An information theory–based approach for optimal model reduction of biomolecules

Marco Giulini,[†,¶] Roberto Menichetti,[†,¶] M. Scott Shell,[‡] and Raffaello Potestio[*,†,¶]

†*Physics Department, University of Trento, via Sommarive, 14 I-38123 Trento, Italy*

‡*Department of Chemical Engineering, University of California Santa Barbara, Santa Barbara, California 93106, USA*

¶*INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, I-38123 Trento, Italy*

E-mail: raffaello.potestio@unitn.it

Abstract

In the theoretical modelling of a physical system a crucial step consists in the identification of those degrees of freedom that enable a synthetic, yet informative representation of it. While in some cases this selection can be carried out on the basis of intuition and experience, a straightforward discrimination of the important features from the negligible ones is difficult for many complex systems, most notably heteropolymers and large biomolecules. We here present a thermodynamics-based theoretical framework to gauge the effectiveness of a given simplified representation by measuring its information content. We employ this method to identify those reduced descriptions of proteins, in terms of a subset of their atoms, that retain the largest amount of information from the original model; we show that these highly informative representations share common features that are intrinsically related to the biological

properties of the proteins under examination, thereby establishing a bridge between protein structure, energetics, and function.

1 Introduction

The quantitative investigation of a physical system relies on the formulation of a *model* of it, that is, an abstract representation of its constituents and the interactions among them in terms of mathematical constructs. In the realisation of the simplest model that entails all the relevant features of the system under investigation, one of the most crucial aspects is the determination of its level of detail. The latter can vary depending on the properties and processes of interest: the quantum mechanical nature of matter is explicitly incorporated in *ab initio* methods,[1] while effective classical interactions are commonly employed in the the all-atom force fields used in all-atom (AA) molecular dynamics (MD) simulations.[2,3] Representations of a molecular system whose resolution level is lower than the atomistic one are commonly dubbed *coarse-grained* (CG) models:[4–8] in this case, the fundamental degrees of freedom, or effective interaction centroids, are representatives of groups of atoms, and the interactions among these CG sites are parametrised so as to reproduce equilibrium properties of the reference system.

An important distinction should be made between *reproducing* a given property, and *describing* it. For example, it is evident that the explicit incorporation of the electronic degrees of freedom in the model of a molecule is necessary to reproduce, with qualitative and quantitative accuracy, its vibrational spectrum; on the other hand, the latter can be measured and described from the knowledge of the nuclear coordinates alone, i.e. from the inspection of a *subset* of the system's degrees of freedom. This is a general feature, in that the *understanding* of a complex system's properties and behaviour can typically be achieved in terms of a reduced set of variables: statistical mechanics provides some of the most recognisable examples of this, such as the description of systems composed of an Avogadro

number of atoms or molecules in terms of a handful of thermodynamical parameters.

In computer-aided studies, and particularly in the fields of computational biophysics and biochemistry, recent technological advancements–most notably massive parallelisation,[9] GPU computing,[10] and tailor-made machines such as ANTON[11]–have extended the range of applicability of atomistic simulations to molecular complexes composed of millions of atoms;[12–14] even in absence of such impressive resources, it is now common practice to perform microseconds-long simulations of relatively large systems, up to hundred thousands atoms. However, a process of filtering, dimensionality reduction, or feature selection is anyhow required in order to distill the physically and biologically relevant information from the immense amount of data it is buried in.

The problem is thus to identify the most synthetic picture of the system that entails all and only its important properties: an optimal balance is sought between parsimony and informativeness. This objective can be pursued making use of the language and techniques of bottom-up coarse-grained modelling;[5,15] in this context, in fact, one defines a *mapping operator* $\mathbf{M}$ that performs a transformation from a high-resolution configuration $\mathbf{r}_i$, $i = 1, ..., n$ of the system described in large detail to a simpler, *coarser* configuration $\mathbf{R}_I$, $I = 1, ..., N < n$ at lower resolution:

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

where n and N are the number of atoms in the system and the number of CG sites chosen, respectively. The linear coefficients c_{Ii} in Eq. 1 are constant, positive and subject to the normalisation condition $\sum_i c_{Ii} = 1$ to preserve translational invariance. Furthermore, coefficients are generally taken to be *specific* to each site,[15] that is, an atom i taking part to the definition of CG site I will not be involved in the construction of another site J ($c_{Ji} = 0 \; \forall \; J \neq I$).

Once the *mapping* $\mathbf{M}$ is chosen, the interactions among CG sites must be determined. In this respect, several methodologies have been devised in the past decades to parametrise such

CG potentials.[4–8] Some approaches aim at reproducing as accurately as possible the *exact* effective potential obtained through the integration of the microscopic degrees of freedom of the system, that is, the multi-body potential of mean force (MB-PMF); this is achieved in practice by tuning the CG interactions so as to reproduce specific, low-resolution structural properties of the reference systems.[16,17] Recently, other methods have been proposed that target not only the structure, but also the energetics.[18,19]

In this work we do not tackle the issue of parametrising approximate CG potentials, but rather focus on the consequences of the simplification of the system's description even if the underlying physics is the same, i.e. configurations are sampled with the reference, all-atom probability. In other words, we focus purely on the effect of projecting the all-atom conformational ensemble onto a coarse-grained configurational space using the mapping as a filter.

Inevitably, in fact, a CG representation loses information about the high-resolution reference,[5,20] and the amount of information lost depends only on the number and selection of the retained degrees of freedom. In coarse-grained modelling, the mapping is commonly chosen based on general and intuitive criteria: for example, it is rather natural to represent a protein in terms of one single centroid per amino acid (usually the choice falls on the α carbon of the backbone).[21] However, it is by no means assured that a given representation that is natural and intuitive to the human eye is also the one that allows the CG model to retain the largest amount of information about the original, higher-resolution system.[22,23] A quantitative criterion to assess how much detail is lost upon structural coarsening is thus needed in order to perform a sensible choice.

In the past few years, various methods have been developed that target the problem of the automated construction of a simplified protein's representation at a resolution level lower than atomistic. In a pioneering work Gohlke and Thorpe proposed to partition a protein in few, size-wise diverse blocks, distributing the amino acids among the different domains so as to minimise the degree of internal flexibility of the latter.[24] This picture of

a protein subdivided in *quasi-rigid domains*, which has been further developed by several other authors,[25–31] is founded on the notion of a simplified model where groups of atoms are not assigned to coarse-grained sites according to their chemistry (e.g., one residue - one site), but rather based on the local properties of the specific molecule under examination. These partitioning methods, however, only employ structural information, in that they aim at minimising each block's internal strain, while the energetics of the system is neglected.

Some approaches systematically reduce the number of atoms in a system's representation by grouping them according to graph-theoretical procedures. For example, De Pablo and coworkers map the static structure on a graph and hierarchically decimate it by clustering together the "leaves";[32] alternative methods lump residues in effective sites based on a spectral analysis of the graph Laplacian.[33] More recently, Li et al.[34] developed a graph neural network-based method to match a dataset of manually annotated CG mappings.

Alternatively, it was proposed to retain only those atoms that guarantee the set of new interactions to quantitatively reproduce the MB-PMF.[22,35] These methods, though, are based on linearised elastic network models[36–41] that have the remarkable advantage of being exactly solvable, thus allowing a direct comparison between CG potential and MB-PMF, but cannot be taken as significant representations of the system's highly nonlinear interactions.

It follows that all these pioneering approaches rely either on purely geometrical/topological information obtained from a single, static structure; or on an ensemble of structures, neglecting energetics and thermodynamics; or on extremely simplified representations of both structure and interactions, that do not guarantee general applicability to systems of great complexity.

Here we tackle the issue of the automated, unsupervised construction of the most informative simplified representation of biological macromolecules in purely statistical mechanical terms, that is, in the language that is most naturally employed to investigate such systems. Specifically, we search for the mapping operator that, for a given number of atoms retained from the original all-atom model, provides a description whose information content is as close

as possible to the reference. In this context, then, the term "coarse-grained representation" should not be interpreted as a system with effective interactions whose scope is to reproduce a certain property, phenomenon, or behaviour; rather, the representations we discuss here are simpler *pictures* of the reference system evolving according to the reference microscopic Hamiltonian, but *looked at* in terms of fewer degrees of freedom. Our objective is thus the identification of the most informative simplified picture among those possible.

To this end, we make use of the concept of *mapping entropy*, S_{map},[17,42–44] a quantity that measures the quality of a CG representation in terms of the "distance" between probability distributions—the Boltzmann distribution of the reference, all-atom system, and the equivalent distribution when the AA probabilities are projected into the CG coordinate space. The mapping entropy is ignorant of the parametrisation of the effective interactions of the simplified model: S_{map} effectively compares the reference system, described through all its degrees of freedom, to the same system in which configurations are viewed through "coarse-graining lenses". The difference between these two representations only lies in the resolution, not in the microscopic physics.

Recently, the introduction of a mapping entropy-related metric proved to be a powerful instrument for determining the optimal CG'ing resolution *level* for a biological system.[44] Applied to a set of model proteins, this method was capable of identifying the number of sites that need to be employed in the simplified CG picture to preserve the maximum amount of thermodynamic information about the microscopic reference. However, such analysis was carried out at a fixed CG resolution *distribution*, with a homogeneous placement of sites along the protein sequence. Moreover, calculations were performed relying on an exactly solvable, yet very crude approximation to the system's microscopic interactions, namely a Gaussian Network Model.

In the following we take the moves from these results to develop a computationally effective protocol that enables the approximate calculation of the mapping entropy for an arbitrarily complex system. We employ this novel scheme to explore the space of the sys-

tem's possible CG representations, varying the resolution level *as well as* distribution, with the objective of identifying the ones featuring the lowest mapping entropy—that is, allowing for the smallest amount of information loss upon resolution reduction. The method is applied to three proteins of substantially different size, conformational variability, and biological activity. We show that the choice of retained degrees of freedom, guided by the objective of preserving the largest amount of information while reducing the complexity of the system, highlights biologically meaningful and *a priori* unknown structural features of the proteins under examination, whose identification would otherwise require computationally more intensive calculations or even wet lab experiments.

2 Results

In this section we report the main findings of our work. Specifically, (*i*) we outline the theoretical and computational framework that constitutes the basis for the calculation of the mapping entropy; (*ii*) we illustrate the biological systems on which we apply the method; and (*iii*) we describe the results of the mapping entropy minimisation for these systems and the properties of the associated mappings.

2.1 Theory

The concept of mapping entropy as a measure of the loss of information inherently generated by performing a CG'ing procedure on a system was first introduced by one of us in the framework of the relative entropy method,[17] and subsequently expanded in Refs.[42–44] For the sake of brevity, we here omit the formal derivation connecting relative entropy and mapping entropy as well as a discussion of the former. A brief summary of the relevant theoretical results presented in Refs.[17,42–44] is provided in Appendix A.

In the following we restrict our analysis to the case of decimation mappings $\mathbf{M}$, in which a subset of $N < n$ atoms of the original system is retained while the remaining ones are

integrated out, so that

$$\begin{aligned} &\mathbf{M}_I(\mathbf{r}) = \sigma_i \mathbf{r}_i,\ \sigma_i = 1 \text{ for one } I, \text{ 0 otherwise,} \\ &\sum_{i=1}^{n} \sigma_i = N. \end{aligned} \tag{2}$$

In this case, the mapping entropy S_{map} reads (see Appendix A)[42]

$$\begin{aligned} S_{map} &= k_B \times D_{KL}(p_r(\mathbf{r})||\bar{p}_r(\mathbf{r})) \\ &= k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})}\right], \end{aligned} \tag{3}$$

that is, a Kullback-Leibler (KL) divergence D_{KL}[45] between the probability distribution $p_r(\mathbf{r})$ of the high-resolution system and the distribution obtained by observing the latter through "coarse-graining glasses", $\bar{p}_r(\mathbf{r})$. Following the notation of Ref.,[42] $\bar{p}_r(\mathbf{r})$ is defined as

$$\bar{p}_r(\mathbf{r}) = p_R(\mathbf{M}(\mathbf{r}))/\Omega_1(\mathbf{M}(\mathbf{r})), \tag{4}$$

where $p_R(\mathbf{R})$ is the probability of the CG macrostate $\mathbf{R}$, given by

$$\begin{aligned} p_R(\mathbf{R}) &= \int d\mathbf{r}\ p_r(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \\ &= \frac{1}{Z} \int d\mathbf{r}\ e^{-\beta u(\mathbf{r})} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \\ Z &= \int d\mathbf{r}\ e^{-\beta u(\mathbf{r})}, \end{aligned} \tag{5}$$

while $\Omega_1(\mathbf{R})$ is defined as

$$\Omega_1(\mathbf{R}) = \int d\mathbf{r}\ \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \tag{6}$$

which is the degeneracy of the macrostate—how many microstates map onto the CG configuration $\mathbf{R}$. In Eq. 5 $\beta = 1/k_B T$, $u(\mathbf{r})$ is the microscopic potential energy of the system, and Z its canonical partition function.

The calculation of S_{map} in Eq. 3 thus amounts at determining the distance (in the KL sense) between two, although both microscopic, conceptually very different distributions. In contrast to $p_r(\mathbf{r})$, Eq. 4 displays that $\bar{p}_r(\mathbf{r})$ associates, to all configurations that map onto the same CG macrostate $\mathbf{R}$, the same probability; the latter is given by the average of the original probabilities of these microstates. Importantly, $\bar{p}_r(\mathbf{r})$ represents the high-resolution description of the system that would be accessible *only starting* from its low-resolution one—i.e., $p_R(\mathbf{R})$. Grouping together configurations into a CG macrostate has the effect of flattening the detail of their original probabilistic weights. An attempt to revert the CG'ing procedure and restore an atomistic resolution by reintroducing the mapping operator $\mathbf{M}$ in $p_R(\mathbf{R})$ can only result in microscopic configurations that are uniformly distributed within each macrostate.

Due to the smearing in probabilities, the CG'ing transformations constitute a semi-group.[46] This irreversible character highlights a fundamental consequence of CG'ing strategies: a loss of information about the system. The definition, based on the KL divergence, presented in Eq. 3 is useful for practical purposes. A more direct understanding of this information loss and how it is encoded in the mapping entropy, however, can be obtained by considering the non-ideal configurational entropies of the original and CG representation,

$$s_r = -k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln(V^n p_r(\mathbf{r})) \tag{7}$$

$$s_R = -k_B \int d\mathbf{R}\; p_R(\mathbf{R}) \ln(V^N p_R(\mathbf{R})) \tag{8}$$

respectively quantifying the information contained in the associated probability distributions, $p_r(\mathbf{r})$ and $p_R(\mathbf{R})$:[47] the higher the entropy, the more uniform the distribution, which we associate to a smaller amount of information content. By virtue of Gibbs' inequality, from Eq. 3 one has $S_{map} \geq 0$. Furthermore, see Appendix A

$$s_R - s_r = S_{map} \geq 0, \tag{9}$$

so that the entropy of the CG representation is always larger than the reference, microscopic one, implying that a loss of information occurs in decreasing the level of resolution.[42,44] Critically, the difference between the two information contents is precisely the mapping entropy.

The information that is lost in the CG'ing process through S_{map} only depends on the mapping operator $\mathbf{M}$—in our case, on the choice of the retained sites. This paves the way for the possibility of assessing the quality of a CG mapping based on the amount of information it is able to *retain* about the original system, a qualitative advancement with respect to the more common a priori selection of CG representations.[21] Unfortunately, Eq. 3 or 9 do not allow—except in very simple microscopic models, see Ref. 44—a straightforward computational estimate of S_{map} for a system arising from a choice of its CG mapping, as the observables to be averaged involve logarithms of high-dimensional probability distributions, and ultimately configuration-dependent free energies. However, having introduced the loss of information per macrostate $S_{map}(\mathbf{R})$ defined by the relation[42,44]

$$S_{map} = \int d\mathbf{R}\; p_R(\mathbf{R}) S_{map}(\mathbf{R}), \tag{10}$$

in Appendix B we show that this problem can be overcome by further subdividing microscopic configurations that map to a given macrostate according to their potential energy. Let us define the conditional probability $P_\beta(U|\mathbf{R})$ for the system, thermalised at inverse temperature β, to have energy U provided that is in macrostate $\mathbf{R}$ as

$$\begin{aligned} P_\beta(U|\mathbf{R}) &= \frac{p_R(U,\mathbf{R})}{p_R(\mathbf{R})} \\ &= \frac{1}{p_R(\mathbf{R})} \int d\mathbf{r} p_r(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \delta(u(\mathbf{r}) - U), \end{aligned} \tag{11}$$

so that $S_{map}(\mathbf{R})$ can be *exactly* rewritten as (see Appendix B):

$$S_{map}(\mathbf{R}) = k_B \ln \left[\int dU' P_\beta(U'|\mathbf{R}) \; e^{\beta(U' - \langle U \rangle_{\beta|\mathbf{R}})} \right], \tag{12}$$

where $\langle U \rangle_{\beta|\mathbf{R}}$ is the average of the potential energy restricted to the CG macrostate $\mathbf{R}$,

$$\langle U \rangle_{\beta|\mathbf{R}} = \int dU P_\beta(U|\mathbf{R}) U. \tag{13}$$

This derivation enables a direct estimate of the mapping entropy S_{map} from configurations sampled according to the microscopic probability distribution $p_r(\mathbf{r})$. For a given mapping, the histogram of these configurations with respect to CG coordinates $\mathbf{R}$ and energy U approximates the conditional probability $P_\beta(U|\mathbf{R})$ and, consequently, $S_{map}(\mathbf{R})$, see Eq. 12; the total mapping entropy can thus be obtained as a weighted sum of the latter over all CG macrostates, Eq. 10.

The only remaining difficulty consists in obtaining accurate estimates of the exponential average in Eq. 12, which are prone to numerical errors. As often in these cases—see e.g. free-energy calculations through Jarzynski's equality or the free-energy perturbation method[48,49]—it is possible to rely on a cumulant expansion of Eq. 12, which truncated at second order provides

$$S_{map}(\mathbf{R}) \simeq k_B \frac{\beta^2}{2} \langle (U - \langle U \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}. \tag{14}$$

Inserting Eq. 14 in Eq. 10 results in a *total* mapping entropy given by:

$$S_{map} \simeq k_B \frac{\beta^2}{2} \int d\mathbf{R} p_R(\mathbf{R}) \langle (U - \langle U \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}. \tag{15}$$

For a CG representations to exhibit an exactly zero mapping entropy, it is required that all microstates $\mathbf{r}$ that map onto a given macrostate $\mathbf{R} = \mathbf{M}(\mathbf{r})$ have the same energy in the reference system. Indeed, in this case one has $P_\beta(U|\mathbf{R}) = \delta(U - \bar{u}_{\mathbf{R}})$ in Eq. 12, with $\bar{u}_{\mathbf{R}}$ being the potential energy common to all microstates within macrostate $\mathbf{R}$, and

consequently $S_{map}(\mathbf{R}) = 0$. Eq. 14 highlights that deviations from this condition result in a loss of information associated to a particular CG macrostate that is proportional to the variance of the potential energy of all the atomistic configurations that map to $\mathbf{R}$. The overall mapping entropy is an average of these energy variances over all macrostates, each one weighted with the corresponding probability.

In the numerical implementation we thus seek to identify those mappings that cluster together atomistic configurations having the same, or at least very close energy, so as to minimise the information loss arising from CG'ing. With respect to Eq. 15, we further approximate S_{map} to its discretised counterpart (see Methods),

$$\tilde{S}_{map} = k_B \frac{\beta^2}{2} \sum_{i=1}^{N_{cl}} p_R(\mathbf{R}_i) \langle (U - \langle U \rangle_{\beta|\mathbf{R}_i})^2 \rangle_{\beta|\mathbf{R}_i} \tag{16}$$

where we identify N_{cl} discrete CG macrostates $\mathbf{R}_i$, each of which contributes to $\tilde{S}_{map}$ with its own probability $p_R(\mathbf{R}_i)$ taken as the relative population of the cluster. We then employ an algorithmic procedure to estimate and efficiently minimise, over the possible mappings, a cost function (Eq. 25 of the Methods section)

$$\Sigma \equiv \langle \tilde{S}_{map} \rangle \tag{17}$$

defined as an average of values of $\tilde{S}_{map}$ computed over different CG configuration sets, each of these being associated to a given number of conformational clusters N_{cl}.

Finally, it is interesting to note that the mapping entropy in the form presented in Eq. 15 appears in the dual-potential approach recently developed by Lebold and Noid.[18,19] In these works, the authors obtain an approximate CG energy function $E(\mathbf{R})$ able to accurately reproduce the *exact* energetics of the low resolution system—i.e. the average energy $\langle U \rangle_{\beta|\mathbf{R}}$ in macrostate $\mathbf{R}$, see Eq. 13. This is achieved by minimising the functional

$$\chi^2[E] = \langle |E(\mathbf{M}(\mathbf{r})) - u(\mathbf{r})|^2 \rangle \tag{18}$$

with respect to the force-field parameters contained in E, the average in Eq. 18 being performed over the microscopic model. By expressing $\langle U \rangle_{\beta|\mathbf{R}}$ as a function of $\mathbf{r}$ through the mapping $\mathbf{M}$, $\chi^2[E]$ can be decomposed as[18,19]

$$\chi^2[E] = \langle |\langle U \rangle_{\beta|\mathbf{R}}(\mathbf{M}(\mathbf{r})) - u(\mathbf{r})|^2 \rangle + \langle |E(\mathbf{M}(\mathbf{r})) - \langle U \rangle_{\beta|\mathbf{R}}(\mathbf{M}(\mathbf{r}))|^2 \rangle. \tag{19}$$

Minimising $\chi^2[E]$ on $E(\mathbf{R})$ for a given, *fixed* mapping as in Refs. 18,19 is tantamount to minimising the second term of Eq. 19, with the objective of reducing the error introduced by approximating $\langle U \rangle_{\beta|\mathbf{R}}$ through $E(\mathbf{R})$.

However, a comparison of Eqs. 15 and 19 displays that S_{map} coincides with the first term of Eq. 19. Critically, the latter depends only on the mapping $\mathbf{M}$ and would be nonzero also in the case of an *exact* parametrisation of E, if $E(\mathbf{R}) \equiv \langle U \rangle_{\beta|\mathbf{R}}$. The approach illustrated in the present work goes in a direction complementary to that of Refs. 18,19, as we concentrate on identifying those mappings that minimise the one contribution to $\chi^2[E]$ that is due to, and depends only on, the CG representation $\mathbf{M}$.

2.2 Biological structures

It is worth stressing that the results of the previous section are completely general and independent of the specific features of the underlying system. Of course, characteristics of the input such as the force field quality, the simulation duration, the number of conformational basins explored etc. will impact the outcome of the analysis, as it is necessarily the case in any computer-aided investigation; nonetheless, the applicability of the method is not prevented or limited by these features or other system properties, e.g. the specific molecule under examination, its complexity, its size, or its underlying all-atom modelling.

To illustrate the method in its generality, we here focus our attention on three proteins we chose to constitute a small yet representative set of case studies. These molecules cover

a size range spanning from $\sim$ 30 to $\sim$ 400 residues and a similarly broad spectrum of conformational variability and biological function, and can be taken as examples of several classes of enzymatic as well as non-enzymatic proteins.

Each protein is simulated for 200 ns in the NVT ensemble with physiological ion concentration. Out of 200 ns, snapshots every 20 ps are extracted from each trajectory, for a total 10^4 AA configurations per protein employed throughout the analysis. Details about the simulation parameters, a quantitative inspection of MD trajectories, characteristic features of each protein's results, as well as the validation of the latter with respect to the duration of the MD trajectory employed, can be found in the Supplemental Material. Hereafter we provide a description of each molecule, along with a brief summary of its behaviour as observed along MD simulations.

[**TAM**] A recently released[50] 31-residue *tamapin* mutant (PDB code 6D93). Tamapin is the toxin produced by the Indian red scorpion. It features a remarkable selectivity towards a peculiar calcium-activated potassium channel (SK2), whose potential use in the pharmaceutical context has made it a preferred object of study during the past decade.[51,52] Throughout our simulation almost every residue is highly solvent-exposed. Side chains fluctuate substantially, thus giving rise to an extreme structural variability.

[**AKE**] *Adenylate Kinase* (PDB code 4AKE). It is a 214 residue-long phosphotransferase enzyme that catalyses the interconversion between adenine diphosphate (ADP) and monophosphate (AMP) and their energetically rich complex, Adenine triphosphate (ATP).[53] It can be subdivided in three structural domains, CORE, LID, and NMP.[54] The CORE domain is stable, while the other two undergo large conformational changes.[55] Its central biochemical role in the regulation of the energetic balance of the cell and relatively small size, combined with the possibility to observe conformational transitions over timescales easily accessible by plain MD,[56] make it the ideal candidate to test and validate novel computational methods.[22,57,58] In our MD simulation the protein displays many rearrangements in the two motile domains, which occur to be quite close at many points. Nevertheless, the protein does not

undergo a full *open* $\leftrightarrow$ *closed* conformational transition.

[AAT] $\alpha-1$ *antitrypsin* (PDB code 1QLP). With 5934 atoms (372 residues), this protein is almost two times bigger than adenylate kinase. $\alpha-1$ antytripsin is a globular biomolecule and it is well known to exhibit a conformational rearrangement over the timescales of the minutes.[59–61] During our simulated trajectory the molecule experiences fluctuations particularly localised in correspondence of the most solvent-exposed residues. The protein bulk appears to be very rigid, and there is no sign of a conformational rearrangement.

2.3 Minimisation of the mapping entropy and characterisation of the solution space

The algorithmic procedure described in the Methods section and Appendix B enables one to quantify the information loss experienced by a system as a consequence of a *specific* decimation of its degrees of freedom. This quantification, which is achieved through the approximate calculation of the associated mapping entropy, opens the possibility of minimising such measure in the space of CG representations, so as to identify the mapping that, for a given number of CG sites N, is able to preserve as much information as possible about the AA reference.

In the following we allow CG sites to be located only on heavy atoms, thus reducing the maximum number of possible sites to N_{heavy}. We then investigate the properties of various kinds of CG mappings having different numbers of retained sites N. Specifically, we consider three chemically-intuitive values of N for each biomolecule: (i) N_α, i.e., the number of C_α atoms of the structure (equal to the number of amino acids); (ii) $N_{\alpha\beta}$, the number of C_α and C_β atoms; and (iii) N_{bkb}, which results from counting all the heavy atoms belonging to the main chain of the protein. The values of N for mappings (i)-(iii) in the case of TAM, AKE and AAT are listed in Tab. 1, together with the corresponding N_{heavy}.

Even restricting N to $N_\alpha, N_{\alpha\beta}$ and N_{bkb}, the combinatorial dependence of the number of possible decimation mappings on the amount of retained sites and N_{heavy} makes their

exhaustive exploration unfeasible in practice (see Methods). To identify the CG representations that minimise the information loss we thus rely on a Monte Carlo simulated annealing approach (SA, see Methods).[62,63] For each analysed protein and value of N, we perform 48 independent optimisation runs, i.e., minimisations of the mapping entropy with respect to the CG site selection; we then store the CG representation characterised by the lowest value of Σ in each run, thus resulting in a pool of *optimised* solutions. In order to assess their statistical significance and properties, we also generate a set of random mappings and calculate the associated Σ's, which constitute our reference values.

Fig. 1 displays, for each value of N considered, the distribution of mapping entropies obtained from a random choice of the CG representation of TAM, AKE, and AAT together with each protein's optimised counterpart. For $N = N_{bkb}$ and $N = N_\alpha$ in Fig. 1 we also report the values of Σ associated to physically-intuitive choices of the CG mapping that are commonly employed in the literature: the backbone mapping ($N = N_{bkb}$), which neglects all atoms belonging to the side chains; and the C_α mapping ($N = N_\alpha$), in which we only retain the C_α atoms of the structures. The first is representative of united-atom CG models, while the second is a ubiquitous and rather intuitive choice to represent a protein in terms of a single bead per amino acid.[21]

Table 1: Values of N_α, $N_{\alpha\beta}$, N_{bkb} and N_{heavy} (see text) for each analysed protein.

Protein	N_α	$N_{\alpha\beta}$	N_{bkb}	N_{heavy}
Tamapin (TAM)	31	59	124	230
Adenylate Kinase (AKE)	214	408	856	1656
$\alpha - 1$ antytripsin (AAT)	372	723	1488	2956

The optimality of a given mapping with respect to a random choice of the CG sites can be quantified in terms of the Z-score

$$Z = \frac{\Sigma_{opt} - \mu}{\sigma}, \tag{20}$$

where μ and σ represent mean and standard deviation of the distribution of Σ over randomly sampled mappings, respectively. Table 2 summarises the values of Z found for each N for the proteins under examination, including $Z[backbone]$ and $Z[\mathrm{C}_\alpha]$, which are computed with respect to the random distribution generated with $N = N_{bkb}$ and $N = N_\alpha$ respectively.

Table 2: Table of Z scores of each analysed protein. We report the mean and standard deviation of the distribution of Z values of the optimised solutions, $\overline{Z}$, for all values of N investigated. Results for the standard mappings—$Z[backbone]$ for backbone atoms only and $Z[\mathrm{C}_\alpha]$ for C_α atoms only—are also included.

N	TAM	AKE	AAT
$\overline{Z}[N_\alpha]$	-2.22 ± 0.06	-7.85 ± 1.14	-6.96 ± 1.03
$\overline{Z}[N_{\alpha\beta}]$	-2.38 ± 0.08	-6.09 ± 0.79	-6.64 ± 0.84
$\overline{Z}[N_{bkb}]$	-2.65 ± 0.09	-5.55 ± 0.62	-7.24 ± 0.85
$Z[backbone]$	4.37	5.65	4.31
$Z[\mathrm{C}_\alpha]$	0.87	3.36	3.28

As for the physically intuitive CG representations, Fig. 1 shows that the value of Σ associated to the backbone mapping is very high for all structures. For TAM in particular, the amount of information retained is so low that the mapping entropy falls 4.37 standard deviations higher than the reference distribution of random mappings, see Table 2. This suggests that neglecting the side chains in a CG representation of a protein is detrimental, at least as far as the structural resolution is concerned. In fact, the backbone of the protein undergoes relatively minor structural rearrangements when exploring the neighbourhood of the native conformation, thereby inducing negligible energetic fluctuations; for side chains, on the other hand, the opposite is true, with comparatively larger structural variability and a similarly broader energy range associated to it. Removing side chains from the mapping induces the clustering of atomistically different structures with different energies onto the same coarse-grained configuration, the latter being solely determined by the backbone. The corresponding mapping entropy is thus large—worse than a random choice of the retained atoms—since it is related to the variance of the energy in the macrostate.

Calculations employing the C_α mapping for the three structures show that this provides Σ values that are very close to the ones we find with the backbone mapping, thus suggesting that C_α atoms retain about the same amount of information that is encoded in the backbone. This is reasonable, given the rather limited conformational variability of the atoms along the peptide chain. However, a comparison of the random case distributions for a number N_α and N_{bkb} of retained atoms in Fig. 1 reveals that the former generally has a broader spread than the latter, due to the lower number of CG sites; consequently, the Σ of the C_α atoms mapping is closer to the bulk of the distribution of the random case than that of the backbone mapping.

We now discuss the case of optimised mappings, that is, CG representations retaining the maximum amount of information about the AA reference. Each of the 48 minimisation runs, which have been carried out for each protein in the set and value of N considered, provided an optimal solution—a deep local minimum in the space of CG mappings; the corresponding Σ's spread over a compact range of values that are systematically lower than, and do not overlap with, those of the random case distributions (Fig. 1).

Optimal solutions for AKE and AAT span a wide interval of values of Σ; when $N = N_\alpha$ in particular, the support of this set and of the corresponding random reference have comparable sizes. A quantitative measure of this broadness is displayed in the distributions of Z scores of optimal solutions presented in Table 2. In both proteins, we observe that the Σ's associated to optimal mappings increase with the degree of CG'ing N; this is a consequence of keeping the number of CG configurations of each system (conformational clusters, see Sec. 4.2) constant across different resolutions. As N increases, the available CG conformational clusters are populated by more energetically diverse conformations, thereby incrementing the associated energy fluctuations. On the other hand, TAM shows narrowly peaked distributions of optimal values of Σ, whose position does not vary with the amount of retained sites. Both effects can be ascribed to the fact that most of the energy fluctuations in TAM—and consequently the mapping entropy—are due to a subset of atoms that are almost always maintained in each

optimal mapping (see Sec. 2.4) in contrast to a random choice of the CG representation. At the same time, the associated Z scores are lower than the ones of the bigger proteins for all values of N under examination, as TAM conformations generally feature a lower variability in energy than the other molecules.

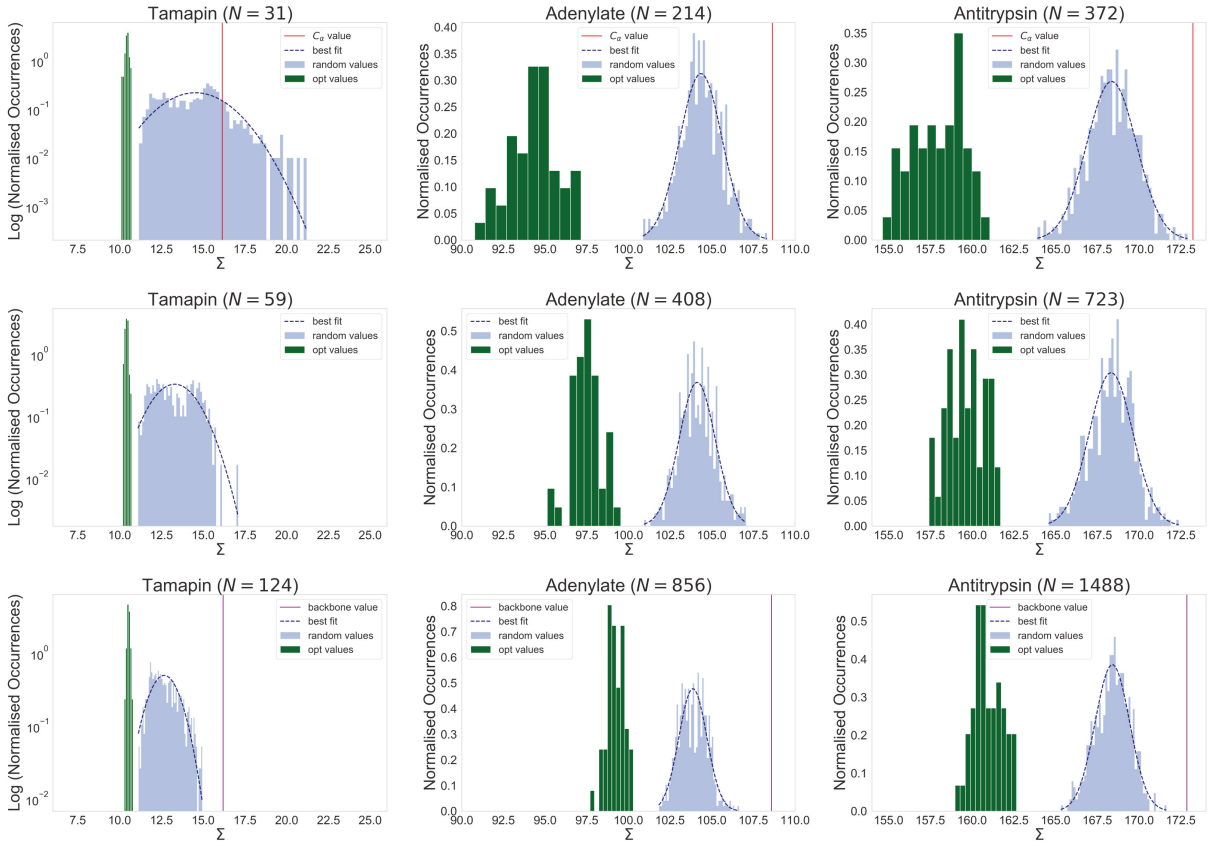

Figure 1: Distributions of the values of mapping entropy Σ $[kJ/\mathrm{mol}/K]$ in Eq. 17 for random mappings (light blue histograms) and optimised solutions (green histograms). Dark blue dashed lines show the best fit with normal distributions over the random cases. Each column corresponds to an analysed protein, each row to a given number N of retained atoms. In the first and last rows, corresponding to numbers of CG sites equal to the number of C_α atoms and of backbone atoms, N_α and N_{bkb} respectively, the values of the mapping entropy associated to the physically-intuitive choice of the CG sites (see text) is indicated by vertical lines (red for $N = N_\alpha$, purple for $N = N_{bkb}$). Note that the S_{map} ranges have the same width in all plots.

For all the investigated proteins, the absence of an overlap between the distributions of Σ associated to random and optimised mappings raises some relevant questions. First, one might wonder what kind of structure the *solution space* has, that is, if the identified

solutions lie at the bottom of a rather flat vessel or, on the contrary, each of them is located in a narrow well, neatly separated one form the other. Second, it is reasonable to ask whether some degree of similarity exists between these quasi-degenerate solutions of the optimisation problems and, in case, what significance this has.

In order to answer these questions, for each structure we select four pairs of mapping operators $\mathbf{M}^{opt}$ that result in the lowest values of Σ. We then perform 100 independent transitions between these solutions, constructing intermediate mappings by randomly swapping two non-overlapping atoms from the two solutions at each step and calculating the associated mapping entropy. Fig. 2 shows the results of this analysis for the pair of mappings with the lowest Σ, all the other transitions being reported in Fig. S2 of the Supplemental Material. It is interesting to notice that the endpoints (that is, the optimised mappings) correspond to the lowest values of Σ along each transition path; by increasing the size of the proteins, the values of Σ for intermediate mappings get closer to the average of Σ_{random}. We cannot rule out, by this analysis, the absence of lower minima over all the possible paths, although it seems quite unlikely given the available sampling.

Finally, it is interesting to observe the pairwise correlations of the site conservation probability within a pool of solutions, as it is informative of the existence of atom pairs that are, in general, simultaneously present, simultaneously absent, or mutually exclusive. As reported in detail in the Supplemental Material (Figs. S6-S7), no clear evidence is available that conserving a given atom can increase or decrease in a statistically relevant manner the conservation probability of another: this behaviour supports the idea that the organisation internal to a given optimal mapping is determined in a nontrivial manner by the intrinsically multi-body nature of the problem at hand.

These analyses thus address the first question by showing that at least the deepest solutions of the optimisation procedure are distinct from each other. It is not possible to (quasi)continuously transform an optimal mapping into another through a series of steps keeping the value of the mapping entropy low. Each of the inspected solutions is a small

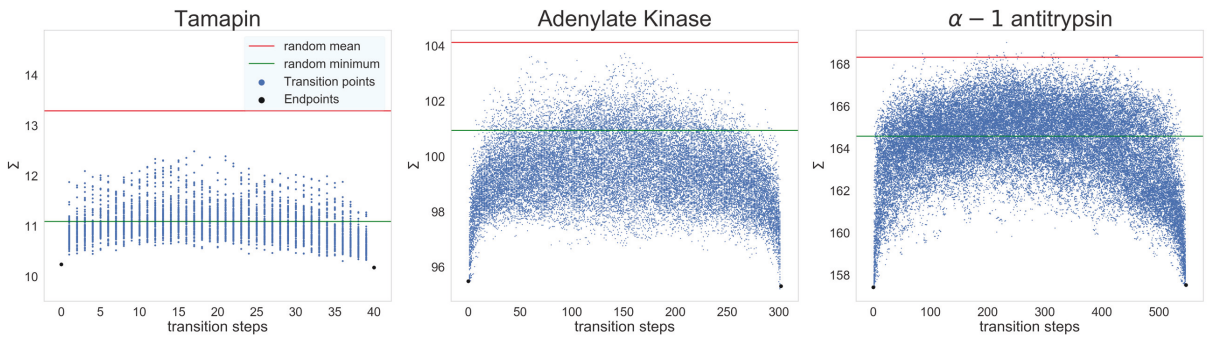

Figure 2: Values of the mapping entropy Σ $[kJ/\mathrm{mol}/K]$ of mappings connecting two optimal solutions. In each plot, one per protein under examination, the two lowest-Σ mappings are taken as initial and final endpoints (black dots) for paths constructed by swapping pairs of atoms between them (blue dots). For each protein, 100 independent paths at given $N = N_{\alpha\beta}$ are constructed and the mapping entropy of each intermediate point is computed. In each plot, horizontal lines represent the mean (red) and minimum (green) S_{map} obtained from the corresponding distribution of random mappings presented in Fig. 1.

town surrounded by high mountains in each direction, isolated from the others with no valley connecting them.

The second question, namely what similarity, if any, exists among these disconnected solutions, is tackled in the following section.

2.4 Biological Significance

The degree of similarity between the optimal mappings can be assessed by a simple average, returning the frequency with which a given atom is retained in the 48 solutions of the optimisation problem.

Fig. 3 shows the probability P_{cons} of conserving each heavy atom, separately for each analysed protein and degree of coarse-graining N investigated, computed as the fraction of times it appears in the corresponding pool of optimised solutions. One can notice the presence of regions that appear to be more or less conserved. Quantitative differences can be observed between the three cases under examination: while the heat map of TAM shows narrow and pronounced peaks of conservation probability, optimal solutions for AKE feature a more uniform distribution, where the maxima and minima of P_{cons} extend over secondary

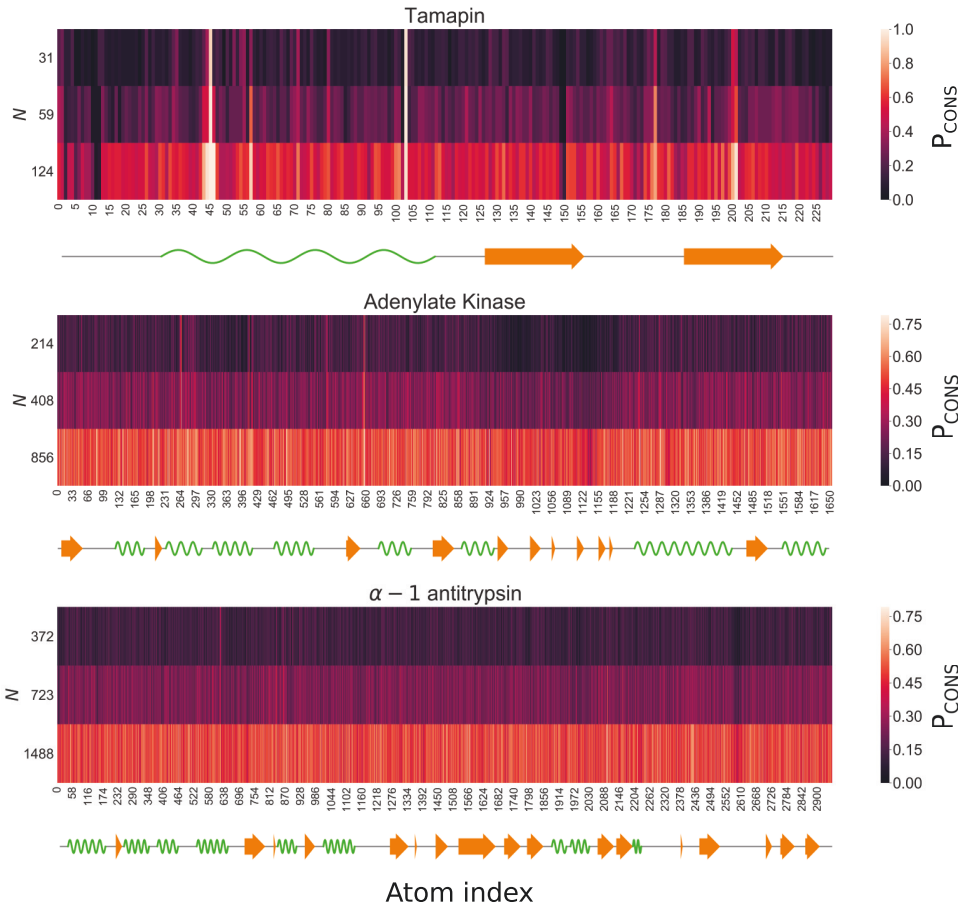

Figure 3: Probability P_{cons} that a given atom is retained in the optimal mapping at various numbers N of CG sites and for each analysed protein, expressed as a function of the atom index. Atoms are ordered according to their number in the PDB file. The secondary structure of the proteins is depicted using Biotite:[64] green waves represent alpha helices and orange arrows correspond to beta strands.

structure fragments rather than small sets of atoms. The distribution gets even more blurred for AAT.

As index proximity does not imply spatial proximity in a protein structure, we mapped the aforementioned probabilities to the three-dimensional configurations. Results for TAM are shown in Fig. 4, while the corresponding ones for AKE and AAT are provided in the Supplemental Material (Fig. S3). From the distribution of P_{cons} at different number of retained sites N it is possible to infer some relevant properties of optimal mappings.

For what concerns TAM (Fig. 4), it seems that, at the highest degree of CG ($N = N_\alpha$), only two sites are always conserved, namely two nitrogen atoms belonging to ARG6 and ARG13 residues ($P_{\mathrm{cons}}(\mathrm{NH1}, \mathrm{ARG6}) = 0.92$, $P_{\mathrm{cons}}(\mathrm{NH2}, \mathrm{ARG13}) = 0.96$). The atoms that constitute the only other arginine residue, ARG7, are well conserved but with lower probability. By increasing the resolution ($N = N_{\alpha\beta}$), i.e., employing more CG sites, we see that the atoms in the side chain of LYS27 appear to be retained more than average together with atoms of GLU24 ($P_{\mathrm{cons}}(\mathrm{NZ}, \mathrm{LYS27}) = 0.65$, $P_{\mathrm{cons}}(\mathrm{OE2}, \mathrm{GLU24}) = 0.75$). At $N = 124$ the distribution becomes more uniform, but still sharply peaked around terminal atoms of ARG6 and ARG13.

Interestingly, ARG6 and ARG13 have been identified to be the main actors involved in the TAM-SK2 channel interaction:[65–67] Andreotti *et al.*[65] suggested that these two residues strongly interact with the channel through electrostatics and hydrogen bonding. Furthermore, Ramírez-Cordero *et al.*[67] showed that mutating one of the three arginines of TAM dramatically decreases its selectivity towards the SK2 channel.

It thus appears that the mapping entropy minimisation protocol was capable of singling out the two residues that are crucial for a complex biological process. The rationale for this can be found in the fact that such atoms strongly interact with the remainder of the protein, so that small variations of their relative coordinates have a large impact on the value of the overall system's energy. Retaining these atoms, and fixing their position in the coarse-grained conformation, thus enables the model to discriminate effectively a macrostate

from another.

We note that this result was achieved solely relying on data obtained in standard MD simulations. This aspect is particularly relevant as the simulation was performed in absence of the channel, whose size is substantially larger than that of TAM. Consequently, we stress that valuable biological information, otherwise obtained *via* large-scale, multi-complex simulations, bioinformatic approaches, or experiments, can be retrieved by means of straightforward simulations of the molecule of interest in absence of its substrate.

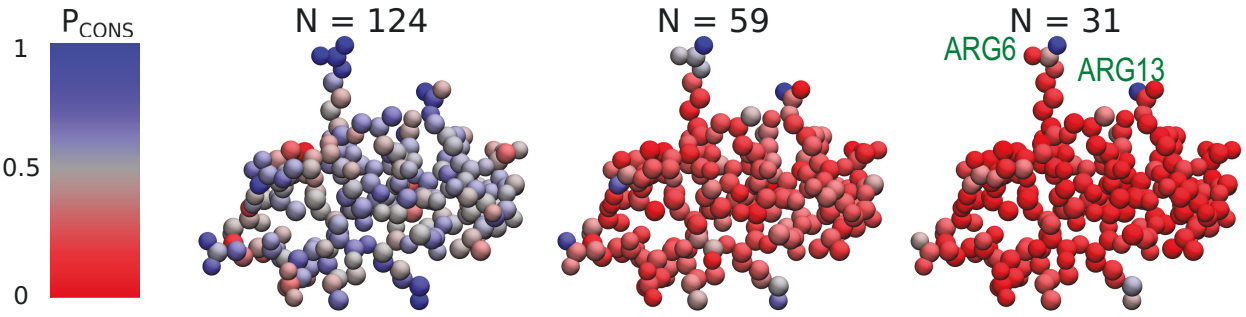

Figure 4: Structure of tamapin (one bead per atom) coloured according to the probability P_{cons} for each atom to be retained in the pool of optimal mappings. Each structure corresponds to a different number N of retained CG sites. Residues presenting the highest retainment probability across N (ARG6 and ARG13) are highlighted.

For the AKE (Fig. S3 in the Supplemental Material) we have that when $N = N_\alpha$ the external, solvent-exposed part of the LID domain is heavily coarse-grained, while its internal region is more conserved. The CORE region of the protein is always largely retained, without noteworthy peaks in probability. Such peaks, on the contrary, appear in correspondence of some terminal nitrogens of ARG36, LYS57 and ARG88 ($P_{\text{cons}}(\text{NH2}, \text{ARG36}) = 0.52$, $P_{\text{cons}}(\text{NZ}, \text{LYS57}) = 0.48$, $P_{\text{cons}}(\text{NH2}, \text{ARG88}) = 0.58$). The two arginine amino acids are located in the internal region of the NMP arm, at the interface with the LID domain. ARG88 is known to be the most important residue for catalytic activity,[68,69] being central in the process of phosphoryl transfer.[70] Phenylglyoxal,[71] a drug that mutates ARG88 to a glycine, has been shown to substantially hamper the catalytic capacity of the enzyme.[70] ARG36 is also bound to phosphate atoms.[69] Finally, LYS57 is on the external part of NMP and has been identified to play a pivotal role in collaboration with ARG88 to block the release of

adenine from the hydrophobic pocket of the protein.[72] More generally, this amino acid is crucial for stabilising the closed conformation of the kinase,[73,74] which was never observed throughout the simulation. The overall probability pattern persists as N increases, even though less pronounced.

As for AAT, Fig. S3 shows that the associated optimisations heavily coarse-grain the reactive center loop of the protein. On the other hand, two of the most conserved residues in the pool of optimised mappings, MET358 and ARG101, are central to the biological role of this serpin. MET358 ($P_{\text{cons}}(\text{CE}, \text{MET358}) = 0.31$) constitutes the reactive site of the protein.[75] Being extremely inhibitor-specific, mutations or oxidation of this amino acid lead to severe diseases. In particular, heavy oxidation of MET358 is one of the main causes of emphysema.[76] The AAT *Pittsburgh* variant shows MET358–ARG mutation, which leads to diminished anti-elastase activity but markedly increased antithrombin activity.[59,75,77] In turn, ARG101 ($P_{\text{cons}}(\text{CZ}, \text{ARG101}) = P_{\text{cons}}(\text{NH1}, \text{ARG101}) = P_{\text{cons}}(\text{NH2}, \text{ARG101}) = 0.35$) has a crucial role is due to its connection to mutations that lead to severe AAT deficiency.[60,61]

In summary we observe that, in all the proteins investigated, the presented approach identifies biologically relevant residues. Most notably, these residues, which are known to be biologically active in presence of other compounds, are singled out *from substrate-free MD simulations*. With the exception of MET358 of AAT, the most probably retained atoms belong to amino acids that are charged and highly solvent-exposed. To quantify the statistical significance of the selection operated by the algorithm, we note that the latter detects those fragments out of a pool of 8, 69 and 100 charged residues for TAM, AKE and AAT, respectively. If we account for solvent exposition, these numbers reduce to 7, 32 and 40 considering amino acids with solvent accessible surface area (SASA) higher than 1 nm^2.

Another aspect worth mentioning is the fact that several atoms pinpointed as highly conserved in optimal mappings are located in the side chains of relatively large residues, such as arginine, lysine and methionine. It is thus legitimate to wonder whether a correlation might exist between an amino acid size and the probability of one or more of its atoms to

be present in a low S_{map} reduced representation. An inspection of the RMSF values of the three proteins' atoms *vs.* their conservation probability (see Fig. S4 in the Supplemental Material) shows no significant correlation for low or intermediate values of P_{cons}; highly conserved atoms, on the other hand, tend to be located on highly mobile residues because a relatively large conformational variability is a prerequisite for an atom to be determinant in the mapping. In conclusion, highly mobile residues are not necessarily highly conserved, while the opposite is more likely.

3 Discussion and Conclusions

In this work we have addressed the question of identifying the subset of atoms of a macromolecule, specifically a protein, that entails the largest amount of information about its conformational distribution while employing a reduced number of degrees of freedom with respect to the reference. The motivation behind this objective is to provide a synthetic yet informative representation of a complex system, simulated in high resolution but observed in low resolution, thus rationalising its properties and behaviour in terms of relatively few important variables–namely the positions of the retained atoms.

This goal was pursued making use of tools and concepts largely borrowed from the field of coarse-grained modelling, in particular bottom-up coarse-graining. The latter term identifies a class of theoretical and computational strategies employed to construct a simplified model of a system that, if treated in terms of a high-resolution description, would otherwise be too onerous to simulate. Coarse-graining methods make use of the configurational landscape of the reference high-resolution model to construct a simplified representation that retains its large-scale properties. The interactions among effective sites are parametrised by directly integrating out (in an exact or approximate manner) the higher-resolution degrees of freedom, and imposing the equality of the probability distributions of the coarse-grained degrees of freedom in the two representations.[5]

These approaches have a long and successful history in the field of statistical mechanics and condensed matter, the most prominent, pioneering example probably being Kadanoff's spin block transformations of ferromagnetic systems.[78] This process, which lies at the heart of real-space renormalisation group (RG) theory, allows the relevant variables of the system to naturally emerge out of a (potentially infinite) pool of fundamental interactions, thus linking microscopic physics to macroscopic behaviour.[79,80]

The generality of the concepts of renormalisation group and coarse-graining has naturally taken them outside of their native environment,[81–83] the whole field of coarse-grained modelling of soft matter being one of the most fruitful offsprings of this cross-fertilisation.[5] However, a straightforward application of RG methods in this latter context is severely restricted by fundamental differences between the objects of study. Most notably, the crucial assumptions of self-similarity and scale invariance, which justify the whole process of renormalisation at the critical point, clearly do not apply to, say, a protein, in that the latter does certainly not resemble itself upon resolution reduction. Furthermore, scaling laws cannot be applied to a system such as a biomolecule that is intrinsically finite, for which the thermodynamic limit is not defined.

Additionally, one of the key consequences of self-similarity at the critical point is that the filtering process put forward by the renormalisation group turns out to be largely independent of the specific coarse-graining prescription: the set of relevant macroscopic variables emerges as such for almost whatever choice of mapping operator is taken to bridge the system across different length scales.[84] In the case of biological matter, where the organisation of degrees of freedom is not fractal, rather hierarchical—from atoms, to residues, to secondary structure elements, and so on—the mapping operator acquires instead a central role in the "renormalisation" process. The choice of a particular transformation rule, projecting an atomistic conformation of a molecule to its coarse-grained counterpart, more severely implies an external—i.e. not *emergent*—selection of which variables are relevant in the description of the system, and which others are redundant. In this way, what should

be the main outcome of a genuine coarse-graining procedure is demeaned to be one of its ingredients.

It is only recently that the central importance of the resolution distribution, i.e., the definition of the CG representation, has gradually percolated in the field of biomolecular modelling.[22,44] Moving away from an *a priori* selection of the effective interaction sites,[21] few different strategies have been developed that rather aim at the automatic identification of CG mappings. These techniques rely on specific properties of the system under examination: examples include quasi-rigid domain decomposition,[24–31] or graph theory–based model construction methods that attempt at creating CG representations of chemical compounds based only on their static graph structure;[32,33,85] other approaches aim at selecting those representations that closely match the high resolution model's energetics.[22,35] Finally, more recent strategies rooted in the field of machine learning generate discrete CG variables by means of variational autoencoders.[86] All these methods take into account the system structure, or its conformational variability, or its energy, but none of them integrates these complementary properties in a consistent framework embracing topology, structure, dynamics, and thermodynamics.

In this context, information-theoretical measures, such as the mapping entropy,[17,42–44] can bring novel and potentially very fruitful features.[87] In fact, this quantity associates structural and thermodynamical properties, so that both the conformational variability of the system and its energetics are accurately taken into account. Making use of the advantages offered by the mapping entropy, we developed a protocol to identify, in an automated, unsupervised manner, the low-resolution representation of a molecular system that maximally preserves the amount of thermodynamical information contained in the corresponding higher-resolution model.

The results presented here suggest that the method may be capable of identifying not only thermodynamically consistent, but also biologically informative mappings. Indeed, a central result we reported is that those atoms consistently retained with high probability

across various lowest-S_{map} mappings at different CG site numbers tend to be located in amino acids that play a relevant role in the function of the three proteins under examination. Most importantly, these key residues, whose biological activity consists in binding with other molecules, have been singled out on the basis of plain MD simulations of the substrate-free molecules in explicit solvent. In general, the vast majority of available techniques for the identification of putative binding or allosteric sites in proteins rely, explicitly or implicitly, on the analysis of the interaction between the molecule of interest and its partner—be that a small ligand, another protein, or else.[88–93] This is the case, for example, of binding site prediction servers,[94,95] which perform a structural comparison between the target protein and those archived in a precompiled, annotated database; other bioinformatic tools make use of machine learning methods[96–99]—with all pros and cons that come with the training over a possibly vast, but certainly finite dataset of known cases.[100] To the best of our knowledge, the remaining alternative methods perform a structural analysis of the protein in search of binding pockets based on purely geometrical criteria.[101,102] The results obtained in the present work, on the contrary, suggest that a significant fraction of biologically relevant residues, whose function is intrinsically related to the interaction with other molecules, might be identified as such from the analysis of simulations *in absence of the substrate*. This observation would imply that a substantial amount of information about functional residues, even those that exploit their activity through the interaction with a partner molecule, is entailed in the protein's own structure and energetics. In the past few decades, the successful application of extremely simplified representations of proteins such as elastic network models has shown that the key features of a protein's large-scale dynamics are encoded in its native structure;[27,36–41,103–107] in analogy with this, we hypothesise that the mapping entropy minimisation protocol is capable of bringing to light those *relational* properties of proteins—namely the interaction with a substrate—from the thermodynamics of the single molecule, in absence of its partner.

The mapping entropy minimisation protocol establishes a quantitative bridge between

a molecule's representation—and hence its information content—on the one side, and the structure-dynamics-function relationship on the other. This method might represent a novel and useful tool in various fields of applications, e.g. for the identification of important regions of proteins, such as druggable sites and allosteric pockets, relying on simple, substrate-free MD simulations, and efficient analysis tools. In this study, a first exploration of the method's capabilities, limitations, and potential developments has been carried out, and several perspectives lie ahead that deserve further exploration. Among the most pressing and interesting ones we mention the investigation of how the optimised mappings depend on the conformational space sampling; the relation of the mapping entropy minimisation with more established schemes such as the maximum entropy method; and the viability of a machine learning-based implementation of the protocol, e.g. making use of deep learning tools that have proven to be strictly related to coarse-graining, dimensionality reduction, and feature extraction. All these avenues are the object of ongoing study.

In conclusion, it is our opinion that the proposed automated selection of coarse-grained sites entails a great potential for further development, being at the nexus between molecular mechanics, statistical mechanics, information theory, and biology.

4 Methods

In this section we describe the technical preliminaries and the details of the algorithm we employ to obtain the CG representation $\mathbf{M}$, see Eq. 2, that minimises the loss of information inherently generated by a CG'ing procedure—that is, the mapping entropy.

Eq. 15 provides us with a way of measuring the mapping entropy of a biomolecular system associated to any particular choice of decimation of its atomistic degrees of freedom. One can visualise a decimation mapping (Eq. 2) as an array of bits, where 0 and 1 correspond to *not retained* and *retained* atoms, respectively. Order matters: swapping two bits produces a different mapping operator. Applying this procedure, one finds that the total number

of possible CG representations of a biomolecule, irrespectively of how many atoms N are selected out of n, is

$$\sum_{N=0}^{n} \frac{n!}{N!\ (n-N)!} = 2^n, \tag{21}$$

which is astronomical even for the smallest proteins. In this work we restrict the set of possibly retained sites to the N_{heavy} heavy atoms of the compound—excluding hydrogens—thus significantly reducing the cardinality of the space of mappings. Nonetheless, finding the global minimum of Eq. 15 for a reasonably large molecule would be computationally intractable whenever N is different from $1, 2$ and $N_{heavy} - 1, N_{heavy} - 2$. As an example, there are 2.4×10^{38} CG representations of tamapin with 31 sites ($N = N_\alpha$) and 9.6×10^{887} for antitrypsin with 1488 sites ($N = N_{bkb}$).

Hence, it is necessary to perform the minimisation of the mapping entropy through a Monte Carlo-based optimisation procedure, and we specifically rely on the simulated annealing (SA) protocol.[62,63] As it is typically the case with this method, the computational bottleneck consists in the calculation of the observable (the mapping entropy) at each SA step.

We develop an approximate method that is able to obtain the mapping entropy of a biomolecule by analysing a MD trajectory that can contain up to tens of thousands of frames. At each SA step, that is, for each putative mapping, the algorithm calculates a similarity matrix among all the generated configurations. The entries of this matrix are given by the root mean square deviation (RMSD) between structure pairs, the latter being defined only in terms of the retained sites associated to the CG mapping, and aligned accordingly; we then identify CG macrostates by clustering frames based on the distance matrix, making use of bottom-up hierarchical clustering (UPGMA[108]). Finally, we determine the observable of interest from the variances of the atomistic intramolecular potential energy of the protein corresponding to the frames that map onto the same CG conformational cluster, see Eq. 16.

The protocol is initiated with the generation of a mapping such that the overall number of retained sites is equal to N. Then, at each SA step, the following operations are performed:

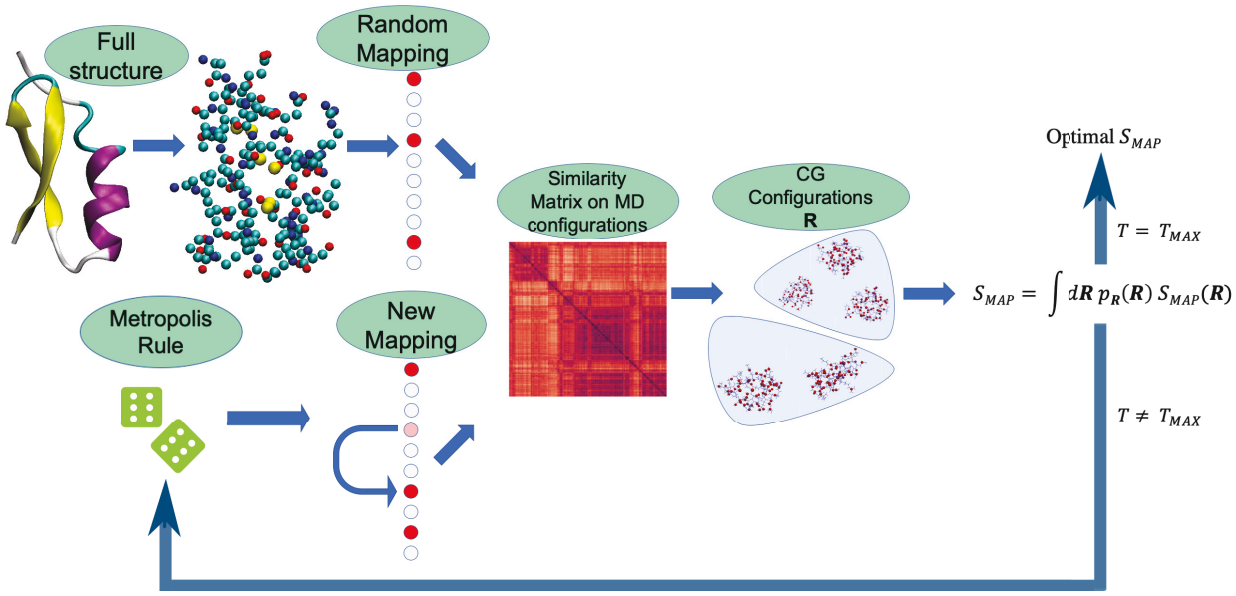

Figure 5: Schematic representation of the algorithmic procedure described in the text that we employ to minimise the mapping entropy, the latter being calculated by means of Eq. 25. The full similarity matrix is computed once every T_K steps, while in the intermediate steps we resort to the approximation given by Eq. 23. T_K depends both on the protein and on N. T_{MAX} is the number of simulated annealing steps, $T_{MAX} = 2 \times 10^4$.

1. swap a retained site ($\sigma_i = 1$) and a removed site ($\sigma_j = 0$) in the mapping;

2. compute a similarity matrix among CG configurations using the RMSD;

3. apply a clustering algorithm on the RMSD matrix in order to identify the CG macrostates **R**;

4. compute $\tilde{S}_{map}$ using Eq. 16.

Once the new value of $\tilde{S}_{map}$ is obtained, the move is accepted/rejected using a Metropolis-like rule. The overall workflow of the algorithm is schematically illustrated in Fig. 5.

For the sake of the accuracy of the optimisation, the more exhaustive the sampling the better, hence the number of sampled atomistic configurations should be at least of the order of the tens of thousands. However, in that case step 2 would require to align a huge number of structure pairs for each proposed CG mapping, which in turn would dramatically slow down the entire process. This problem is circumvented performing a reasonable approximation in

the calculation of the CG RMSD matrix.

4.1 RMSD matrix calculation

The RMSD between two *superimposed* structures $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$RMSD(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{y}_i)^2}, \tag{22}$$

where n is the number of sites in the system, being they atomistic or CG, and $\boldsymbol{x}_i$, $\boldsymbol{y}_i$ represent the cartesian coordinates of the i-th element in the two sets. According to Kabsch[109,110] it is possible to find the superimposition that minimises this quantity, namely the rotation matrix U that has to be applied to $\boldsymbol{x}$ for a given $\boldsymbol{y}$ in order to reach the minimum of the RMSD.

The aforementioned procedure is not computationally heavy *per se*; in our case, however, we would have to repeat this alignment for all configuration pairs in the MD trajectory every time a new CG mapping is proposed along the Monte Carlo process, thus making the overall workflow inctractable in terms of computational investment.

The simplest solution to this problem is to discard the differences in the Kabsch alignment between two CG structures differing by a pair of swapped atoms. This assumption is particularly appealing from the point of view of speed and memory, since the expensive and relatively slow alignment procedure produces a result (a rotation matrix) that can be stored with negligible use of resources. In order to take advantage of this simplification without losing accuracy, for each structure and degree of CG we select an interval of Simulated Annealing steps T_K in which we consider rotation matrices constant. After these steps, the full Kabsch alignment is applied again.

This approximation results in a substantial reduction of the number of operations that we have to execute at each Monte Carlo step. At first, given the initial random mapping operator $\mathbf{M}$, we build the sets of coordinates that have been conserved by the mapping

operator $\Gamma(\mathbf{M}) = \mathbf{M}(\mathbf{r})$. Then we compute the overall RMSD matrix between every pair of aligned structures Γ_α and Γ_β, $RMSD(\Gamma_\alpha(\mathbf{M}), \Gamma_\beta(\mathbf{M}))$, where α and β run over the MD configurations. For all moves $\mathbf{M} \to \mathbf{M}'$ within a block of T_K Monte Carlo steps, $\mathbf{M}$ and $\mathbf{M}'$ only differing in a pair of swapped atoms, this quantity is then updated with the simple rule

$$MSD(\Gamma_\alpha(\mathbf{M}'), \Gamma_\beta(\mathbf{M}')) = MSD(\Gamma_\alpha(\mathbf{M}), \Gamma_\beta(\mathbf{M})) - \frac{1}{N} MSD(\Gamma_\alpha(\boldsymbol{s}), \Gamma_\beta(\boldsymbol{s})) + \frac{1}{N} MSD(\Gamma_\alpha(\boldsymbol{a}), \Gamma_\beta(\boldsymbol{a})), \tag{23}$$

where $\boldsymbol{s}$ and $\boldsymbol{a}$ are the removed (substituted) and added atom, respectively, and MSD is the Mean Squared Deviation.

This approach clearly represents an approximation to the correct procedure; it has to be emphasised, however, that the impact of such approximation is increasingly perturbative as the size of the system grows. Furthermore, the computational gain that the described procedure enables is sufficient to counterbalance the fact that the exact protocol would be so inefficient to make the optimisation impossible. For example, choosing $T_K = 1000$ for AAT with $N = N_{bkb}$ our approximation gives a speed-up factor of the order of 10^3.

4.2 Hierarchical Clustering of Coarse Grained configurations

Several clustering algorithms exist that have been applied to group molecular structures based on RMSD similarity matrices.[111,112] Many such algorithms have been developed and incorporated in the most common libraries for data science. Among the various available methods we choose to resort on the agglomerative bottom-up hierarchical clustering with average linkage (UPGMA algorithm[108]). We here briefly recapitulate the basics underpinnings of this procedure.

1. At the first step, the minimum of the similarity matrix is found and the two corresponding entries x, y (*leaves*) are merged together in a new cluster k;

2. k is placed in the middle of its two constituents. The distance matrix is updated to

take into account the presence of the new cluster in place of the two *close* structures: $d(k,z) = (d(x,z) + d(y,z))/2$;

3. Steps 1. and 2. are iterated until one *root* is found. The distance among clusters k and w is generalised as follows:

$$d(k,w) = \sum_{i \in k} \sum_{j \in w} \frac{d(k[i], w[j])}{|k| \times |w|} \tag{24}$$

where $|k|$ and $|w|$ are the populations of the clusters and $k[i]$ and $w[j]$ their elements;

4. The actual division in clusters can be performed by cutting the tree (*dendrogram*) using a threshold value on the inter-clusters distance or taking the first value of distance that gives rise to a certain number of clusters N_{cl}. In both cases it is necessary to introduce a hyperparameter. In our case the latter is a more viable choice to reduce the impact of roundoff errors. Indeed, the first criterion would push the optimisation to create as many clusters as possible, in order to minimise the energy variance inside them (a cluster with one sample has zero variance in energy).

This algorithm, whose implementation[113,114] is available in Python Scipy,[115] is simple, relatively fast ($O(n^2 \, log \, n)$), and completely deterministic: given the distance matrix, the output dendrogram is unique.

Although this algorithm scales well with the size of the dataset, it may not be robust with respect to small variations along the optimisation trajectory. In fact, even the slightest modifications of the dendrogram may lead to abrupt changes in $\tilde{S}_{map}$. This is perfectly understandable from an algorithmic point of view, but it is deleterious for the stability of the optimisation procedure. Furthermore, the aforementioned choice of N_{cl} is somehow arbitrary. Hence, we perform the following analysis in order to enhance the robustness of $\tilde{S}_{map}$ at each MC move and to provide a quantitative criterion to set the hyperparameter:

1. Compute the RMSD similarity matrix between all the heavy atoms of the biological

system under consideration;

2. Apply UPGMA algorithm to this object, retrieving the all-atom dendrogram;

3. Impose lower and upper bounds (see Table 3) on the inter-clusters distance depending on the conformational variability of the structure;

4. Visualise the cut dendrogram to identify the number of different clusters available at each of the two values of the threshold (N_{cl}^{+} and N_{cl}^{-}) (Table 3);

5. Build a list CL of five integers selecting three (intermediate) values between N_{cl}^{-} and N_{cl}^{+};

6. Define the observable as the average over the values of $\tilde{S}_{map}$ (see Eq. 16) computed choosing different N_{cl}:

$$\Sigma = \frac{1}{|\mathrm{CL}|} \sum_{N_{cl} \in \mathrm{CL}} \tilde{S}_{map}(N_{cl}) \tag{25}$$

where |CL| is the cardinality of the list we chose.

Table 3: Bounds on inter-clusters distance and correspondent number of clusters.

Protein	Upper bound (nm)	Lower bound (nm)	N_{cl}^{+}	N_{cl}^{-}
Tamapin	0.20	0.18	91	34
Adenylate Kinase	0.25	0.20	147	29
$\alpha - 1$ antytripsin	0.20	0.15	96	7

The overall procedure amounts at identifying many different sets of CG macrostates $\mathbf{R}$ on which $\tilde{S}_{map}$ can be computed, assuming that the average of this quantities can be used effectively as driving observable inside the optimisation. This trivial assumption allows to increase the robustness of the SA optimisation and to keep in memory all the values of $\tilde{S}_{map}$ calculated at different distances from the root of the dendrogram.

4.3 Simulated Annealing

We use Monte Carlo simulated annealing to stochastically explore the space of the possible decimation mappings associated to each degree of CG'ing. We here briefly describe the main features of our implementation of the SA algorithm, referring the reader to a few excellent reviews for a comprehensive description of the techniques that can be employed in the choice of temperature decay and parameter estimation.[116,117]

We run the optimisation for 2×10^3 MC epochs, each of which is composed by 10 steps. This amounts at keeping the temperature constant for 10 steps and then decreasing it according to an exponential law. For the i-th epoch we have that $T(i) = T_0 \; e^{-i/\nu}$.

The hyperparameters T_0 and ν are crucial for a well-behaved MC optimisation. We choose $\nu = 300$ so that the temperature at $i = 2000$ is approximately $T_0/1000$. In order to feed our algorithm with reasonable values of T_0, for each of 100 random mappings we perform 10 MC stochastic moves, measuring $\Delta\Sigma$, namely the difference between the observables computed at two consecutive steps. Then we estimate T_0 so that a move that leads to an increment of the observable equal to the average of $\Delta\Sigma$ would possess an acceptance probability of 0.75 at the first step.

4.4 Data available

For each analysed protein, the raw data about all the CG representations investigated in this work including random, optimised and transition mappings are freely available on the Zenodo repository `https://zenodo.org/record/3776293` together with the associated mapping entropies. We further provide all the scripts we employed to analyse such data and construct all the figures presented in this work.

Acknowledgement

The authors thank Attilio Vargiu for critical reading of the manuscript and useful comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758588). MSS acknowledges funding from the U.S. National Science Foundation through award CHEM-1800344.

5 Supporting information

The Supporting Information file contains detailed information on the following topics:

- a quantitative analysis of the all-atom MD simulations of the three proteins investigated in this work
- additional figures about the CG representations that minimise the mapping entropy
- an analysis of the relation between the size and mobility of residues and the conservation probability of their atoms
- an assessment of the results' stability with respect to the duration of the MD trajectory.

This information is available free of charge via the Internet at http://pubs.acs.org

A Relative and mapping entropy

Bottom-up coarse-graining approaches aim at constructing effective, low-resolution representations of a system that reproduce as accurately as possible the equilibrium statistical mechanical properties of the underlying, high-resolution reference. In particular, this problem is phrased in terms of the parametrisation of a CG potential that approximates the

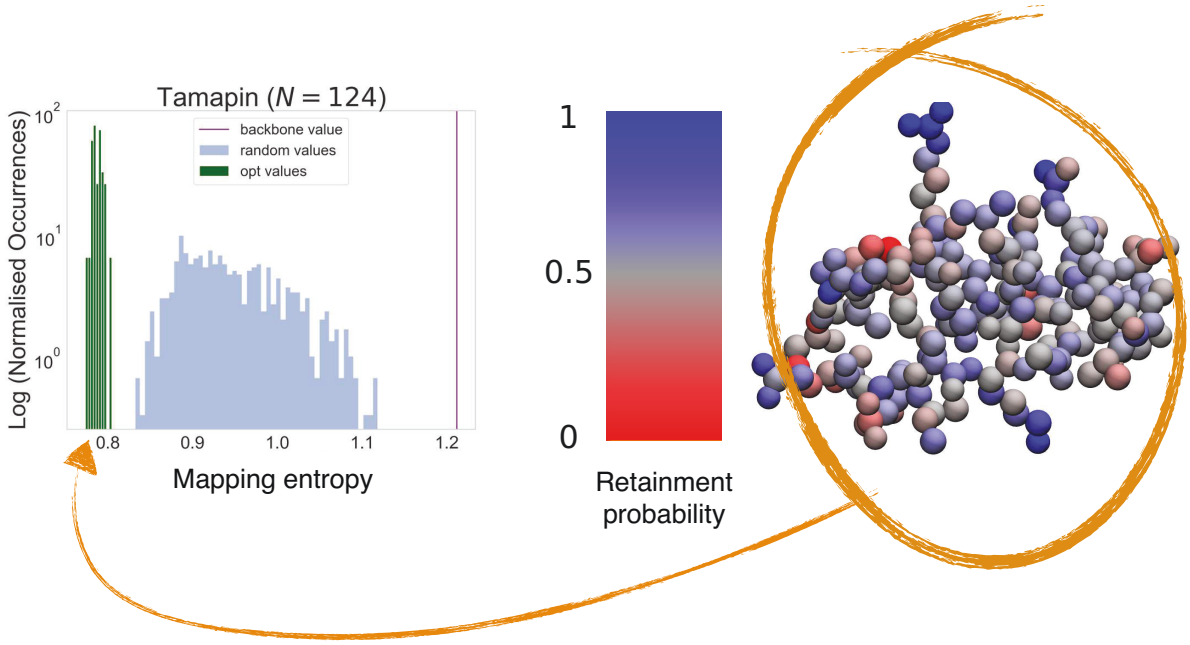

Figure 6: Table of Content figure.

reference system's multi-body potential of mean force (PMF) U^0,

$$U^0 = -k_B T \ln(V^N p_R(\mathbf{R})) + const, \tag{26}$$

where $p_R(\mathbf{R})$ is the probability for the atomistic model to sample a specific CG configuration $\mathbf{R}$. In the canonical ensemble, one has

$$\begin{aligned} p_R(\mathbf{R}) &= \int d\mathbf{r}\ p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \\ &= \frac{1}{Z}\int d\mathbf{r}\ e^{-\beta u(\mathbf{r})}\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \end{aligned} \tag{27}$$

where $\beta = 1/k_B T$, $u(\mathbf{r})$ is the microscopic potential energy of the system, $p_r(\mathbf{r}) \propto \exp(-\beta u(\mathbf{r}))$ is the Boltzmann distribution and Z the associated configurational partition function.

From Eqs. 26 and 27 it follows that a computer simulation of the low-resolution system performed with the potential U^0 (more precisely, a free energy) would allow the CG sites to sample their configurational space with the same probability as they would do in the

reference system. Unfortunately, the intrinsically multi-body nature of U^0 is such that its exact determination is largely unfeasible in practice.[118] Considerable effort has thus been devoted to devise increasingly accurate methods to approximate the PMF with a CG potential U;[16,17,119,120] however, the latter is in general defined in terms of a necessarily incomplete set of basis functions.[4–7] It is thus natural to look for quantitative measures of a CG model's quality with respect to U^0.

In this respect, one of the most notable examples of such metrics is the relative entropy,[17,42–44]

$$\begin{aligned} S_{rel} &= k_B \times D_{KL}(p_r(\mathbf{r})||P_r(\mathbf{r}|U)) \\ &= k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln\left[\frac{p_r(\mathbf{r})}{P_r(\mathbf{r}|U)}\right], \end{aligned} \tag{28}$$

where $D_{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence between two probability distributions,[45] with $S_{rel} \geq 0$ by virtue of Gibbs' inequality. In Eq. 28, $p_r(\mathbf{r})$ is the atomistic probability distribution of the system, see Eq. 27, while $P_r(\mathbf{r}|U)$ is defined as a product of probabilities over CG and AA configurational spaces,[42,44]

$$P_r(\mathbf{r}|U) = \frac{p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))} P_R(\mathbf{M}(\mathbf{r})|U). \tag{29}$$

The term $P_R(\mathbf{R}|U) \propto \exp(-\beta U(\mathbf{R}))$ in Eq. 29 runs over CG configurations, and describes the probability that a CG model with approximate potential $U(\mathbf{R})$ samples the CG configuration $\mathbf{R}$. Then, to obtain $P_r(\mathbf{r}|U)$ it is sufficient to multiply $P_R(\mathbf{R}|U)$ by the atomistic probability $p_r(\mathbf{r})$ of sampling $\mathbf{r}$, normalised by the Boltzmann weight $p_R(\mathbf{R})$ of the CG configuration $\mathbf{R}$ (see Eq. 27).

KL divergences quantify the information loss between probability distributions; specifically, $D_{KL}(s(\mathbf{r})||t(\mathbf{r}))$ represents the information that is lost by representing a system originally described by a probability distribution $s(\mathbf{r})$ through a distribution $t(\mathbf{r})$.[45] Given a CG mapping $\mathbf{M}$, the relative entropy S_{rel} in Eq. 28 implicitly measures the loss that arises as a

consequence of approximating the exact CG potential of mean force U^0 of a system by an effective potential U, that is, the error introduced by using incorrect interactions to describe the low-resolution system. By replacing Eq 29 in Eq. 28 and introducing $1 = \int d\mathbf{R}\ \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$, one indeed obtains

$$S_{rel} = k_B \int d\mathbf{R}\ p_R(\mathbf{R}) \ln \left[\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)} \right], \tag{30}$$

a KL divergence $D_{KL}(p_R(\mathbf{R})||P_R(\mathbf{R}|U))$ between the *exact* and *approximate* probability distributions in the CG configuration space, with no explicit connection to the underlying microscopic reference. As such, S_{rel} is a measure of an approximate CG model's quality. However, it is possible to expand S_{rel} as a difference between two information losses (the one due to U and the one due to U^0) calculated with respect to the atomistic system,

$$\begin{aligned} S_{rel} &= k_B \times D_{KL}(p_r(\mathbf{r})||V^{N-n} P_R(\mathbf{M}(\mathbf{r})|U)) \\ &- k_B \times D_{KL}(p_r(\mathbf{r})||V^{N-n} p_R(\mathbf{M}(\mathbf{r}))) \\ &= k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln \left[\frac{V^n}{V^N} \frac{p_r(\mathbf{r})}{P_R(\mathbf{M}(\mathbf{r})|U)} \right] \\ &- k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln \left[\frac{V^n}{V^N} \frac{p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))} \right], \end{aligned} \tag{31}$$

where n and N denote the number of atomistic and CG sites, respectively.

Both KL divergences in Eq. 31 are positive defined due to Gibbs' inequality, with $D_{KL}(p_r(\mathbf{r})||V^{N-n} P_R(\mathbf{M}(\mathbf{r})|U)) \geq D_{KL}(p_r(\mathbf{r})||V^{N-n} p_R(\mathbf{M}(\mathbf{r})))$ as $S_{rel} \geq 0$; the second one is called mapping entropy [1] S_{map},[17,42,44]

$$S_{map} = k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln \left[\frac{V^n}{V^N} \frac{p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))} \right] \geq 0, \tag{32}$$

[1] In this work we employ a different sign convention for the mapping entropy S_{map} with respect to Refs.,[42,44] and consistent with the one in Ref.[17] On one hand, this enables the mapping entropy to be directly related to a loss of information in the KL sense—a *positive* KL divergence implies a *loss* of information. On the other hand, it allows the relative entropy in Refs.[42,44] to be considered a difference of information losses—those of U and U^0, see Eq. 31—calculated with respect to the atomistic system, so that the vanishing of S_{rel} for $U = U^0$ in Refs.[42,44] effectively amounts at recalibrating the zero of the relative entropy as originally defined in Ref.[17]

which noteworthy does not depend on the approximate CG force field U but only on the mapping operator $\mathbf{M}$.

In multi-scale modelling applications, one seeks to minimise the relative entropy with respect to coefficients in terms of which the coarse-grained potential $U(\mathbf{R})$ is parametrised, for a *given* mapping.[17,42–44] The aim is to generate CG configurations that sample the *atomistic* conformational space with the same microscopic probability $p_r(\mathbf{r})$, see Eq. 28. However, since the model can only generate configurations in the CG space, minimising Eq. 28 is tantamount to minimise Eq. 30, that is, the "error" introduced by approximating U_0 with U; furthermore, in the minimisation with respect to U the contribution of the mapping entropy vanishes, because the latter does not depend on the coarse-grained potential. In this context, then, S_{map} only represents a constant shift of the KL distance between the all-atom and the coarse-grained models, and a minimisation of the first term of Eq. 31 is equivalent to that of Eq. 28.

When taken *per se*, on the other hand, the mapping entropy provides substantial information about the modelling of the system. In fact, this quantity represents the loss of information that would be inherently generated by reducing the resolution of a system even in the case of an *exact* CG'ing procedure, in which $U = U^0$ and $S_{rel} = 0$.[42] In the calculation of S_{map}, the reference AA density is compared to a distribution in which probabilities are smeared out and redistributed equally to all the microscopic configurations $\mathbf{r}$ inside each CG macrostate.

Starting from Eq. 32, Rudzinski *et al.* further divide S_{map} into a sum of two terms,[42]

$$\begin{aligned} S_{map} &= -k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{V^N}{V^n} \Omega_1(\mathbf{M}(\mathbf{r})) \right] \\ &+ k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})} \right], \end{aligned} \tag{33}$$

where the first one is purely geometrical while the second one accounts for the smearing in probability generated by the CG'ing procedure. In Eq. 33, $\Omega_1(\mathbf{M}(\mathbf{r})) = \int d\mathbf{r} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$ is

the degeneracy of the CG macrostate $\mathbf{R}$—i.e., how many microstates map onto a given CG configuration—and

$$\bar{p}_r(\mathbf{r}) = p_R(\mathbf{M}(\mathbf{r}))/\Omega_1(\mathbf{M}(\mathbf{r})) \tag{34}$$

is the average probability of all microstates that map to the macrostate $\mathbf{R} = \mathbf{M}(\mathbf{r})$.

The geometric term in Eq. 33 does not vanish in general.[42] However, if the mapping takes the form of a decimation, see Eq. 2, one has

$$\Omega_1(\mathbf{M}(\mathbf{r})) = V^{n-N}, \tag{35}$$

and the first logarithm in Eq. 33 is identically zero, so that

$$S_{map} = k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln \left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})} \right]. \tag{36}$$

In the case of decimation mappings, moreover, a direct relation holds between the mapping entropy S_{map} as expressed in Eq. 36 and the non-ideal configurational entropies of the original and CG systems,[42,44]

$$s_r = -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln(V^n p_r(\mathbf{r})), \tag{37}$$

$$s_R = -k_B \int d\mathbf{R}\ p_R(\mathbf{R}) \ln(V^N p_R(\mathbf{R})). \tag{38}$$

Indeed, by introducing Eq. 27 in Eq. 38 s_R can be rewritten as

$$\begin{aligned} s_R &= -k_B \int d\mathbf{R} \left[\int d\mathbf{r}\ p_r(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})) \right] \ln(V^N p_R(\mathbf{R})) \\ &= -k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln(V^N p_R(\mathbf{M}(\mathbf{r}))). \end{aligned} \tag{39}$$

Subtracting Eq. 37 and 39 results in

$$s_R - s_r = k_B \int d\mathbf{r}\ p_r(\mathbf{r}) \ln \left(\frac{V^{n-N} p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))} \right), \tag{40}$$

and by virtue of Eq. 34 and 35, one finally obtains

$$s_R - s_r = S_{map}, \tag{41}$$

further highlighting that the mapping entropy represents the difference in information content between the distribution obtained by reducing the level of resolution at which the system is observed, $p_R(\mathbf{R})$, and the original, microscopic reference, $p_r(\mathbf{r})$.

B Explicit calculation of the mapping entropy

We here provide full detail of our derivation of the mapping entropy, as in Eqs. 10-12, and its cumulant expansion approximation, Eq. 15, starting from Eq. 36.

In the case of CG representations obtained by decimating the number of original degrees of freedom of the system, the mapping entropy S_{map} in Eq. 36 vanishes if the probabilities of the microscopic configurations that map onto the same CG one are the same.[42,44] In the canonical ensemble, the requirement is that those configurations must possess the same energy. This can be directly inferred by writing the negative of the average in Eq 36 as

$$\begin{aligned} \left\langle \ln \left[\frac{\bar{p}_r(\mathbf{r})}{p_r(\mathbf{r})} \right] \right\rangle = \int d\mathbf{r}\ p_r(\mathbf{r}) \times \\ \ln \left[\frac{\int d\mathbf{r}' \exp[-\beta(u(\mathbf{r}') - u(\mathbf{r}))] \delta(\mathbf{M}(\mathbf{r}') - \mathbf{M}(\mathbf{r}))}{\int d\mathbf{r}' \delta(\mathbf{M}(\mathbf{r}') - \mathbf{M}(\mathbf{r}))} \right], \end{aligned} \tag{42}$$

so that if $u(\mathbf{r}') = u(\mathbf{r})\ \forall\ \mathbf{r}'$ s.t. $\mathbf{M}(\mathbf{r}') = \mathbf{M}(\mathbf{r})$, the argument of the logarithm is unity and the right-hand side of Eq. 42 vanishes.

Importantly, this implies that no information on the system is lost along the coarse-graining procedure if CG macrostates are generated by grouping together microscopic configurations characterised by having the same energy. In our case, this translates into the search for *isoenergetic mappings*.

By introducing $1 = \int d\mathbf{R}\ \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$ in Eq. 42, one obtains

$$S_{map} = -k_B \int d\mathbf{R} \int d\mathbf{r}\ p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \times \tag{43}$$
$$\ln\left[\frac{\int d\mathbf{r}' \exp[-\beta(u(\mathbf{r}') - u(\mathbf{r}))]\delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}{\int d\mathbf{r}'\delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}\right]$$
$$= \int d\mathbf{R}\ p_R(\mathbf{R})S_{map}(\mathbf{R}), \tag{44}$$

so that the overall mapping entropy is decomposed as a weighted average over the CG configuration space of the mapping entropy $S_{map}(\mathbf{R})$ of a *single* CG macrostate,

$$S_{map}(\mathbf{R}) = -\frac{k_B}{p_R(\mathbf{R})} \int d\mathbf{r}\ p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \times \tag{45}$$
$$\ln\left[\frac{\int d\mathbf{r}' \exp[-\beta(u(\mathbf{r}') - u(\mathbf{r}))]\delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}{\int d\mathbf{r}'\delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}\right].$$

Eq. 45 shows that determining $S_{map}(\mathbf{R})$ for a given macrostate $\mathbf{R}$ involves a comparison of the energies of all pairs of microscopic configurations that map onto it. A further identity $1 = \int dU'\delta(u(\mathbf{r}') - U')$ fixing the energy of configuration $\mathbf{r}'$ can be inserted in the logarithm of Eq. 45 to switch from a configurational to an energetic integral. This provides:

$$\ln\left[\frac{\int d\mathbf{r}' \exp[-\beta(u(\mathbf{r}') - u(\mathbf{r}))]\delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}{\int d\mathbf{r}'\delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})}\right] =$$
$$\ln \int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - u(\mathbf{r}))], \tag{46}$$

where

$$P(U'|\mathbf{R}) = \frac{\int d\mathbf{r}'\delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})\delta(u(\mathbf{r}') - U')}{\int d\mathbf{r}'\delta(\mathbf{M}(\mathbf{r}') - \mathbf{R})} \tag{47}$$

is the microcanonical (unweighted) conditional probability of possessing energy U' given that the CG macrostate is $\mathbf{R}$. It is possible to write it as $\Omega_1(U', \mathbf{R})/\Omega_1(\mathbf{R})$, that is, the multiplicity of AA configurations such that $\mathbf{M}(\mathbf{r}) = \mathbf{R}$ and $u(\mathbf{r}') = U'$ normalised by the multiplicity of configurations that map to $\mathbf{R}$.

A second identity $1 = \int dU \delta(u(\mathbf{r}) - U)$ on the energies provides the following expression for $S_{map}(\mathbf{R})$:

$$\begin{aligned}
&S_{map}(\mathbf{R}) = -k_B \int d\mathbf{r} \frac{p_r(\mathbf{r})}{p_R(\mathbf{R})} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \times \\
&\ln \left[\int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - u(\mathbf{r}))] \right] \\
&= -k_B \int dU \ln \left[\int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - U)] \right] \times \\
&\int d\mathbf{r} \frac{p_r(\mathbf{r})}{p_R(\mathbf{R})} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \delta(u(\mathbf{r}) - U). \qquad (48)
\end{aligned}$$

The last integral in Eq 48, which we dub $P_\beta(U|\mathbf{R})$,

$$P_\beta(U|\mathbf{R}) = \int d\mathbf{r} \frac{p_r(\mathbf{r})}{p_R(\mathbf{R})} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \delta(u(\mathbf{r}) - U) \qquad (49)$$

is now the canonical—i.e., Boltzmann-weighted—conditional probability of possessing energy U provided that $\mathbf{M}(\mathbf{r}) = \mathbf{R}$, namely $p_R(U, \mathbf{R})/p_R(\mathbf{R})$. One thus obtains:

$$\begin{aligned}
&S_{map}(\mathbf{R}) = -k_B \int dU P_\beta(U|\mathbf{R}) \times \qquad (50) \\
&\ln \left[\int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - U)] \right] \\
&= -k_B \ln \left[\int dU' P(U'|\mathbf{R}) \exp[-\beta(U' - \langle U \rangle_{\beta|\mathbf{R}})] \right],
\end{aligned}$$

where

$$\langle U \rangle_{\beta|\mathbf{R}} = \int dU P_\beta(U|\mathbf{R}) U \qquad (51)$$

is the canonical average of the microscopic potential energy over the CG macrostate $\mathbf{R}$.

A direct calculation of $S_{map}(\mathbf{R})$ starting from the last line of Eq. 50 requires to perform an average over the microcanonical distribution $P(U'|\mathbf{R})$, which is not straightforwardly accessible in NVT simulations. However, there is a connection between $P(U|\mathbf{R})$ in Eq. 47 and $P_\beta(U|\mathbf{R})$ in Eq. 49: if one writes $p_R(\mathbf{R})$ as $\int dU' \exp[-\beta(U')] \Omega_1(U', \mathbf{R})$ and $p_R(U, \mathbf{R})$ as

$\exp[-\beta(U)]\Omega_1(U,\mathbf{R})$, standard reweighing provides

$$P(U|\mathbf{R}) = \frac{P_\beta(U|\mathbf{R})\exp[\beta U]}{\int dU' P_\beta(U'|\mathbf{R})\exp[\beta U']}. \tag{52}$$

Eq. 52 enables one to convert the microcanonical average in Eq. 50 to a canonical one, so that

$$S_{map}(\mathbf{R}) = k_B \ln\left[\int dU' P_\beta(U'|\mathbf{R})\; e^{\beta(U'-\langle U\rangle_{\beta|\mathbf{R}})}\right]. \tag{53}$$

Finally, by means of a second order cumulant expansion of Eq. 12 one obtains

$$S_{map}(\mathbf{R}) \simeq k_B \frac{\beta^2}{2}\langle (U-\langle U\rangle_{\beta|\mathbf{R}})^2\rangle_{\beta|\mathbf{R}}, \tag{54}$$

that inserted in Eq. 44 results in a *total* mapping entropy given by Eq. 15.

References

(1) Car, R.; Parrinello, M. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

(2) Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics* **1959**, *31*, 459–466.

(3) Karplus, M. Molecular Dynamics Simulations of Biomolecules. *Accounts of Chemical Research* **2002**, *35*, 321–323.

(4) Takada, S. Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol.* **2012**, *22*, 130–137.

(5) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics* **2013**, *139*, 090901.

(6) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **2013**, *42*, 73–93.

(7) Potestio, R.; Peter, C.; Kremer, K. Computer Simulations of Soft Matter: Linking the Scales. *Entropy* **2014**, *16*, 4199–4245.

(8) D'Adamo, G.; Menichetti, R.; Pelissetto, A.; Pierleoni, C. Coarse-graining polymer solutions: A critical appraisal of single-and multi-site models. *The European Physical Journal Special Topics* **2015**, *224*, 2239–2267.

(9) https://foldingathome.org.

(10) http://www.gpugrid.net.

(11) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. Millisecond-scale Molecular Dynamics Simulations on Anton. Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. New York, NY, USA, 2009; pp 65:1–65:11.

(12) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **2006**, *14*, 437–449.

(13) Bock, L. V.; Blau, C.; Schröder, G. F.; Davydov, I. I.; Fischer, N.; Stark, H.; Rodnina, M. V.; Vaiana, A. C.; Grubmüller, H. Energy barriers and driving forces in tRNA translocation through the ribosome. *Nat Struct Mol Biol* **2013**, *20*, 1390 – 1396.

(14) others,, et al. Atoms to Phenotypes: Molecular Design Principles of Cellular Energy Metabolism. *Cell* **2019**, *179*, 1098–1111.

(15) Noid, W. G. *Biomolecular Simulations*; Springer, 2013; pp 487–531.

(16) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of chemical physics* **2008**, *128*, 244114.

(17) Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.

(18) Lebold, K. M.; Noid, W. G. Dual approach for effective potentials that accurately model structure and energetics. *The Journal of Chemical Physics* **2019**, *150*, 234107.

(19) Lebold, K. M.; Noid, W. Dual-potential approach for coarse-grained implicit solvent models with accurate, internally consistent energetics and predictive transferability. *The Journal of Chemical Physics* **2019**, *151*, 164113.

(20) Jin, J.; Pak, A. J.; Voth, G. A. Understanding Missing Entropy in Coarse-Grained Systems: Addressing Issues of Representability and Transferability. *The Journal of Physical Chemistry Letters* **2019**, *10*, 4549–4557.

(21) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-grained protein models and their applications. *Chemical reviews* **2016**, *116*, 7898–7936.

(22) Diggins IV, P.; Liu, C.; Deserno, M.; Potestio, R. Optimal coarse-grained site selection in elastic network models of biomolecules. *Journal of chemical theory and computation* **2018**, *15*, 648–664.

(23) Khot, A.; Shiring, S. B.; Savoie, B. M. Evidence of information limitations in coarse-grained models. *The Journal of Chemical Physics* **2019**, *151*, 244105.

(24) Golhlke, H.; Thorpe, M. F. A natural coarse graining for simulating large biomolecular motion. *Biophysical Journal* **2006**, *91*, 2115–2120.

(25) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophysical Journal* **2008**, *95*, 5073 – 5083.

(26) Zhang, Z.; Pfaendtner, J.; Grafmller, A.; Voth, G. A. Defining Coarse-Grained Representations of Large Biomolecules and Biomolecular Complexes from Elastic Network Models. *Biophysical Journal* **2009**, *97*, 2327 – 2337.

(27) Potestio, R.; Pontiggia, F.; Micheletti, C. Coarse-grained description of proteins' internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. *Biophys J* **2009**, *96*, 4993–5002.

(28) Aleksiev, T.; Potestio, R.; Pontiggia, F.; Cozzini, S.; Micheletti, C. PiSQRD: a web server for decomposing proteins into quasi-rigid dynamical domains. *Bioinformatics* **2009**, *25*, 2743–2744.

(29) Zhang, Z.; Voth, G. A. Coarse-Grained Representations of Large Biomolecular Complexes from Low-Resolution Structural Data. *Journal of Chemical Theory and Computation* **2010**, *6*, 2990–3002.

(30) Sinitskiy, A. V.; Saunders, M. G.; Voth, G. A. Optimal number of coarse-grained sites in different components of large biomolecular complexes. *The Journal of Physical Chemistry B* **2012**, *116*, 8363–8374.

(31) Polles, G.; Indelicato, G.; Potestio, R.; Cermelli, P.; Twarock, R.; Micheletti, C. Mechanical and assembly units of viral capsids identified via quasi-rigid domain decomposition. *PLoS computational biology* **2013**, *9*, 1–13.

(32) Webb, M. A.; Delannoy, J.-Y.; de Pablo, J. J. Graph-Based Approach to Systematic Molecular Coarse-Graining. *Journal of Chemical Theory and Computation* **2019**, *15*, 1199–1208.

(33) Ponzoni, L.; Polles, G.; Carnevale, V.; Micheletti, C. SPECTRUS: A Dimensionality Reduction Approach for Identifying Dynamical Domains in Protein Complexes from Limited Structural Datasets. *Structure* **2015**, *23*, 1516 – 1525.

(34) Li, Z.; Wellawatte, G. P.; Chakraborty, M.; Gandhi, H. A.; Xu, C.; White, A. D. Graph neural network based coarse-grained mapping prediction. *Chemical Science* **2020**,

(35) Koehl, P.; Poitevin, F.; Navaza, R.; Delarue, M. The renormalization group and its applications to generating coarse-grained models of large biological molecular systems. *Journal of chemical theory and computation* **2017**, *13*, 1424–1438.

(36) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Paramter Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.

(37) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **1997**, *2*, 173–181.

(38) Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins* **1998**, *33*, 417–429.

(39) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anistropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80*, 505–515.

(40) Delarue, M.; Sanejouand, Y. H. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J Mol Biol* **2002**, *320*, 1011–1024.

(41) Micheletti, C.; Carloni, P.; Maritan, A. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. *Proteins: Structure, Function, and Bioinformatics* **2004**, *55*, 635–645.

(42) Rudzinski, J. F.; Noid, W. G. Coarse-graining entropy, forces, and structures. *The Journal of Chemical Physics* **2011**, *135*, 214101.

(43) Shell, M. S. Systematic coarse-graining of potential energy landscapes and dynamics in liquids. *J. Chem. Phys.* **2012**, *137*, 084503.

(44) Foley, T. T.; Shell, M. S.; Noid, W. G. The impact of resolution upon entropy and information in coarse-grained models. *The Journal of chemical physics* **2015**, *143*, 243104.

(45) Kullback, S.; Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics* **1951**, *22*, 79–86.

(46) Fisher, M. E. Renormalization group theory: Its basis and formulation in statistical physics. *Reviews of Modern Physics* **1998**, *70*, 653.

(47) Shannon, C. E. A mathematical theory of communication. *Bell system technical journal* **1948**, *27*, 379–423.

(48) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynskis equality. *The Journal of chemical physics* **2003**, *119*, 3559–3566.

(49) Chipot, C.; Pohorille, A. *Free energy calculations*; Springer, 2007.

(50) Mayorga-Flores, M.; Chantôme, A.; Melchor-Meneses, C. M.; Domingo, I.; Titaux-Delgado, G. A.; Galindo-Murillo, R.; Vandier, C.; del Río-Portilla, F. Novel blocker of onco SK3 channels derived from scorpion toxin tamapin and active against migration of cancer cells. *ACS Medicinal Chemistry Letters* **2020**, *11*, 1627–1633.

(51) Pedarzani, P.; D'hoedt, D.; Doorty, K. B.; Wadsworth, J. D. F.; Joseph, J. S.; Jeyaseelan, K.; Kini, R. M.; Gadre, S. V.; Sapatnekar, S. M.; Stocker, M.; Strong, P. N. Tamapin, a Venom Peptide from the Indian Red Scorpion (Mesobuthus tamulus)

That Targets Small Conductance Ca2+-activated K+ Channels and Afterhyperpolarization Currents in Central Neurons. *Journal of Biological Chemistry* **2002**, *277*, 46101–46109.

(52) Gati, C. D.; Mortari, M. R.; Schwartz, E. F. Towards therapeutic applications of arthropod venom K+-channel blockers in CNS neurologic diseases involving memory acquisition and storage. *Journal of toxicology* **2012**, *2012*, 756358.

(53) Müller, C. W.; Schlauderer, G. J.; Reinstein, J.; Schulz, G. E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **1996**, *4*, 147–56.

(54) Shapiro, Y. E.; Kahana, E.; Meirovitch, E. Domain mobility in proteins from NMR/SRLS. *The Journal of Physical Chemistry B* **2009**, *113*, 12050–12060.

(55) Formoso, E.; Limongelli, V.; Parrinello, M. Energetics and Structural Characterization of the large-scale Functional Motion of Adenylate Kinase. *Scientific reports* **2015**, *5*, 8425.

(56) Whitford, P. C.; Miyashita, O.; Levy, Y.; Onuchic, J. N. Conformational transitions of adenylate kinase: switching by cracking. *Journal of molecular biology* **2007**, *366*, 1661–1671.

(57) Wang, J.; Peng, C.; Yu, Y.; Chen, Z.; Xu, Z.; Cai, T.; Shao, Q.; Shi, J.; Zhu, W. Exploring Conformational Change of Adenylate Kinase by Replica Exchange Molecular Dynamic Simulation. *Biophysical Journal* **2020**, *118*, 1009–1018.

(58) Seyler, S. L.; Beckstein, O. Sampling large conformational transitions: adenylate kinase as a testing ground. *Molecular Simulation* **2014**, *40*, 855–877.

(59) Scott, C. F.; Carrell, R. W.; Glaser, C. B.; Kueppers, F.; Lewis, J. H.; Colman, R. W.

Alpha-1-antitrypsin-Pittsburgh. A potent inhibitor of human plasma factor XIa, kallikrein, and factor XIIf. *The Journal of clinical investigation* **1986**, *77*, 631–634.

(60) Nukiwa, T.; Brantly, M. L.; Ogushi, F.; Fells, G. A.; Crystal, R. G. Characterization of the gene and protein of the common alpha 1-antitrypsin normal M2 allele. *American journal of human genetics* **1988**, *43*, 322–330.

(61) Luisetti, M.; Seersholm, N. α1-antitrypsin deficiency · 1: Epidemiology of α1-antitrypsin deficiency. *Thorax* **2004**, *59*, 164–169.

(62) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.

(63) Černỳ, V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of optimization theory and applications* **1985**, *45*, 41–51.

(64) Kunzmann, P.; Hamacher, K. Biotite: a unifying open source computational biology framework in Python. *BMC bioinformatics* **2018**, *19*, 1–8.

(65) Andreotti, N.; di Luccio, E.; Sampieri, F.; Waard, M. D.; Sabatier, J.-M. Molecular modeling and docking simulations of scorpion toxins and related analogs on human SKCa2 and SKCa3 channels. *Peptides* **2005**, *26*, 1095 – 1108.

(66) Quintero-Hernndez, V.; Jimnez-Vargas, J.; Gurrola, G.; Valdivia, H.; Possani, L. Scorpion venom components that affect ion-channels function. *Toxicon* **2013**, *76*, 328 – 342.

(67) Ramrez-Cordero, B.; Toledano, Y.; Cano-Snchez, P.; Hernndez-Lpez, R.; Flores-Solis, D.; Saucedo-Yez, A. L.; Chvez-Uribe, I.; Brieba, L. G.; del Ro-Portilla, F. Cytotoxicity of Recombinant Tamapin and Related Toxin-Like Peptides on Model Cell Lines. *Chemical Research in Toxicology* **2014**, *27*, 960–967.

(68) Thach, T. T.; Luong, T. T.; Lee, S.; Rhee, D.-K. Adenylate kinase from Streptococcus pneumoniae is essential for growth through its catalytic activity. *FEBS open bio* **2014**, *4*, 672–682.

(69) Bellinzoni, M.; Haouz, A.; Graña, M.; Munier-Lehmann, H.; Shepard, W.; Alzari, P. M. The crystal structure of Mycobacterium tuberculosis adenylate kinase in complex with two molecules of ADP and Mg2+ supports an associative mechanism for phosphoryl transfer. *Protein science* **2006**, *15*, 1489–1493.

(70) Reinstein, J.; Gilles, A.-M.; Rose, T.; Wittinghofer, A.; Saint Girons, I.; Bârzu, O.; Surewicz, W. K.; Mantsch, H. H. Structural and catalytic role of arginine 88 in Escherichia coli adenylate kinase as evidenced by chemical modification and site-directed mutagenesis. *Journal of Biological Chemistry* **1989**, *264*, 8107–8112.

(71) Akbari, A. Phenylglyoxal. *Synlett* **2012**, *23*, 951–952.

(72) Matsunaga, Y.; Fujisaki, H.; Terada, T.; Furuta, T.; Moritsugu, K.; Kidera, A. Minimum Free Energy Path of Ligand-Induced Transition in Adenylate Kinase. *PLOS Computational Biology* **2012**, *8*, 1–12.

(73) Gur, M.; Madura, J. D.; Bahar, I. Global transitions of proteins explored by a multiscale hybrid methodology: application to adenylate kinase. *Biophysical journal* **2013**, *105*, 1643–1652.

(74) Halder, R.; Manna, R. N.; Chakraborty, S.; Jana, B. Modulation of the Conformational Dynamics of Apo-Adenylate Kinase through a π–Cation Interaction. *The Journal of Physical Chemistry B* **2017**, *121*, 5699–5708.

(75) Schapira, M.; Ramus, M.-A.; Jallat, S.; Carvallo, D.; Courtney, M. Recombinant alpha 1-antitrypsin Pittsburgh (Met 358—Arg) is a potent inhibitor of plasma kallikrein and activated factor XII fragment. *The Journal of clinical investigation* **1986**, *77*, 635–637.

(76) Taggart, C.; Cervantes-Laurean, D.; Kim, G.; McElvaney, N. G.; Wehr, N.; Moss, J.; Levine, R. L. Oxidation of either Methionine 351 or Methionine 358 in α1-Antitrypsin Causes Loss of Anti-neutrophil Elastase Activity. *Journal of Biological Chemistry* **2000**, *275*, 27258–27265.

(77) Owen, M. C.; Brennan, S. O.; Lewis, J. H.; Carrell, R. W. Mutation of Antitrypsin to Antithrombin. *New England Journal of Medicine* **1983**, *309*, 694–698.

(78) Kadanoff, L. P. Scaling laws for Ising models near T_c. *Physics Physique Fizika* **1966**, *2*, 263.

(79) Ma, S.-K. *Modern theory of critical phenomena*; Routledge, 2018.

(80) Zinn-Justin, J. *Phase transitions and renormalization group*; Oxford University Press on Demand, 2007.

(81) Schäfer, L. *Excluded volume effects in polymer solutions: as explained by the renormalization group*; Springer Science & Business Media, 2012.

(82) Cavagna, A.; Di Carlo, L.; Giardina, I.; Grandinetti, L.; Grigera, T. S.; Pisegna, G. Dynamical Renormalization Group Approach to the Collective Behavior of Swarms. *Physical Review Letters* **2019**, *123*, 268001.

(83) Antonov, N. V.; Kakin, P. I. Scaling in landscape erosion: Renormalization group analysis of a model with infinitely many couplings. *Theoretical and Mathematical Physics* **2017**, *190*, 193–203.

(84) Van Enter, A. C.; Fernández, R.; Sokal, A. D. Regularity properties and pathologies of position-space renormalization-group transformations: Scope and limitations of Gibbsian theory. *Journal of Statistical Physics* **1993**, *72*, 879–1167.

(85) Chakraborty, M.; Xu, C.; White, A. D. Encoding and selecting coarse-grain mapping operators with hierarchical graphs. *The Journal of Chemical Physics* **2018**, *149*, 134106.

(86) Wang, W.; Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials* **2019**, *5*, 125.

(87) Lenggenhager, P. M.; Gökmen, D. E.; Ringel, Z.; Huber, S. D.; Koch-Janusz, M. Optimal Renormalization Group Transformation from Information Theory. *Phys. Rev. X* **2020**, *10*, 011037.

(88) Ghersi, D.; Sanchez, R. EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **2009**, *25*, 3185–3186.

(89) Ngan, C.-H.; Hall, D. R.; Zerbe, B.; Grove, L. E.; Kozakov, D.; Vajda, S. FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **2012**, *28*, 286–287.

(90) Kuzmanic, A.; Bowman, G. R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Accounts of Chemical Research* **2020**, *53*, 654–661.

(91) Brady, G. P.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *Journal of computer-aided molecular design* **2000**, *14*, 383–401.

(92) Laurie, A. T.; Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics* **2005**, *21*, 1908–1916.

(93) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling* **1997**, *15*, 359–363.

(94) Yang, J.; Roy, A.; Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research* **2012**, *41*, D1096–D1103.

(95) Yang, J.; Roy, A.; Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29*, 2588–2595.

(96) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *33*, 3036–3042.

(97) Song, K.; Liu, X.; Huang, W.; Lu, S.; Shen, Q.; Zhang, L.; Zhang, J. Improved Method for the Identification and Validation of Allosteric Sites. *Journal of Chemical Information and Modeling* **2017**, *57*, 2358–2363.

(98) Krivák, R.; Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics* **2018**, *10*, 39.

(99) Jendele, L.; Krivak, R.; Skoda, P.; Novotny, M.; Hoksza, D. PrankWeb: a web server for ligand binding site prediction and visualization. *Nucleic acids research* **2019**, *47*, W345–W349.

(100) Yang, J.; Shen, C.; Huang, N. Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. *Frontiers in Pharmacology* **2020**, *11*, 69.

(101) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics* **2009**, *10*, 168.

(102) Zhu, H.; Pisabarro, M. T. MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics* **2011**, *27*, 351–358.

(103) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics* **1993**, *17*, 412–425.

(104) Pontiggia, F.; Colombo, G.; Micheletti, C.; Orland, H. Anharmonicity and self-similarity of the free energy landscape of protein G. *Phys Rev Lett* **2007**, *98*, 048102–048102.

(105) Hensen, U.; Meyer, T.; Haas, J.; Rex, R.; Vriend, G.; Grubmüller, H. Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function. *PLOS ONE* **2012**, *7*, 1–16.

(106) Nussinov, R.; Wolynes, P. G. A second molecular biology revolution? The energy landscapes of biomolecular function. *Phys. Chem. Chem. Phys.* **2014**, *16*, 6321–6322.

(107) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chemical Reviews* **2016**, *116*, 6516–6551.

(108) Sokal, R. R.; Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull* **1958**, *28*, 14091438.

(109) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **1976**, *32*, 922–923.

(110) Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **1978**, *34*, 827–828.

(111) Kim, H.; Jang, C.; Yadav, D. K.; Kim, M.-h. The comparison of automated clustering algorithms for resampling representative conformer ensembles with RMSD matrix. *Journal of Cheminformatics* **2017**, *9*, 21.

(112) Fraccalvieri, D.; Pandini, A.; Stella, F.; Bonati, L. Conformational and functional

analysis of molecular dynamics trajectories by Self-Organising Maps. *BMC bioinformatics* **2011**, *12*, 158.

(113) Müllner, D. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378* **2011**,

(114) Bar-Joseph, Z.; Gifford, D. K.; Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **2001**, *17*, S22–S29.

(115) others,, et al. SciPy: Open source scientific tools for Python. 2001–; `http://www.scipy.org/`.

(116) Park, M.-W.; Kim, Y.-D. A systematic procedure for setting parameters in simulated annealing algorithms. *Computers & Operations Research* **1998**, *25*, 207 – 217.

(117) Connolly, D. An Improved Annealing Scheme for the QAP. *European Journal of Operational Research* **1990**, *46*, 93 – 100.

(118) Dijkstra, M.; van Roij, R.; Evans, R. Phase diagram of highly asymmetric binary hard-sphere mixtures. *Physical Review E* **1999**, *59*, 5744.

(119) Rudzinski, J. F.; Noid, W. G. A generalized-Yvon-Born-Green method for coarse-grained modeling. *The European Physical Journal Special Topics* **2015**, *224*, 2193–2216.

(120) Menichetti, R.; Pelissetto, A.; Randisi, F. Thermodynamics of star polymer solutions: A coarse-grained study. *The Journal of chemical physics* **2017**, *146*, 244908.