define the ED subspace according to Eq. (57) and $\mathbf{c}_{\mathrm{ED}} = \mathbb{P}_{\mathrm{ED}} \mathbf{c} \mathbb{P}_{\mathrm{ED}}$ as before. The EDCG residual then becomes:

$$\chi^2(\mathbf{M}) = \frac{1}{N} \sum_{I=1}^{N} \sum_{i \in S_I} \sum_{j \geq i \in S_I} (c_{\mathrm{ED};ii} - 2c_{\mathrm{ED};ij} + c_{\mathrm{ED};jj}) \qquad [62]$$

6. Additional results

A. Model Proteins. The main text focuses on results for a 40 residue three-helix bundle protein with PDBID 2ERL. In this Supporting Information (SI) document, we present similar results for an additional 6 proteins with varying size and structure. Table 1 lists these proteins. Supporting Figures S1 and S2 characterize these 6 proteins. The left panels of these figures present the corresponding three-dimensional folded structures. The right panels combine the Kirchoff matrix, $\boldsymbol{\kappa}$, with the (scaled) covariance matrix, $\beta \Gamma \mathbf{c} = \boldsymbol{\kappa}^{-1}$, for each protein. Supporting Figures S3 and S4 present the DoS's for $\ln \Omega(I)$ and $\ln \Omega(\mathcal{Q})$, respectively, for the three smaller proteins described by Supporting Fig. S1. Supporting Figures S5 and S6 present the DoS's for $\ln \Omega(I)$ and $\ln \Omega(\mathcal{Q})$, respectively, for the three larger proteins described by Supporting Fig. S2. The DoS's for the larger proteins are only determined for positive temperatures.

B. Characterizing optimal maps. The present subsection provides further analysis of "optimal" maps. Supporting figures S7, S8, and S9 present the optimal maps for CG models of the proteins 3HJD, 1IJU, and 3E7R, respectively, with the indicated number of sites. In these three figures, the top and bottom rows present the maps that maximize $\mathcal{Q}$ and I, respectively. Supporting figures S10, S11, and S12 present the maps for CG models that maximize $\mathcal{Q}$ for the proteins 1UG4, 2V1Q, and 1UBQ, respectively, with the indicated number of sites. These figures reinforce the results for 2ERL that are presented in Fig. 3 of the main text. The maps that maximize $\mathcal{Q}$ tend to form compact, localized sites, while maps that maximize I tend to form loose, distributed sites.

Supporting figures S13, S14, and S15 present the 10 maps with maximal spectral fitness in $N = 2$, 4, and 8 site representations. These figures demonstrate that the optimal ten clusterings correspond to similar clusterings, although there is notable variation.

C. Correlations with spectral fitness. The present subsection provides insight into the characteristic properties of "good" maps. Specifically, we present scatter plots indicating the correlations that are observed among sampled maps. Supporting figure S17 presents the correlation of $\mathcal{Q}$ with the "size" of each map as defined by the $\mathrm{R_G}(\mathbf{M})$ metric, which is defined in subsection 5.A. Supporting figure S17 presents the correlation of $\mathcal{Q}$ with the distance $d_0(\mathbf{M}) = \mathrm{VI}(\mathbf{M}, \mathbf{M}_0)$ of a map, $\mathbf{M}$, from the "ground state" map, $\mathbf{M}_0$, that maximizes $\mathcal{Q}$, which is defined in subsection 5.B. Supporting figure S18 presents the correlation of $\mathcal{Q}$ with the modularity, $Q(\mathbf{M})$, of the associated clustering, which is defined in subsection 5.C. Table S2 presents the best fit lines and R^2 values that characterize each correlation.

The spectral quality, $\mathcal{Q}$, of a map is strongly anti-correlated with both its size, $\mathrm{R_G}$, and also with its distance, d_0, from the ground state. Interestingly, the $\mathcal{Q}$ - d_0 correlation appears to indicate two different regimes. In particular, the slope of this correlation becomes steeper very near the ground state map. Moreover, as $\mathcal{Q}$ approaches it maximum value for the given resolution, the d_0 distribution broadens significantly and develops a long "tail" towards the ground state map, $\mathbf{M}_0$. Thus, there is considerable variation among clusterings with high spectral quality, as suggested by Supporting figures S13-S15. This also explains the maximum in $\mathrm{var}(d_0)$ at very low temperature, which is presented in Fig. 4 of the main text. Conversely, the spectral quality is positively correlated with the modularity, Q, except at the highest resolution, for which there appears to be little correlation. As the resolution decreases, the slope of $\mathcal{Q}$ - Q correlation systematically increases. Therefore, it appears that $\mathrm{R_G}$ and Q may prove most useful for identifying the mapping with optimal spectral quality.

D. Relation to essential-dynamics coarse-graining. It is instructive to compare the spectral quality to the metric, χ^2, that is adopted by the EDCG methodology.(19) Subsection 5.D defines χ^2 and describes the EDCG methodology in detail. Supporting figure S19 presents a scatter plot indicating the correlation between $\mathcal{Q}$ and χ^2 for 4 different model proteins and various resolutions $R = n/N$. Clearly, χ^2 is strongly anti-correlated with $\mathcal{Q}$ at all but the highest resolutions. Table S2 quantifies this correlation. Thus, maps with high spectral quality define atomic groups that move rigidly within the essential dynamics subspace. Moreover, the slope characterizing this correlation becomes increasingly steep with increased coarsening, R.

Based upon the correlation between $\mathcal{Q}$ and χ^2, we employed our sampled maps to estimate the density of states for the EDCG metric, $\Omega(\chi^2)$. Specifically, given the maps sampled from MC simulations employing $\mathcal{E} = 1 - \mathcal{Q}$ as an energy function, we constructed histograms for χ^2 based upon the statistical weight for each sampled map in the $T_{\mathcal{Q}} \to \infty$ limit. Supporting figure S20 presents the resulting estimate for the density of states, $\Omega(\chi^2)$ for different model proteins. These densities of states also demonstrate noticeable inflection points. This suggests that similar phase transitions would be observed if χ^2 were adopted as the primary metric for characterizing the landscape of CG representations. Thus, we expect that the findings of the main text will be quite robust and apply for a wide variety of metrics that are employed to identify coherently moving atomic groups.

E. Sensitivity to cutoff. The microscopic GNM employed a cut-off $R_c = 7.5$Å to determine the microscopic contact matrix, θ_{ij}. This subsection investigates the sensitivity of our findings to this cut-off.

Specifically, we constructed two additional microscopic GNM's for the model protein 2ERL, which employed cut-offs of $R_c = 6.0$ Å and $R_c = 10.0$Å. We performed corresponding MC simulations for both of these new GNM's. Supporting figure S21 characterizes these additional simulations. The GNM with the longest cut-off (right column) undergoes phase transitions at the resolutions $R = 20, 10$ and 8, which are precisely the same resolutions for which phase transitions are observed in the original GNM with $R_c = 7.5$Å. The GNM with the shortest cut-off (left column) only undergoes phase transitions at the resolutions $R = 20$ and $R = 10$. We hypothesize that this effect is due to the fact that there is less information present in the $R_c = 6.0$Å model, and, consequently, a coarser resolution is required in order to distinguish between the two phases. Nevertheless, all three GNM's for 2ERL demonstrate qualitatively similar phase behavior with only minor changes in the critical resolution. Thus, we conclude that the results of the main text are robust with respect to minor variations in the definition of the microscopic GNM.

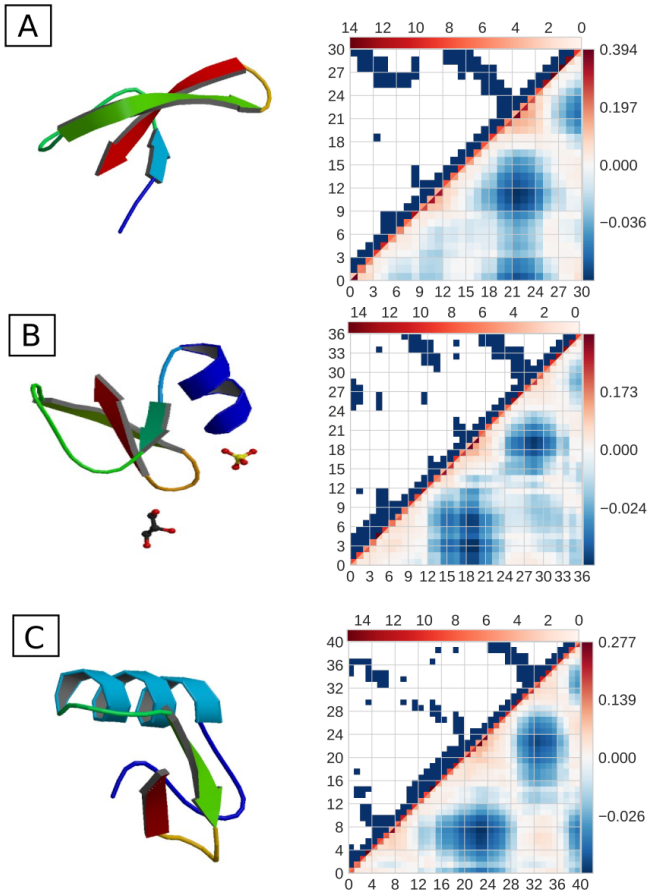

Fig. S1. Characterization of the additional small proteins: 3HJD (A), 1IJU (B) and 3E7R (C). (Left) Cartoon representations of the equilibrium folded structures. (Right) Intensity plots of the upper and lower halves of the symmetric connectivity, $\boldsymbol{\kappa}$, and covariance, $\mathbf{c} = \boldsymbol{\kappa}^{-1}$, matrices, respectively.

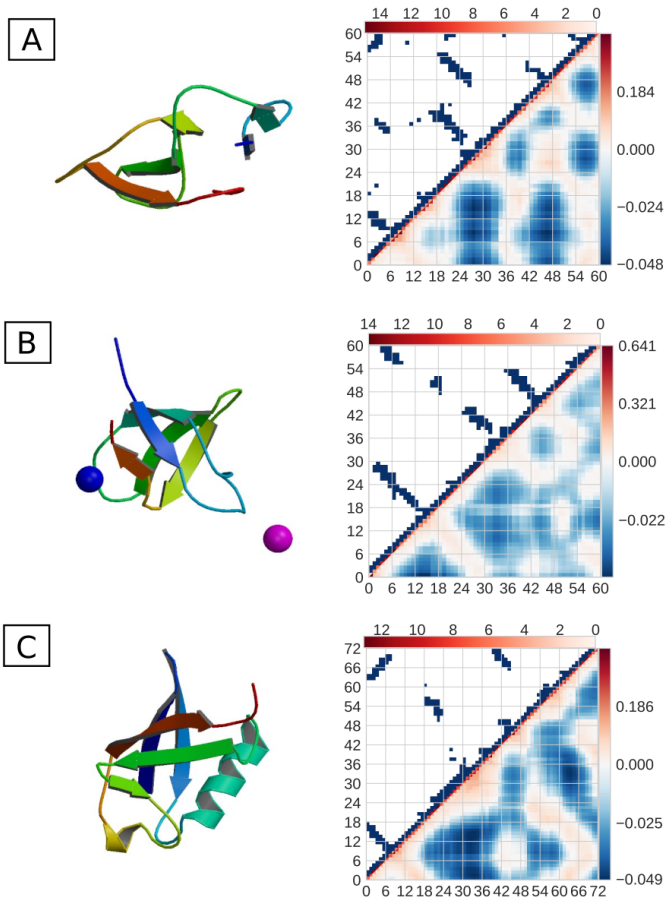

Fig. S2. Characterization of the large proteins: 1UG4 (A), 2V1Q (B) and 1UBQ (C). (Left) Cartoon representations of the equilibrium folded structures. (Right) Intensity plots of the upper and lower halves of the symmetric connectivity, $\boldsymbol{\kappa}$, and covariance, $\mathbf{c} = \boldsymbol{\kappa}^{-1}$, matrices, respectively.

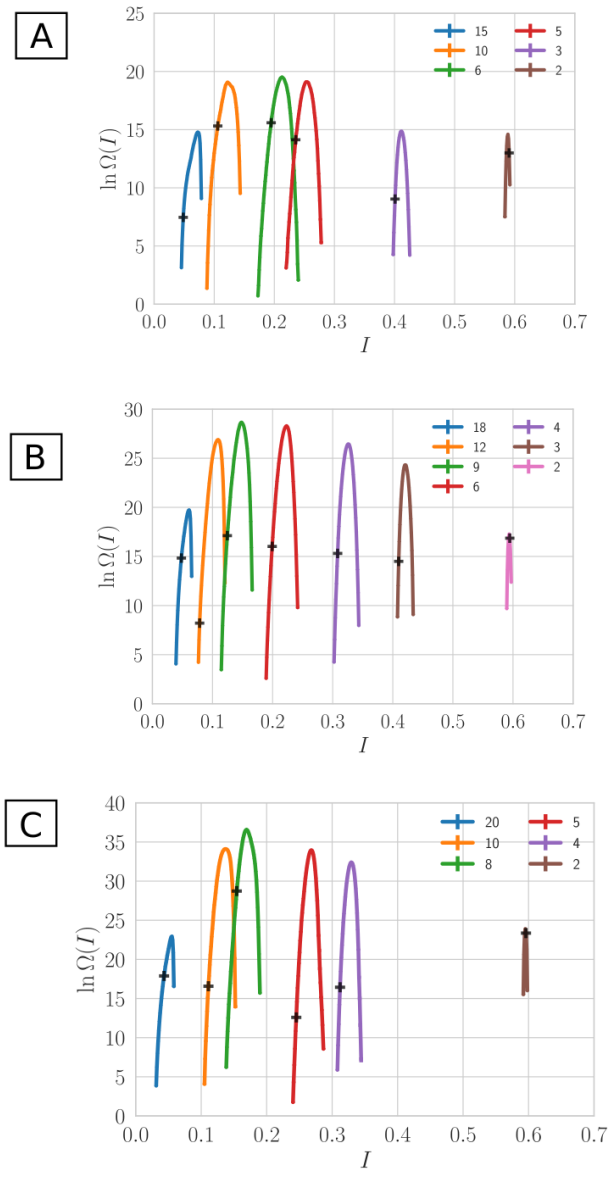

Fig. S3. Statistical analysis of mapping space for the additional small proteins: 3HJD (A), 1IJU (B) and 3E7R (C). The (logarithm of) the density of states Ω quantifying the number of maps, $\mathbf{M}$, with given information content, I at varying resolutions, $R = n/N$, indicated by the colors of the legend. The black crosses ('+') indicate I for the block map at each resolution.

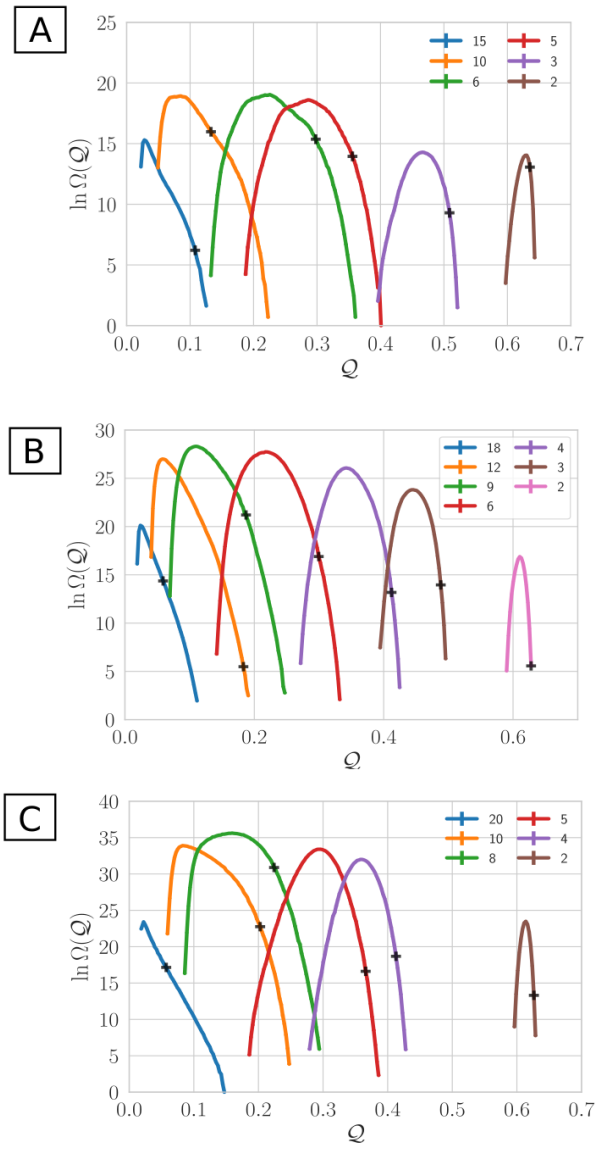

Fig. S4. Statistical analysis of mapping space for the additional small proteins: 3HJD (A), 1IJU (B) and 3E7R (C). The (logarithm of) the density of states Ω quantifying the number of maps, $\mathbf{M}$, with given spectral quality, $\mathcal{Q}$, at varying resolutions, $R = n/N$, indicated by the colors of the legend. The black crosses ('+') indicate $\mathcal{Q}$ for the block map at each resolution.

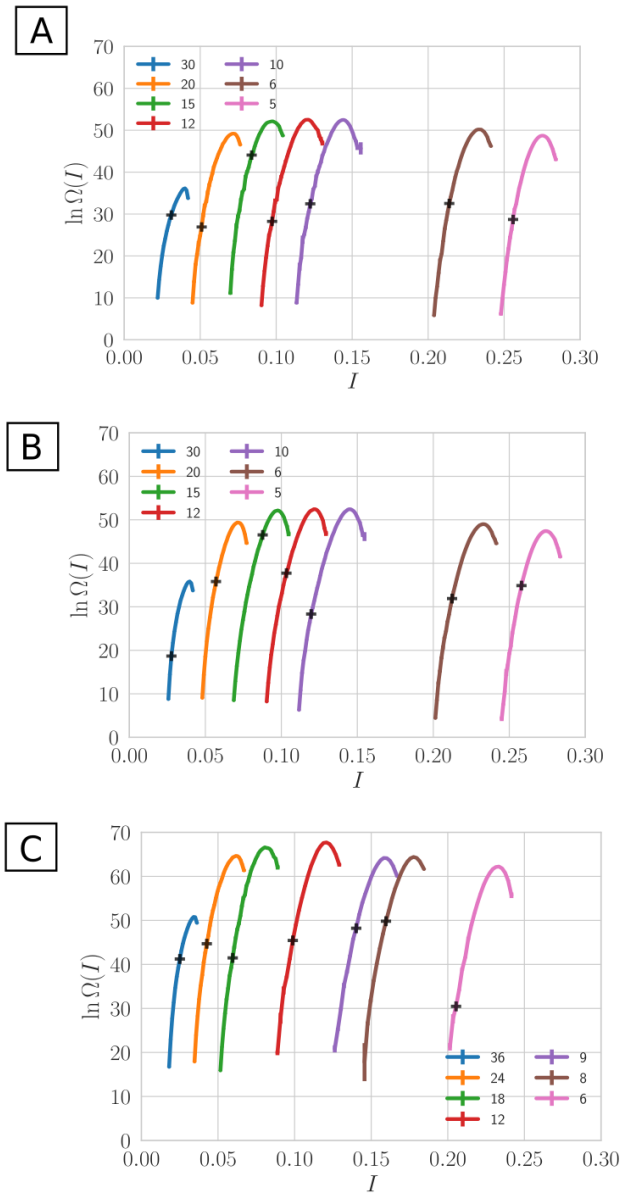

Fig. S5. Statistical analysis of mapping space for the large proteins: 1UG4 (A), 2V1Q (B) and 1UBQ (C). The (logarithm of) the density of states Ω quantifying the number of maps, $\mathbf{M}$, with given information content, I at varying resolutions, $R = n/N$, indicated by the colors of the legend. The black crosses ('+') indicate I for the block map at each resolution.

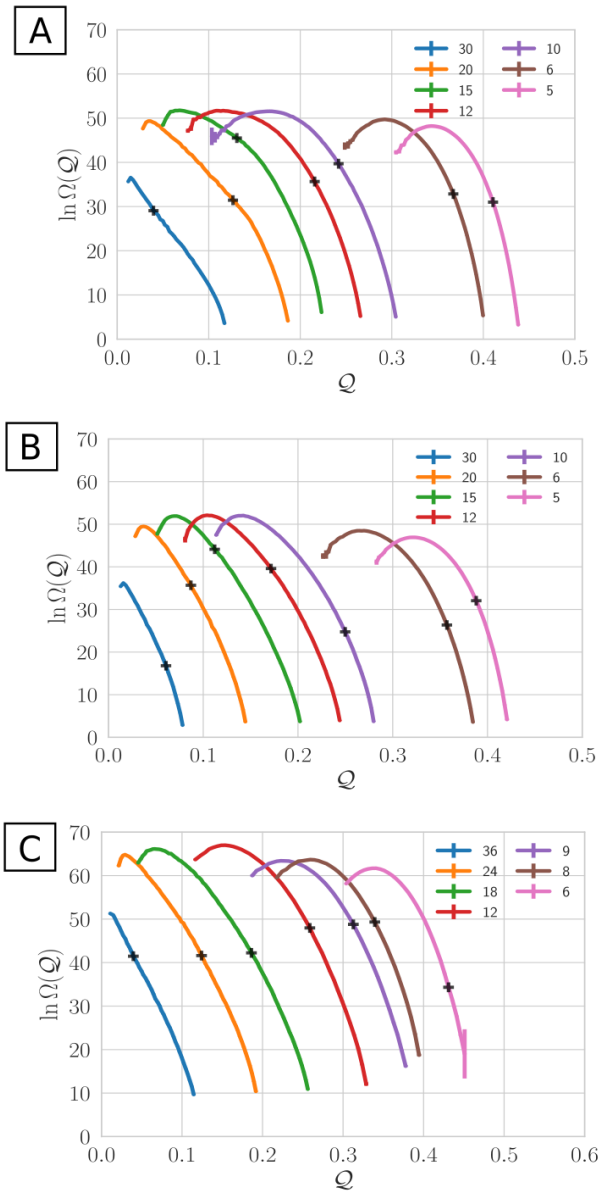

Fig. S6. Statistical analysis of mapping space for the large proteins: 1UG4 (A), 2V1Q (B) and 1UBQ (C). The (logarithm of) the density of states Ω quantifying the number of maps, $\mathbf{M}$, with given spectral quality, $\mathcal{Q}$, at varying resolutions, $R = n/N$, indicated by the colors of the legend. The black crosses ('+') indicate $\mathcal{Q}$ for the block map at each resolution.

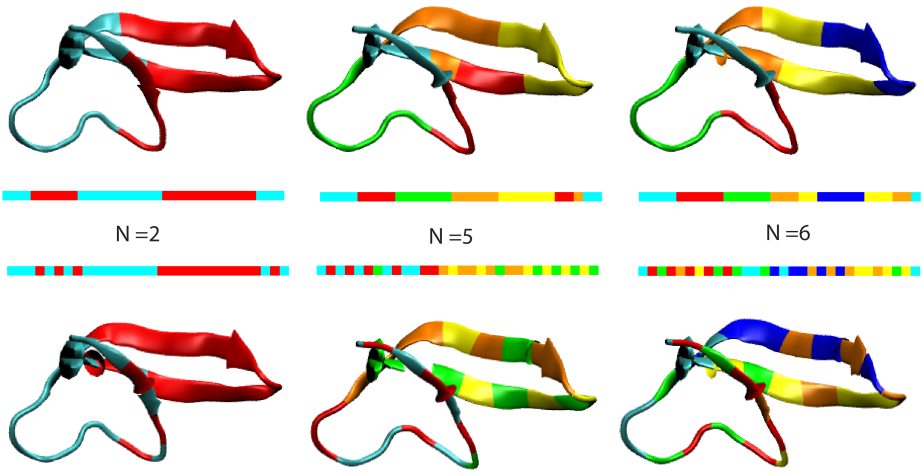

Fig. S7. CG representations with maximal $\mathcal{Q}$ (top) and I (bottom) for CG models of 3HJD with $N = 2$, 5, and 6 CG sites. The representations are indicated by assigning the same color to each residue in the same CG site. The bar graphs indicate the linear sequence of the protein, while the cartoons indicates its equilibrium three-dimensional structure.

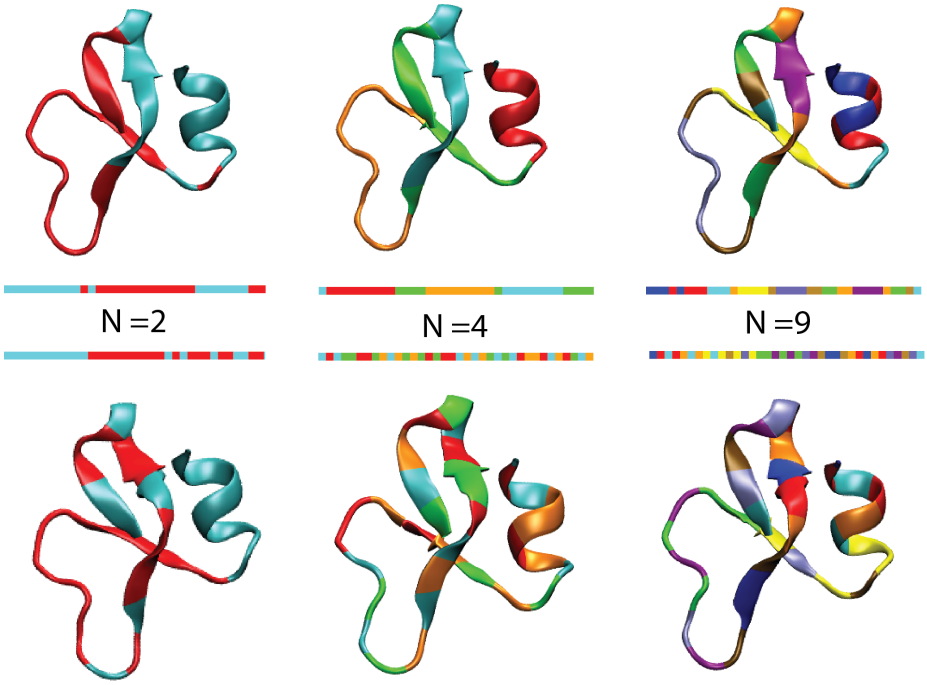

Fig. S8. CG representations with maximal $\mathcal{Q}$ (top) and I (bottom) for CG models of 1IJU with $N = 2$, 4, and 9 CG sites. The representations are indicated by assigning the same color to each residue in the same CG site. The bar graphs indicate the linear sequence of the protein, while the cartoons indicates its equilibrium three-dimensional structure.

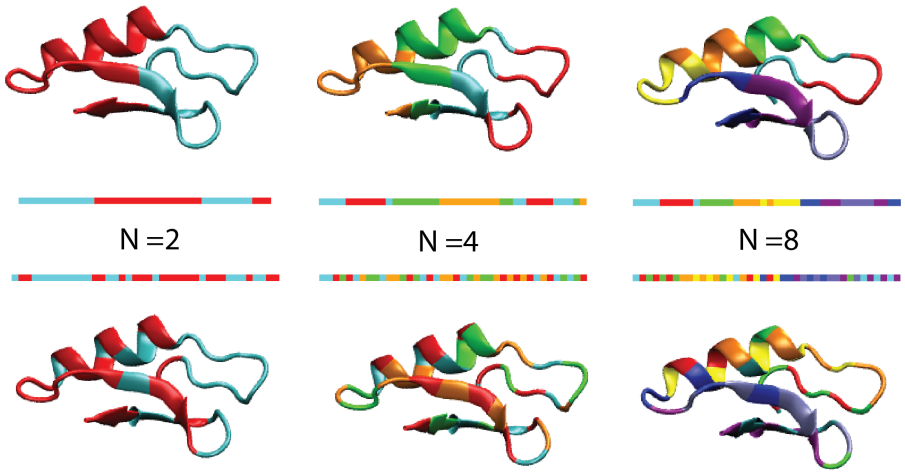

Fig. S9. CG representations with maximal $\mathcal{Q}$ (top) and I (bottom) for CG models of 3E7R with $N = 2$, 4, and 8 CG sites. The representations are indicated by assigning the same color to each residue in the same CG site. The bar graphs indicate the linear sequence of the protein, while the cartoons indicates its equilibrium three-dimensional structure.

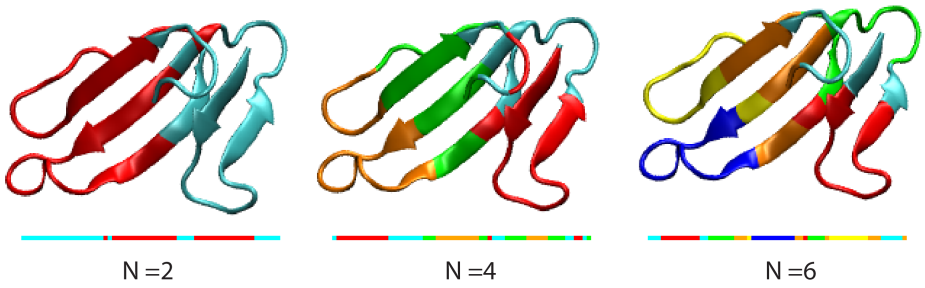

Fig. S10. CG representations with maximal $\mathcal{Q}$ for CG models of 1UG4 with $N = 2$, 4, and 6 CG sites. The representations are indicated by assigning the same color to each residue in the same CG site. The bar graphs indicate the linear sequence of the protein, while the cartoons indicates its equilibrium three-dimensional structure.

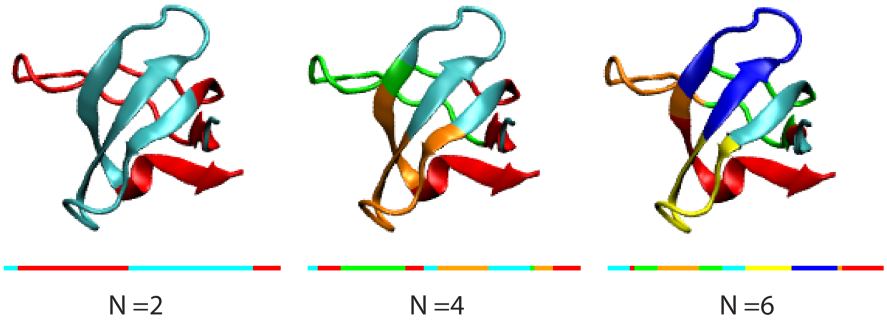

Fig. S11. CG representations with maximal $\mathcal{Q}$ for CG models of 2V1Q with $N = 2$, 4, and 6 CG sites. The representations are indicated by assigning the same color to each residue in the same CG site. The bar graphs indicate the linear sequence of the protein, while the cartoons indicates its equilibrium three-dimensional structure.

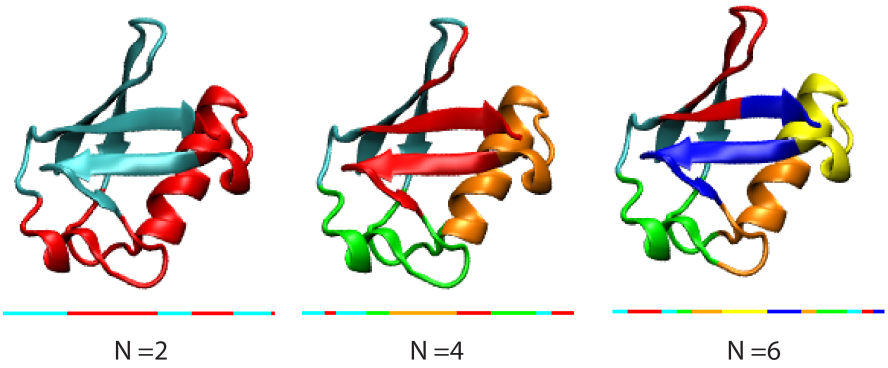

Fig. S12. CG representations with maximal $\mathcal{Q}$ for CG models of 1UBQ with $N = 2$, 4, and 6 CG sites. The representations are indicated by assigning the same color to each residue in the same CG site. The bar graphs indicate the linear sequence of the protein, while the cartoons indicates its equilibrium three-dimensional structure.

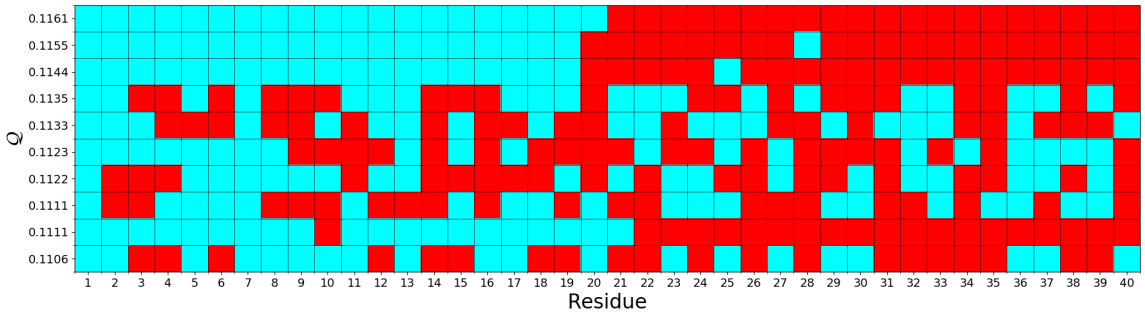

Fig. S13. Sequence alignment of the 10 maps with $N = 2$ sites that provide maximal spectral quality, $\mathcal{Q}$, for the model protein 2ERL. Each row corresponds to a single map, $\mathbf{M}$, with $\mathcal{Q}(\mathbf{M})$, indicated along the y-axis. The x-axis indicates the atomic sequence, while the colors indicate the site assignment of each atom.

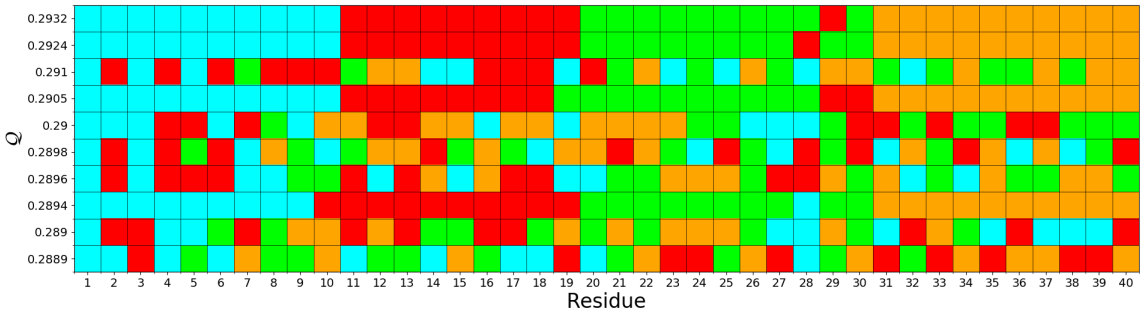

Fig. S14. Sequence alignment of the 10 maps with N = 4 sites that provide maximal spectral quality, $\mathcal{Q}$, for the model protein 2ERL. Each row corresponds to a single map, $\mathbf{M}$, with $\mathcal{Q}(\mathbf{M})$, indicated along the y-axis. The x-axis indicates the atomic sequence, while the colors indicate the site assignment of each atom.

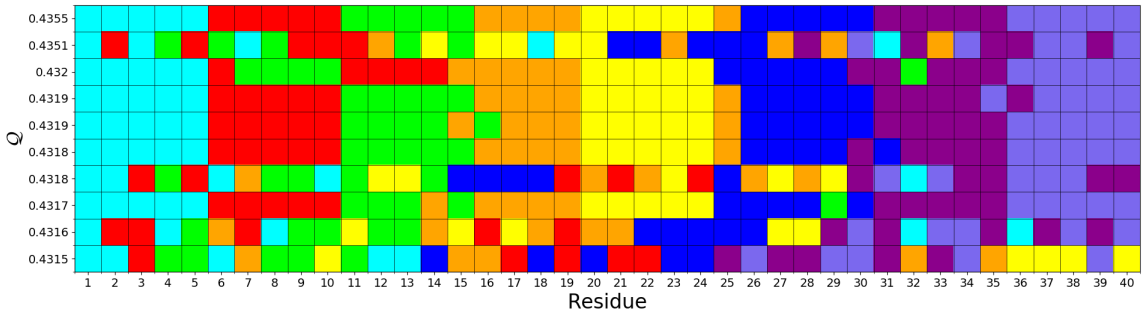

Fig. S15. Sequence alignment of the 10 maps with N = 8 sites that provide maximal spectral quality, $\mathcal{Q}$, for the model protein 2ERL. Each row corresponds to a single map, $\mathbf{M}$, with $\mathcal{Q}(\mathbf{M})$, indicated along the y-axis. The x-axis indicates the atomic sequence, while the colors indicate the site assignment of each atom.

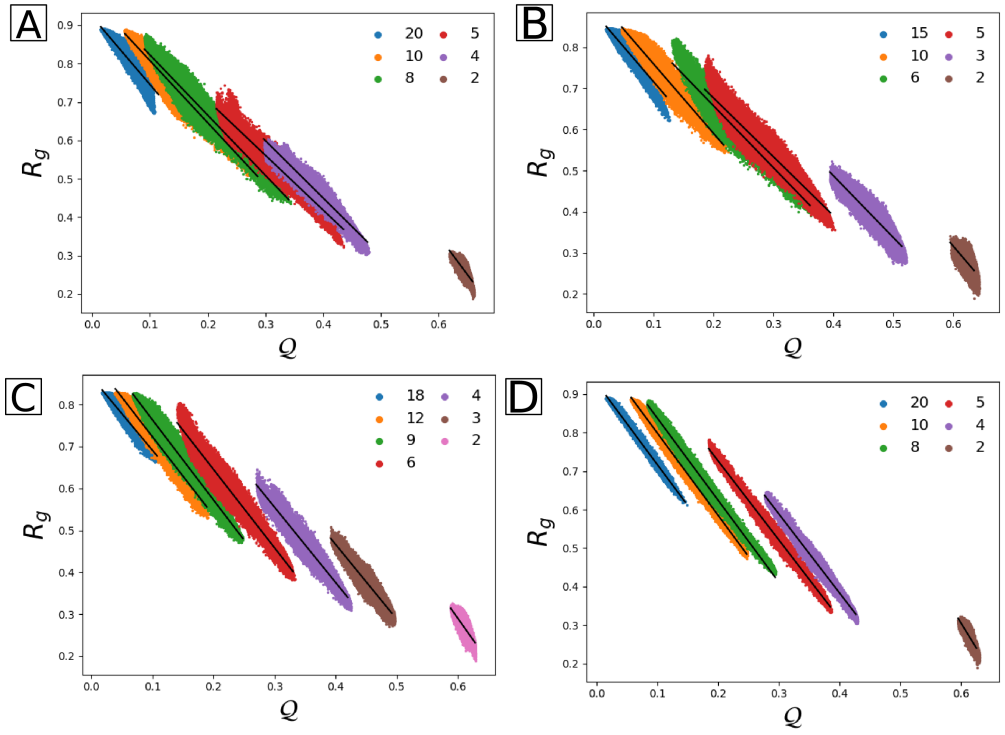

Fig. S16. Scatter plot of $(\mathcal{Q}, \mathrm{R_G})$ among sampled maps for (A) 2ERL, (B) 3HJD, (C)1IJU, and (D) 3E7R at the resolutions $R = n/N$, indicated by the colors in the legend.

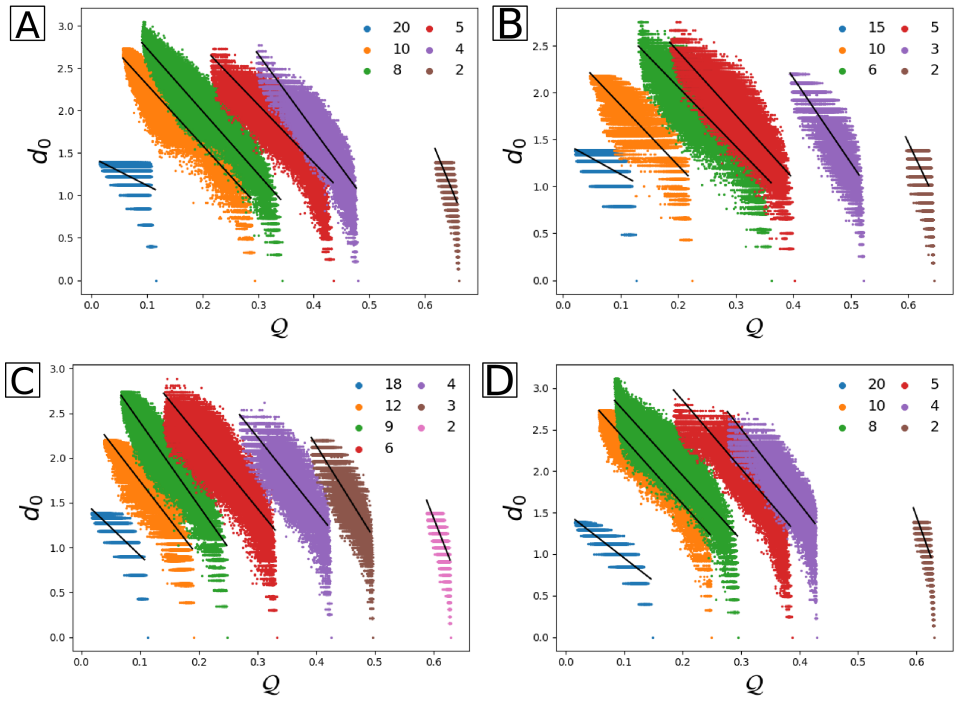

Fig. S17. Scatter plot of $(\mathcal{Q}, d_0)$ among sampled maps for (A) 2ERL, (B) 3HJD, (C)1IJU, and (D) 3E7R at the resolutions $R = n/N$, indicated by the colors in the legend.

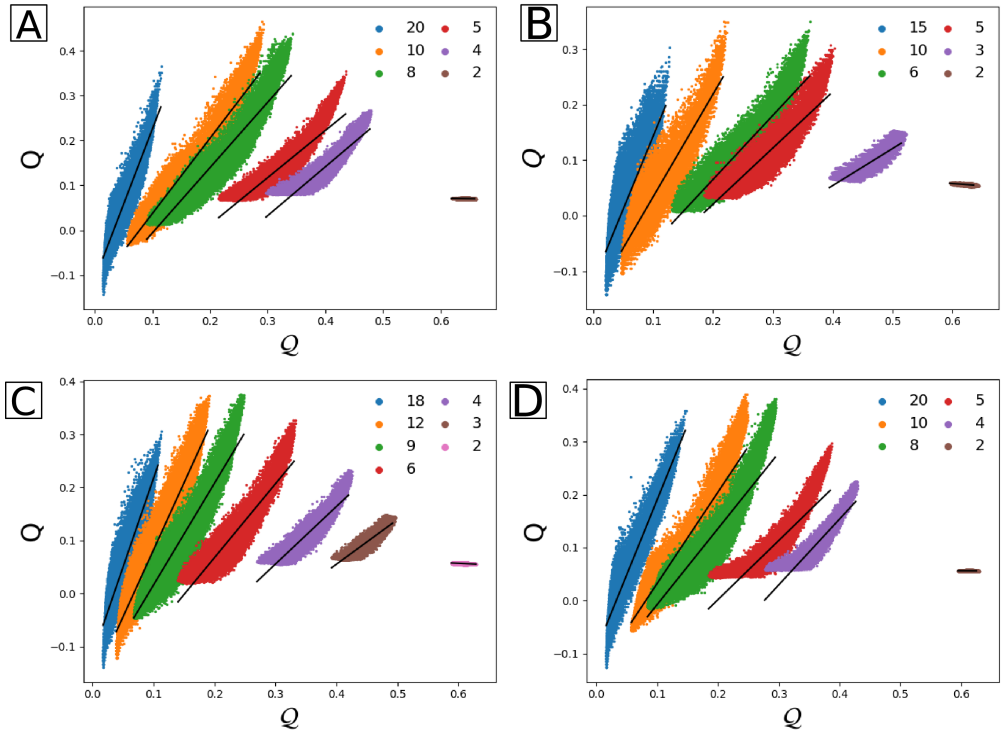

Fig. S18. Scatter plot of ($\mathcal{Q}$, Q) among sampled maps for (A) 2ERL, (B) 3HJD, (C)1IJU, and (D) 3E7R at the resolutions $R = n/N$, indicated by the colors in the legend.

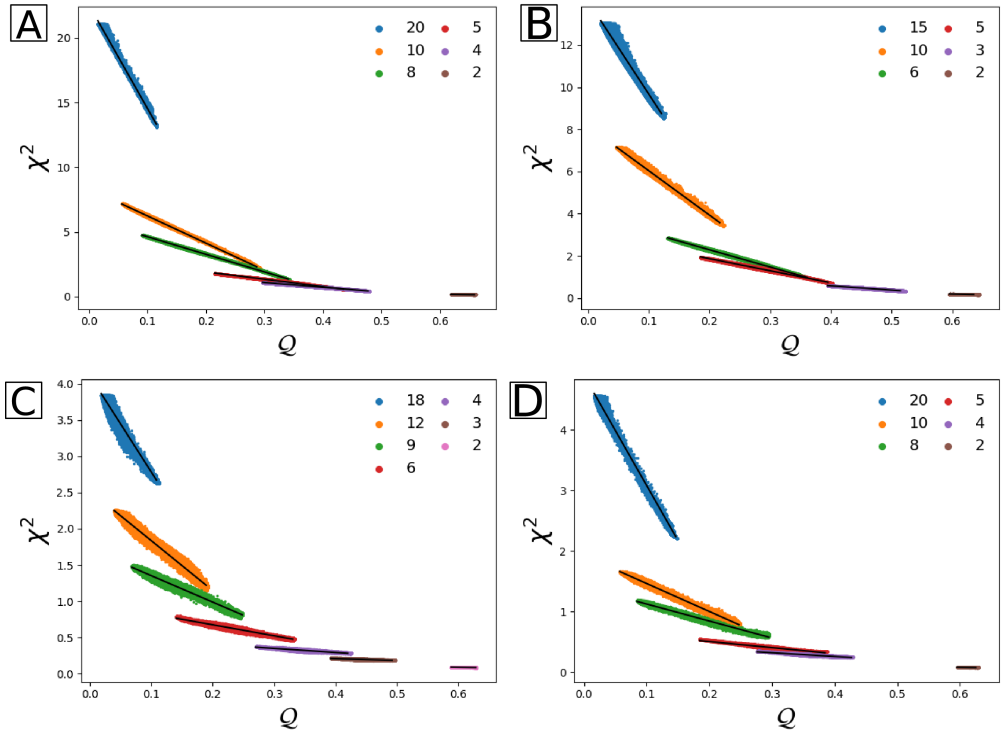

Fig. S19. Scatter plot of $(\mathcal{Q}, \chi^2)$ among sampled maps for (A) 2ERL, (B) 3HJD, (C)1IJU, and (D) 3E7R at the resolutions $R = n/N$, indicated by the colors in the legend.

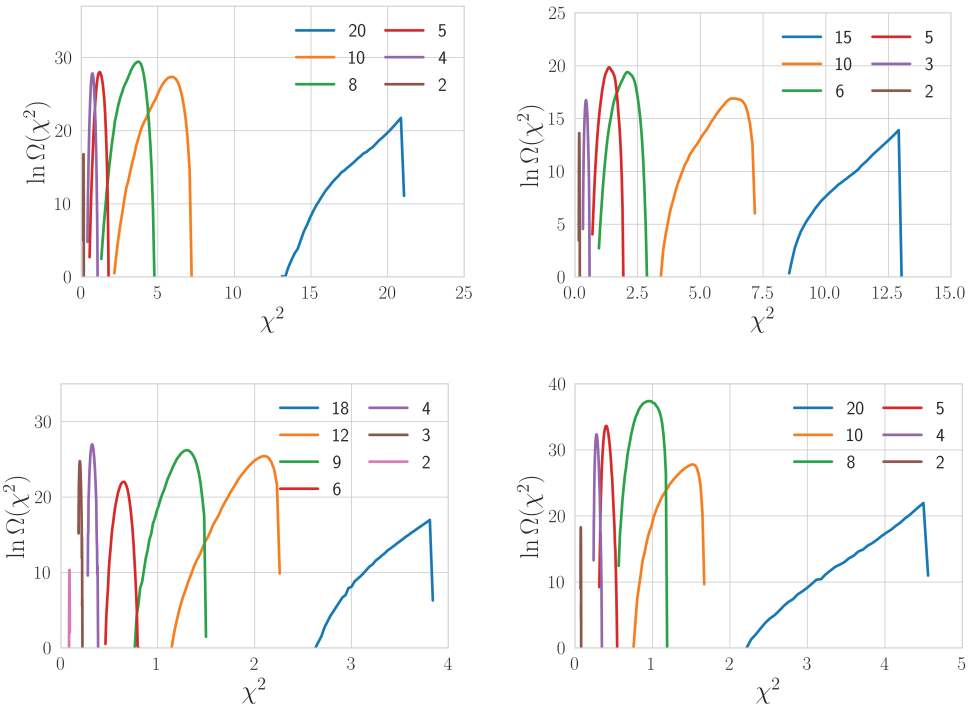

Fig. S20. Numerical estimate for the natural logarithm of the density of states, $\ln \Omega(\chi^2)$, quantifying the number of maps, $\mathbf{M}$, with given value of χ^2 at the resolutions, $R = n/N$, indicated by the colors of the legend.

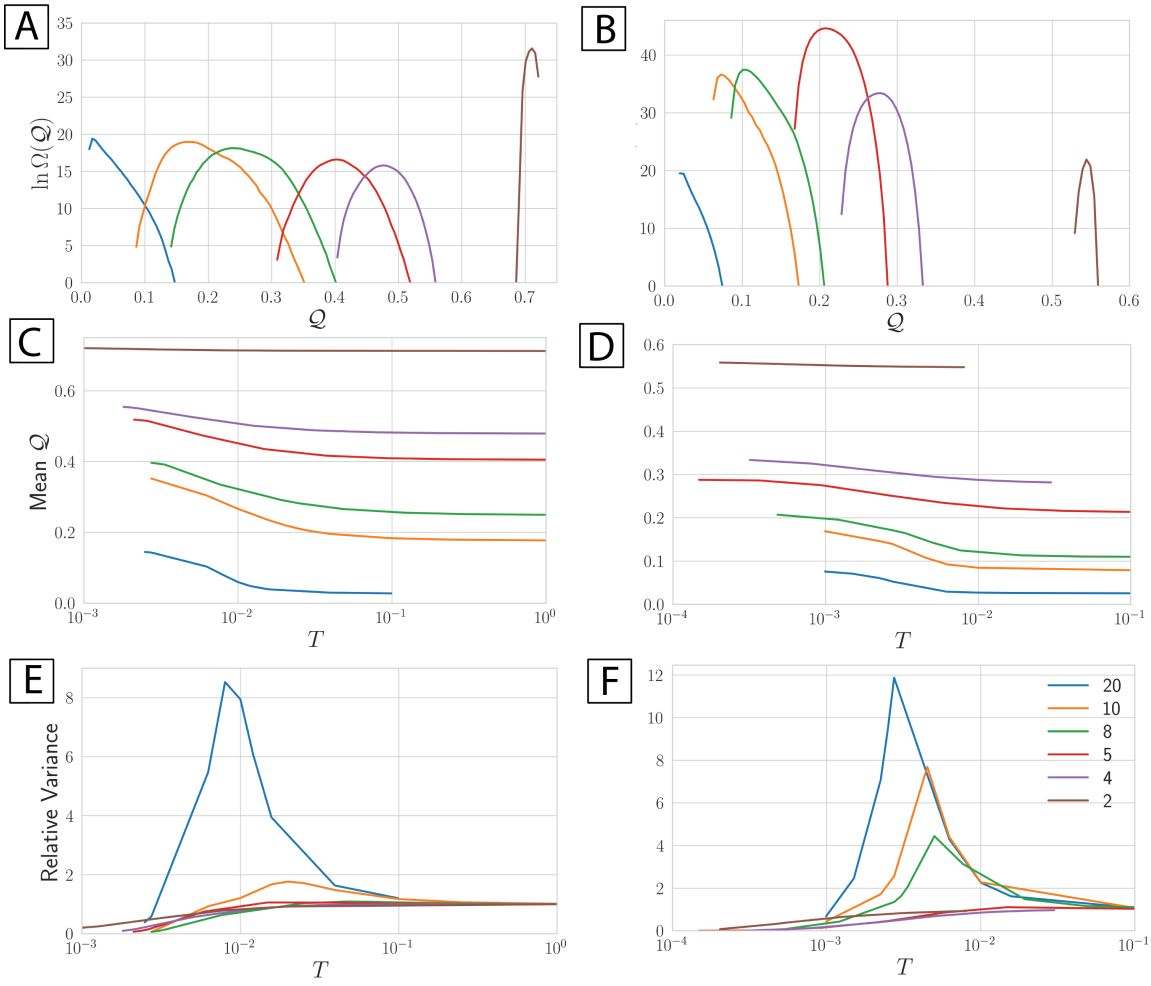

Fig. S21. Statistical analysis of mapping space for GNM's with the cutoff $R_c = 6.0$Å (left column) and $R_c = 10.0$ Å (right column). The top row presents the natural logarithm of the density of states, $\ln \Omega(\mathcal{Q})$, quantifying the number of maps, $\mathbf{M}$, with given spectral quality, $\mathcal{Q}$, at the resolutions, $R = n/N$, indicated by the colors of the legend. The middle and bottom rows represent the mean and relative variance of the spectral quality as a function of the temperature $T = T_{\mathcal{Q}}$ conjugate to $\mathcal{E} = 1 - \mathcal{Q}$. As in the main text, the variance is scaled with respect to the variance in the high temperature limit, $T_{\mathcal{Q}} \to \infty$.

Table S1. Model proteins. For each protein, we indicate the PDBID of the equilibrium structure, $\mathbf{r}^*$, the number of amino acids, n, that are treated in the high-resolution GNM, as well as any residues that are neglected by the GNM. The PDB structures for 3HJD and 2V1Q correspond to symmetric dimers, while the the PDB structure for 1IJU is a symmetric tetramer. In these three cases, the GNM is defined by the structure of chain A in the PDB file. In the case of 1UBQ, the last 4 residues correspond to a flexible tail that is trimmed from the GNM.

PDBID	number of residues (n)	residues trimmed
2ERL	40	0
3HJD	30	31-60
1IJU	36	37-144
3E7R	40	0
1UG4	60	0
2V1Q	60	61-120
1UBQ	72	73-76

Table S2. Table of lines of best fit correlations of four metrics with $\mathcal{Q}$

Protein	R	χ^2			Q			d_0			R_g		
		Slope	Intercept	r^2	Slope	Intercept	r^2	Slope	Intercept	r^2	Slope	Intercept	r^2
2ERL	20	-79.966	22.536	1.0	- 3.356	-0.11	0.93	-3.321	1.45	0.45	-1.766	0.922	0.93
	10	-21.031	8.335	1.0	1.677	-0.13	0.93	-7.164	3.022	0.84	-1.617	0.969	0.95
	8	-13.598	5.983	1.0	1.454	-0.151	0.9	-7.392	3.47	0.86	-1.567	0.979	0.93
	5	-5.659	3.036	1.0	1.049	-0.197	0.81	-6.796	4.109	0.72	-1.427	0.989	0.91
	4	-3.693	2.196	0.99	1.088	-0.293	0.83	-8.927	5.344	0.76	-1.492	1.047	0.9
	2	-0.589	0.534	0.73	-0.014	0.08	0.02	-15.938	11.415	0.33	-2.021	1.564	0.55
3HJD	15	-44.037	14.073	0.99	2.627	-0.119	0.82	-3.397	1.47	0.61	-1.699	0.886	0.95
	10	-21.034	8.14	1.0	1.849	-0.15	0.89	-6.445	2.512	0.79	-1.686	0.928	0.95
	6	-8.187	3.926	1.0	1.156	-0.166	0.86	-6.321	3.324	0.72	-1.503	0.958	0.91
	5	-5.724	3.011	1.0	1.013	-0.182	0.83	-6.784	3.795	0.73	-1.429	0.962	0.89
	3	-2.029	1.392	0.98	0.668	-0.213	0.77	-9.054	5.778	0.68	-1.503	1.089	0.86
	2	-0.591	0.544	0.77	-0.079	0.106	0.32	-12.97	9.248	0.28	-1.704	1.34	0.43
1IJU	18	-13.245	4.106	0.99	3.35	-0.118	0.9	-6.286	1.542	0.81	-1.744	0.865	0.97
	12	-6.875	2.523	0.99	2.528	-0.171	0.94	-8.48	2.593	0.88	-1.878	0.911	0.98
	9	-3.644	1.717	0.97	2.528	-0.171	0.94	-9.302	3.332	0.88	-1.91	0.954	0.97
	6	-1.521	0.98	0.97	1.393	-0.211	0.86	-8.039	3.852	0.72	-1.869	1.019	0.94
	4	-0.567	0.52	0.84	1.086	-0.27	0.85	-8.224	4.703	0.64	-1.794	1.093	0.91
	3	-0.259	0.313	0.72	0.841	-0.281	0.82	-10.523	6.355	0.68	-1.78	1.178	0.89
	2	-0.122	0.163	0.56	-0.053	0.089	0.22	-16.699	11.359	0.47	-2.054	1.523	0.58
3E7R	20	-18.18	4.902	1.0	2.826	-0.092	0.94	-5.499	1.504	0.91	-2.125	0.93	0.99
	10	-4.619	1.925	0.99	1.722	-0.14	0.9	-7.863	3.18	0.91	-2.138	1.013	0.99
	8	-2.809	1.405	0.98	1.431	-0.15	0.85	-7.774	3.504	0.81	-2.132	1.051	0.99
	5	-1.005	0.705	0.91	1.116	-0.221	0.74	-8.214	4.496	0.66	-2.055	1.138	0.98
	4	-0.622	0.505	0.83	1.241	-0.342	0.83	-8.998	5.209	0.64	-2.058	1.206	0.96
	2	-0.135	0.159	0.38	-0.005	0.059	0.0	-19.787	13.33	0.4	-2.573	1.848	0.58

References

1. PJ Flory, M Gordon, NG McCrum, Statistical thermodynamics of random networks [and discussion]. Proc. Roy. Soc. Lond. A: Math. Phys. Sci. **351**, 351–380 (1976).
2. T Haliloglu, I Bahar, B Erman, Gaussian dynamics of folded proteins. Phys. Rev. Lett. **79**, 3090–3093 (1997).
3. I Bahar, TR Lezon, A Bakan, IH Shrivastava, Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. Chem. Rev. **110**, 1463–1497 (2010).
4. A Bakan, LM Meireles, I Bahar, Prody: Protein dynamics inferred from theory and experiments. Bioinformatics **27**, 1575–1577 (2011).
5. BE Eichinger, Configuration statistics of gaussian molecules. Macromolecules **13**, 1–11 (1980).
6. JG Kirkwood, Statistical mechanics of fluid mixtures. J. Chem. Phys. **3**, 300–313 (1935).
7. CN Likos, Effective interactions in soft condensed matter physics. Phys. Rep. **348**, 267 – 439 (2001).
8. LP Kadanoff, Statistical Physics. (WORLD SCIENTIFIC), (2000).
9. JF Rudzinski, WG Noid, Coarse-graining entropy, forces, and structures. J. Chem. Phys. **135**, 214101 (2011).
10. TT Foley, MS Shell, WG Noid, The impact of resolution upon entropy and information in coarse-grained models. J. Chem. Phys. **143**, 243104 (2015).
11. NJH Dunn, TT Foley, WG Noid, Van der waals perspective on coarse-graining: progress toward solving representability and transferability problems. Acc. Chem. Res. **49**, 2832–2840 (2016).
12. JM Harris, JL Hirst, MJ Mossinghoff, Combinatorics and graph theory. (Springer), (2010).
13. S Fortunato, Community detection in graphs. Phys. Rep. **486**, 75–174 (2010).
14. D Frenkel, B Smit, Understanding Molecular Simulation: From Algorithms to Applications. (Academic Press, San Diego, CA USA), Second edition, (2002).
15. MR Shirts, JD Chodera, Statistically optimal analysis of samples from multiple equilibrium states. J. Chem. Phys. **129**, 124105 (2008).
16. M Meilă, Comparing clusterings—an information based distance. J. Multivar. Anal. **98**, 873–895 (2007).
17. TM Cover, JA Thomas, Elements of Information Theory. (Wiley Interscience), 2 edition, (2006).
18. MEJ Newman, M Girvan, Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004) Publisher: American Physical Society.
19. ZY Zhang, et al., A systematic methodology for defining coarse-grained sites in large biomolecules. Biophys. J. **95**, 5073–5083 (2008).
20. A Amadei, ABM Linssen, HJC Berendsen, Essential dynamics of proteins. Proteins **17**, 412 – 425 (1993).