Soft Matter

PAPER

View Article Online
View Journal | View Issue

Cite this: *Soft Matter*, 2022, **18**, 7064

Information-theoretical measures identify accurate low-resolution representations of protein configurational space†

Margherita Mele, [a] Roberto Covino [b] and Raffaello Potestio *[ac]

The steadily growing computational power employed to perform molecular dynamics simulations of biological macromolecules represents at the same time an immense opportunity and a formidable challenge. In fact, large amounts of data are produced, from which useful, synthetic, and intelligible information has to be extracted to make the crucial step from knowing to understanding. Here we tackled the problem of coarsening the conformational space sampled by proteins in the course of molecular dynamics simulations. We applied different schemes to cluster the frames of a dataset of protein simulations; we then employed an information-theoretical framework, based on the notion of *resolution* and *relevance*, to gauge how well the various clustering methods accomplish this simplification of the configurational space. Our approach allowed us to identify the level of resolution that optimally balances simplicity and informativeness; furthermore, we found that the most physically accurate clustering procedures are those that induce an ultrametric structure of the low-resolution space, consistently with the hypothesis that the protein conformational landscape has a self-similar organisation. The proposed strategy is general and its applicability extends beyond that of computational biophysics, making it a valuable tool to extract useful information from large datasets.

Received 16th May 2022,
Accepted 26th July 2022

DOI: 10.1039/d2sm00636g

rsc.li/soft-matter-journal

1 Introduction

A celebrated quote attributed to Aristotle states that "the whole is more than the sum of its parts". This statement effectively encapsulates the defining characteristic of complex systems, whose global properties generally cannot be traced back to those of their individual constituents, but rather *emerge* from the interplay of the latter.

Among those systems that most clearly show this behaviour, a prominent example is represented by biological macromolecules such as proteins: these, being composed of several thousands of interacting atoms, display a rich and sophisticated phenomenology over a broad range of length and time scales, which cannot be naively predicted or anticipated from the knowledge of their structure. In order to generate, inspect, and comprehend the properties and behaviour of these systems, computational, *in silico* methods have been developed, most notably molecular dynamics[1–4] (MD) simulations, that serve the purpose, among others, of sampling the conformational space of the molecule. Once a dataset of sampled conformations, or frames, is available, however, one faces the problem of extracting useful and intelligible information out of it, separating the relevant feature from the irrelevant detail.

This task can be carried out through dimensionality reduction[5] or clustering schemes. These methods rely on some notion of similarity – usually a structural similarity – between distinct conformations to group together those whose differences are negligible, while a much larger discrepancy exists from other frames or groups of frames. It might appear desirable to devise these clustering schemes taking advantage of a preexisting knowledge about the system, in order to steer the algorithm towards physically sensible partitions of the sampled conformational space. It can be the case, however, that an undesired bias is introduced in the process, with potentially detrimental consequences for the interpretation of the results; alternatively, one might hope for a completely unsupervised procedure,[6–8] so as to let the system itself dictate how to cluster its data points, and allow the intrinsic organisation of the conformational space to emerge.

A recently developed information-theoretical approach, the resolution-relevance framework,[9,10] holds the promise to carry out this task of identifying intrinsically informative low-dimensional representations of the system in an unbiased

[a] *Physics Department, University of Trento, via Sommarive, 14 I-38123 Trento, Italy. E-mail: raffaello.potestio@unitn.it*
[b] *Frankfurt Institute for Advanced Studies, 60438 Frankfurt am Main, Germany*
[c] *INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, I-38123 Trento, Italy*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2sm00636g

manner. This approach relies on distinct measures of the information content of a dataset to group the instances of the latter in a way that optimally separates information from noise, and allows the extraction of the largest amount of information about the generative process that underlies the data points. The method, however, operates on the basis of a predefined clustering procedure, whose impact cannot be neglected in the assessment of the resulting partition's quality and physical soundness: in fact, the values of these information metrics for a given arrangement of the data points in clusters only make sense relative to the strategy employed to perform the grouping.

In this work, we tackle the issue of investigating if, and to what extent, different strategies to carry out the clustering of protein MD trajectory frames affect the intrinsic quality of the resulting partitions, and if the resolution-relevance framework can be employed to make sense of these results. We apply this strategy to a dataset of 12 structurally dissimilar proteins as well as to a specific case study, making use of agglomerative clustering strategies with 7 different linkage criteria. Our results support the hypothesis that the resolution-relevance analysis can select those linkage methods giving rise to low-resolution representations of the protein conformational space that reproduce the high-resolution reference with the highest degree of fidelity; furthermore, we propose that this capacity of performing a sensible clustering is a direct consequence of the clustering method being capable of preserving the intrinsically hierarchic structure and ultrametricity of the protein conformational space.

2 The relevance-resolution framework and the impact of the underlying clustering method

The resolution-relevance framework, or critical variable selection, is a recently developed method[11] for identifying important variables without any prior knowledge of, or assumption on, their nature. The idea at the heart of the approach is that the information on the generative model that underlies the elements of an empirical sample is contained in the distribution of their frequencies, that is to say, in the number of times different outcomes occur in the data set. It can be shown[12,13] that the entropy of the outcome distribution, dubbed *resolution*, quantifies the overall information content of the sample, while the entropy of the frequency distribution, dubbed *relevance*, measures the amount of important information. In this section we provide a synthetic review of this approach, specialising the formulation for the application in the context of computational biophysics.

The output of a molecular dynamics simulation consists of a collection of M configurations, or frames, $\hat{s} = (s^{(1)}, \ldots, s^{(M)})$; these can be thought of as the realisations of a stochastic sampling process, where each element takes the values of one of the possible system states $s = (s_1, \ldots, s_n)$, with $n \gg M$. In spite of absolute structural differences, two distinct configurations might result equivalent for a practical purpose; for example, if the relative position of a few atoms in two frames differs by less than a given tolerance, they might be considered essentially equivalently representative of the same overall organisation of the molecule. In analysing the outcomes of a simulation it is thus crucial to filter out redundant details by grouping together structures that can be safely associated to the same state; hence, one has to perform a *clustering*.

The most trivial level of clustering consists in identifying each frame as a distinct cluster (assuming that no pair of exactly identical configurations exists in the sample). Such a representation clearly allows the highest level of detail in the description of the dataset, but it bears no use in making sense of it; the number of clusters thus has to be reduced, and frames that in principle describe distinct structural organisations have to be grouped together if their distance (as quantified by an appropriate measure) takes values below a predefined threshold. In so doing, the number of clusters is reduced from $K = M$ to values $K < M$, which correspond to increasingly less resolved representations of the system's configuration space.

For each partition of the dataset it is possible to compute the corresponding values of the aforementioned resolution and relevance. Resolution is defined as (note that we employ logarithms in units of M, or Mats, so that $\log_M M = 1$):

$$H[s] = -\sum_{s=1}^{K} \frac{k_s}{M} \log_M \frac{k_s}{M} \tag{1}$$

where k_s is the number of frames associated to the cluster with label s, and k_s/M is the empirical probability that a randomly chosen frame from the data set belongs to cluster s. The normalisation condition $\sum_s k_s = M$ ensures that k_s/M is indeed at most unity.

Since all frames in a cluster are indistinguishable at the level of detail employed, the lowest resolution value $H[s] = 0$ is obtained when all frames are gathered in the same cluster; similarly, the largest value $H[s] = \log_M M = 1$ is attained when each frame is a singleton cluster. Both extremes are equally little informative: on one hand, when the resolution is too low, potentially different conformations are grouped in the same cluster; on the other hand, discriminating all M states as distinct is equivalent to associate to each of them the same probability, which does not provide useful information to infer the underlying generative process. Hence, resolution alone is not sufficient to pinpoint an optimal level of detail at which the system should be inspected, and a second measure has to be employed to this end. Such measure is the relevance $H[k]$, given by:

$$H[k] = -\sum_{k=1}^{M} \frac{km_k}{M} \log_M \frac{km_k}{M} \tag{2}$$

where m_k is the number of outcomes s for which $k_s = k$, and km_k/M is the empirical probability that a frame chosen at random from the data set is associated to a state with (un-normalised) frequency k.

The relevance is null for both extreme values of the resolution: in the case $H[s] = 0$ all frames are in the cluster with $k = M$,

which gives $m_M = 1$, $m_{k\neq M} = 0$, and hence $\frac{km_k}{M}\ln_M\frac{km_k}{M} = 0\forall k$; in the case $H[s] = 1$ all clusters only contain one frame, hence $m_1 = M$, $m_{k\neq 1} = 0$ and $\frac{km_k}{M}\ln_M\frac{km_k}{M} = 0\forall k$ as well.

As the relevance is nonnegative and equal to zero at the extremes of the resolution range, it follows that the relevance as a function of the resolution has to have a maximum; thence, there must be one representation, with intermediate resolution and positive relevance value, that more than the others allows an informative characterisation of the underlying probability distribution.[14] The partitions at the right of the maximum are in what is called the *under-sampling regime*, $M \ll n$, in which the statistics of the data is relatively poor and several frames associated to distinct states can happen to appear the same number of times. For a given value of the resolution in this region, those partitions that maximise the relevance – the *most informative samples* – feature a frequency distribution that follows a power law, $m_k \sim k^{-\mu-1}$ with $\mu > 0$, such that each value of the frequency is associated to a distinct number of clusters. In particular, the partition for which the quantity $H[s] + H[k]$ is the largest has $\mu = 1$: this corresponds to Zipf's law, $m_k \sim k^{-2}$, which is associated to the point of optimal tradeoff between parsimony of the representation (low resolution) and its informativeness (high relevance).[15,16] This is the case, for example, for the frequency of the words in a language,[17] and the spike patterns of neuron populations[18] (even though the occurrence of Zipf's law in the latter case, *e.g.* in neural data from the retina, might be a consequence of the statistics of input – the underlying visual scene – rather than of the neuronal dynamics itself[19–21]).

In a context of complete ignorance, *i.e.* in absence of any information about the data except their empirical probability k/M based on some pre-defined classification, the frequency is the only label that can be employed to distinguish between frames in distinct states.[15,16] The frequency thus constitutes a *minimally sufficient representation*, which, in absence of additional information about the data, allows one to write the resolution $H[s]$ as:

$$H[s] = H[k] + H[s|k] \tag{3}$$

where the information content of a given partition in states s is decomposed in the relevance and a noise term, $H[s|k]$. The latter is larger for a partition based on the frequency than for any other partition, and it constitutes a measure of the degeneracy of the distinct classifications that produce the same frequency distribution.[16] In fact, different classifications that preserve the number of elements in each state are fully equivalent, as relevance and resolution solely depend on the partition's combinatorics: that is, *how many* clusters are there and *how many* elements there are in each cluster; both quantities are blind to *which* elements are included in a given cluster.

This is a crucial aspect, which shows that the implication *high relevance therefore informative representation* is not necessarily true. The concept of *informative representation*, obtained as maximization of the relevance, is independent of what the sample represents. Indeed, a random clustering of the system may produce partitions with high relevance values that are informative about some generative process but devoid of any significant information on the *specific* model producing the data under study (Fig. 1). Consequently, complementing the relevance-resolution framework with a sensible strategy to group elements into clusters based on the physical properties (geometric, structural, energetic, *etc.*) of the sample, is crucial to steer the generation of empirical probabilities that have maximum relevance consistent with the imposed boundaries.

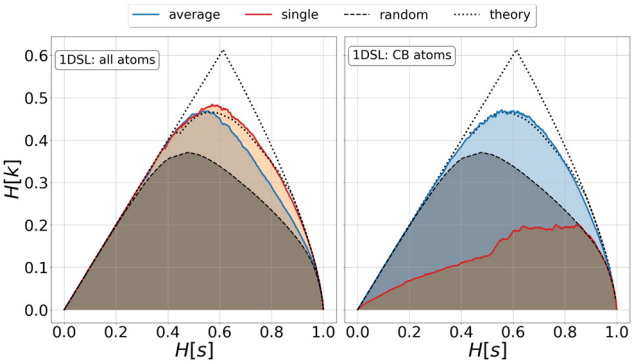

Fig. 1 Relevance-resolution curves obtained partitioning the simulation data of protein 1DSL with two clustering protocols, average linkage and single linkage. The two panels differ by the atom selection adopted: all atoms on the left and C_β atoms on the right. In both panels the random curve is also present, obtained by randomly partitioning the structures into groups (see Methods). Each ($H[s]$, $H[k]$) point corresponds to a fixed number of clusters in which the frames of the whole trajectory are grouped. The theoretical highest $H[k]$ for given $H[s]$ is also plotted (black dotted line); specifically, the dotted lines show the upper and lower bounds to the theoretical maximum.[12,13] All curves show the expected characteristic trend: zero relevance at the lowest (all frames in a cluster) and highest (every frame in a single cluster) resolution values.

In the case of agglomerative clustering of molecular structures, the clustering procedure relies on the specific functions defining the inter-frame and inter-cluster distance (Fig. 1). The former defines the property in terms of which the similarity of two configurations is quantified (structure, compactness, energy, *etc.*), while the latter is the metric employed to measure the distance between clusters. The latter, which is referred to as *linkage criterion*, thus determines the protocol employed for agglomerative clustering, and different choices result in different partitions of the system. The ability of a given protocol to return a meaningful partitioning naturally depends on the specific dataset under examination. For example, in single linkage the similarity of two clusters is equivalent to that of their most similar members; this protocol is effective in identifying compact and separate clusters, but it is strongly subject to the *chaining effect*: two close-by points can form a bridge between two clusters, causing them to merge and resulting in an elongated cluster. Without any prior characterisation of the explored configurational space, the goodness of the partition can be assessed only *a posteriori*.

In this work, we employed various linkage criteria and investigated the most informative partitions obtained with each

of them, with two objectives: first, to identify those linkage methods that are most appropriate for a meaningful and physically sound clustering of molecular structure data; second, to infer general properties of the configurational space explored by proteins in molecular dynamics simulations. In the next section we report the results of our study; the list and definitions of the employed distance measures and linkage criteria are provided in the Appendix.

3 Results

We investigated the impact of choosing seven different linkage methods on the relevance-resolution curves obtained clustering protein structures sampled in molecular dynamics simulations (see Methods). To this end, we employed two sets of systems: a dataset of 12 structurally distinct molecules, and a specific protein, adenylate kinase,[22,23] that was taken as a case study. The similarity of configurations was quantified in terms of the root mean square distance (RMSD) between frame pairs, after the removal of global translations and rotations, and we analysed configurations at three levels of structural resolution: all-atom, which accounts for all heavy atoms of the molecule; alpha carbon (C_α); and beta carbon (C_β). The last two consist in reduced representations with a single coarse-grained site per amino acid. Resolution-relevance curves were computed in all aforementioned cases, however the overall informativeness of a given representation was assessed in terms of the multi-scale relevance[24] (MSR), *i.e.* the area under the relevance curve: this is a global measure of how a specific clustering scheme performs at all levels of resolution.

The data in Fig. 1 suggest that it is possible to characterise the ability of a clustering method to identify informative partitions across various levels of structural detail. Indeed, it can be seen that, although all the curves in the figure were obtained by clustering the same trajectory frames, the MSR values obtained are rather different. In particular, the selection of atoms drastically influences the MSR value obtained through a given clustering protocol. The results of this analysis, carried out for 7 linkage criteria and 12 proteins at varying levels of structural resolution (all, C_α, and C_β atoms), are summarised in Fig. 2, where the deviation of the MSR relative to the random reference ($\overline{\mathrm{MSR}}$, see eqn (4)) is plotted against the mean value of the RMSD matrices used for the clustering procedure. Fig. 2 additionally allows comparison of the obtained MSR values with the optimal one, namely the area under the theoretical curve maximizing relevance at each level of resolution (the colored region marks the upper and lower limits of the theoretical maximum[12,13]). For some linkage criteria, the $\overline{\mathrm{MSR}}$ values are always positive (*i.e.* larger than the random value $\mathrm{MSR_R}$) and close to the optimal one regardless of the system and structural selection employed; for other methods, the performance depends on the system or its representation. In particular, the plot shows that the performance of some methods correlates with the mobility of the protein as quantified by the average RMSD. The linkage criteria can be divided in two groups: those for which the MSR value correlates with the mean value of the RMSD matrix ('centroid' and 'median') and those for which it does not ('average', 'complete', 'weighted' and 'ward'). Consequently, it can be argued that the former are less effective and reliable than the latter: in fact, their ability to identify simplified but meaningful representations strongly depends on the specific data under examination, and their applicability is restricted to high-resolution (all-atom) descriptions. This represents a substantial shortcoming, in that these methods are not adequate for the analysis of all-atom data in less detailed terms.

The single linkage criterion shows a peculiar pattern, and deserves to be discussed separately. Here, the distance between clusters is the minimum pairwise distance between their

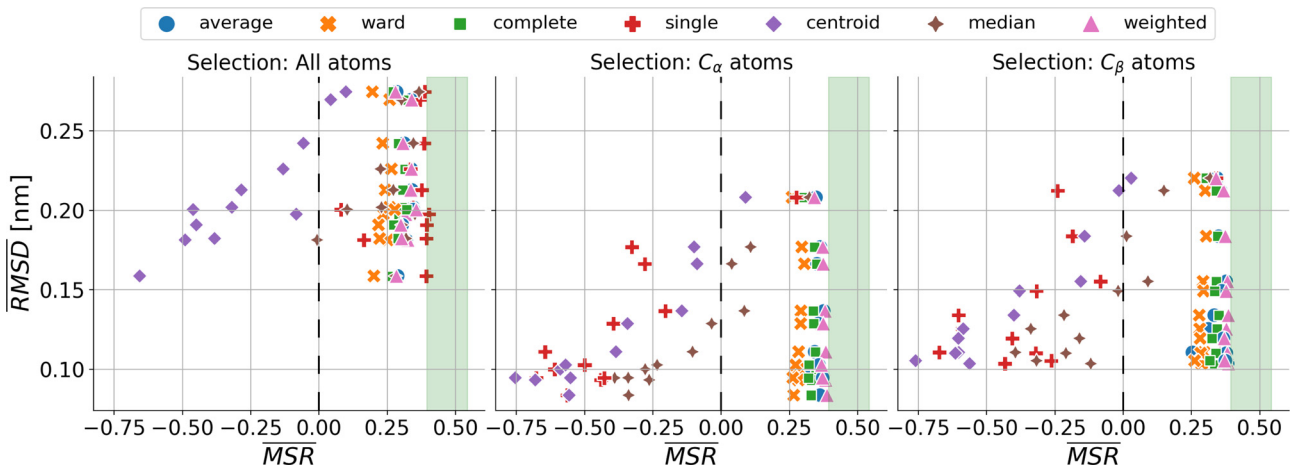

Fig. 2 Performance of clustering algorithms, as quantified by the relative $\overline{\mathrm{MSR}}$, related to the protein mobility, given by the mean value of the RMSD matrix ($\overline{\mathrm{RMSD}}$). Each point corresponds to one of the 12 proteins of the dataset. The coloured region highlights the range of MSR values associated to the relevance-resolution curves falling between the theoretical limits to the curve of maximal relevance. The three panels differ by the atom selection adopted for clustering: all atoms on the left, C_α atoms in the middle, and C_β atoms on the right. The colours and the markers identify the linkage criterion used in the clustering procedure as shown in the legend.

elements. Hence, in the all-atom case, as the matrix elements are widely spread, this algorithm manages to form differently-populated clusters. Conversely, the C_α and C_β selections implement a coarse-graining that "blurs" the structural differences from the outset; therefore, the algorithm tends to form highly populated clusters by putting together even frames that are relatively different from each other (chaining effect), and provides a rather uninformative representation of the system.

It can be shown[25] that some of the hierarchical clustering algorithms induce a monotonic hierarchy, *i.e.* the values in the inter-cluster distance matrix increase monotonically during agglomerative clustering. Algorithms that induce a monotonic hierarchy lead to an ultrametric in the cluster space:[26] this implies that the metric distance satisfies an inequality stronger than the triangular one.[27,28] In our analysis, it turns out that clustering protocols that satisfy these qualities coincide with those showing a consistently positive $\overline{\mathrm{MSR}}$; the only exception to this trend is single linkage, which, although inducing an ultrametric in cluster space, still shows negative $\overline{\mathrm{MSR}}$ values when coarse-grained representations of the system are employed. In this case, however, the clustering protocol is severely limited by the chaining effect, plausibly producing uninformative partitions of the system and consequently obtaining a lower or comparable MSR value with respect to the random case. Taken together, these results suggest that protein structures sampled in the course of a molecular dynamics simulation populate the configurational space according to an ultrametric structure, which is consistent with the self-similar organisation of the free energy landscape observed in previous works;[29–33] additionally, the MSR appears to be capable of capturing, in a parameter-free and unbiased manner, the effectiveness of a clustering method in finding informative representations of a biomolecule's configurational variability at different scales of resolution, in that MSR correlates with the method's capacity to preserve the ultrametric structure of the reference configurational space.

Since relevance and resolution are not sensitive to the features of the elements gathered in the clusters and their relative similarity, it is crucial to validate *a posteriori* that partitions with a higher relevance are indeed more informative than the others. This task can be achieved through dimensionality reduction techniques. In particular, we use diffusion maps,[34–36] which project the high-dimensional trajectory of the molecule in Cartesian coordinate space onto a low-dimensional manifold of collective coordinates called diffusion coordinates (DCs). We thus performed a comparison of the distribution of points (frames or cluster centroids) in the space spanned by the first two DCs obtained from the high-resolution (HR) or low-resolution (LR) representation of the system (see Methods). It is reasonable to expect that a meaningful partition gathers, in the same cluster, frames close in the HR space, and that the distribution of centroids resembles the HR distribution, thus allowing the same information to be extracted. In order to assess and compare the goodness of partitions we resort to the decomposition of the covariance matrix in its inter- and intra-state contributions (see eqn (5)–(7) in the Methods section). In fact, a key property of an informative LR representation of a system is to capture more information in the retained data than what is left in the discarded ones; we thus expect that the trace of the inter-cluster covariance will be significantly higher than the intra-cluster one.

We thus proceeded to investigate in greater detail the relationship between linkage method and informativeness of the resulting LR representation of a protein's conformational space. To this end, we focused on a specific case study, that of adenylate kinase: the configurations obtained from a 800 ns long simulation, reduced to the positions of the sole C_β atoms, were grouped with the single, average, and random clustering methods at various levels of cluster numbers, corresponding to 22 different resolution values in the range [0,1]; the spacing in resolution between the first 7 representations is $\sim$0.06, for the remaining ones is $\sim$0.04. For each LR partition we computed, and reported in Fig. 3, the resolution-relevance plots (left panel), the trace of the inter-state and intra-state matrix (middle panel), and the value of the Pearson correlation coefficient (PCC) between the first two DC in the HR and LR space (right panel); the last two sets of quantities are plotted against the number of clusters K employed in the representation.

It is possible to observe that, at a fixed level of resolution, the LR representations obtained through the average linkage are simpler and more informative, as the corresponding number of clusters K is lower and the relevance value higher than those obtained by single linkage; the latter also produces a relevance curve that lies very close to that of the random partition, while the average linkage curve closely approaches the lower bound to the maximum. This hierarchy in performance is also confirmed by the trends of the trace of the covariance matrices and PCC: already with a small number of clusters ($K \sim 10$) average linkage identifies LR representations in which the inter-cluster contribution is significantly higher than the intra-cluster one. In contrast, at the same number of clusters the single linkage algorithm produces partitions for which the two terms are of the same order of magnitude, or even ranked oppositely (the intra-state contribution is larger than the inter-state).

These results show that the LR representations obtained with single linkage clustering do not fully capture the information contained in the data and destroy a comparable, or even larger, amount of information than what is maintained. In general, the inter-state (resp. intra-state) contribution to the covariance for average linkage is always significantly higher (resp. lower) than that obtained with single linkage, and the outcomes of the two linkage methods are comparable only when more than half of the frames are retained. Results from dimensionality reduction also support the observation that the average linkage identifies more informative LR representations than those produced by single linkage in that, coherently with the trend of MSR, the PCC value is consistently higher in the former case than in the latter.

In the three panels of Fig. 3 the yellow stars indicate the representations that, for each method, maximise the relevance. Interestingly, in both the graph of covariance matrices and of the PCC, these representations are at the elbow of the curve. Further analyses in support of this interpretation, specifically of the location of the point of slope $\mu = 1$ and of the inter-cluster

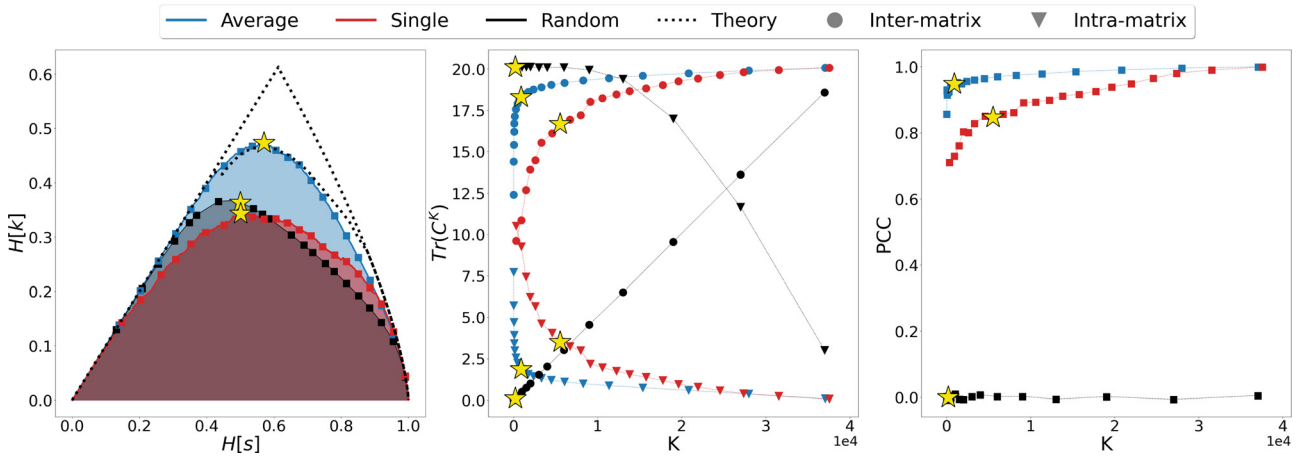

Fig. 3 Comparison of the different LR representations obtained from clustering the MD trajectory of adenylate kinase using the average linkage (blue), single linkage (red), and random clustering (black). The system was analysed using a coarse representation in which only C_β atoms are considered. Left panel: Relevance-resolution curves drawn by the different protocols; the square points mark the 22 low-resolution representations under examination. The theoretical limits of the maximal relevance curve are also shown as dotted lines for comparison. Middle panel: Traces of the inter-state (circles) and intra-state (triangles) correlation matrix plotted against the number of clusters K. Right panel: Pearson correlation coefficient (PCC) between the first two DCs in the HR and LR representation, plotted against the number of clusters K. The yellow star in each of the graphs indicates the representation that maximises the relevance for the corresponding clustering method (K = {901, 5511, 200} for average, single or random protocol respectively).

distance as a function of the cluster number, are provided as ESI.†

The observed behaviour is thus suggestive of the fact that further increases in resolution lead to an increased model complexity that is no longer balanced by information gain: the tradeoff between complexity and informativeness turns in favour of the former, consistently with the interpretation of relevance as a measure of useful information content. The resolution-relevance framework allows for the identification of optimally chosen representative configurations of the dataset; any observable can be computed on these configurations, provided that their statistical weight, proportional to the size of the corresponding cluster, has been accounted for. This results in a significant gain in computational cost: in fact, assuming a linear scaling with respect to the size of the dataset, for the computation of the observable in question there will be a gain proportional to the ratio of the number of data and that of the clusters; in the case of the representation identified by average linkage, this ratio is $\sim$50. We note, in passing, that the aforementioned strategy is insensitive to any temporal ordering of the data points, if any; hence, in order to perform any measure that relies on such order (*e.g.* diffusion and transport coefficients) one has to explicitly include this information in the clustering procedure, specifically associating a time stamp to the representative frames and keeping track of the identity of particles.

To gain further insight in the statistical significance of these results, we compared the data obtained for average and single linkage with those of the random clustering. The latter has a very close relevance curve to that of single linkage, and the MSR values associated with single linkage (MSR_S = 0.235) and random clustering (MSR_R = 0.233) differ only at the third decimal place; in spite of that, the usefulness of the partitions obtained with an information-driven protocol is incomparably greater than that returned by random clustering: the trace of the inter-cluster covariance matrix of the latter is always lower than the intra-cluster one until we consider representations in which about 2/3 of the original frames are preserved, and the PCC between reference and random partition DC is almost zero at any level of resolution. These observations further support the idea that the relevance alone cannot be taken as an absolute measure of the informativeness of a given low-resolution representation, however this quantity in combination with the appropriate classification method proves extremely effective in identifying protocols that maximise the emergence of useful information.

We then looked in detail at the three representations that maximise the relevance for each of the clustering methods under examination. In the right-hand side of Fig. 4 the distributions of centroids in LR representations are compared with the frame distributions in the HR ones, as the points in each panel are coloured according to the value taken from the first DC in the LR representation. A visual inspection of these data shows that the distribution of average linkage centroids in the LR DC space is consistent with that of the HR frames; in both graphs it is possible to recognise a colour gradient along the x axis, showing that neighbouring frames in the HR space are grouped together in the LR space. As for the linkage criterion, the LR representation maximising the relevance produces a slightly different distribution of points than that of the HR frames; furthermore, looking at the colour of points in both spaces it appears that distant frames in the diffusion space are associated to the same cluster. This is even more evident when correlating the values assumed by the DCs in the HR and LR representation, as shown in the bottom-left corner of Fig. 4. For both linkage measures (average and single) it is possible to identify a strong correlation between the first DCs in LR and HR: the Pearson correlation coefficient is 0.95 for

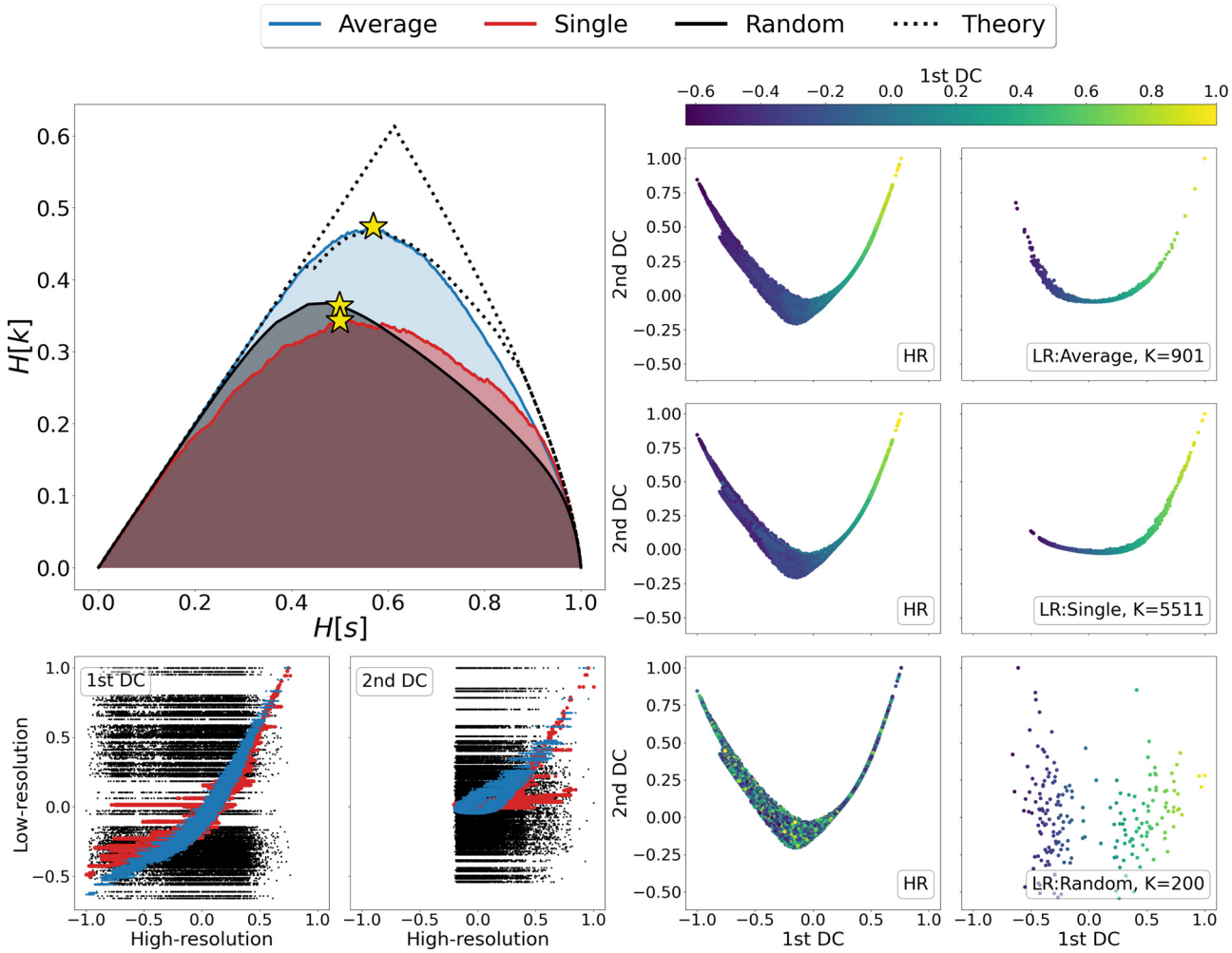

Fig. 4 Detailed analysis of the highest-relevance partitions of adenylate kinase trajectory data obtained with single linkage, average linkage, and random clustering. Upper left-hand panel: Relevance-resolution curves obtained from clustering the MD trajectory of adenylate kinase *via* average linkage (blue), single linkage (red), and random clustering (black). The system was analysed using a coarse-grained representation in which only the C_β atoms are considered. The yellow stars indicate the position, on the corresponding curve, of the low-resolution representations maximising the relevance obtained by partitioning the system into K clusters. The dotted lines show the theoretical upper and lower bounds to the maximal relevance curve. Right panel: Trajectory frames or cluster points projected onto the space spanned by the first two diffusion coordinates obtained by diffusion maps. The panel's left column shows the low-dimensional manifold resulting from high-resolution (HR) representation where each point is a frame of the MD simulation; the right column shows the two-dimensional manifold resulting from the low-resolution (LR) representation, where each point is the centroid of a cluster. All HR diffusion maps were calculated using the RMSD between configurations as a distance. For the LR manifold, instead, the distance between points is the linkage measure that produced the partition for the information-driven clustering (average linkage in the first row and single linkage in the second row), or the RMSD between the clusters centroid for the random clustering (third row). In both HR and LR spaces, the points are coloured according to the value taken by the first diffusion coordinate in the LR space as reported in the colour bar at the top. Lower left-hand panel: Scatter plot of the first (left) and second (right) diffusion coordinates of the LR space plotted against the corresponding coordinates of the HR space; in each graph, we report the points of the representations obtained through random clustering (black), single linkage (red), and average linkage (blue). Note that the compared LR representations display close resolution values ($H[s]$) but significantly different numbers of clusters (K).

average linkage and 0.85 for single linkage; nevertheless, in the case of single linkage, some clusters contain frames with a wide distribution of HR diffusion coordinate values, *i.e.*, frames carrying very different information are mistakenly lumped in the same bin. Last but not least, we observe, as expected, a total lack of correlation – both in terms of point distribution and cluster composition – between the DCs of the random partition and the reference HR DCs.

Finally, a detailed analysis of the relevance-resolution curve, in the upper left corner of Fig. 4, shows that the relevance obtained from average linkage is significantly higher than that obtained from single linkage; additionally, it is comparable with its theoretical maximum. Indeed, the average linkage curve lies close to the area enclosed by the upper and lower limits to the theoretical maximum of the relevance.[12,13] Since relevance proved capable of capturing the informativeness of representations, one might think of modifying the clustering outcome by shifting points between neighboring groups to further increase the relevance value. In this respect, the results obtained through average linkage would represent an ideal

starting point; indeed, being already so close to the theoretical maximum, it would allow to further increase the relevance with perturbative changes that, while violating the rules of the clustering, would preserve its general structure.

4 Conclusions

The steady increase of available computational power offers impressive opportunities in the investigation of biological macromolecules; at the same time, the corresponding growth of the pile of data produced by *in silico* studies requires the application of coarse-graining and dimensionality reduction techniques that allow one to discriminate between signal and noise in the dataset, and extract simple, useful, and intelligible information out of it. In fact, whatever the aim of the analysis of a dataset, it is of ever growing importance to have access to a synthetic representation of it, that retains the salient *general* features while reducing the weight of the unimportant details; this compression allows at a time the storage of a smaller amount of data and a faster analysis of the latter, as well as the extraction of global properties of the system in the process, as we have shown in this work.

To this end, the resolution-relevance framework represents a novel, powerful instrument to construct informative simplified representations of a molecule's conformational space; however, a blind and black-box application of this approach bears the risk of giving high-relevance partitions more credit than they deserve, in that the quality of said partitions cannot be disentangled by the specific classification method employed to construct them.

In the present work we have tackled this issue through the systematic, dataset-wide application of the resolution-relevance framework to a number of structurally distinct proteins, making use of state-of-the-art agglomerative clustering methods. Our results show that the clustering strategies, and more specifically the particular definitions of inter-cluster distance, employed to group together "similar" frames into structurally homogeneous clusters return different values of the multi-scale relevance, a global measure of the relevance at various levels of resolution. We find that the partitions having higher values of the MSR are those that produce the most physically sensible partitions, as quantified in terms of intra- and inter-cluster covariance, as well as the correlation between the collective diffusion coordinates computed in the reference, high-resolution space and those of the low-resolution representation. Most interestingly, a positive correlation emerges between high values of MSR and the efficacy of a clustering method in reconstructing a low-resolution representation that features an ultrametric structure: this observation is suggestive of the fact that the configurational space spanned by a protein in the course of a molecular dynamics simulation is intrinsically organised in a hierarchical manner, which is consistent with the hypothesis, proposed and verified in the literature, that the free energy landscape of proteins is effectively self-similar.

In conclusion, we propose that the clustering method employed in the dimensionality reduction of a dataset could be not only employed as a tool to preprocess the data *in order to analyse them*, but also treated as an analysis tool itself: in fact, through the joint usage with the general, parameter-free resolution-relevance framework it is possible to discriminate among partitioning approaches that produce low-resolution models more or less representative of the salient qualities of the high-resolution reference. The combination of these algorithms can thus pave the way to an even more fruitful deployment of clustering approaches in computational biophysics, bringing further insight in the behaviour of complex macromolecules.

5 Methods

5.1 Protein selection

For the exploratory analysis it was essential to employ a set of structurally distinct and uncorrelated proteins, in order to draw general conclusions. To this end, a dataset of 107 proteins, including many of the known folds and structure classes, was constructed and clustered based on their dynamics.[37] For each protein, the first 10 normal modes of fluctuation were analysed using an elastic network model,[38] and superimposed by means of the ALADYN[39] protocol, which performs a hybrid structural/dynamical alignment. The similarity between the essential spaces spanned by the first 10 normal modes was quantified by means of the root mean square inner product[40] (RMSIP). The distance between the essential dynamics of two aligned proteins was defined as $d_{ij} = 1 - \mathrm{RMSIP}_{ij}$; this distance was employed to perform a hierarchical clustering. The resulting dendrogram allowed us to identify 12 clusters, each of which contains proteins whose dynamics are similar (RMSIP above 0.5). The 12 proteins used in this work are the centroids of these clusters, and their PDB codes are: 1DSL, 1NOA, 1SNO, 1UNE, 1XWL, 1IGD, 1HYP, 2FGF, 1KNT, 1QKE, 2EXO, 1KOE.

Two specific proteins were used for the second part of the analysis. The protein adenylate kinase (PDB code AKE4) because of its relatively small size and the possibility to observe conformational transitions over time scales easily achievable by means of plain MD. The second system is the humanised IgG4 monoclonal antibody (PDB code 5DK3). This system was chosen because of its large size and higher structural and dynamical complexity.[41] As the results obtained in the two cases are consistent, for the sake of clarity we only reported the data pertaining the adenylate kinase in the main text, while those of the antibody are provided as ESI.†

5.2 Simulation setup

For all simulations, the Gromacs 2018[42,43] software was employed, and the topology was defined through the AMBER99SB-ILDN[44] force field. The simulations were performed in explicit solvent, the latter being TIP3P water;[45] Na^+ and Cl^- ions at the concentration of 0.15 M were added to neutralise the global electric charge and mimic physiological ion concentration in the cell. Energy minimisation was performed until the maximum force reached a specific value, F_{min} = 1000 kJ mol^{-1} nm^{-1} for the 12-protein dataset and F_{min} = 500 kJ mol^{-1} nm^{-1} in case of the adenylate

kinase protein and the humanised IgG4 monoclonal antibody. *NVT* and *NPT* equilibrations were performed using the velocity-rescale thermostat[46] and the Parrinello–Rahman barostat.[47] With respect to the interaction, a cut-off was used for van der Waals interaction and for the short-range component of the Coulomb one. The long-range component of the Coulomb force, instead, was computed with the Particle Mesh Ewald algorithm. The LINCS algorithm[48] was employed to define the constraints on the hydrogen-containing bonds and allows an integration time of 2 fs. The dynamic of each system in the 12-protein dataset has been simulated for 300 ns; regarding the adenylate kinase and the humanised IgG4 monoclonal antibody their dynamic were simulated for 800 ns and 2 ms, respectively.

5.3 Clustering

For each protein the RMSD matrix was computed[49] for three atom selection: all atoms, C_α only, C_β only. From each of these matrices, 7 dendrograms were constructed exploring the 7 different definitions of inter-cluster distance supported by the python module Syipy,[50] employed for the clustering analysis. Each dendrogram was cut at different levels, ranging the number of clusters from 1 to the number of frames M by steps of 10. For each resulting partition, the corresponding resolution and relevance values were computed following eqn (1) and (2). In this way, for each system 21 different curves are obtained: 3 representations of the system (all, C_α, C_β) times 7 measures of distance between clusters.

5.4 Multi-scale relevance and random curves

In addition to information-driven clustering, random clustering is also possible. Given a number of clusters K, a label vector can be randomly generated by iteratively sampling M elements from a list containing all integers from 1 to K. For a given K, 10^4 vectors of labels were generated, and for each of them the corresponding values of relevance and resolution were calculated. The points on the random curve were obtained by averaging the relevance and resolution values obtained for a given number of clusters K, and varying the number of clusters from 1 to M by steps of 10. The MSR value associated to this curve was used to normalise the MSRs resulting from the hierarchical clustering algorithms:

$$\overline{\mathrm{MSR}}_\mathrm{i} = \frac{\mathrm{MSR}_\mathrm{i} - \mathrm{MSR}_\mathrm{R}}{\mathrm{MSR}_\mathrm{R}} \tag{4}$$

where MSR_i is the area under the relevance-resolution curve drawn by algorithm type i, and MSR_R is the one obtained by the random procedure.

5.5 Diffusion maps

Diffusion Maps is a nonlinear dimensionality reduction method.[34–36] The algorithm was employed to compare the manifold obtained from the whole trajectory (high-resolution representation) and those obtained considering only the centroids of some partitions (low-resolution representation). In the high-resolution representation, the inter-frame distance is the RMSD matrix, as for the clustering. In the low-resolution one, each cluster is represented by its centroid and the distance between clusters is given by the linkage measure adopted. The terms high-resolution and low-resolution are used here in connection with the number of frames retained and not to the selection of atoms, which is the C_β atom selection in both scenarios. The algorithm also requires a threshold parameter to determine what is near or far in the source data set. To be consistent in its approach, the quantile of order 0.1 of the distribution of distances was always chosen. It is possible to relate the points of the HR space to those of the LR one, since on one side there are the elements of the clusters, and on the other the centroids. To make this visual inspection easier, data were rescaled so that all points were contained in the square $[-1,1]^2$.

5.6 Covariance matrix

The covariance matrix of the positions of the C_β atoms along the trajectory can be subdivided in two contributions: inter- and intra-clusters.[29,51,52] The correlation between two elements of the system (in this case two atoms C_β) is given by:

$$C_\mathrm{ij} = C_\mathrm{ij}^\mathrm{intra} + C_\mathrm{ij}^\mathrm{inter} \tag{5}$$

$$C_\mathrm{ij}^\mathrm{intra} = \sum_l \omega_l \left\langle \left[\vec{r}_i - \langle \vec{r}_i \rangle_l\right]\left[\vec{r}_j - \langle \vec{r}_j \rangle_l\right]\right\rangle_l \tag{6}$$

$$C_\mathrm{ij}^\mathrm{inter} = \sum_l \omega_l \left[\langle \vec{r}_i \rangle_l - \langle \vec{r}_i \rangle\right]\left[\langle \vec{r}_j \rangle_l - \langle \vec{r}_j \rangle\right] \tag{7}$$

where l runs over a cluster C_l; ω_l is the weight of the state l, which is the fraction of simulation time spent by the system in it; $\langle\rangle_l$ denotes the average over the conformations belonging to the cluster l. The decomposition is performed in analogy with the "jumping among minima" model.[52] The first term in the decomposition of the covariance matrix is the contribution arising from structural fluctuations within clusters. The second term arises from the structural differences of the clusters centroids. Consequently, a good partition will have a high inter-cluster term and a low intra-cluster one.

Data availability

The raw data produced and analysed in this work are freely available on the Zenodo repository with DOI: **10.5281/zenodo.6554498**.

Author contributions

RP and RC conceptualised and supervised the study. MM developed the methodology, performed the simulations, wrote the analysis software and the original draft. MM, RP and RC wrote, reviewed, and edited the final manuscript.

Conflicts of interest

There are no conflicts to declare.

Appendix

• Single linkage: the distance between a pair of clusters is determined by the two closest objects belonging to the different clusters.

$$D(C_i,C_j) = \min\{d(x_i,x_j) \text{ for } x_i \in C_i \text{ and } x_j \in C_j\} \quad (8)$$

Single linkage clustering tends to produce elongated clusters, which causes the chaining effect. Two points that form a bridge between two clusters cause the single-link clustering to join these two clusters into one.

• Complete linkage: it consider the distance between two clusters to be equal to the largest distance from any member of one cluster to any member of the other cluster.

$$D(C_i,C_j) = \max\{d(x_i,x_j) \text{ for } x_i \in C_i \text{ and } x_j \in C_j\} \quad (9)$$

This procedure tends to form smaller and more compact clusters.

• Average linkage: it considers the distance between two clusters as the average distance between all pairs of points coming from the different groups.

$$D(C_i, C_j) = \frac{1}{|C_i||C_j|}\sum_{\mathbf{x}_i\in C_i}\sum_{\mathbf{x}_j\in C_j} d(\mathbf{x}_i, \mathbf{x}_j) \quad (10)$$

where $|.|$ stands for the cardinality of set, *i.e.* the number of items pertaining to it. This approach can cause the splitting of elongated clusters and the merging of portions of neighbouring elongated clusters.

• Weighted linkage: also in this case, the protocol takes as cluster distance the average distance from any member of one cluster to any member of the other one. The difference is that the distance between the new cluster and another is weighted with respect to the number of data in each cluster. Consequently, the distance between the cluster $C_k = C_i \cup C_j$ and a third cluster C_l, not involved in the definition of C_k, is:

$$\begin{aligned} D(C_k, C_l) &= \frac{1}{2|C_i||C_l|}\sum_{\mathbf{x}_i\in C_i}\sum_{\mathbf{x}_l\in C_l} d(x_i, x_l) + \\ &+\frac{1}{2|C_j||C_l|}\sum_{\mathbf{x}_j\in C_j}\sum_{\mathbf{x}_l\in C_l} d(x_j, x_l) \end{aligned} \quad (11)$$

• Centroid linkage: in this case, two clusters are merged based on the distance of their centroids. The definition of centroids is:

$$\mu_i = \frac{1}{|C_i|}\sum_{\mathbf{x}_i\in C_i} x_i \quad (12)$$

Consequently, the distance between clusters results the Euclidean distance between the centroids:

$$D\,(C_i,C_j) = \|\mu_i - \mu_j\|_2 \quad (13)$$

The centroid of the resulting cluster $C_k = C_i \cup C_j$ is recomputed according to eqn (12) considering all the points belonging to it.

• Median linkage: the procedure is similar to the centroid linkage, except that the centroid of the resulting cluster μ_k is the average of the centroid of the merged ones:

$$\mu_k = \frac{1}{2}(\mu_i + \mu_k) \quad (14)$$

This is equivalent to giving the same weight to merged clusters regardless of the number of elements in them.

• Ward linkage: the methods aims to minimise the increase of the intra-cluster sum of squared errors:

$$E = \sum_{k=1}^{K}\sum_{\mathbf{x}_i\in C_k} \|x_i - \mu_k\|_2^2 \quad (15)$$

where K is the number of clusters and μ_k is the centroid of the cluster C_k (defined by eqn (12)). Merging clusters C_i and C_j produces an increase in variance of

$$\Delta E = \frac{n_i n_j}{n_i + n_j}\left\|\mu_i - \mu_j\right\|_2^2. \quad (16)$$

Consequently, the distance between the new cluster $C_k = C_i \cup C_j$ and an unused cluster C_l is given by the recursive equation:

$$\begin{aligned} D(C_l, C_k) &= \frac{|C_i| + |C_l|}{|C_i| + |C_j| + |C_l|}D(C_l, C_i) + \\ &+\frac{|C_j| + |C_l|}{|C_i| + |C_j| + |C_l|}D(C_l, C_j) + \\ &-\frac{|C_l|}{(|C_i| + |C_j|)^2}D(C_i, C_j) \end{aligned} \quad (17)$$

All definitions of distance between clusters can be summarised by the recursive relation proposed by Lance and Williams:[53]

$$\begin{aligned} D(C_l,C_i\cup C_j) = \alpha_i D(C_l,C_i) + \alpha_j D(C_l,C_j) + \beta D(C_i,C_i) \\ + \gamma|D(C_l,C_i) - D(C_l,C_j)| \end{aligned} \quad (18)$$

where α_i, α_j, β and γ are coefficients that takes different values depending on the protocol used.

Acknowledgements

The authors are indebted with Thomas Tarenzi for valuable help in the construction of the protein dataset, and with Luca Tubiana for a critical and insightful reading of the manuscript. R.C. acknowledges the support of the Frankfurt Institute for Advanced Studies and the LOEWE Center for Multiscale Modelling in Life Sciences of the state of Hesse. This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant 758588).

References

1 M. Karplus and G. A. Petsko, *Nature*, 1990, **347**, 631–639.
2 M. González, *École thématique de la Société Française de la Neutronique*, 2011, **12**, 169–200.

3 A. C. Pan, T. M. Weinreich, S. Piana and D. E. Shaw, *J. Chem. Theory Comput.*, 2016, **12**, 1360–1367.
4 S. A. Adcock and J. A. McCammon, *Chem. Rev.*, 2006, **106**, 1589–1615.
5 G. A. Tribello and P. Gasparotto, *Front. Mol. Biosci.*, 2019, **6**, 46.
6 A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé and A. Laio, *Chem. Rev.*, 2021, **121**, 9722–9758.
7 F. Noé and C. Clementi, *Curr. Opin. Struct. Biol.*, 2017, **43**, 141–147.
8 A. Glielmo, C. Zeni, B. Cheng, G. Csanyi and A. Laio, arXiv preprint arXiv:2104.15079, 2021, 8.
9 C. Battistin, B. Dunn and Y. Roudi, *Curr. Opin. Syst. Biol.*, 2017, **1**, 122–128.
10 M. Marsili and Y. Roudi, *Phys. Rep.*, 2022, **963**, 1–43.
11 S. Grigolon, S. Franz and M. Marsili, *Mol. BioSyst.*, 2016, **12**, 2147–2158.
12 M. Marsili, I. Mastromatteo and Y. Roudi, *J. Stat. Mech.: Theory Exp.*, 2013, **2013**, P09003.
13 A. Haimovici and M. Marsili, *J. Stat. Mech.: Theory Exp.*, 2015, **2015**, P10013.
14 J. Song, M. Marsili and J. Jo, *J. Stat. Mech.: Theory Exp.*, 2018, **2018**, 123406.
15 R. J. Cubero, M. Marsili and Y. Roudi, *Entropy*, 2018, **20**, 755.
16 R. J. Cubero, J. Jo, M. Marsili, Y. Roudi and J. Song, *J. Stat. Mech.: Theory Exp.*, 2019, **2019**, 063402.
17 G. K. Zipf, *Selected studies of the principle of relative frequency in language*, Harvard university press, 2013.
18 G. Tkačik, T. Mora, O. Marre, D. Amodei, S. E. Palmer, M. J. Berry and W. Bialek, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 11508–11513.
19 J. Tyrcha, Y. Roudi, M. Marsili and J. Hertz, *J. Stat. Mech.: Theory Exp.*, 2013, **2013**, P03005.
20 D. J. Schwab, I. Nemenman and P. Mehta, *Phys. Rev. Lett.*, 2014, **113**, 068102.
21 L. Aitchison, N. Corradi and P. E. Latham, *PLoS Comput. Biol.*, 2016, **12**, e1005110.
22 M. I. Ionescu, *Proteins*, 2019, **38**, 120–133.
23 E. Formoso, V. Limongelli and M. Parrinello, *Sci. Rep.*, 2015, **5**, 1–8.
24 R. J. Cubero, M. Marsili and Y. Roudi, *J. Comput. Neurosci.*, 2020, **48**, 85–102.
25 G. W. Milligan, *Psychometrika*, 1979, **44**, 343–346.
26 S. C. Johnson, *Psychometrika*, 1967, **32**, 241–254.
27 N. Jardine and R. Sibson, *Math. Biosci.*, 1968, **2**, 465–482.
28 H. Fushing, H. Wang, K. van der Waals, B. McCowan and P. Koehl, *PLoS One*, 2013, **8**, e56259.
29 F. Pontiggia, G. Colombo, C. Micheletti and H. Orland, *Phys. Rev. Lett.*, 2007, **98**, 048102.
30 A. Volkhardt and H. Grubmüller, *Phys. Rev. E*, 2022, **105**, 044404.
31 M. J. Pandya, S. Schiffers, A. M. Hounslow, N. J. Baxter and M. P. Williamson, *Front. Mol. Biosci.*, 2018, **5**, 115.
32 K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus and D. Kern, *Nature*, 2007, **450**, 913–916.
33 K. Henzler-Wildman and D. Kern, *Nature*, 2007, **450**, 964–972.
34 J. De la Porte, B. Herbst, W. Hereman and S. Van Der Walt, Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa, 2008, 15-25.
35 S. Lafon and A. B. Lee, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, 1393–1403.
36 B. Nadler, S. Lafon, R. R. Coifman and I. G. Kevrekidis, *Appl. Comput. Harmon. Anal.*, 2006, **21**, 113–127.
37 T. Tarenzi, G. Mattiotti, M. Rigoli and R. Potestio, *Appl. Sci.*, 2022, **12**, 7157.
38 C. Micheletti, P. Carloni and A. Maritan, *Proteins: Struct., Funct., Bioinf.*, 2004, **55**, 635–645.
39 R. Potestio, T. Aleksiev, F. Pontiggia, S. Cozzini and C. Micheletti, *Nucleic Acids Res.*, 2010, **38**, W41–W45.
40 A. Amadei, M. A. Ceruso and A. Di Nola, *Proteins: Struct., Funct., Bioinf.*, 1999, **36**, 419–424.
41 T. Tarenzi, M. Rigoli and R. Potestio, *Sci. Rep.*, 2021, **11**, 1–12.
42 H. Bekker, H. Berendsen, E. Dijkstra, S. Achterop, R. Vondrumen, D. Vanderspoel, A. Sijbers, H. Keegstra and M. Renardus, 4th International Conference on Computational Physics (PC 92), 1993, pp. 252–256.
43 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1**, 19–25.
44 K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins: Struct., Funct., Bioinf.*, 2010, **78**, 1950–1958.
45 R. W. Hockney and J. W. Eastwood, *Computer simulation using particles*, CRC Press, 2021.
46 G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.
47 M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182–7190.
48 B. Hess, H. Bekker, H. Berendsen and J. Fraaije, LINCS: A linear constraint solver for molecular simulations, *J. Comput. Chem.*, 1997, **18**, 1463–1472.
49 R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane and V. S. Pande, *Biophys. J.*, 2015, **109**, 1528–1532.
50 J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow and P. Barth, *et al.*, *Nat. Methods*, 2020, **17**, 665–680.
51 F. Pontiggia, A. Zen and C. Micheletti, *Biophys. J.*, 2008, **95**, 5901–5912.
52 A. Kitao, S. Hayward and N. Go, *Proteins: Struct., Funct., Bioinf.*, 1998, **33**, 496–517.
53 G. N. Lance and W. T. Williams, *Comput. J.*, 1967, **9**, 373–380.

