






# A journey through mapping space: characterising the statistical and metric properties of reduced representations of macromolecules

Roberto Menichetti<sup>1,2</sup> , Marco Giulini<sup>1,2</sup> , and Raffaello Potestio<sup>1,2,a</sup> 

<sup>1</sup> Physics Department, University of Trento, via Sommarive, 14, 38123 Trento, Italy

<sup>2</sup> INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, via Sommarive, 14, 38123 Trento, Italy

Received 15 June 2021 / Accepted 13 September 2021

© The Author(s) 2021

**Abstract.** A mapping of a macromolecule is a prescription to construct a simplified representation of the system in which only a subset of its constituent atoms is retained. As the specific choice of the mapping affects the analysis of all-atom simulations as well as the construction of coarse-grained models, the characterisation of the *mapping space* has recently attracted increasing attention. We here introduce a notion of scalar product and distance between reduced representations, which allows the study of the metric and topological properties of their space in a quantitative manner. Making use of a Wang–Landau enhanced sampling algorithm, we exhaustively explore such space, and examine the qualitative features of mappings in terms of their squared norm. A one-to-one correspondence with an interacting lattice gas on a finite volume leads to the emergence of discontinuous phase transitions in mapping space, which mark the boundaries between qualitatively different reduced representations of the same molecule.

## 1 Introduction

The research area of computational molecular biophysics has experienced, in the past few decades, impressive advancements in two complementary and strictly intertwined fields: on the one hand, the steadily growing and increasingly cheaper computational power has enabled the simulation of ever larger systems with atomistic resolution [1, 2]; on the other hand, there has been an explosion of diverse coarse-grained (CG) models [3–5], i.e. simpler representations of molecules in terms of relatively few sites interacting through effective potentials: these have filled several gaps between the length- and time-scales of interest and the current capability of all-atom methods to cover them. The scientific efforts making use of one or both these techniques have cracked several important problems open, ranging from protein folding to cell growth [6–8].

The development of a successful CG model is strongly dependent on the choice of the reduced representation, or CG mapping, and on the correct parametrization of the effective interactions [4, 8]. The latter challenge has received an enormous amount of attention, leading, e.g. in the case of proteins, to extremely accurate and sophisticated CG potentials such as OPEP [9, 10], AWSEM [11, 12] and UNRES [13, 14]. The former task

has been the object of a smaller number of works, however its centrality in and beyond the process of coarse graining has recently started to emerge [8, 15]; indeed, a deep relationship exists between the degrees of freedom one selects to *construct* a CG model of the system, and those one employs to *analyse* the system's behaviour from a more detailed representation.

On the one hand, high-resolution, fully atomistic models are necessarily required to let the properties and behaviour of complex biomolecular systems emerge; on the other hand, the interpretation and understanding of this behaviour requires a reduction of the mountain of *data* and its synthesis in a smaller amount of *information*. In a nutshell, while the generative process has to be high-resolution to be useful, its outcome has to be low-resolution to be intelligible. An intuitive example of this concept is given by the representation of a protein structure in terms of its  $C_\alpha$ 's, i.e. the alpha carbons of the backbone: this mapping is not only extensively employed in the development of CG models [16, 17] (that is, models in which the whole amino acid is represented as a single bead whose position coincides with that of the  $C_\alpha$ ), but it is also extremely common in the analysis of structures sampled in fully atomistic simulations [18, 19].

A few different strategies have been developed that aim at identifying the optimal CG mapping to describe a molecule, which differ most notably in the observable used to drive the optimisation. There exists a first class of algorithms that rely on a completely static, graph-based description of the system [20, 21], such as the

<sup>a</sup>e-mail: [raffaello.potestio@unitn.it](mailto:raffaello.potestio@unitn.it) (corresponding author)

recent one proposed by Webb et al. in which a bottom-up approach iteratively aggregates separate nodes of the molecular graph in CG sites [20]. A second group of approaches makes use of the dynamics of the system, obtained through models with more [22, 23] or less [24, 25] detailed force fields. For instance, a recent protocol proposed by us [23] revolves around the analysis of an all-atom molecular dynamics (MD) [26, 27] simulation trajectory of a protein in terms of a subset of the molecule's atoms; a physics-driven choice of the latter allows one to identify the one or few mappings that return the most parsimonious yet informative simplified description of the system.

Each of these methods can be the most appropriate to investigate specific properties of the system at hand; at the same time, the majority of them performs the search for solutions of an optimisation problem within the overwhelmingly large space of all possible CG representations that can be assigned to the system. As an exhaustive exploration of this *mapping space* is hardly feasible in practice, the outcome of such schemes is often a *pool* of optimal CG representations, that is, an ensemble of local minima in the—likely—rugged landscape of the cost function that defines the mapping optimisation procedure.

Given the complexity of this problem, several natural questions arise: how degenerate is the space of solutions? Are its elements all markedly distinct from one another, or do they only represent mostly neutral and equivalent variations of a “typical” optimal CG mapping? Furthermore, how are the solutions *distributed* across the space of possible mappings? How much do they differ from, e.g. randomly chosen CG representations?

To be in the position of answering these questions, the most basic instrument one needs is a *meter*, i.e. a tool to measure distances in mapping space and assess the degree of “similarity” among its elements in a quantitative manner. The definition of such metric should be independent of the choice of the function subsequently employed to quantify the *quality* of a given reduced representation, in the same manner, e.g. in which the Euclidean distance separating two particles is independent of the interaction potential acting between them.

The objective of the present work is, thus, to contribute a tool for the exploration and characterisation of the mapping space, its metric and topological properties, and the relations among its instances. To this end, we introduce a notion of scalar product, and consequently of norm and distance, between reduced representations of a system. We first make use of these instruments in the exploration of some basic, *bare* metric and topological properties of mappings of a single, static molecular structure, i.e. without reference to its interactions and/or conformational sampling, but rather solely considering the coordinates of its constituent atoms. This provides a notion of mapping distance based on purely geometric properties of the molecule. Through the application of an enhanced sampling algorithm, namely the Wang–Landau method [28, 29], we characterise this mapping space in its entirety, and asso-

ciate its properties to structural features of the underlying molecule. Furthermore, the isomorphism between the problem of exploring the possible mappings of a molecule and that of a lattice gas in a finite volume enables to highlight the emergence of first-order phase transitions in the latter, distinguishing CG representations with qualitatively different properties. We then investigate the topology of the mapping space making use of the distance between reduced representations, which enables a low-dimensional embedding that highlights its general features. This analysis is performed both in absence and in presence of a cost function, namely the mapping entropy [23, 24, 30–32], which gauges the quality of a given CG representation according to an information theoretical measure. Finally, we suggest a possible manner to extend the tools we introduced to characterise the mapping space in the static case so as to incorporate information about the reference system's exploration of conformational space and thus, indirectly, about its energetics as well.

The paper is organised as follows: in Sect. 2 we develop a scalar product between decimation mappings of a macromolecular structure in a static conformation, and derive from it a notion of norm and distance in the mapping space; in Sect. 3 we study CG representations in terms of the distribution of values of the squared norm for mappings having a given number of retained sites  $N$ , first through random sampling, then making use of the Wang–Landau enhanced sampling method; in Sect. 4 we exploit a duality between the problem of mappings of a macromolecule and that of an interacting lattice gas in a finite volume to investigate the properties the molecule's reduced representations; in Sect. 5 we discuss some topological features of the mapping space; in Sect. 6 we discuss an extension of the structure-based definition of the norm that includes information about the system's energetics; in Sect. 7 we sum up the results of this work and discuss its future perspectives.

## 2 Theory

The construction of a CG model for a macromolecular system starts with the selection of a *mapping*  $M$ , that is, the projection operator connecting a microscopic, detailed configuration  $\mathbf{r}_i$ ,  $i = 1, \dots, n$  to a low-resolution one  $\mathbf{R}_I$ ,  $I = 1, \dots, N < n$ ,

$$M = \{\mathbf{M}_I(\mathbf{r}), I = 1, \dots, N\},$$

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

where  $n$  and  $N$  are the number of atoms in the system and the number of effective interaction sites employed in its CG simplified picture, respectively. In Eq. 1, the weights  $c_{Ii}$  are positive, spatially homogeneous—i.e. independent of the configuration  $\mathbf{r}$ —and subject to the normalization condition  $\sum_{i=1}^n c_{Ii} = 1$  to preserve trans-

lational invariance [4]. While a particular choice of these coefficients corresponds to a specific CG representation of the system, by varying them, along with changing the degree of CG'ing  $N$ , one spans the mapping space  $\mathcal{M}$ , whose elements are all the possible low-resolution descriptions that can be assigned to a macromolecule.

In the perspective of quantitatively characterising the properties of such space, the cardinality of  $\mathcal{M}$  in the continuous definition presented in Eq. 1 makes its thorough exploration, although appealing, hard to handle in practice. In this work, we, thus, restrict our analysis to the *discrete* subspace of CG representations that can be obtained for a system through a *decimation* [33,34] of its microscopic degrees of freedom: a subset of  $N$  constituent atoms is retained while the remaining ones are neglected.

By selectively discarding a subset of the system's atoms, this structural simplification procedure can be applied to systems of arbitrary size or environments, e.g. molecules in solution, embedded in lipid membranes, or in the gas phase; the set of potentially retained atoms can be extended to the solvent as well, so as to expand the definition of the system under examination to include e.g. water molecules that form hydrogen bonds with the solute. Clearly, the filtering brought forward by the decimation mapping can be applied to a single configuration as well as to an ensemble of conformations sampled, e.g. in a molecular dynamics simulation. The outcome would consist, in the first case, in a single subset of coordinates, which thus preserves no information about the system other than the positions of the retained atoms; in the second case, the filtering would result in an ensemble of reduced representations whose distribution in configuration space reverberates the intra- and intermolecular interactions of the underlying high-resolution sample.

In this work, we will focus on a particular type of molecular structure, namely a protein, and restrict our analysis to decimation mappings in which retained atoms are only selected among those composing the molecule. Initially, calculations will be performed only considering the static, crystallographic configuration of the molecule; subsequently, its energetics will be explicitly accounted for in Sect. 6, where we provide some preliminary results on the extension of the theory to an ensemble of conformations generated through a molecular dynamics simulation of the protein in explicit solvent. Also in this latter case, the environment is not explicitly retained by the CG mapping—again, only protein degrees of freedom are considered—but its effects are implicitly encoded in the distributions of conformations sampled by the molecule.

Let us, thus, consider a protein composed by  $n$  constituent atoms; the number of representations  $\Omega_N$  that can be constructed by selecting  $N$  of them as effective CG sites is

$$\Omega_N = \frac{n!}{N!(n-N)!}, \tag{2}$$

so that the *total* number of possible decimation mappings  $\Omega$  reads

$$\Omega = \sum_{N=1}^n \Omega_N = \sum_{N=1}^n \frac{n!}{(n-N)!N!} = 2^n - 1, \tag{3}$$

which becomes prohibitively large as the size of the system increases. Consequently, in the following we only consider the heavy atoms of the molecule as candidate CG sites, indicating with  $\mathcal{M}$  the subspace of mappings obtained according to these prescriptions.

The investigation of the topological structure of  $\mathcal{M}$  calls for the introduction of a distance  $\mathcal{D}(M, M')$ ,  $M, M' \in \mathcal{M}$ , able to quantify the “separation” between pairs of points  $M$  and  $M'$  belonging to the space of decimation mappings, that is, pairs of CG representations employed to represent the system that differ in the choice of the retained atoms. Such distance must be equipped with all the associated metric properties, namely identity, symmetry, and triangle inequality.

To construct  $\mathcal{D}(M, M')$ , we consider a *static* configuration of the molecule, namely the crystallographic one, with (heavy) atoms located in positions  $\mathbf{r}_i$ ,  $i = 1, \dots, n$  and a set of selection operators  $\chi_{M,i}$ ,  $i = 1, \dots, n$  defining mapping  $M$ ,

$$\chi_{M,i} = \begin{cases} 1 & \text{if atom } i \text{ is retained,} \\ 0 & \text{if atom } i \text{ is not retained,} \end{cases} \tag{4}$$

$$\sum_{i=1}^n \chi_{M,i} = N(M), \tag{5}$$

where  $N(M)$  is the number of retained atoms in the mapping. Taking inspiration from the Smooth Overlap of Atomic Positions method (SOAP) developed by Csányi et al. [35,36], we associate with each  $M \in \mathcal{M}$  an element  $\phi_M(\mathbf{r})$  of the Hilbert space of square-integrable real functions  $L_2(\mathbb{R}^3)$  as

$$\phi_M(\mathbf{r}) = \sum_{i=1}^n \phi_{M,i}(\mathbf{r}) = \sum_{i=1}^n C e^{-(\mathbf{r}-\mathbf{r}_i)^2/2\sigma^2} \chi_{M,i}, \tag{6}$$

obtained by centering a three-dimensional Gaussian—whose normalization factor  $C$  will be fixed in the following—on the position of each atom of the macromolecule retained in the mapping.

The inner product  $\langle \phi_M, \phi_{M'} \rangle$  of  $L_2(\mathbb{R}^3)$  between two mappings  $M$  and  $M'$ ,

$$\langle \phi_M, \phi_{M'} \rangle = \int d\mathbf{r} \phi_M(\mathbf{r}) \phi_{M'}(\mathbf{r}), \tag{7}$$

induces a norm  $\|\phi_M\|$  for mapping  $M$ , with

$$\mathcal{E}(M) = \|\phi_M\|^2 = \langle \phi_M, \phi_M \rangle, \tag{8}$$

starting from which the distance  $\mathcal{D}(M, M')$  can be defined as

$$\begin{aligned} \mathcal{D}(M, M') &= \|\phi_M - \phi_{M'}\| \\ &= \langle \phi_M - \phi_{M'}, \phi_M - \phi_{M'} \rangle^{\frac{1}{2}}, \end{aligned} \tag{9}$$

$\mathcal{D}(M, M')$  satisfying all the aforementioned metric properties.<sup>1</sup>

By inserting Eq. 6 in Eq. 7, the inner product  $\langle \phi_M, \phi_{M'} \rangle$  between mappings generated by two distinct selection operators  $\chi_M$  and  $\chi_{M'}$  becomes

$$\langle \phi_M, \phi_{M'} \rangle = \sum_{i,j=1}^n J_{ij} \chi_{M,i} \chi_{M',j}, \quad (10)$$

while the associated distance  $\mathcal{D}(M, M')$  in Eq. 9 reads

$$\begin{aligned} \mathcal{D}(M, M') &= (\mathcal{E}(M) + \mathcal{E}(M') - 2\langle \phi_M, \phi_{M'} \rangle)^{\frac{1}{2}} \\ &= \left( \sum_{i,j=1}^n J_{ij} \chi_{M,i} \chi_{M,j} + \sum_{i,j=1}^n J_{ij} \chi_{M',i} \chi_{M',j} + \right. \\ &\quad \left. - 2 \sum_{i,j=1}^n J_{ij} \chi_{M,i} \chi_{M',j} \right)^{\frac{1}{2}}. \end{aligned} \quad (11)$$

In Eqs. 10 and 11, the coupling constant  $J_{ij} = J_{ij}(\mathbf{r}_i, \mathbf{r}_j)$  between two atoms  $i$  and  $j$  is given by

$$J_{ij}(\mathbf{r}_i, \mathbf{r}_j) = C^2 \int d\mathbf{r} e^{-[(\mathbf{r}-\mathbf{r}_i)^2 + (\mathbf{r}-\mathbf{r}_j)^2]/2\sigma^2}, \quad (12)$$

with

$$J_{ij}(\mathbf{r}_i, \mathbf{r}_j) = J_{ij}(|\mathbf{r}_i - \mathbf{r}_j|) = J_{ij}(r_{ij}). \quad (13)$$

due to translational and rotational invariance. By introducing polar coordinates in Eq. 12, one has

$$\begin{aligned} J_{ij}(r_{ij}) &= 2\pi C^2 \int dr d\theta r^2 \sin \theta e^{-\frac{1}{2\sigma^2}(2r^2 + r_{ij}^2 - 2rr_{ij} \cos \theta)} \\ &= \frac{4\pi\sigma^2}{r_{ij}} C^2 e^{-r_{ij}^2/2\sigma^2} \int dr r e^{-r^2/\sigma^2} \sinh\left(\frac{rr_{ij}}{\sigma^2}\right), \end{aligned} \quad (14)$$

and a chain of Gaussian integrals provides

$$J_{ij}(r_{ij}) = \pi^{3/2} C^2 \sigma^3 e^{-r_{ij}^2/4\sigma^2} = e^{-r_{ij}^2/4\sigma^2}, \quad (15)$$

<sup>1</sup> In contrast to the original definition of the SOAP measure—which enables to quantify the similarity between two molecular structures [35, 36]—we here aim at determining the overlap between different CG representations of a *single* compound. As such, with respect to SOAP: (i) in Eq. 6 we do not employ local densities representing the chemical environment of a *specific* atom (which would afterwards require, e.g. to average over all pairs of atoms for the calculation of the total similarity kernel [36]), but rather global ones associated to the molecule as a whole; and (ii) in Eq. 7 we do not introduce an additional integral over rotations of one of the two structures. Indeed, there is no ambiguity in defining the alignment of different CG representations, as this is dictated by the original, full-atom reference.

where the last equality has been obtained by setting, without loss of generality,

$$C^2 = \frac{1}{\pi^{3/2} \sigma^3}. \quad (16)$$

Finally, by combining Eqs. 10 and 15 the inner product  $\langle \phi_M, \phi_{M'} \rangle$  reads

$$\langle \phi_M, \phi_{M'} \rangle = \sum_{i,j=1}^n e^{-r_{ij}^2/4\sigma^2} \chi_{M,i} \chi_{M',j}, \quad (17)$$

i.e. a sum of Gaussian factors over the positions of all pairs of atoms retained in the two mappings. Notably, the factorization with respect to the operators  $\chi_M$  and  $\chi_{M'}$  in Eqs. 10 and 17 enables the inner product (and therefore the distance  $\mathcal{D}$  and the squared norm  $\mathcal{E}$ ) to be determined starting from a matrix  $J_{ij}$  that can be calculated *a priori* over the static structure of the molecule.

One might ask what kind of information the previously defined quantities provide about the possible CG representations of a system. To answer this question, we first focus on the squared norm of a mapping  $\mathcal{E}(M)$ , see Eqs. 8 and 17,

$$\mathcal{E}(M) = \langle \phi_M, \phi_M \rangle = \sum_{i,j=1}^n e^{-r_{ij}^2/4\sigma^2} \chi_{M,i} \chi_{M,j}. \quad (18)$$

For a given retained atom  $i$ , the sum over  $j$  in Eq. 18,

$$Z_i(M) = \sum_{j=1}^n e^{-r_{ij}^2/4\sigma^2} \chi_{M,j}, \quad (19)$$

approximately represents its CG coordination number, that is, the number of retained atoms in the mapping that are located within a sphere of radius  $\sqrt{2}\sigma$  from  $i$ . By fixing the degree of coarse-graining  $N$ ,  $\mathcal{E}(M)$  scales as

$$\mathcal{E}(M) = N \bar{Z}(M), \quad (20)$$

$$\bar{Z}(M) = \frac{1}{N} \sum_{i=1}^n Z_i(M) \chi_{M,i} \quad (21)$$

showing that the dependence of the norm on the specific selection of atoms is dictated by  $\bar{Z}(M)$ , the *average* CG coordination number. Let us now consider two limiting cases: (i) extremely sparse and homogeneous CG representations, in which each retained atom does not have any retained neighbour within a radius of order  $\sqrt{2}\sigma$ —this condition can only be fulfilled provided that  $N$  is not too large, *vide infra*, or  $\sigma$  is much smaller than the typical interatomic distance. In this case, one has  $\bar{Z}(M) \approx 1$  and consequently  $\mathcal{E}(M) \approx N$ ; (ii) globular mappings characterised by densely populated (i.e. almost atomistic) regions of retained sites surrounded

by “empty” ones. In this case, the average coordination number  $\bar{Z}(M)$  will roughly resemble its atomistic counterpart, the latter being defined as

$$\bar{z} = \frac{1}{n} \sum_{i,j=1}^n e^{-r_{ij}^2/4\sigma^2}, \tag{22}$$

and thus  $\mathcal{E}(M) \approx N\bar{z}$ . It follows that the squared norm  $\mathcal{E}(M)$  captures the average homogeneity of a CG representation, that is, whether the associated retained atoms are uniformly distributed across the macromolecule or are mainly localized in well-defined regions of it. In Fig. 1, we report examples of CG mappings extracted for these two extreme categories in the case of adenylate kinase (see Sect. 3 for further details on this protein) together with a CG representation in which the retained atoms are randomly selected.

An analogous discussion can be performed for the inner product  $\langle \phi_M, \phi_{M'} \rangle$  in Eq. 17, calculated between two mappings  $M$  and  $M'$  that respectively retain  $N$  and  $N'$  atoms of the system. For a given atom  $i$  in mapping  $M$ ,

$$T_i(M') = \sum_{j=1}^n e^{-r_{ij}^2/4\sigma^2} \chi_{M',j}, \tag{23}$$

approximately counts the number of neighbours  $j$  in mapping  $M'$  located within a sphere of radius  $\sqrt{2}\sigma$  from  $i$ . The inner product scales as

$$\langle \phi_M, \phi_{M'} \rangle = N\bar{T}(M, M'), \tag{24}$$

$$\bar{T}(M, M') = \frac{1}{N} \sum_{i=1}^n T_i(M') \chi_{M,i}, \tag{25}$$

where  $\bar{T}(M, M')$  is again the *average* number of neighbours an atom in mapping  $M$  has that belong to mapping  $M'$ . Eqs. 23, 24 and 25 provide a very intuitive explanation of the orthogonality of mappings,  $\langle \phi_M, \phi_{M'} \rangle \approx 0$ : it is sufficient that each atom in mapping  $M$  does not have any neighbour in  $M'$  (and obviously vice-versa). As such, orthogonal mappings cover complementary regions of the system.

In general, the existence of an inner product enables the definition of an angle  $\theta_{M,M'}$  between mappings, whose cosine reads

$$\cos \theta_{M,M'} = \frac{\langle \phi_M, \phi_{M'} \rangle}{(\mathcal{E}(M)\mathcal{E}(M'))^{\frac{1}{2}}}. \tag{26}$$

While the orthogonality of mappings ( $\cos \theta_{M,M'} \approx 0$ ) has a relatively straightforward interpretation in terms of their spatial complementarity, the condition of parallelism,  $\cos \theta_{M,M'} \approx 1$ , is a bit less intuitive. If the mappings  $M$  and  $M'$  have the same number of atoms  $N$ , by inserting Eqs. 20 and 24 in Eq. 26 one obtains

$$\cos \theta_{M,M'} = \frac{\bar{T}(M, M')}{(\bar{Z}(M)\bar{Z}(M'))^{\frac{1}{2}}}. \tag{27}$$

If furthermore the two mappings show also roughly the same “globularity”,  $\bar{Z}(M) \approx \bar{Z}(M')$ , their parallelism requires

$$\bar{T}(M, M') \approx \bar{Z}(M), \tag{28}$$

that is, the average number of neighbors one atom of  $M$  has that belong to mapping  $M'$  has to be equal to the average number of neighbors the atom has that belong to  $M$  itself. This means that the two mappings must place retained atoms across the macromolecule in a similar fashion. Examples of approximately parallel and orthogonal CG representations for adenylate kinase are presented in Fig. 1.

It follows that while  $\mathcal{E}(M)$  quantifies the average sparseness of a CG representation,  $\langle \phi_M, \phi_{M'} \rangle$ —or equivalently  $\cos \theta_{M,M'}$ —characterises the average degree of spatial similarity between two different decimations of the microscopic degrees of freedom of the system. The distance  $\mathcal{D}(M, M')$  in Eq. 11 combines these two notions to extract how “far” a pair of CG representations is in the space of possible mappings  $\mathcal{M}$ .

Based on these observations, we implemented a slight modification to the inner product  $\langle \phi_M, \phi_{M'} \rangle$ —and hence to the squared norm  $\mathcal{E}(M)$  and distance  $\mathcal{D}(M, M')$ —with respect to the definition originally presented in Eq. 17, which, however, does not change its overall properties or interpretation. We have previously discussed how in the limiting cases of extremely sparse and globular mappings one respectively obtains  $\mathcal{E}(M) \approx N$  and  $\mathcal{E}(M) \approx N\bar{z}$ , where  $\bar{z}$  is the atomistic coordination number in Eq. 22. As the number of CG sites  $N$  increases, however, it will be extremely hard for a retained site not to have any retained neighbor within a sphere of radius of order  $\sigma$ , so that the exact scaling of  $\mathcal{E}(M)$  on the degree of CG’ing  $N$  in the case of sparse mappings will be hardly observed. We thus divide the inner product in Eq. 17 by the average atomistic coordination number, and define

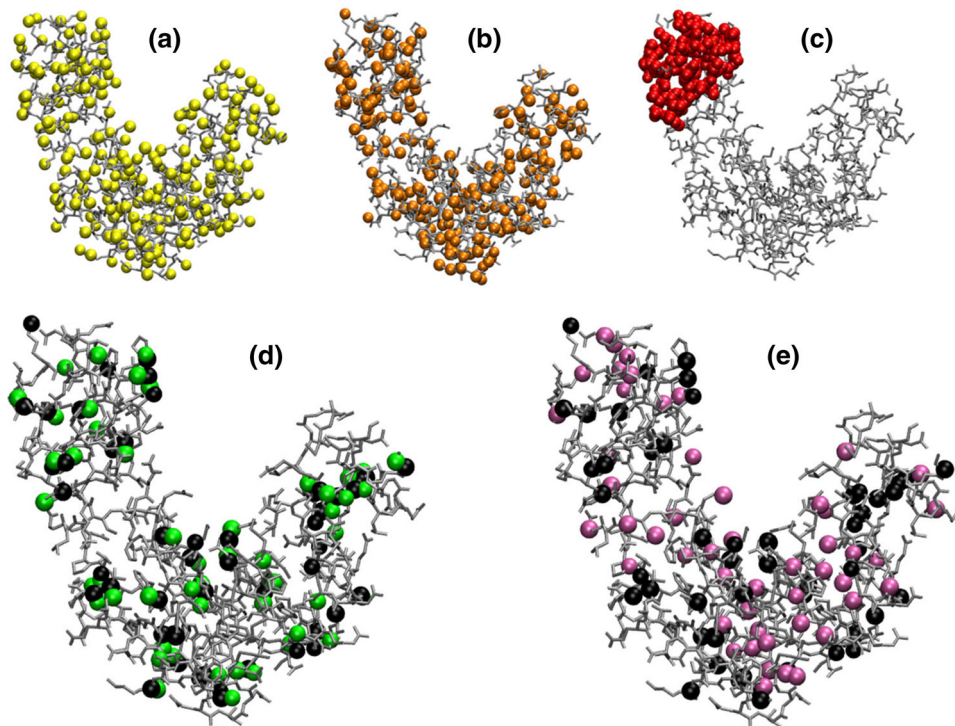
$$\langle \phi_M, \phi_{M'} \rangle_{\bar{z}} = \frac{1}{\bar{z}} \langle \phi_M, \phi_{M'} \rangle. \tag{29}$$

Consequently, one has

$$\mathcal{E}_{\bar{z}}(M) = \frac{1}{\bar{z}} \mathcal{E}(M), \tag{30}$$

$$\mathcal{D}_{\bar{z}}(M, M') = \frac{1}{\sqrt{\bar{z}}} \mathcal{D}(M, M'), \tag{31}$$

while the cosine between two mappings  $\cos \theta_{M,M'}$  is not affected by the rescaling. With this choice, globular mappings are now associated to  $\mathcal{E}(M)_{\bar{z}} \approx N$ , which can always be observed also in the case of low degrees of CG’ing, that is, high  $N$ . Note that the definition of  $\langle \phi_M, \phi_{M'} \rangle_{\bar{z}}$  in Eq. 29 corresponds to a rescaling of the coupling constant  $J_{ij}$  in Eq. 15 to



**Fig. 1** Top row: Example of possible CG representations for adenylate kinase with  $N = 214$  sites (represented as beads) characterised by a low (a), intermediate (b) and high (c) mapping squared norm  $\mathcal{E}$ . By increasing  $\mathcal{E}$  we move from maximally homogeneous to extremely globular CG representations. Bottom row: Examples of CG mappings with  $N = 53$  sites that are approximately parallel (d) and orthogonal (e) to a given one. The atoms composing the reference CG representation are represented as black beads. Parallel (resp. orthogonal) mappings tend to displace CG sites on similar (resp. complementary) regions of the system

$$J_{ij} = \frac{1}{\bar{z}} e^{-r_{ij}^2/4\sigma^2}. \quad (32)$$

For notational convenience, in the following, we will omit the subscript  $\bar{z}$  and refer to  $\mathcal{E}(M)_{\bar{z}}$ ,  $\langle \phi_M, \phi_{M'} \rangle_{\bar{z}}$  and  $\mathcal{D}_{\bar{z}}(M, M')$  as  $\mathcal{E}(M)$ ,  $\langle \phi_M, \phi_{M'} \rangle$  and  $\mathcal{D}(M, M')$ , respectively.

### 3 Exploration of the mapping space

Starting from the definitions introduced in Sect. 2, we now proceed to perform a quantitative analysis of the high-dimensional space  $\mathcal{M}$  of CG representations that can be constructed for a macromolecule through a decimation of its atomistic degrees of freedom. As a testbed system we consider *adenylate kinase* (AKE), a 214 residue-long phosphotransferase enzyme catalysing the interconversion between adenine nucleotides, namely adenine diphosphate (ADP), adenine monophosphate (AMP), and the adenine triphosphate complex (ATP) [37]. The structure of adenylate kinase can be divided in three main building blocks [38, 39], with the mobile LID and NMP domains exhibiting a conformational rearrangement around a hinge, the stable CORE domain, which results in an overall *open*  $\leftrightarrow$  *closed* transition of the enzyme [40, 41]. Our calculations require in input

only the value of the  $\sigma$  parameter and a static configuration  $\mathbf{r}_i$ ,  $i = 1, \dots, n$  of the system to determine the set of Gaussian couplings  $J_{ij}$  in Eq. 32. We here set  $\sigma = 1.9\text{\AA}$  (that is, half the separation between two consecutive  $\alpha$  carbons), and rely on the *open* crystal conformation of adenylate kinase (PDB code 4AKE), excluding from the analysis all hydrogens composing the biomolecule, resulting in a total of 1656 heavy atoms.

The investigation of the topological structure of the decimation mapping space of AKE calls for an extensive characterisation of the relational properties among its points, achievable by analysing the behaviour of the distance  $\mathcal{D}(M, M')$  over an ensemble of prototypical CG representations extracted from  $\mathcal{M}$ . The discussion carried out in Sect. 2, however, highlighted that  $\mathcal{D}(M, M')$  strictly depends on two factors: the globularity of each mapping—encoded in the squared norm  $\mathcal{E}(M)$ —and their mutual spatial complementarity—that is, the inner product  $\langle \phi_M, \phi_{M'} \rangle$  or equivalently the cosine  $\cos \theta_{M, M'}$ . It is then useful to first focus on these one- and two-“body” ingredients before combining them into the distance  $\mathcal{D}(M, M')$ . As such, in Sects. 3.1, 3.2 and 4 we will respectively discuss the behaviour of  $\mathcal{E}(M)$  and  $\cos \theta_{M, M'}$  across the mapping space of AKE; the analysis of the distance  $\mathcal{D}$ , and hence of the topology of  $\mathcal{M}$ , will be presented in Sect. 5.

### 3.1 Norm distributions

Let us first consider the squared norm  $\mathcal{E}(M)$  of a CG representation  $M$  defined in Eq. 30. As previously discussed, this quantity provides information about the spatial homogeneity of a mapping with a given degree of CG in  $N$ ; that is to say, it recapitulates how the retained atoms are distributed across the molecular structure, from uniformly scattered ( $\mathcal{E}(M) \approx N/\bar{z}$ ) to mainly concentrated in well-defined, almost atomistic domains emerging out of a severely CG'ed background ( $\mathcal{E}(M) \approx N$ ).

It is important to stress that mappings belonging to the two aforementioned extreme cases are routinely employed by the CG'ing community in the description of a biomolecular system. In proteins, examples from the homogeneous class include physically intuitive, residue-based CG representations of the molecule in terms of its  $\alpha$  carbons or backbone atoms [5,8]; homogeneity, on the other hand, is often abruptly broken in chemically informed, multiscale mappings, in which a higher level of detail, up to the atomistic one, is sharply localized on the biologically/chemically relevant regions of the system—e.g. the active sites of the protein—while the reminder is treated at extremely low resolution [8]. Furthermore, moving away from these limiting cases, an increasing attention is being posed in employing CG descriptions in which the level of detail is, although inhomogeneously, quasi-continuously modulated throughout the molecular structure [8].

Be they fully homogeneous, markedly inhomogeneous, or smoothly interpolating between these two classes, the CG representations that are usually adopted in the literature to simplify a biomolecule are often selected *a priori* by relying on general and intuitive criteria. Critically, such representations only constitute elements, isolated instances extracted from the high-dimensional mapping space  $\mathcal{M}$  of the system. One natural question follows: how representative are these “common” mappings of the diversity of the space  $\mathcal{M}$ ? In other words, how spatially homogeneous are the possible CG descriptions that can be designed for a macromolecule when no prior knowledge about its chemical structure or biological function is exploited to guide the mapping construction?

To answer this question, we start by introducing the number of mappings  $\Omega_N(\mathcal{E})$  that attain a particular value  $\mathcal{E}$  of the squared norm for a given number of CG sites  $N$ , which is given by:

$$\Omega_N(\mathcal{E}) = \sum_{M \in \mathcal{M}} \delta(N(M), N) \delta(\mathcal{E}(M), \mathcal{E}) \quad (33)$$

with

$$\sum_{M \in \mathcal{M}} \mathcal{O}(M) = \sum_{\chi_1=0,1} \dots \sum_{\chi_n=0,1} \mathcal{O}(\{\chi_i\}), \quad (34)$$

where  $\mathcal{O}$  is a generic observable that depends on the mapping through the operators  $\chi_i$ . Normalizing Eq. 33 by the total number of mappings with  $N$  sites,  $\Omega_N$ , we

define the *conditional probability* of having a mapping with norm  $\mathcal{E}$  given that the degree of coarse-graining is  $N$ , that is

$$P_N(\mathcal{E}) = \frac{\Omega_N(\mathcal{E})}{\Omega_N}, \quad (35)$$

which satisfies the normalization condition

$$\sum_{\mathcal{E}} P_N(\mathcal{E}) = 1 \quad (36)$$

regardless of the number of retained sites.  $P_N(\mathcal{E})$  can be rewritten as

$$P_N(\mathcal{E}) = \left( \frac{n!}{(n-N)!N!} \right)^{-1} \sum'_{M \in \mathcal{M}} \delta(\mathcal{E}(M), \mathcal{E}), \quad (37)$$

where the primed sum runs over all mappings with fixed resolution  $N$ , i.e. over all values of the set of operators  $\chi_i = 0, 1, i = 1, \dots, n$  satisfying

$$\sum_{i=1}^n \chi_i = N. \quad (38)$$

By providing direct insight on the degree of spatial uniformity characterising the ensemble of all possible CG descriptions of a macromolecular system,  $P_N(\mathcal{E})$  represents a first important ingredient in the investigation of the structure of the mapping space  $\mathcal{M}$ . We, thus, aimed at analysing the behaviour of the conditional probability  $P_N(\mathcal{E})$  across the decimation mapping space  $\mathcal{M}$  of AKE for a set of 16 values of  $N$  ranging from  $N = 53$  to 1605, see Table 1. However, even restricted to these cases, an exhaustive enumeration of all possible CG representations of the system is unfeasible in practice: for example, in the case of AKE ( $n = 1656$ ), roughly  $10^{276}$  possible CG representations can be constructed that describe the enzyme in terms of a subset of  $N = 214$  heavy atoms (one for each residue). This number grows to  $10^{496}$  for  $N = 856$  (four heavy atoms per residue on average), that is, close to the maximum of the binomial coefficient, obtained for  $N = n/2$ , see Eq. 2.

To overcome this combinatorial challenge, for each degree of CG'ing we generated  $\tilde{\Omega}_{\text{tot}} = 2 \times 10^6$  uniformly distributed random mappings as strings  $\chi_i, i = 1, \dots, n$  of zeros and ones compatible with Eq. 38, and calculated the associated squared norm  $\mathcal{E}$ . Results for each  $N$  were then binned along the  $\mathcal{E}$  axis in intervals of  $\delta\mathcal{E} = 0.1$ , and the corresponding  $P_N(\mathcal{E})$  was estimated as

$$P_N(\mathcal{E}) = \frac{1}{\delta\mathcal{E}} \frac{\tilde{\Omega}_N(\mathcal{E})}{\tilde{\Omega}_{\text{tot}}}, \quad (39)$$

where  $\tilde{\Omega}_N(\mathcal{E})$  is the number of sampled mappings with squared norm falling between  $\mathcal{E}$  and  $\mathcal{E} + \delta\mathcal{E}$ . Note that in this way we are approximately treating as continuous the intrinsically discrete, unevenly spaced spectrum

**Table 1** Average mapping squared norm  $\langle \mathcal{E} \rangle_N$  and associated standard deviation  $\sigma_{\mathcal{E},N}$  at different degrees of coarse-graining  $N$ , calculated over the mapping space  $\mathcal{M}$  of AKE. We present random sampling results (RS), as well as those obtained from a saddle-point approximation to the density of states  $\Omega_N(\mathcal{E})$  determined through the Wang–Landau method (WL-SP), see text

$N$	$\langle \mathcal{E} \rangle_N$		$\sigma_{\mathcal{E},N}$	
	RS	WL-SP	RS	WL-SP
53	5.41	–	0.31	–
107	14.15	–	0.63	–
214	41.14	40.82	1.32	1.32
321	80.95	–	2.03	–
428	133.58	133.17	2.74	2.74
535	199.04	–	3.45	–
642	277.33	276.93	4.12	4.11
749	368.44	–	4.74	–
856	472.39	471.95	5.29	5.29
963	589.16	–	5.74	–
1070	718.76	718.29	6.06	6.07
1177	861.18	–	6.22	–
1284	1016.43	1016.14	6.16	6.17
1391	1184.51	–	5.79	–
1498	1365.42	1365.05	4.94	4.94
1605	1559.15	–	3.09	–

of possible norms, and the density  $P_N(\mathcal{E})$ —and consequently  $\Omega_N(\mathcal{E})$ —as piecewise constant. In this “continuous” limit, the normalization condition of  $P_N(\mathcal{E})$  becomes

$$1 = \sum_{\mathcal{E}} P_N(\mathcal{E}) \delta \mathcal{E} \simeq \int d\mathcal{E} P_N(\mathcal{E}). \quad (40)$$

The set of distributions  $P_N(\mathcal{E})$  obtained from our random sampling of the mapping space of AKE are displayed in Fig. 2. We observe that, for each value of the CG resolution  $N$ ,  $P_N(\mathcal{E})$  is unimodal and narrowly peaked around its average squared norm,

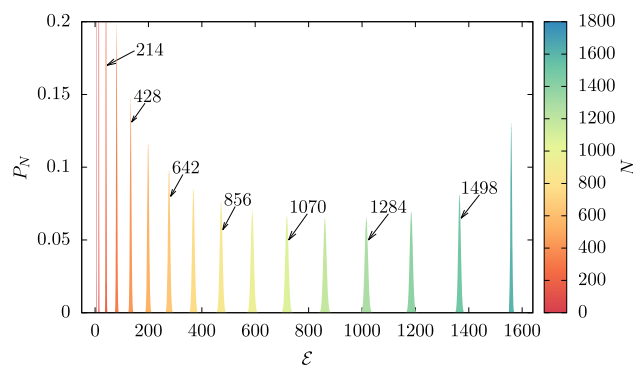
$$\langle \mathcal{E} \rangle_N = \int d\mathcal{E} P_N(\mathcal{E}) \mathcal{E}, \quad (41)$$

$\langle \mathcal{E} \rangle_N$  being an increasing function of  $N$ . On the other hand, the standard deviation  $\sigma_{\mathcal{E},N}$ ,

$$\sigma_{\mathcal{E},N} = \left( \int d\mathcal{E} P_N(\mathcal{E}) (\mathcal{E} - \langle \mathcal{E} \rangle_N)^2 \right)^{\frac{1}{2}}, \quad (42)$$

is non-monotonic in the degree of CG’ing: starting from extremely small values in the case of few retained atoms (e.g.  $N = 53, 107$  and  $214$ ),  $\sigma_{\mathcal{E},N}$  increases roughly up to  $N \approx 3n/4$  and then starts to decrease, reaching zero for  $N = n$ —in this case, only one possible mapping exists, namely the atomistic one. These features are further highlighted in Table 1 and Fig. 3, in which we report the dependence of  $\langle \mathcal{E} \rangle_N$  and  $\sigma_{\mathcal{E},N}$  on the degree of CG’ing  $N$  as obtained from the distributions  $P_N(\mathcal{E})$  in Fig. 2.

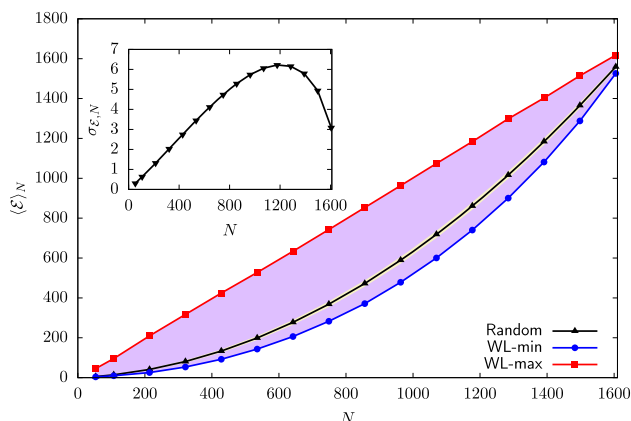
$\langle \mathcal{E} \rangle_N$  quantifies the average spatial homogeneity of the ensemble of CG representations that can be randomly assigned to AKE at a specific resolution. As pre-



**Fig. 2** Probability  $P_N(\mathcal{E})$  of the norm of the mapping  $\mathcal{E}$  for AKE calculated at various degrees of CG’ing  $N$ , as obtained from a random sampling of the mapping space  $\mathcal{M}$ . Arrows indicate the values of  $N$  for which a reconstruction of the density of states  $\Omega_N(\mathcal{E})$  through the Wang–Landau algorithm has been performed

viously discussed, maximally inhomogeneous mappings, in which a chiseled chunk of the biomolecule is treated atomistically while the remainder is almost neglected, are characterised by  $\mathcal{E} \approx N$ . Critically, Fig. 3 displays that such linear scaling lies always above the average  $\langle \mathcal{E} \rangle_N$  for all degrees of coarse-graining investigated. The deviation between the two curves is non-monotonic, with a maximum obtained for  $N = n/2$ , and only vanishes for  $N \rightarrow n$ , where mappings become very dense as they collapse towards the atomistic representation. As a consequence, the CG representations one encounters by randomly probing the mapping space  $\mathcal{M}$  tend to be “sparse” rather than compact. Furthermore, the difference between the squared norm of the globular case and  $\langle \mathcal{E} \rangle_N$  is always (but for  $N \approx n$ ) one or two orders of magnitudes larger than the standard deviation of the corresponding  $P_N(\mathcal{E})$ , see Fig. 3. It follows that inhomoge-





**Fig. 3** Inset: Standard deviation  $\sigma_{\mathcal{E},N}$  of the mapping norm  $\mathcal{E}$  as a function of the degree of CG'ing  $N$  obtained from a random sampling of the mapping space  $\mathcal{M}$  of AKE. Main plot:  $N$ -dependence of the average squared norm  $\langle \mathcal{E} \rangle_N$  (“Random”, black line) and associated  $3\sigma_{\mathcal{E},N}$  confidence interval (khaki area) as obtained from a random sampling of the mapping space of AKE, superimposed to the region covered by the set of single-window, preliminary WL runs (purple area). The minimum (“WL-min”, blue line) and maximum (“WL-max”, red line) squared norms reached by the preliminary runs are highlighted. “WL-max” also corresponds to the scaling  $\mathcal{E} \approx N$  obtained in the case of inhomogeneous, globular mappings

neous mappings lie extremely far away in the right tails of the distributions displayed in Fig. 2, thus constituting an exponentially vanishing subset of the space  $\mathcal{M}$ .

The suppression of the statistical weight associated with high-norm, globular CG representations of AKE in the space of all possible ones is not surprising, and is solely driven by entropic effects. Indeed, at least for small and intermediate  $N$ , it is extremely unlikely that a completely random selection of retained atoms across the biomolecule will result in their dense confinement within sharply defined spatial domains of the system, just as it is unlikely for a gas to occupy only a small fraction of the volume in which it is enclosed. Interestingly, this latter analogy can be pushed further by noting that the squared norm  $\mathcal{E}(M)$ , see Eqs. 30 and 18, is akin to the negative configurational energy of a lattice gas living on the irregular grid defined by the protein's conformation, whose particle interact via a hard-core, short-range potential followed by an attractive Gaussian tail. In this context, the selection operators  $\chi_{M,i} = 0, 1, i = 1, \dots, n$  of a mapping  $M$  with  $N$  retained atoms can be interpreted as the set of occupation numbers describing a distribution of the  $N$  particles of the gas on the  $n$  available lattice sites. It follows that compact CG representations of AKE, located in the large- $\mathcal{E}$  limit of  $P_N(\mathcal{E})$ , are just as challenging to randomly sample within the space  $\mathcal{M}$  as are the low-energy configurations of the gas in which the  $N$  particles spontaneously occupy only a fraction of the available volume. The implications of this analogy will be thoroughly explored in Sect. 4.

The strongly entropy-driven distribution of mappings calls for the introduction of enhanced sampling tech-

niques to boost the exploration of the mapping space; in this work, we resort to the algorithm proposed by Wang and Landau (WL) [28, 29, 42, 43]. For each CG resolution  $N$ , the aim is to obtain a *uniform* sampling of the possible mapping norms  $\mathcal{E}$  across the space  $\mathcal{M}$ , in contrast to the set of narrowly peaked probability distributions displayed in Fig. 2. In principle, this is attained by setting up a Markov chain Monte Carlo simulation in which a transition between two subsequent mappings  $M$  and  $M'$ —both retaining  $N$  atoms—is accepted with probability  $\alpha$  given by [28]

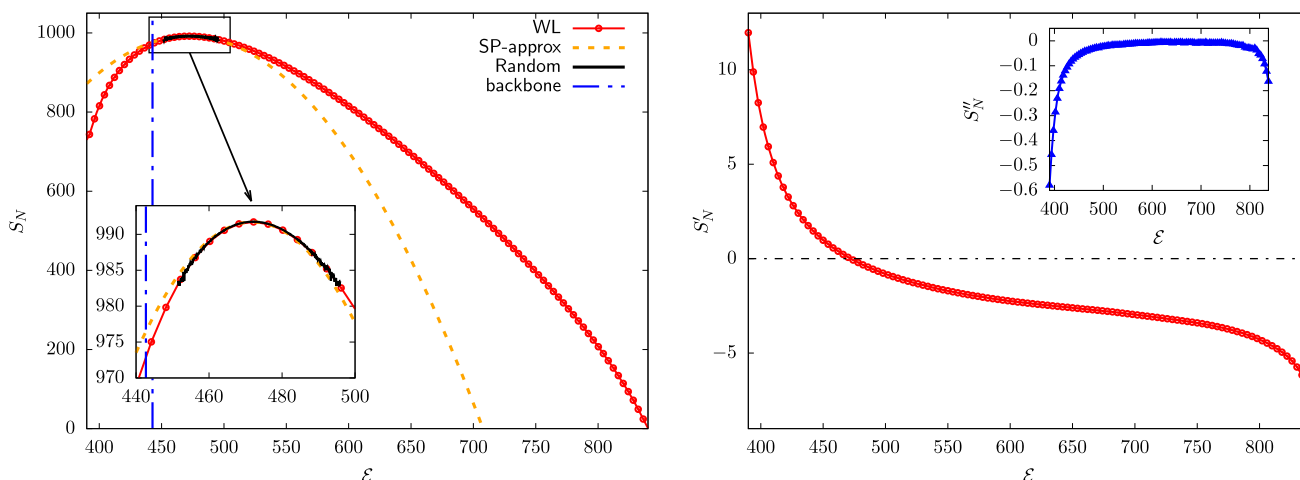
$$\alpha_{M \rightarrow M'} = \min \left[ 1, \frac{\Omega_N(\mathcal{E}(M))}{\Omega_N(\mathcal{E}(M'))} \right] = \min [1, \exp(-[S_N(\mathcal{E}(M')) - S_N(\mathcal{E}(M))])], \quad (43)$$

where  $\Omega_N(\mathcal{E})$  is the density of states defined in Eq. 35 while  $S_N(\mathcal{E}) = \ln[\Omega_N(\mathcal{E})]$  is the corresponding micro-canonical entropy. When compounded with a symmetric proposal probability  $\pi$  for the attempted move,  $\pi_{M \rightarrow M'} = \pi_{M' \rightarrow M}$ , the Markov chain in Eq. 43 would generate, after an initial relaxation transient, CG representations distributed according to  $p(M) \sim 1/\Omega_N(\mathcal{E}(M))$  [28], resulting in a flat histogram  $P_N(\mathcal{E})$  of visited norms *over the whole range of possible ones* [43].

In practice, however, the density of states in Eq. 43 is not known a priori. The power of WL approach resides in its ability to self-consistently obtain  $\Omega_N(\mathcal{E})$  through a sequence  $k = 1, \dots, K$  of non-equilibrium simulations in which increasingly accurate approximations  $\hat{\Omega}_N^k(\mathcal{E})$  to the exact result are generated, iterations being stopped when the desired precision is achieved [28, 29]. For the sake of brevity, we here omit an exhaustive discussion of the general algorithmic workflow behind WL sampling as well as an in-depth description of the specific implementation employed in this work; these details are provided in Appendix A.

In the WL reconstruction of a density of states such as  $\Omega_N(\mathcal{E})$ , knowledge of the sampling boundaries proves extremely beneficial to the accuracy and rate of convergence of the self-consistent scheme [44]. For each degree of CG'ing investigated, we, thus, initially performed a preliminary, non-iterative WL run to approximately locate the minimum and maximum mapping norms  $\mathcal{E}_{\min}(N)$  and  $\mathcal{E}_{\max}(N)$  achievable for AKE at that specific CG resolution, and consequently bound the support of the corresponding  $\Omega_N(\mathcal{E})$ .

The results for  $\mathcal{E}_{\min}(N)$  and  $\mathcal{E}_{\max}(N)$  obtained from this analysis are presented in Fig. 3 and Table 2 of Appendix A. We observe that the mapping norms visited by the set of preliminary WL runs extend, for all values of  $N$ , over a significantly wider range compared to the one obtained by random sampling. Remarkably, the maximum norm  $\mathcal{E}_{\max}(N)$  exhibits a linear dependence on  $N$  that is fully compatible with the one associated to globular CG representations,  $\mathcal{E}_{\max}(N) \approx N$ , highlighting that the WL approach succeeds in exploring this entropically suppressed region of the mapping space. Furthermore, Fig. 3 displays that the minimum



**Fig. 4** Left: Logarithm of the density of states  $\Omega_N(\mathcal{E})$  of AKE,  $S_N(\mathcal{E}) = \ln[\Omega_N(\mathcal{E})]$ , for  $N = 856$ . We report results obtained via (i) Wang–Landau sampling (“WL”, red dotted line), vertically shifting the data so that the minimum of  $S_N$  over the range of investigated norms is zero; (ii) a saddle-point approximation of the WL predictions (“SP-approx”, orange dashed line); and (iii) a random drawing of CG representations (“Random”, black line), in this latter case shifting the curve so that its maximum coincides with the one of the WL profile. In the figure we also report the squared norm associated to the mapping in which all the heavy atoms composing the backbone of AKE are retained (“backbone”, dashed blue line), a CG representation that is commonly employed when CG’ing a protein [5, 8]. Right: First (main plot) and second (inset) derivatives  $S'_N(\mathcal{E})$  and  $S''_N(\mathcal{E})$  of the entropy  $S_N(\mathcal{E})$  determined via WL sampling for  $N = 856$

norm  $\mathcal{E}_{\min}(N)$  identified by the preliminary runs lies always below the average  $\langle \mathcal{E} \rangle_N$  for all values of  $N$ . In contrast to globular mappings, CG representations living in this low  $\mathcal{E}$  limit are *maximally homogeneous*, that is, retained atoms are scattered throughout the molecular structure as uniformly as possible. This class constitutes another exponentially vanishing subset of the mapping space: in the gas picture, it would correspond to the ensemble of configurations in which gas particles are *regularly* distributed within the available volume.

Having approximately identified the range of mapping norms achievable for AKE at each CG resolution, we subsequently moved to the determination of the associated densities of states  $\Omega_N(\mathcal{E})$  via the iterative WL scheme, see Appendix A for all technical details. Calculations were only performed for a subset of degrees of CG’ing, namely those in which the number of retained atoms  $N$  is an integer multiple of the number of residues composing the biomolecule,  $N = i \cdot 214$ ,  $i = 1, \dots, 7$ .

To speed-up convergence of the algorithm, for each  $N$  we slightly reduced the range of norms  $[\mathcal{E}_{\min}, \mathcal{E}_{\max}]$  with respect to the one predicted by the explorative WL runs, see Table 2 in Appendix A. This interval was then divided into a set of overlapping windows in which independent WL simulations were performed [29]. The resulting partial densities of states were a posteriori combined to determine the cumulative  $\Omega_N(\mathcal{E})$  up to a global multiplicative factor, or, in our case, the entropy  $S_N(\mathcal{E}) = \ln[\Omega_N(\mathcal{E})]$  up to an additive constant.

WL estimates of the entropy  $S_N(\mathcal{E})$  are presented in Fig. 4 for  $N = 856$ , while results for all the other degrees of CG’ing are reported in Fig. 12 of Appendix A. In all cases, we observe that the behaviour of  $S_N$  is non-monotonic in  $\mathcal{E}$ , exhibiting a unique maximum as the

mapping norm moves from the left to right boundary of the range of investigated ones—that is, in transitioning from extremely homogeneous to maximally globular CG representations. As  $\Omega_N(\mathcal{E}) = \exp[S_N(\mathcal{E})]$ , this result confirms how these two limiting classes of mappings constitute regions of exponentially vanishing size within the broad space  $\mathcal{M}$ . At the same time, the overall shape of  $S_N$  strongly depends on the degree of CG’ing: while for high  $N$  entropy profiles are nearly symmetric around their maximum, they become increasingly skewed as fewer and fewer atoms are employed to represent the macromolecule, see Fig. 12. This asymmetry becomes apparent by performing, for each CG resolution, a quadratic expansion of  $S_N$  around its maximum,

$$S_N(\mathcal{E}) \simeq S_N(\tilde{\mathcal{E}}(N)) + \frac{1}{2} S''_N(\tilde{\mathcal{E}}(N)) (\mathcal{E} - \tilde{\mathcal{E}}(N))^2, \tag{44}$$

where  $\tilde{\mathcal{E}}(N)$  is the norm at which the first derivative  $S'_N$  of the entropy vanishes, and  $S''_N(\tilde{\mathcal{E}}(N))$  is the corresponding second derivative—the dependence of  $S'_N$  and  $S''_N$  on  $\mathcal{E}$  being displayed in Fig. 4 for  $N = 856$ . The accuracy of this parabolic, symmetric approximation in reproducing the exact  $S_N$  over the whole  $\mathcal{E}$ -range increases with the number of retained atoms, see Figs. 4 and 12, especially as far as the limit of high mapping norms is concerned.

Finally, it is interesting to test the predictions of WL sampling against the results obtained via a completely random exploration of the mapping space. To this end, Fig. 4 and Fig. 12 include a comparison between the WL entropies  $S_N$  and their random counterparts  $S_N^{ran}$ , the latter defined as  $S_N^{ran}(\mathcal{E}) = \ln[P_N(\mathcal{E})] + C_N$ , where  $P_N(\mathcal{E})$  are the probability densities presented in Fig. 2

and the constants  $C_N$  are set so that the maxima of  $S_N^{\text{ran}}$  and  $S_N$  coincide. For each value of  $N$  the two profiles are in perfect agreement, thus confirming the accuracy of the self-consistent WL scheme in determining the density of states of a system. Critically, results for  $S_N^{\text{ran}}$  only extend over a very narrow range of mapping norms, centred around the value  $\tilde{\mathcal{E}}(N)$  for which the maximum of the entropy is attained. It is, therefore, largely unfeasible, by randomly drawing CG representations, to exhaustively explore the mapping space  $\mathcal{M}$  of a macromolecule. In this respect it is worth to inspect the position, on the  $\mathcal{E}$  axis, of the  $C_\alpha$  and backbone mappings (which in AKE retain  $N = 214$  and  $N = 856$  sites, respectively), two reduced representations that are routinely employed for CG'ing proteins [5, 8]. These turn out to be located in the vicinity of the class of “prototypical” random ones, for which the entropy  $S_N$  reaches its maximum; however, their intrinsic regularity, dictated by the position of the retained sites on the peptide chain, makes these mappings slightly more homogeneous than the random ones, see Figs. 4 and 12.

To provide a more quantitative measure of the consistency between random and WL sampling results, for each degree of CG'ing, we recalculated the average and variance of the mapping norm, see Eqs. 41 and 42, starting from the WL entropies  $S_N$ . These are used to compute  $P_N(\mathcal{E})$  making use of a saddle-point approximation of Eq. 35, namely

$$P_N(\mathcal{E}) = \frac{\Omega_N(\mathcal{E})}{\Omega_N} = \frac{\exp[S_N(\mathcal{E})]}{\int d\mathcal{E} \exp[S_N(\mathcal{E})]} \simeq \left( \frac{|S_N''(\tilde{\mathcal{E}}(N))|}{2\pi} \right)^{\frac{1}{2}} \exp \left[ \frac{1}{2} S_N''(\tilde{\mathcal{E}}(N)) (\mathcal{E} - \tilde{\mathcal{E}}(N))^2 \right], \quad (45)$$

where in the last step of Eq. 45 we made use of the quadratic expansion of  $S_N$  defined in Eq. 44. Within the saddle point approximation, one has  $\langle \mathcal{E} \rangle_N = \tilde{\mathcal{E}}(N)$ ,  $\tilde{\mathcal{E}}(N)$  being the position of the maximum of  $S_N$ , and  $\sigma_{\mathcal{E},N} = |S_N''(\tilde{\mathcal{E}}(N))|^{-\frac{1}{2}}$ : these predictions are found to be in perfect agreement with their random sampling counterparts, results being presented in Table 1.

### 3.2 Inner product distributions

We now proceed to the description of the mapping space  $\mathcal{M}$  from the perspective of the inner product between its elements. Following the same scheme of Sect. 3.1, we here focus on the cosine between mappings that are constrained to share the same resolution  $N$ , and introduce the probability  $P_{NN}(\cos \theta)$  of observing a value of  $\cos \theta$  provided that this constraint is satisfied:

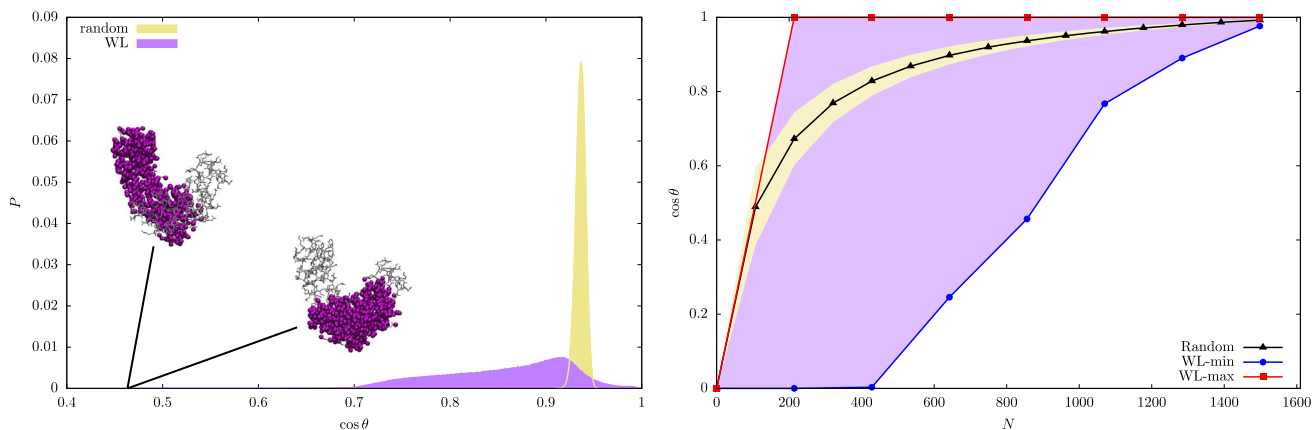
$$P_{NN}(\cos \theta) = \frac{\Omega_{NN}(\cos \theta)}{\Omega_N^2}, \quad (46)$$

that is, the ratio between the number of mapping pairs whose cosine is equal to  $\cos \theta$ ,  $\Omega_{NN}^2(\cos \theta)$ , and the total number of possible pairs  $\Omega_N^2$ . We can now investigate how the average *degree of parallelism* between two mappings changes when considering randomly selected mappings or more peculiar elements of  $\mathcal{M}$ .

In this section, we compare two data sets, each one containing  $10^6$  elements: the first one was obtained by computing the cosine between two mappings in which the retained sites were picked randomly; the second data set was instead constructed in a more sophisticated manner, making use of the WL sampling scheme to collect mappings that uniformly span the range  $[\mathcal{E}_{\min}, \mathcal{E}_{\max}]$  of accessible values of  $\mathcal{E}$  identified in the previous section. More specifically, we started a WL exploration as in Sect. 3.1 over this range and, when all the reference bins were visited at least once, we began saving a mapping every 1656 Monte Carlo moves. Mappings were saved in different macro-bins, each one covering an interval of amplitude 20 (in terms of units of  $\mathcal{E}$ ). Sampling ended when 5000 mappings were saved in each box, without considering the convergence of the WL algorithm. The data set was then generated by computing the cosine (Eq. 26) between randomly selected pairs of mappings extracted through this procedure. Importantly, the WL sampling scheme produces a pool of potentially correlated mappings, so that the chance of collecting similar elements of  $\mathcal{M}$  cannot be excluded.

The normalised histograms of  $\cos(\theta)$  values obtained from the two datasets are displayed in Fig. 5a for  $N = 856$ . We observe that while the random cosine distribution displays a narrow peak around its average the WL histogram is more smeared, reflecting the increased diversity of the data set. Indeed, the latter histogram spans values that range from  $\approx 1$ , obtained when two mappings are perfectly parallel, to 0.457, when two mappings are as orthogonal as possible given the properties of the lattice and the selected number of retained sites. In Fig. 5a, we also report a graphical rendering of the two maximally orthogonal mappings, which possess a high value of  $\mathcal{E}$  ( $\mathcal{E} = 847.32$  and  $\mathcal{E} = 843.82$ , respectively) and cover different regions of the enzyme's structure.

In Fig. 5b, we extend these considerations to different values of  $N$ , namely those employed in Sect. 3.1. The random distribution is always confined in a narrow interval of values of  $\cos \theta$ , while WL data sets are capable of spanning a much wider range. In particular, for sufficiently small values of  $N$ , it is possible to retrieve maximally parallel ( $\cos \theta = 1$ ) and maximally orthogonal ( $\cos \theta = 0$ ) mappings inside the WL dataset. Reaching orthogonality is made possible by the fact that, at such low values of  $N$ , it is possible to confine retained sites in two separate regions of the protein structure.



**Fig. 5** Left: histogram of cosine values extracted from random (yellow) and WL CG mappings (purple, see main text) for AKE with  $N = 856$  sites. Elements of  $\mathcal{M}$  with the lowest value of the cosine ( $\cos \theta = 0.457$ ) are shown; such value corresponds to an angle of 63.25 degrees. Right: range of cosine values covered by the two data sets when  $N$  is changed. The dotted black line shows the average value of  $\cos \theta$  over the different random data sets and the yellow region represents the points within  $3\sigma$  from the mean. The red (blue) dotted lines report the maximum (minimum) values of  $\cos \theta$  inside WL data sets, respectively

### 4 Lattice gas analogy and phase transitions

As anticipated in Sect. 3, the reduced representation discussed in the present work, in which a mapping is defined in terms of a *decimation* of the atoms available on the molecular structure, suggests the analogy with a lattice gas. Also in this case, in fact, we have a number  $n$  of nodes that can be occupied by  $N \leq n$  sites, each node being accessible to a single site at a time—thus implementing a hard-core repulsion. This analogy is a classic of statistical mechanics, and enables one, e.g. to map an Ising model to a gas of interacting particles, thus making it manifest that the spontaneous magnetisation in the former and the liquid–gas phase transition in the latter belong to the same universality class [45]. Here, we investigate the consequences of the lattice gas interpretation of reduced representations in order to tackle the issue of characterising the mapping space from a different perspective. Specifically, we mutuate concepts from equilibrium statistical mechanics to show that sharp transitions can occur that separate one or more phases corresponding to classes of reduced representations endowed with markedly distinct structural properties. While the previously performed analysis of the smooth and continuous densities of states  $\Omega_N(\mathcal{E})$  already suggested the existence of such classes, see Sect. 3, for particular numbers of retained sites these are shown to be as distinct as two or more phases of a fluid can be when observed through the perspective of this statistical mechanical analogue.

The role of the energy can be played by the norm of the mapping: in analogy with a lattice gas, we expect that if two retained sites are close to each other, they feel an attractive interaction, thereby reducing the energy. We thus define the energy of the system as

$$E(M) = -\mathcal{E}(M). \tag{47}$$

In the previous sections, we obtained the density of states in terms of the mapping norm,  $\Omega_N = \Omega_N(\mathcal{E})$ . Making use of Eq. 47 we can, thus, write

$$\Omega_N(E) = \Omega_N(-\mathcal{E}). \tag{48}$$

Let us now consider a system governed by the lattice Hamiltonian in Eq. 47 at equilibrium with a reservoir at temperature  $T = \beta^{-1}$ . The partition function of such system can be expressed in terms of  $\Omega_N(E)$  via

$$\begin{aligned} \mathcal{Z}_N(\beta) &= \int dE e^{-\beta E} \Omega_N(E) \\ &\equiv \int dE e^{-(\beta E - S_N(E))}, \end{aligned} \tag{49}$$

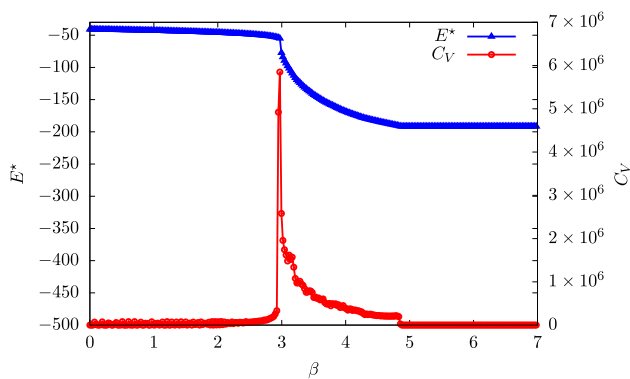
where we used the relation  $S_N(E) = \ln \Omega_N(E)$  to define the entropy. Equation 49 enables us to compute the dimensionless Helmholtz free energy as

$$\begin{aligned} \beta F_N(\beta) &= -\ln \mathcal{Z}_N(\beta) \\ &= -\ln \int dE e^{-(\beta E - S_N(E))}. \end{aligned} \tag{50}$$

While the logarithm of the integral can be theoretically and numerically cumbersome to compute, it is possible to obtain a reasonable estimate of  $\beta F_N$  through a saddle point approximation. Specifically, we can expect that the integral is approximately equal to the largest integrand, so that

$$\int dE e^{-(\beta E - S_N(E))} \simeq C \max_E \left( e^{-(\beta E - S_N(E))} \right), \tag{51}$$

where  $C$  is an immaterial constant. This approximation provides us with a definition of the free energy that is



**Fig. 6** Heat capacity  $C_V$  (red circles, right ordinate) and value of the energy  $E^*$  corresponding to the minimum of the free energy (blue triangles, left ordinate) as functions of the inverse temperature  $\beta$  for the system with  $N = 214$ .  $E^*$  decreases monotonically with  $\beta$ , indicating that higher temperatures correspond to higher values of the average internal energy of the lattice gas, as expected; however, a jump discontinuity in  $E^*$  appears in correspondence of the same value  $\beta_{gl}$  for which the heat capacity features a sharp peak, suggesting the occurrence of a first-order phase transition that separates two distinct phases: a gas (low  $\beta$ ) from a liquid (high  $\beta$ ) for the lattice gas model, and, correspondingly, a sparse phase from a dense, localised phase in the case of mappings

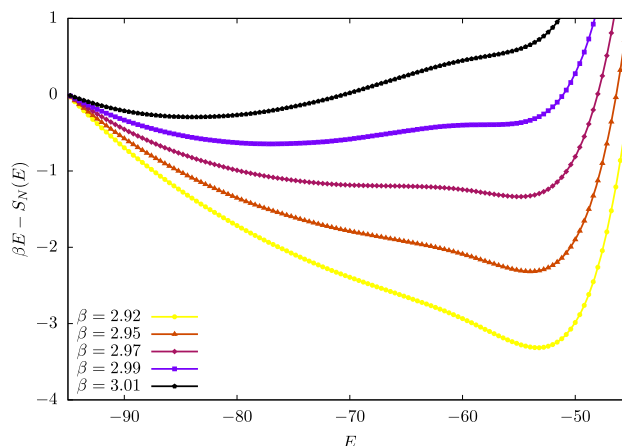
equivalent to the Legendre-Fenchel transform:

$$\beta F_N(\beta) \simeq \min_E (\beta E - S_N(E)). \tag{52}$$

The thermodynamics of the lattice gas at thermal equilibrium can thus be retrieved computing Eq. 52 for a given value of  $N$  at all values of  $\beta$ .

It is particularly instructive to investigate the temperature dependence of  $E^*$ , defined as the value of the energy for which  $\beta E - S(E)$  reaches its minimum. In Fig. 6 (blue curve, left ordinate), we report this function for  $N = 214$ : it is possible to observe that  $E^* = E^*(\beta)$  decreases monotonically, i.e. the lower the temperature, the lower the value of the energy—which corresponds to higher values of the mapping norm. At a particular value  $\beta_{gl}$  of the inverse temperature, however,  $E^*$  drops abruptly: in this context, such behaviour is suggestive of a first-order, discontinuous phase transition.

To gain further insight, we computed the shapes of  $\beta E - S(E)$  for values before and after  $\beta_{gl}$ . These functions, reported in Fig. 7, indeed show two minima separated by a relatively low barrier; increasing  $\beta$ , the absolute minimum shifts from the right to the left, crossing a point for which the two are essentially degenerate. This is the point of coexistence of two distinct “phases” of our lattice gas: a low density one corresponding to distributed mappings (high energy), and one ascribable to more dense, compact conglomerates of sites (low energy). The critical nature of the transition from one regime to the other is confirmed by the



**Fig. 7** Helmholtz free energy  $\beta F$  of the lattice gas as a function of the energy for different  $\beta$  values of the inverse temperature  $\beta$ . For low values of  $\beta$  the curves have a unique and absolute minimum; however, as  $\beta$  increases, a metastable minimum appears that, for a particular value of the inverse temperature, becomes degenerate with the previous one. The presence of a small but appreciable barrier between the two minima makes the position of the absolute minimum,  $E^*$ , shift abruptly from one to the other, as can be seen in Fig. 6, thus making  $E^*(\beta)$  discontinuous

inspection of the heat capacity, computed as

$$C_V = -\beta^2 \frac{\partial^2(\beta F)}{\partial \beta^2} \tag{53}$$

and reported in Fig. 6 (red curve, right ordinate). The sharp, asymmetric peak in  $C_V$ , located at the value  $\beta_{gl}$  of the inverse temperature, shows that the lattice gas crosses a phase transition between a gas and a liquid phase.

A crucial role in this behaviour is played by the number of coarse-grained sites. In fact, as  $N$  increases, the system acquires the possibility of crossing a second phase transition: for example, in the case of  $N = 1070$ , besides the gas–liquid one, it is possible to observe a second, even sharper discontinuity in  $E^*$  for a value of the inverse temperature  $\beta_{ls} > \beta_{gl}$ . This temperature separates the liquid from the solid phase: when the lattice gas particles are sufficiently many, and the temperature sufficiently low, the system can “freeze” in particularly dense mappings with very low entropy. Also in this case, the inspection of the heat capacity (Fig. 13 in Appendix B) supports the interpretation of this as a phase transition. Finally, if the number of sites is too large (e.g.  $N = 1498$ ) no transition is observed, see Fig. 13.

The observations reported in this section resonate with those made by Foley and collaborators in a recent work [15]: there, they observed a phase transition in a system whose degrees of freedom were the retained sites of a reduced model of proteins. In that case, the energy of a given mapping was obtained from the calculation of the spectral quality of the associated model,

a quantity related to the sum of the eigenvalues of the covariance matrix obtained integrating exactly a Gaussian network model (GNM). While apparently very distinct, the spectral quality and the norm of the mapping might bear substantial similarities: in fact, the former entails information about a very simple model, whose mechanical and thermodynamical properties are completely determined by the contact matrix of the underlying protein structure. It is, thus, reasonable to guess that the mapping norm provides, in an effective and efficient manner, information akin to that entailed in the spectral quality about the sparsity or localisation of the retained sites in a given mapping. If and up to which degree these two quantities are related, and how intimately this relation depends on the Gaussian nature of the GNM, requires further investigations that will be the object of future studies.

In conclusion of this section, we note that the observed phase transitions separate mappings so structurally diverse that they can be associated to qualitatively different phases. It is, thus, natural to wonder if and how these phases are organised in the metric space induced by the norm of the mapping, and what information the exploration of the latter can bring about the system it is applied to. To provide an answer to these questions, the next section is devoted to the topological characterisation of the mapping space.

## 5 Topology

In the previous sections, we analyzed the mapping space  $\mathcal{M}$  in terms of the mapping norm  $\mathcal{E}$  and of the cosine between its constituent elements. Here, we discuss the distance  $\mathcal{D}$  (Eqs. 11, 31) between members of  $\mathcal{M}$  with the aim of showing, once again, that a *peculiar* choice of retained CG sites, i.e. one impossible to obtain with random sampling, displays non-trivial statistical properties that reflect in the topological organization of the mapping space.

### 5.1 Topology of the mapping norm space

Without loss of generality,<sup>2</sup> we restrict our investigation to the case  $N = 214$ , namely the number of amino acids of adenylate kinase. We generated a data set of mappings following the protocol explained in Sect. 3.2; in this case, the range of values of  $\mathcal{E}$  was narrower and only 10 macro-bins of amplitude 20 were explored. The data set was constructed by randomly selecting 100 elements for each of the macro-bins, resulting in 1000 CG mappings that homogeneously span the accessible values of  $\mathcal{E}$ .

The sketch map algorithm [46, 47] was employed to embed 1000 points from the high-dimensional space of mappings  $\mathcal{M}$  into a two-dimensional plane, at the same

time preserving as faithfully as possible the relative distances among them—that is to say that nearby points in the mapping space are mapped onto nearby points on the 2D space, see Fig. 8. The two critical parameters of the algorithm are  $\sigma_d$  and  $\sigma_D$ , which modulate how *far* and *close* points are in the low and high resolution space, respectively [46]. To provide the reader with a feeling of the impact that these parameters have on the structure of the low-dimensional representation, we report the embeddings obtained for a low (Fig. 8a) and high (Fig. 8b) value of  $\sigma_d$  and  $\sigma_D$ .

In the first case, presented in Fig. 8a and referring to low values of the  $\sigma$  parameters, data points are in general very sparse and uniformly distributed on the plane, with the exception of a group of points that accumulate in a denser cluster: these are particularly compact mappings localised in a specific region of the molecule. Such mappings remain close to each other even when the  $\sigma$  parameters are increased, thus “squeezing” all points in the low-D embedding, see Fig. 8b. At the same time, we observe that, in this latter scenario, points corresponding to low- $\mathcal{E}$ , uniform mappings collapse in a small region of the embedding space. Furthermore, a third group of points corresponding to compact mappings appears, distinct from the ones previously discussed, and absent in the low- $\sigma$  embedding.

The high- $\sigma$  embedding, thus, highlights two relevant features: first, the presence of specific regions with qualitatively distinct mapping properties; these are either sparse, but necessarily similar one to the other (Fig. 8d), or dense, with atoms localised in different domains of the molecule (Figs. 8c, e). The distance among the latter is necessarily large, since the retained sites cover non-overlapping regions.

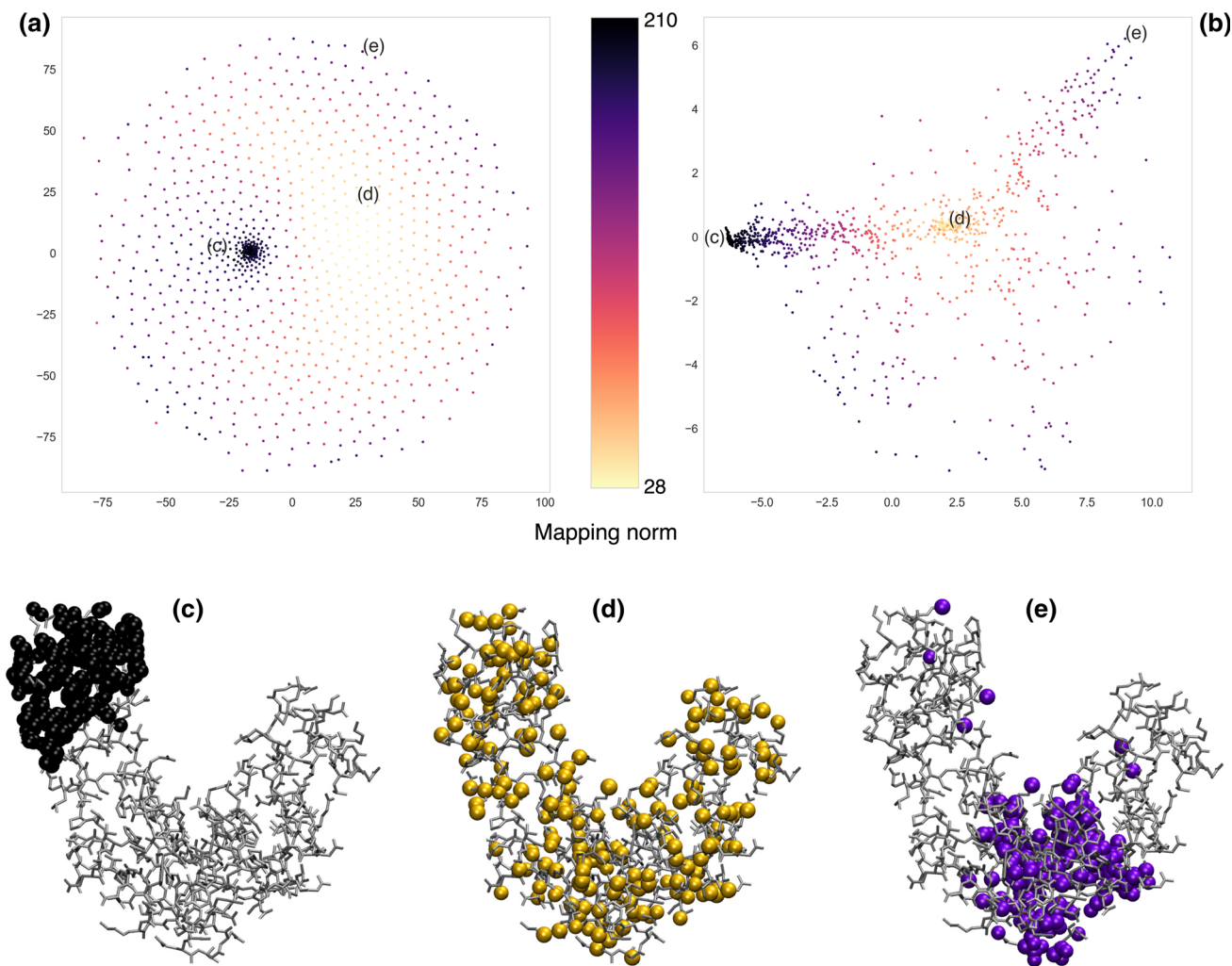
The second relevant feature is that different groups of points, associated to qualitatively distinct types of mappings, can be connected one to the other only “passing through” a third one, as in the case, see Fig. 8, of mapping  $c$  going to  $e$  through  $d$ . This is suggestive of the presence of *routes* in mapping space that join points having the same value of the norm, which, however, cannot be connected through “iso- $\mathcal{E}$ ” paths: to transform mappings such as that in  $c$  into that in  $e$  through a sequence of single-site changes (i.e. one retained atom is discarded, a formerly discarded one is now retained) one cannot but increase or decrease the value of the norm.

### 5.2 Topology of mapping entropy space

While the mapping norm  $\mathcal{E}$  can be employed to investigate the structure of  $\mathcal{M}$  itself, the *quality* of a CG representation can be determined by means of an appropriate cost function. One such function is, e.g. the mapping entropy  $S_{\text{map}}$  [23, 24, 30–32], which is a measure of the intrinsic information loss that is inherent to the process of dimensionality reduction operated by a mapping. This quantity is defined as

$$S_{\text{map}}(M) = k_B \int d\mathbf{r} p_r(\mathbf{r}) \ln \left[ \frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})} \right], \quad (54)$$

<sup>2</sup> The general validity of the discussion presented here is supported by the results obtained for the case  $N = 856$ , which are reported in Fig. 14 of the Appendix.



**Fig. 8** Top: topology of the mapping space  $\mathcal{M}$  in 2D obtained with the sketch map algorithm [46, 47]. The algorithm requires six parameters, namely  $\sigma_d, a_d, b_d$  in the low resolution space and  $\sigma_D, a_D, b_D$  in the original, high resolution one. We select  $\sigma_D = \sigma_d = 2$  for subfigure (a) and  $\sigma_D = \sigma_d = 20$  for subfigure (b), while  $a_d = b_d = 2$  and  $a_D = b_D = 5$  in both cases. Mappings are depicted with different colors depending on their norm  $\mathcal{E}$ . We note that a different choice for  $\sigma_D$  and  $\sigma_d$  results in a completely different 2D embedding (see [46] for a detailed explanation). Bottom: three different mappings located in three separated regions of the plane in (a, b). Mappings in subfigures (c) and (e) possess very high values of  $\mathcal{E}$  and are localised in different domains of the protein. It is interesting to notice that sparse mappings, such as the one in subfigure (d), are clustered in the same region in (b) but not in (a)

where  $p_r(\mathbf{r}) \propto \exp(-\beta u(\mathbf{r}))$  is the Boltzmann weight associated to the atomistic configuration  $\mathbf{r}$ , while  $\bar{p}_r(\mathbf{r})$  represents the “smeared” weight of  $\mathbf{r}$  upon coarse-graining the system by means of a CG mapping  $M$ . More specifically, one introduces the probability of sampling the CG configuration  $\mathbf{R}$ , given by

$$p_R(\mathbf{R}) = \int d\mathbf{r} p_r(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \quad (55)$$

where  $\mathbf{M}(\mathbf{r})$  is the projection operator defined in Eq. 1, as well as the number of high-resolution microstates  $\mathbf{r}$  mapping onto it,

$$\Omega_1(\mathbf{R}) = \int d\mathbf{r} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}). \quad (56)$$

The probability  $\bar{p}_r(\mathbf{r})$  is then defined as [31]

$$\bar{p}_r(\mathbf{r}) = p_R(\mathbf{M}(\mathbf{r})) / \Omega_1(\mathbf{M}(\mathbf{r})). \quad (57)$$

Critically, while both  $p_r(\mathbf{r})$  and  $\bar{p}_r(\mathbf{r})$  are functions of the atomistic coordinates, they differ in assigning the probability to a given high-resolution configuration or microstate, in that  $\bar{p}_r(\mathbf{r})$  associates the same probability with all microstates that map onto the same macrostate  $\mathbf{R}$ . Minimising the mapping entropy  $S_{\text{map}}$  in the space of possible CG representations of the system thus implies maximising the consistency between the reconstructed probability distribution  $\bar{p}_r(\mathbf{r})$  and the all-atom one. In Ref. [23] we derived an approximate expression for Eq. 54, which allows one to compute this observable provided a set of configurations and their

energies are available, e.g. sampled from the canonical ensemble by means of a MD simulation:

$$S_{\text{map}} \simeq k_B \frac{\beta^2}{2} \int d\mathbf{R} p_R(\mathbf{R}) \langle (u - \langle u \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}, \quad (58)$$

where  $\langle (u - \langle u \rangle_{\beta|\mathbf{R}})^2 \rangle_{\beta|\mathbf{R}}$  is the variance of the energies of the atomistic microstates mapping onto macrostate  $\mathbf{R}$ .

While the norm  $\mathcal{E}$  depends only on the geometric properties of a single protein conformation,  $S_{\text{map}}$  is calculated from an ensemble of configurations sampled according to the Boltzmann distribution;  $S_{\text{map}}(M)$ , thus, contains more information than  $\mathcal{E}(M)$ , since it makes explicit use of the average structural and thermodynamical properties of the system.

Here we employ a data set of 1968 CG mappings of AKE with  $N = 214$  generated by us in a previous work [48] and covering a wide range of values of  $S_{\text{map}}$ ; the relations among these mappings are then quantified in terms of their distance  $\mathcal{D}$ , taking the enzyme crystal structure as a reference. With respect to this, it is worth keeping in mind that  $\mathcal{D}$  intimately depends on this reference, and mappings that lie close to each other when a given structure is employed might turn out to be closer or further away from each other when a different conformation is used.

Figure 9 shows that the two-dimensional embedding obtained through the application of the sketch map algorithm separates the CG mappings according to a gradient of  $S_{\text{map}}$ . In particular, the  $x$  component of the sketch map and the mapping entropy  $S_{\text{map}}$  display a clear anticorrelation. The results suggest that highly informative mappings, characterised by low values of  $S_{\text{map}}$ , share geometrical features that are not present in less informative (high  $S_{\text{map}}$ ) representations. In other words, the peculiar resolution distribution found in low- $S_{\text{map}}$  mappings separates them from the other elements of  $\mathcal{M}$ . The relevant features that the mapping entropy highlights thus reverberate in the merely structural characterisation provided by the mapping distance; this connection among the norm  $\mathcal{E}$ , the distance  $\mathcal{D}$ , and highly informative representations is potentially interesting and deserves to be further investigated.

## 6 Extension of the theory to equilibrium sampling: preliminary results

Insofar, our analysis of the mapping space has relied on a definition of a scalar product between CG representations based on a single, static structure of the reference protein. Proteins and other biologically relevant macromolecules, however, are not static objects, but rather flexible entities which, in a typically aqueous environment, undergo fluctuations and deformations. It is therefore natural to extend our metric to incorporate such structural variability; in this Section, we will,

thus, present some preliminary results obtained by performing such an extension, restricting, for the sake of brevity, the discussion to the case of the mapping norm  $\mathcal{E}$ .

We assume our high-resolution (i.e. atomistic) system, constituted by the protein (whose atomic coordinates are indicated with  $\mathbf{r}$ ) and its environment (indicated with  $\mathbf{s}$ ), to be subject to an interaction potential  $u(\mathbf{r}, \mathbf{s})$ . In the canonical ensemble the probability density to sample a given configuration is proportional to the Boltzmann weight, that is,

$$p_r(\mathbf{r}, \mathbf{s}) = \frac{e^{-\beta u(\mathbf{r}, \mathbf{s})}}{Z}, \quad (59)$$

where  $Z = \int d\mathbf{r} d\mathbf{s} e^{-\beta u(\mathbf{r}, \mathbf{s})}$  is the configurational partition function of the system.

The norm  $\mathcal{E}$  of a mapping in Eqs. 30 and 8 only depends on a single conformation of the molecule under examination; however, one can straightforwardly extend the definition of  $\mathcal{E}$ —and analogously of the scalar product and the distance between mappings—to account for the whole conformational space sampled by the system, in that the canonical average of the norm is taken:

$$\begin{aligned} \langle \mathcal{E} \rangle &= \int d\mathbf{r} d\mathbf{s} p_r(\mathbf{r}, \mathbf{s}) \mathcal{E}(\mathbf{r}) \\ &= \int d\mathbf{r} d\mathbf{s} p_r(\mathbf{r}, \mathbf{s}) \frac{1}{\bar{z}(\mathbf{r})} \left( \sum_{i,j=1}^n e^{-r_{ij}^2/4\sigma^2} \chi_{M,i} \chi_{M,j} \right) \\ &= \sum_{i,j=1}^n \langle J_{ij} \rangle \chi_{M,i} \chi_{M,j}. \end{aligned} \quad (60)$$

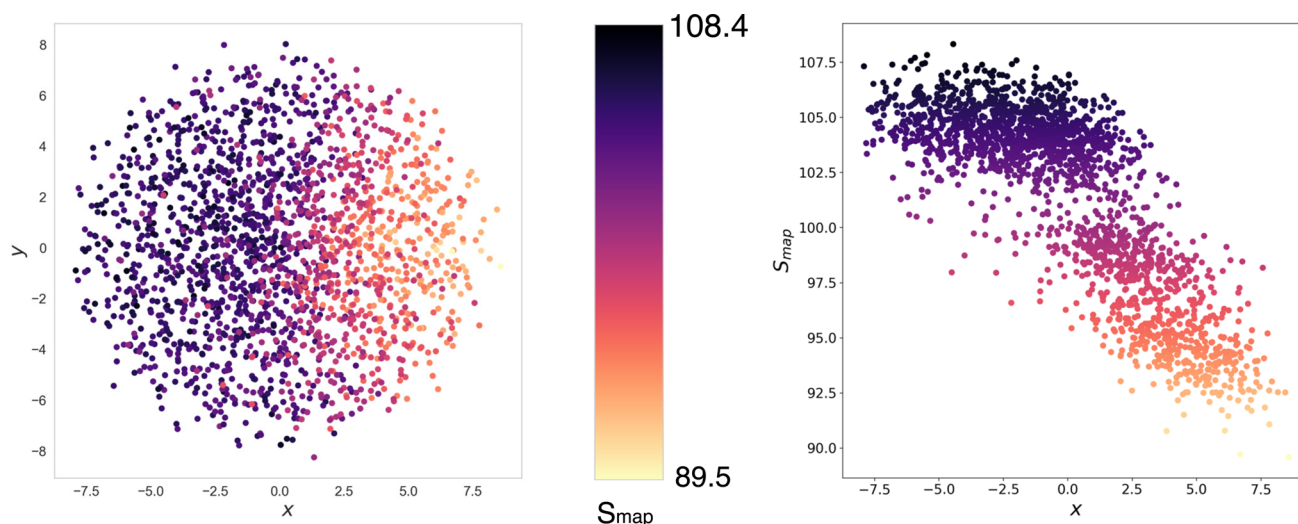
Note that the average is carried out both over the protein and environment degrees of freedom; at the same time, for mappings that only retain protein degrees of freedom, the couplings  $J_{ij}$ —and thus the norm  $\mathcal{E}$ —only depend on the latter. The linearity of the norm with respect to the couplings allows one to first compute their thermal average, that is,

$$\begin{aligned} \langle J_{ij} \rangle &= \int d\mathbf{r} d\mathbf{s} p_r(\mathbf{r}, \mathbf{s}) J_{ij}(\mathbf{r}) \\ &= \int d\mathbf{r} d\mathbf{s} \frac{e^{-\beta u(\mathbf{r}, \mathbf{s})}}{Z} \frac{1}{\bar{z}(\mathbf{r})} e^{-r_{ij}^2/4\sigma^2}, \end{aligned} \quad (61)$$

and subsequently employ them for the calculation of norms, scalar products, and distances, in the same manner as it was done insofar. In this case, however, the resulting values entail information about the conformational space sampled by the whole system, including the environment, described in terms of a high-resolution model.

To investigate the effect that accounting for the conformational variability of the system has on the norm of a mapping, Fig. 10 displays a comparison between the value of  $\mathcal{E}$  computed on the crystal structure of



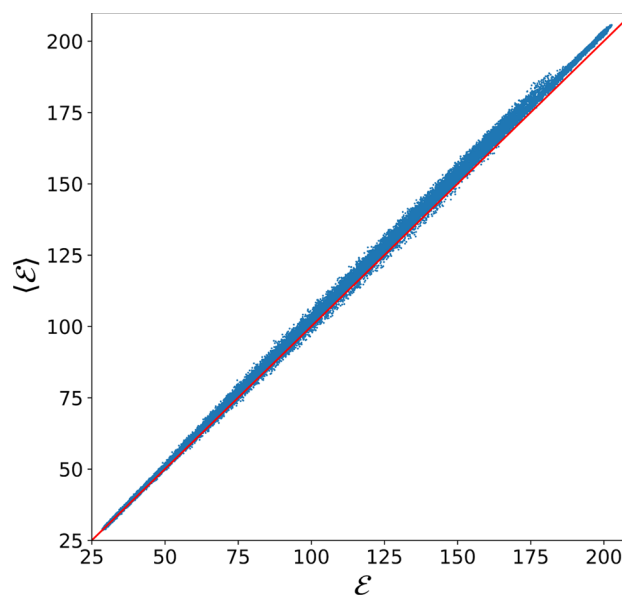


**Fig. 9** Application of the sketch map algorithm to a distance matrix obtained calculating  $\mathcal{D}$  (Eqs. 11 and 30) over a data set of 1968 mappings [48] that span a wide range of values of mapping entropy. The  $x$  component separates very well the data points according to their value of  $S_{\text{map}}$ , thus suggesting that informative mappings can be distinguished among the elements of  $\mathcal{M}$  according to a measure of geometrical similarity such as  $\mathcal{D}$ . The parameters fed to the algorithm are the following:  $\sigma_D = \sigma_d = a_D = b_D = 5$ ,  $a_d = b_d = 2$

AKE and its canonical average  $\langle \mathcal{E} \rangle$  obtained through molecular dynamics sampling. Each point in the plot represents a  $\mathcal{E} - \langle \mathcal{E} \rangle$  pair out of  $5 \times 10^4$  mappings with  $N = 214$  extracted so as to homogeneously span all the possible values of  $\mathcal{E}$ , see Sect. 3.2. The ensemble average is performed over  $10^4$  configurations of a 200 ns long NVT simulation, the technical details of which are available in the SI of Ref. [23].

Interestingly, points are very narrowly dispersed along the diagonal, with a Pearson correlation coefficient very close to unity. This suggests that, at least in this case, the canonical average of  $\mathcal{E}$  is robust to structural changes: we ascribe this behavior to the fact that at the outset of the simulation the protein is in its native state and, due to the strong constraints present in the molecule, the local environment of each atom generally performs small-amplitude fluctuations about a well-defined average. In this particular case, the couplings computed explicitly accounting for the energetics of the system do not induce significant deviations in the value of the norm with respect to their static-structure counterparts it is hence reasonable to expect that the same will hold for the metric and topological properties of the mapping space discussed insofar.

However, this consistency will not be observed when secondary and tertiary structures heavily change, as, e.g. in the case of protein folding: the value of  $\mathcal{E}$  calculated over the unfolded polypeptide chain will not match its canonical average performed over a sample containing folded, more globular configurations. A more detailed understanding of how equilibrium sampling can change the metric properties of the mapping space—especially in the presence of large-amplitude conformational rearrangements—is required, and will be the subject of future work.



**Fig. 10** Scatter plot of the single-conformation mapping norm  $\mathcal{E}$ , calculated on the crystal structure of AKE, against its canonical average  $\langle \mathcal{E} \rangle$  (Eq. 60) for  $5 \times 10^4$  CG mappings with  $N = 214$ . The red straight line with slope one serves as a guide to the eye. The Pearson correlation coefficient is 0.9997

## 7 Conclusions

In this work we have addressed the problem of defining a measure to quantify the distance between two low-resolution representations of a macromolecule, and to “explore” the metric space induced by it.

The recent advances in the computational investigation of soft and biological matter have provided us with the tools to perform large-scale simulations of large and complex systems; however, due to the sheer size of the data produced, one has to filter out the large amount of detail with which the system is described [8], thus relying on a coarse-grained description of it.

Decimation mappings offer a simple and intuitive way of applying this filter, in that only a subset of a molecule's atoms is retained; however, not all mappings entail or deliver the same amount of information, and the identification of the most informative ones allows one to highlight relevant properties of the system. Various methods have been devised [20–25] to identify the most informative mappings as the solution to an optimisation problem, which thus relies on the definition of an appropriate cost function. Since the landscape induced by the latter is typically a rather rugged one, as it is often the case in the field of complex systems [49, 50], it is to be expected that more than one “optimal” solution will be found. Hence, to understand the relationship among such solutions, as well as between structural representation and physical properties in general, it is of fundamental importance to possess an instrument to measure the difference, or distance, among mappings.

The metrics proposed here, which builds on the SOAP measure proposed by Csányi and coworkers [35, 36], has been employed to quantify the number, dissimilarity, and structural features of different mappings of a macromolecule in a static conformation, thereby providing the basis for quantitative analysis of the aforementioned relationship.

The exploration of the mapping space relied on the application of the Wang–Landau enhanced sampling algorithm [28, 29], which allowed us to compute the (logarithm of the) density of states for mappings with a given number of CG sites, as a function of their squared norm. On the one hand, these calculations brought to the surface information about “special” (i.e. atypical) representations that, just due to their lower number with respect to randomly sampled ones, are exponentially suppressed; on the other hand, we made use of the densities of states to implement a lattice-gas analogy in terms of which we have interpreted mappings of qualitatively different types as different phases of the same physical system undergoing a phase transition. We have then made use of the distance between mappings to investigate the properties of optimal reduced representations obtained by minimising the mapping entropy, a measure of the amount of information that a given mapping can return about the underlying system at thermal equilibrium: this last analysis has shown that optimal mappings are markedly distant, and therefore qualitatively different, from randomly sampled ones, thus corroborating the idea that the former belong to a particular subregion of the mapping space endowed with non-trivial properties. Finally, we proposed a possible extension of the theory to *samples* of conformations at thermal equilibrium, focusing on the case of the mapping norm. In this manner, the  $J_{ij}$  couplings are weighted

with the probabilities associated to each configuration, thus indirectly accounting for the system's energy.

A number of questions remain open, which could not be addressed in this work. As a first thing, in the application of the theory to the system under examination we have observed a substantial consistency between the values of the mapping norm computed with single-structure couplings and their averaged counterparts; however, it is reasonable to expect that this won't be a general case. Consequently, the inclusion of the reference system's conformational variability might lead to interesting outcomes in the analysis of structural and topological properties of the mapping space. This relevant line of research is currently under investigation.

A second open issue, partially related to the former, is that the phase transitions that we observed are analogous to the ones reported in a previous work [15], where explicit reference to the molecule's free energy was made: this connection might entail important insights in the relationship between the properties of mappings as elements of the mapping space and the functional characteristics of the underlying system, and it certainly deserves to be further inspected.

Finally, the general character of the tools developed in this work make them suitable to be easily combined with other methods. For example, they can be employed to quantitatively gauge similarities and differences among the solutions to the mapping optimisation problem obtained making use of the various protocols proposed in the literature; or to boost the accurate determination of cost functions profiles, whose computation can be accelerated by a preliminary exploration of the mapping space followed by a cycle of biased enhanced sampling simulations [48].

In conclusion, the mathematical, biophysical, and computational methods developed and applied in this work have served to start gathering the treasure of information buried in the relationship between how we look at a system and the properties it is endowed with, of which we think that what has been reported here has just scratched the surface.

**Acknowledgements** The authors thank Giovanni Mattiotti and Virginia Agostiniani for a 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).

## Author contributions

RP and RM elaborated the study. RM and RP developed the scalar product between mappings. RM developed the software for the mapping norm calculations and the Wang–Landau sampling. MG produced the sketch maps. RM and MG performed the analysis. All authors contributed to the interpretation of the results and the writing of the manuscript.

**Funding** Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement.

**Data availability** This manuscript has associated data in a data repository. [Authors' comment: The data associated to the paper ARE freely available on the Zenodo repository <https://doi.org/10.5281/zenodo.4954580>.]

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Appendix A: Wang–Landau sampling

For the set of selected degrees of coarse graining  $N$  reported in Table 1, the corresponding density of states  $\Omega_N(\mathcal{E})$  defined in Eq. 33—that is, the number of possible CG representations in the mapping space  $\mathcal{M}$  that retain  $N$  atoms and have a squared norm of  $\mathcal{E}$ —was determined by relying on the protocol proposed by Wang and Landau (WL) [28, 29, 42, 43].

WL sampling enables to self-consistently determine  $\Omega_N(\mathcal{E})$ , or, for computational convenience, the associated entropy  $S_N(\mathcal{E}) = \ln[\Omega_N(\mathcal{E})]$ , through a sequence  $k = 0, \dots, K$  of nonequilibrium Monte Carlo (MC) simulations that provide an increasingly accurate approximation to the correct result [28, 29]. Given a partition of the ensemble of possible norms  $\mathcal{E}$  in bins of width  $\delta\mathcal{E}$ , the pivotal ingredients of the WL iterative scheme are, respectively: (i) the MC estimate of the entropy  $\bar{S}_N(\mathcal{E})$ ; (ii) the histogram of visited norms at iteration  $k$ ,  $H_N^k(\mathcal{E})$ ; and (iii) the modification factor  $\ln(f_k)$  governing convergence of the algorithm—for  $k = 0$ , one typically sets  $\bar{S}_N(\mathcal{E}) = 0$  and  $\ln(f_0) = 1$ .

At the beginning of each iteration  $k$ , the histogram  $H_N^k(\mathcal{E})$  is set to zero. Subsequently, a series of MC moves is performed in which a transition between two mappings  $M$  and  $M'$ , respectively with norms  $\mathcal{E}$  and  $\mathcal{E}'$ , is accepted with probability, see Eq. 43,

$$\alpha_{M \rightarrow M'} = \min [1, \exp(-[\bar{S}_N(\mathcal{E}') - \bar{S}_N(\mathcal{E})])] . \quad (\text{A1})$$

In our case, both mappings have  $N$  sites but differ by the retainment of a single atom. If the move  $M \rightarrow M'$  is accepted, the histogram  $H_N^k$  and entropy  $\bar{S}_N$  are updated according to

$$H_N^k(\mathcal{E}') = H_N^k(\mathcal{E}') + 1, \quad (\text{A2})$$

$$\bar{S}_N(\mathcal{E}') = \bar{S}_N(\mathcal{E}') + \ln(f_k), \quad (\text{A3})$$

while in case of rejection one has to replace  $\mathcal{E}'$  with  $\mathcal{E}$  in Eqs. A2 and A3. As highlighted by Eqs. A1 and A3, the early

stages of the WL scheme tend to “push away” the sampling from already visited regions of the mapping space, thus significantly boosting its exploration compared to randomly drawing CG representations. The algorithm then evolves to generate a “random walk” in the space of possible norms [43].

The series of MC moves within iteration  $k$  is interrupted when the histogram of sampled norms  $H_N^k(\mathcal{E})$  is “flat”, meaning that each of its entries does not exceed a threshold distance from the average of the histogram  $\langle H_N^k \rangle$ . A typical requirement is  $p_{\text{flat}} \times \langle H_N^k \rangle < H_N^k(\mathcal{E}) < (2 - p_{\text{flat}}) \times \langle H_N^k \rangle$  for every value of  $\mathcal{E}$ ,  $p_{\text{flat}}$  being a predefined flatness parameter. When the flatness condition is satisfied, iteration  $k + 1$  of the algorithm begins with a reduced modification factor—in our case, we set  $\ln(f_{k+1}) = \frac{1}{2} \ln(f_k)$ . Finally, iterations over  $k$  are stopped when  $\ln(f_k) < \ln(f_{\text{end}}) \ll 1$ ,  $\ln(f_{\text{end}})$  being another control parameter provided in input to the WL protocol. Up to an additive constant, the MC estimate of the entropy  $\bar{S}_N(\mathcal{E})$  reproduces the exact result  $S_N(\mathcal{E})$  with an accuracy of order  $\ln(f_{\text{end}})$  [51].

In WL sampling, knowledge of the boundaries of the domain of the density of states  $\Omega_N(\mathcal{E})$  (equivalently, of the entropy  $S_N$ ) plays a crucial role in the convergence of the iterative scheme, e.g. for checking the flatness of the histogram  $H_N(\mathcal{E})$  throughout the simulation [44]. In contrast to “more traditional” systems such as Ising ferromagnets on a lattice [28], this information is not readily available in our case. As such, we initially performed a set of explorative, non-iterative—i.e. without updating the modification factor  $\ln(f_k)$ —WL runs so as to approximately locate the minimum and maximum norms  $\mathcal{E}_{\text{min}}(N)$  and  $\mathcal{E}_{\text{max}}(N)$  achievable at each degree of CG'ing. To mitigate the effect of bins that are only visited at a very late stage of the simulation, thus risking to temporarily “trap” the mapping space exploration, we followed the protocol described in Ref. [44]: every time a bin  $[\mathcal{E}_i, \mathcal{E}_i + \delta\mathcal{E}]$  was populated for the first time, it was marked as “visited”, the corresponding entropy was initialised to the minimum of  $\bar{S}_N(\mathcal{E})$  over the previously visited bins, and the histogram  $H_N(\mathcal{E})$  was reset. The results obtained from these preliminary runs for  $\mathcal{E}_{\text{min}}(N)$  and  $\mathcal{E}_{\text{max}}(N)$  as a function of  $N$  are displayed in Fig. 3 and summarised in Table 2.

Having identified the range of possible norms for each investigated degree of CG'ing, we subsequently moved to the determination of the corresponding entropies  $S_N(\mathcal{E})$  via the iterative WL scheme. To boost convergence of the algorithm, for each  $N$  we slightly reduced the interval of norms  $[\mathcal{E}_{\text{min}}(N), \mathcal{E}_{\text{max}}(N)]$  with respect to the one predicted by the explorative runs, and divided this spectrum in a total of  $W_N$  overlapping windows of equal width, see Table 2 [29]. The overlap between two consecutive windows was fixed to half their size. Within each window, we then performed a separate WL simulation in which confinement of the range of norms was achieved by rejecting all mapping moves  $M \rightarrow M'$  that would bring the exploration outside the  $\mathcal{E}$  interval of interest. In discarding these moves, we concurrently updated the histogram and entropy of the current state according to Eqs. A2 and A3 to avoid boundary effects [52]. Furthermore, also in these production runs we kept track of the norm bins that were sampled during the course of the simulation, resetting the histogram every time a new bin was populated, the entropy of which was initialised to the minimum of  $\bar{S}_N(\mathcal{E})$  over the previously visited ones. All

**Table 2** Lower and upper bound of the norms  $\mathcal{E}_{\min}$  and  $\mathcal{E}_{\max}$  identified by the set of preliminary WL runs for each degree of CG'ing  $N$ , see Fig. 3, and corresponding values  $\bar{\mathcal{E}}_{\min}$  and  $\bar{\mathcal{E}}_{\max}$  employed in the reconstruction of the entropy  $S_N(\mathcal{E})$  through the iterative WL scheme. For boosting convergence of the algorithm, the interval  $[\mathcal{E}_{\min}, \mathcal{E}_{\max}]$  was divided in  $W_N$  windows overlapping by half their width; the associated simulations were performed with a flatness parameter  $p_{\text{flat}} = 0.90$ , assuming convergence of the iterations when the modification factor  $\ln(f_k)$  became smaller than  $\ln(f_{\text{end}}) = 10^{-6}$

$N$	$\mathcal{E}_{\min}$	$\mathcal{E}_{\max}$	$\bar{\mathcal{E}}_{\min}$	$\bar{\mathcal{E}}_{\max}$	$W_N$
214	25.6	209.8	28	192	3
428	92.4	424.4	98	410	7
642	206.2	633.8	218	618	15
856	371.2	853.0	390	840	17
1070	600.2	1074.2	612	1062	17
1284	900.2	1298.4	910	1290	18
1498	1287.8	1514.6	1296	1504	12

WL simulations were performed setting  $p_{\text{flat}} = 0.90$ , and checking the histogram flatness over the visited bins every  $3 \cdot 10^6$  “single spin” MC moves that involved the swap of a retained and a non-retained atom in the mapping. We interrupted the iterative scheme when the modification factor  $\ln(f_k)$  became smaller than  $\ln(f_{\text{end}}) = 10^{-6}$ .

For each degree of CG'ing  $N$ , the outcome of the converged WL protocol is a set of entropies  $\bar{S}_{N,i}(\mathcal{E})$ ,  $i = 1, \dots, W_N$ , restricted to bounded and overlapping  $\mathcal{E}$  domains that need to be combined to provide the complete  $S_N(\mathcal{E})$  over the whole range of investigated norms. These  $\bar{S}_{N,i}(\mathcal{E})$  differ—besides numerical uncertainties that are inherent to the self-consistent scheme [53]—from the exact results  $S_{N,i}(\mathcal{E})$  by additive constants  $C_{N,i}$  that are not uniform across the different WL windows. Rather than determining the relative shifts that most accurately superimpose the various  $\bar{S}_{N,i}(\mathcal{E})$  profiles within the overlapping regions—see e.g. Ref. [42]—in this work, we directly considered the (numerical) derivatives of  $S'_{N,i}(\mathcal{E})$  in each WL window,

$$\bar{S}'_{N,i}(\mathcal{E}) = \frac{d\bar{S}_{N,i}(\mathcal{E})}{d\mathcal{E}} = \frac{1}{T}, \quad (\text{A4})$$

where  $T$  is the “temperature” of the system. These derivatives are not affected by the constants  $C_{N,i}$ , so that each  $\bar{S}'_{N,i}(\mathcal{E})$  is approximately equal to its exact counterpart  $S'_{N,i}(\mathcal{E})$ . One can thus combine all the derivatives of the different WL windows in a global derivative  $S'_N(\mathcal{E})$  that extends over the whole range of analysed norms, from which the overall entropy  $S_N(\mathcal{E})$  can be calculated as

$$S_N(\mathcal{E}) = S_N(\mathcal{E}_{\min}(N)) + \int_{\mathcal{E}_{\min}(N)}^{\mathcal{E}} S'_N(\mathcal{E}') d\mathcal{E}', \quad (\text{A5})$$

where  $\mathcal{E}_{\min}(N)$  is the lowest norm sampled at degree of CG'ing  $N$ . Note that in contrast to systems as the ferromagnetic Ising model, we do not *a priori* know the value of  $S_N(\mathcal{E}_{\min})$ , so that the entropy  $S_N(\mathcal{E})$  will be only determined up to a constant.

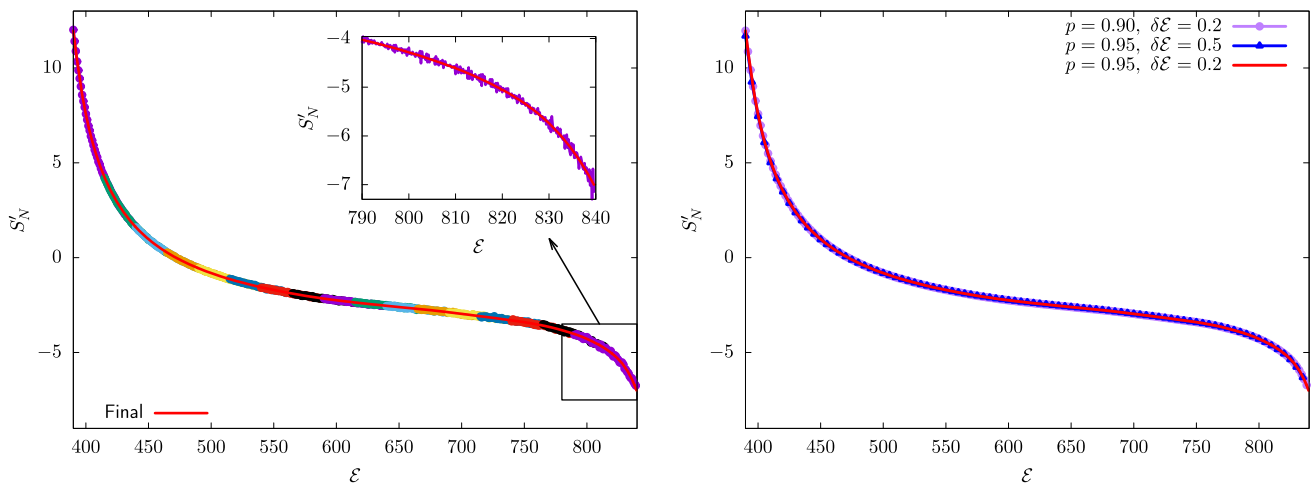
To merge the set of derivatives and reconstruct  $S'_N(\mathcal{E})$  for each degree of CG'ing, we first applied a Savitzky–Golay filter [54] to the WL estimates of the entropies  $\bar{S}_{N,i}(\mathcal{E})$  so as to reduce the amount of noise in the simulation results, and consequently smoothen the derivative  $\bar{S}'_{N,i}(\mathcal{E})$  of each window. A comparison of the derivatives obtained in presence or absence of the filter, see Fig. 11, highlights how this only applies a tiny correction to the original data, which nonetheless significantly improves the quality of the set of  $S'_{N,i}(\mathcal{E})$ .

Despite this refinement, the presence of residual numerical fluctuations leave room to a certain degree of arbitrariness in how, within the overlap region of two consecutive windows, the combined derivative should be constructed. At the same time, these fluctuations appear to be marginal in the vicinity of the center of a window, while tend to slightly increase if we move towards its boundaries (data not shown). Exploiting this observation, we thus tackled the problem of merging the derivatives of two consecutive windows  $i$  and  $i+1$  within their overlap region as follows: first, we divided the region in three separate intervals, the central one being roughly double the size of the other two. Given that the windows overlap by half their width, it follows that the first interval will be located close to the center of window  $i$ , where the derivative  $S'_{N,i}(\mathcal{E})$  is numerically more stable, but close to the boundary of window  $i+1$ , where  $S'_{N,i+1}(\mathcal{E})$  is slightly more noisy. The opposite holds for the last interval. As such, in the first and last regions, we considered the combined derivative  $S'_N(\mathcal{E})$  to be equal to  $S'_{N,i}(\mathcal{E})$  and  $S'_{N,i+1}(\mathcal{E})$ , respectively. Within the central interval, by increasing  $\mathcal{E}$  we move from the vicinity of the center of window  $i$  to that of window  $i+1$ . In this latter region, we, thus, set the final derivative  $S'_N$  to a weighted average of the derivatives of the two windows, namely

$$S'_N(\mathcal{E}) = (1 - \alpha(\mathcal{E})) S'_{N,i}(\mathcal{E}) + \alpha(\mathcal{E}) S'_{N,i+1}(\mathcal{E}), \quad (\text{A6})$$

where  $\alpha(\mathcal{E})$  is a mixing parameter that linearly increases from zero to one as  $\mathcal{E}$  moves from the left to the right boundary of the interval.

Repeating this interpolation for all the set of  $W_N$  windows—note that in the first (resp. last) half of the first (resp. last) window no mixing applies—provided us, for each of the analysed degrees of CG'ing, with a global derivative  $S'_N(\mathcal{E})$  that extends over the whole range of sampled norms. Figure 11 displays a comparison between  $S'_N(\mathcal{E})$  and the original, piecewise derivatives for the case  $N = 856$ , highlighting the accuracy of our approach. This accuracy is further confirmed by the smooth behavior of the second derivative  $S''_N(\mathcal{E})$  calculated from the reconstructed  $S'_N$ , that we display in Fig. 4 for  $N = 856$ . Starting from the set of  $S'_N(\mathcal{E})$ , the corresponding entropies  $S_N(\mathcal{E})$  were subsequently obtained via direct integration, see Eq. A5, producing the profiles presented in Fig. 4 and in Fig. 12. In these figures, entropies were shifted so that their minimum value is zero.



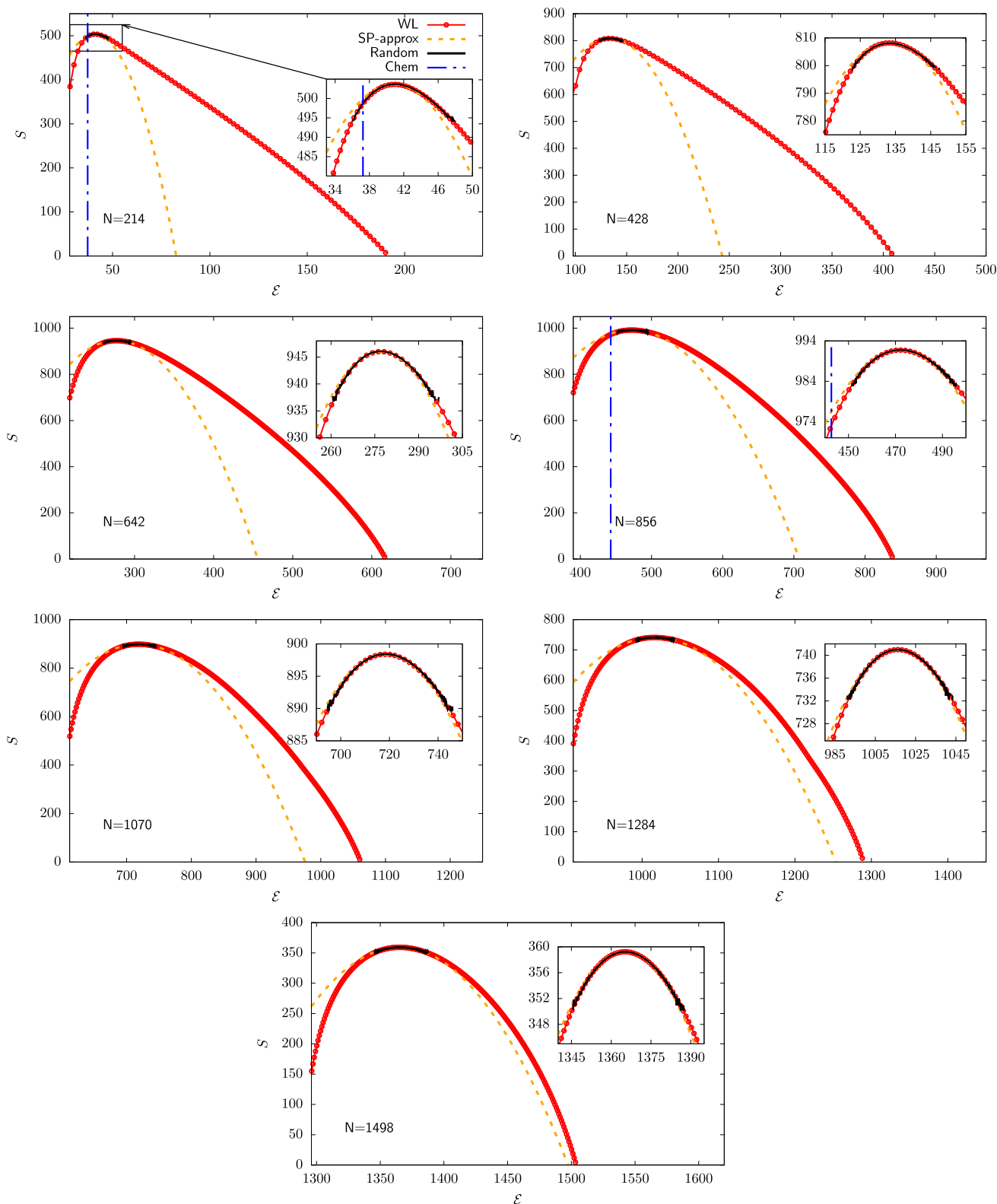
**Fig. 11** Left: Main figure: Comparison between the piecewise entropy derivatives  $S'_{N,i}(\mathcal{E}), i = 1, \dots, W_N$  of AKE (colored dots) obtained from the set of independent WL simulations performed over the  $W_N$  overlapping windows, and the final, reconstructed derivative  $S'_N$  (“Final”, red line) calculated through the mixing procedure of the  $S'_{N,i}$  described in the text. We report results for  $N = 856$ . Inset: Behaviour of the derivative  $S'_{N,i}(\mathcal{E})$  for the last WL window before and after the application of the Savitzky–Golay filter to the raw simulation results for the entropy  $\bar{S}_{N,i}(\mathcal{E})$ . Right: Reconstructed derivatives  $S'_N(\mathcal{E})$  for  $N = 856$  obtained by varying a subset of the input parameters of the WL protocol. Specifically, we test the sensitivity of the results to a change in the flatness parameter  $p_{\text{flat}}$  as well as in the bin width  $\delta\mathcal{E}$ , considering as reference profile the derivative  $S'_N$  obtained by setting  $p_{\text{flat}} = 0.95$  and  $\delta\mathcal{E} = 0.2$  (red full line)

Finally, it is interesting to test the dependence of our results on the input parameters of the WL protocol. While initially all MC simulations were performed with a flatness condition  $p_{\text{flat}} = 0.90$ , for the case  $N = 856$ , we repeated the calculations using  $p_{\text{flat}} = 0.95$  finding a perfect agreement of the reconstructed  $S'_N(\mathcal{E})$ , see Fig. 11. The same sensitivity analysis was performed for the bin size  $\delta\mathcal{E}$  dictating the discretisation of the mapping norms: while in all simulations we employed  $\delta\mathcal{E} = 0.2$ , by repeating the calculations for  $N = 856$  with a bin width of  $\delta\mathcal{E} = 0.5$  we again observed excellent agreement of the results, see Fig. 11.

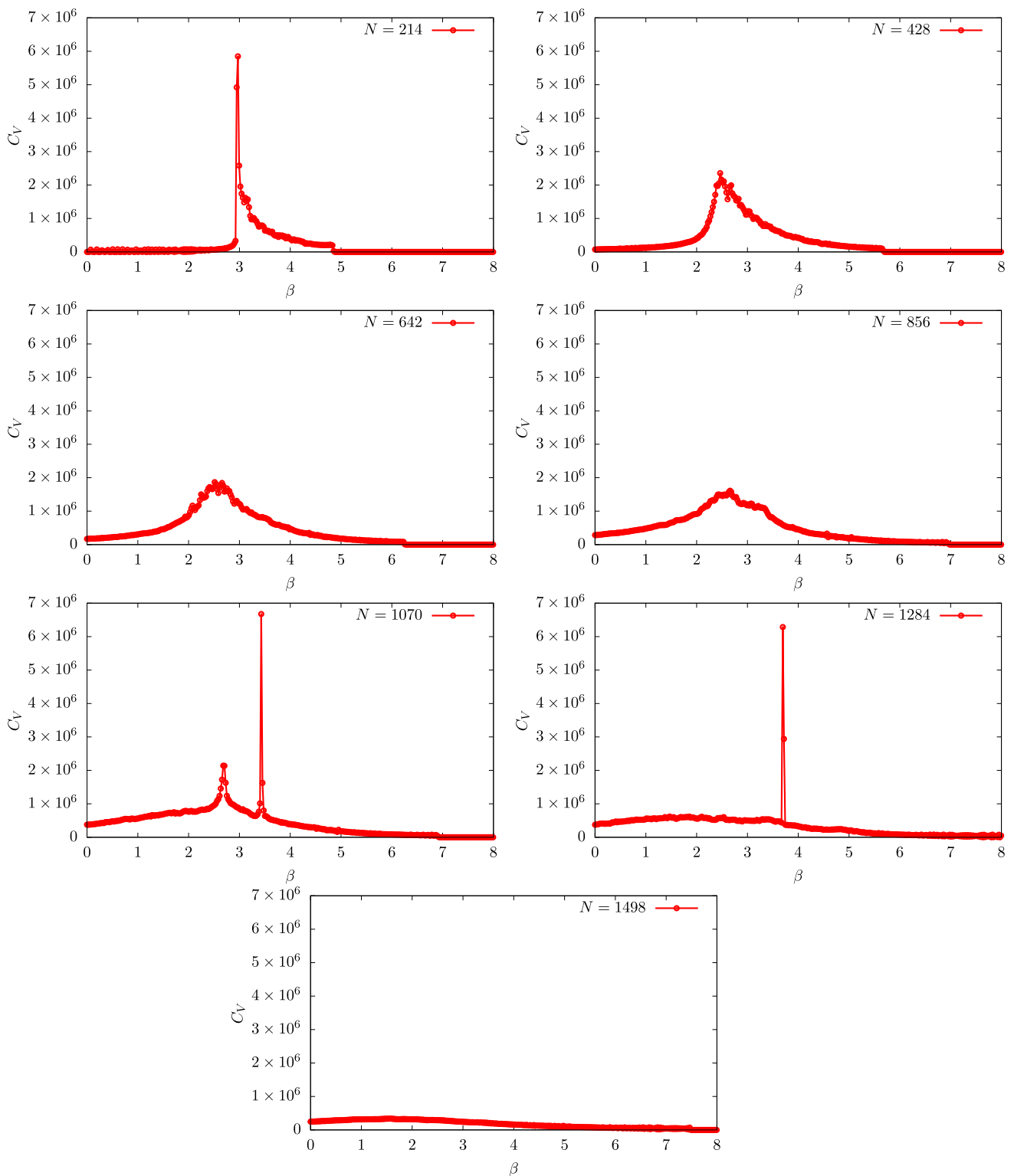
### Appendix B: Heat capacity of the lattice gas

In this Appendix we provide additional information about the phase transitions observed in the lattice gas analogue

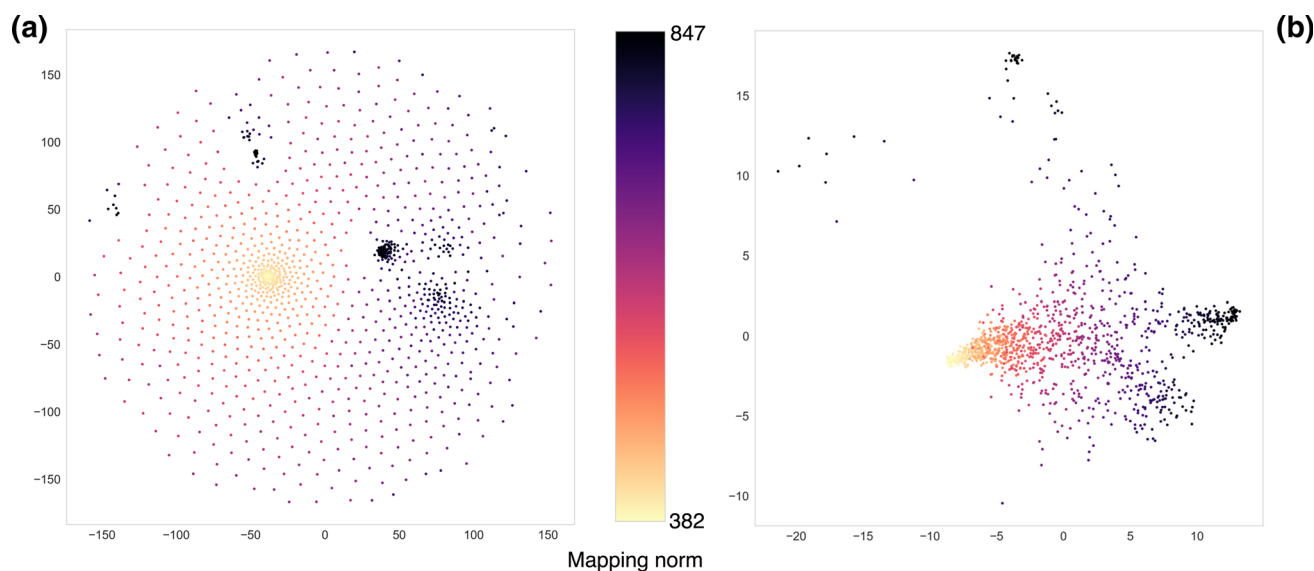
of the mapping norm. Specifically, Fig. 13 displays the heat capacity  $C_V$  of the lattice gas, see Eq. 53 in the main text, as a function of the inverse temperature  $\beta$  for different degrees of CG'ing  $N$ , calculated from the Legendre-Fenchel transform  $\beta F_N(\beta)$  of the WL entropies  $S_N(\mathcal{E})$ , see Eq. 52. While for the highest degree of CG'ing investigated,  $N = 1438$ , the heat capacity has a smooth dependence on  $\beta$ , for  $N = 1284$   $C_V$  develops a sharp peak for low temperatures, which suggests the presence of a solid–liquid transition in the system. By further decreasing the degree of CG'ing, this solid–liquid peak gets initially flanked by a shoulder located at higher temperatures, and finally disappears. The shoulder, on the other hand, grows in magnitude as fewer and fewer sites are retained, and becomes a discontinuity for  $N = 214$ , suggesting the appearance of a liquid–gas transition.



**Fig. 12** Behavior of the entropy  $S_N(\mathcal{E})$  of AKE for different degrees of CG'ing. For each  $N$ , we report results obtained via (i) Wang–Landau sampling (“WL”, red dotted lines), shifting the data so that the minimum of  $S_N$  over the range of investigated norms is zero; (ii) a saddle-point approximation of the WL predictions (“SP-approx”, orange dashed lines); and (iii) a random drawing of CG representations (“Random”, black lines), in this latter case shifting the curve so that its maximum coincides with the one of the corresponding WL profile. For  $N = 214$  (resp.  $N = 856$ ) we further report the squared norm associated to the  $C_\alpha$  (resp. backbone) mapping (“Chem”, blue dashed line), a CG representation that is routinely employed while CG'ing a protein system [5, 8]



**Fig. 13** Dependence of the heat capacity  $C_V$  on the inverse temperature  $\beta$  for the lattice gas analogue of the mapping norm of AKE calculated at several degrees of CG'ing. Sharp peaks in  $C_V$  at high (resp. low) values of  $N$  suggest the presence of a solid–liquid (resp. liquid–gas) transition in the system. It should be noted that the scales of  $\beta$  and  $C_V$  are the same in all plots



**Fig. 14** Application of the sketch map algorithm to the mapping space  $\mathcal{M}$ : case of  $N = 856$ . We employed the same set of parameters described in Fig. 8 of the main text, where CG mappings have  $N = 214$ , with the exception of  $\sigma_D$  and  $\sigma_d$  in subfigure (a), which are equal to 5. The two-dimensional embedding shown here displays similar properties to the one in the main manuscript; specifically, if  $\sigma_D$  and  $\sigma_d$  have low values, essentially all the data points are depicted as isolated instances in  $\mathcal{M}$  and only the extremely sparse and globular mappings are capable of forming recognisable clusters. With a higher value of these parameters, all sparse mappings collapse in a well-defined region of the plane, from which several routes depart, each one directed towards globular mappings covering different domains of the protein structure

## References

1. A. Singharoy, C. Maffeo, K.H. Delgado-Magnero, D.J. Swainsbury, M. Sener, U. Kleinekathöfer, J.W. Vant, J. Nguyen, A. Hitchcock, B. Isralewitz et al., Atoms to phenotypes: molecular design principles of cellular energy metabolism. *Cell* **179**(5), 1098–1111 (2019)
2. M.I. Zimmerman, J.R. Porter, M.D. Ward, S. Singh, N. Vithani, A. Meller, U.L. Mallimadugula, C.E. Kuhn, J.H. Borowsky, R.P. Wiewiora et al., Sars-cov-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat.e Chem* **13**, 651–659 (2021)
3. S. Takada, Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol.* **22**(2), 130–137 (2012)
4. W.G. Noid, Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **139**(9), 090901 (2013)
5. S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A.E. Dawid, A. Kolinski, Coarse-grained protein models and their applications. *Chem. Rev.* **116**(14), 7898–7936 (2016)
6. C. Clementi, Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **18**(1), 10–15 (2008)
7. T. Sun, V. Minhas, N. Korolev, A. Mirzoev, A.P. Lyubartsev, L. Nordenskiöld, Bottom-up coarse-grained modeling of dna. *Front. Mol. Biosci.* **8**, 1–17 (2021)
8. M. Giulini, M. Rigoli, G. Mattiotti, R. Menichetti, T. Tarenzi, R. Fiorentini, R. Potestio, From system modelling to system analysis: the impact of resolution level and resolution distribution in the computer-aided investigation of biomolecules. *Front. Mol. Biosci.* **8**, 460 (2021)
9. J. Maupetit, P. Tuffery, P. Derreumaux, A coarse-grained protein force field for folding and structure prediction. *Proteins Struct. Funct. Bioinform.* **69**(2), 394–408 (2007)
10. F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragolini, Y. Chebaro, J.-F. St-Pierre, M. Kalimeri, A. Barducci et al., The opep protein model: from single molecules, amyloid formation, crowding and hydrodynamics to dna/rna systems. *Chem. Soc. Rev.* **43**(13), 4871–4893 (2014)
11. A. Davtyan, N.P. Schafer, W. Zheng, C. Clementi, P.G. Wolynes, G.A. Papoian, Awsem-md: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **116**(29), 8494–8503 (2012)
12. M. Chen, X. Lin, W. Zheng, J.N. Onuchic, P.G. Wolynes, Protein folding and structure prediction from the ground up: the atomistic associative memory, water mediated, structure and energy model. *J. Phys. Chem. B* **120**(33), 8557–8565 (2016)
13. A. Liwo, M. Baranowski, C. Czaplowski, E. Gołaś, Y. He, D. Jagieła, P. Krupa, M. Maciejczyk, M. Makowski, M.A. Mozolewska et al., A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions. *J. Mol. Model.* **20**(8), 2306 (2014)
14. A. Liwo, C. Czaplowski, A.K. Sieradzan, E.A. Lubecka, A.G. Lipska, Ł. Golon, A. Karczyńska, P. Krupa, M.A. Mozolewska, M. Makowski et al., Scale-consistent approach to the derivation of coarse-grained force fields for simulating structure, dynamics, and thermodynam-



- ics of biopolymers. *Prog. Mol. Biol. Transl. Sci.* **170**, 73–122 (2020)
15. T.T. Foley, K.M. Kidder, M.S. Shell, W. Noid, Exploring the landscape of model representations. *Proc. Natl. Acad. Sci.* **117**(39), 24061–24068 (2020)
  16. C. Clementi, H. Nymeyer, J.N. Onuchic, Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.* **298**(5), 937–953 (2000)
  17. A.R. Atilgan, S. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80**(1), 505–515 (2001)
  18. K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold. *Science* **334**(6055), 517–520 (2011)
  19. A. Grottesi, N. Bešker, A. Emerson, C. Manelfi, A.R. Beccari, F. Frigerio, E. Lindahl, C. Cerchia, C. Talarico, Computational studies of sars-cov-2 3clpro: Insights from md simulations. *Int. J. Mol. Sci.* **21**(15), 5346 (2020)
  20. M.A. Webb, J.-Y. Delannoy, J.J. de Pablo, Graph-based approach to systematic molecular coarse-graining. *J. Chem. Theory Comput.* **15**(2), 1199–1208 (2019)
  21. Z. Li, G.P. Wellawatte, M. Chakraborty, H.A. Gandhi, C. Xu, A.D. White, Graph neural network based coarse-grained mapping prediction. *Chem. Sci.* **11**(35), 9524–9531 (2020)
  22. W. Wang, R. Gómez-Bombarelli, Coarse-graining auto-encoders for molecular dynamics. *npj Comput. Mater.* **5**(1), 125 (2019)
  23. M. Giulini, R. Menichetti, M.S. Shell, R. Potestio, An information-theory-based approach for optimal model reduction of biomolecules. *J. Chem. Theory Comput.* **16**(11), 6795–6813 (2020)
  24. T.T. Foley, M.S. Shell, W.G. Noid, The impact of resolution upon entropy and information in coarse-grained models. *J. Chem. Phys.* **143**(24), 243104 (2015)
  25. P. Diggins IV., C. Liu, M. Deserno, R. Potestio, Optimal coarse-grained site selection in elastic network models of biomolecules. *J. Chem. Theory Comput.* **15**(1), 648–664 (2018)
  26. B.J. Alder, T.E. Wainwright, Studies in molecular dynamics. i. general method. *J. Chem. Phys.* **31**(2), 459–466 (1959)
  27. M. Karplus, Molecular dynamics simulations of biomolecules. *Acc. Chem. Res.* **35**(6), 321–323 (2002)
  28. F. Wang, D. Landau, Determining the density of states for classical statistical models: a random walk algorithm to produce a flat histogram. *Phys. Rev. E* **64**(5), 056101 (2001)
  29. F. Wang, D.P. Landau, Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86**(10), 2050 (2001)
  30. M.S. Shell, The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **129**(14), 144108 (2008)
  31. J.F. Rudzinski, W.G. Noid, Coarse-graining entropy, forces, and structures. *J. Chem. Phys.* **135**(21), 214101 (2011)
  32. M.S. Shell, Systematic coarse-graining of potential energy landscapes and dynamics in liquids. *J. Chem. Phys.* **137**(8), 084503 (2012)
  33. L.P. Kadanoff, Notes on Migdal’s recursion formulas. *Ann. Phys.* **100**(1–2), 359–394 (1976)
  34. J.V. José, L.P. Kadanoff, S. Kirkpatrick, D.R. Nelson, Renormalization, vortices, and symmetry-breaking perturbations in the two-dimensional planar model. *Phys. Rev. B* **16**(3), 1217 (1977)
  35. A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments. *Phys. Rev. B* **87**(18), 184115 (2013)
  36. S. De, A.P. Bartók, G. Csányi, M. Ceriotti, Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**(20), 13754–13769 (2016)
  37. C.W. Müller, G.J. Schlauderer, J. Reinstein, G.E. Schulz, Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4**, 147–56 (1996)
  38. F. Pontiggia, A. Zen, C. Micheletti, Small and large scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics. *Biophys. J.* **95**(12), 5901–5912 (2008)
  39. R. Potestio, F. Pontiggia, C. Micheletti, Coarse-grained description of proteins’ internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. *Biophys. J.* **96**(12), 4993–5002 (2009)
  40. Y.E. Shapiro, E. Kahana, E. Meirovitch, Domain mobility in proteins from nmr/srls. *J. Phys. Chem. B* **113**(35), 12050–12060 (2009)
  41. E. Formoso, V. Limongelli, M. Parrinello, Energetics and structural characterization of the large-scale functional motion of adenylate kinase. *Sci. Rep.* **5**, 8425 (2015)
  42. M.S. Shell, P.G. Debenedetti, A.Z. Panagiotopoulos, Generalization of the Wang-Landau method for off-lattice simulations. *Phys. Rev. E* **66**(5), 056703 (2002)
  43. L.Y. Barash, M. Fadeeva, L. Shchur, Control of accuracy in the Wang-Landau algorithm. *Phys. Rev. E* **96**(4), 043307 (2017)
  44. T. Wüst, D. Landau, The hp model of protein folding: a challenging testing ground for Wang-Landau sampling. *Comput. Phys. Commun.* **179**(1–3), 124–127 (2008)
  45. P. Beale, *Statistical Mechanics* (Elsevier Science, Amsterdam, 2011)
  46. M. Ceriotti, G.A. Tribello, M. Parrinello, Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci.* **108**(32), 13023–13028 (2011)
  47. M. Ceriotti, G.A. Tribello, M. Parrinello, Demonstrating the transferability and the descriptive power of sketch-map. *J. Chem. Theory Comput.* **9**(3), 1521–1532 (2013)
  48. F. Errica, M. Giulini, D. Bacciu, R. Menichetti, A. Micheli, R. Potestio, A deep graph network-enhanced sampling approach to efficiently explore the space of reduced representations of proteins. *Front. Mol. Biosci.* **8**, 136 (2021)
  49. D. Sherrington, Landscape paradigms in physics and biology: Introduction and overview. *Phys. D Nonlinear Phenom.* **107**, 117–121 (1997)
  50. W. Janke, *Rugged Free Energy Landscapes: Common Computational Approaches to Spin Glasses, Structural Glasses and Biological Macromolecules. Lecture Notes in Physics* (Springer, Berlin, 2007)
  51. D. Landau, S.-H. Tsai, M. Exler, A new approach to Monte Carlo simulations in statistical physics: Wang-landau sampling. *Am. J. Phys.* **72**(10), 1294–1302 (2004)

52. B. Schulz, K. Binder, M. Müller, D. Landau, Avoiding boundary effects in Wang-Landau sampling. *Phys. Rev. E* **67**(6), 067102 (2003)
53. R. Belardinelli, V. Pereyra, Fast algorithm to calculate density of states. *Phys. Rev. E* **75**(4), 046701 (2007)
54. A. Savitzky, M.J. Golay, Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**(8), 1627–1639 (1964)