JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

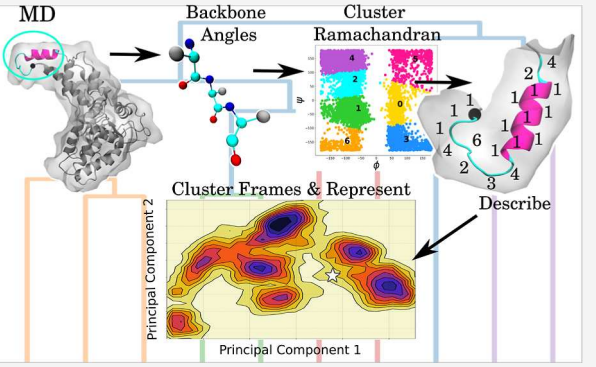
Searching for Structure: Characterizing the Protein Conformational Landscape with Clustering-Based Algorithms

Amanda C. Macke, Jacob E. Stump, Maria S. Kelly, Jamie Rowley, Vageesha Herath, Sarah Mullen, and Ruxandra I. Dima*

Cite This: J. Chem. Inf. Model. 2024, 64, 470−482 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The identification and characterization of the main conformations from a protein population are a challenging and inherently high-dimensional problem. Here, we evaluate the performance of the Secondary sTructural Ensembles with machine LeArning (StELa) double-clustering method, which clusters protein structures based on the relationship between the φ and ψ dihedral angles in a protein backbone and the secondary structure of the protein, thus focusing on the local properties of protein structures. The classification of states as vectors composed of the clusters' indices arising naturally from the Ramachandran plot is followed by the hierarchical clustering of the vectors to allow for the identification of the main features of the corresponding free energy landscape (FEL). We compare the performance of StELa with the established root-mean-squared-deviation (RMSD)-based clustering algorithm, which focuses on global properties of protein structures and with Combinatorial Averaged Transient Structure (CATS), the combinatorial averaged transient structure clustering method based on distributions of the φ and ψ dihedral angle coordinates. Using ensembles of conformations from molecular dynamics simulations of intrinsically disordered proteins (IDPs) of various lengths (tau protein fragments) or short fragments from a globular protein, we show that StELa is the clustering method that identifies many of the minima and relevant energy states around the minima from the corresponding FELs. In contrast, the RMSD-based algorithm yields a large number of clusters that usually cover most of the FEL, thus being unable to distinguish between states, while CATS does not sample well the FELs for long IDPs and fragments from globular proteins.



INTRODUCTION

In many enzymes, conformational changes can occur due to the binding of ligands, other proteins, or protein mutations. These are fundamental processes with important functional implications.[1] Such conformational changes sometimes correspond to a global rearrangement of the protein. This is observed in the quaternary structure of the microtubule (MT) tubulin dimers where the straight guanosine 5′-triphosphate (GTP)-bound tubulin becomes curved when the hydrolyzable GTP is chemically converted into guanosine 5′-diphosphate.[2−4] Other times, they correspond to local changes at the secondary level or to a transition from ordered to disordered (or vice versa) as seen in MT-severing enzymes when the substrate-binding loops become ordered in the presence of the MT substrate.[1,5,6] Great effort has been expended to predict the steps that lead to both global and local levels of conformational changes and the corresponding states. Experimental methods, such as X-ray crystallography and cryogenic electron microscopy (cryo-EM), are useful tools for identifying conformational transitions by solving structures in the presence or absence of cofactors of interest.[7] Computational tools, such as the bioinformatics-based PSIPRED and the AI-based AlphaFold, are employed to predict protein structures from protein sequences using machine learning.[8−12] However, all of these methods struggle to characterize regions or entire proteins that are "fuzzy" or that are heterogeneous due to a lack of order. They also have difficulty capturing structural transitions over time.

All-atomistic computational tools, such as molecular dynamics (MD) simulations, allow for a more detailed exploration of the conformational landscape. The challenge then becomes the extraction of the most dominant sampled states from a large and complex data set, particularly when characterizing how environments or allosteric modulators affect the conformational space. The aforementioned proteins that lack a well-defined structure are known as intrinsically disordered proteins (IDPs). Despite the usage of the term

Received: September 19, 2023
Revised: December 22, 2023
Accepted: December 22, 2023
Published: January 4, 2024

"disorder", IDPs do not sample their conformational space in a completely random manner.[13] An IDP favors certain conformations over others, albeit to a much less extreme degree than structured globular proteins.[14] IDPs sample a large number of diverse conformations while carrying out their functions in the cell, in apparent contrast to the traditional "structure−function" paradigm.[15−17] A well-known example of an IDP is the MT-stabilizing protein found primarily in neurons called tau.[18] Tau is made up of an N-terminal projection domain, a proline-rich domain, a combination of repeat domains (R1−R4 and a pseudorepeat R′), and a C-terminal domain that are spliced together in various isoforms. The misregulation of tau is associated with neurodegenerative diseases such as Alzheimer's disease where tau is found to aggregate as a result of a change to its conformational space.[18−22] Describing these changes is an essential piece for understanding neurodegenerative diseases, and therefore MD is commonly employed to probe the underlying conformational space of IDPs.[17,20,23]

IDPs are not the only types of proteins characterized by a large or changing conformational space that could be challenging to explore. Regions of proteins that experience conformational changes due to the allosteric influence of a modulator are also difficult to characterize. These types of transitions have been captured with the above-mentioned experimental methods, as well as with computational tools. In our previous work, using all-atomistic MD simulations to study lower-order oligomers of the MT-severing enzyme katanin, we identified a ligand-dependent conformational transition in a region of the protein from a loop−helix to a helix−helix structure.[24] Results from PSIPRED predicted only a structure resembling that reported in the cryo-EM structure (loop−helix).[8,24] As the helix−helix structure was also proposed based on the X-ray structure of the monomeric form of katanin,[25] our finding of the conformational switch in simulations was essential for understanding the allosteric contribution of the binding cofactors in katanin and for assessing the stability of the lower-order oligomers. Additional transient and flexible structures were identified in dimers and trimers in the presence and absence of binding ligands. MD simulations were also used to identify a nucleotide-dependent conformational transition in secondary structure from a loop in the G-actin filament to a helix in the F-actin filament.[26−28] Another classic example of extensive ligand-induced allosteric control over a quaternary protein structure is hemoglobin.[29] A single-point mutation (E6V) in the β chains of hemoglobin leads to sickle cell anemia.[30,31] This mutation causes the aggregation of hemoglobin fibrils in the absence of O_2 (T) that leads to the deformation of red blood cells; however, when O_2 is present (R), the fibrils disintegrate.[31,32] Having a means of characterizing the conformational landscape is crucial for identifying states for these more heterogeneous proteins/regions, which are essential for understanding enzymatic mechanisms or the effects of allosteric modulators and of disease-related mutations.

Unsupervised machine-learning methods such as clustering algorithms have been used to extract significantly populated states from the sampled space by grouping together structures identified as similar according to a given criterion. Previous studies have emphasized the impact that the choice of the clustering methodology has on the predicted states based on both the size and intrinsic properties of the system being characterized.[33,34] For the identification of more global changes, approaches such as root-mean-squared-deviation (RMSD)-based algorithms that determine similarity using an RMSD cutoff have been found to be particularly effective. Structures within this cutoff are grouped together, whereas a neighbor outside of the cutoff is assigned to a distinct group.[14] While additional similarity measurements have been applied to RMSD-based clustering to explain localized changes, such as within a ligand-binding pocket, studies found that the RMSD criterion produces larger errors in structure separation, which would result in inaccurate characterization of more subtle conformational changes.[35,36] To address these more subtle changes, Cheung and Ezerski focused on the identification of local structures with similar patterns, especially in IDPs, when introducing a new algorithm, Combinatorial Averaged Transient Structure (CATS), which first characterizes structures based on an internal collective variable before determining similarity.[37] Inspired by their use of the descriptive internal coordinate, we developed our in-house double-clustering algorithm, StELa (Secondary sTructural Ensembles with machine LeArning), to probe for a secondary structural transition in a local heterogeneous region of severing enzymes referred to as the helical bundle domain tip ("HBD tip") in our previously mentioned study of oligomeric species from MT severing enzymes.[24] Here, we put our algorithm through rigorous testing to map its performance on a number of protein systems. Namely, we performed clustering of the states from the sampled conformational space of R2/Tau fragments, R4−R′/Tau fragments, and the HBD tip fragments, all taken from the corresponding MD simulations, using each of the three algorithms: RMSD-based, CATS, and StELa.[14,20,24,37] We evaluated the differences in the free energy landscapes (FELs) between short and long fragments of IDPs (R2/Tau and R4−R′/Tau), as well as between IDPs and the transient, but ordered, region from a globular protein in a tertiary and quaternary assembly (HBD tip of each of the katanin protomers). Mapping out the clusters of states on the FEL in the principal component (PC) space, which is the most general illustration of the energy space of a protein,[38,39] allowed us to evaluate the ability of each method to sample the conformational space. We conclude with a comparison of the advantages and pitfalls of each of the methods.

METHODS

MD Simulations. To address how the three clustering methods discussed in the Introduction handle different systems, we used MD simulations for a short fragment of an IDP (R2/Tau),[20,37] a longer fragment of an IDP (R4−R′/Tau), a short fragment from a tertiary globular protein structure (katanin), and the same fragment from its quaternary form (ABC/katanin).[24] The MD simulations for the R2 fragment of tau (14 units), computed with the GROMACS 4.6.1, detailed in Table S1, were from an earlier paper.[17,20] In that study, the authors tested how different solvents induce the formation of aggregation-prone states of R2/Tau: urea to mimic a denatured state, water for a more standard state, and TMAO for a more aggregation-prone state.[20,40−42] The models used for TMAO (2 M) and urea (5 M) were previously developed by Weerasinghe and Smith and Larini and Shea.[43,44] For their simulations, the authors used the all-atom-optimized potentials for liquid simulation (OPLS) force field with three-site transferable intermolecular potential rigid water.[20,45,46] The RMSD for the provided simulations is shown in Figure S1, and the corresponding global averages are provided in Table

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

S2. The structural changes observed in these simulations, resulting from an altered conformational landscape, were characterized by the Cheung and Ezerski using two different clustering methods: the RMSD-based algorithm and their in-house developed algorithm, CATS.[37]

We carried out all-atom MD simulations for R4−R′/Tau (121 residues) using GROMACS 2022 and the AMBER99SB-ILDN force field, for which system details are reported in Table S3.[40−42,47] We created the structure used in the simulations by applying the Modeler plug-in from ChimeraX visualization software to the PDB 7PQC configuration (without the tubulin).[48−50] We then centered this structure in a dodecahedral box with periodic boundary conditions (PBCs), solvated with TIP3P water molecules, and neutralized with NaCl.[51] We performed energy minimization for 50,000 steps using the steepest descent algorithm and the Verlet cutoff scheme.[52] We did *NVT* equilibration for 500 ps, followed by *NPT* equilibration for an additional 500 ps. Both equilibrations employed the leapfrog integrator, Verlet cutoff scheme, and velocity-rescaling thermostat. The *NPT* equilibration used also the Parrinello−Rahman pressure coupling.[53,54] We ran four production trajectories, each 200 ns long, using the same temperature and pressure coupling methods as in the *NPT* equilibration. Coulombic interactions were treated using the Particle Mesh Ewald (PME) algorithm, while other nonbonded interactions were computed using the Verlet cutoff scheme.[52,55] Equilibration was monitored by using both RMSD and the DCCM convergence test, as shown in Figure S2.

The MD simulations used for the analysis of conformational behaviors of lower-order oligomers of katanin in our previous studies[24] were carried out with GROMACS 2019 using the GROMOS96 54a7 force field and cubic boxes with SPC solvent, as described in Table S3.[40,42,56−58] Utilizing PBC, these systems were minimized with the steepest descent algorithm and the Verlet cutoff-scheme for 50,000 steps with a criterion of the maximum force value less than 23.9006 kcal/mol/Å to account for steric clashes.[52] An *NVT* ensemble was used to bring the system to 300 K with the velocity-rescaling thermostat and the leapfrog integrator for 500 ps.[53] Then, the *NPT* ensemble with the leapfrog integrator and the Parrinello−Rahman pressure coupling scheme was used to keep the system at 1.0 bar for 500 ps.[54] The LINCS algorithm was used with an integration step of 2 fs.[59] The PME algorithm was used for the electrostatic interactions, and a cutoff of 10.0 Å was used to define nonbonded interactions.[55,56] The starting structures corresponded to select chains from the katanin spiral (PDB: 6UGD) and ring (PDB: 6UGE) hexamers modeled in our previous study.[56] The missing residues from disordered loops (residues 183−187 and 324−331) in the cryo-EM structure were modeled in our previous work with Modeler (version 9.23).[56,60] We note that the spiral conformation corresponds to the prehydrolysis assembly of katanin, characterized by a 40 Å gap between the terminal protomers (A and F) where adenosine 5′-triphosphate (ATP) is present in each of the six protomers. In turn, the ring conformation is a posthydrolysis state where the nucleotide is missing from protomer A resulting in the closing of the gap between the end protomers due to induced flexibility in its nucleotide-binding domain.[5,6] The automated topology builder server was used to parametrize the ATP molecules based on the GROMOS54a7 force field.[56,61]

In the study of lower-order oligomers, we probed monomers from the spiral conformation as well as dimers (AB and BC) and trimers (ABC) from the spiral and ring conformations of katanin, as described in Figure S3. To understand the effects of binding the ligands, the monomer was simulated with both the ATP and polyglutamate MT minimal substrate (COMPLEX), as found in the solved cryo-EM structure,[6] with ATP only (NUCLEOTIDE), with the minimal substrate (SUBSTRATE), and in the absence of both ligands (APO). The trimers were simulated in the COMPLEX and APO states. The RMSD over time for each of the described systems is shown in Figures S4 and S5 with their accompanying DCCM convergence tests, following eq S1, to show that the systems were appropriate for analysis. Global averages were calculated and are provided in Table S4 to assess overall stability. We observed that a region in the helical bundle domain, which we dubbed the "HBD tip" (see Figure S3), consisting of amino acids ASP417-LEU437 (21 peptide bonds), showed significant structural fluctuations in several of the setups. Our hot spot analysis determined that only three regions from katanin experience allosteric changes upon the binding of the ATP, the MT substrate, and the formation of the various interprotomer interfaces. The HBD tip was one of these regions.[24] We developed our in-house algorithm, StELa, to characterize such structural changes from the extracted HBD tip region conformations of each protomer in a simulation. Clustering with StELa showed that this region undergoes a structural change in the monomer upon the binding of ATP and the minimal MT substrate. In the ring ABC trimer, we found a similar structural change associated with the binding of cofactors to that seen in the monomer. However, this time, the concave interface made between the ith and the $(i + 1)$th protomers and/or the convex interface formed between the ith and the $(i - 1)$th protomers, described in Figure S3, also caused interesting changes to the conformational space of the HBD tip in the ith protomer (B).[5,6] In this oligomer, protomer A has only the concave interface, formed between both its nucleotide-binding domain (NBD) and HBD regions and the NBD of protomer B. Protomer C has only the convex interface, formed between its NBD region and the NBD of protomer B. Our analysis of the PC motions showed that protomer A is highly flexible due to its absence of the nucleotide, so much so that its NBD detached from the NBD of protomer B. We determined that, while protomers B and C would likely remain bound to each other, protomer A would dissociate from protomer B due to the motions leading to the detachment of its HBD from the NBD of protomer B. This indicated that the NBD−HBD contacts are more important for the stability of the interprotomer interface and the overall stability of the machine than the NBD−NBD interfaces. Importantly, the NBD−HBD contacts are formed with the long HBD helix and the loop/helix of the HBD tip making the region particularly important to characterize.[24]

Current Methods for Clustering Protein Fragments. The RMSD-based clustering algorithm[14] implemented with GROMACS[42] (**gmx cluster-gromos**) has long been considered the gold standard for evaluating the structures explored in MD trajectories. This algorithm uses the RMSD to evaluate how similar a data point, or structure, is to its neighbors within a specified cutoff.[14,37] We determined an appropriate cutoff for each system by carrying out clustering using a range of cutoffs dependent on the system. For example, we took increments of 0.01 between 0.1 and 0.25 for R2/Tau, whereas for the R4−R′/Tau system, we took values ranging from 0.14 to 0.50. The cutoff was chosen by evaluating the distribution of the RMSD

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

between the cluster center of the topmost populated clusters for each cutoff and all the frames.[62] The actual cutoff was selected based on the idea that a robust clustering should result in clusters for which all the cluster members are closely distributed around the cluster center, while any other frames belonging to other clusters are well separated from that center ($RMSD_{center}$). To this end, we chose cutoffs that resulted in well-defined first peaks of the RMSD distributions, followed by a decrease to as close to 0 as possible.[62] Examples of such distributions for the R4−R′/Tau and the katanin monomer-APO are shown in Figures S6 and S7. The chosen cutoffs for all systems are listed in Table S5. The RMSD-based clustering is based on the Cartesian coordinate space of the structure. For representation purposes in the FEL space, we used as cluster representatives the cluster centers, which are the structures with the largest number of neighbors. Cheung and Ezerski used this method as a control for testing the performance of their in-house algorithm, CATS.[37]

The CATS algorithm (https://github.com/Cheung-group/CATS) converts the structure found in each frame of a trajectory to a vector, prior to determining similarity using clustering. The authors chose the φ and ψ dihedral backbone angles as descriptive collective variables (CVs), instead of the Cartesian space-based CVs used in the RMSD-based algorithm.[37,63] This was significant, as the dihedral angles provide a detailed internal descriptor of the secondary structure of proteins and are known to capture conformational changes.[64] The user provides the φ and ψ backbone angles as the input, which we extracted using the **gmx rama** function in GROMACS.[42] Distributions of each residue's φ and ψ angles are fitted with Gaussian curves in order to identify one to three main peaks. Next, CATS creates representative vectors for each trajectory frame consisting of a series of integer labels corresponding to the Gaussian curve in which each angle was found, resulting in two labels per residue (Figure S8). CATS then groups together identical representative vectors, which could result in a large number of small clusters, as seen with the RMSD-based algorithm; however, this makes it difficult to analyze/identify the types of important structures sampled by the simulations. To address this issue, we added a centroid-based clustering step at the end of our implementation of CATS, which we refer to as "CATS+", to group similar clusters. This also makes for a more fair comparison between the CATS+ and StELa results. Our version of CATS was written in Python, while the original CATS algorithm used a combination of MATLAB, C++, and TCL.

Clustering Protein Fragments with StELa. Inspired by the use of φ and ψ dihedral backbone angles as a CV for describing the structure before determining similarity, we developed our own in-house algorithm, StELa. The first version of this algorithm was described in our previous work where we used StELa to characterize the conformations of the HBD tip in katanin.[24] StELa is a double-clustering algorithm, written in Python, which uses libraries such as SciPy and sklearn.[65,66] Similar to CATS, StELa first characterizes the input structures per sample frame by using the calculated φ/ψ backbone torsion angles extracted from the MD trajectories. This set of angles is the ideal descriptive reaction coordinate for characterizing types of protein structures because it defines the geometry observed in specific secondary structure motifs.[37,67,68] These backbone angles are plotted together on the Ramachandran plot, a classic tool for describing the secondary structure. It has been well established that residues in secondary structures populate specific regions of this plot, in particular α helices and β strands.[63,69] We take advantage of this finding by applying the first round of clustering directly to the φ/ψ dimensions with the centroid based k-means algorithm. In doing so, we determine the number of clusters based on whether or not they split these well-defined regions associated with α helices and β strands/sheets, as described in Figure S9. We then convert each structure from a given sampling of frames from MD into a representative vector composed of cluster labels describing each region of the Ramachandran plot. In doing so, we describe the local structure of a residue in the protein fragment with a single integer. In the previously reported version of our algorithm, we then performed an automated biochemical correction of the vector geometries to ensure that an α helix cannot be shorter than four consecutive residues with at least two consecutive residues between separated helices.[67] In addition, we now apply a similar check to the β sheet region of the plot, such that β strands cannot be shorter than three consecutive residues.[68] To this corrected set of representative vectors, we then apply complete linkage agglomerative hierarchical clustering with the Euclidean distance metric as our second clustering step. In complete linkage agglomerative clustering, structures are sequentially grouped into larger clusters based on the maximum distance between elements. The number of clusters was chosen using statistical-based measurements called the silhouette score, calculated with eq S2, and the Calinski−Harabasz index, calculated with eq S3.[65,70−72] For representation of the resulting clusters in the FEL space, we chose the representative structure by calculating the most probable vector per position of the vectors for a given cluster and then identifying a matching frame.

Evaluation of the Clustering Methods. Testing the performance of the three methods on each of the protein systems requires the use of a common measure. As previously discussed, clustering of frames from MD simulations is employed to extract the most dominant (largest Boltzmann weight) states, with the goal of providing a more detailed view of the conformational landscape. Therefore, the natural measure of performance of a clustering approach is how well the representative structures for each cluster describe the respective protein FEL.[21] Previous work[37] expressed the FEL in the radius of gyration (R_{gyr}) and the end-to-end distance (R_{EE}) space. This choice, depicted in Figure S10a,c for the R2/Tau and the HBD tip in a protomer, generally resulted in single minima representations and the lumping of unique structures, which indicates that R_{gyr} and R_{EE} are not good choices of reaction coordinates. The customarily used representation of the FEL in the literature is in the PC space. For example, the representation of the FEL in the principal components 1 (PC1) and 2 (PC2) space has been used in the past couple of years for various systems from globular proteins to IDPs (including tau).[38,39,56,73−77] Therefore, to evaluate the performance of the clustering methods, we chose to plot the FELs in the (PC1, PC2) space. We evaluated the PCs based on the C-α atom of each residue using the GROMACS[42] command **gmx anaeig**. The first two PCs are representative of the most significant motions from a given data set and are found to be distinct. The variance covered by the first three PCs for each system, reported in Tables S6 and S8, varied between 40% (for the R2/Tau) and 74% (for the HBD tip). The explained variance of the first two PCs for R2/Tau covered around 30%, for R4−R′/Tau covered over 40%,

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

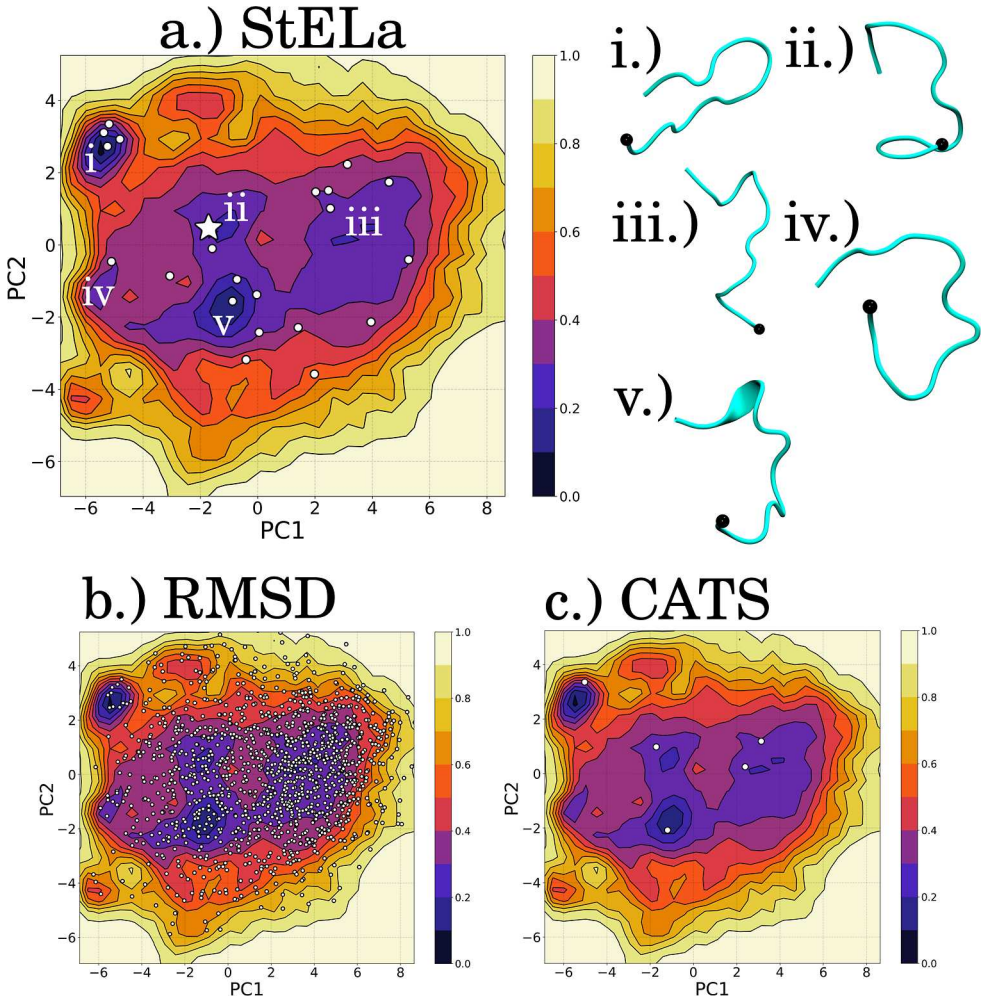


Figure 1. Centers of the identified clusters for the R2/Tau in water plotted on the FEL in the (PC1, PC2) space for each method: (a) StELa, (b) RMSD, and (c) CATS+. The white star in (a) indicates the structure at the beginning of the simulation. Representative structures (i−v) are for the minima indicated in (a). The N-terminal end of the structure is indicated with a black bead.

and for the HBD tip of katanin, it covered over 50%. To determine whether the first two PCs are sufficient for describing the FEL of tau, we plotted the FEL of tau fragments projected onto the first three PC spaces (see Figure S11). These plots show that the addition of PC3 (covering 7.0 and 8.5% of the variance for the R2/Tau and R4−R′/Tau systems, respectively) does not describe any new minima compared to the projection onto the (PC1, PC2) space. Importantly, the FEL in the (PC1, PC2) space, for example, in Figure S10b,d, showed additional minima and regions between them compared to the (R_{gyr}, R_{EE}) space. The FELs in the (PC1, PC2) space for all the systems tested in this work are shown in Figures 1−4 and S14−S23. By plotting the cluster centers or representatives resulting from each method on the corresponding FEL, we can compare the degree to which a method captures the conformational space. It is important to note that we do not apply clustering to the FEL as done in other studies.[31,73] The evaluation of how well a clustering method performs is based on how well the resulting cluster centers cover the FEL minima and the states around and between minima, without populating the entire space, i.e., while still differentiating between the input conformations.

RESULTS

StELa Selects an Optimal Number of Clusters by Using Statistical Checkpoints in a Second Clustering Step to Determine the Number of Unique Structures. One of the main challenges encountered when using clustering algorithms, which stems from their unsupervised learning origin, is the decision on the number of selected clusters. For example, in the case of secondary structures, this means how many unique structures are found in the data set. In our StELa algorithm, this challenge appears in the second step, when using the complete linkage hierarchical clustering. To address this challenge, we relied on statistical-based measures such as the silhouette score and the Calinski−Harabasz index which yield a score versus the number of clusters.[65,70−72] A maximum (usually local) in either of these two scores signals the number that results in the best separation of the clusters, thus indicating the optimal number of clusters. More often than not, we found the two scoring methods to be in agreement in their selection of the number of clusters, as shown in Figure S12. An additional checkpoint consists of the analysis of the corresponding dendrogram, an example of which is shown in Figure S13, to determine if the breakdown of the selected number of clusters agrees with the organization of the dendrogram.[65] The final checkpoint consists of the analysis

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

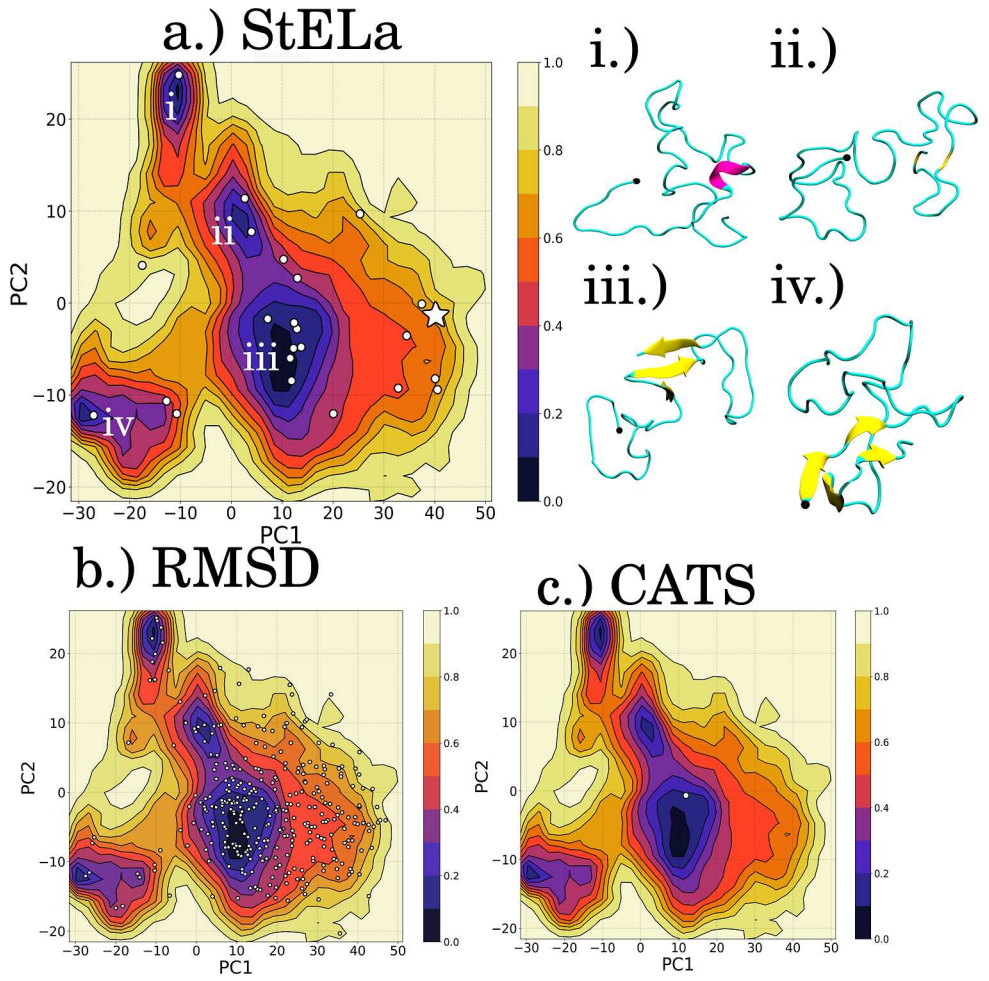


Figure 2. Centers of the identified clusters for the R4−R′/Tau in water plotted on the FEL in (PC1, PC2) space for each method: (a) StELa, (b) RMSD, and (c) CATS+. The white star in (a) indicates the structure at the beginning of the simulation. Representative structures (i−iv) are for the minima indicated in (a). The N-terminal end of the structure is indicated with a black bead.

of each of the resulting clusters to ensure that the selected number of clusters adequately separates unique structures. The number of clusters identified by each algorithm for the various systems probed is reported in Tables S7−S10. Across all of the katanin protomers included in this study, we found the number of clusters to be around 13 for the HBD tip. In turn, for the R2/Tau, we found around 20 clusters, and for R4−R′/Tau, we needed 23 clusters to describe the respective conformational spaces. Our results show that, in general, the more disordered the fragment, the higher the number of clusters that get selected by StELa. Importantly, for all of the systems, the number of clusters is small enough to make it manageable for future analysis. In turn, the RMSD-based algorithm for katanin identified around 30 clusters for each of the monomers, around 20 clusters for protomers A and B, regardless of the presence of the cofactors, and around 75 clusters for protomer C. For R2/Tau, under each of the solvent conditions, the RMSD-based algorithm identified over 1500 clusters. Finally, the CATS+ algorithm selected fewer than 15 clusters across all of the systems probed in our study. Next, we discuss in detail the results for each protein system probed with the three clustering algorithms.

Selection of Representative States from Small Intrinsically Disordered Protein Fragments: R2/Tau. Previous work on the R2 fragment of tau focused on studying the effects of different solvents on the conformational space of this peptide, which can influence its aggregation propensity.[20,37] The FELs in the PC space, shown in Figures 1, S14, and S15, provide insight into how the different solvents affect the conformational landscapes. We found that all of the landscapes are generally characterized by one main basin with multiple minima separated by low energy barriers. In water, (a) three main minima were identified, separated by comparatively higher barriers than those found for the TMAO (b) or urea (c) solvents. R2/Tau in TMAO, the solvent used to induce aggregation, sampled three distinct minima. In contrast, R2/Tau in urea, the solvent used to prevent the formation of secondary structure, was largely confined to a single minimum.

The cluster centers or representatives found with each clustering method for the R2/Tau fragment in water are listed in Figure 1. The structures corresponding to the five potential minima of interest from the FEL, indicated in panel (a), are shown in the, respective, (i−v) snapshots.

The FEL plot shows two deeper minima, one in the upper region (i), which corresponds to a bent loop, and one toward the middle region (v), which corresponds to a loop with some helical characteristics. The RMSD-based algorithm, depicted in panel (b), struggled to find any similarity between the probed loop structures, resulting in a large number of clusters that cover most of the space.[37] The CATS+ algorithm from panel

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

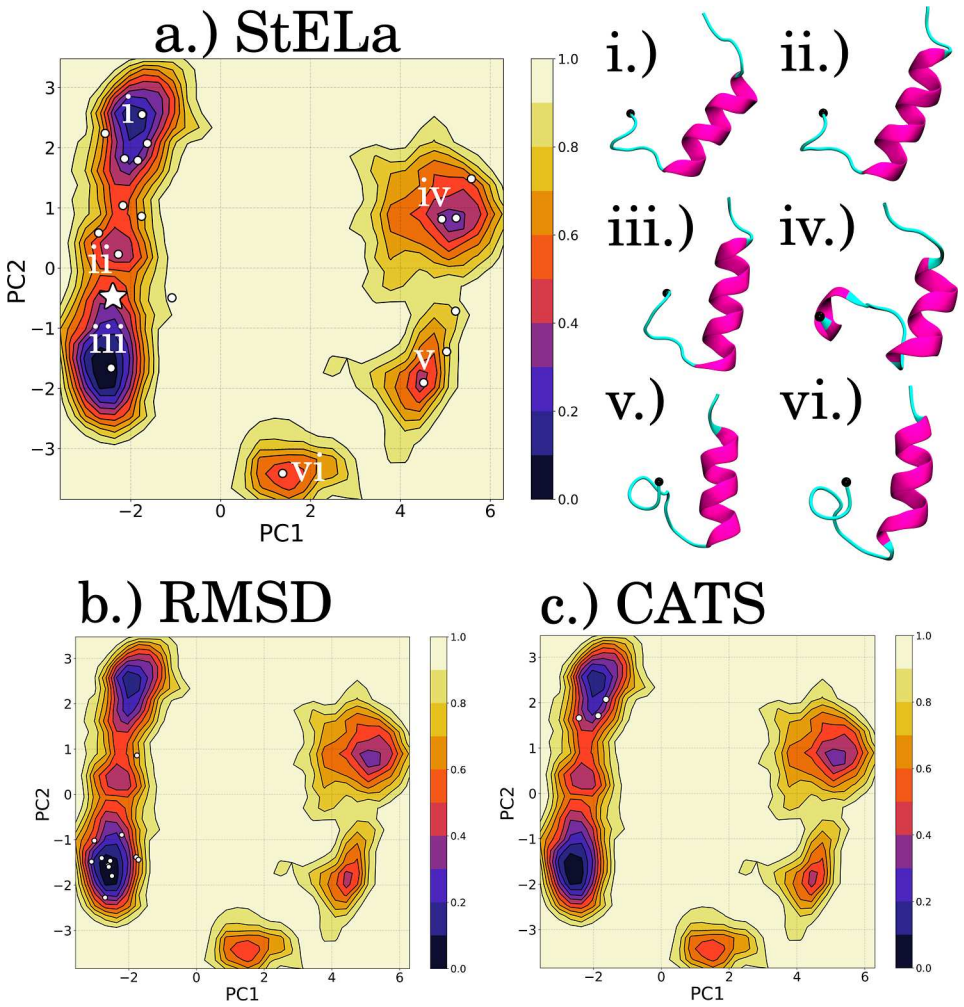


Figure 3. Centers of the identified clusters for the HBD tip of the monomer COMPLEX setup plotted on the FEL in (PC1, PC2) space for each method: (a) StELa, (b) RMSD, and (c) CATS+. The white star in (a) indicates the structure at the beginning of the simulation. Representative structures (i–vi) are for the minima indicated in (a). The N-terminal end of the structure is indicated with a black bead.

(c) identified all but one minimum; however, the resulting populations, shown in Table 1, used to represent the two deeper minima are <1% for (i) and (v), which is unexpected considering their low energy, thus their high Boltzmann weight. The majority of the population is located in the clusters from region (iii), with the remaining population contributing primarily to region (ii). In addition, CATS+ did not identify any states between the minima.

The StELa clusters identified each of the five regions and with populations closer to the expected range, such that the majority of the structures populated regions (i) and (ii). The remaining population was located in clusters representing structures from higher-energy states, located around and between the minima. The above trends found in the RMSD-based and CATS+ algorithms persisted across the different solvent systems, as shown in Figure S14, for TMAO and in Figure S15 for urea. Interestingly, the FEL from the TMAO setup showed that region (i) corresponds to a folded β-sheet. CATS+ did identify a cluster near this area characterized by the folded β-sheet conformation; however, the population was only 0.3%, whereas StELa found it to be 15%. The most striking difference between the clustering for R2/Tau according to the three methods was in the populations of the clusters, most notably of the largest cluster, as detailed in Tables S9–S11. The population of the largest cluster for the RMSD-based algorithm was 4%, that for CATS+ was 85%, and that for StELa was 28%. This cluster for the RMSD-based and StELa algorithms is located in region (ii), while for CATS+, it is found in region (iii). The very large population of the largest cluster for CATS+ is indicative of a clustering algorithm that would result in a distorted conformational space, centered on one deep minimum, while most of the space is left unrepresented. These results illustrate the difficulty in grouping together structures from the MD data sets.

Selection of Representative States from Large Intrinsically Disordered Protein Fragments: R4–R′/Tau. We tested the performance of the clustering algorithms when applied to a longer IDP: a 121-residue long portion of Tau, consisting of the fourth repeat in the MT-binding domain and the adjacent pseudorepeat located toward the N-terminal. The FEL in the (PC1, PC2) space (Figure 2) shows four minima (i–iv), with (ii) and (iii) located in a shared basin and separated by a relatively low energy barrier. Minima (i) and (iv) are separated from this basin and each other by larger barriers. This is notable compared to the shorter R2/Tau system, whose FEL exhibited only low energy barriers between the minima. Figure 2 depicts the cluster centers from each of the three algorithms plotted on the FEL. The populations of

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470–482

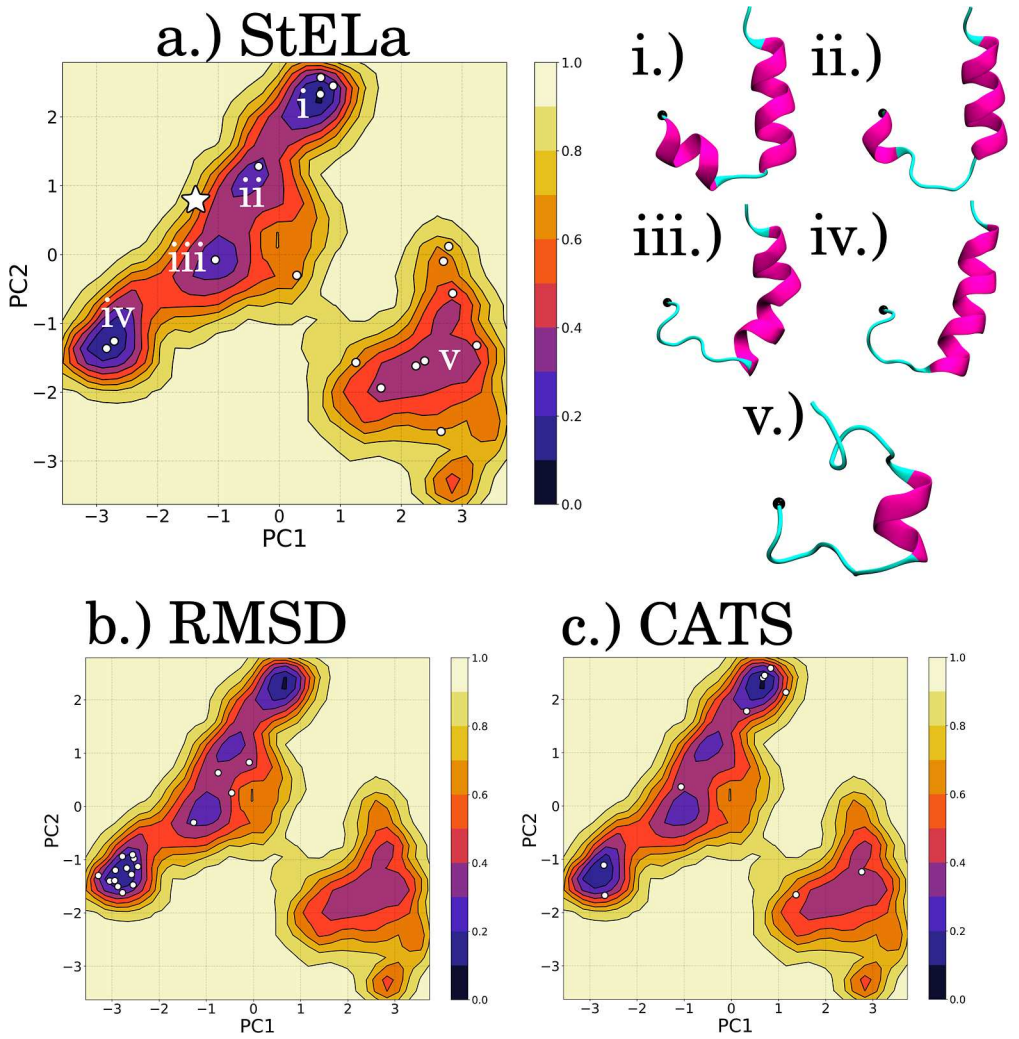


Figure 4. Centers for the identified clusters for the HBD tip of protomer B from the ring−ABC−APO setup plotted on the FEL in PC1/PC2 space for each method: (a) StELa, (b) RMSD, and (c) CATS+. The white star in (a) indicates where the structure is at the beginning of the simulation. Representative structures (i−v) are for the minima indicated in (a). The N-terminal end of the structure is indicated with a black bead.

Table 1. Collective Populations for the Clusters Found in Each Minimum from Figure 1

FEL region	CATS+ (%)	StELa (%)
i	0.2	20
ii	5	28
iii	95	9
iv	-	0.8
v	0.3	12

the top five largest clusters for each of the three methods are given in Table S12. This clearly shows that CATS+ finds only a single cluster, with a representative frame in minimum (iii). The RMSD-based algorithm identifies over 300 clusters, with the centers being distributed across most of the FEL. The largest cluster has a population of 18%. In contrast, StELa is the only algorithm able to successfully locate a reasonable amount (23) of well-sized clusters that are spread across all four minima and nearby states.

States from a Tertiary Protein Fragment: the HBD Tip of the Katanin Monomer. In a previous study, we determined the effects of binding the associated cofactors, ATP and the MT minimal substrate, on the structural stability of lower-order oligomers of katanin.[24] The HBD tip fragment, taken from the tertiary structure of the katanin monomer as described in the Methods, has well-defined FELs in the various setups (Figures 3 and S16−S18). The fragment from the monomer in the NUCLEOTIDE setup (b) resulted in relatively more shallow energy barriers in comparison to the other setups but still high in comparison to the barriers found in the R2/Tau FEL. The FEL for the COMPLEX (a) and SUBSTRATE (c) setups resulted in minima separated by high and wide energy barriers along the PC1. We noticed that the FELs for these katanin monomers have a similar number of well-defined regions, with the change in energy barriers reflecting their dependence on the presence of the cofactors.

The FEL for the COMPLEX setup of the katanin monomer has six minima, as indicated in Figure 3a, with the representative structures plotted in (i−vi). The deepest minima are (i) and (iii), which are characterized by structures differentiated by the length of the helix and the orientation of the loop region. The more shallow region (ii) corresponds to the C-terminal helix being similar in length to that in region (iii) but with the orientation of the N-terminal loop being more similar to the one found in region (i). Thus, this region of the FEL could correspond to a potential intermediate state

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

between (i) and (iii). There is a large and wide barrier described by PC1 to another group of minima in the landscape where structures (iv−vi) are found. The structure in region (iv) corresponds to a state where the N-terminal loop transitioned into a 3−10 helix, according to the STRIDE assignment with VMD.[78] The structures that represent regions (v) and (vi) correspond to loop regions and a helix of varying length. For this well-defined tertiary structure of the fragment from the katanin monomer, the RMSD-based algorithm in (b) finds a smaller number of clusters in comparison to those found for the tau fragments (Tables S9−S12), but the corresponding cluster centers populate only the (iii) region. CATS+, shown in (c), only identifies three clusters, all of which are localized in region (i). In contrast, StELa, shown in (a), yields clusters whose centers cover all the minima and the states between them, including the potentially intermediate from region (ii). The largest cluster for the RMSD-based algorithm was 58% and corresponded to region (iii), for CATS +, it was 56% and corresponded to region (i), and for StELa, it was 30% and corresponded to region (iii). For the other setups of the katanin monomer, shown in Figures S20−S22, similar sets of representative structures were identified, corresponding to the various minima. The RMSD-based algorithm sampled two of the minima found for the NUCLEOTIDE setup, one for the SUBSTRATE setup, and four for the APO setup. CATS + identified three minima for the NUCLEOTIDE setup, two for the SUBSTRATE setup, and two for the APO setup. StELa did fail to characterize one minimum in the APO setup for the katanin monomer that was captured with the RMSD-based algorithm; however, the RMSD-based algorithm failed to represent three other well-defined areas as well.

Selection of Representative States from a Protein Fragment in Quaternary Assemblies: The HBD Tip of Each Protomer from the Katanin Ring Trimer. In our previous study, we characterized the effects of the interprotomer interface formation ($i - 1$, i and $i + 1$, i) on the conformational flexibility and allostery of severing enzyme protomers in dimers and trimers.[24] Our analysis showed that the ring ABC trimer was the least-stable quaternary configuration, characterized by the dissociation of protomer A due to the increase in disorder of its NBD due to the lack of the nucleotide.[5,6] This movement resulted in bending and conformational changes in the HBD tip of protomer B, which in turn allowed it to preserve its contacts with protomer C. The lack of the concave interface in C led to more flexible structures in its HBD tip. The characterization of the HBD tip region in each of the protomers was particularly important for understanding the inner working of the katanin oligomers: the persistence of contacts between the HBD tip fragment of protomer i and the NBD of protomer $i + 1$ is essential for the stability of any quaternary assembly. It thus comes as no surprise that the FELs for each protomer of the ring ABC trimer in the COMPLEX and APO setups are distinct, depending on the bound cofactors and the presence of different types of interfaces (Figures 4 and S19−S23). Moreover, these profiles are also distinct from the corresponding FELs of the katanin monomer, which supports the idea that the HBD tip experiences conformational changes due to the presence of the specific interprotomer interfaces of the quaternary ensembles.

The FEL for protomer B in the APO setup, from Figure 4, is particularly interesting due to the unique structures corresponding to a bent helix in the HBD tip. In this FEL, region (iv) contains a structure resembling the starting structure, and region (i) corresponds to the formation of a helix at the N-terminal end of the fragment. Regions (ii) and (iii) are separated by low-energy barriers from regions (i) and (iv). The structure associated with region (ii) has a single turn helix at the N-terminal end and a helix at the C-terminal end. The structure associated with region (iii) has a loop at the N-terminal end and a dramatically bent helix at the C-terminal end. As pointed out above, based on the putative intermediate state in the COMPLEX setup of the monomer, these regions likely correspond to two intermediate states between the loop−helix structure and the helix−helix structure. Region (v) is broad but relatively shallow, and it is separated by a high energy barrier from the other regions. This region corresponds to a flexible structure with a single-turn helix in the middle of the fragment. The RMSD-based algorithm (b) identified regions (iv) and (iii), while CATS+ (c) identified clusters from (i), (iv) and (v) but did not identify the described intermediate regions (ii) and (iii). In contrast, StELa (a) identified each indicated region. The results obtained using StELa were key in determining that the unique behavior of the conformational space of protomer B is due to the persistance of its contacts with the NBD of protomer C.[24] In the COMPLEX state, shown in Figure S21, the FEL for protomer B presents five regions of interest. Regions (i) and (ii) are close to each other in space and configurations, being separated by medium energy barriers and corresponding to various N-terminal loop orientations and C-terminal helix lengths. Region (v) represents a similar C-terminal helix, but it is higher in energy due to a tight turn and coiled loop arrangement. Region (iv) corresponds to a bent C-terminal helix, similar to that found in the APO setup. StELa identified clusters corresponding to each of these regions, while the RMSD-based algorithm only identified two and CATS+ identified three regions. In summary, the analysis of the FELs in the (PC1, PC2) space for the HBD tip of the lower-order katanin oligomers showed that only the use of a clustering algorithm such as StELa, which employs an accurate representation of the unique types of secondary structures adopted by protein fragments, allows for the identification of all of the important energy states. This in turn provided useful insight into the influence of the cofactors and of the interfaces for the structural and functional behavior of katanin. A full description and comparison of the results for each of the protomers in the ring trimer in the presence and absence of the binding cofactors can be found in the Supporting Information.

DISCUSSION AND CONCLUSIONS

Unsupervised Machine Learning Offers Powerful Tools for Describing and Extracting Representative Protein Structures. Computational tools such as MD simulations allow us to access the conformational landscape of proteins in atomistic detail under dynamic equilibrium conditions.[14] The challenge of identifying the rich structural variety from MD data is due to the sheer size of the system, which makes the identification of the number and the identity of the unique states a very challenging problem. Because usually there is little or no prior knowledge regarding the number of the representative states, unsupervised learning methods such as clustering are the only types of machine learning tools that can be employed to identify subgroups from a data set.[79] Effective clustering approaches, however, are highly dependent on whether the intention is to study more

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

global structural changes of larger systems or focus more on local shifts such as with IDPs, where the vast array of sampled conformations require more detailed CVs.[33,80] Characterizing these disordered or flexible regions, often associated with functional regions, allows for the description of changes to the conformational landscape in different environments or due to ligand binding, which is crucial to understanding enzymatic mechanisms. Accessing these states is a starting point for identifying functionally relevant conformations, determining structures appropriate for docking, or for accelerated simulations to further interrogate the conformational space of the protein of choice.[81] Additional challenges associated with clustering are the determination of the number of clusters/states. Importantly, it is also possible to think of approaches that perform a clustering of the initial set of clusters to further reduce the data complexity and collectively describe the structures from a simulation.[82]

RMSD-Based Clustering Captures Large Global Changes but Struggles to Determine Similarity for IDPs and Secondary Structures in Parts of Globular Proteins. We found that the standard RMSD-based algorithm favors one large primary cluster that usually gravitates toward a folded native structure, which can be particularly useful for well-defined protein structures that generally maintain that same shape, as observed for many tertiary structures of globular proteins.[2,3,14] This is particularly useful when the goal is to describe global tertiary or quaternary sampled states, as found in previous studies.[14,82] Unlike globular proteins, IDPs are characterized by broad conformational landscapes and shallow energy barriers. As a result, they tend to populate a rich variety of structures, which makes it difficult for algorithms to appropriately identify the similarity between states. The RMSD-based algorithm, which determines similarity based on an RMSD cutoff, struggled to describe short or long IDP fragments, as it produced a large number of small clusters.[37] This is further demonstrated by the significantly overlapping RMSD distributions that we found for R2/Tau (Figure S24). The RMSD algorithm has better performance when applied to the HBD tip region from the tertiary and quaternary protein structures and results in similar populations for the largest clusters, as listed in Tables S13−S18; however, it only characterized one or two of the identified minima in each system, which were usually the lowest in energy. While these minima described states considered to be the most populated, it failed to describe the full conformational space as it missed or over-represented key regions of interest. The minima that corresponded to the helix−helix conformation, that we previously associated with ligand binding, were not well represented by the RMSD-based algorithm. For example, this state was missing in the trimer-A-COMPLEX setup (Figure S19b) and was over-represented in the monomer-SUBSTRATE setup (Figure S17b). In conclusion, the RMSD-based clustering does not reflect the atomistic detail required for determining similarity in IDP states, as previously reported,[37] or for characterizing local secondary structures in folded proteins, as seen in the monomers and trimers of katanin.

Reducing the Backbone Torsion Angles to a Single Dimension Allows for Better Structural Characterization Prior to Clustering of Protein Structures. To address the challenge of finding a more useful similarity metric, Cheung and Ezerski created the CATS algorithm, based on the φ and ψ backbone torsion angles, which is a well-known and detailed internal descriptor for the secondary structures of proteins.[37] Prior to determining similarity, the user describes the structure with a chosen number of labels from Gaussian distributions of the φ angle and then the ψ angle. The algorithm then defines similarity based on whether or not two vectors are found to be identical, which takes additional time.[37] The longer the protein, the longer the time it takes to make these decisions. One of the drawbacks they observed for the clustering of R2/Tau was that, similar to the RMSD-based algorithm, CATS identifies a large number of small clusters. When an additional clustering step, using *k*-means, is applied to the resulting clusters, CATS+ identifies no more than 10 similar clusters for the tau data sets. Even more striking is the fact that CATS+ only identified one cluster for the longer tau fragment (R4−R′) (Figure 2c). Similar to the RMSD-based algorithm, we found that the clusters extracted with CATS+ resulted in a poor representation of the FEL regions and, at times, did not identify regions of the landscape at all, as seen in the R2/Tau TMAO setup (Figure S14c). When clustering the katanin data sets, CATS+ provided a relatively reasonable set of clusters compared to the tau results, but it still struggled with misrepresenting and identifying the minima and states around them, as observed in more challenging landscapes from the COMPLEX setup of the monomer (Figure 3c) of protomer B in the APO setup (Figure 4c) and of protomer C in the COMPLEX and APO setups (Figures S22c and S23c). Additionally, one of the reported limitations of CATS was its ability to characterize the regions associated with β strands and sheets that are found in abundance in the R2/Tau simulations.[37] The Ramachandran plots for R2/Tau had a relative higher density in the β region in comparison to the HBD tip of Katanin, as shown in Figure S25, signaling the importance of finding a better way to describe β strands and sheets than the approach employed by CATS. By considering the φ and ψ angles together and reducing the two dimensions into one using centroid cluster labels, our StELa approach was able to overcome this shortcoming of CATS in characterizing all the major protein secondary structures.

StELa Identifies Unique Regions from the Free Energy Landscape for IDPs and Secondary Structures. Characterizing the protein fragments prior to clustering, as done by Cheung and Ezerski, opens up exciting doors for probing local conformational transitions.[37] Inspired by CATS, StELa uses the Ramachandran plot more holistically to describe proteins and defined helices as at least four helical angles and a β strand as at least three angles in the β region. StELa then determined the number of states for each setup using statistical scoring methods and applied hierarchical clustering directly to these enhanced descriptive vectors. In our algorithm, we employed the maximum Euclidean distance between vectors to signal similarity (complete linkage). The KMeans and hierarchical clustering functions from sklearn and SciPy are quite efficient at handling the data set.[65,66] The resulting clusters from StELa were striking. The clusters for the tau data sets more appropriately represented and identified unique regions from the FEL in each of the solvent environments. This was particularly true with regard to the significantly longer R4−R′ fragment. Similarly, the clusters identified from the katanin data sets sampled the landscape well, only missing a minimum of interest in the monomer-APO setup (Figure S18a) and the trimer−C−COMPLEX setup (Figure S22a), although neither the RMSD-based nor CATS+ captured these two minima. For each system, the cluster

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

centers from CATS+ oftentimes were found to occupy main minima. The centers from RMSD-based clustering were spread out for the IDP systems, identifying high- and low-energy states. However, for the katanin systems, the centers generally fell into main minima for the katanin systems, often ignoring other regions identified as significant by the FEL. In contrast, the cluster centers from StELa identified the majority of the minima, as well as the higher-energy regions between the minima for both the IDP and katanin systems.

In our previous study, we used StELa to characterize the observed changes in the HBD tip for various lower-order oligomers of katanin.[24] The results of this analysis identified a state that resembled the structure described by cryo-EM (loop−helix) and a state that was unique to the binding ligands in the monomer (helix−helix). In the dimers and trimers, we identified additional more flexible and potential intermediate states that were unique to the protein−protein interactions of the protomers. This analysis was the key to understanding the stability and allostery of these assemblies as well as the overall hexamer. This would have been a challenge, or even impossible, to characterize and understand using the RMSD-based algorithm or CATS, as the resulting cluster centers do not cover the energy landscape well, being unable to identify all of the minima and the states between them.

Dihedral Angle-Based Clustering Algorithms Are Challenged by the Presence of Substantial β-Sheet Structures. The ability of StELa to correctly cluster structures depends on how well the initial centroid-based clustering step separates the regions in the Ramachandran plot, which we found to generally perform better in katanin than in tau. Notably, the fewer centroids used, the less distinct the resulting vectors are going to be. Furthermore, similar to CATS, StELa found it challenging to characterize the β region. The β region of the Ramachandran plot includes straight β strands as well as sheets, and StELa has no way of identifying a difference between consecutive straight β strands and the sheet conformation, which is tertiary in nature, although it does a very good job of characterizing and identifying α helices.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c01511.

Additional computational details and methods, additional details of clustering the quaternary assemblies, details of results, convergence tests, choices of cut-offs, and clustering examples (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ruxandra I. Dima − *Department of Chemistry, University of Cincinnati, Cincinnati, Ohio 45221, United States;* orcid.org/0000-0001-6105-7287; Phone: +1 (513) 556-3961; Email: ruxandra.dima@uc.edu; Fax: +1 (513) 556-9239

Authors

Amanda C. Macke − *Department of Chemistry, University of Cincinnati, Cincinnati, Ohio 45221, United States*

Jacob E. Stump − *Department of Chemistry, University of Cincinnati, Cincinnati, Ohio 45221, United States*

Maria S. Kelly − *Department of Chemistry, University of Cincinnati, Cincinnati, Ohio 45221, United States*

Jamie Rowley − *Department of Chemistry, University of Cincinnati, Cincinnati, Ohio 45221, United States*

Vageesha Herath − *Department of Chemistry, University of Cincinnati, Cincinnati, Ohio 45221, United States; Department of Chemistry, Emory University, Atlanta, Georgia 30322, United States*

Sarah Mullen − *Department of Chemistry, The College of Wooster, Wooster, Ohio 44691, United States; Department of Chemistry, Virginia Tech, Blacksburg, Virginia 24061, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01511

Notes

The authors declare no competing financial interest.

The MD trajectory input files for the various systems simulated by the Dima group and tested in this work are available on our GitHub page. The code for StELa can be found on our GitHub page at https://github.com/DimaUClab/StELa-Protein-Structure-Clustering-Algorithm.

■ ACKNOWLEDGMENTS

We thank Joan-Emma Shea and Margaret Cheung for providing the MD simulations files for TAU/R2 as well as Rohith Anand Varikoti for the MD simulation files for the katanin monomers. This research was funded by the National Science Foundation (NSF) MCB-1817948 (to RID). S.M. was supported through the NSF Research Experience for Undergraduates in Chemistry grant CHE-1950244. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) through allocation TG-BIO210094 to R.I.D.

■ REFERENCES

(1) Ahmad, E.; Rabbani, G.; Zaidi, N.; Khan, M. A.; Qadeer, A.; Ishtikhar, M.; Singh, S.; Khan, R. H. Revisiting ligand-induced conformational changes in proteins: essence, advancements, implications and future challenges. *J. Biomol. Struct. Dyn.* **2013**, *31*, 630−648.

(2) Bailey, M. E.; Jiang, N.; Dima, R. I.; Ross, J. L. Invited review: Microtubule severing enzymes couple atpase activity with tubulin GTPase spring loading. *Biopolymers* **2016**, *105*, 547−556.

(3) Alushin, G. M.; Lander, G. C.; Kellogg, E. H.; Zhang, R.; Baker, D.; Nogales, E. High-Resolution Microtubule Structures Reveal the Structural Transitions in $\alpha\beta$-Tubulin upon GTP Hydrolysis. *Cell* **2014**, *157*, 1117−1129.

(4) Zhang, R.; Alushin, G. M.; Brown, A.; Nogales, E. Mechanistic Origin of Microtubule Dynamic Instability and Its Modulation by EB Proteins. *Cell* **2015**, *162*, 849−859.

(5) Zehr, E. A.; Szyk, A.; Piszczek, G.; Szczesna, E.; Zuo, X.; Roll-Mecak, A. Katanin spiral and ring structures shed light on power stroke for microtubule severing. *Nat. Struct. Mol. Biol.* **2017**, *24*, 717−725.

(6) Zehr, E. A.; Szyk, A.; Szczesna, E.; Roll-Mecak, A. Katanin Grips the β-Tubulin Tail through an Electropositive Double Spiral to Sever Microtubules. *Dev. Cell* **2020**, *52*, 118−131.e6.

(7) Russell, R. B.; Alber, F.; Aloy, P.; Davis, F. P.; Korkin, D.; Pichaud, M.; Topf, M.; Sali, A. A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.* **2004**, *14*, 313−324.

(8) McGuffin, L. J.; Bryson, K.; Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **2000**, *16*, 404−405.

(9) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.;

Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(10) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439−D444.
(11) Ruff, K. M.; Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433*, 167208.
(12) Miskei, M.; Horvath, A.; Vendruscolo, M.; Fuxreiter, M. Sequence-Based Prediction of Fuzzy Protein Interactions. *J. Mol. Biol.* **2020**, *432*, 2289−2303.
(13) Receveur-Brechot, V.; Durand, D. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr. Protein Pept. Sci.* **2012**, *13*, 55−75.
(14) Daura, X.; van Gunsteren, W. F.; Mark, A. E. Folding-Unfolding Thermodynamics of a B-Heptapeptide From Equilibrium Simulations. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 269−280.
(15) Wright, P. E.; Dyson, H. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321−331.
(16) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114*, 6589−6631.
(17) Levine, Z. A.; Shea, J.-E. Simulations of disordered proteins and systems with conformational heterogeneity. *Curr. Opin. Struct. Biol.* **2017**, *43*, 95−103.
(18) Venkatramani, A.; Panda, D. Regulation of neuronal microtubule dynamics by tau: Implications for tauopathies. *Int. J. Biol. Macromol.* **2019**, *133*, 473−483.
(19) Eschmann, N. A.; Do, T. D.; LaPointe, N. E.; Shea, J.-E.; Feinstein, S. C.; Bowers, M. T.; Han, S. Tau Aggregation Propensity Engrained in its Solution State. *J. Phys. Chem. B* **2015**, *119*, 14421−14432.
(20) Levine, Z. A.; Larini, L.; LaPointe, N. E.; Feinstein, S. C.; Shea, J.-E. Regulation and aggregation of intrinsically disordered peptides. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 2758−2763.
(21) Jahn, T. R.; Radford, S. E. The Yin and Yang of protein folding. *FEBS J.* **2005**, *272*, 5962−5970.
(22) Grundke-Iqbal, I.; Iqbal, K.; Tung, Y.; Quinlan, M.; Wisniewski, H.; Binder, L. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 4913−4917.
(23) Lyons, A. J.; Gandhi, N. S.; Mancera, R. L. Molecular dynamics simulation of the phosphorylation-induced conformational changes of a tau peptide fragment. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 1907−1923.
(24) Macke, A. C.; Kelly, M. S.; Varikoti, R. A.; Mullen, S.; Groves, D.; Forbes, C.; Dima, R. I. Microtubule severing enzymes oligomerization and allostery: a tale of two domains. *J. Phys. Chem. B* **2022**, *126*, 10569−10586.
(25) Nithianantham, S.; McNally, F. J.; Al-Bassam, J. Structural basis for disassembly of Katanin heterododecamers. *J. Biol. Chem.* **2018**, *293*, 10590−10605.
(26) Pfaendtner, J.; Lyman, E.; Pollard, T. D.; Voth, G. A. Structure and Dynamics of the Actin Filament. *J. Mol. Biol.* **2010**, *396*, 252−263.
(27) Pfaendtner, J.; De La Cruz, E.; Voth, G. A. Actin filament remodeling by actin depolymerization factor/cofilin. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 7299−7304.
(28) Mani, S.; Katkar, H. H.; Voth, G. A. Compressive and Tensile Deformations Alter ATP Hydrolysis and Phosphate Release Rates in Actin Filaments. *J. Chem. Theory Comput.* **2021**, *17*, 1900−1919.
(29) Lukin, J. A.; Kontaxis, G.; Simplaceanu, V.; Yuan, Y.; Bax, A.; Ho, C. Quaternary structure of hemoglobin in solution. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 517−520.
(30) Ingram, V. Gene Mutations in Human Hæmoglobin: the Chemical Difference Between Normal and Sickle Cell Hæmoglobin. *Nature* **1957**, *180*, 326−328.
(31) Maity, D.; Pal, D. Molecular Dynamics of Hemoglobin Reveals Structural Alterations and Explains the Interactions Driving Sickle Cell Fibrillation. *J. Phys. Chem. B* **2021**, *125*, 9921−9933.
(32) Henry, E. R.; Bettati, S.; Hofrichter, J.; Eaton, W. A. A tertiary two-state allosteric model for hemoglobin. *Biophys. Chem.* **2002**, *98*, 149−164.
(33) de Souza, V. C.; Goliatt, L.; Capriles Goliatt, P. V. Z. Clustering algorithms applied on analysis of protein molecular dynamics. *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*; IEEE, 2017; Vol. *1*−*6*.
(34) Peng, J.-h.; Wang, W.; Yu, Y.-q.; Gu, H.-l.; Huang, X. Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems. *Chin. J. Chem. Phys.* **2018**, *31*, 404−420.
(35) Kufareva, I.; Abagyan, R. Methods of protein structure comparison. *Methods Mol. Biol.* **2012**, *857*, 231−257.
(36) De Paris, R.; Quevedo, C. V.; Ruiz, D. D. A.; Norberto de Souza, O. An Effective Approach for Clustering InhA Molecular Dynamics Trajectory Using Substrate-Binding Cavity Features. *PLoS One* **2015**, *10*, No. e0133172.
(37) Ezerski, J. C.; Cheung, M. S. CATS: A Tool for Clustering the Ensemble of Intrinsically Disordered Peptides on a Flat Energy Landscape. *J. Phys. Chem. B* **2018**, *122*, 11807−11816.
(38) Yuan, Y.; Deng, J.; Cui, Q. Molecular Dynamics Simulations Establish the Molecular Basis for the Broad Allostery Hotspot Distributions in the Tetracycline Repressor. *J. Am. Chem. Soc.* **2022**, *144*, 10870−10887.
(39) Janson, G.; Valdes-Garcia, G.; Heo, L.; Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **2023**, *14*, 774.
(40) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS:Fast, flexible and free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(41) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−437.
(42) Abraham, M.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*−*2*, 19−25.
(43) Weerasinghe, S.; Smith, P. E. A Kirkwood-Buff derived force field for sodium chloride in water. *J. Chem. Phys.* **2003**, *119*, 11342−11349.
(44) Larini, L.; Shea, J.-E. Double resolution model for studying TMAO/water effective interactions. *J. Phys. Chem. B* **2013**, *117*, 13268−13277.
(45) Kaminski, G. A.; Stern, H. A.; Berne, B.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. Development of a polarizable force field for proteins via ab initio quantum chemistry: First generation model and gas phase tests. *J. Comput. Chem.* **2002**, *23*, 1515−1531.
(46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(47) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950−1958.

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482

(48) Webb, B.; Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinf.* **2016**, *54*, 5−6.
(49) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **2021**, *30*, 70−82.
(50) Brotzakis, Z. F.; Lindstedt, P. R.; Taylor, R. J.; Rinauro, D. J.; Gallagher, N. C.; Bernardes, G. J.; Vendruscolo, M. A structural ensemble of a tau-microtubule complex reveals regulatory tau phosphorylation and acetylation mechanisms. *ACS Cent. Sci.* **2021**, *7*, 1986−1995.
(51) Berendsen, H. J.; Postma, J. P.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. *Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, israel, April 13−16, 1981*; 1981; pp 331−342.
(52) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98−103.
(53) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(54) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(55) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(56) Damre, M.; Dayananda, A.; Varikoti, R. A.; Stan, G.; Dima, R. I. Factors underlying asymmetric pore dynamics of disaggregase and microtubule-severing AAA+ machines. *Biophys. J.* **2021**, *120*, 3437−3454.
(57) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40*, 843−856.
(58) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces*; Springer Netherlands, 1981; pp 331−342.
(59) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(60) Eswar, N.; Eramian, D.; Webb, B.; Shen, M.-Y.; Sali, A. *Protein Structure Modeling with MODELLER*; Humana Press, 2008.
(61) Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E. An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **2011**, *7*, 4026−4037.
(62) Daura, X.; Conchillo-Solé, O. On Quality Thresholds for the Clustering of Molecular Structures. *J. Chem. Inf. Model.* **2022**, *62*, 5738−5745.
(63) Hovmöller, S.; Zhou, T.; Ohlson, T. Conformations of amino acids in proteins. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 768−776.
(64) Srinivasan, R.; Rose, G. The T-to-R transformation in hemoglobin: a reevaluation. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 11113−11117.
(65) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(66) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, I.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261−272.
(67) Pauling, L.; Corey, R. B.; Branson, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 205−211.
(68) Pauling, L.; Corey, R. B. Two Rippled-Sheet Configurations of Polypeptide Chains, and a Note about the Pleated Sheets. *Proc. Natl. Acad. Sci. U.S.A.* **1953**, *39*, 253−256.
(69) Mannige, R. V.; Kundu, J.; Whitelam, S. The Ramachandran Number: An Order Parameter for Protein Geometry. *PLoS One* **2016**, *11*, No. e0160023.
(70) Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1−27.
(71) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.
(72) Shahapure, K. R., Nicholas, C. Cluster Quality Analysis Using Silhouette Score. *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*; IEEE, 2020; pp 747−748.
(73) Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys.* **2019**, *150*, 150901.
(74) Tribello, G. A.; Gasparotto, P. Using Dimensionality Reduction to Analyze Protein Trajectories. *Front. Mol. Biosci.* **2019**, *6*, 46.
(75) Capelli, R.; Bochicchio, A.; Piccini, G.; Casasnovas, R.; Carloni, P.; Parrinello, M. Chasing the Full Free Energy Landscape of Neuroreceptor/Ligand Unbinding by Metadynamics Simulations. *J. Chem. Theory Comput.* **2019**, *15*, 3354−3361.
(76) Leander, M.; Yuan, Y.; Meger, A.; Cui, Q.; Raman, S. Functional plasticity and evolutionary adaptation of allosteric regulation. *Proc. Natl. Acad. Sci.* **2020**, *117*, 25445−25454.
(77) Krishnan, K.; Kassab, R.; Agajanian, S.; Verkhivker, G. Interpretable Machine Learning Models for Molecular Design of Tyrosine Kinase Inhibitors Using Variational Autoencoders and Perturbation-Based Approach of Chemical Space Exploration. *Int. J. Mol. Sci.* **2022**, *23*, 11262.
(78) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(79) Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O. P.; Tiwari, A.; Er, M. J.; Ding, W.; Lin, C.-T. A Review of Clustering Techniques and Developments. *Neurocomputing* **2017**, *267*, 664−681.
(80) Teletin, M.; Czibula, G.; Albert, S.; Bocicor, I. Using unsupervised learning methods for enhancing protein structure insight. *Procedia Comput. Sci.* **2018**, *126*, 19−28.
(81) Karamzadeh, R.; Karimi-Jafari, M. H.; Sharifi-Zarchi, A.; Chitsaz, H.; Salekdeh, G. H.; Moosavi-Movahedi, A. A. Machine Learning and Network Analysis of Molecular Dynamics Trajectories Reveal Two Chains of Red/Ox-specific Residue Interactions in Human Protein Disulfide Isomerase. *Sci. Rep.* **2017**, *7*, 3666.
(82) Oruganti, B.; Lindahl, E.; Yang, J.; Amiri, W.; Rahimullah, R.; Friedman, R. Allosteric enhancement of the BCR-Abl1 kinase inhibition activity of nilotinib by cobinding of asciminib. *J. Biol. Chem.* **2022**, *298*, 102238.

https://doi.org/10.1021/acs.jcim.3c01511
J. Chem. Inf. Model. 2024, 64, 470−482